Appendix 2: Ambiguous Instruction Generation （Prompt）

Task Description

In contrast to well-designed, clear, complete, and specific high-level instructions, human commands in natural language are often ambiguous. An **Embodied Ambiguous Instruction** is defined as follows: an instruction that must be converted into a sequence of specific, executable sub-tasks by relying on an analysis of the current scene's visual information, combined with reasoning based on context and general knowledge. These sub-tasks must have clear target objects and actions.

I will provide you with an image of a scene. You need to combine the visual information to generate multiple embodied ambiguous instructions that a human might ask a robot to perform in that scene. These instructions must conform to the definition of an embodied ambiguous instruction provided above.

Images as follows: {images}

Chain of Thought

You need to follow the thinking process below:

1. **Extract objects and their features:** Extract all objects from the image. Combining the image content with general knowledge, describe the properties and features of these objects (must include the object's length, width, height/thickness, purpose, interactive parts, state, color, position, and relative spatial relationships). In this step, output the names of the extracted objects and their detailed feature descriptions.
2. **Operability Analysis:** Based on the detailed feature descriptions of the objects extracted in the previous step, determine if they can be manipulated by a two-fingered gripper robot and in what ways.

   The parameters for the two-fingered gripper are: maximum payload of 1kg, gripping force of 20-140N, stroke width of 0-110mm, and single-arm

operation. Achievable actions include grasp, place, push, pull, rotate, and press.

Operating Principles:

- A. Must adhere to Asimov's Three Laws of Robotics.
- B. Objects graspable by the gripper must have a certain degree of hardness and a defined shape. Liquids or granular materials must be handled using containers or tools. Fixed, large objects cannot be moved.
- C. Do not operate on fragile items like glass, touch sharp objects like knives, or interact with electrified objects like power outlets.
- D. Do not make contact with humans or other living beings.

In this step, output the object's name, whether it is operable, and a description of the ways it can be operated.

3. **Scene and Human Intent Analysis:** Combining the outputs from the first two steps, analyze all potential human needs and intentions in the current scene. Analyze what tasks a human might ask a robot to perform in this context. Output a detailed analysis of human intent within the scene.
4. **Unambiguous Instruction Generation and Ambiguity Introduction:** (Loop through 15 types of embodied ambiguous instructions, generating one for each category).

   {A Common-Sense/Habit-Dependent Ambiguous Instruction is an ambiguous instruction whose execution needs to conform to human life habits. For example, "set the table" typically follows the "fork on the left, knife on the right" convention in Western dining. Another example is "put the milk in the refrigerator," which usually means placing it in the cooling section, not the freezer.}

   For all human intents analyzed in the previous step, generate one clear and unambiguous robot instruction for each. Then, based on the definition of a {Common-Sense/Habit-Dependent} ambiguous instruction, introduce this type of ambiguity to convert each unambiguous instruction into an embodied ambiguous instruction.
5. **Filter Embodied Ambiguous Instructions:** Using reasoning that combines the image, object descriptions, operability analysis, scene and intent analysis, and ambiguity analysis, filter all the embodied ambiguous

instructions generated in the previous step. Select the instructions that are reasonable for the current scene and conform to the definition of an embodied ambiguous instruction.