

RESEARCH ARTICLE

ChatGPT Empowered Long-Step Robot Control in Various Environments: A Case Application

NAOKI WAKE¹, ATSUSHI KANEHIRA, KAZUHIRO SASABUCHI¹, JUN TAKAMATSU¹, (Member, IEEE), AND KATSUSHI IKEUCHI¹, (Life Fellow, IEEE)

Applied Robotics Research, Microsoft, Redmond, WA 98052, USA

Corresponding author: Naoki Wake (naoki.wake@microsoft.com)

ABSTRACT This paper introduces a novel method for translating natural-language instructions into executable robot actions using OpenAI's ChatGPT in a few-shot setting. We propose customizable input prompts for ChatGPT that can easily integrate with robot execution systems or visual recognition programs, adapt to various environments, and create multi-step task plans while mitigating the impact of token limit imposed on ChatGPT. In our approach, ChatGPT receives both instructions and textual environmental data, and outputs a task plan and an updated environment. These environmental data are reused in subsequent task planning, thus eliminating the extensive record-keeping of prior task plans within the prompts of ChatGPT. Experimental results demonstrated the effectiveness of these prompts across various domestic environments, such as manipulations in front of a shelf, a fridge, and a drawer. The conversational capability of ChatGPT allows users to adjust the output via natural-language feedback. Additionally, a quantitative evaluation using VirtualHome showed that our results are comparable to previous studies. Specifically, 36% of task planning met both executability and correctness, and the rate approached 100% after several rounds of feedback. Our experiments revealed that ChatGPT can reasonably plan tasks and estimate post-operation environments without actual experience in object manipulation. Despite the allure of ChatGPT-based task planning in robotics, a standardized methodology remains elusive, making our work a substantial contribution. These prompts can serve as customizable templates, offering practical resources for the robotics research community. Our prompts and source code are open source and publicly available at <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

INDEX TERMS Task planning, robot manipulation, large language models, ChatGPT.

I. INTRODUCTION

Recent advances in natural language processing have yielded large language models (LLMs) with significantly improved abilities to understand and generate language. As a result of learning vast amounts of data, some LLMs can be fine-tuned given a small set of sample data as instructions (i.e., few-shot learning [1]). ChatGPT [2] is a representative example of such an LLM. One exciting application of ChatGPT is in the field of robotics, where it can be used for generating executable robot programs (i.e., task planning).

Task planning from natural-language instructions is a research topic in robotics, and there are many existing

studies [3], [4], [5], some of which are built on top of LLMs [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18] ([19] for review). However, most of them were developed within a limited scope of operations, such as pick-and-place [9], [15], [20], [21], are hardware-dependent, or lack the functionality of human-in-the-loop [10], [14], [16], [22], [23]. Additionally, most of these studies rely on specific datasets [6], [7], [8], [12], [13], [24], [25], [26], [27], [28], [29], which necessitate data recollection and model retraining when transferring or extending these to other robotic settings.

In contrast to these pioneering studies, a significant advantage of utilizing most recent LLMs, such as ChatGPT, is their adaptability to various operational settings. This adaptability is facilitated by few-shot learning, which eliminates the

The associate editor coordinating the review of this manuscript and approving it for publication was Wai-Keung Fung¹.

need for extensive data collection or model retraining in customizing the scope of operations. Additionally, the recent LLMs' superior ability to process language allows for safe and robust task planning, as it can efficiently reflect user feedback in a human-in-the-loop manner.

In this study, we aim to demonstrate a specific but extensible use case of ChatGPT for task planning (Fig. 1), employing ChatGPT as an example of the most recent LLMs. Although interest has been growing in the potential of ChatGPT in the field of robotics [30], its practical application is still in its early stages, and no standardized methodology has yet been proposed. We design customizable prompts to meet the following requirements that are common to many practical robotic applications:

- 1) Easy integration with robot execution systems or visual recognition programs.
- 2) Applicability to various home environments.
- 3) The ability to provide multi-step instructions while mitigating the impact of token limit imposed on ChatGPT.

To meet these requirements, prompts are designed to have ChatGPT accomplish the following:

- 1) Output a sequence of user-defined robot actions with explanations in an easy-to-parse JSON format.
- 2) Explicitly handle the environmental information, enabling task planning considering the spatial relationships between objects.
- 3) Estimate the post-operation environment as a hint for subsequent task planning, reducing the burden of holding lengthy conversation histories for multi-step instructions.

Through experiments, we demonstrate that ChatGPT succeeds in estimating action sequences for multi-step instructions in various environments. Additionally, we show that the conversational capability of ChatGPT allows users to adjust the output through natural-language feedback, which is crucial for safe and robust task planning. Quantitative tests using the VirtualHome environment [31] show that the proposed prompts result in both executable and correct task planning after a few rounds of feedback, suggesting the effectiveness of our approach.

While previous research has validated the utility of ChatGPT within specific environments and scenarios [30], we explored whether ChatGPT can operate effectively across diverse environments and scenarios. This attempt expands the practical applicability of ChatGPT, paving the way for broader and more flexible applications in various real-world settings. Our proposed prompts can serve as customizable templates and are open source and available at <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>. Depending on the specifications of robot actions, environmental representations, and object names, users can easily modify them to meet their requirements. The contributions of this paper are threefold: we demonstrate the applicability of ChatGPT to multi-step task planning with a focus on robot action granularity,

propose a set of customizable prompts adaptable to various environments, and make these prompts publicly accessible as a practical resource for the robotics research community.

II. ChatGPT PROMPTS

The details of the designed prompts are explained in this section. The prompts consist of 1) an explanation of the role of ChatGPT, 2) a definition of robot actions, 3) an explanation of how to represent the environment, 4) an explanation of how to format the output, 5) examples of input and output, and 6) a specific instruction from the user.

In every instance of task planning with ChatGPT, the prompts one to five are loaded from pre-prepared text files, while the sixth prompt is dynamically generated based on the user's instructions and environmental information. Notably, through preliminary experimentation, we found that ChatGPT appears to operate more robustly when we input the six prompts as a conversation consisting of six turns (see Section II-F for details), rather than bundling them into a single prompt. All prompts and their output examples are available online <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>, and anyone can try them out through OpenAI's API or a web browser.

The prompts shown in this section assumed that the robot has at least one arm, sufficient degrees of freedom, and reachability to execute the desired task in the working environment. Additionally, we assume that each instruction is given at the granularity of grasp-manipulation-release, which involves handling a single object from grasping to releasing. Challenges and discussions on extending our approach to more general-purpose robotic systems are discussed in Section IV.

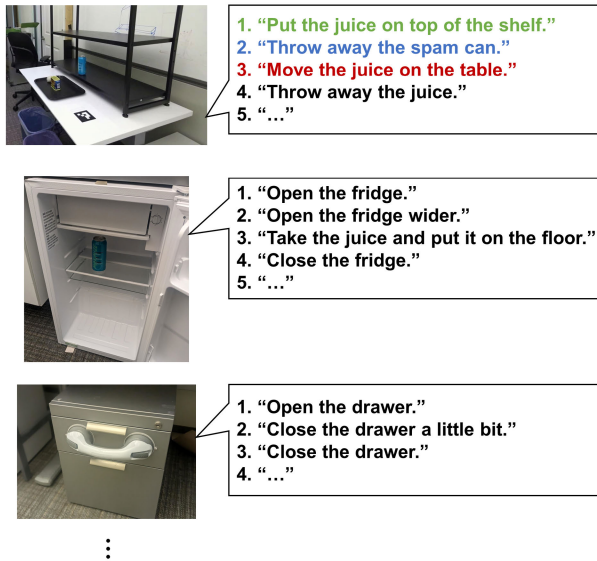
A. THE ROLE OF ChatGPT

In the first prompt, we provide ChatGPT with a context for the specific activity of task planning by explaining the role that ChatGPT should play (Fig. 2). To accommodate multiple prompts, we include a sentence instructing ChatGPT to wait for the next prompt until all the prompts are input.

B. THE DEFINITION OF ROBOT ACTIONS

In this prompt, we define a set of robot actions. Since an appropriate set of robot actions depends on the application and implementation of the robotic software, this prompt should be customized by experimenters. In Fig. 3, we show an example of robot actions based on our in-house learning-from-observation application [32], [33], in which robot actions are defined as functions that change the motion constraints on manipulated objects based on the Kuhn-Tucker theory [34]. This definition allows us to theoretically establish a necessary and sufficient set of robot actions for object manipulation. Experiments in Section III are conducted using these robot actions, except for an experiment in Section III-C, in which we defined a set of actions that were prepared for VirtualHome.

Multi-step instructions in various environments



Readable sequence of object manipulation (JSON files)

ChatGPT



Prompt templates

```
{
  "task_sequence": [
    "move_hand()",
    "grasp_object()",
    "move_object()",
    "move_object()",
    "put_down_object()",
    "release_object()"
  ],
  "step_instructions": [
    "move the hand near the juice",
    "grasp the juice",
    "take the juice from the bottom shelf",
    "move the juice upward",
    "place the juice on top of the shelf",
    "release the juice"
  ],
  "object_states_before": {
    "<spam>": "on_something(<table>)",
    "<juice>": "on_something(<shelf_bottom>)"
  },
  "object_states_after": {
    "<spam>": "on_something(<table>)",
    "<juice>": "on_something(<shelf_top>)"
  },
  "instruction_summary": "put juice on top of shelf"
}
```

FIGURE 1. This paper presents practical prompts designed for various environments. The prompts enable ChatGPT to translate multi-step human instructions into sequences of executable robot actions.

You are an excellent interpreter of human instructions for household tasks. Given an instruction and information about the working environment, you break it down into a sequence of robot actions. Please do not begin working until I say "Start working." Instead, simply output the message "Waiting for next input." Understood?

FIGURE 2. The prompt for explaining the role of ChatGPT.

Necessary and sufficient robot actions are defined as follows:
 """"
 "ROBOT ACTION LIST"
 - move_hand(): Move the robot hand from one position to another with/without grasping an object.
 - grasp_object(): Grab an object.
 - release_object(): Release an object in the robot hand.
 ...
 - wipe_on_plane(): This action can only be performed if an object is grabbed. Move an object landing on a plane along two axes along that plane. For example, when wiping a window with a sponge, the sponge makes this motion.
 """"

FIGURE 3. The prompt explaining a set of robot actions. See Fig. 20 in Appendix A for the full action list.

C. REPRESENTATION OF THE ENVIRONMENTS

This prompt defines the rule for representing working environments (Fig. 4). In this specific prompt, all physical entities are classified into non-manipulable obstacles, referred to as *assets*, such as shelves and tables, and manipulable objects, referred to as *objects*, such as cans and handles. These two classes are defined to differentiate between the entities that may be manipulated and those that cannot. As a hint for task planning, the spatial relationships between entities

are described as *states*, which are chosen from a "STATE LIST." Through preliminary experimentation, items in the STATE LIST were identified as providing sufficient hints for ChatGPT to work effectively. Notably, the STATE LIST is customizable, and in Section III-C, we define different states to meet the specifications of VirtualHome.

D. THE FORMAT OF THE OUTPUT PRODUCED BY ChatGPT

This prompt defines the format of the output produced by ChatGPT (Fig. 5). To facilitate easy integration with other pipelines, such as robot control systems and visual recognition programs, we encourage ChatGPT to output a Python dictionary that can be saved as a JSON file. Additionally, we encourage ChatGPT to include not only the sequence of robot actions, but also explanations of each action step and supplementary information on the updated environment after executing the actions. These additional pieces of information help the user debug whether ChatGPT correctly processes the input information.

