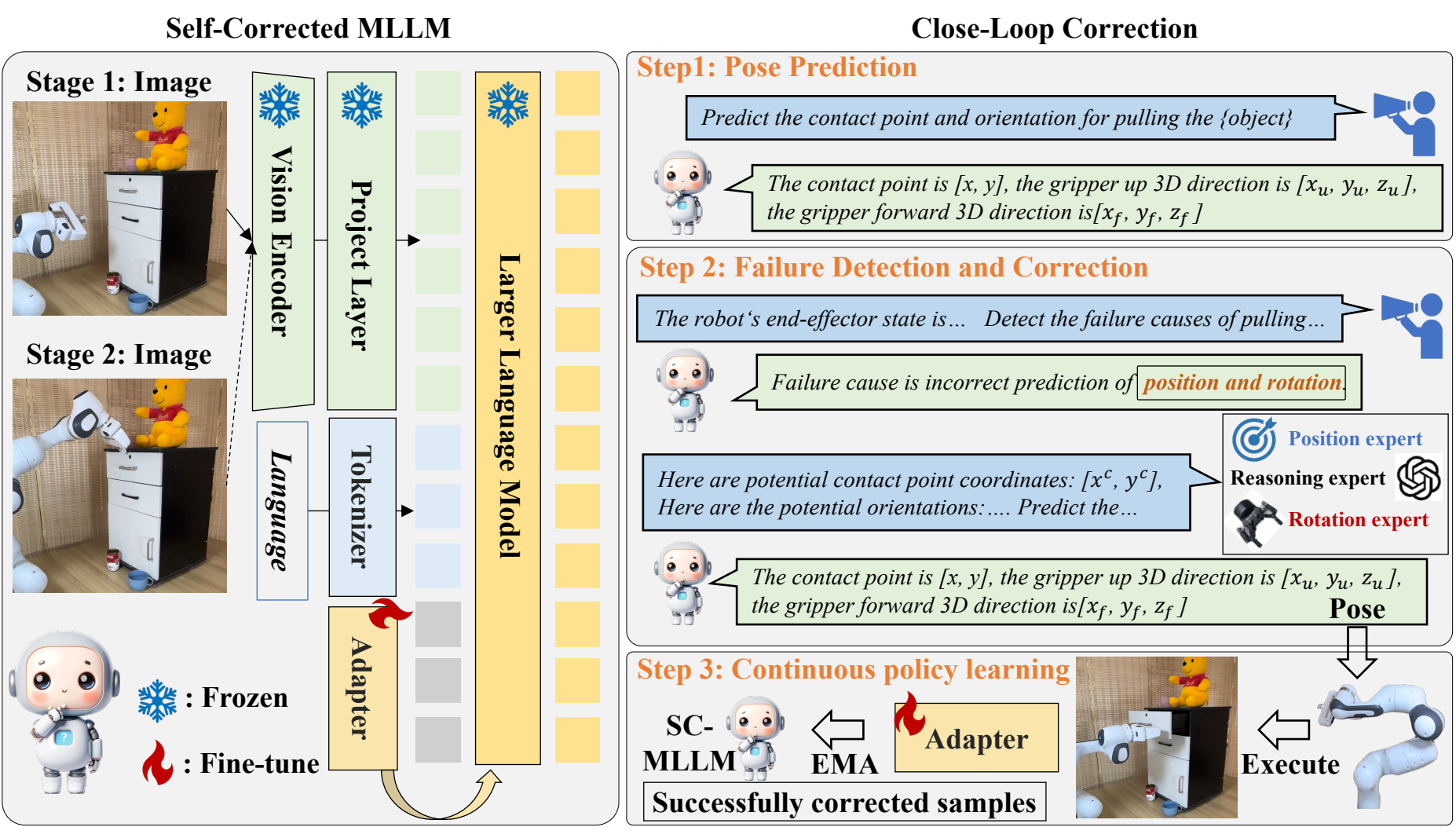
Self-Corrected MLLM
Close-Loop Correction
Stage 1: Image
Stage 2: Image
Vision Encoder
Project Layer
Language
Tokenizer
Larger Language Model
Adapter
: Frozen
: Fine-tune
Step1: Pose Prediction
Predict the contact point and orientation for pulling the {object}
The contact point is [x, y], the gripper up 3D direction is [x_u, y_u, z_u], the gripper forward 3D direction is[x_f, y_f, z_f]
Step 2: Failure Detection and Correction
The robot's end-effector state is… Detect the failure causes of pulling…
Failure cause is incorrect prediction of position and rotation.
Here are potential contact point coordinates: [x^c, y^c], Here are the potential orientations:…. Predict the…
Position expert
Reasoning expert
Rotation expert
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]
Pose
Step 3: Continuous policy learning
SC-MLLM
EMA
Adapter
Successfully corrected samples
Execute