When Code Reaches for the World



Exploring Spatial and Embodied Intelligence with Coding Agents

Spatial intelligence

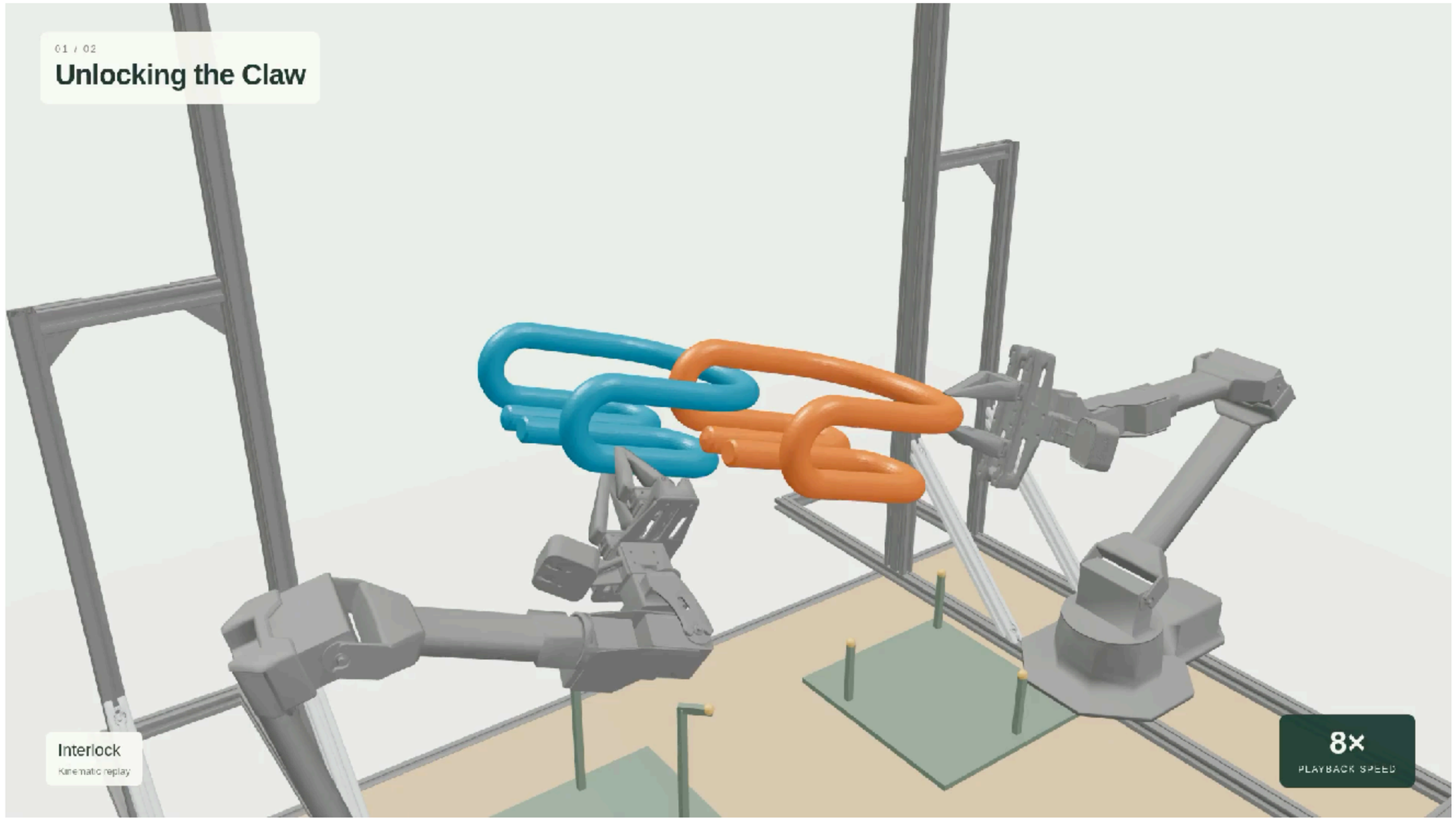
World modeling

Embodied intelligence

Qineng Wang

PhD Student · Northwestern University

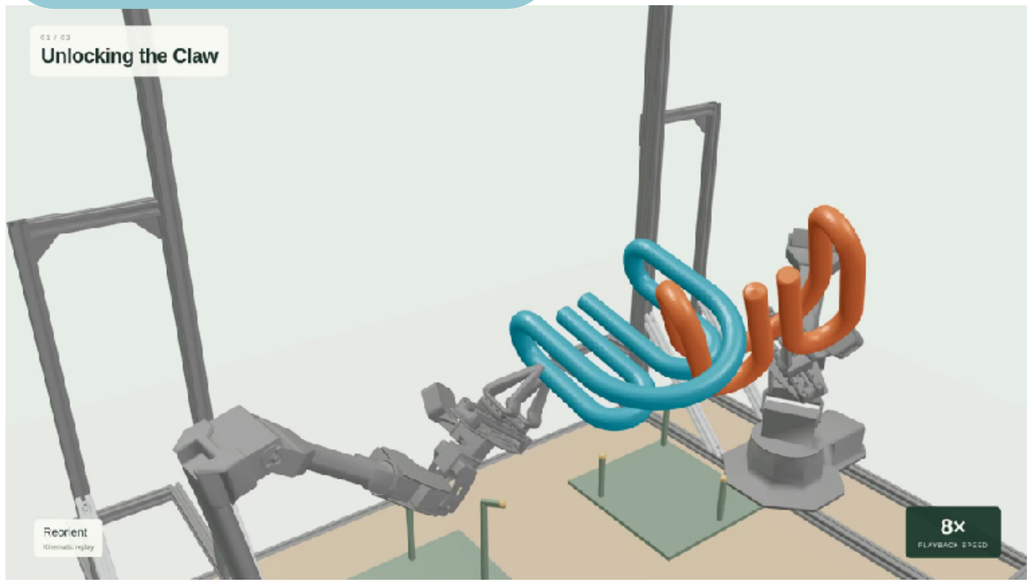
GPT-6 Astra meets spatial constraints



Qineng Wang · September 15, 2026 · Accelerated replay: 8× / 16×

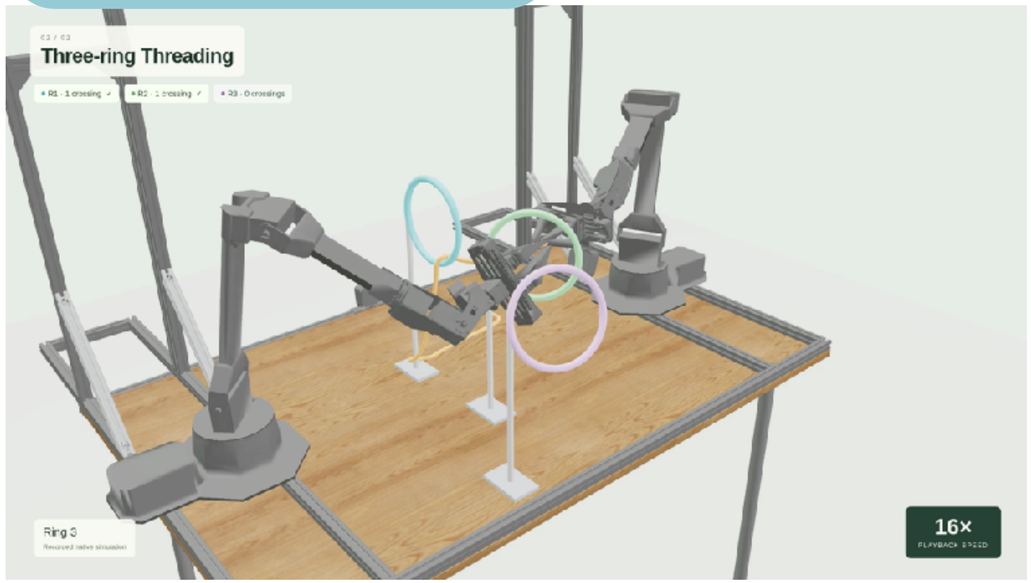
What had to be right?

Rigid configuration



Avoid collisions along the motion

Flexible object



Preserve the route through all three rings

Success is a property of the whole trajectory.