E. EXAMPLES OF INPUT AND OUTPUT

This prompt provides examples of the expected inputs and outputs (Fig. 6). We found that providing more examples helps ChatGPT generate the desired sequence and thus minimizes the effort users need to expend to correct the output through conversations.

```

Information about environments and objects are given as Python dictionary.
Example:
"""
{
  "environment": {
    "assets": ["<table>", "<shelf_bottom>", "<shelf_top>",
    "<trash_bin>", "<floor>"],
    "asset_states": {"<shelf_bottom>": "on_something(<table>)",
    "<trash_bin>": "on_something(<floor>)"},
    "objects": ["<spam>", "<juice>"],
    "object_states": {"<spam>": "on_something(<table>)",
    "<juice>": "on_something(<shelf_bottom>)"}}
}
"""
Asset states and object states are represented using those state sets:
"""
"STATE LIST"
- on_something(<something>): Object is located on <something>
- inside_something(<something>): Object is located inside <something>
- inside_hand(): Object is being grasped by a robot hand
- closed(): Object can be opened
- open(): Object can be closed or kept opened
"""
<something> should be one of the assets or objects in the environment.
    
```

FIGURE 4. The prompt for defining the rules for representing working environments.

```

You divide the actions given in the text into detailed robot actions and put them
together as a Python dictionary. The dictionary has five keys:
"""
- dictionary["task_cohesion"]: A dictionary containing information about the
robot's actions that have been split up.
- dictionary["environment_before"]: The state of the environment before the
manipulation.
- dictionary["environment_after"]: The state of the environment after the
manipulation.
- dictionary["instruction_summary"]: contains a brief summary of the given
sentence.
- dictionary["question"]: If you cannot understand the given sentence, you can
ask the user to rephrase the sentence. Leave this key empty if you can
understand the given sentence.
"""
Three keys exist in dictionary["task_cohesion"].
"""
- dictionary["task_cohesion"]["task_sequence"]: Contains a list of robot
actions. Only the behaviors defined in the "ROBOT ACTION LIST" will be
used.
- dictionary["task_cohesion"]["step_instructions"]: contains a list of
instructions corresponding to dictionary["task_cohesion"]["task_sequence"].
- dictionary["task_cohesion"]["object_name"]: The name of the manipulated
object. Only objects defined in the input dictionary will be used for the object
name.
"""
    
```

FIGURE 5. The prompt for defining the format of the output produced by ChatGPT.

F. SPECIFIC INSTRUCTION FROM THE USER

While the previous five prompts are fixed, the sixth prompt is dynamically generated in every instance of task planning by editing a template prompt (Fig. 7). This prompt is generated by replacing [INSTRUCTION] with the given instruction and [ENVIRONMENT] with the corresponding environmental information. Notably, the user is required to provide the environmental information in the initial instance of task planning using a separate process (e.g., manual preparation). However, this effort is unnecessary for subsequent instances because we can reuse an updated environment incorporated in the last output of ChatGPT (orange parts in Fig. 8). This approach facilitates task planning based on the most

```

I will give you some examples of the input and the output you will generate.
"""
Example 1:
"""
- Input:
{
  ... (environmental information)...
  "instruction": "Put the juice on top of the shelf"
}
- Output:
{
  "task_cohesion": {
    "task_sequence": [
      "move_hand()",
      "grasp_object()",
      ... ,
      "attach_to_plane()",
      "release_object()"
    ],
    "step_instructions": [
      "move the hand near the juice",
      "grasp the juice",
      ... ,
      "place the juice",
      "release the juice"
    ]
  },
  "object_name": "<juice>"
},
"environment_before": ... ,
"environment_after": ... ,
"instruction_summary": "put the juice on top of the shelf",
}
"""
Example 2:
"""
...
    
```

FIGURE 6. The prompt providing examples of desired inputs and outputs. The full information is available at the URL provided in the text.

```

"""
{"environment": [ENVIRONMENT]}
"""
The instruction is as follows:
"""
{"instruction": [INSTRUCTION]}
"""
The dictionary that you return should be formatted as python dictionary.
Follow these rules:
1. The first element should be move_hand() to move the robot hand closer to
the object. Always end with releasing the object.
2. Make sure that each element of the ["step_instructions"] explains
corresponding element of the ["task_sequence"]. Refer to the "ROBOT
ACTION LIST" to understand the elements of ["task_sequence"].
...
9. Make sure that you output a consistent manipulation as a single arm robot.
For example, grasping an object should not occur in successive steps.
Adhere to the output format I defined above. Follow the nine rules. Think step
by step.
    
```

FIGURE 7. The user input template and examples of the actual input used. The user is assumed to provide environmental information. Multi-step task planning can be realized by reusing the environmental information that ChatGPT outputs in the following task planning.

recent environment, eliminating the need for extensive historical records that exceed ChatGPT's token limit. In our experiments, we practically included as much of the historical record as the token limit of ChatGPT allowed, ranging from the most recent to the oldest conversation history (Fig. 8).

As a specific usage of this task planner within a robot system, we assume that the output of ChatGPT is checked by the user in every instance of task planning. If the user

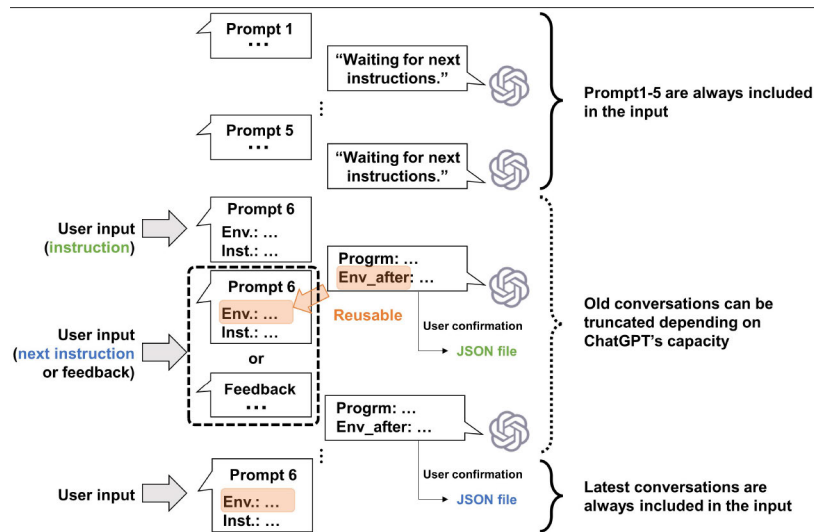


FIGURE 8. The entire structure of the conversation with ChatGPT in task planning.

confirms that there is no further need for adjustment, the output is then saved as a JSON file. In Appendix C, we have provided more details regarding how the proposed task planner is integrated and operated within a robot system.

III. EXPERIMENTS

We tested the prompts to verify whether ChatGPT behaves in accordance with the specified requirements. We used a fixed GPT model provided by Azure OpenAI (gpt-3.5-turbo) in our experiments. Some experimental results are not fully presented in order to save space, but all results, including parameters for ChatGPT inference, instructions, and environment definitions, can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

A. MULTI-STEP MANIPULATION OF THE ENVIRONMENT

We tested the applicability of the proposed prompts to multi-step instructions in various environments. As examples of household tasks, we conducted role-plays instructing the rearrangement and disposal of objects placed on tables and shelves, retrieving objects from refrigerators and drawers, and cleaning tables and windows with a sponge. The instructions and feedback texts were prepared in a style that resembles the way humans communicate with each other. The environmental information in the initial instance of task planning was prepared manually for each scenario. The output of ChatGPT was manually checked by the authors at every instruction step. Specifically, we conducted a visual inspection to qualitatively confirm whether the generated action sequences were executable and whether they accompanied reasonable environment estimations. In summary, the results shown below suggest that ChatGPT can translate multi-step human instructions into adequate sequences of executable robot actions.

1) RELOCATION OF OBJECTS ON A TABLE

The task involves manipulating a can of juice situated on the bottom shelf of a two-shelf structure and a can of spam positioned on a table (refer to the top panel in Fig. 1 for the scene). First, the juice is relocated from the bottom to the top shelf. Subsequently, the spam is discarded into a trash bin. Thereafter, the juice is moved from the top shelf to the table. Finally, the juice, too, is discarded into the trash bin. The output of ChatGPT, which demonstrates successful task planning, is shown in Fig. 9.

2) OPEN A FRIDGE/DRAWER DOOR

Next, we tested the scenario of opening a refrigerator door, opening the door slightly wider, removing a juice from the refrigerator and placing it on the floor, and finally closing the refrigerator (see the middle panel in Fig. 1 for the scene). The output of ChatGPT is shown in Fig. 10, indicating a successful task planning. Similar results were obtained for the scenario of sliding a drawer open (data not shown).

3) WIPE A WINDOW WITH A SPONGE, AND THROW IT AWAY

Next, we tested the scenario of taking a sponge from the desk, wiping the window with the sponge, and returning it to the table. Following the operation, a user throws the sponge into a trash bin. The output of ChatGPT is shown in Fig. 11, indicating a successful task planning. Similar results were obtained for the scenario of wiping the table with a sponge (data not shown).

B. ADJUSTMENT OF THE OUTPUT PRODUCED BY CHATGPT THROUGH USER FEEDBACK

Since ChatGPT does not always generate complete action sequences, it is important for users to review and correct errors to ensure safe and robust operation. With this in mind,

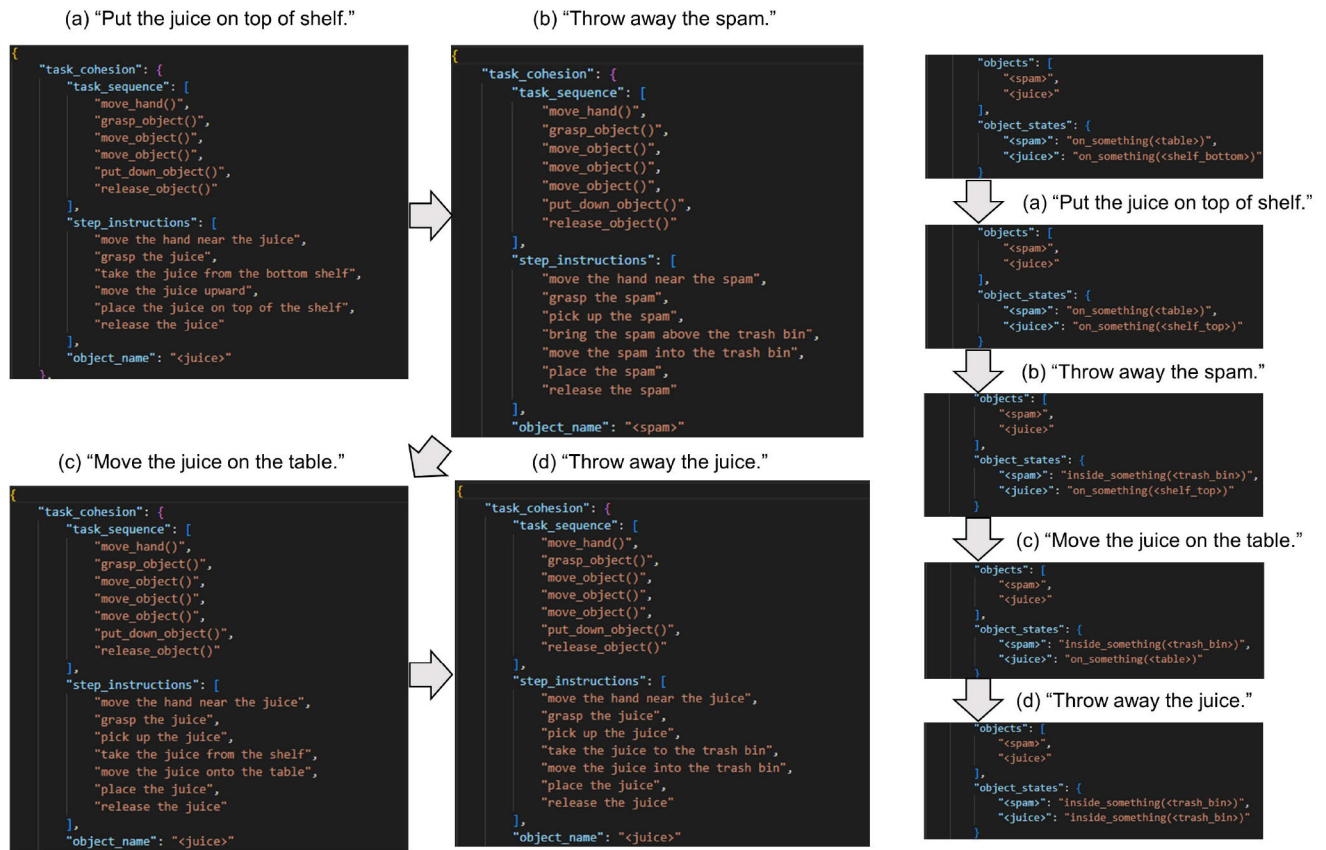


FIGURE 9. An example of the output produced by ChatGPT for the task of relocating objects. (Left panel) Robot actions broken down for each natural language. (Right panel) The state of the environment that is output by ChatGPT. A part of the JSON output is shown for each file. All the results, including the representation of the environment can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

we tested the ability of ChatGPT to adjust the output through natural-language feedback.

