Vision Encoder

Project Layer

Language questions:

Pose prediction prompt
Failure detection prompt
Failure correction prompt

Tokenizer

Larger Language Model

The contact point at [x , y] , the gripper up ⋮

Adapter

Inject

: Frozen

: Fine-tune

Pose prediction prompt:

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

1) The object in the image is a small wooden cabinet
2) You would typically pull the door from the right side.
3) The contact point is [x, y], the gripper up 3D direction is $[x_u, y_u, z_u]$*, the gripper forward 3D direction is* $[x_f, y_f, z_f]$

Failure detection prompt:

1) The robot's end-effector state is at [x, y], with the gripper's up 3D direction $[x_u, y_u, z_u]$ *and the gripper's forward 3D direction* $[x_f, y_f, z_f]$*. Detect the failure causes of pulling the cabinet.*

1) Failure cause is incorrect prediction of position and rotation.

Failure correction prompt:

1) Here are potential contact point coordinates: $[x^c, y^c]$*, Here are the potential orientations: the gripper's upward 3D direction is* $[x_u^c, y_u^c, z_u^c]$*, and the gripper's forward 3D direction is* $[x_f^c, y_f^c, z_f^c]$*. Predict the contact point and orientation for pulling the cabinet.*