

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

Naoki Wake¹, Atsushi Kanehira¹, Kazuhiro Sasabuchi¹, Jun Takamatsu¹, and Katsushi Ikeuchi¹

Abstract

This paper aims to provide a specific example of how OpenAI's ChatGPT can be used in a few-shot setting to convert natural language instructions into a sequence of executable robot actions (Fig. 1). Generating programs for robots from natural language instructions is an attractive goal, but the practical application using ChatGPT is still in its early stages, and there is no established methodology yet. Here, we have designed easy-to-customize input prompts for ChatGPT that meet common requirements in many practical applications, including: 1) easy integration with robot execution systems or visual recognition programs, 2) applicability to various environments, and 3) the ability to provide long-step instructions while minimizing the impact of ChatGPT's token limit. Specifically, the prompts encourage ChatGPT to 1) output a sequence of predefined robot actions with explanations in a readable JSON format, 2) represent the operating environment in a formalized style, and 3) infer and output the updated state of the operating environment as the result of each operation, which will be input with the next instruction to allow ChatGPT to work based solely on the memory of the latest operations. Through experiments, we confirmed that the proposed prompts allow ChatGPT to act in accordance with the requirements in various environments. Additionally, we observed that ChatGPT's conversational ability allows users to adjust its output with natural language feedback, which is crucial for developing an application that is both safe and robust while providing a user-friendly interface. Users can easily customize the prompts as templates. The contribution of this paper is to provide and publish the prompts, which are generic enough to be easily modified to fit the requirements of each experimenter, thereby providing practical knowledge to the robotics research community. Our prompts and source code for using them are open-source and publicly available at <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

Muti-step instructions in various environments



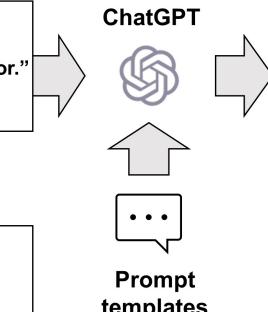
1. “Put the juice on top of the shelf.”
2. “Throw away the spam can.”
3. “Move the juice on the table.”
4. “Throw away the juice.”
5. ...



1. “Open the fridge.”
2. “Open the fridge wider.”
3. “Take a juice and put it on the floor.”
4. “Close the fridge.”
5. ...



1. “Open the drawer.”
2. “Close a drawer a little bit.”
3. “Close the drawer.”
4. ...



Readable sequence of object manipulation (JSON files)

```
{
  "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"
}
```

Fig. 1. This paper shows practical prompts for ChatGPT to generate for translating a sequences of executable robot actions from multi-step human instructions in various environments.

I. INTRODUCTION

Recent advances in natural language processing have enabled the development of large language models (LLMs) that can understand and generate human-like language. As a result of learning vast amount of data, some LLMs can be fine-tuned for specific tasks in a few-shot manner through conversations. ChatGPT [1] is a prime example of such an LLM. One exciting

¹Applied Robotics Research, Microsoft, Redmond, WA 98052, USA naoki.wake@microsoft.com

application of ChatGPT is in the field of robotics, where it can be used to convert natural language instructions into executable programs for controlling robots.

Generating programs for robots, or task plans, from natural language instructions is an attractive goal, and many existing works exist (for example, [2]–[4]), some of which are built on top of LLMs ([5]–[17]; please refer to [18] for a review). However, most of them were developed within a limited scope of operations such as pick-and-place (for example, [8], [14], [19], [20]), hardware-dependent, or lack the functionality of human-in-the-loop (for example, [9], [13], [15], [21], [22]). Additionally, most of these studies rely on specific datasets (for example, [5]–[7], [11], [12], [23]–[28]), which necessitate data recollection and model retraining when transferring or extending them to other robotic settings. From a practical application standpoint, an ideal robotic solution would be one that can be easily applied to other applications or operational settings without requiring extensive data collection or model retraining.

The advantage of using ChatGPT for robotic applications is that we can utilize its language recognition and interaction capabilities as an interface, and begin with a small amount of sample data to tune up the model for specific applications. While the potential of ChatGPT for robotic application is attracting attention [29], the practical application is still in its early stages, and there is no established methodology yet.