Fig. 12 shows the result when a user asked ChatGPT to add/remove a task in the output sequence. ChatGPT changed the output following the semantic content of the feedback, suggesting the functionality for making the necessary adjustments.

C. QUANTITATIVE EVALUATION OF TASK DECOMPOSITION PERFORMANCE OF ChatGPT

The previous sections qualitatively demonstrated that the proposed prompts achieve successful task planning using an action set from our in-house learning-from-observation system. In this section, we quantitatively evaluate the performance of task planning using a general-use simulation environment called VirtualHome [31]. Specifically, we made ChatGPT generate task plans from a single instruction for several household operation scenarios, and tested whether the resulting action sequences were valid in terms of executability in simulation and correctness upon visual inspection. All the source codes and prompts used for the experiment can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

1) EXPERIMENTAL SETUP

VirtualHome is software that simulates interactions between an agent and various home environments. The agent can navigate and interact within these environments by executing sequences of commands using a Python-based API. This API provides a set of pre-defined atomic actions (see Table 4 in Appendix B for the action list), which represent the smallest units of action in VirtualHome. The environment consists of typical household objects (e.g., a plate, a microwave, and a table), each associated with unique IDs. The relationships between objects are represented as a graph that can be accessed through the API.

In the experiments, we selected a kitchen as a representative home environment and defined test scenarios for household chores that could be implemented in VirtualHome. To mitigate bias in the process of scenario preparation, we utilized ChatGPT to generate scenario candidates.¹ We then manually selected the test scenarios that satisfied the following criteria:

- The scenario can be realized by executing multiple actions in sequence.

¹ChatGPT was provided with the list of the kitchen objects and the pre-defined atomic actions to generate the candidates.

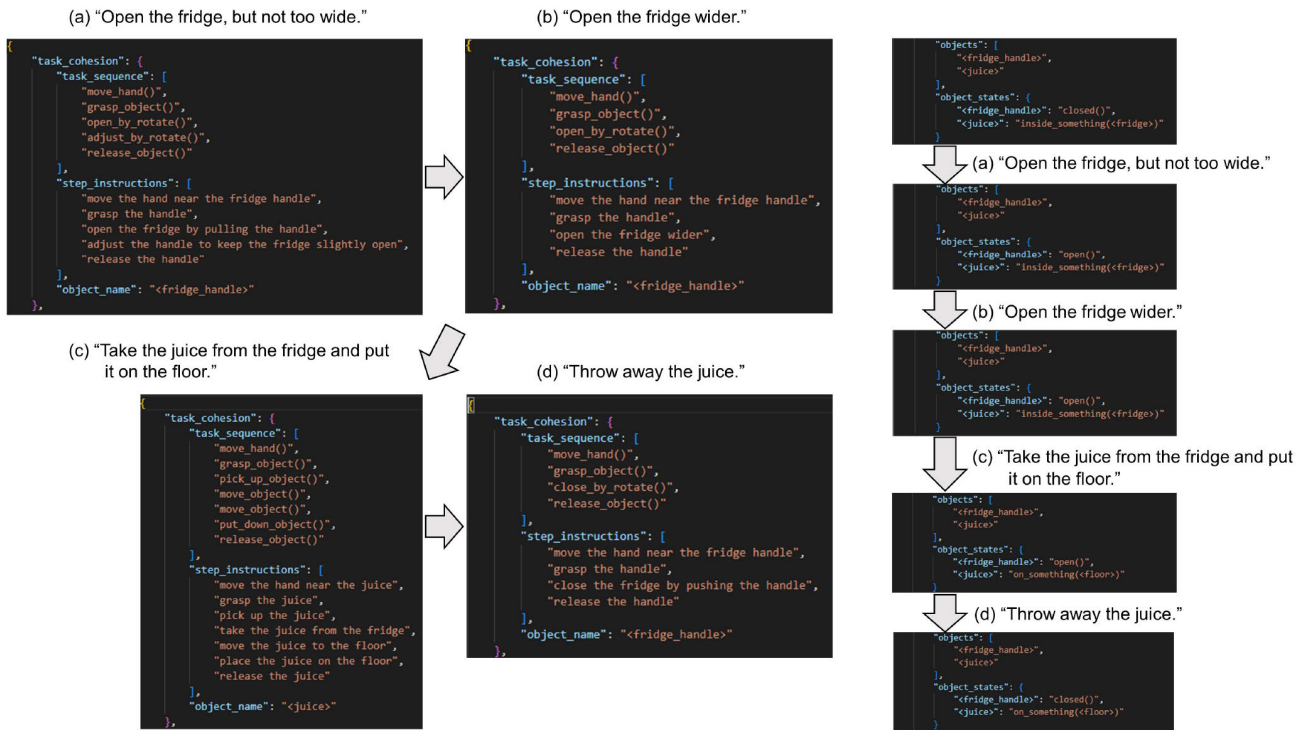


FIGURE 10. An example of the output produced by ChatGPT for the task of opening a refrigerator and retrieving juice. (Left panel) Robot actions broken down for each natural language. (Right panel) The state of the environment that is output by ChatGPT. A part of the JSON output is shown for each file. All the results, including the representation of the environment can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

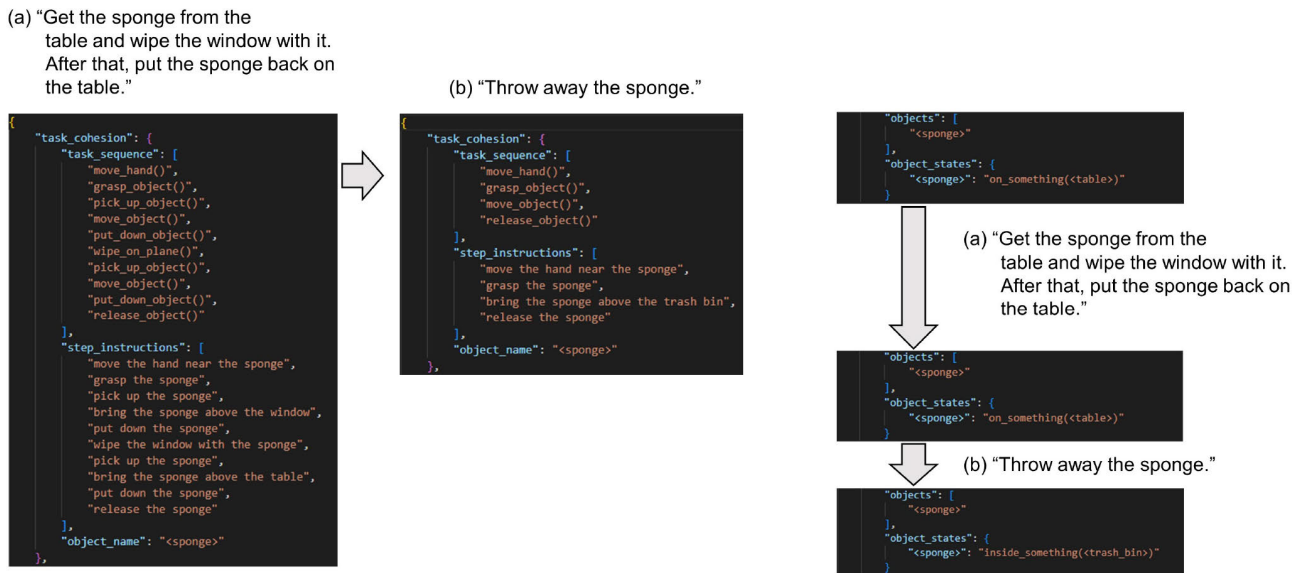


FIGURE 11. An example of the output produced by ChatGPT for the task of wiping a window with a sponge (Left panel) Robot actions broken down for each natural language. (Right panel) The state of the environment that is output by ChatGPT. A part of the JSON output is shown for each file. All the results, including the representation of the environment can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

- The scenario involves at least one instance of object manipulation, i.e., grasping or releasing an object.
- The tasks in the scenario are relevant to everyday activities.

Fourteen scenarios were prepared as test scenarios (Table 1). For these scenarios, we manually identified action sequences to achieve the scenarios along with the list of objects involved with their IDs. The action sequences identified are provided in Table 5 in Appendix B.

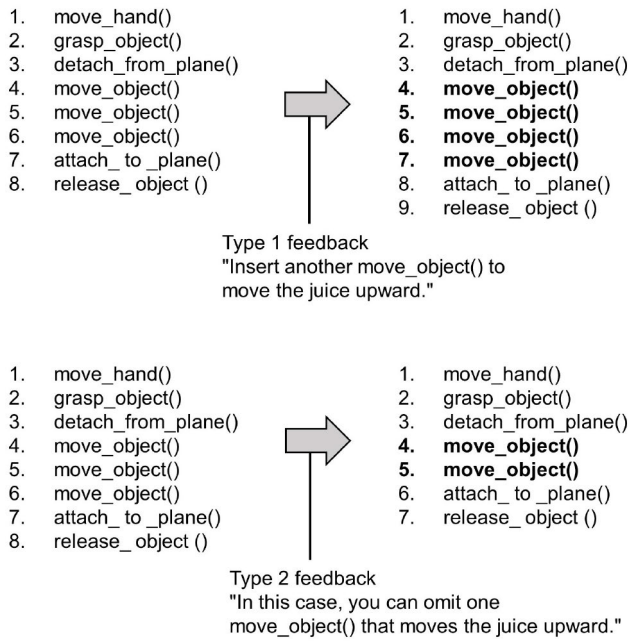


FIGURE 12. An example of adjusting an output sequence through natural-language feedback. The initial instruction was to move a juice from the bottom shelf to the top shelf. (Top panel) After the feedback of “Insert another move_object() to move the juice upward,” an action of move_object() was added to the sequence. (Bottom panel) After the feedback of “In this case, you can omit one move_object() that moves the juice upward,” an action of move_object() was deleted from the sequence.

