

Grounding Language Models with Semantic Digital Twins for Robotic Planning

Mehreen Naeem¹, Andrew Melnik¹, Michael Beetz¹.

¹Institute for Artificial Intelligence, University of Bremen, Germany

Corresponding author: mnaeem@uni-bremen.de

Abstract—We introduce a novel framework that integrates Semantic Digital Twins (SDTs) with Large Language Models (LLMs) to enable adaptive and goal-driven robotic task execution in dynamic environments. The system decomposes natural language instructions into structured action triplets, which are grounded in contextual environmental data provided by the SDT. This semantic grounding allows the robot to interpret object affordances and interaction rules, enabling action planning and real-time adaptability. In case of execution failures, the LLM utilizes error feedback and SDT insights to generate recovery strategies and iteratively revise the action plan. We evaluate our approach using tasks from the ALFRED benchmark, demonstrating robust performance across various household scenarios. The proposed framework effectively combines high-level reasoning with semantic environment understanding, achieving reliable task completion in the face of uncertainty and failure.

I. INTRODUCTION

Exploring dynamic environments has revolutionized how robots interpret and execute complex tasks by leveraging advanced natural language processing capabilities [10]. This allows robots to decompose high-level instructions into actionable steps and adapt dynamically [24], as seen in frameworks like RePLan [23][25], which enables online re-planning in response to unexpected obstacles. However, challenges such as safety concerns and model hallucinations (where models generate plausible but incorrect outputs) persist. To address these, some work incorporates safety checks into the planning process[4][12], while others use scene graphs to ground LLMs in the physical environment[13], enhancing plan feasibility by connecting language models to real-world contexts through sensor data or computer vision.

Traditional Digital Twins [16] primarily focus on geometric and physical representations of the environment, such as objects, meshes, and features, whereas Semantic Digital Twins (SDTs) incorporate real-time data [26]. It mirrors a building's systems and components of DT by adding layer of rich contextual and semantic information about dynamic surroundings by integrating the objects' text descriptions, rules, and interaction properties[9][6]. Various approaches have been used to SDTs with LLM agent in affordance-based scene representations to enable facilitate large-scale task planning in complex environments[28][3].

The fusion of LLM and SDTs offers a promising path to creating robotic systems capable of operating in dynamic

environments. The relevance of this collaboration is to support a robot's ability by effectively adjusting robot actions that rely on real-time environmental feedback[18][25]. However, the main challenge is to align the LLM response with the semantic structures represented in SDTs, ensuring seamless communication between the two systems. Our framework addresses these challenges essential for bridging high-level reasoning with real-world execution in dynamic environments. This paper discusses the necessary preliminaries, including related work on integrating LLMs and SDTs, workflow construction, experimental setup, and evaluation results. Finally, it concludes with potential directions for future research.

II. RELATED WORK

A. Leveraging Semantic Digital Twins for Task Planning

Recently, there have been advancements in how robots plan tasks, especially when they don't have all the information [15, 27, 19]. AutoGPT+P [5] is an example of robots using object affordances and large language models to make plans even when certain objects are missing from the environment. However, this method can be slow because it relies on external LLM models for affordance mapping. In contrast, our approach employs Semantic Digital Twins (SDTs) that encapsulate object-action rules. This allows agents to plan tasks in real time without the need for training or external object reasoner, enhancing responsiveness and reducing reliance on potentially unreliable external data sources.

B. Integrating LLMs and SDTs in Complex Environments

Approaches like SayPlan utilize 3D scene graphs to ground LLM-generated plans in large-scale, multi-room environments[20, 21, 2]. Similarly, the SMART-LLM framework demonstrates using LLMs for multi-agent task planning, translating high-level instructions into coordinated actions[11]. These methods, while scalable, often depend on extensive pre-training and may struggle with real-time adaptability[1]. Our method diverges by using SDTs that inherently understand object-action relationships [17]. These enable agents to plan and adapt in complex environments, facilitating immediate responsiveness to dynamic changes and enhancing operational efficiency.