In this paper, we provide a specific example of how ChatGPT can be used in a few-shot setting to convert natural language instructions into a sequence of actions that a robot can execute (Fig.1). In designing the prompts, we tried to ensure that they meet the requirements common to many practical applications while also being structured in a way that they can be easily customizable. The requirements we defined for this paper are:

- 1) Easy integration with robot execution systems or visual recognition programs.
- 2) Applicability to various home environments.
- 3) The ability to provide an arbitrary number of natural language instructions while minimizing the impact of ChatGPT's token limit.

To meet these requirements, we designed input prompts to encourage ChatGPT to:

- 1) Output a sequence of predefined robot actions with explanations in a readable JSON format.
- 2) Represent the operating environment in a formalized style.
- 3) Infer and output the updated state of the operating environment, which can be reused as the next input, allowing ChatGPT to operate based solely on the memory of the latest operations.

We conducted experiments to test the effectiveness of our proposed prompts in inferring appropriate actions for multi-stage language instructions in various environments. Additionally, we observed that ChatGPT's conversational ability allows users to adjust its output with natural language feedback, which is crucial for developing an application that is both safe and robust while providing a user-friendly interface.

The proposed prompts can be used as templates and are easily customizable, including the set of robot operations, representation of environments, and object names. The contribution of this paper is to provide and publish generic prompts that can be easily modified to fit the requirements of each experimenter, thereby providing practical knowledge to the robotics research community. Our prompts and source code for using them are open-source and publicly available at <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

II. CHATGPT PROMPTS

In this section, we introduce the actual prompts that we have designed. 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) specific input from the user.

In the sixth prompt, the user inputs a pair of environment information and language instructions. Note that we considered including these prompts as a single prompt, but it worked more robustly when we input them through a conversation consisting of six turns (see Fig. 2 for how to make ChatGPT process several prompts before the actual operation begins). We describe each prompt below. All the prompts and their output examples are available online, and anyone can try them out through OpenAI's API or a web browser: <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

Please note that the specific prompts are designed under the assumption 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 a user's instructions are 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 will be presented in Section IV.

A. The role of ChatGPT

At the first prompt, we provided ChatGPT with context for this task by explaining the role that ChatGPT should play. The specific details are shown in Fig. 2. To accommodate multiple prompts, we included a sentence instructing ChatGPT to wait for the next prompt before beginning the actual operation.

```
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 robotic actions.
Please do not begin working until I say "Start working." Instead, simply output the message "Waiting for next input."
Understood?
```

Fig. 2. Prompt for explaining the role of ChatGPT.

B. The definition of robot actions

In the next prompt, we define and describe the set of robot actions. The appropriate set of robot actions should depend on the application and the implementation of robotic software, and thus should be customized by experimenters. In the case of our in-house learning-from-observation application [30], [31], we define robot actions as functions that change the motion constraints on the objects being manipulated, based on the Kuhn-Tucker theory [32]. This application enables us to prepare a necessary and sufficient set of robot actions for object manipulation in theory. Among the action set, we defined the most commonly appearing robot actions (e.g., move_hand, pick_up, or rotate) in the prompt as shown in Fig.3. Again, users can add other functions, including non-manipulative functions as needed.

```
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 widen 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.
"""

-----
The texts above are part of the overall instruction. Do not start working yet:
```

Fig. 3. Prompt explaining the robotic functions.

C. Representation of the environments

In the next prompt, we define the representation method for the environment (Fig. 4). We describe the environment by dividing it into non-manipulable obstacles, which we refer to as *assets*, such as shelves and tables, and manipulable objects, which we refer to as *objects*, such as cans and handles. The environment was represented as a dictionary containing the lists of assets, objects, and their states. The states explained the spatial relationships between assets and objects. The environment information presented in this prompt is merely an example, and the actual environment will be provided with text instructions in the sixth prompt. The environment name can be arbitrary, but we define labels that are compatible with general object recognition to facilitate the integration with visual recognition systems. The spatial relationships, represented as the "STATE LIST" in Fig. 4, are described in language that is easy to specify, rather than using numerical values. However, in practice, this level of expression has proven to be sufficient for the system to function effectively. Of course, the contents of this state list can also be modified to meet the user's needs.

```

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.
-----
The texts above are part of the overall instruction. Do not start working yet:

```

Fig. 4. Prompt for the representation of environments.