ChatGPT generated an action sequence intended to complete the scenario, given the proposed prompts, environmental information, and an instruction provided in the right column of Table 1. The environmental information, which corresponds to each scenario, was derived from the graph. Because of redundancy in representing all kitchen objects, only those objects involved in each scenario were considered. The action sequence generated by ChatGPT was then converted into a format that VirtualHome could interpret and executed in a step-by-step manner through the API. An action sequence was considered successful when the following two conditions were met:

- Executability: The simulator was able to execute all steps without encountering any errors.
- Correctness: Upon visual inspection, it was determined that the proposed action steps could successfully complete the scenario.

We incorporated visual inspections in our criteria because a successful execution in the simulator does not necessarily guarantee that the final goal is achieved [23].

2) RESULTS

We first tested whether the generated action sequences were successful without feedback. Since we conducted multiple trials, we set the temperature parameter to its maximum to ensure trial-to-trial variations in the output of ChatGPT. Table 2 shows the results, with a success rate

TABLE 1. The list of scenarios used in the experiment.

Scenario	Textual instruction
Scenario1	Take the bread from the toaster on the kitchen counter and place it on the plate on the table.
Scenario2	Take the frying pan from the counter and place it in the sink.
Scenario3	Take the pie from the table and put it in the microwave.
Scenario4	Take the condiment shaker from the bookshelf and place it on the table.
Scenario5	Take the book from the table and put it on the bookshelf.
Scenario6	Take the water glass from the table and drink from it.
Scenario7	Take the salmon on top of the microwave and put it in the fridge.
Scenario8	Turn on the TV.
Scenario9	Put a plate that is on the table into the sink.
Scenario10	Take the pie on the table and warm it using the stove.
Scenario11	Put the sponge in the sink and wet it with water.
Scenario12	Take the knife from the table and move it to another place on the table.
Scenario13	Take the plate from the table and move it to another place on the table.
Scenario14	Take the condiment bottle from the bookshelf and put it on the table.

of approximately 36% (5 out of 14 scenarios), with only a minimal variation observed between trials.

Upon investigating the unsuccessful cases, we identified two failure patterns in ChatGPT:

- Incorrect verb selection: In VirtualHome, the simulator raises errors when it fails to select an action applicable to an object. For example, when the task involves “placing an object,” the action “PutIn” should be selected when placing the object inside a container, while the action “Put” should be selected when placing it on a flat surface. Despite these verb selection rules being part of the prompts, ChatGPT sometimes confused the actions.
- Omission of necessary steps: Some outputs skipped essential steps necessary for successfully completing a scenario, such as opening a container before placing an object inside it.

Following this analysis, we investigated whether adjustments could be made with a reasonable amount of effort, given appropriate feedback from a user who is familiar with task planning. To this end, we prepared an automatic feedback system as an objective method that detects these types of errors. This system was designed to simulate a user who is knowledgeable in task planning, instead of relying on manual feedback. The output of ChatGPT was checked by the system in every instance of task planning, and if an error was detected, an error message was automatically generated and fed back to ChatGPT. Using this feedback system, we examined the number of rounds of feedback needed to reach a successful sequence, or whether it was possible at all, across 14 scenarios. We set the temperature

TABLE 2. Executability of the output action sequence across trials. "1" indicates success, and "0" indicates failure.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Trial 1	0	0	0	0	1	1	0	1	0	0	0	1	1	0
Trial 2	0	0	0	0	1	1	0	1	0	0	0	1	1	0
Trial 3	0	0	0	0	1	1	0	1	0	0	0	1	1	0
Trial 4	0	0	0	0	1	1	0	1	0	0	0	1	1	0
Trial 5	0	0	0	0	1	1	0	1	0	0	0	1	1	0

TABLE 3. The number of rounds of feedback needed to reach a successful sequence.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of feedback	1	1	3	1	0	0	1	0	1	2	1	0	0	1

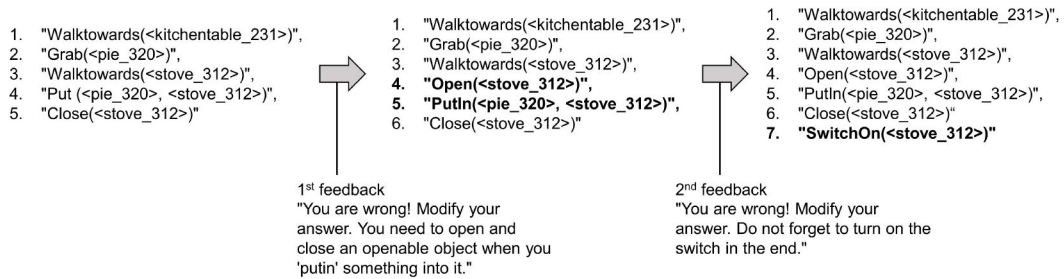


FIGURE 13. Example of adjusting an output sequence through auto-generated feedback. The output for the scenario 10 is shown (i.e., Take the pie on the table and warm it using the stove.) All the results, including the representation of the environment, can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

parameter to its minimum to ensure the reproducibility of the results.

Table 3 shows the results. ChatGPT was able to produce successful action sequences in all scenarios after receiving several rounds of feedback. Fig. 13 shows an example where auto-generated feedback texts guide ChatGPT towards a successful action sequence, suggesting that ChatGPT is capable of reflecting the semantic content of the feedback in its output and making the necessary adjustments.

Our proposed prompts aim to estimate the post-operation environment as a hint for subsequent task planning, which enables multi-step task planning beyond the token limit imposed on ChatGPT. Therefore, we visually inspected the output of ChatGPT in Table 2 and examined whether the post-operation environment was accurately estimated. As a result, we found that for all 14 scenarios and five trials, the estimation by ChatGPT was accurate, regardless of the success or failure of the action sequence. This result suggests that our proposed prompts can be adopted for multi-step task planning beyond a single scenario. The results can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

IV. DISCUSSION: TOWARDS MORE GENERAL ROBOTIC APPLICATIONS

In this study, we focused on task planning of robot actions from multi-step instructions. We designed prompts for ChatGPT to meet three requirements: 1) easy integration with robot execution systems or visual recognition programs,

2) applicability to various environments, and 3) the ability to provide multi-step instructions while mitigating the impact of token limit imposed on ChatGPT. Through experiments, we confirmed that the proposed prompts work for multi-step instructions in various environments, and that ChatGPT enables the user to adjust the output interactively. Based on these results, we believe that the proposed prompts are practical resources that can be widely used in the robotics research community.

It is noteworthy that ChatGPT is capable of performing task planning without any actual experience in object manipulation, relying solely on few-shot data. This ability may be attributed to the fact that the model acquires knowledge of object manipulation and the temporal relationships between cohesively occurring actions during its training on a vast amount of data. In fact, the ability of ChatGPT to generate recipes from a menu suggests that it implicitly learns procedural steps [30]. Nevertheless, we cannot access the inner computations of ChatGPT, thus the computation process for task planning, including the estimation of the post-operation environment, remains unclear.

The quantitative analysis using VirtualHome showed that ChatGPT produced action sequences with a success rate of 36% without feedback, which approached 100% after several rounds of feedback. Although we used a different task set, our results align with those of previous studies that used LLMs for task planning in VirtualHome. Huang et al. [23] reported that 35.23% of an LLM's outputs were both executable

(a) Example of an environment and instruction

```
environment = {"assets": ["<table>", "<shelf_bottom>", "<shelf_top>",
"<trash_bin>", "<floor>"],
"asset_states": {"<shelf_bottom>": "on_something(<table>)",
"<trash_bin>": "on_something(<floor>)"},
"objects": ["<spam>", "<juice>"],
"object_states": {"<spam>": "on_something(<table>)",
"<juice>": "on_something(<shelf_bottom>)"}}
instructions = ["Take the spam, and throw it away if the our-of-date date is
expired. Otherwise, put it on the shelf."]
```

(b) Example of ChatGPT's output

```
{
  "instruction_summary": "Take the spam, and throw it away if the
our-of-date date is expired. Otherwise, put it on the shelf.",
  "question": "",
  "python_script": """
# Grasp the spam and detach it from the table
move_hand() # move hand to the spam; <spam>: on_something(<table>)
grasp_object() # grasp the spam; <spam>: inside_hand()
detach_from_plane() # detach the spam from the table; <spam>:
inside_hand()

# Move the spam near the camera
move_hand() # move the spam near the camera; <spam>: inside_hand()

# Check the best-by date of the spam
check_best_by_date() # <spam>: inside_hand()

# If the best-by date is expired, throw it away
if not check_best_by_date():
    move_hand() # move hand near the trash bin; <spam>: inside_hand
    ()
    release_object() # release the spam to drop it in the trash
    bin; <spam>: inside_something(<trash_bin>)
# If the best-by date is not expired, put it on the shelf
else:
    move_hand() # move hand to the shelf; <spam>: inside_hand()
    move_hand() # move the spam to the shelf; <spam>: inside_hand()
    attach_to_plane() # place the spam on the shelf; <spam>:
    on_something(<shelf_top>)
    release_object() # release the spam; <spam>: on_something
    (<shelf_top>)
"""
}
```

FIGURE 14. An example demonstrating the feasibility of using ChatGPT to generate control programs that include conditional branching. A part of the prompts is shown. Note that we encouraged ChatGPT to add comments at every line to track the state of objects, as the final state may vary according to the conditional branching. We also added a non-manipulative function (i.e., `check_best_by_date()`) in the robot action set. All the results, including the representation of the environment, can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

and correct from a human perspective. Raman et al. [17] showed an improvement in both task execution and correctness through re-prompting based on precondition error information. Thus, we emphasize that our findings attest to the effectiveness of the proposed prompts, which aligns with the existing research.

The prompts were designed under the assumption that the robot has at least one arm, sufficient degrees of freedom, and reachability to execute the desired task in a given environment. Additionally, we assume that each instruction is given at the granularity of grasp-manipulation-release. However, these assumptions may be restrictive for some scenarios in general robotic manipulations. In the following sections, we discuss several strategies to effectively integrate our task planner with practical robotic applications.

(a) Prompt

```
You are an agent that plans the actions of a robot. There are two
files available, throw_away_can.json and throw_away_paper.json, which
the robot can execute by calling the function execute(filename) to
find and dispose of an empty can on the desk into the recycling bin
and a paper cup into the burning garbage bin respectively. Now,
consider the situation where there are multiple empty cans and
multiple paper cups on the table, and the robot needs to dispose of
all of them. The function find_object() searches for an object on the
desk and returns its name if found. If the object is not found, it
returns None. Write a Python source code using these functions to
have the robot dispose of all the empty cans and paper cups on the
desk into their respective garbage bins.
```

(b) Example of ChatGPT's output

```
Sure, here's an example Python source code that uses the provided
functions to dispose of all the empty cans and paper cups on the desk
into their respective garbage bins:
```python
def dispose_all_garbage():
 while True:
 # Check if there is any cans left on the desk
 can = find_object("empty_can")
 if can is None:
 break
 else:
 # Dispose of the can in the recycling bin
 execute("throw_away_can.json")

 while True:
 # Check if there is any paper cups left on the desk
 cup = find_object("paper_cup")
 if cup is None:
 break
 else:
 # Dispose of the paper cup in the burning garbage bin
 execute("throw_away_paper.json")
...
```
```

FIGURE 15. An example of using separate ChatGPT process to generate higher-level conditional branching control by reading out stored task plans.

A. HANDLING OF CONDITIONAL BRANCHING

Some manipulations may require selecting actions based on the recognition results (e.g., disposing of a food item if it is recognized as out of date), or require repeating actions until certain conditions are met (e.g., wiping a table until it is spotless). It is known that LLMs can generate programs that include conditional branching [35]. It has also been suggested that ChatGPT can handle conditional branching for robotic applications [30]. Consistent with these ideas, we confirmed that small modifications to the prompts enabled ChatGPT to generate a Python code that included conditional branching (Fig. 14). Additionally, we verified that employing a separate ChatGPT process enables higher-level conditional branching by composing sets of task plans (Fig. 15). These results suggest the feasibility of extending the proposed task planner to handle conditional branching.

