

(a) Current gaps

Open-loop video QA



Q: Explain the demonstrated steps.
A: Pick up the mug, navigate to the cabinet, and place it inside.



Fixed video,
no decision-dependent feedback

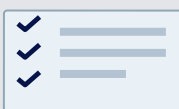
Limited closed-loop settings



Sparse post-action views
Occasional scene observations.



Idealized execution
Assumes perfect execution.



Step-by-step instructions
Scripted steps with limited scene-grounded planning.



Degenerate to language decomposition
Weak scene and outcome dependence.

(b) SCOPER-bench

Scene-evidence tasks

Goal: Tidy this room

Discover evidence before choosing actions



Running faucet
Turn off



Open cabinet
Close

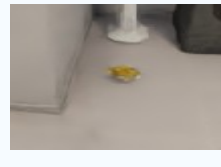


Plate on floor
Pick up

Ongoing visual feedback

Scene observations throughout execution



Observe scene changes and action outcomes

Shared rule-based System 1



System 2

Calls a common action interface



Explore



Navigate



Manipulate

Available during evaluation

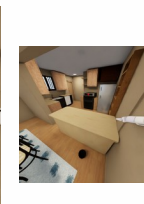
Controllable failure injection

Inject failures into System 1 execution

Expected: pot held



Observed: pot falls



Evaluate recovery from execution failures

Scene-grounded decisions & closed-loop robustness

Challenges revealed



Evidence retention

Remember discovered scene evidence



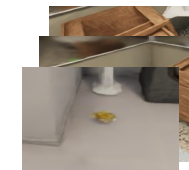
Progress & recovery

Verify outcomes and recover

Motivates

(c) SCOPER-harness

Evidence Memory & Progress Verification



Evidence Memory

Retain scene evidence
and update the agenda

Faucet: running
Cabinet: open
Plate: on the ground



Progress Verification

Check outcomes and update progress



→ Pick up
again

Detect failure and revise the next action



Anti-deadlock

Prevent repeated checks
from stalling progress

Observe

Check

Act

Empirical evidence



Improved task completion



Gains with real-world observations