D. The format of ChatGPT's output

In the following prompt, we explain how to format the output (Fig. 5). To facilitate easy integration with other processes, such as robot control systems and visual recognition programs, we encourage ChatGPT to output a Python dictionary, which can be saved as a JSON file. As a key point, we also encourage ChatGPT to output not only the sequence of robot actions but also the linguistic descriptions of each action and supplementary information on environmental changes before and after each operation. This information helps the user to debug whether ChatGPT is correctly understanding the input language and inferring environmental operations. Furthermore, these supplementary pieces of information are also useful for integration as an application. For example, the system can leverage the linguistic descriptions of each action to ask the user to perform the specific actions required to collect visual information needed to complete the task (see [33] for an actual application, an interactive learning-from-observation system). Additionally, the system can reuse the updated environmental information in the next input to enable multi-step instructions.

```

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 for the
robot corresponding to the list of robot actions.
- 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.
"""

-----
The texts above are part of the overall instruction. Do not start working yet:

```

Fig. 5. Prompt explaining the format of ChatGPT's output.

E. Examples of input and output

In the following prompt, we provide examples of the expected inputs and outputs (Fig. 6) to achieve effective few-shot results [34]. We found that providing more examples helps ChatGPT generate the desired sequence and minimizes the effort required by users to correct the output during conversations.

```
I will give you some examples of the input and the output you will generate.
Example 1:
"""
- Input:
{
    "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>)"},
    "instruction": "Put the juice on top of the shelf"
}
- Output:
```
{
 "task_cohesion": {
 "task_sequence": [
 "move_hand()", "grasp_object()", "detach_from_plane()", "move_object()", "move_object()", "move_object()", "attach_to_plane()", "release_object()"
],
 "step_instructions": [
 "move the hand near the juice", "grasp the juice", "pick up the juice", "take the juice from the bottom of the shelf", "move the juice upward", "move the juice above the upper shelf", "place the juice", "release the juice"
],
 "object_name": "<juice>"
 },
 "environment_before": {
 "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>)"}
 },
 "environment_after": {
 "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_top>)"}
 }
}
instruction_summary: "put the juice on top of the shelf"
question:""
```
"""

```

Fig. 6. Prompt providing examples of desired input and output. A part of the prompt is shown.

F. Specific input from the user

Lastly, we format the user input that contains the environmental and natural language instructions specific to the use case (Fig. 7). We have provided a template for the user input, as shown in (a) of Fig. 7, which can be reused by replacing [INSTRUCTION] with the user's text input and [ENVIRONMENT] with the user's specified environment, as illustrated in (b) of Fig. 7. It is worth noting that since ChatGPT's outputs the post-operation environment as the result of the former action sequence, the user only needs to input the environment at the beginning. For the second and subsequent instructions, ChatGPT's next response was created including all previous turns of conversation in order to allow ChatGPT to make corrections, if requested by the user, based on its own previous output and user feedback. If the number of input tokens exceeded the allowable number that ChatGPT could process, we adjusted the token size by truncating the first part of the sixth prompt (Fig. 8).

(a) Instruction at the last part of the instruction

```
Start working. Resume from the environment below.
"""
{"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.
2. Make sure that each element of the ["step_instructions"] explains corresponding element of the ["task_sequence"]. Refer to "ROBOT ACTION LIST" to understand the elements of ["task_sequence"].
3. The length of the ["step_instructions"] list must be the same as the length of the ["task_sequence"] list.
4. Even though objects disappear, do not delete them from the environment and use the "STATE LIST" to keep track of all the objects.
5. Make sure that you output a consistent manipulation. For example, grasping should not occur in successive steps.
6. Never left ',' at the end of the list.
7. All keys of the dictionary should be double-quoted.
8. Insert ``` at the beginning and the end of the dictionary to separate it from the rest of your response.
Adhere to the output format I defined above. Think step by step, being aware of what the left/right hand is grabbing or not
grabbing.]
```