B. COLLABORATION OF MULTIPLE ARMS AND ROBOTS

A robot with multiple arms may need to coordinate its arms to perform a task. We confirmed that small modifications to the prompts enabled ChatGPT to generate an action sequence involving the arms (Fig. 16). Additionally, we verified that employing a separate ChatGPT process enables the coordination of multiple arms by composing sets of task plans

(a) Example of an environment and instruction

```
environment = {"assets": ["<fridge>", "<floor>"],
              "asset_states": {"<fridge>": "on_something(<floor>)"},
              "objects": ["<fridge_handle>", "<juice>"],
              "object_states": {"<fridge_handle>": "closed()",
                               "<juice>": "inside_something(<fridge>)"}}
instructions = ['Open the fridge with the right arm, take the juice and put
it on the floor with the left arm, and close the fridge',]
```

(b) Example of ChatGPT's output

```
{
  "task_cohesion": {
    "task_sequence": [
      "move_hand(right)",
      "grasp_object(right)",
      "open_by_rotate(right)",
      "move_hand(left)",
      "grasp_object(left)",
      "detach_from_plane(left)",
      "move_hand(left)",
      "move_hand(left)",
      "move_hand(left)",
      "attach_to_plane(left)",
      "release_object(left)",
      "close_by_rotate(right)",
      "release_object(right)"
    ],
    "step_instructions": [
      "move the right hand near the fridge handle",
      "grasp the fridge handle with the right hand",
      "open the fridge by pulling the handle with the right hand",
      "move the left hand near the juice",
      "grasp the juice with the left hand",
      "pick up the juice with the left hand",
      "take the juice out of the fridge with the left hand",
      "move the juice to the floor with the left hand",
      "move the left hand down to the floor",
      "place the juice on the floor with the left hand",
      "release the juice with the left hand",
      "close the fridge by pushing the handle with the right hand",
      "release the fridge handle with the right hand"
    ],
    "object_names": ["<fridge_handle>", "<juice>"]
  },
}
```

FIGURE 16. An example demonstrating the feasibility of ChatGPT in generating control programs that involve multiple arms or robots cooperating. Note that we included hand laterality in every function and outputted all the objects to be manipulated, as multiple objects can be handled during the grasp-manipulation-release operations of both hands. All the results, including the representation of the environment, can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

(Fig. 17). These results suggest the feasibility of extending the proposed task planner to handle multiple arms and robots.

C. MANAGING ENVIRONMENTAL CHANGES

One unique aspect of our approach is that we explicitly handle changes in environmental information by incorporating it as part of the input to and output of ChatGPT, respectively. In the context of Minsky's frame theory [36], environmental information can serve as "frames" that guide ChatGPT in selecting the most appropriate plan among a multitude of task planning options. Moreover, enabling ChatGPT to be aware of environmental information may enhance its ability to output consistent task plans [37]. However, a limitation of this approach is the necessity to prepare environmental information, specifically for the initial instance of task planning (Fig. 8). In future studies, we aim to explore a separate ChatGPT process to prepare this information based on a symbolic scene understanding given

(a) Prompt

```
You are an agent that plans the actions of a robot. There are three files available: open_fridge.json (to open the fridge using an arm), close_fridge.json (to close the fridge using an arm), and take_juice.json (to find and take out a juice can from the opened fridge and place it on the table). The robot can execute these files by calling a function: execute(filename, arm_laterality), where arm_laterality specifies whether to use the left or right arm. Now, using these functions, please write a Python source code that has the robot take out a juice can using the left arm while opening the door with the right arm.
```

(b) Example of ChatGPT's output

```
Here's the Python source code that has the robot take out a juice can using the left arm while opening the door using the right arm:
python
# Open the fridge using the right arm
execute("open_fridge.json", "right")

# Take out the juice can using the left arm
execute("take_juice.json", "left")

# Close the fridge using the right arm
execute("close_fridge.json", "right")
```

FIGURE 17. An example demonstrating the feasibility of ChatGPT in generating control programs that involve multiple arms or robots coordinating. This type of planning is beyond the scope of this paper.

either by a vision encoder or through human explanation (Fig. 18).

Additionally, our current approach assumes static environments, where changes are attributed solely to the robot's actions, and the environment remains consistent from task planning to execution. However, real-world scenarios frequently involve dynamic changes, such as the movement, introduction, or disappearance of objects, including people. Addressing such dynamic environments in task planning is an important direction for future research.

D. CONNECTION WITH VISION SYSTEMS AND ROBOT CONTROLLERS

Among recent experimental attempts that used ChatGPT for task planning, our work is unique in its focus on the generation of robot action sequences, addressing the "what-to-do" aspect, and we consciously avoid redundant language instructions related to visual and physical parameters, such as how to grab [39], [40], [41], where to focus [42], and what posture to adopt [32], [43], which are part of the "how-to-do" aspect. Both types of information are vital for robot operation [44], yet the "how-to-do" aspect is often more effectively demonstrated visually. Therefore, our design approach is such that the "what-to-do" is obtained by a vision system or a robot system following task planning, which is outside the scope of this paper.

As part of our efforts to develop a realistic robotic operation system, we have integrated our proposed task planner with a learning-from-observation system (Fig. 19) incorporating a speech interface [45], [46], a visual teaching interface [47], a reusable robot skill library [48], [49], and a simulator [50]. The code for the teaching system is available at: <https://github.com/microsoft/cohesion-based-robot-teaching-interface>. For reference, details of the robotic

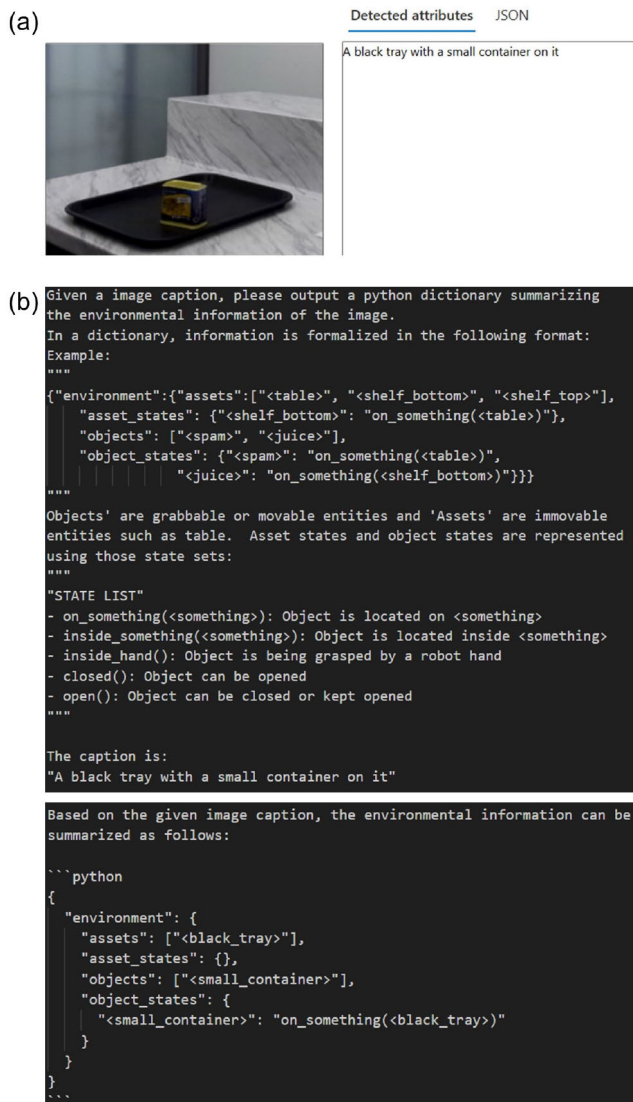


FIGURE 18. An example of leveraging an image description model and ChatGPT to obtain environmental information from a scene. (a) The employment of a commercially-available image description model [38] to gain symbolic comprehension of the environment. (b) Utilizing ChatGPT to format the information derived from the image description.

system—including how the output of ChatGPT are specifically translated into robot actions that are quantitatively controlled, how the system handles errors or unanticipated situations, and the timing for user feedback within the overall system—are provided in Appendix C.

V. METHODOLOGICAL CONSIDERATIONS

A. TOKEN LIMIT

Our proposed prompts aim to estimate the post-operation environment as a hint for subsequent task planning. This approach can alleviate the impact of the token limit imposed on ChatGPT, as it reduces the burden of maintaining lengthy conversation histories for multi-step instructions. However, the issue of the token limit is not completely eliminated, and it might affect the scalability of the system.

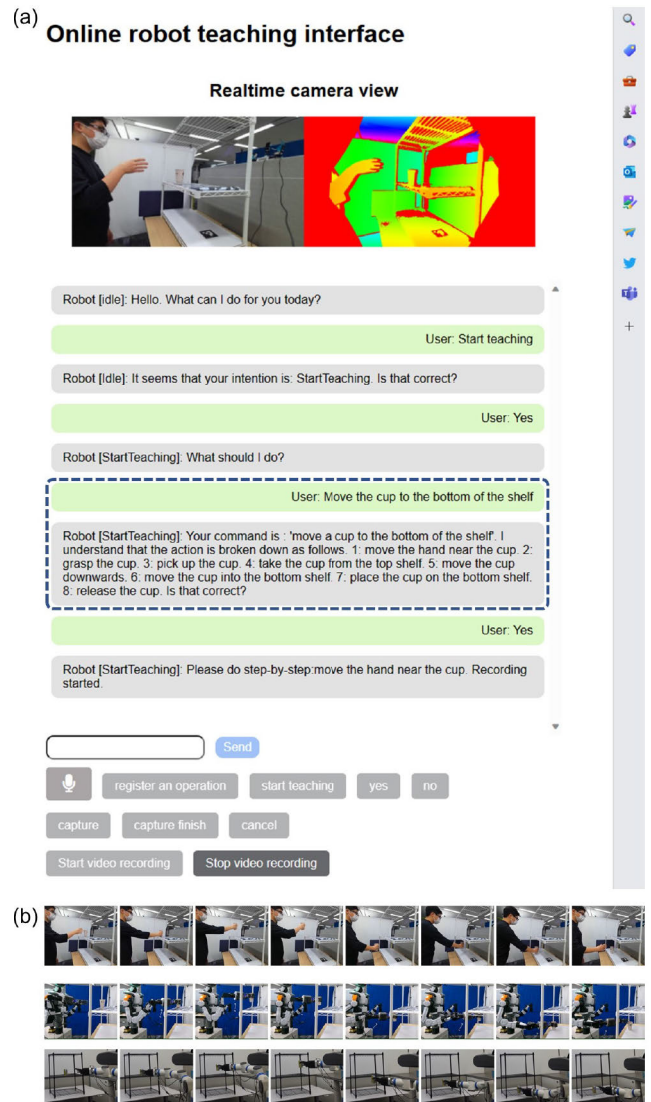


FIGURE 19. An example of integrating the proposed ChatGPT-empowered task planner into a robot teaching system. (a) A teaching system that incorporates the task planner (indicated by the dashed box). Following task planning, the system asks the user to visually demonstrate the tasks in a step-by-step manner. “How-to-do” parameters are then extracted from this visual demonstration. (b) (Top) The step-by-step demonstration corresponding to the planned tasks. (Middle and Bottom) Execution of the tasks by two different types of robot hardware.