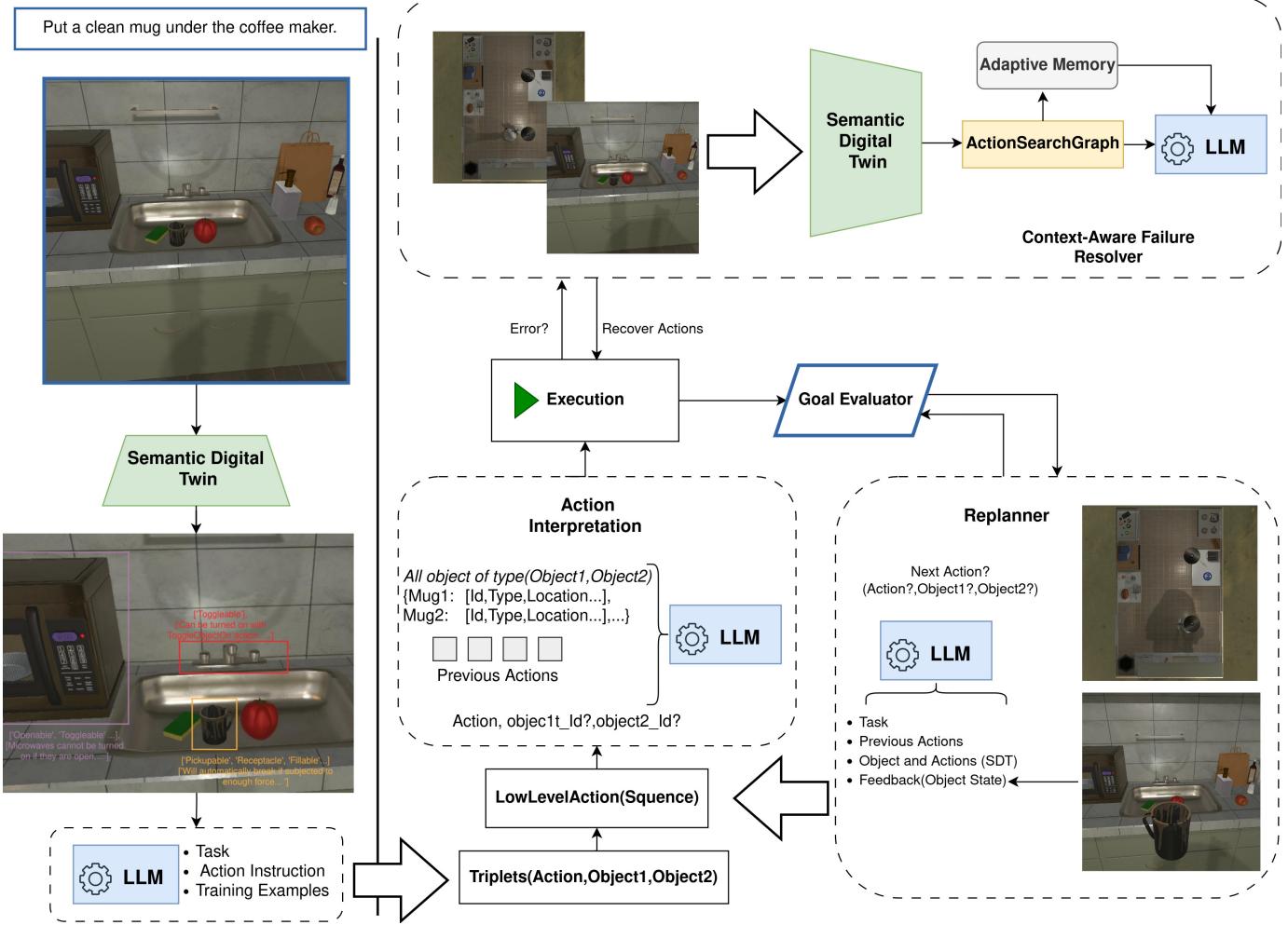


Fig. 1: LLM-based task planning with a Semantic Digital Twin. The SDT provides real-time context, enabling the LLM to generate and adapt action for goal-directed planning and execution.

C. Failure Detection and Recovery

Traditional failure recovery systems in robotics often combine LLMs with SDTs to detect and adapt to errors[12]. For example, REFLECT [14] uses vision-language models (VLMs) [8] conditioned on failure-related queries to generate summaries and re-plan tasks. Recent work such as LoTa-Bench has brought significant attention to evaluating language-based task planners in failure-prone environments [7]. LoTa-Bench provides a benchmark to assess how well LLM-based agents detect, describe, and recover from execution failures. It emphasizes querying large models during execution to reason about error types and generate recovery strategies. While LoTa-Bench has made valuable contributions in standardizing failure scenarios and offering metrics for recovery evaluation, its reliance on external querying and LLM-based introspection introduces latency and potential inconsistency during run-time execution.

III. METHODOLOGY

Enabling LLMs to operate in dynamic surroundings, reacting to changing robotic states and environments without excessive computational latency, our framework leverages a Semantic Digital Twin (SDT) to prompt a Large Language Model (LLM), enabling it to decompose the natural language description of a target task into a sequence of structured action triplets. These triplets act as step-by-step execution guidelines for the robot. Next, the agent grounds each action triplet by selecting the most contextually appropriate parameters to ensure smooth execution of the corresponding trajectory. The agent also handles potential failure cases at this step by leveraging an object-action knowledge semantic map to predict suitable recovery actions.

A. Semantic Digital Twin

Our framework consists of two key Semantic Digital Twin components: “Rules and Interaction Properties” and “Textual Descriptions.” These components represent objects’ responses to specific actions based on the surrounding context and

conditions. Specifically, we utilize a structured set of rules describing object affordances and the consequences of particular interactions (Figure 10,Figure 11). For example: Bottle, “Will break if subjected to sufficient force. Will fill up with water if placed under a running water source.” We also define a structured set of permissible actions. For example, the object “Bottle” is annotated with the action properties [“Pickupable”, “Fillable”, “Breakable”]. These affordances indicate that the bottle can be picked up, filled (e.g., with water), or broken under certain conditions. Performing such a link explicitly to objects with corresponding action capabilities enables the agent to create a concept map of a given task. By grounding the task description with a concept map of SDT using LLMs, we predicted the set of action triplets and later handled the failure cases followed by the replanner. By grounding the task description using a conceptual map derived from the Semantic Digital Twin (SDT) and leveraging the reasoning capabilities of Large Language Models (LLMs), we first predicted a structured set of action triplets for task execution and later handled the failure cases.

B. LLM-Powered Adaptive Planner

To provide Large Language Models (LLMs) with structured input for triplet prediction, we design an LLM-based agent that operates in a multi-step process. First, the agent scans the environment and filters out only the objects that are relevant to the task at hand. Once these task-relevant objects are isolated, the agent queries a SDT, which supplies information about the possible actions that can be performed on or with each object, along with the actionable properties associated with them (e.g., whether an object is fillable, openable, or can be picked up). With this contextual understanding, comprising the filtered set of objects, their actionable properties, and the predefined rules or instructions that govern permissible actions along with previous task examples, the LLM agent is equipped to generate a structured output. Specifically, it predicts a sequence of action triplets, where each triplet *Triplets(HighLevelAction, Object1, Object2)* typically represents an action, the object it applies to, and any associated parameters. The training examples are task-context examples used to construct the prompt, providing the LLM agent with guidance on the expected type of response. Alongside this sequence, the agent also predicts the final goal condition these actions aim to achieve.

After establishing general guidelines for the task in the form of triplets, the next crucial step involves resolving the specific parameters of the predicted action triplets. In this context, resolving parameters means determining the most appropriate objects to associate with each high-level action in the sequence. To accomplish this, the LLM agent performs step-by-step reasoning during action execution in the Action Interpretation Engine. At each step, the agent takes into account three key inputs: (1) the current state SDT of the environment, (2) the task description, and (3) the history of previously executed actions. Using this information, the agent constructs a context-aware query to determine which object in

the environment is the most suitable for the following high-level action in the triplet sequence.

C. Context-Aware Failure Handling and Replanning

If an action fails during execution, the Failure Resolver intervenes and attempts to resolve the issue by predicting alternative actions. It explores the environment and uses Action Search Graph to generate a map of available action pairs as shown in Figure 2 and Figure 3. The Failure Resolver Engine also incorporates adaptive memory, which tracks previously attempted solutions to avoid repeating the same ones for a specific failure point.