From a robot video to a simulated world

Recorded robot

Reconstructed simulation



Evidence

Multi-view video + actions

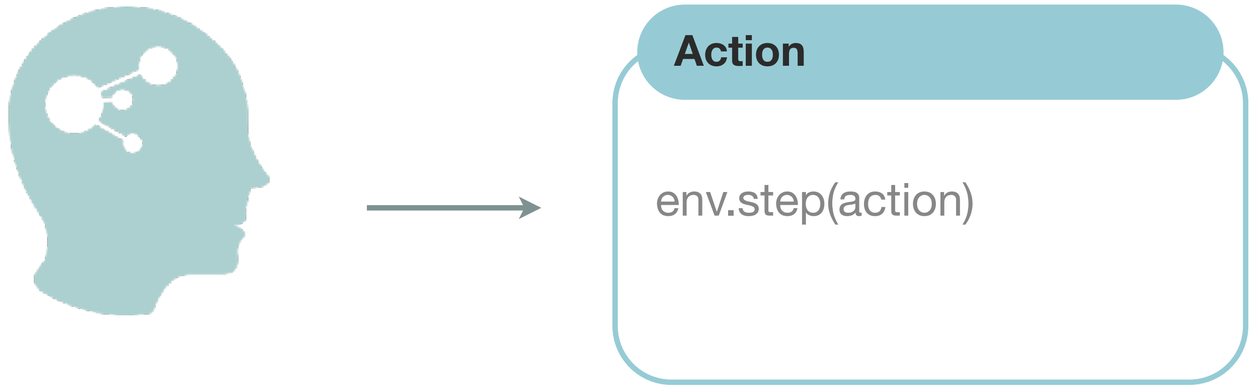
Model

Fit the scene and contacts

Test

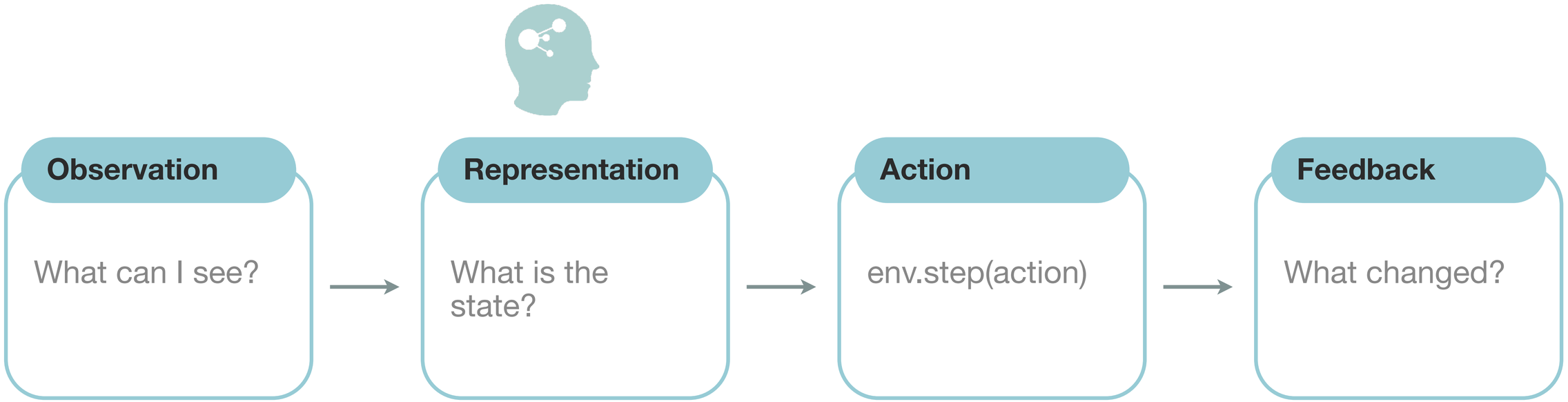
Replay and compare

Is executable code enough?



A valid call. A successful task?

Three capabilities in one agent loop



Spatial intelligence

Ground the representation.

World modeling

Predict the effects of actions.

Embodied intelligence

Act, observe, and adapt.

NEURAL INFORMATION PROCESSING SYSTEMS

Embodied Agent Interface

Benchmarking LLMs for Embodied Decision Making


Manling Li*


Shiyu Zhao*


Qineng Wang*


Kangrui Wang*


Yu Zhou*


Sanjana Srivastava


Cem Gokmen


Tony Lee


Li Erran Li


Ruohan Zhang


Weiyu Liu


Percy Liang


Fei-Fei Li


Jiayuan Mao


Jiajun Wu

* Equal contribution

What is a generalist Embodied Agent



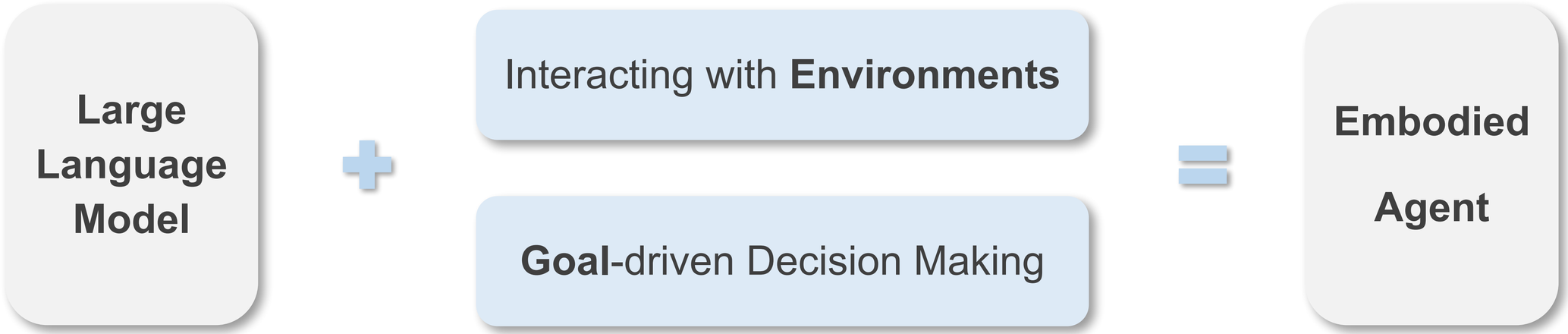
Having an agent that can do **many tasks**, across **many environments**.

Goal-driven Decision Making

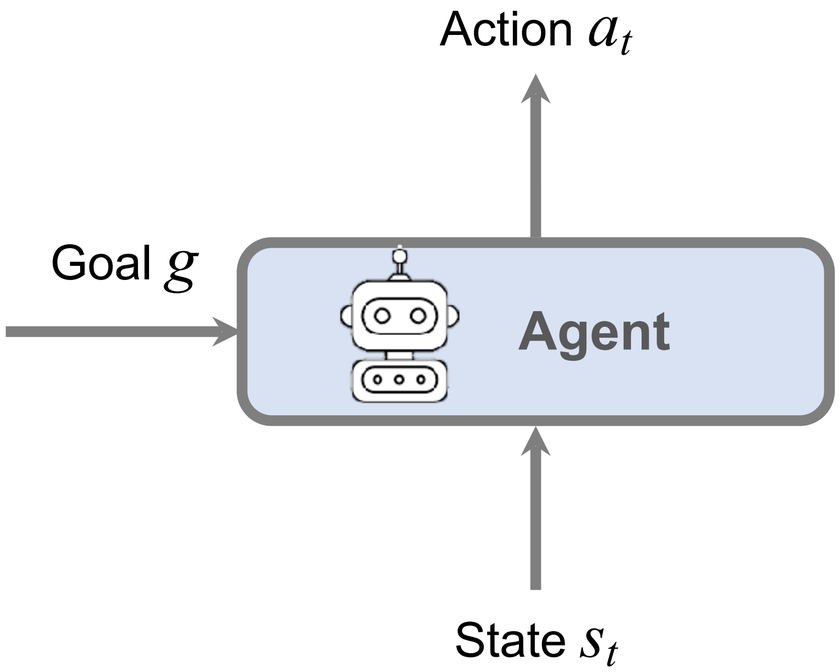
Interacting with **Environments**

Figure credits: Jiayuan Mao

Building Embodied Agents



Markov Decision Process



Output

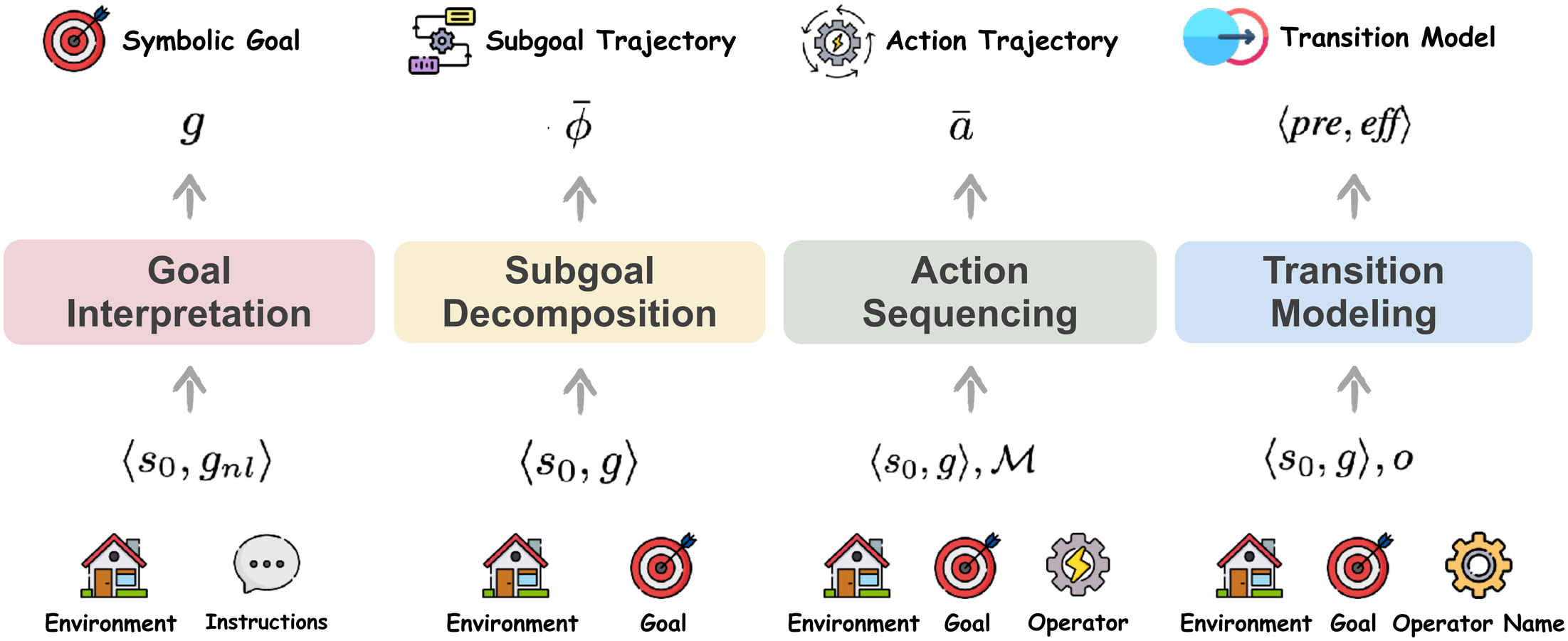
$<a_1, a_2, \ldots, a_t>$ **Action Sequencing**

