WCM: World-Cognition Model for Generalizable Human-Robot Interaction

Yuzhen Chen[1] KC Zhou[1,2]

[1] Harvard University [2] MIT

ROBOTICS SCIENCE AND SYSTEMS
RSS 2026 | INTERNATIONAL CONVENTION CENTRE AND UNIVERSITY OF TECHNOLOGY SYDNEY
13 - 17 JULY 2026 | AUSTRALIA

Paper Website

A Human-Centric, Reasoning-First Framework:
Robots should understand from the human's perspective, act, and explain its decisions.

TEACHING MODE

Hi I am Sage, ready to help out	I will teach you how to take the banana from me and put into the drawer	I gave the banana to you, next align yourself to the gray cavity	I align to the cabinet	I off handed the banana	I look for the cabinet handle	I opened the cabinet drawer	I picked up the off-handed item	I placed it inside the drawer	I closed the drawer
Turn on teaching mode	Sure, I will reach out to you and please hand me the banana	Hold the banana, you can release your hand	Put down the item currently in your possession.	Look for the handle on the gray cabinet	Open the cabinet by using the cabinet handle	Pick up the off-handed item	Place the item inside of the open drawer	Close the drawer using the handle	Ok, you finished the task

RESULTS

TABLE I

Spoken instruction	SR (%)	Trials	Time (s)
"Grab me the popcorn."	85	11/13	72
"Hand me the honey."	79	11/14	34
"Hand me the jacket."	67	6/9	54
"Hand me the green ratchet."	64	7/11	36
"Hold the BBQ tongs."	86	12/14	48
"Grab me the water bottle."	83[1]	10/12	30
"Take the screwdriver to the drawer."	69[2]	9/13	165
"Put the cup in the trash can." (Outdoor)	67[1]	10/15	46
"Put the bottle in the trash can."	64[1]	9/14	34

Nine live dialogue-plus-handover tasks; WCM only. Trials = successes/attempts, SR their ratio, Time = end-to-end run (s). [1] zero-shot (held out). [2] learned via teaching (pre-distillation).

TABLE II

	Movement time (s)↓		GPU busy time (s)↓		Success (%)↑
Task	WCM	X-Sq.	WCM	X-Sq. (est.)	WCM
ColorPlateSort	**25.0**	109.0	**19.4**	109.0	92
LabelRecognition	**36.0**	116.0	**18.4**	116.0	90
PickUpFruits	**16.6**	18.5	**6.4**	18.5	90
ThrowTrash	**80.5**	258.0	**22.5**	258.0	86
PickUpCube	**16.1**	43.2	**5.6**	43.2	92
SolveMathProblem	**57.0**	76.0	**59.9**	76.0	78
Average	**38.5**	103.5	**22.0**	103.5	88.0

Case study: WCM vs. X-Square (WALL-OSS), six tasks. WCM moves faster and is far more GPU-efficient: it uses the GPU only intermittently (freeing it for other robots), while a VLA occupies it every control step. Success reported for WCM only (baseline gave none). Arrows = better direction; bold = better.

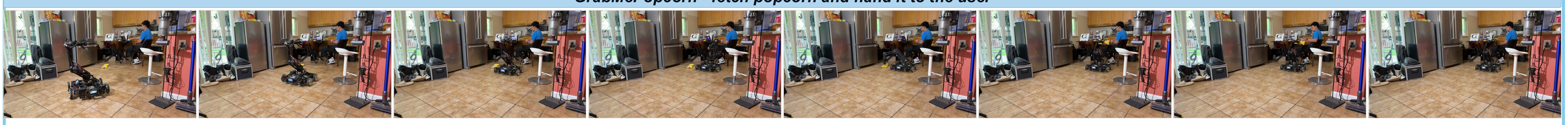
TABLE III

	Time (s)↓		Success (%)↑	
Task	WCM	w/o Async	WCM	w/o CyberView
"Hand me the green ratchet."	**36**	61	**64**	18
"Grab me the water bottle."	**30**	47	**83**	27
"Take the screwdriver to the drawer."	**165**	254	**69**	0
"Put the cup in the trash can." (Outdoor)	**46**	63	**67**	18
"Put the bottle in the trash can."	**34**	49	**64**	9

Ablations, five tasks. Removing the asynchronous runtime forces the robot to halt and think between actions, inflating end-to-end time; replacing CyberView with a plain VLM + 2D grounding (no metric interactable points) collapses success. Bold = full WCM.