The Context-Aware Failure Resolver is designed to analyze errors and suggest appropriate recovery actions to address the failure. Action Search Graph takes a set of object **O**, a list of actions **A**, and filters the actions based on a condition function **Ca**.

$$A' = \{a \in A \mid C_a(O, a) = T\} \quad (1)$$

Here **O** be a list of object descriptions generated from SDT, **A**=[a1,a2,...,an] be a finite set of possible actions. **Ca(O,a)** is a boolean-valued condition function, which returns true if action a is valid for object **O** and false otherwise. The filtered action set **A'** consists of all actions a from the original set **A** for which the condition function **Ca(O,a)** evaluates to true. After generating all possible action pairs, they are sent to both the adaptive memory and the query generator. Adaptive memory is responsible for tracking the action pairs that have been attempted and providing feedback on them. The Query Generator then formulates a query (as illustrated in the Figure 6) to the LLM agent to identify a potential solution. If the suggested action pair fails to resolve the error, a new query is generated, incorporating feedback from the previous attempt.

Once all predicted triplets have been executed successfully, the agent evaluates the task by comparing the current state with the goal state. If the goal has not yet been achieved, it gathers information on the actions performed so far, the current state of the objects, the task description, and feedback on the unmet goal conditions. This information is then used to prompt the LLM agent to generate additional action steps (in the form of triplets) needed to fulfill the remaining goal conditions.

IV. EXPERIMENT

We systematically evaluate our approach using examples from the ALFRED dataset[22] build on ai2thor, a benchmark designed for embodied AI agents that plan and execute low-level actions to accomplish household tasks. These tasks include actions such as cleaning a mug, cutting vegetables, and cooling an apple in the refrigerator. For our evaluation, we selected tasks from three categories: Clean&Place, Heat & Place, and Cool & Place, as they present greater challenges in terms of reasoning and interaction complexity. The examples are drawn from the valid-seen split, which includes various interactive actions such as SliceObject, OpenObject, and ToggleOnObject. The SDT is generated from the behavior of each type

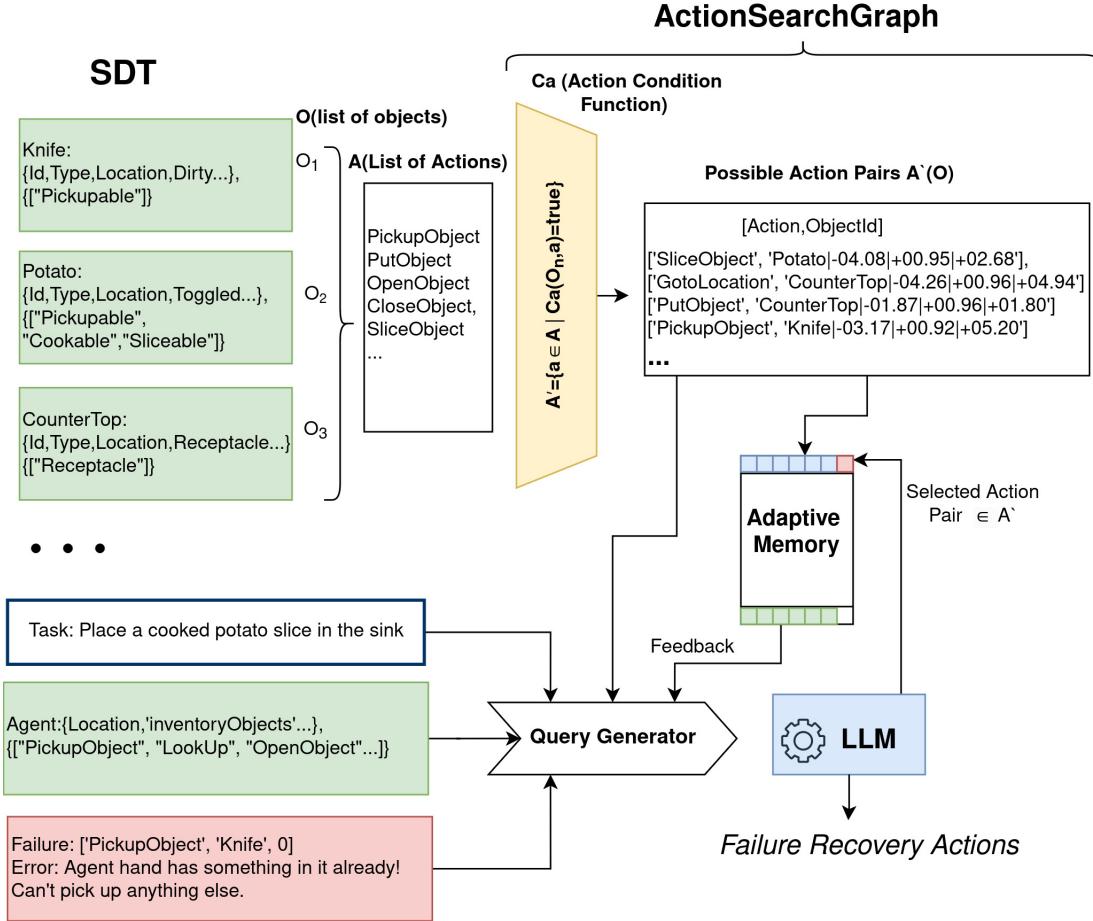


Fig. 2: Overview of the Context-Aware Failure Resolver system: SDT Object descriptions are processed with a condition function to filter possible actions from the action set (A). Filtered action pairs are passed to the Adaptive Memory and Query Generator.

of object that appears in ai2thor. <https://ai2thor.allenai.org/> We use an SDT-assisted LLM to break down the language task description into triplets. Failure conditions are generated by altering the environment’s ground truth from the dataset, such as making an object dirty or placing it inside a closed container.