$<s_1, s_2, \ldots, s_t>$ **Subgoal Decomposition**

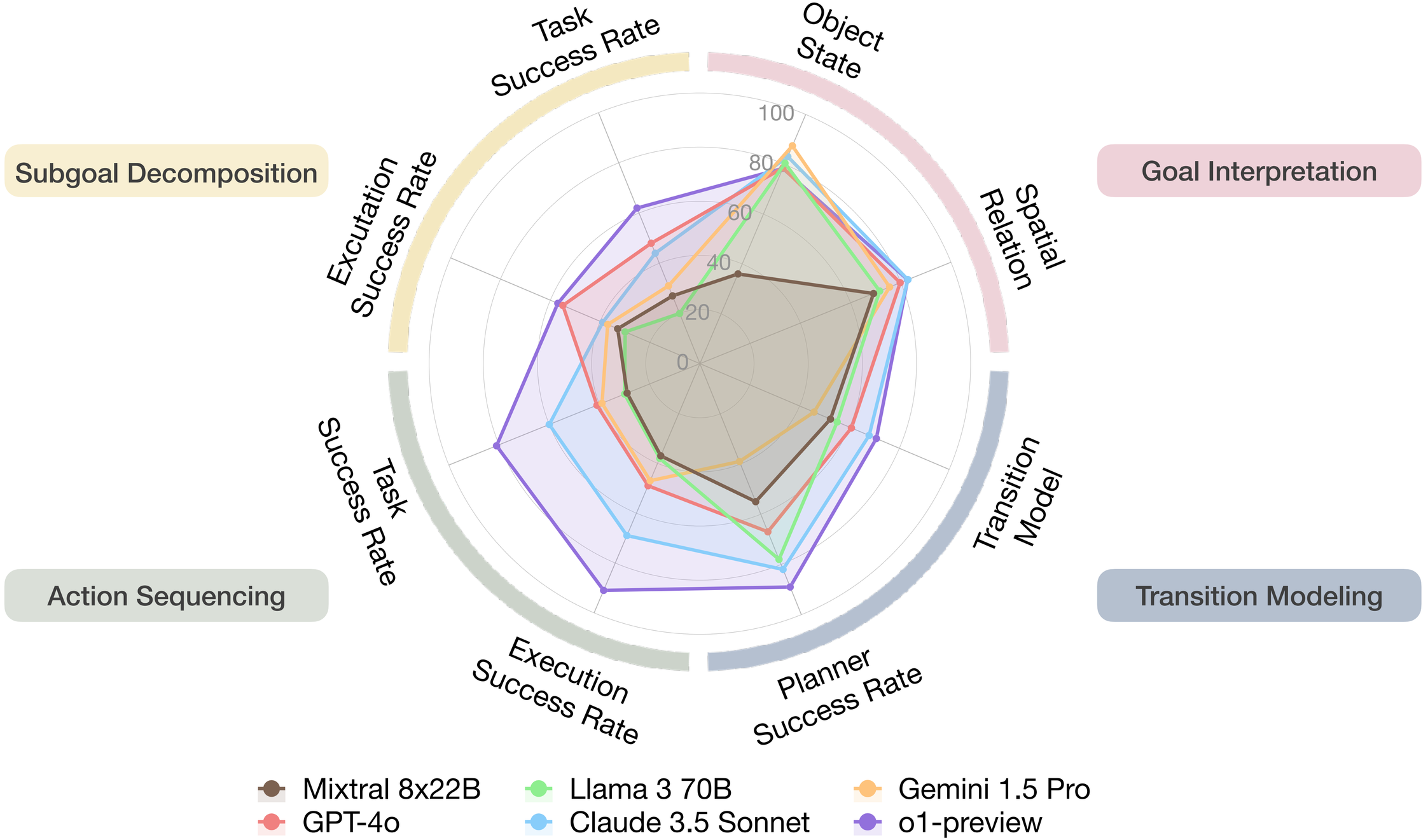
Learn
$(s_t, a_t) \rightarrow s_{t+1}$ **Transition Modeling**

Input
$<g, s_0>$ **Goal Interpretation**

Standardization of modules and interfaces



Transition modeling and subgoal decomposition remain difficult



EmbodiedBench

ICML 2025 · Oral

Comprehensive Benchmarking Multi-modal Large Language Models for Vision-Driven Embodied Agents

UIUC Northwestern University of Toronto Toyota Technological Institute at Chicago



Rui Yang*



Hanyang Chen*



Junyu Zhang*



Mark Zhao*‡



Cheng Qian



Kangrui Wang



Qineng Wang

TVK

Teja Venkat Koripella

MM

Marziyeh Movahedi‡



Manling Li



Heng Ji



Huan Zhang†



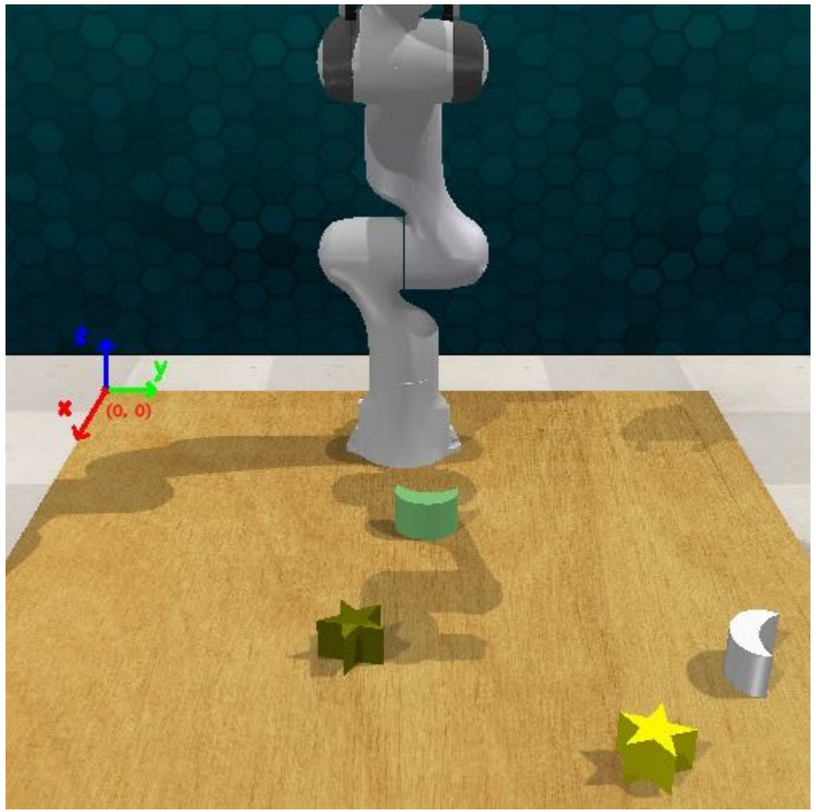
Tong Zhang†

* Equal contribution. † Equal advising. ‡ Work done during internship at UIUC.

What should the robot do next?

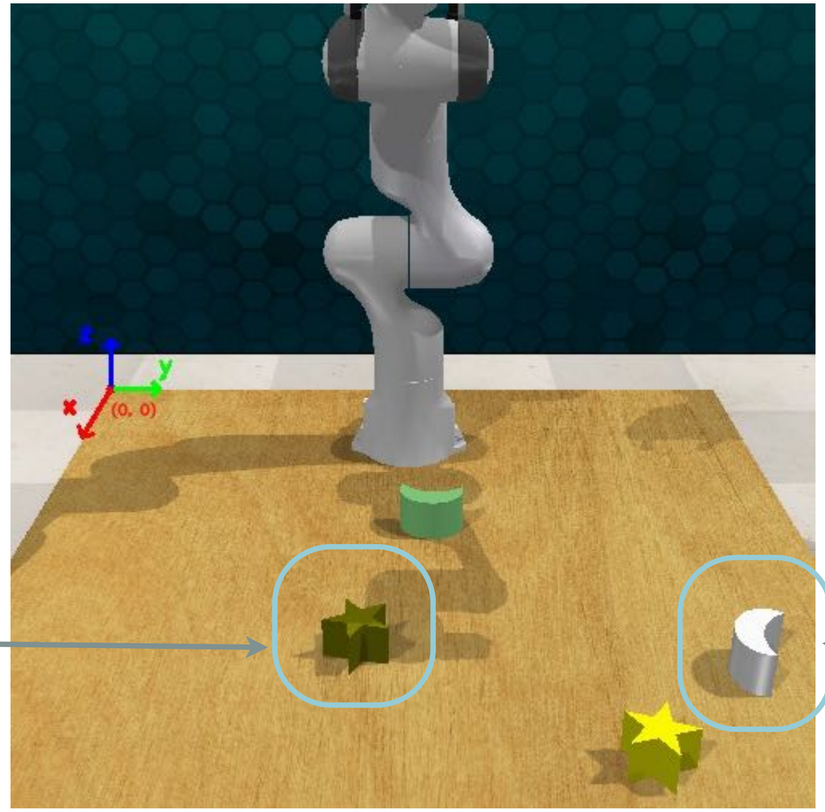
Stack the right moon on top of the left star.

First: which two objects?