(b) 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>)"}}
instruction = 'Put the juice on top of the shelf'
```

Fig. 7. The user input template and examples of actual input used. The user is assumed to define the environment information in advance in some way. By inputting the environmental information after the operation output by ChatGPT as the next input, continuous environmental operations can be realized.

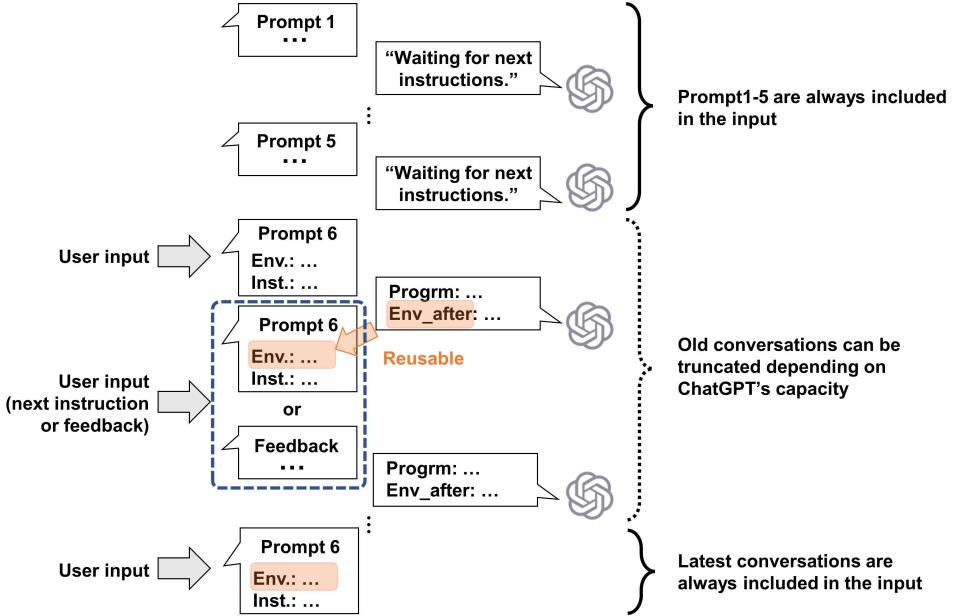


Fig. 8. The entire structure of the conversation that will be inputted into ChatGPT for generating a response.

III. EXPERIMENTS

In this section, we tested the prompts to check if ChatGPT can operate as expected. We used the fixed GPT model provided by Azure OpenAI (gpt-3.5-turbo) in our experiments. Due to space limitations, some experimental results are not presented here as figures. However, all results including instructions and environment definitions can be found at <https://github.com/microsoft/ChatGPT-Robot-Manipulation-Prompts>.

A. Multi-step manipulation of the environment

We tested the applicability of the proposed prompts for multi-step instructions in various environments. Specifically, 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, while assuming the operation in a household environment. To ensure usability, we provided ChatGPT with instructions and feedback in natural language that resembled the way humans communicate with each other. The results suggest that ChatGPT can effectively convert natural language instructions into robot actions. Furthermore, by reusing the state of the environment after each manipulation, which was output by ChatGPT, as the initial state for the subsequent manipulation (see Fig. 8), we were able to manipulate objects along the instruction sequences without losing track of the environment's state. Note that some of the results shown in this section contain the outcomes generated by ChatGPT after receiving user feedback.

1) *Relocation of objects on a table:* A juice can placed on the bottom shelf of two shelves and a Spam can placed on the table are manipulated by a multi-step natural language (Fig.9). First, the juice can on the bottom shelf is moved to the top shelf. Next, the Spam can is thrown into the trash. Next, the juice can is moved to the table top. Finally, the juice can is placed in the trash.



Fig. 9. Example of ChatGPT’s output for step-by-step natural language teaching, including multi-step manipulation of the environment. (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 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>.

2) *Open a fridge/drawer door:* Next, we tested the following actions: rotating the refrigerator door open, opening the refrigerator a little wider, removing the juice from the refrigerator and placing it on the floor, and finally closing the refrigerator (Fig.9). Similarly, the task could be broken down by ChatGPT with the action of sliding the drawer open (Figure is not shown)



Fig. 10. Example of ChatGPT’s output for opening fridge and taking out a 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 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>.

3) *Wipe a window with a sponge, and throw it away:* Next, we tested the action of taking the sponge on the desk, wiping the window, and returning it to the table. Following the operation, a robot throws the sponge into the trash (Fig.11). Similarly, the task could be broken down by ChatGPT with the action of wiping the table with a sponge (Figure is not shown).

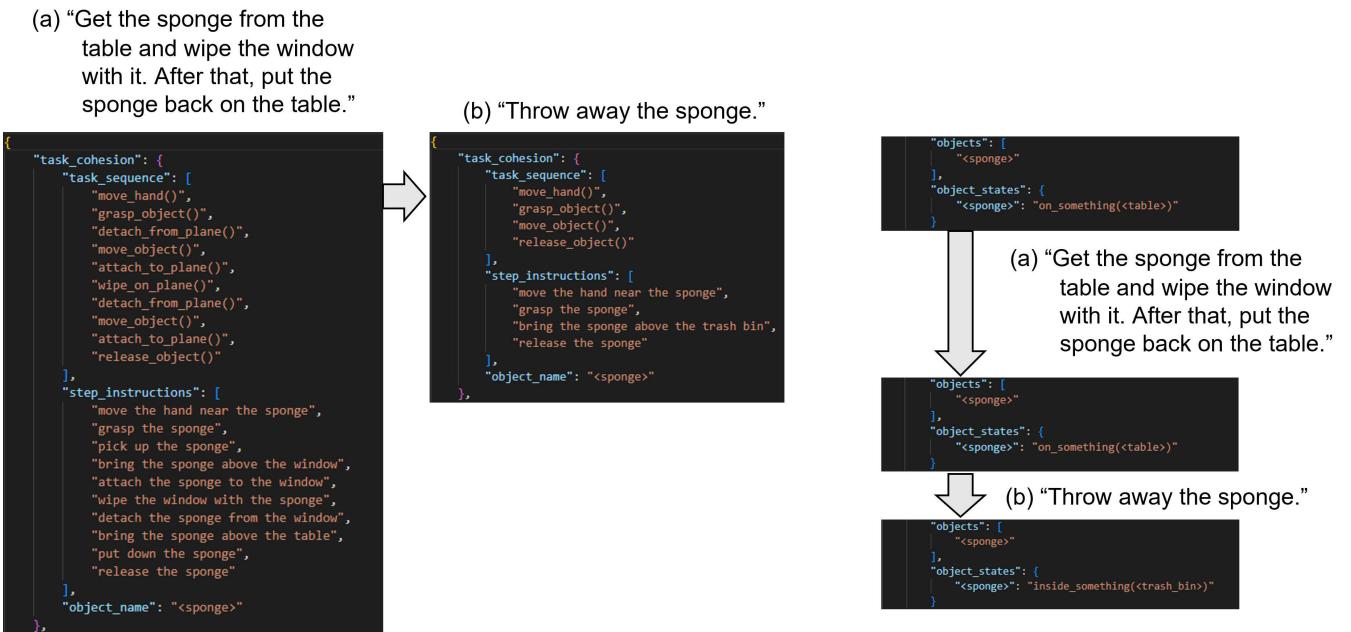


Fig. 11. Example of ChatGPT’s output for wiping a window with a sponge. (Left panel) Robot actions broken down for each natural language. (Right panel) The state of the environment that is output by ChatGPT. A part of JSON output is shown for each file. All the results, including the representation of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

B. Correction of ChatGPT's output through user feedback