V. RESULT AND DISCUSSION

We selected a subset of 14 tasks for evaluation that are enactable by an agent. All tasks were successfully completed, demonstrating the robustness of the framework. Success was defined as achieving the specified goal condition. Although all tasks ultimately succeeded, several experienced failure cases during initial attempts. The system used an iterative correction mechanism for each failure case to refine the plan. For example, the task: Place a cooked potato slice in the sink; experienced two failure cases, each of them resolved in 1 iteration. We recorded the number of failure cases per task and the number of iterations required to resolve each failure across all tasks as shown in Table I. This allowed us to assess the planner’s adaptability and the LLM’s ability

to generate corrective actions. Replanner iteration specifies how often the agent invoked the replanner after completing the sequence of predicted triplets in cases where the goal condition was still unmet. This does not refer to replanning after individual triplet failures but after executing all planned steps and failing to reach the desired end state. A zero value means that the predicted triplets were sufficient to achieve the goal. The following Figure 4 compares subgoal success rates across three setups: using triplets generated by the planner alone, using triplets with the failure resolver, and using triplets with the failure resolver followed by the replanner.

Based on the results, we observe that most failure cases stem from incorrect triplet predictions, while others are due to inappropriate object selection. For instance in task “Place a rinsed knife inside a drawer”, the agent attempts to place a object on target but fails because the chosen target lacks sufficient space. In such scenarios, the agent needs to reconsider and revise its object selection. For tasks where the goal condition was not satisfied after initial execution, the Replanner Module was triggered. This component analyzed the current state and generated additional corrective actions using the LLM. We

TABLE I: Summary of Task Execution Performance on for Various Household Tasks

Task IDs	Task Description	No. Failure	Iteration Per Failure	Replanner Iteration	Success
1	Place a cooked potato slice in the sink	2	2	0	Yes
2	Put a cooked piece of potato in the sink.	0	0	2	Yes
3	Place a rinsed knife inside a drawer.	1	1	0	Yes
4	Slice an apple, cook it and set it on the counter	1	1	0	Yes
5	Place a clean knife in the drawer	0	0	0	Yes
6	Put a warm apple slice on the counter.	2	2	0	Yes
7	To cook a sliced tomato to throw it in the trash.	1	1	0	Yes
8	Put a chilled plate on the counter left of the sink.	2	2	0	Yes
9	Set a chilled bottle of wine on the table.	1	4	0	Yes
10	Put a clean mug under the coffee maker.	0	0	0	Yes
11	Put a clean cloth on the back of the toilet.	0	0	0	Yes
12	Put a wet sponge on the counter.	1	2	0	Yes
13	Cool a slice of bread and put it in the microwave.	0	0	0	Yes
14	Cut an apple, cool a piece and bring it to the table	1	1	2	Yes

recorded the number of replanning iterations required per task. Notably, some tasks (such as the heating object examples) did not require any replanning, indicating that the initial plan and failure resolver were sufficient. However, if a goal condition is not met—for example, the agent performs the "heat the potato" action but misses the "slice" action, even though the goal requires the object to be both hot and sliced—then the replanner module steps in. It suggests the agent pick up a knife and perform the slicing action, thereby fulfilling the goal condition.

The purpose of our structured "Rules and Interaction Properties" goes beyond encoding general object behavior. These rules are explicitly designed to reflect the capabilities and constraints of specific robot agents and simulation environments. For example. While an LLM may know that a bottle can be picked up, filled, or broken, the actual feasibility of

these interactions depends on the agent's embodiment and the simulation's physics fidelity. A bottle labeled as "Pickupable" in our framework means that it is semantically and mechanically feasible for a particular robot in a given simulator to perform the action — accounting for reachability, gripper type, force limits, etc. Our framework captures contextual dependencies and interaction consequences that may not be static or universally known. For example, whether an object breaks may depend on the surface it's dropped on, or whether it fills depends on proximity to a simulated water source — details that require structured specification rather than reliance on general commonsense priors. Thus, the rules provide grounded, operationalizable knowledge that is critical for agents operating in simulated or real environments where affordances are not just about what "should" happen but what can happen under defined system constraints.

In earlier research, failures were typically addressed by summarizing the actions taken and then replanning the entire task from the beginning[14]. In contrast, LOTA-Benchmark [7] examined six failure cases and aimed to resolve them through demonstration-based feedback and example-driven replanning. None of these approaches consider SDT based on object-actionable properties to predict context-aware actions. Our proposed method addresses failure cases such as the absence of visual grounding (e.g., attempting to interact with invisible or

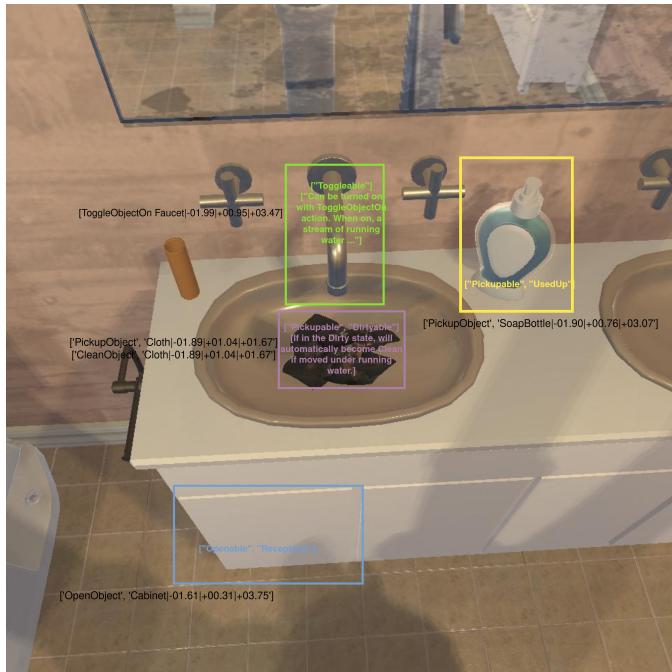


Fig. 3: Semantic Digital Twin based on "Rules and Interaction Properties" and "Textual Descriptions"

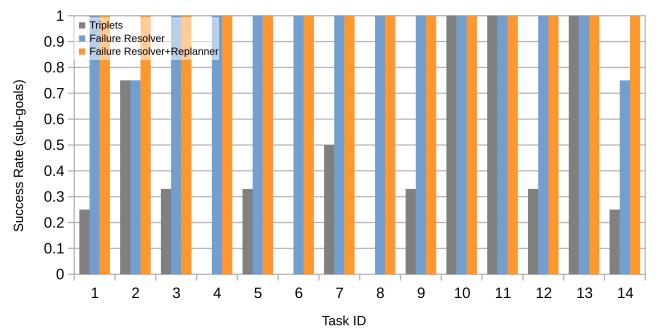


Fig. 4: Comparison of task (in TableI) success with and without Failure Resolver and Re-planner

non-present objects) and object selection errors (e.g., choosing the wrong object). Moreover, the Action interpretation Engine helps execute the action triplets by determining whether interacting with an object logically follows previous actions. Its purpose is to ensure that each triplet is valid on its own and contributes meaningfully to accomplishing the overall task. This is achieved through the use of a Semantic Digital Twin-based ActionSearch Graph, which enhances the agent's understanding of object properties and contextual relevance.

