Multi-Modal HRI Instruction

The goal of our proposed system is to achieve the most intuitive and most natural HRI. Imagine that there are multiple similar objects on a table and you want a robot to give you one of them. What would you do?
You might say please give me the one on the left, but what if there are five objects? How do you tell left and right if your relative pose with the robot is not fixed?

Therefore, we choose to combine the two most natural forms of interaction, speech and gaze, even for patients with limb disability (Hawking was able to move his eyes wen he was completely paralyzed).

In our scenario, you just have to talk to the robot like you would talk to a human. Say your command and look naturally at the target object. Usually the user does not look at the target all the time, only at certain moments will use the gaze to select a specific target. For example, when a user says ‘please give me this cup’, the user usually looks at the cup when he says ‘this’, while before this word the user just surveys the entire workplace. In our proposed system, we utilize LLM to inference the user’s command and predict this word.

Here is how you can try this system:
1. Take on the glasses and wait.
2. Start SLAM with a slow, slight left-to-right, back-and-forth movement.
3. When I said ‘OK’, name the task you want the robot to do. To save time, you can try the following three levels of tasks.
    a. Single Step task: ‘Please pick up this cup’
    b. Multiple Steps Task: ‘Please put this cup on the plate’
    c. Causal Task: ‘Please put this strange fruit on the plate and pour something from this cup on it’ (The action of the front has an impact on the back, you only need to determine the position of the fruit, plate and cup by looking at it, and the LLM action planner will automatically choose the pour position).