Since ChatGPT does not always generate complete action sequences, it is important for users to be able to review and correct any errors in order to ensure safe and robust operation. With this in mind, We tested the ability to adjust ChatGPT's output by providing natural language feedback (see Fig. 12).



Fig. 12. Example of adjusting an output sequence with natural language feedback. The initial instruction to ChatGPT is to move a juice from the bottom shelf to the upper shelf. (Left panel) After the feedback of "Insert another move_object() to in moving the the juice upward." an action of move_object() was added to the sequence. (Right 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.

IV. DISCUSSION: TOWARDS MORE GENERAL ROBOTIC APPLICATIONS

In this paper, we focused on generating a sequence of basic robot actions from natural language commands given by a human. 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 long-step instructions while minimizing the impact of ChatGPT's token limit. Using these prompts, we conducted experiments to evaluate the effectiveness of ChatGPT in generating robot actions from natural language commands. Through the experiments, we confirmed that the proposed prompts work as expected. Additionally, we checked that ChatGPT enables the user to correct the output interactively, making the interface robust and useful. Based on these results, we consider that the proposed prompts provide a practical example that can be widely used in the robotics research community.

It seems not trivial that ChatGPT is able to decompose given instructions without experiencing actual object manipulation, only based on few-shot data. We consider that this ability may be attributed to the fact that the model learns about knowledge related to object manipulation and the temporal relationships between cohesively occurring actions within the vast amount of data it is trained on. In fact, ChatGPT can generate recipes when provided with a menu, which suggests that procedural steps are learned within the language model.

The specific prompts are designed under the assumption that the robot has at least one arm, sufficient degrees of freedom, and reachability to execute the desired task in the given environment. Additionally, we assume that a user's instructions are given

at the granularity of grasp-manipulation-release, which involves handling a single object from grasping to releasing. However, these assumptions may be too strong when considering more general robot manipulations. We will discuss several issues that arise when integrating the proposed approach into more general robotic applications, such as dealing with multi-object manipulations, and propose ideas such as incorporating visual feedback to address these challenges.

1) *Handling of higher-level logic:* Some manipulations may require changing actions depending on the recognition results or the environment, or repeating actions until certain conditions are met. For example, when performing a task such as throwing away trash, it may be necessary to use different garbage cans depending on the object or to wipe the table until it is clean. It is known that tasks such as generating programs from natural language, including high-level logic such as conditional branching, can be achieved by large-scale language models such as Codex ([35]). In robot applications, it is also suggested that logic can be constructed based on abstracted action sequences in a similar manner ([29]). Consistent with these ideas, we confirmed that small modifications to the prompts enabled ChatGPT to generate Python-style source codes that include logic (see Fig. 13). Additionally, we checked that it is also possible to prepare higher-level logic that controls the generated JSON files by preparing a separate ChatGPT interface responsible for logic and reading these JSON files (see Fig. 14 for an example). These results suggest that the proposed approach can be relatively easily connected to applications requiring higher-order logic.

(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>", "<cjuice>"],
"object_states": {"<spam>": "on_something(<table>)",
"<cjuice>": "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>)
"
}
```