The instruction has to become spatial.

Stack the right moon on top of the left star.

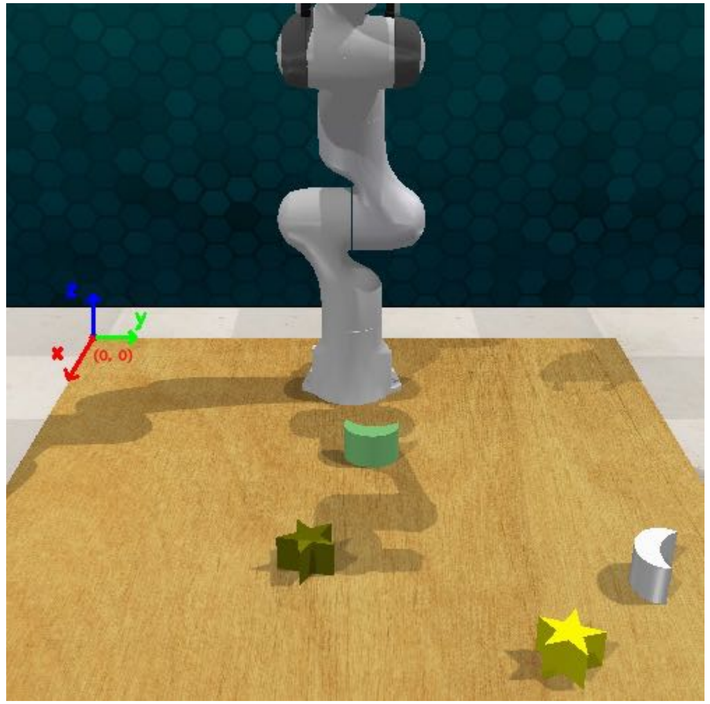


left star

right moon

Now make that relation true.

Stack the right moon on top of the left star.



Initial observation

on top of



Successful final state

The interface changes the agent's job.

EB-ALFRED

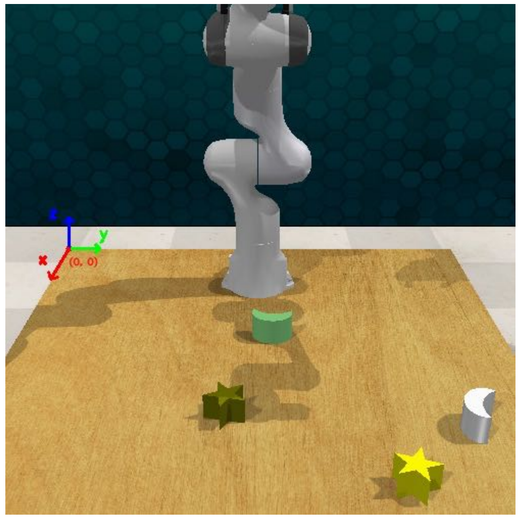


Place a sauce pan with a spatula on the kitchen counter.

find • pick up • put down

Choose and sequence available skills.

EB-Manipulation



Stack the right moon on top of the left star.

[x, y, z, roll, pitch, yaw, gripper]

Specify movement, rotation, and gripper state.

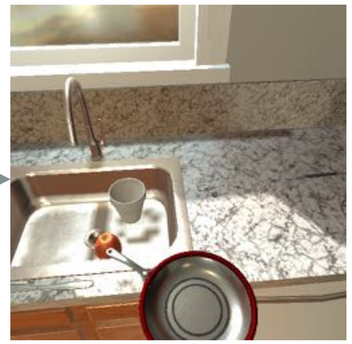
The action failed. What happens next?

Place a sauce pan with a spatula on the kitchen counter.

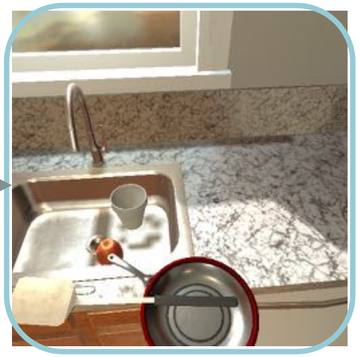
Step 1



Step 7



Step 11



Failed action

Stop? Repeat? Replan?

Observe again. Replan. Continue.

Place a sauce pan with a spatula on the kitchen counter.

Step 1 → Step 7 → Step 11 → Step 14 → Step 20



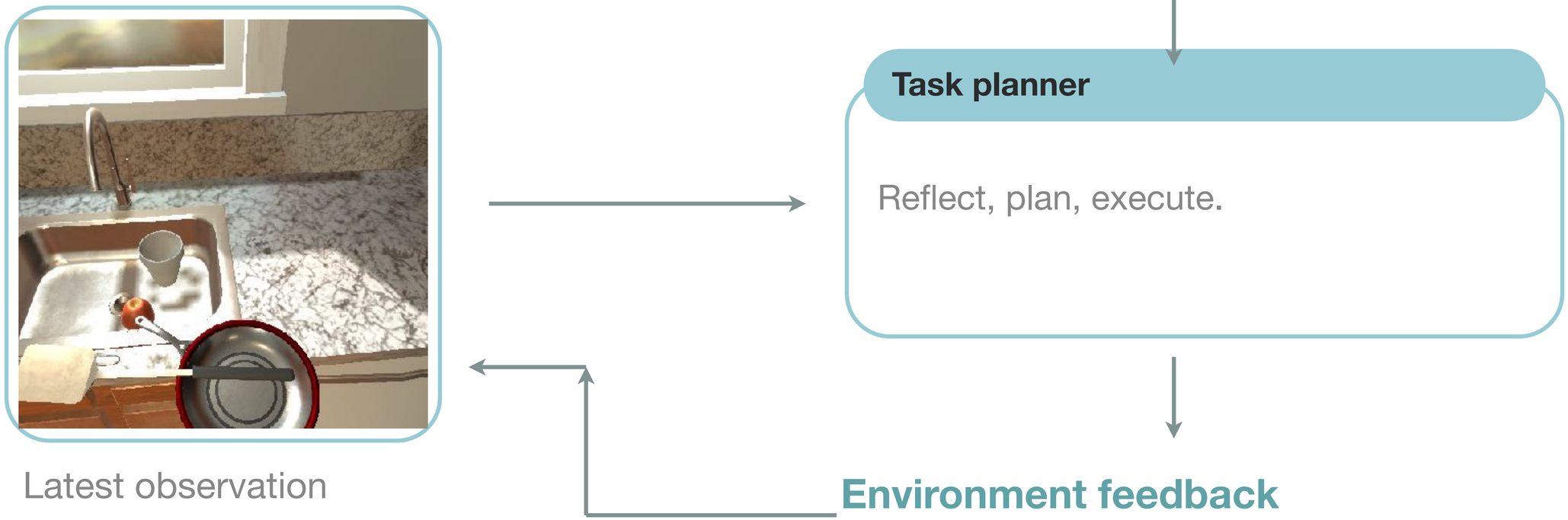
Failed action

Replanning

Success

Use the new observation to revise the plan.

Keep the goal, history, and feedback.



The next action uses the updated state.

Did it move two different books?

Move two books from the bed to the desk.



First book on desk

Find a book again

Back at the desk

Two pick-ups. Two different objects?

It moved the same book again.

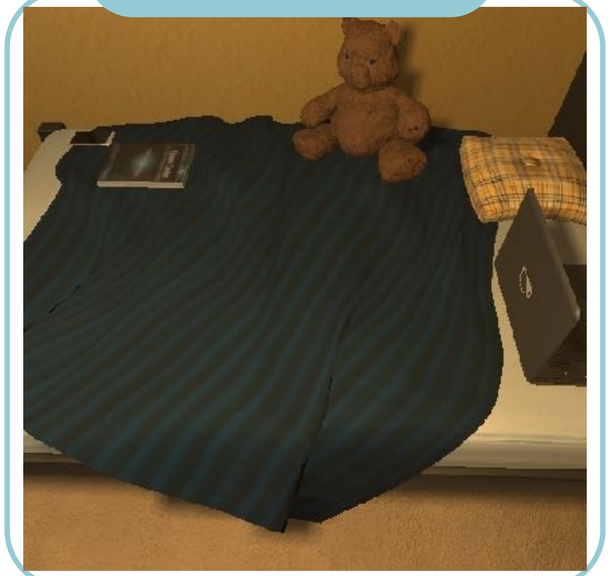
Move two books from the bed to the desk.

Book 1



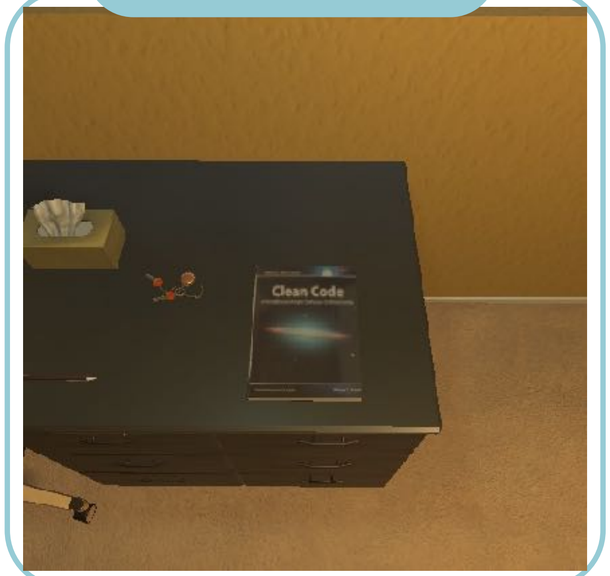
First book on desk

Book 2



Find a book again

Book 1 again



Back at the desk

Object identity belongs in the state.