For example, if many actions need to be explained in detail, or if an environmental description becomes lengthy as the result of including information about numerous objects, the prompts may use a significant portion of the total available tokens for the entire conversation. In such cases, one might need to accept the system limitations imposed by the token limit and adapt a strategy accordingly. This could involve simplifying the environmental descriptions or reducing the number of defined actions, in accordance with the specific scenarios being addressed (e.g., kitchen tasks, bedroom tasks, or factory assembly tasks).

Furthermore, if a long action step is anticipated from an instruction, the need for user feedback in adjusting the output

of ChatGPT may lead to increased consumption of available tokens in the conversation. In such situations, truncation of the conversation to accommodate the token limit could result in the loss of human intent included in earlier feedback (see Fig. 8). This means that the token limit imposes certain restrictions on the length of actions that can be taught in a single instruction. If an instruction is likely to result in a long action sequence, the instruction may need to be broken down into smaller segments. This could help the task planning of ChatGPT, and thus reduce the amount of required feedback.

B. OPTIMAL PROMPTING STRATEGIES

Through our experiment with VirtualHome, we identified two failure patterns in ChatGPT: incorrect verb selection and omission of necessary steps.

Incorrect verb selection may be partially attributable to the naming conventions used for actions. In the experiment, we adhered to the original action names in VirtualHome, such as “Put” (an action of placing an object on another object) and “PutIn” (an action of placing an object inside a container with a door, such as a microwave). While these terms denote distinct actions, their similar names could potentially lead to confusion. To verify this hypothesis, we conducted a follow-up experiment where we renamed the actions to “PutSurface” and “PutContainerWithDoor,” respectively, to reflect their definitions more accurately. This modification led to a reduction in this type of error (data not shown²), underscoring the importance of precise action naming when instructing ChatGPT—a finding that aligns with prior research [30].

Omission of necessary steps may partially stem from ChatGPT’s difficulty in interpreting the granularity of defined actions. The granularity of actions implied by language is often ambiguous. For example, an instruction such as “Put food in the microwave” could be perceived either as a single action or a sequence of more detailed actions, such as “open the microwave, put food in it, close it.” Despite having provided action definitions in the prompt, the inherent ambiguity in language may lead to the omission of necessary actions in task planning. To address this failure pattern, providing more examples could effectively guide ChatGPT to decompose intended actions at the desired level of granularity. Although our initial experiments with VirtualHome included one pick-and-place example in the prompt, a follow-up experiment confirmed that including an example of placing food in a microwave reduced the occurrence of step omission—specifically, the omission of opening and closing actions (data not shown²).

C. VARIATIONS IN THE EXPRESSION OF INSTRUCTIONS

In our experiments, the instructions used were relatively explicit expressions that directly specified the actions to be performed. While the use of LLMs in processing such

²The results of the follow-up experiments can be found here: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>

expressions might not yield significantly greater benefits compared to conventional machine learning techniques (e.g., [51]) for text processing, one advantage of LLMs is suggested to lie in their ability to handle high-level texts [52], [53]. Therefore, as a follow-up experiment, we adjusted our instructions to focus on the desired outcomes and objectives of the operation (e.g., “Let’s watch TV.” instead of “Turn on the TV.”, see Table 5 in Appendix B for details). As a result, we observed a performance level consistent with that of the original instructions (Table 6 in Appendix B). Additionally, we tested the task planner with various instructions that contained similar intent but were worded differently for a given scenario (e.g., “Take the book from the table and put it on the bookshelf.” and “Grab the book from the table and place it on the bookshelf.”), and confirmed that the performance level remained consistent across variations in instruction (data not shown²).

These results do not imply that our task planner can accommodate any forms of variation in expression, yet suggest its effectiveness to a reasonable extent. Nevertheless, we reiterate that the robustness and soundness of the proposed task planning can be supported more by the functionality allowing for necessary adjustments through user feedback, rather than the performance of single-shot task planning.

D. DATA PRIVACY AND SECURITY

In response to emerging concerns regarding data privacy and security, careful data handling is crucial for systems that rely on LLMs. To address this issue, we operate ChatGPT via Azure OpenAI, a service provided by Microsoft. This service enables us to handle data in compliance with various legal regulations and standards related to data security and privacy, ensuring the proper protection of user instructions and information [54]. Furthermore, Azure OpenAI includes features for content filtering and abuse monitoring, which aid in mitigating risks associated with misuse. Thus, we believe that our task planning system can operate while meeting industry standards for data privacy and security. However, it is essential for future work to continually assess and improve these protective measures as our understanding of potential risks evolves.

E. FUTURE DIRECTIONS

Among the pioneering studies for task planning from natural language, a significant advantage of utilizing the most recent LLMs is their adaptability to various operational settings through few-shot learning and user feedback. These functionalities not only remove the need for extensive data collection or model retraining but also enable user adjustments, thereby facilitating safe and robust task planning.

While we use ChatGPT as an example of such an LLM, these capabilities are not confined to any specific model. The ability to perform few-shot learning is considered a result of increased model sizes [1] and extended training on large datasets [55]. Furthermore, the capacity to effectively

TABLE 4. The action list defined for the experiment in Section III-C.

| Action | Description |
|-------------------|---|
| Walktowards(arg1) | Walks some distance towards a room or object. |
| Grab(arg1) | Grabs an object. |
| Open(arg1) | Opens an object. |
| Close(arg1) | Closes an object. |
| Put(arg1, arg2) | Puts an object on another object. |
| PutIn(arg1, arg2) | Puts an object inside another container. |
| SwitchOn(arg1) | Turns an object on. |
| SwitchOff(arg1) | Turns an object off. |
| Drink(arg1) | Drinks from an object. |

accommodate user feedback could be partially attributed to learning methods that align model behavior with human intent, known as reinforcement learning from human feedback [56]. In fact, other models that utilize similar training techniques, such as GPT-4 [57] and Llama2-chat [58], have been reported to possess these features. Future research will explore whether other models can yield results comparable to those found in this study when applied to task planning.

Regarding the adjustment capability, our experiments suggested that the output of ChatGPT can be adjusted through a reasonable amount of feedback. ChatGPT's ability to reflect the semantic content of user feedback provides a means for users to convey their intentions to the system. Thus, we consider that this aspect contributes to the foundation of a user-friendly system. However, this study did not delve into how this adjustment capability directly contributes to user-friendliness of the system. Future research areas include user studies focusing on usability and comparisons with other adjustment methods, such as directly editing the output of ChatGPT.

VI. CONCLUSION

This paper presented a practical application of OpenAI's ChatGPT for translating multi-step instructions into executable robot actions. We designed input prompts to meet the common requirements in practical applications, specifically encouraging ChatGPT to output a sequence of robot actions in a readable format and explicitly handle the environmental information before and after executing the actions. Through experiments, we tested the effectiveness of our proposed prompts in various environments. Additionally, we observed that the conversational capability of ChatGPT allows users to adjust the output through natural-language feedback, which is crucial for safe and robust task planning. Our prompts and source code are open source and publicly available. We hope that this study will provide practical resources to the robotics research community and inspire further developments in this research area.

APPENDIX A DETAILED PROMPT FOR DEFINING ROBOT ACTIONS

Fig. 20 provides the unabridged prompt that is exemplified in Section II-B. It includes the list of robot actions and their definitions.

```

Necessary and sufficient robot actions are defined as follows:
""""
"ROBOT ACTION LIST"
- move_hand(): Move the robot hand from one position to another with/without grasping an object.
- grasp_object(): Grab an object.
- release_object(): Release an object in the robot hand.
- move_object(): Move the object grabbed by the robot hand from one position to another. move_object() is allowed only when the object is not physically constrained by the environment. For example, if the robot hand moves an object on the floor to another location, move_object() is not allowed because the object is constrained by the floor.
- detach_from_plane(): This action can only be performed if an object is grabbed. Move the grabbed object from a state in which it is constrained by a plane to a state in which it is not constrained by any plane. For example, detach_from_plane() is used when a robot hand picks up an object on a table.
- attach_to_plane(): This action can only be performed if an object is grabbed. The opposite operation of detach_from_plane().
- open_by_rotate(): This action can only be performed if an object is grabbed. Open something by rotating an object that is rotationally constrained by its environment along its rotation. For example, when opening a refrigerator, the refrigerator handle makes this motion. Also, when opening the lid of a plastic bottle, the lid makes this motion.
- adjust_by_rotate(): This action can only be performed if an object is grabbed. Rotate an object that is rotationally constrained by its environment along its rotation. For example, when adjusting the temperature of a refrigerator, the temperature knob makes this motion.
- close_by_rotate(): This action can only be performed if an object is grabbed. The opposite operation of open_by_rotate().
- open_by_slide(): This action can only be performed if an object is grabbed. Moves an object that is translationally constrained in two axes from its environment along one unbounded axis. For example, when opening a sliding door or drawer, the handle makes this movement.
- adjust_by_slide(): This action can only be performed if an object is grabbed. Slide an object that is translationally constrained in two axes from its environment along one unbounded axis. For example, when widening the gap between a sliding door and the wall, the handle makes this movement.
- close_by_slide(): This action can only be performed if an object is grabbed. The opposite operation of open_by_slide().
- wipe_on_plane(): This action can only be performed if an object is grabbed. Move an object landing on a plane along two axes along that plane. For example, when wiping a window with a sponge, the sponge makes this motion.
""""

```

FIGURE 20. The prompt explaining the robotic functions.

APPENDIX B SUPPLEMENTARY INFORMATION FOR VIRTUALHOME EXPERIMENT

This section provides supplementary information for the VirtualHome experiment discussed in Section III-C. Table 4 displays a list of pre-defined atomic actions in VirtualHome, which represent the smallest units of action. Table 5 illustrates fourteen scenarios used for the experiment. The "Textual instruction" column indicates instructions that were fed into the task planner. The "Action sequence" column shows the manually identified action sequences to achieve the scenarios. The "Higher-level textual instruction" column displays instructions that have been adjusted to emphasize desired outcomes and objectives of the operation. These instructions were used in a follow-up experiment that tested ChatGPT's ability to understand and respond to higher-level texts (See Section V-C). Table 6 shows the results.