Fig. 13. 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 track the state of objects at every line, as the final state can change according to the conditional branching. We also added a non-manipulative function 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>.



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.



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 are 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 are 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')
```



Fig. 14. An example showing the feasibility of using ChatGPT to generate control programs that include conditional branching. Such logic generation is outside the scope of this paper, which focuses on the generation of a sequence of robot actions for a single object manipulation (corresponding to the JSON files in the figure).

2) *Collaboration of multiple arms/robots:* When a robot has multiple arms, it is possible to coordinate them to achieve a task. The planner can treat the left and right arms as independent manipulators and read separate JSON files to manipulate a single object. Although this is beyond the scope of this study, any planner, including ChatGPT, can achieve this. We confirmed that small modifications to the prompts enabled ChatGPT to generate a task sequence that involves both arms (Fig. 15). Additionally, we checked that it is also possible to prepare another logic that controls multiple arms by calling the generated JSON files through a separate ChatGPT interface (Fig. 16). These results suggest that the proposed approach can be relatively easily connected to applications requiring multiple arms/robots to coordinate.

(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>"]
  },
}

```

Fig. 15. 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>.



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 hand_laterality specifies the left or right arm to use. Now, using these functions, please write Python source code that has the robot take out a juice can using the left arm while opening the door using the right arm.