VI. CONCLUSION

This work introduces an adaptive framework that synergistically combines Semantic Digital Twins (SDTs) with Large Language Models (LLMs) to enhance robotic task planning and execution in dynamic, real-world environments. By representing tasks as structured action triplets grounded in semantic and contextual information, our system enables execution, real-time error recovery, and iterative replanning. Experimental evaluations using the ALFRED dataset validate the effectiveness of our approach in handling complex household tasks, even in scenarios involving uncertainty or unexpected failures. Unlike prior methods, our framework leverages the rich semantics of SDTs to inform context-aware decision-making, eliminating reliance on external affordance mapping or retraining. This integration bridges the gap between high-level language reasoning and low-level robotic control, setting a new direction for autonomous agents capable of reliable, interpretable, and resilient behavior. Future work will explore expanding the semantic representation of environments, and incorporating multimodal sensory feedback for more nuanced task understanding.

ACKNOWLEDGMENTS

This research was partially funded by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) “EASE - Everyday Activity Science and Engineering”, University of Bremen.

REFERENCES

- [1] Reo Abe, Akifumi Ito, Kanata Takayasu, and Satoshi Kurihara. Llm-mediated dynamic plan generation with a multi-agent approach, 2025. URL <https://arxiv.org/abs/2504.01637>.
- [2] PS Arjun, Andrew Melnik, and Gora Chand Nandi. Cognitive planning for object goal navigation using generative ai models. In *NeurIPS 2024 Workshop on Open-World Agents*.
- [3] Alessio Baratta, Antonio Cimino, Lucia Gazzaneo, Letizia Nicoletti, and Vittorio Solina. Conceptual modeling for a simulation-based digital twin in human-robot collaboration. *Procedia Computer Science*, 253:3247–3256, 01 2025. doi: 10.1016/j.procs.2025.02.049.
- [4] Vineet Bhat, Ali Umut Kaypak, Prashanth Krishnamurthy, Ramesh Karri, and Farshad Khorrami. Grounding llms for robot task planning using closed-loop state feedback, 2024. URL <https://arxiv.org/abs/2402.08546>.
- [5] Timo Birr, Christoph Pohl, Abdelrahman Younes, and Tamim Asfour. Autogpt+p: Affordance-based task planning using large language models. In *Robotics: Science and Systems XX*, RSS2024. Robotics: Science and Systems Foundation, July 2024. doi: 10.15607/rss.2024.xx.112. URL <http://dx.doi.org/10.15607/RSS.2024.XX.112>.
- [6] Tianxing Chen, Yao Mu, Zhixuan Liang, Zanxin Chen, Shijia Peng, Qiangyu Chen, Mingkun Xu, Ruizhen Hu, Hongyuan Zhang, Xuelong Li, and Ping Luo. G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation, 2025. URL <https://arxiv.org/abs/2411.18369>.
- [7] Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. Lota-bench: Benchmarking language-oriented task planners for embodied agents, 2024. URL <https://arxiv.org/abs/2402.08178>.
- [8] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023. URL <https://arxiv.org/abs/2303.03378>.
- [9] 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>.
- [10] Siyuan Huang, Zhengkai Jiang, Hao Dong, Yu Qiao, Peng Gao, and Hongsheng Li. Instruct2act: Mapping multi-modality instructions to robotic actions with large language model, 2023. URL <https://arxiv.org/abs/2305.11176>.
- [11] Shyam Sundar Kannan, Vishnunandan L. N. Venkatesh, and Byung-Cheol Min. Smart-llm: Smart multi-agent robot task planning using large language models, 2024. URL <https://arxiv.org/abs/2309.10062>.
- [12] Azal Ahmad Khan, Michael Andrev, Muhammad Ali Murtaza, Sergio Aguilera, Rui Zhang, Jie Ding, Seth Hutchinson, and Ali Anwar. Safety aware task planning via large language models in robotics, 2025. URL <https://arxiv.org/abs/2503.15707>.
- [13] Yuchen Liu, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello. Delta: Decomposed efficient long-term robot task planning using large language models, 2025. URL <https://arxiv.org/abs/2404.03275>.
- [14] Zeyi Liu, Arpit Bahety, and Shuran Song. Reflect: Summarizing robot experiences for failure explanation and correction, 2023. URL <https://arxiv.org/abs/2306.15724>.
- [15] Andrew Melnik, Michael Büttner, Leon Harz, Lyon Brown, Gora Chand Nandi, Arjun PS, Gaurav Kumar