EmbodiedBench · EB-ALFRED failure example

Four environments. Two action levels.

High-level skills

EB-ALFRED



Put the books on the desk.

EB-Habitat



Find an orange on the TV stand and move it to the sink.

Low-level actions

EB-Navigation



Navigate to the laptop and stay close.

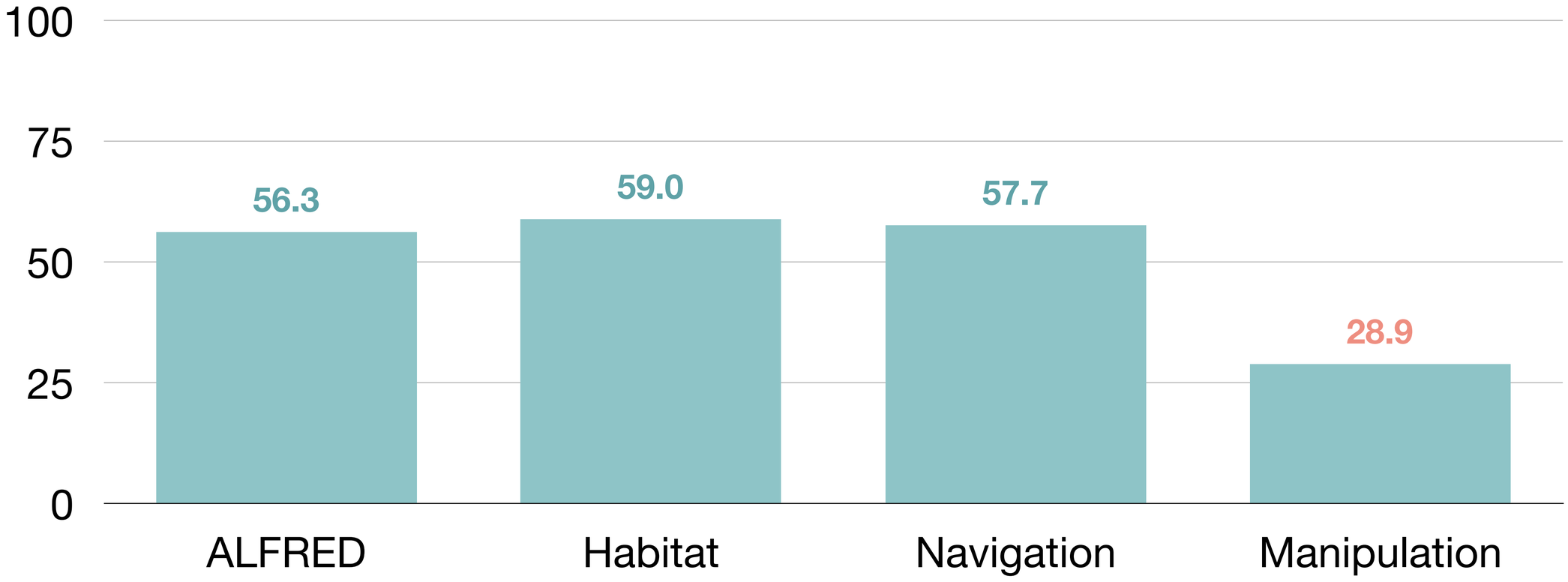
EB-Manipulation



Pick up the star and place it into the silver container.

EmbodiedBench: 1,128 task instances; reported ICML 2025 benchmark.

Manipulation remains the bottleneck.



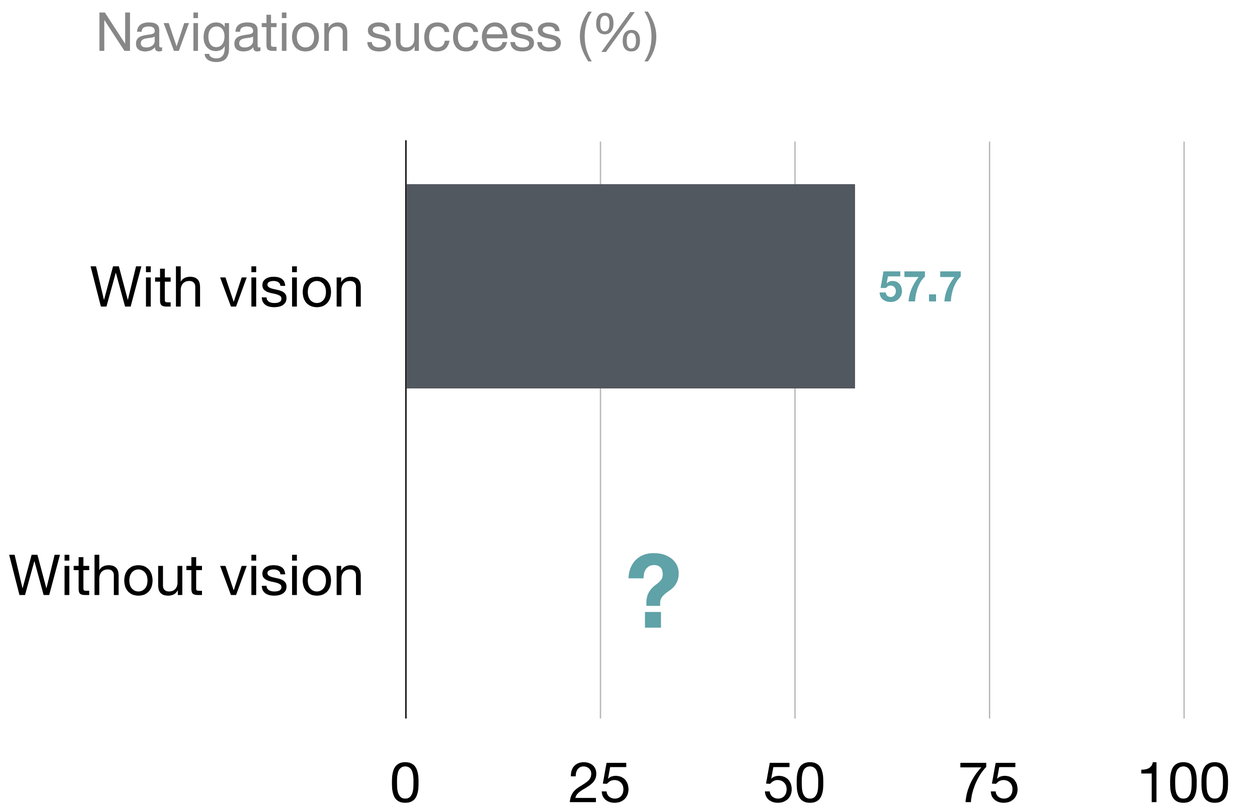
28.9% manipulation success

EmbodiedBench · GPT-4o · overall task success (%) · ICML 2025

What happens if we remove vision?



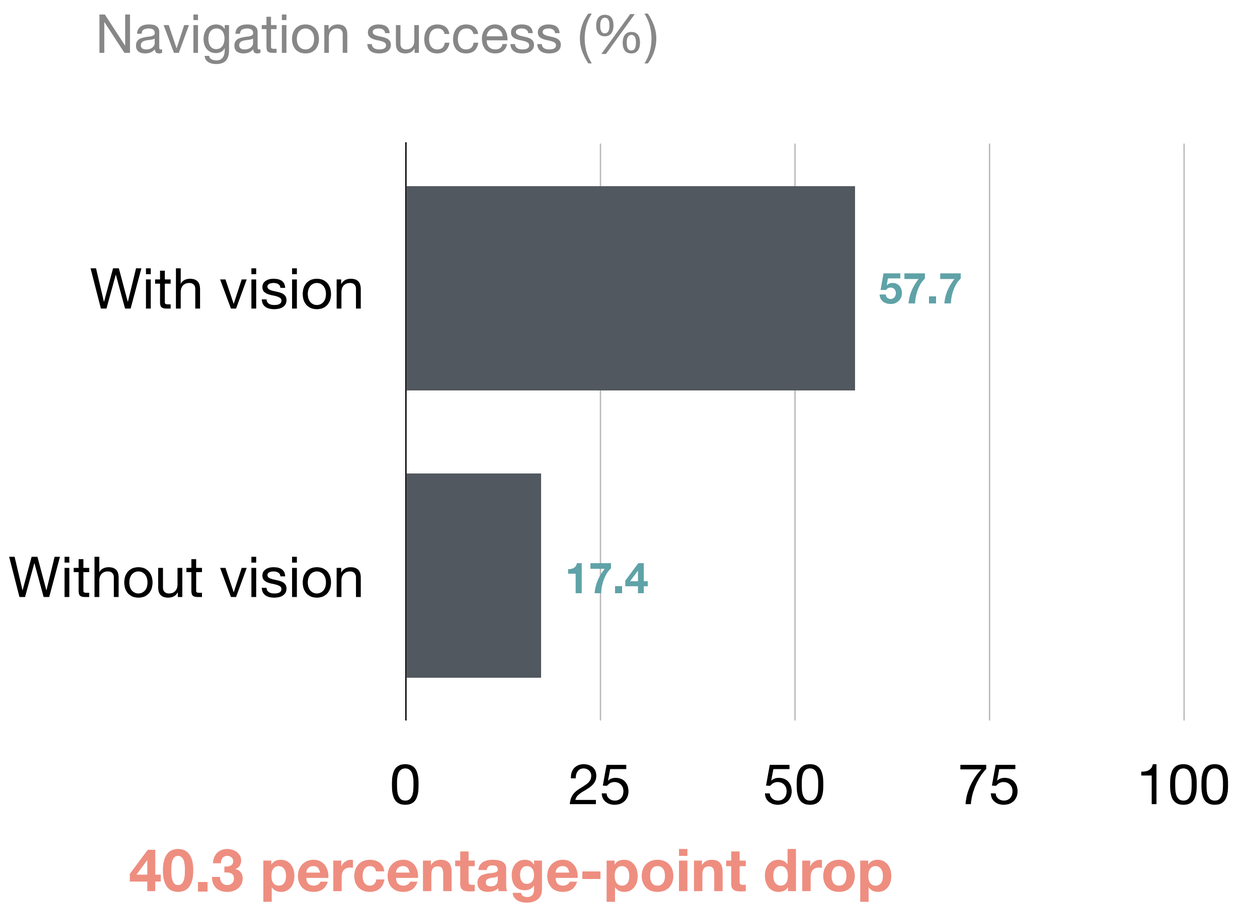
Navigate to the laptop and stay close.