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")

```



Fig. 16. 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.

3) Optimization of motion: Optimizing the motion of a robot is desirable for performing tasks safely and efficiently. For instance, the number of "move_hand()" calls and the trajectory of the hand can be optimized during manipulation. We recognize that language inference has limitations in inferring parameters related to the interaction between the environment and the robot hardware, since it requires physical experience of object interaction. For example, humans switch among a variety of grasping

strategies depending on the type of object to manipulate [36]. This is rooted in the symbol grounding problem. Our system is built on the philosophy of 90% AI, where humans always check the operation to ensure safe, predictable, and robust performance. Possible solutions include: 1) teaching the system optimal strategies through verbal correction (Section III-B) or visual demonstration (e.g., using our in-house learning-from-observation vision systems, where humans visually demonstrate how to move the body [30], [37] and where to focus [38], and how to grasp efficiently [39], [40]), and 2) designing robots to learn optimal policies with the help of humans (e.g., through reinforcement learning [41]).

4) *Adapting to environmental changes:* Our prompts aimed to teach environmental information to ChatGPT through the Prompt and to make ChatGPT output environmental information after the operation. Having ChatGPT output environmental information contributes to mitigating the effects of the token limit while allowing long-step task planning (Fig. 8). Additionally, the design of prompts that are aware of environmental information may contribute to ChatGPT's ability to output consistent task plans [42]. The working environment of a robot may change while a task is being performed. For example, objects may move within the workspace. By outputting these environmental changes to ChatGPT, along with the robot's action sequence and self-referencing prompts, the robot can track the changes it causes to its environment. However, a limitation of this paper is that humans need to provide ChatGPT with the environmental information in a formatted way. In future studies, we plan to consider integrating a vision (or multimodal) encoder to output the formatted environmental information.

5) *Connection with vision systems and robot controllers:* Among recent experimental attempts to generate robot manipulation from natural language using ChatGPT, our work is unique in its focus on the generation of robot action sequences (i.e., "what-to-do"), while avoiding redundant language instructions to obtain visual and physical parameters (i.e., "how-to-do"), such as how to grab, how high to lift, and what posture to adopt. Although both types of information are essential for operating a robot in reality [43], the latter is often better presented visually than explained verbally. Therefore, we have focused on designing prompts for ChatGPT to recognize what-to-do, while obtaining the how-to-do information from human visual demonstrations and a vision system during robot execution.

As part of our efforts to develop a realistic robotic operation system, we have integrated the proposed system with a learning-from-observation system (Fig. 17) that includes a speech interface [44], [45], a visual teaching interface [33], a reusable library of robot actions [46], and a simulator for testing robot execution [47]. Please refer to the respective papers for the results of robot execution, as it is beyond the scope of this paper. The code for the teaching interface is available at: <https://github.com/microsoft/cohesion-based-robot-teaching-interface>.



Fig. 17. An example of integrating the proposed ChatGPT prompts into a robot teaching system. The system breaks down natural language input instructions into a sequence of robot actions, and then obtains the necessary parameters for robot execution (i.e., how to perform the actions) by prompting a human to visually demonstrate each step of the decomposed action sequence.

V. CONCLUSION

This paper presents a practical application of OpenAI's ChatGPT for converting natural language instructions into a sequence of executable robot actions. We designed input prompts to meet common requirements in practical applications, specifically encouraging ChatGPT to output a sequence of robot actions in a readable format, represent the operating environment in a formalized style, and output the updated state of the environment. Through experiments, we tested the effectiveness of our proposed prompts in various environments. Additionally, we observed that ChatGPT's conversational ability allows users to adjust its output with natural language feedback, which is crucial for developing an application that is both safe and robust

while providing a user-friendly interface. Our prompts and source code are open-source and publicly available. We hope this paper provides practical knowledge to the robotics research community and inspires further development in this research area.

VI. NOTE

This paper was written by the authors, and ChatGPT was used for proofreading.

REFERENCES

- [1] OpenAI, "Chatgpt." <https://openai.com/blog/chatgpt>.
- [2] P. Pramanick, H. B. Barua, and C. Sarkar, "Decomplex: Task planning from complex natural instructions by a collocating robot," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6894–6901, IEEE, 2020.
- [3] S. G. Venkatesh, R. Upadrashta, and B. Amrutur, "Translating natural language instructions to computer programs for robot manipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1919–1926, IEEE, 2021.
- [4] I. Yanokura, 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 *2022 IEEE/SICE International Symposium on System Integration (SII)*, pp. 367–374, IEEE, 2022.