- Yadav, Rahul Kala, and Robert Haschke. Uniteam: Open vocabulary mobile manipulation challenge. *arXiv preprint arXiv:2312.08611*, 2023.
- [16] Andrew Melnik, Benjamin Alt, Giang Nguyen, Artur Wilkowski, Qirui Wu, Sinan Harms, Helge Rhodin, Manolis Savva, Michael Beetz, et al. Digital twin generation from visual data: A survey. *arXiv preprint arXiv:2504.13159*, 2025.
- [17] Yusuke Mikami, Andrew Melnik, Jun Miura, and Ville Hautamäki. Natural language as policies: Reasoning for coordinate-level embodied control with llms. *arXiv preprint arXiv:2403.13801*, 2024.
- [18] Soroush Nasiriany, Fei Xia, Wenhao Yu, Ted Xiao, Jacky Liang, Ishita Dasgupta, Annie Xie, Danny Driess, Ayzaan Wahid, Zhuo Xu, Quan Vuong, Tingnan Zhang, Tsang-Wei Edward Lee, Kuang-Huei Lee, Peng Xu, Sean Kirmani, Yuke Zhu, Andy Zeng, Karol Hausman, Nicolas Heess, Chelsea Finn, Sergey Levine, and Brian Ichter. Pivot: Iterative visual prompting elicits actionable knowledge for vlms, 2024. URL <https://arxiv.org/abs/2402.07872>.
- [19] Arjun PS, Andrew Melnik, and Gora Chand Nandi. Splatr: Experience goal visual rearrangement with 3d gaussian splatting and dense feature matching. *arXiv preprint arXiv:2411.14322*, 2024.
- [20] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning, 2023. URL <https://arxiv.org/abs/2307.06135>.
- [21] Krishan Rana, Andrew Melnik, and Niko Sünderhauf. Contrastive language, action, and state pre-training for robot learning. *arXiv preprint arXiv:2304.10782*, 2023.
- [22] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks, 2020. URL <https://arxiv.org/abs/1912.01734>.
- [23] Marta Skreta, Zihan Zhou, Jia Lin Yuan, Kourosh Darvish, Alán Aspuru-Guzik, and Animesh Garg. Re-plan: Robotic replanning with perception and language models, 2024. URL <https://arxiv.org/abs/2401.04157>.
- [24] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: Towards unleashing infinite data for automated robot learning via generative simulation, 2024. URL <https://arxiv.org/abs/2311.01455>.
- [25] Chengxing Xie and Difan Zou. A human-like reasoning framework for multi-phases planning task with large language models, 2024. URL <https://arxiv.org/abs/2405.18208>.
- [26] Senqiao Yang, Jiaming Liu, Ray Zhang, Mingjie Pan, Zoey Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Yandong Guo, and Shanghang Zhang. Lidar-llm: Exploring the potential of large language models for 3d lidar under-
- standing, 2023. URL <https://arxiv.org/abs/2312.14074>.
- [27] Sriram Yenamandra, Arun Ramachandran, Mukul Khanna, Karmesh Yadav, Jay Vakil, Andrew Melnik, Michael Büttner, Leon Harz, Lyon Brown, Gora Chand Nandi, et al. Towards open-world mobile manipulation in homes: Lessons from the neurips 2023 homerobot open vocabulary mobile manipulation challenge. *arXiv preprint arXiv:2407.06939*, 2024.
- [28] Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S. Yu. Large language models for robotics: A survey, 2023. URL <https://arxiv.org/abs/2311.07226>.

APPENDIX

```

Generate a list of action triplets planner for executing the following Task:#TASK DESCRIPTION#
Only use the actions listed here: ['PickupObject', 'PutObject', 'OpenObject', 'CloseObject', 'SliceObject', 'ToggleObjectOn', 'ToggleObjectOff', 'EmptyLiquidFromObject']
Select appropriate object w.r.t to action from the Object_List by analyzing their Object_actionable_properties and Object_rules
-----
Where as:
For clearing,rinse or washing, put object in the SinkBasin first and then toggled the Faucet.
[Object_List][#LIST OF OBJECTS TYPES#]
-----
[Object_actionable_properties][#OBJECTS_ACTIONABLE_PROP#]
-----
[Object_rules][#OBJECTS_PROP#]
-----
[Actions_rules]Basic Instructions about actions selection w.r.t to objects actionable properties
PickupObject: allows the agent to pick up an interactable object specified by its objectid. Compatible objects have "Pickupable" as actionable property
PutObject: attempts to put an object the agent is holding onto or into the target receptacle. Valid target receptacle objects have "Receptacle" as actionable property.
OpenObject: attempts to open an object having "Openable" as actionable property.
CloseObject: attempts to close an object having "Closeable" as actionable property.
SliceObject: attempt to slice an object having "Sliceable" as actionable property. If an object, such as an Apple, is successfully sliced, there may include several new AppleSliced objects in the scene. So Apple Changed to AppleSliced after SliceObject.
ToggleObjectOn: toggles an object on if object is "Toggleable"
ToggleObjectOff: toggles an object off if object is "Toggleable"
EmptyLiquidFromObject: attempts to empty the liquid from an object which has "Fillable" as actionable property and fill with liquid previously.
-----
Example:
Task1:#TASK_DESCRIPTION#
Output1:
[#TRIPLETS#]
Task2:#TASK_DESCRIPTION#
Output2:
[#TRIPLETS#]
Return a Python-style list containing only the list of triplets e.g. [[Action', 'Object1', 'Object2']]. Do not include any explanations.

```

Fig. 5: Action-Triplets Prompting

```

[Query] ---
Response should be like this Reason(string),ActionPair(list)
Example#Reason:The agent needs to go to the Fridge.,ActionPair['GotoLocation', 'Fridge|-00.32|00.00|+03.60']
---
Agent Fails to execute Action Pair ('PickupObject', 'Knife|-03.17|+00.92|+05.20') and
got Error: Agent hand has something in it already! Can't pick up anything else..
Where as Knife|-03.17|+00.92|+05.20 is not visible, is in/on('CounterTop|-04.26|+00.96|+04.94',)
and agent is near to Drawer|-04.09|+00.78|+04.66,Fork|-03.20|+00.92|+05.46,Knife|-03.17|+00.92|+05.20,Potato|-04.08|+00.95|+02.68,
Spatula|-03.91|+05.17,Spoon|-03.17|+00.91|+05.37, and holding Potato|-04.08|+00.95|+02.68.
---
Please select replanner action pair from given list : [['SliceObject', 'Potato|-04.08|+00.95|+02.68'],
['PickupObject', 'Knife|-03.17|+00.92|+05.20'], ['Crouch', 'Knife|-03.17|+00.92|+05.20'],
['Stand', 'Knife|-03.17|+00.92|+05.20'], ['PutObject', 'CounterTop|-04.26|+00.96|+04.94'],
['GotoLocation', 'CounterTop|-04.26|+00.96|+04.94']]
---
[Feedback]Already tried action pairs are : [['SliceObject', 'Potato|-04.08|+00.95|+02.68']]

