

Step 1: Pose prediction


Initial state:



Language prompts:

1. What is the category of the object in the image?
2. How to open the {object}?
3. Predict the contact point and orientation for pulling the {object}

Input

 **MCC-MLLM**

Pose prediction:

Contact point: (x, y) ,
gripper's upward direction: (x_u, y_u, z_u) ,
gripper's forward direction: (x_f, y_f, z_f) .

Step 2: Failure detection and correction

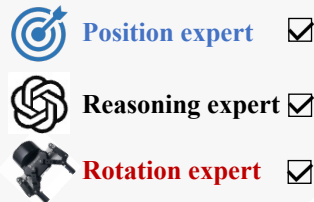
End state:



Language prompts:

The robot end-effector state is
Detect the failure causes of pulling the {object}.

Correction experts:



Error type

Prompt feedbacks



Manipulation and Correction Collaborative (MCC)-MLLM

Failure causes:

The failure cause is {*position* // *rotation* // *position and rotation*}.

Corrected pose prediction:

Contact point: (x^c, y^c) ,
gripper's upward

Step 3: Continuous policy learning

Corrected state:



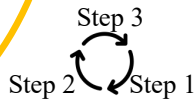
Corrected pose label:

Contact point: (x^c, y^c) ,
gripper's upward

Train samples



MCC-MLLM



Close-loop Correction



: Fine-tune



: Frozen