LEAP Workshop, CoRL 2026

Evaluating Long-Horizon Robot Planning under Uncertainty

Harsh Muriki*, Gayathri Aravindan, Yuchen Zheng, Sonia Chernova
Georgia Institute of Technology *vmuriki3@gatech.edu

*How do **initial robot memory** and **changes in the environment** affect long-horizon planning performance?*

Motivation

A robot's memory of a home can become outdated. Objects can be moved, replaced, or fully removed. Different scenarios require the robot to take different actions: search, substitute, replan, or report the goal as infeasible.

Benchmark

Sim: Meta Habitat[1] , 12 High-level Skills (pick, place, navigate, open, close, etc.)

Tasks: T1 water fetching, **T2** package preparation, **T3** book collection, **T4** grocery organization, **T5** guest table preparation, **T6** meal delivery, **T7** household cleanup.

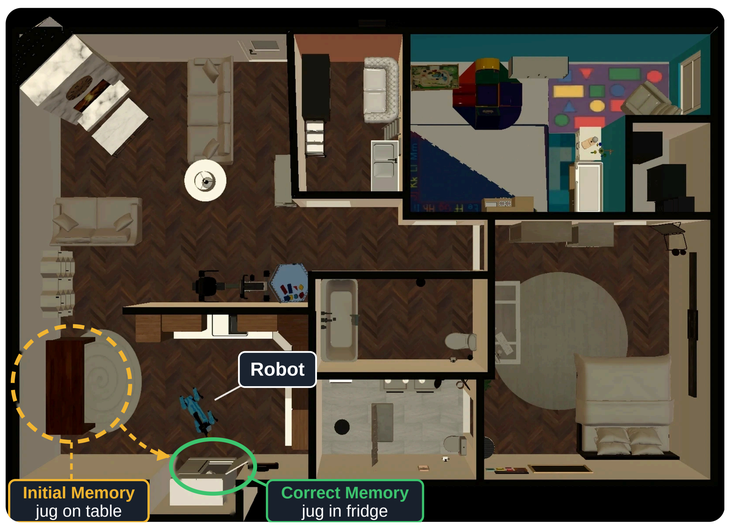
Variants = world × memory (up to 15 per task)

External Environment Changes: what is true in the environment

BASE	SUB	ABS	HID	DIS
baseline	substitute obj available	object absent	object hidden	distractors exist

Initial Memory: what the robot believes at the start

Accurate	Incomplete	Outdated
correct information	some information omitted	stale, no longer true



Example:
T1 under **HID** × **Outdated**

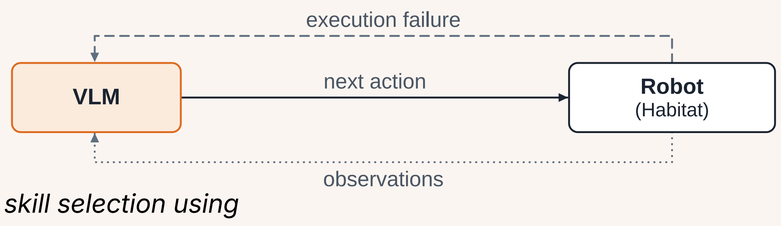
"Turn off the master-bedroom lights and bring a full jug of water to the living-room table."

The jug is **in the fridge**, but the robot remembers it **on the dining table**.

Baseline Planners

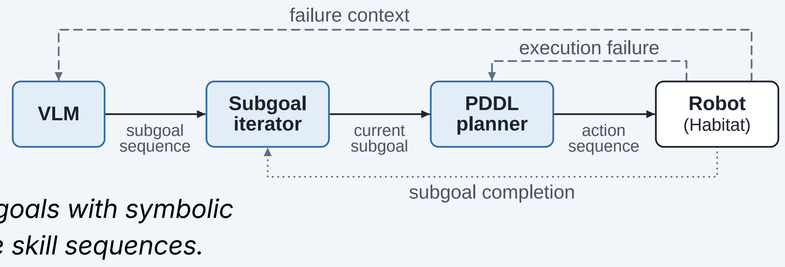
All algorithms have some shared inputs: memory, goal, scene graph, images (except Tru-POMDP). *GPT-5.6 Luna (High) as VLM.*
—— plan / action --- failure signal ····· observation / progress

ReAct[2]



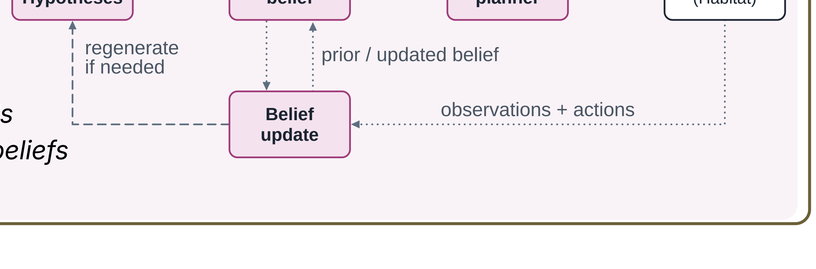
Interleaves VLM reasoning and skill selection using visual observations, memory, and execution feedback.