```

[Response]

Reason: The agent needs to free its hand by putting down the potato before attempting another pickup.
ActionPair: ['PutObject', 'CounterTop|-04.26|+00.96|+04.94']

Fig. 6: Example of Failure Query



Fig. 7: Robot fails to locate bottle in a Fridge (top-left) then it locate the bottle (top-right) after Failure resolver

Task: Set a chilled bottle of wine on the table. Action-Triplets:[['PickupObject', 'WineBottle', 0], ['OpenObject', 'Fridge', 0], ['PutObject', 'WineBottle', 'Fridge'], ['CloseObject', 'Fridge', 0], ['OpenObject', 'Fridge', 0], ['PickupObject', 'WineBottle', 0], ['CloseObject', 'Fridge', 0], ['PutObject', 'WineBottle', 'DiningTable']]. Agent got an Error: “Target object not found within the specified visibility...” during ['PickupObject', 'WineBottle', 0] Failure Resolver suggested solution actions are: [(Crouch,Fridge| -01.30|+00.01|+00.99),(PickupObject,WineBottle|-01.38|+00.76|+02.20)]



Fig. 8: Robot tries to put knife in a drawer.

TASK : Place a rinsed knife inside a drawer. Robot tries to put a Knife in a small-size drawer with the action (PutObject,Drawer|-00.05|+00.38|-01.32) but encounter a Error: “No valid positions to place object found.” Failure Resolver suggested solution actions are: [(OpenObject,Drawer|-00.86|+00.58|+01.43),PutObject,Drawer|-00.86|+00.58|+01.43)]

Task: Put a cooked piece of potato in the sink.

Triplets: [['PickupObject', 'Potato', 0], ['OpenObject', 'Microwave', 0], ['PutObject', 'Potato', 'Microwave'], ['CloseObject', 'Microwave', 0], ['ToggleObjectOn', 'Microwave', 0], ['ToggleObjectOff', 'Microwave', 0], ['OpenObject', 'Microwave', 0], ['PickupObject', 'Potato', 0], ['CloseObject', 'Microwave', 0], ['PutObject', 'Potato', 'Sink']]

Executed Interactive Actions: [['PickupObject', 'Potato|-04.08|+00.95|+02.68'], ['OpenObject', 'Microwave|-02.69|+01.43|+01.69'], ['PutObject', 'Microwave|-02.69|+01.43|+01.69'], ['CloseObject', 'Microwave|-02.69|+01.43|+01.69'], ['ToggleObjectOn', 'Microwave|-02.69|+01.43|+01.69'], ['ToggleObjectOff', 'Microwave|-02.69|+01.43|+01.69'], ['OpenObject', 'Microwave|-02.69|+01.43|+01.69'], ['PickupObject', 'Potato|-04.08|+00.95|+02.68'], ['CloseObject', 'Microwave|-02.69|+01.43|+01.69'], ['PutObject', 'Sink|-04.19|+00.92|+03.28|SinkBasin'], ['PickupObject', 'ButterKnife|-00.45|+00.92|+04.89'], ['SliceObject', 'Potato|-04.08|+00.95|+02.68']]

Task: Cut an apple, cool a piece and bring it to the table.

Triplets: [['PickupObject', 'Knife', 0], ['SliceObject', 'Apple', 0], ['PutObject', 'Knife', 'CounterTop'], ['PickupObject', 'AppleSliced', 0], ['OpenObject', 'Fridge', 0], ['PutObject', 'AppleSliced', 'Fridge'], ['CloseObject', 'Fridge', 0], ['OpenObject', 'Fridge', 0], ['PickupObject', 'AppleSliced', 0], ['CloseObject', 'Fridge', 0], ['PutObject', 'AppleSliced', 'CounterTop']]

Executed Actions: [['OpenObject', 'Drawer|-00.48|+00.80|+00.52'], ['PickupObject', 'Knife|-00.45|+00.77|+00.48'], ['SliceObject', 'Apple|+01.38|+01.04|+03.33'], ['PutObject', 'CounterTop|+01.39|+01.03|+00.90'], ['PickupObject', 'Apple|+01.38|+01.04|+03.33|AppleSliced_0'], ['OpenObject', 'Fridge|+02.04|+00.00|+05.63'], ['PutObject', 'Fridge|+02.04|+00.00|+05.63'], ['CloseObject', 'Fridge|+02.04|+00.00|+05.63'], ['OpenObject', 'Fridge|+02.04|+00.00|+05.63'], ['PickupObject', 'Apple|+01.38|+01.04|+03.33|AppleSliced_0'], ['CloseObject', 'Fridge|+02.04|+00.00|+05.63'], ['PutObject', 'CounterTop|+01.39|+01.03|+04.72'], ['PickupObject', 'Apple|+01.38|+01.04|+03.33|AppleSliced_0'], ['PutObject', 'DiningTable|-00.16|+00.00|+06.47']]

Fig. 9: Tasks, along with their predicted triplets and summary of actions, leads to success after failure resolver and replanner.