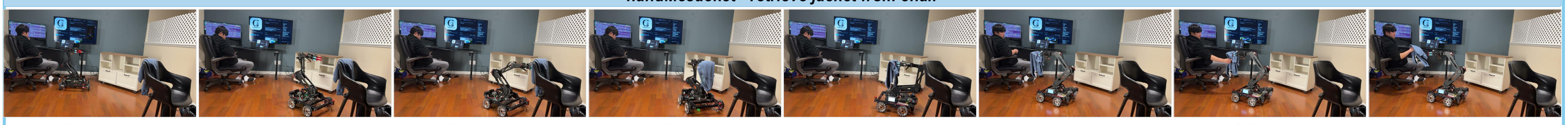
REAL-ROBOT ROLLOUTS · SPOKEN INSTRUCTION · NO TASK-SPECIFIC POST-TRAINING

GrabMePopcorn - fetch popcorn and hand it to the user



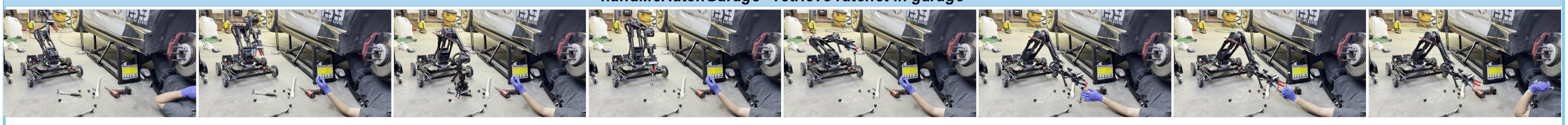
Check user & popcorn | Move to popcorn | Stop by user | Reach for popcorn | Grab popcorn | Bring to user | User takes popcorn | Retract after handoff

handMeJacket - retrieve jacket from chair



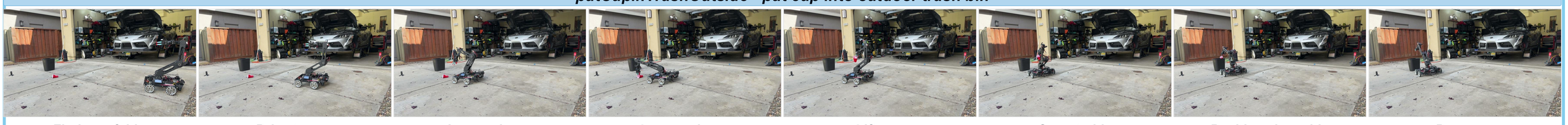
Find jacket | Drive to chair | Reach jacket | Lift jacket | Turn with jacket | Approach user | Offer jacket | Release jacket

handMeRatchGarage - retrieve ratchet in garage



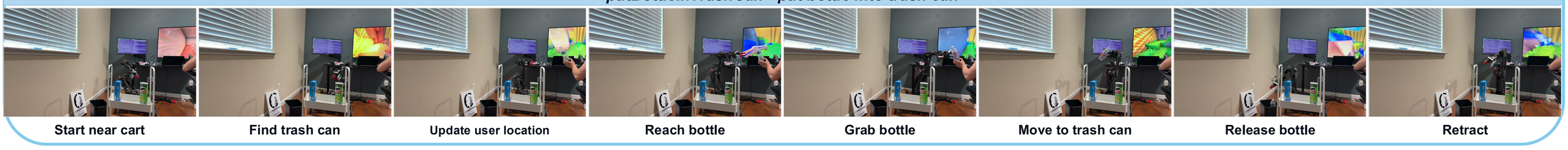
Scan garage | Approach ratchet | Lower gripper | Grab ratchet | Lift for delivery | Extend to user | User takes ratchet | Retract