APPENDIX C AN EXAMPLE OF A ROBOT SYSTEM EXPANDING THE PROPOSED TASK PLANNER

In this supplementary section, we present an overview of our in-house robot teaching system, which serves as an

TABLE 5. The list of scenarios and their action sequences used in the experiment.

| Scenario | Textual instruction | Action sequence | Higher-level textual instruction |
|------------|--|---|--|
| Scenario1 | Take the bread from the toaster on the kitchen counter and place it on the plate on the table. | WalkTowards(toaster),
Grab(breadslice),
WalkTowards(kitchentable),
Put(breadslice, plate) | Serve the toast on the table. |
| Scenario2 | Take the frying pan from the counter and place it in the sink. | WalkTowards(stove),
Grab(fryingpan),
WalkTowards(sink),
Put(fryingpan, sink) | Put away the frying pan into the sink. |
| Scenario3 | Take the pie from the table and put it in the microwave. | WalkTowards(kitchentable),
Grab(pie),
WalkTowards(microwave),
Open(microwave),
Putin(pie, microwave),
Close(microwave),
SwitchOn(microwave) | Heat up the pie using the microwave. |
| Scenario4 | Take the condiment shaker from the bookshelf and place it on the table. | WalkTowards(bookshelf),
Grab(condimentshaker),
WalkTowards(kitchentable),
Put(condimentshaker, kitchentable) | Set out the condiment shaker on the table. |
| Scenario5 | Take the book from the table and put it on the bookshelf. | WalkTowards(kitchentable),
Grab(book),
WalkTowards(bookshelf),
Put(book, bookshelf) | Store the book on the shelf. |
| Scenario6 | Take the water glass from the table and drink from it. | WalkTowards(kitchentable),
Grab(waterglass),
Drink(waterglass),
Put(waterglass, kitchentable) | Drink the water from the glass. |
| Scenario7 | Take the salmon on top of the microwave and put it in the fridge. | WalkTowards(microwave),
Grab(salmon),
WalkTowards(fridge),
Open(fridge),
Putin(salmon, fridge),
Close(fridge) | Chill the salmon in the fridge. |
| Scenario8 | Turn on the TV. | WalkTowards(tvstand),
SwitchOn(tv) | Let's watch TV. |
| Scenario9 | Put a plate that is on the table into the sink. | WalkTowards(kitchentable),
Grab(plate),
WalkTowards(sink),
Put(plate, sink) | Clear away that plate into the sink. |
| Scenario10 | Take the pie on the table and warm it using the stove. | WalkTowards(kitchentable),
Grab(pie),
WalkTowards(stove),
Open(stove),
Putin(pie, stove),
Close(stove)
SwitchOn(stove) | Warm the pie using the stove. |
| Scenario11 | Put the sponge in the sink and wet it with water. | WalkTowards(kitchencounter),
Grab(washingsponge),
WalkTowards(sink),
Put(washingsponge, sink),
SwitchOn(faucet) | Wet the sponge. |
| Scenario12 | Take the knife from the table and move it to another place on the table. | WalkTowards(kitchentable),
Grab(cutleryknife),
WalkTowards(kitchentable),
Put(cutleryknife, kitchentable) | Find another place for the knife. |
| Scenario13 | Take the plate from the table and move it to another place on the table. | WalkTowards(kitchentable),
Grab(plate),
WalkTowards(kitchentable),
Put(plate, kitchentable) | Reposition that plate. |
| Scenario14 | Take the condiment bottle from the bookshelf and put it on the table. | WalkTowards(bookshelf),
Grab(condimentbottle),
WalkTowards(kitchentable),
Put(condimentbottle, kitchentable) | Place the condiment bottle on the table. |

illustrative example of the proposed task planner. Importantly, our system is designed for use under the guidance of experts familiar with robot operations and action definitions,

rather than being an automatic solution for non-experts. The system's main objective is to simplify the robot teaching process, eliminating the need for complex coding

TABLE 6. Executability of the output action sequence across trials (Higher-level textual instruction). "1" indicates success, and "0" indicates failure.

| Scenario | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|----------|---|---|---|---|---|---|---|---|---|----|----|----|----|----|
| Trial 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| Trial 2 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Trial 3 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Trial 4 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| Trial 5 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |

TABLE 7. Robot actions and the parameters obtained from visual demonstrations and during on-site robot execution.

| Robot action | Parameters obtained from corresponding visual demonstrations (RGB-D images) | Parameters obtained during on-site robot execution (RGB-D images and force sensors) |
|---|---|---|
| move_hand() | <ul style="list-style-type: none"> The 3D positions of the manipulating hand at the first and the last video frames. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | NA |
| grasp_object() | <ul style="list-style-type: none"> The 3D position of the object. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. The approach direction of the hand to the object. The grasp type according to a grasping taxonomy. | <ul style="list-style-type: none"> The 3D position of the contact points where the robot's end effectors should engage, as determined by the grasp type. The value of the force sensor attached to the robot's end effectors. |
| release_object() | <ul style="list-style-type: none"> The retreat direction of the hand from the object. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | NA |
| move_object() | <ul style="list-style-type: none"> The 3D positions of the manipulating hand at the first and the last video frames. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | NA |
| detach_from_plane() | <ul style="list-style-type: none"> The displacement when the object in hand detaches from the plane. The 3D position of the manipulating hand at the last video frame. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | <ul style="list-style-type: none"> The value of the force sensor attached to the wrist area near the robot's end effector. |
| attach_to_plane() | <ul style="list-style-type: none"> The displacement when the object in hand attaches to the plane. The 3D position of the manipulating hand at the first video frame. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | <ul style="list-style-type: none"> The value of the force sensor attached to the wrist area near the robot's end effector. |
| open_by_rotate(),
adjust_by_rotate(),
close_by_rotate() | <ul style="list-style-type: none"> The 3D positions of the manipulating hand across all the video frames. The direction of the rotation axis. The position of the rotation center. The angle rotation. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | <ul style="list-style-type: none"> The value of the force sensor attached to the wrist area near the robot's end effector. |
| open_by_slide(),
adjust_by_slide(),
close_by_slide() | <ul style="list-style-type: none"> The 3D positions of the manipulating hand across all the video frames. The displacement of the sliding motion. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | <ul style="list-style-type: none"> The value of the force sensor attached to the wrist area near the robot's end effector. |
| wipe_on_plane() | <ul style="list-style-type: none"> The 3D positions of the manipulating hand across all the video frames. The axis that is vertical to the wiping plane. Whether the manipulating hand is left or right. The arm postures at the first and the last video frames. | <ul style="list-style-type: none"> The value of the force sensor attached to the wrist area near the robot's end effector. |

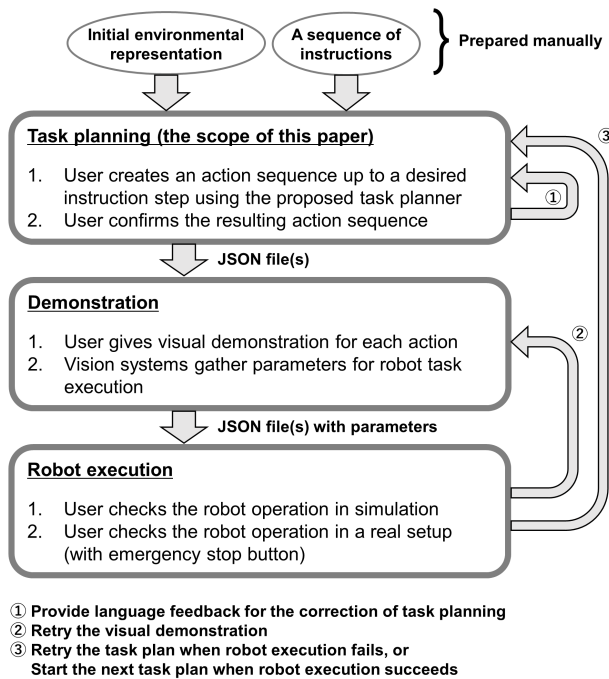


FIGURE 21. Overview of the robot teaching system that integrates the proposed task planner. The process involves three main steps: **Task planning**, where the user employs the task planner to create an action sequence and adjusts the result through feedback as necessary; **Demonstration**, where the user visually demonstrates the action sequence to provide information needed for robot operation; and **Robot Execution**, where the action sequence is first simulated and then tested. If any step fails or shows deficiencies, the previous steps can be revisited as necessary.

by incorporating a method of robot programming that uses multimodal demonstrations and language feedback.

A. PREREQUISITE

The representation of the initial environment and the sequence of instructions are assumed to have been prepared manually.

The robot teaching system operates through the following three steps (Fig. 21):

- 1) **Task planning (the scope of this paper):** The user creates a task plan up to a desired instruction step using the proposed task planner. If any deficiencies are found in the output sequence, the user can provide feedback to the task planner as necessary.
- 2) **Demonstration:** The user visually demonstrates the action sequence to provide information needed for robot operation. Specifically, the system asks the user to demonstrate each task step-by-step in front of an RGB-D camera. The vision system then analyzes the visual demonstration and extracts the parameters needed for the robot to perform each task.
- 3) **Robot execution:** The user first simulates the action sequence and checks the results. The simulation environment is designed to replicate the actual one [50]. If execution fails or leads to an unexpected result, the task planning and demonstration steps are revisited as

necessary. Only when safe operation is confirmed in the simulation does the user test the action sequence in the real setup. For safety, the robot operation is tested under a condition where the user can press the robot's emergency stop switch at any time.

Examples of the parameters required for the robot to execute each task are provided in Table 7. In step 2, the vision system identifies the parameters listed in the table's second column by utilizing third-party pose recognizers and object recognizers. For example, parameters such as the center of rotation, rotation axis, and rotation radius are estimated from the hand's trajectory. To represent arm postures, we have prepared 26 unit vectors that indicate 3D directions. These vectors are used to represent the discrete orientation of upper and lower arms. Specifically, we choose the vector closest to the direction in which each part of the arm is pointing, using it to represent that part's direction [32]. More detailed methods for acquiring parameters are explained in other papers [32], [42].

In Step 3, besides the aforementioned parameters, the robot system controls the robot using data from an RGB-D camera and force sensors mounted on it (see the third column of Table 7). It is assumed that the environment at the start of the demonstration and execution will be identical to the extent that neither the action sequence nor the discrete representation of the posture would be affected. The robot's vision system looks for objects again during execution and corrects slight misalignments. For some tasks, the value of the force sensors attached to and near the end effector is used as force feedback.

The robot system computes the robot's physical movements for tasks such as `move_hand()`, `move_object()`, and `release_object()` by using inverse kinematics with postural constraint [43], following the parameters recognized by the vision system. For other tasks, the movements are computed by pre-trained reinforcement learning policies [48], [49]. Notably, the post-operation environment output by ChatGPT is used only as a hint for subsequent task planning but is not used during robot execution in our robot system. It is worth reiterating that task definitions and execution methods differ depending on the design philosophy, and this table merely illustrates one example of the implementation.

ACKNOWLEDGMENT

This study was conceptualized, conducted, and written by the authors, and an AI (OpenAI's GPT-4 model) was used for proofreading.

REFERENCES

- [1] T. B. Brown et al., "Language models are few-shot learners," in *Proc. Adv. Neur. Inf. Process. Sys.*, vol. 33, 2020, pp. 1877–1901.
- [2] OpenAI. *ChatGPT*. Accessed: May 8, 2023. [Online]. Available: <https://openai.com/blog/chatgpt>
- [3] P. Pramanick, H. B. Barua, and C. Sarkar, "DeComplex: Task planning from complex natural instructions by a collocating robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 6894–6901.
- [4] S. G. Venkatesh, R. Upadrashta, and B. Amrutur, "Translating natural language instructions to computer programs for robot manipulation,"