In the case of Task: Put a cooked piece of potato in the sink, the robot heats the potato and puts it in the sink but forgets to slice it. So, the replanner module includes two more actions(['PickupObject', 'ButterKnife|-00.45|+00.92|+04.89'], ['SliceObject', 'Potato|-04.08|+00.95|+02.68']) at the end of the task to achieve the goal condition. Whereas in Task: Cut an apple, cool a piece and bring it to the table, When robot try to execute actions-triplet ['PickupObject', 'Knife', 0] it encounter the Error: "Target object not found within the specified visibility...". In this case, the Context-Awareness Failure Resolver predicted that the drawer would open first as the knife was in the drawer. The task's goal is for the apple slice to be cold and on the table, but the robot puts the apple slice on the countertop. Replanner suggested the following two actions:['PickupObject', 'Apple|+01.38|+01.04|+03.33|AppleSliced-0'], ['PutObject', 'DiningTable|-00.16|+00.00|+06.47'] to achieve the goal.

```
{
  "Bottle": ["Will break if subjected to enough force. Will fill up with water if moved underneath a running water source."],
  "Bowl": ["Will automatically break if subjected to enough force. Not all bowls are breakable (Plastic-Unbreakable vs Glass-Breakable vs Ceramic-Breakable). If empty, will fill up with water if moved underneath a running water source. If in the Dirty state, will automatically become Clean if moved under running water."],
  "BreadSliced": ["Use PutObject action to place BreadSliced into Toaster objects. If the Toaster is on, the BreadSliced will automatically cook. BreadSliced will also automatically cook if moved over an active StoveBurner object."],
  "Candle": ["Can be lit using the ToggleObjectOn action. If another source of flame (another Candle or a Stove Burner) touches the wick, the Candle will automatically be lit and turn on. If the wick of the Candle touches a source of water, then the Candle will automatically be put out and turned off."],
  "CellPhone": ["Will automatically break if subjected to enough force. When broken, the CellPhone's screen will appear cracked and the On/Off state will be switched to Off. Broken CellPhones will no longer be able to turn on with ToggleObjectOn."],
  "Cloth": ["If in the Dirty state, will automatically become Clean if moved under running water."],
  "CoffeeMachine": ["Coffee Machines are receptacles but only allow objects of type Mug to be placed in them. If a Coffee Machine is On and an empty Mug is placed in it, the Mug will automatically fill with Coffee and have its Temperature changed to Hot"],
  "Cup": ["Will automatically break if subjected to enough force. Not all Cups are breakable (Plastic-Unbreakable vs Glass-Breakable vs Ceramic-Breakable). If empty, will fill up with water if moved underneath a running water source. If in the Dirty state, will automatically become Clean if moved under running water."],
  "Egg": ["Using the Slice action or the Break action will change the Egg to an EggCracked object. Dropping or Throwing the Egg with enough force will also automatically Break the egg into an EggCracked."],
  "EggCracked": ["Placing an EggCracked object onto an active StoveBurner will automatically cook it. Placing an EggCracked object into a Microwave, and then turning on the Microwave will automatically cook it."],
  "Faucet": ["Can be turned on with ToggleObjectOn action. When on, a stream of running water will pour out from the faucet spout. If the Faucet is over a Sink or a Bathtub, the interior of the Sink/Bathtub will fill with standing water. For objects that can be automatically cleaned, the stream of water is a Running Water Source"],
  "Fridge": ["Objects placed in a Fridge will have their temperature automatically changed to Cold"],
  "HandTowelHolder": ["HandTowelHolders are receptacle objects, but only allow HandTowel type objects to be placed on them."],
  "HousePlant": ["Will fill up with water if moved underneath a running water source."],
  "Kettle": ["Will fill up with water if moved underneath a running water source."],
  "Laptop": ["Will automatically break if subjected to enough force. When broken, the Laptop's screen will appear cracked and the On/Off state will be switched to Off. Broken Laptops will no longer be able to turn on with ToggleObjectOn, but they can still be opened or closed with OpenObject or CloseObject."],
  "Microwave": ["Any objects inside a Microwave while it is turned On will have their temperature changed to Hot. Some object will also automatically cook (Potato, Egg). Microwaves cannot be turned on if they are open, and cannot be opened if they are currently on."],
  "Mirror": ["Automatically breaks if subjected to enough force. For example, if an object thrown at it has enough force, the object may break."],
  ...
}
```

Fig. 10: Objects contextual interactions properties from ai2Thor <https://ai2thor.allenai.org/>

```
{
  "AlarmClock": ["Pickupable"],
  "AluminumFoil": ["Pickupable"],
  "Apple": ["Pickupable", "Sliceable"],
  "AppleSliced": ["Pickupable"],
  "ArmChair": ["Receptacle", "Moveable"],
  "BaseballBat": ["Pickupable"],
  "BasketBall": ["Pickupable"],
  "Bathtub": ["Receptacle"],
  "BathtubBasin": ["Receptacle"],
  "Bed": ["Receptacle", "Dirtyable"],
  "Blinds": ["Openable"],
  "Book": ["Openable", "Pickupable"],
  "Boots": ["Pickupable"],
  "Bottle": ["Pickupable", "Fillable", "Breakable"],
  "Bowl": ["Pickupable", "Receptacle", "Fillable", "Breakable", "Dirtyable"],
  "Box": ["Openable", "Pickupable", "Receptacle"],
  "Bread": ["Pickupable", "Sliceable"],
  "BreadSliced": ["Pickupable", "Cookable"],
  "ButterKnife": ["Pickupable"],
  "Cabinet": ["Openable", "Receptacle"],
  "Candle": ["Pickupable", "Toggleable"],
  "CD": ["Pickupable"],
  "CellPhone": ["Pickupable", "Toggleable", "Breakable"],
  "Chair": ["Moveable"],
  "Cloth": ["Pickupable", "Dirtyable"],
  "CoffeeMachine": ["Toggleable", "Receptacle", "Moveable"],
  "CoffeeTable": ["Receptacle", "Moveable"],
  "CounterTop": ["Receptacle"],
  "CreditCard": ["Pickupable"],
  "Cup": ["Pickupable", "Receptacle", "Fillable", "Breakable", "Dirtyable"],
  "Curtains": [],
  "Desk": ["Receptacle", "Moveable"],
  "DeskLamp": ["Toggleable", "Moveable"],
  "DiningTable": ["Receptacle", "Moveable"],
  "DishSponge": ["Pickupable"],
  "DogBed": ["Moveable"],
  "Drawer": ["Openable", "Receptacle"],
  "Dresser": ["Receptacle", "Moveable"],
  "Egg": ["Pickupable", "Sliceable", "Breakable"],
  "EggCracked": ["Pickupable", "Cookable"],
  "Faucet": ["Toggleable"],
  "FloorLamp": ["Toggleable", "Moveable"],
  "Footstool": ["Moveable"],
  "Fork": ["Pickupable"],
  "Fridge": ["Openable", "Receptacle"],
  "GarbageBag": ["Moveable"],
  "GarbageCan": ["Receptacle", "Moveable"],
  "HandTowel": ["Pickupable"],
  "HandTowelHolder": ["Receptacle"],
  ...
}
```

Fig. 11: Objects actionable properties from ai2Thor <https://ai2thor.allenai.org/>