Appendix 1: Classification and Definition of Ambiguous Instructions

1. Ambiguity from Language Itself

This type of ambiguity originates from linguistic issues within the instruction, such as spelling and grammatical errors. Resolving this ambiguity requires reasoning based on visual information from the current scene.

- **Word Choice Error:** The instruction contains incorrectly used words. The ambiguity must be resolved by correcting the words based on visual context.
- **Grammatical Error:** The instruction has an incomplete or incorrect grammatical structure, but its intent can generally be understood. For example, "Give me that, on the table" is understood to mean "Give me that thing on the table."
- **Polysemy:** The instruction includes words with multiple meanings, and the precise meaning must be determined by referencing visual information.
    - **Verb Polysemy:** For example, in the instruction "pick it up," the intended meaning could be to lift an object or to answer a phone. The specific action to be performed must be determined from visual context.
    - **Noun Polysemy:** For example, the instruction "buy an apple" could refer to buying the fruit or an Apple device. The specific object of the operation must be determined from visual context.

2. Omission

This type of instruction lacks key information necessary to complete the task, such as the object to be manipulated or the specific action to be performed. These missing elements must be inferred by analyzing the current environment and context.

- **Omitted Object:** The instruction omits the object of the action. For example, the instruction "wash it" does not specify what to wash. Visual information, such as dirty plates in a kitchen sink, is needed to infer that the plates are the object.
- **Omitted Action:** The instruction omits the specific action to be taken. For example, the instruction "throw away the trash" does not specify the action, which needs to be inferred as "bag the trash and take it to the bin".
- **Omitted Degree of Action:** The instruction does not specify the extent or degree of the action. For example, in "I need to sit there, please pull the chair out for me," the user does not specify how far to pull the chair. This must be inferred from the context that a person needs enough space to sit down at the table.

3. High-Level Instruction

This type of instruction describes a final goal or an abstract task rather than a single, concrete action. It requires task planning capabilities to decompose the high-level goal into an ordered sequence of executable sub-tasks.

- **Explanation:** For instance, "prepare the meeting room" is a typical high-level instruction. It needs to be broken down into a series of specific steps: turn on the lights → turn on the projector → check the connection cables → arrange the chairs → adjust the room temperature. This decomposition process heavily relies on understanding the concept of "preparing a meeting room" and the current state of the room based on visual information. Another example is "season and plate the skewers".

4. Dependence on General Knowledge

Correctly executing this type of instruction requires not only understanding the literal meaning but also relying on common sense and knowledge accepted in human society. This knowledge is typically not explicitly stated in the instruction.

- **Physics Knowledge Dependence:** Executing the task must comply with basic physical laws. For example, when "stacking blocks," one must know to place larger, heavier blocks at the bottom to ensure stability.
- **Common-Sense / Habit Dependence:** The task's execution must align with human life habits and common sense. For example, the instruction "put the

milk in the refrigerator" implies placing it in the cooling compartment, not the freezer. Similarly, "set the table" for a Western meal typically means "fork on the left, knife on the right."

- **Safety Knowledge Dependence:** The task must be performed with safety considerations in mind. For example, if the instruction is "heat this lunch," and the lunch is in a metal container, one must know not to place metal objects directly into a microwave.
- **Math Knowledge Dependence:** The instruction contains concepts that require mathematical calculation to be understood. For instance, "put half of the cookies on the plate" requires first visually counting the total number of cookies and then calculating what constitutes half.

5. Relativity

This type of instruction uses relative descriptions, and its precise meaning depends on the spatial or attributive relationships between objects at the time the instruction is given. It must be interpreted using visual information.

- **Relative Spatial Terms:** Uses relative directional words like "left," "right," "front," and "middle". For example, in "hand me the book to the left of the lamp," one must first locate the "lamp" to determine which book is "to the left of it." An example from the tasks is "Place leftmost object in basin".
- **Relative Object Attributes:** Specifies an object through a comparison of its attributes, such as size, length, color, height, or material. For example, "get the bigger box" requires visually identifying at least two boxes in the scene and comparing their sizes to determine the target. Similarly, "stack the blocks by color" also falls into this category.
- **Relative Time:** "Get the water cup that has been here the longest."

6. Condition-Triggered

This type of instruction contains one or more "condition-action" logical pairs. The robot cannot execute any action directly; it must first use visual information to determine which "condition" is met and then perform the corresponding "action" paired with that condition.

- **Explanation:** For example, the instruction "check the living room light. If it's still on, turn it off." The robot needs to use visual information to determine if

the light is on (the condition), and only if the condition is true does it execute the action of turning it off.

7. Unclear Reference

This type of instruction uses pronouns with ambiguous antecedents, making it impossible to determine the object of the operation even with visual information, context, and general knowledge. In this situation, the robot needs to request clarification from the user by asking a question and can only determine the referent based on the user's response.

- **Explanation:** For example, if there is a book and a cup on a table and the user says, "bring it to me." The pronoun "it" is an unclear reference. The robot needs to ask a clarifying question, such as "Do you mean the cup?" to identify the object based on the user's answer.

8. Subjectivity (Human Preference)

The ambiguity in this type of instruction arises because the user has not clearly expressed the desired final state or standard, which is often subjective. It cannot be resolved solely with visual information, context, and general knowledge. The robot must request clarification from the user to resolve the ambiguity.

- **Unclear Degree of Task Completion:** The user assigns a task but does not define the criteria for completion. For example, "clean the room" could be interpreted as a simple "tidy-up" (e.g., sweeping the floor, taking out the trash) or a "deep clean" (e.g., wiping windows, cleaning carpets). Further inquiry is needed to determine the scope of work.
- **Subjectivity (Human Preference):** The criteria for completing the task depend on personal aesthetics or preferences. For instance, in "arrange these decorations to look nice," there is no objective standard for what "nice" looks like. It is necessary to ask about the user's preferences to resolve the instruction's ambiguity.


Common-Sense / Habit Dependence
Safety Knowledge Dependence
Math Knowledge Dependence
Physics Knowledge Dependence
Word Choice Error
Grammatical Error
Polysemy
Ambiguity from Language Itself
General Knowle-dge
Unclear Reference
Subjectivity (Human Preference)
Condition-Triggered
High-Level Instruction
Relativity
Relative Object Attributes
Relative Spatial Terms
Omission
Omitted Object
Omitted Action
Omitted Degree of Action