VLM-TAMP[3]



Combines VLM-generated subgoals with symbolic planning to produce executable skill sequences.

Tru-POMDP[4]

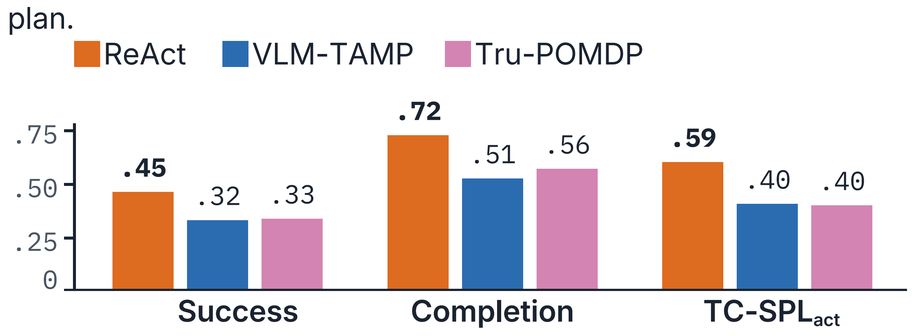


Maintains hypotheses over goals and object locations, updating beliefs through belief-space planning.

Task Success and Efficiency

We evaluate with **task success**, **task completion**, and **TC-SPL$_{act}$**.

TC-SPL$_{act}$ measures subtask completion and action efficiency relative to a human reference plan.



Effect of Initial Memory

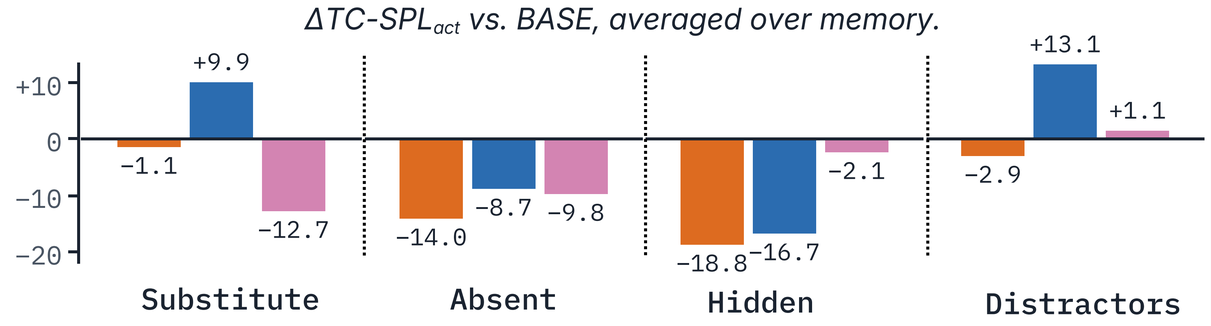
Only ReAct does best with accurate memory; VLM-TAMP and Tru-POMDP perform best with outdated memory.

TC-SPL$_{act}$	Accurate	Incomplete	Outdated
ReAct	**0.71**	0.55	0.52
VLM-TAMP	0.37	0.39	**0.43**
Tru-POMDP	0.39	0.32	**0.47**

Effect of Environment Changes

On average, most environment changes lower metrics.

ΔTC-SPL$_{act}$ vs. BASE, averaged over memory.



Failure Analysis

Rule-based labels on execution traces. Failures: why unsuccessful runs failed. Inefficiencies: wasted actions in any run. Labels overlap.

Failures % of all runs	ReAct	VLM-TAMP	Tru-POMDP
Object not placed	12.9	34.9	30.4
Missing action	20.4	32.7	24.4
Incorrect placement	29.3	26.9	11.1
Object not found	10.9	22.4	12.4
Inefficiencies % of all runs			
Unnecessary actions	26.2	18.4	52.4
Precondition error	46.0	13.8	23.1
Repeated actions	2.4	30.2	41.1

Takeaways

- ReAct does best, outperforming symbolic and belief-space planners in task success and completion.
- Accurate memory does not always improve performance; how planners use and revise knowledge matters.
- Reliable, efficient planning requires knowing when to search, verify task completion, and stop.

1. X. Puig, E. Undersander, A. Szot, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. 2023.
2. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao. ReAct: Synergizing reasoning and acting in language models. *ICLR*, 2023.
3. Z. Yang, C. Garrett, D. Fox, T. Lozano-Pérez, L. P. Kaelbling. Guiding long-horizon task and motion planning with vision language models. ICRA, 2025.
4. W. Tang, X. He, Y. Huang, Y. Xiao, C. Lu, P. Cai. Tru-POMDP: Task planning under uncertainty via tree of hypotheses and open-ended POMDPs. NeurIPS, 2025