Navigation success falls to 17.4%.



Navigate to the laptop and stay close.



40.3 percentage-point drop

Vision dependence differs by task.

Task	With vision	Without vision	Change
EB-ALFRED	56.3%	58.0%	+1.7 pp
EB-Navigation	57.7%	17.4%	**–40.3 pp**

Grounding demands depend on the task and interface.

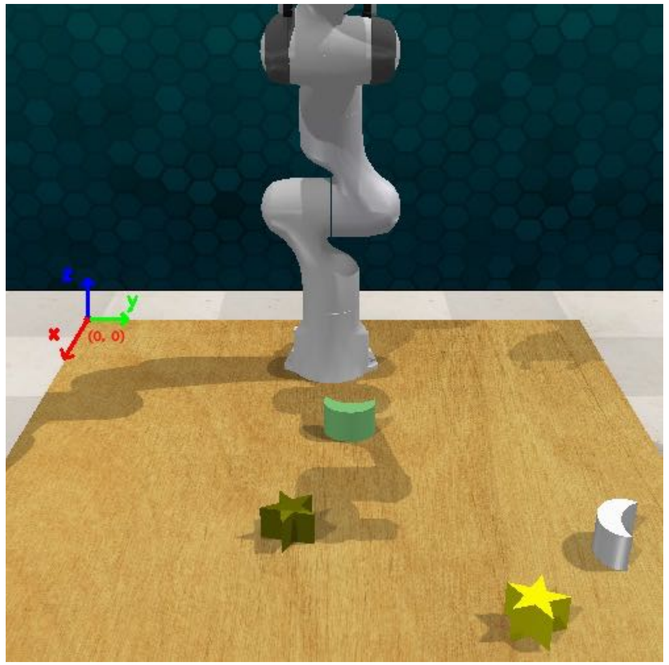
EmbodiedBench · GPT-4o · overall success · ICML 2025

Who does the spatial work?

Where should I act?

What will my action change?

Did it work—and what next?



The action interface decides how much remains for the agent.

MindCube

ICLR

NYU

Spatial Mental Modeling from Limited Views


Qineng Wang*


Baiqiao Yin*


Pingyue Zhang


Jianshu Zhang


Kangrui Wang


Zihan Wang


Jieyu Zhang


Keshigeyan Chandrasegaran


Han Liu


Ranjay Krishna


Saining Xie


Jiajun Wu†


Fei-Fei Li†


Manling Li†

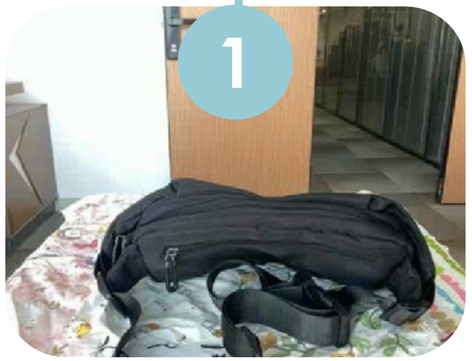
*Alphabetical Order. †Co-Advising.

Can VLMs build **spatial mental models** like humans?


3

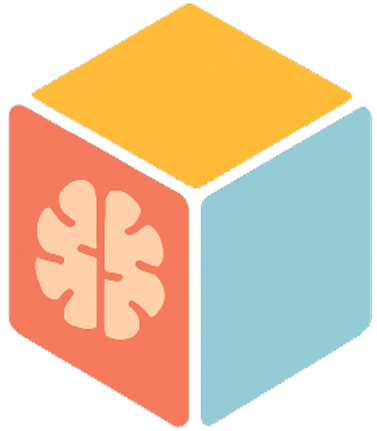

2


4


1

At 4 what is to the **left** of the **black bag** ?

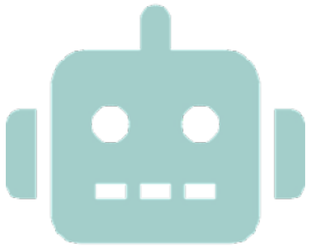
A. Window B. Door C. Desk D. Sofa





At 4 what is to the **left** of the **black bag** ?

A. Window B. Door C. Desk D. Sofa



VLMs

4



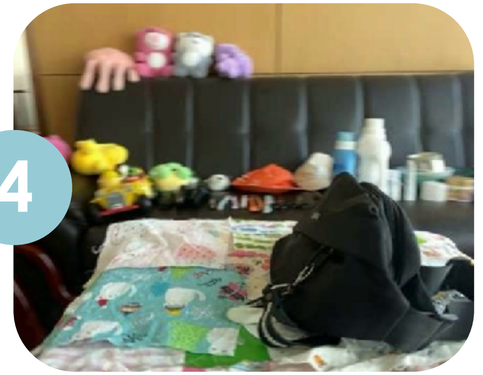
At 4 what is to the **left** of the **black bag** ?

A. Window B. Door C. Desk D. Sofa

?

Cognitive Map

4



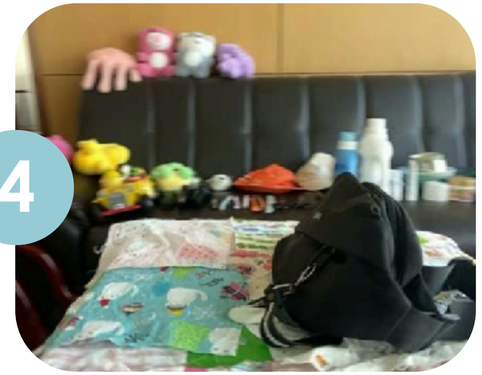

1


4

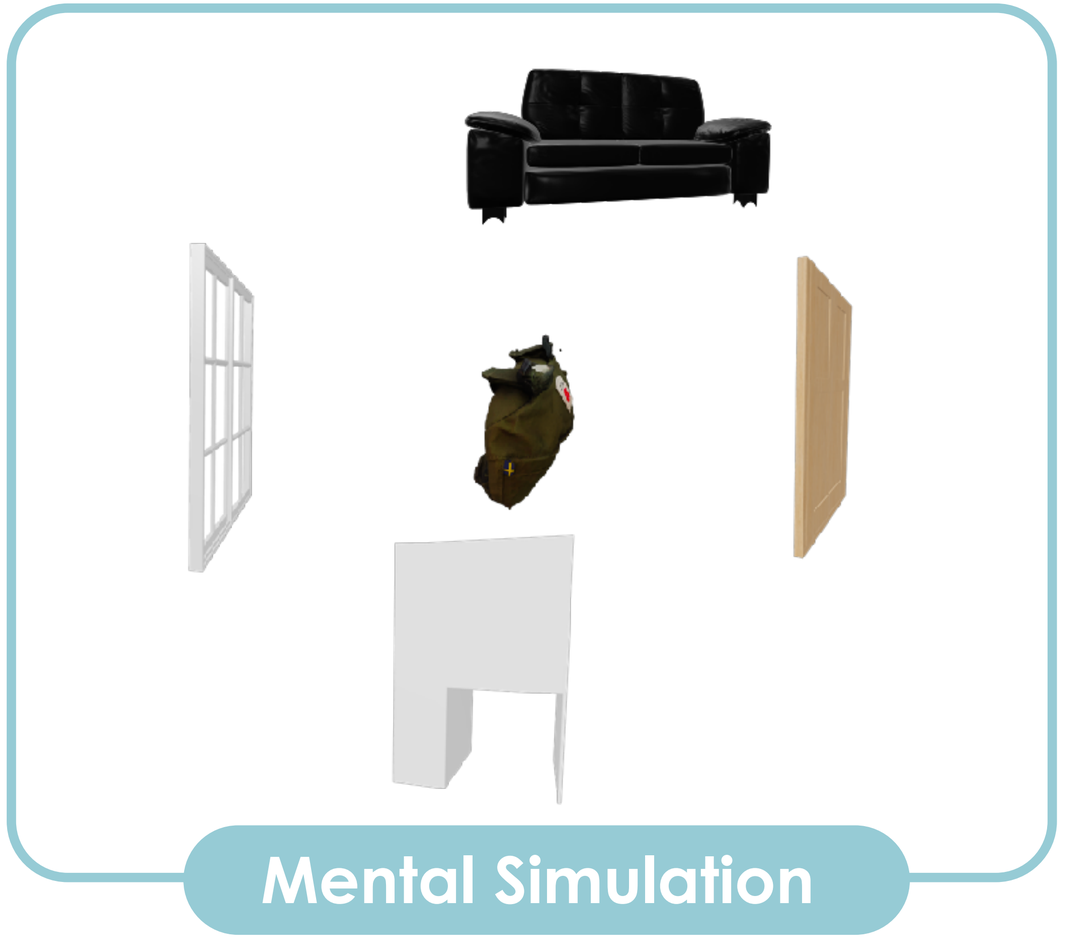
At 4



Mental Modeling



At 4 what is to the **left** of the **black bag** ?

