

Instruction

“Pick the bowl not between the ramekin and plate and place it on the plate.”



FIT-MAC

Visual observation



Planner Agent

- Target: bowl
- Destination: plate
- Relation: not between
- Reference: ramekin, plate



Candidate Detection



Red:
candidate
boxes



Grounder Agent



Yellow:
selected
target

Visual
Information



Textual Information

Reliable
grounding?



VLM Verifier



Rebind Target

Shared Global States: Task | Grounding | Execution | Recovery



Monitor & Diagnosis Agent

State Monitor



Phase

approach--grasp--lift



Gripper

open / closed



Object Motion

displacement / lift

Failure Diagnosis

Target
Failure

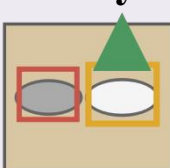
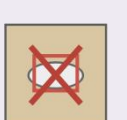


Execution
Failure

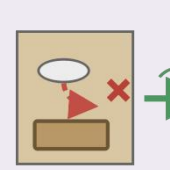


Adaptive Recovery

Target
Rebinding



Action
Rescue



Target + State



Environment



Action Chunk Prediction



Success