- [5] 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," *arXiv preprint arXiv:2210.03094*, 2022.
- [6] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*, pp. 785–799, PMLR, 2023.
- [7] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, et al., "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.
- [8] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar, P. Serbanet, N. Brown, T. Jackson, L. Luu, S. Levine, K. Hausman, and B. Ichter, "Inner monologue: Embodied reasoning through planning with language models," 2022.
- [9] Y. Ding, X. Zhang, C. Paxton, and S. Zhang, "Task and motion planning with large language models for object rearrangement," 2023.
- [10] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, "Progppromt: Generating situated robot task plans using large language models," 2022.
- [11] N. Kalithasan, H. Singh, V. Bindal, A. Tuli, V. Agrawal, R. Jain, P. Singla, and R. Paul, "Learning neuro-symbolic programs for language guided robot manipulation," 2023.
- [12] Z. Zhao, W. S. Lee, and D. Hsu, "Differentiable parsing and visual grounding of natural language instructions for object placement," 2023.
- [13] 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.
- [14] 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.
- [15] 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," 2023.
- [16] S. S. Raman, V. Cohen, E. Rosen, I. Idrees, D. Paulius, and S. Tellex, "Planning with large language models via corrective re-prompting," 2022.
- [17] Y. Xie, C. Yu, T. Zhu, J. Bai, Z. Gong, and H. Soh, "Translating natural language to planning goals with large-language models," 2023.
- [18] A. K. Kovalev and A. I. Panov, "Application of pretrained large language models in embodied artificial intelligence," *Doklady Mathematics*, vol. 106, pp. S85 – S90, 2022.
- [19] M. A. Khan, M. Kenney, J. Painter, D. Kamale, R. Batista-Navarro, and A. Ghalamzan-E, "Natural language robot programming: Nlp integrated with autonomous robotic grasping," 2023.
- [20] 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.
- [21] M. Skreta, N. Yoshikawa, S. Arellano-Rubach, Z. Ji, L. B. 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.
- [22] W. Huang, P. Abbeel, D. Pathak, and I. Mordatch, "Language models as zero-shot planners: Extracting actionable knowledge for embodied agents," 2022.
- [23] C. Lynch and P. Serbanet, "Language conditioned imitation learning over unstructured data," *arXiv preprint arXiv:2005.07648*, 2020.
- [24] J. Pan, G. Chou, and D. Berenson, "Data-efficient learning of natural language to linear temporal logic translators for robot task specification," 2023.
- [25] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," 2023.
- [26] 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," 2023.
- [27] H. Liu, L. Lee, K. Lee, and P. Abbeel, "Instruction-following agents with multimodal transformer," 2023.
- [28] O. Mees, J. Borja-Diaz, and W. Burgard, "Grounding language with visual affordances over unstructured data," 2023.
- [29] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *preprint*, 2023.
- [30] 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 *SII*, IEEE, 2021.
- [31] 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," *arXiv preprint arXiv:2103.02201*, 2021.
- [32] H. T. Kuhn and W. L. Inequalities, "Related systems," *Annals of Mathematic Studies, Princeton Univ. Press. EEUU*, 1956.
- [33] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Interactive learning-from-observation through multimodal human demonstration," *arXiv preprint arXiv:2212.10787*, 2022.
- [34] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.
- [35] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al., "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.
- [36] N. Wake, K. Sasabuchi, and K. Ikeuchi, "Grasp-type recognition leveraging object affordance," *HOBi-RO-MAN Workshop*, 2020.
- [37] K. Sasabuchi, N. Wake, and K. Ikeuchi, "Task-oriented motion mapping on robots of various configuration using body role division," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 413–420, 2020.
- [38] N. Wake, I. Yanokura, K. Sasabuchi, and K. Ikeuchi, "Verbal focus-of-attention system for learning-from-demonstration," in *ICRA*, IEEE, 2021.
- [39] N. Wake, D. Saito, K. Sasabuchi, H. Koike, and K. Ikeuchi, "Object affordance as a guide for grasp-type recognition," *arXiv preprint arXiv:2103.00268*, 2021.
- [40] D. Saito, N. Wake, K. Sasabuchi, H. Koike, and K. Ikeuchi, "Contact web status presentation for freehand grasping in mr-based robot-teaching," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 167–171, 2021.
- [41] D. Saito, K. Sasabuchi, N. Wake, J. Takamatsu, H. Koike, and K. Ikeuchi, "Task-grasping from a demonstrated human strategy," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, pp. 880–887, IEEE, 2022.

- [42] M. Gramopadhye and D. Szafir, "Generating executable action plans with environmentally-aware language models," 2022.
- [43] K. Ikeuchi, J. Takamatsu, K. Sasabuchi, N. Wake, and A. Kanehira, "Applying learning-from-observation to household service robots: three common-sense formulations," *arXiv preprint*, 2023.
- [44] N. Wake, M. Fukumoto, H. Takahashi, and K. Ikeuchi, "Enhancing listening capability of humanoid robot by reduction of stationary ego-noise," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 14, no. 12, pp. 1815–1822, 2019.
- [45] J. Jaroslavceva, N. Wake, K. Sasabuchi, and K. Ikeuchi, "Robot ego-noise suppression with labanotation-template subtraction," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 17, no. 3, pp. 407–415, 2022.
- [46] J. Takamatsu, K. Sasabuchi, N. Wake, A. Kanehira, and K. Ikeuchi, "Learning-from-observation system considering hardware-level reusability," *arXiv preprint arXiv:2212.09242*, 2022.
- [47] 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," *arXiv preprint arXiv:2301.01382*, 2023.