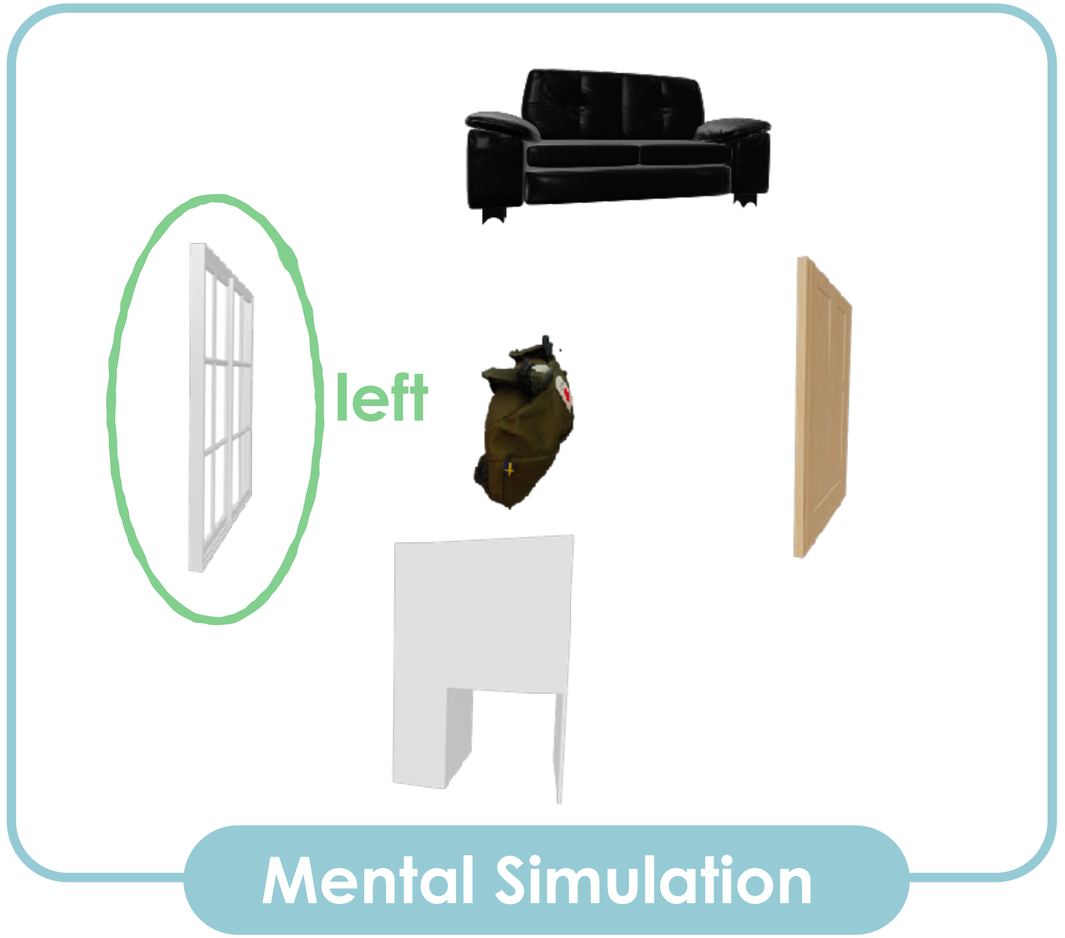


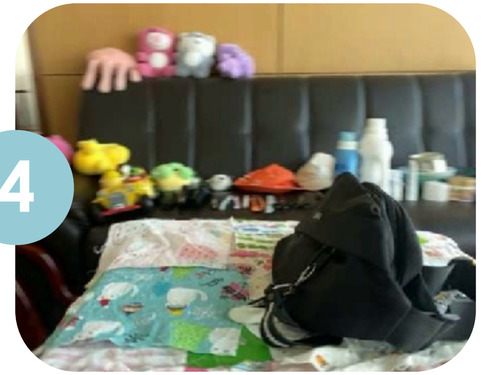
Mental Simulation



At 4 what is to the **left** of the **black bag** ?

A. Window ✓ B. Door C. Desk D. Sofa

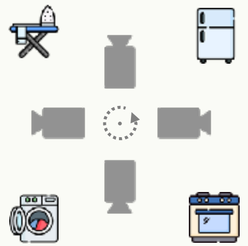





MindCube
21,154 questions
3,268 images

Three Movement Patterns in MindCube

Rotation



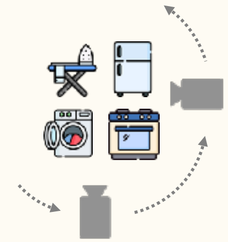








Around

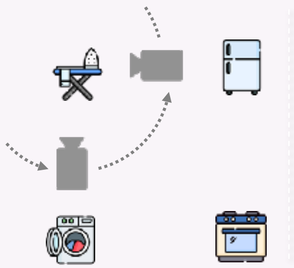








Among



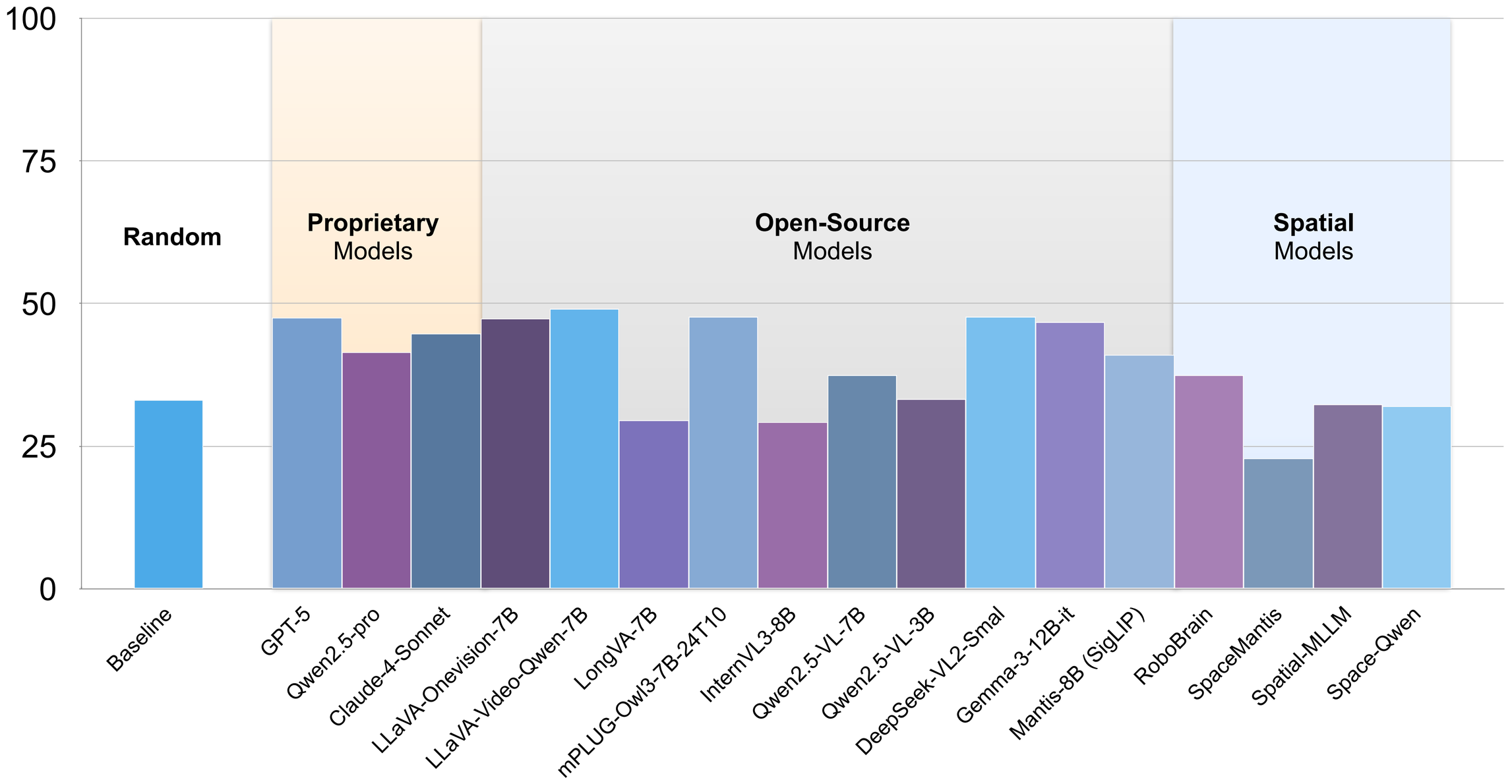








Still Challenging !

