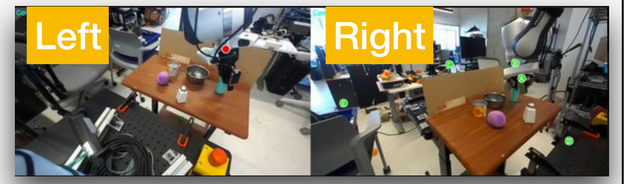
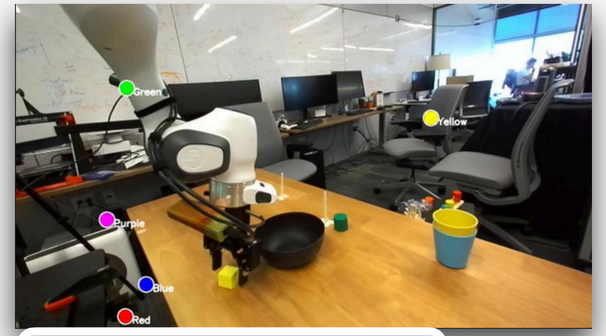


Multiple View

Q: In the left image (ext1 camera), a red dot is marked. Which point is the closest point in the right image (ext2 camera) corresponding to the same 3D location?

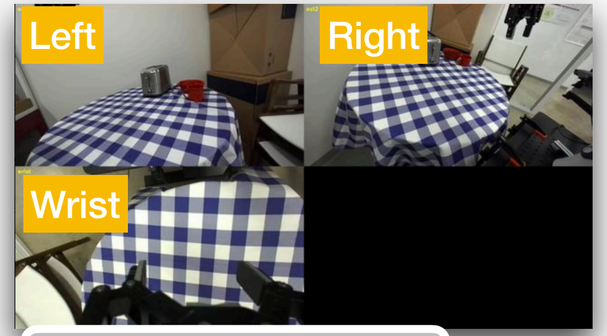
Key Modalities: 2 views, stereo (depth) img



Scene Understanding

Q: In the image from ext2, which colored point is CLOSEST to the camera?

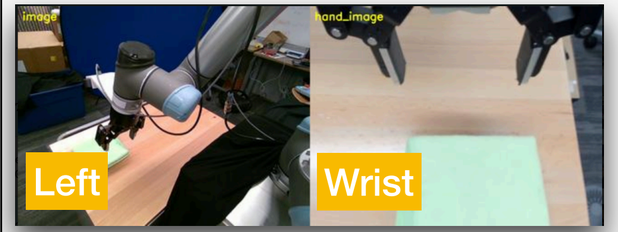
Key Modalities: stereo (depth) images



Task State - Success

Q: The robot is to turn on the toaster. Has the robot successfully completed the task?

Key Modalities: 2 Side View, Wrist View

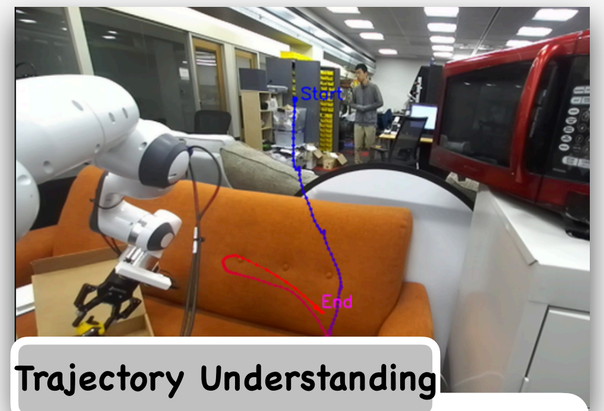


Robot State - Gripper

Q: Is the robot gripper open?

Key Modalities: gripper state, Side+Wrist View

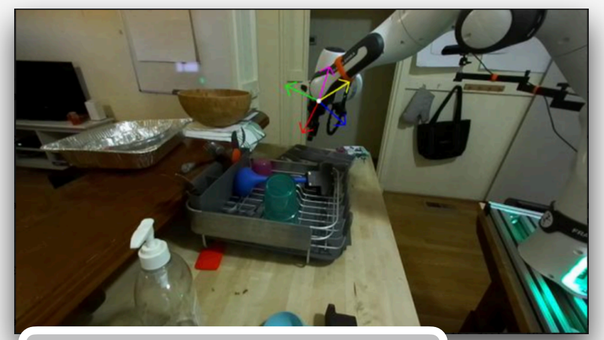
pre-grasp | Immobilize | Contact | Detach | Post-Grasp



Trajectory Understanding

Q: what language instruction best describes the **robot's trajectory** shown in the image?

Key Modalities: End effector pose, language



Spatial Relationship

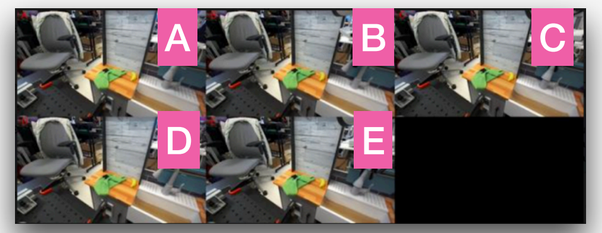
Q: The robot is tasked to move the spoon, which **arrow** shows the most possible direction to move next?

Key Modalities: End effector pose, language

Task State - Grasp

Q: Is the robot grasp Sponge stable?

Key Modalities: language, gripper state



Task State - Goal

Q: The robot is tasked to move the tap, which configuration shows the goal state that the robot show achieve?

Key Modalities: language, gripper state