putCupinTrashOutside - put cup into outdoor trash bin



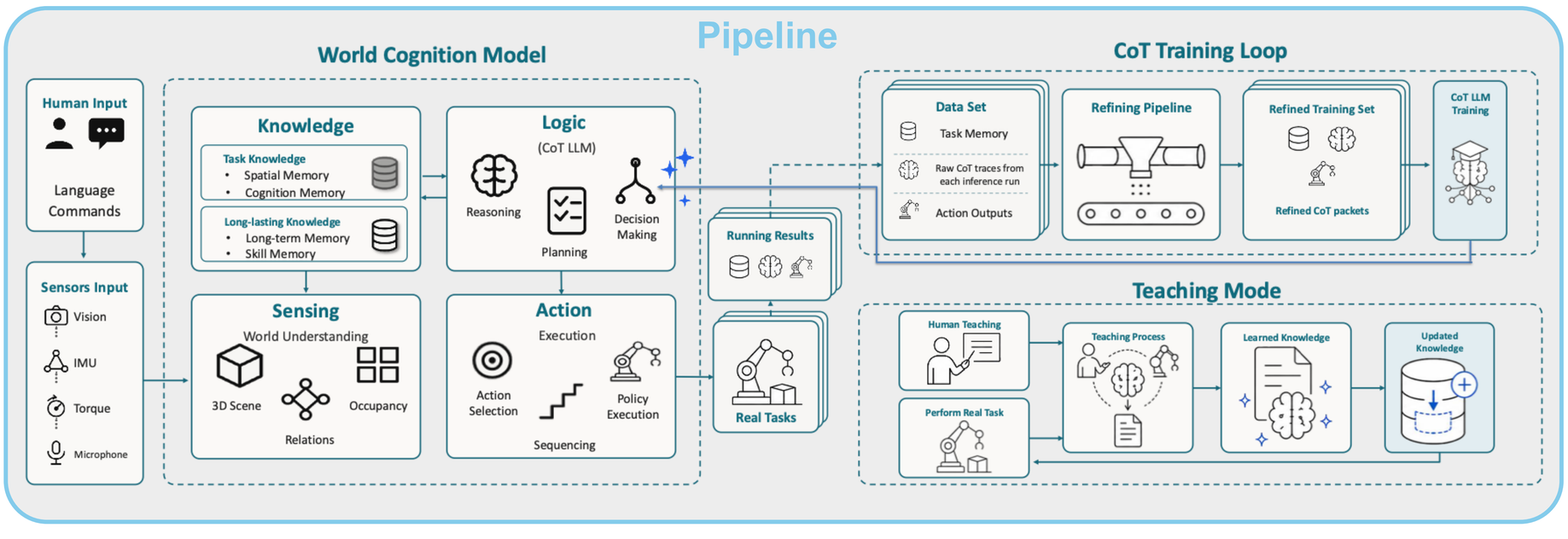
Find cup & bin | Drive to cup | Approach cup | Lower gripper | Lift cup | Carry to bin | Position above bin | Drop cup

putBottleInTrashCan - put bottle into trash can



Start near cart | Find trash can | Update user location | Reach bottle | Grab bottle | Move to trash can | Release bottle | Retract

Pipeline



CONTRIBUTIONS

Architecture
Four decoupled white-box layers (SLAK) + an asynchronous runtime: reason while moving; swap any layer without retraining.

Data Flywheel
A refine pipeline distills real-rollout reasoning back into the robot; teaching mode adds skills by spoken interaction.

Hardware
A capable, low-cost minimal possible hardware that makes reasoning-first embodied HRI affordable.

Autonomy
Act, collect, refine, redeploy as one automated loop; the robot keeps improving with little or no human in the loop.

ACKNOWLEDGMENTS

We thank T. Lihe and Andrew Hoang for their contributions to the system implementation and experiments. This work was funded by Cyberbrain, Inc. as part of its internal R&D program; the system, datasets, and related IP are owned by Cyberbrain, Inc.