- in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 1919–1926.
- [5] I. Yanaokura, N. Wake, K. Sasabuchi, R. Arakawa, K. Okada, J. Takamatsu, M. Inaba, and K. Ikeuchi, “A multimodal learning-from-observation towards all-at-once robot teaching using task cohesion,” in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 367–374.
 - [6] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, “VIMA: General robot manipulation with multimodal prompts,” 2022, *arXiv:2210.03094*.
 - [7] M. Shridhar, L. Manuelli, and D. Fox, “Perceiver-actor: A multi-task transformer for robotic manipulation,” in *Proc. Conf. Robot Learn.*, 2023, pp. 785–799.
 - [8] A. Brohan et al., “Do as I can, not as I say: Grounding language in robotic affordances,” in *Proc. Conf. Robot Learn.*, 2023, pp. 287–318.
 - [9] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Sermanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter, “Inner monologue: Embodied reasoning through planning with language models,” 2022, *arXiv:2207.05608*.
 - [10] Y. Ding, X. Zhang, C. Paxton, and S. Zhang, “Task and motion planning with large language models for object rearrangement,” 2023, *arXiv:2303.06247*.
 - [11] 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,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11523–11530.
 - [12] N. K. H. Singh, V. Bindal, A. Tuli, V. Agrawal, R. Jain, P. Singla, and R. Paul, “Learning neuro-symbolic programs for language guided robot manipulation,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 7973–7980.
 - [13] Z. Zhao, W. S. Lee, and D. Hsu, “Differentiable parsing and visual grounding of natural language instructions for object placement,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11546–11553.
 - [14] Y. Ding, X. Zhang, S. Amiri, N. Cao, H. Yang, C. Esselink, and S. Zhang, “Robot task planning and situation handling in open worlds,” 2022, *arXiv:2210.01287*.
 - [15] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, and P. Florence, “Socratic models: Composing zero-shot multimodal reasoning with language,” 2022, *arXiv:2204.00598*.
 - [16] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, “Code as policies: Language model programs for embodied control,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 9493–9500.
 - [17] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, “Planning with large language models via corrective re-prompting,” 2022, *arXiv:2211.09935*.
 - [18] Y. Xie, C. Yu, T. Zhu, J. Bai, Z. Gong, and H. Soh, “Translating natural language to planning goals with large-language models,” 2023, *arXiv:2302.05128*.
 - [19] A. K. Kovalev and A. I. Panov, “Application of pretrained large language models in embodied artificial intelligence,” *Doklady Math.*, vol. 106, no. S1, pp. S85–S90, Dec. 2022.
 - [20] M. A. Khan, M. Kenney, J. Painter, D. Kamale, R. Batista-Navarro, and A. Ghalamzan, “Natural language robot programming: NLP integrated with autonomous robotic grasping,” 2023, *arXiv:2304.02993*.
 - [21] F. Kaynar, S. Rajagopalan, S. Zhou, and E. Steinbach, “Remote task-oriented grasp area teaching by non-experts through interactive segmentation and few-shot learning,” 2023, *arXiv:2303.10195*.
 - [22] M. Skreta, N. Yoshikawa, S. Arellano-Rubach, Z. Ji, L. Bjørn Kristensen, K. Darvish, A. Aspuru-Guzik, F. Shkurti, and A. Garg, “Errors are useful prompts: Instruction guided task programming with verifier-assisted iterative prompting,” 2023, *arXiv:2303.14100*.
 - [23] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents,” in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 9118–9147.
 - [24] C. Lynch and P. Sermanet, “Language conditioned imitation learning over unstructured data,” 2020, *arXiv:2005.07648*.
 - [25] J. Pan, G. Chou, and D. Berenson, “Data-efficient learning of natural language to linear temporal logic translators for robot task specification,” 2023, *arXiv:2303.08006*.
 - [26] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, “Text2Motion: From natural language instructions to feasible plans,” 2023, *arXiv:2303.12153*.
 - [27] C. Zhao, S. Yuan, C. Jiang, J. Cai, H. Yu, M. Y. Wang, and Q. Chen, “ERRA: An embodied representation and reasoning architecture for long-horizon language-conditioned manipulation tasks,” *IEEE Robot. Autom. Lett.*, vol. 8, no. 6, pp. 3230–3237, Jun. 2023.
 - [28] H. Liu, L. Lee, K. Lee, and P. Abbeel, “Instruction-following agents with multimodal transformer,” 2022, *arXiv:2210.13431*.
 - [29] O. Mees, J. Borja-Diaz, and W. Burgard, “Grounding language with visual affordances over unstructured data,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11576–11582.
 - [30] S. Vemprala, R. Bonatti, A. Buckner, and A. Kapoor, “ChatGPT for robotics: Design principles and model abilities,” *Microsoft Auton. Syst. Robot. Res.*, vol. 2, p. 20, Feb. 2023.
 - [31] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, “VirtualHome: Simulating household activities via programs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8494–8502.
 - [32] N. Wake, R. Arakawa, I. Yanokura, T. Kiyokawa, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, “A learning-from-observation framework: One-shot robot teaching for grasp-manipulation-release household operations,” in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2021, pp. 461–466.
 - [33] K. Ikeuchi, N. Wake, R. Arakawa, K. Sasabuchi, and J. Takamatsu, “Semantic constraints to represent common sense required in household actions for multi-modal learning-from-observation robot,” 2021, *arXiv:2103.02201*.
 - [34] H. T. Kuhn and W. L. Inequalities, “Related systems,” in *Annals of Mathematics Studies*. Princeton, NJ, USA: Princeton Univ. Press, 1956.
 - [35] M. Chen et al., “Evaluating large language models trained on code,” 2021, *arXiv:2107.03374*.
 - [36] M. L. Minsky, “Minsky’s frame system theory,” in *Proc. Workshop Theor. Issues Natural Lang. Process.*, 1975, pp. 104–116.
 - [37] M. Gramopadhye and D. Szafir, “Generating executable action plans with environmentally-aware language models,” 2022, *arXiv:2210.04964*.
 - [38] Microsoft. *Microsoft Azure*. Accessed: Aug. 5, 2023. [Online]. Available: <https://azure.microsoft.com/>
 - [39] N. Wake, K. Sasabuchi, and K. Ikeuchi, “Grasp-type recognition leveraging object affordance,” in *Proc. HOBI-RO-MAN Workshop*, 2020, pp. 1–10.
 - [40] N. Wake, D. Saito, K. Sasabuchi, H. Koike, and K. Ikeuchi, “Text-driven object affordance for guiding grasp-type recognition in multimodal robot teaching,” *Mach. Vis. Appl.*, vol. 34, no. 4, p. 58, Jul. 2023.
 - [41] D. Saito, N. Wake, K. Sasabuchi, H. Koike, and K. Ikeuchi, “Contact web status presentation for freehand grasping in MR-based robot-teaching,” in *Proc. Companion ACM/IEEE Int. Conf. Hum.-Robot Interact.*, Mar. 2021, pp. 167–171.
 - [42] N. Wake, I. Yanokura, K. Sasabuchi, and K. Ikeuchi, “Verbal focus-of-attention system for learning-from-observation,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2021, pp. 10377–10384.
 - [43] K. Sasabuchi, N. Wake, and K. Ikeuchi, “Task-oriented motion mapping on robots of various configuration using body role division,” *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 413–420, Apr. 2021.
 - [44] K. Ikeuchi, J. Takamatsu, K. Sasabuchi, N. Wake, and A. Kanehira, “Applying learning-from-observation to household service robots: Three common-sense formulations,” 2023, *arXiv:2304.09966*.
 - [45] N. Wake, M. Fukumoto, H. Takahashi, and K. Ikeuchi, “Enhancing listening capability of humanoid robot by reduction of stationary ego-noise,” *IEEJ Trans. Electr. Electron. Eng.*, vol. 14, no. 12, pp. 1815–1822, Dec. 2019.
 - [46] J. Jaroslavecva, N. Wake, K. Sasabuchi, and K. Ikeuchi, “Robot ego-noise suppression with labanotation-template subtraction,” *IEEJ Trans. Electr. Electron. Eng.*, vol. 17, no. 3, pp. 407–415, Mar. 2022.
 - [47] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, “Interactive task encoding system for learning-from-observation,” in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatronics (AIM)*, Jun. 2023, pp. 1061–1066.
 - [48] J. Takamatsu, K. Sasabuchi, N. Wake, A. Kanehira, and K. Ikeuchi, “Learning-from-observation system considering hardware-level reusability,” 2022, *arXiv:2212.09242*.

- [49] D. Saito, K. Sasabuchi, N. Wake, J. Takamatsu, H. Koike, and K. Ikeuchi, "Task-grasping from a demonstrated human strategy," in *Proc. IEEE-RAS 21st Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2022, pp. 880–887.
- [50] K. Sasabuchi, D. Saito, A. Kanehira, N. Wake, J. Takamatsu, and K. Ikeuchi, "Task-sequencing simulator: Integrated machine learning to execution simulation for robot manipulation," 2023, *arXiv:2301.01382*.
- [51] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annu. Rev. Control, Robot., Auto. Syst.*, vol. 3, pp. 25–55, Jan. 2020.
- [52] Y. Ye, H. You, and J. Du, "Improved trust in human–robot collaboration with ChatGPT," *IEEE Access*, vol. 11, pp. 55748–55754, 2023.
- [53] Z. G. Cai, D. A. Haslett, X. Duan, S. Wang, and M. J. Pickering, "Does ChatGPT resemble humans in language use?" 2023, *arXiv:2303.08014*.
- [54] Microsoft. *Azure OpenAI—Data Privacy*. Accessed: Aug. 5, 2023. [Online]. Available: <https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy>
- [55] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.
- [56] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," 2022 *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 27730–27744, 2022.
- [57] OpenAI. *GPT-4*. Accessed: Aug. 5, 2023. [Online]. Available: <https://openai.com/research/gpt-4>
- [58] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.



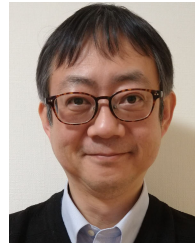
NAOKI WAKE received the B.S. degree in engineering and the Ph.D. degree in information science and technology from The University of Tokyo, in 2014 and 2019, respectively. His research interests include auditory neuroscience, neuro-rehabilitation, and spoken language processing.



ATSUSHI KANEHIRA received the B.S. degree in engineering and the Ph.D. degree in information science and technology from The University of Tokyo, in 2014 and 2019, respectively. His research interests include computer vision and machine learning.



KAZUHIRO SASABUCHI received the Ph.D. degree from the Information Science and Technology Department, The University of Tokyo, in 2019. He is currently a Researcher with Microsoft. He has worked on robot competitions, systems, teaching, and human–robot interaction.



JUN TAKAMATSU (Member, IEEE) received the Ph.D. degree in computer science from The University of Tokyo, Japan, in 2004. From 2004 to 2008, he was with the Institute of Industrial Science, The University of Tokyo. In 2007, he was a Visiting Researcher with Microsoft Research Asia. From 2008 to 2020, he was an Associate Professor with the Nara Institute of Science and Technology, Japan. He was a Visitor with Carnegie Mellon University, in 2012 and 2013, and a Visiting Scientist at Microsoft, in 2018. He is currently a Senior Researcher with Applied Robotics Research, Microsoft. His research interests include robotics, including learning-from-observation, task and motion planning, and physics-based vision.



KATSUSHI IKEUCHI (Life Fellow, IEEE) received the B.E. degree in mechanical engineering from Kyoto University, in 1973, and the Ph.D. degree in information engineering from The University of Tokyo, in 1978.

After working with the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, for three years, the Electrotechnical Laboratory, Ministry of International Trade and Industry, Japanese Government, for five years, and the Robotics Institute, Carnegie Mellon University, for ten years, and the Institute of Industrial Science, The University of Tokyo, for 19 years, he joined Microsoft as a Principal Researcher, in 2015. During this tenure at CMU and UTokyo, he has supervised more than 50 Ph.D. students. His research interests include computer vision, robotics, and computer graphics.

Dr. Ikeuchi was awarded a (Life) Fellow from IEICE, IPSJ, and RSJ, through these research and society service. In his research fields, he received several best paper awards, including the David Marr Prize in computational vision and the IEEE Robotics and Automation Society K. S. Fu Memorial Best Transaction Paper. He received the Distinguished Researcher Award from IEEE-PAMI, the Medal of Honor with Purple Ribbon from Japanese Emperor, the Okawa Prize from Okawa Foundation, and a couple of research achievement awards from Japanese professional societies. His community service includes: the General Chair of IROS95, ITSC99, IV01, ICCV05, ACCV07, and ICCV17; the Program Chair of CVPR96, ICCV03, ICRA09, ICPR12, and ICCV15; the Editor-in-Chief of IJCV, from 2000 to 2017 and IJ ITS, from 2012 to 2014, an Associate Editor of IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION and IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE; and a Distinguished Lecturer of the IEEE Signal Processing Society, from 2000 to 2002, the IEEE Robotics and Automation Society, from 2004 to 2006, and the IEEE Computer Society, from 2008 to 2010.

• • •