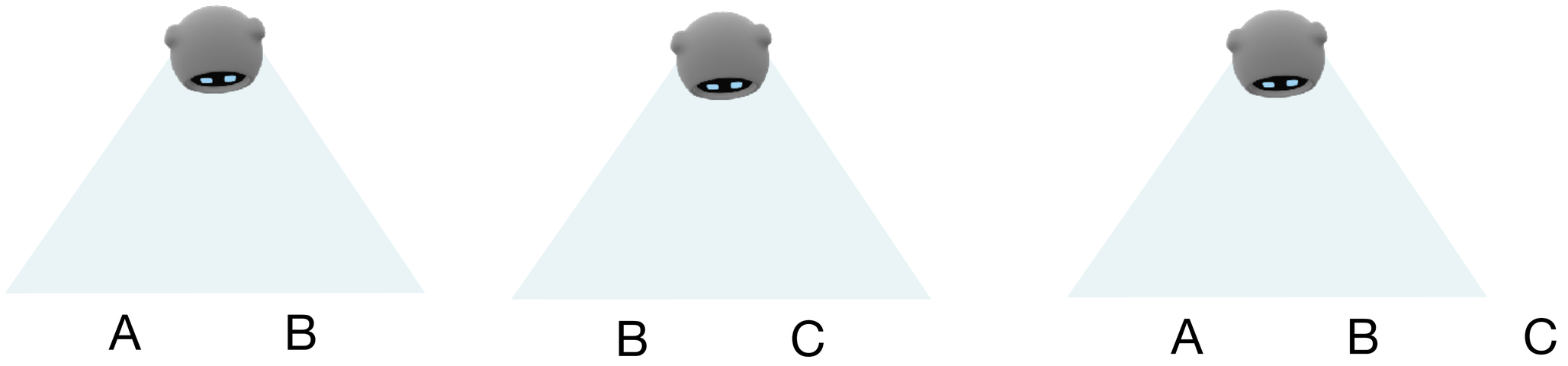


How to teach VLMs to

Approximate Spatial Mental Models?

Spatial Mental Models are not “CONCRETE”

left(A, B) left(B, C) → left(A, C) is True



Approximate Spatial Mental Models in Frozen VLMs

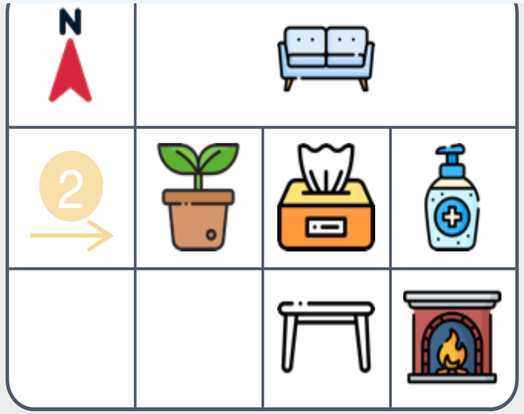
View Interpolation



+0.10%

Reasoning and Cognitive Map are more effective scaffolds

Cognitive Map



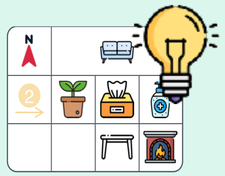

```
{"objects": [{
    "name": "Tissue box",
    "position": [5, 5]
    }, {
    "name": "Hand sanitizer",
    "position": [7, 5]
    }, ...],
 "views": [{
    "name": "View 1",
    "position": [5, 6],
    "facing": "up"
    } ...
```


+3.52%

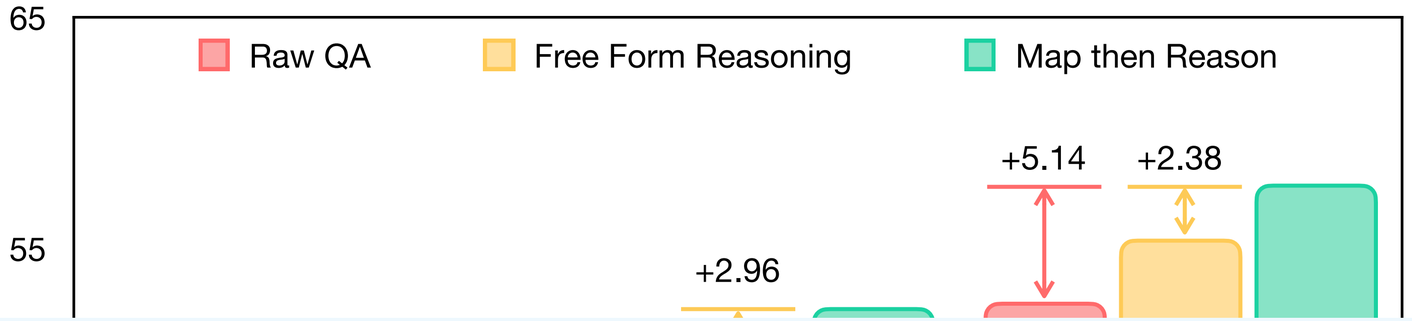
Data Scaffolds

Compare to Raw QA

Approximate Spatial Mental Models



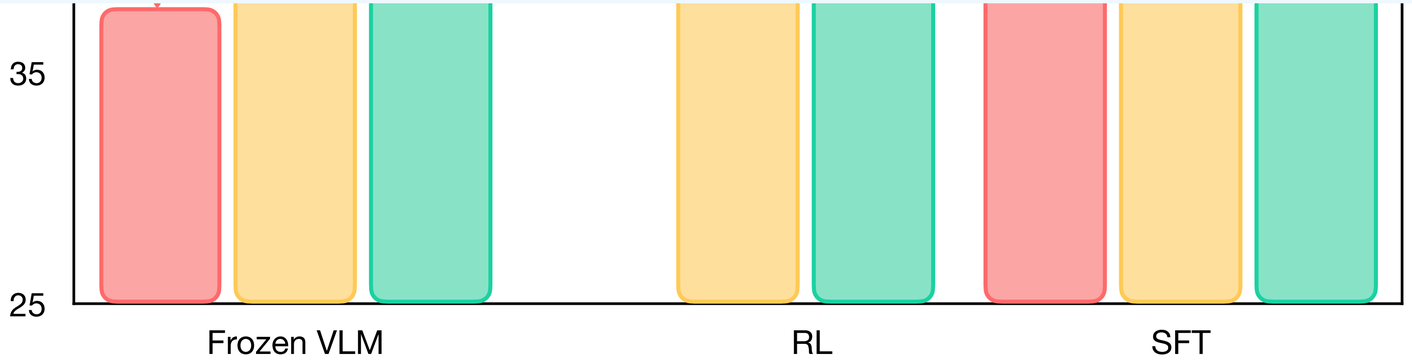
Map then Reason



“Map then Reason” is the most effective approximation



Raw QA



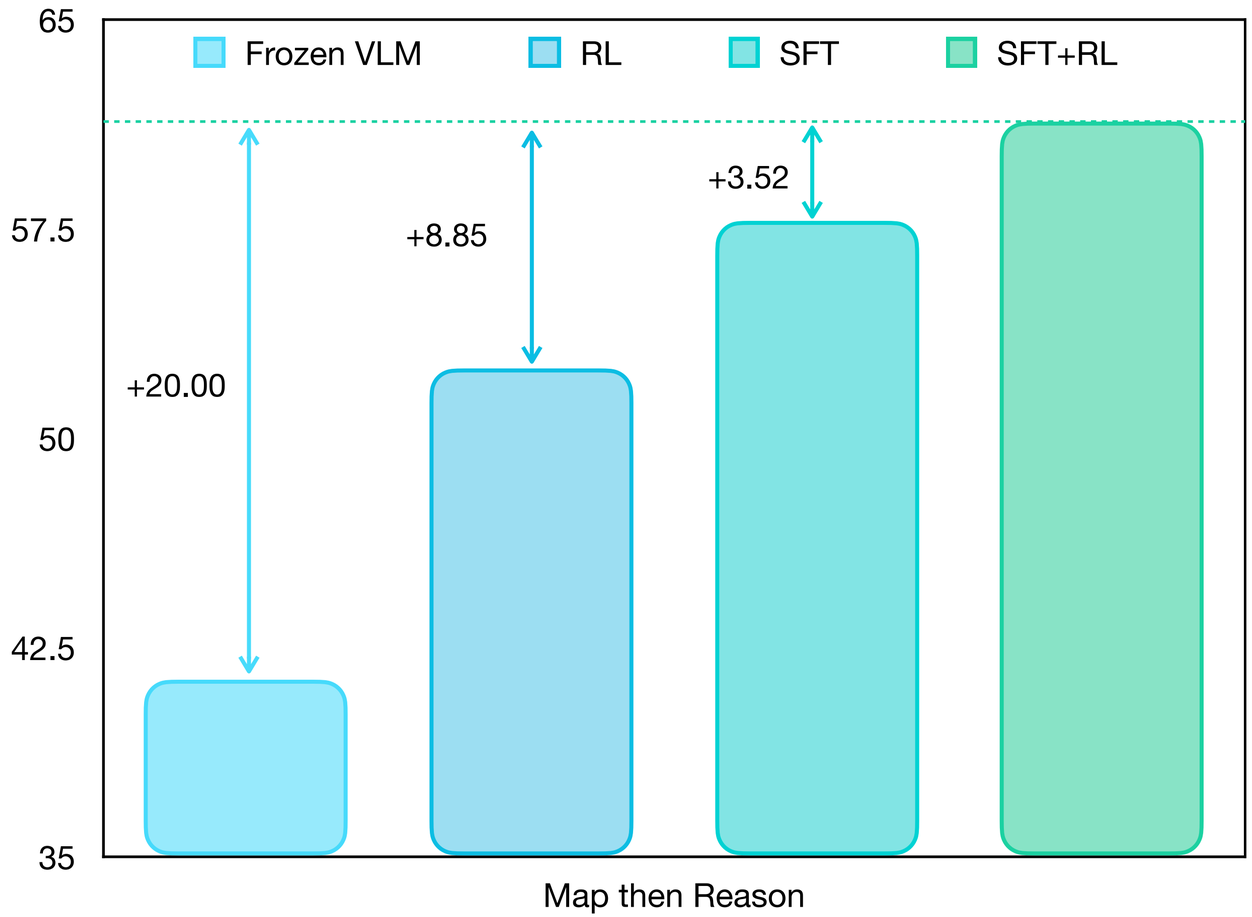
SFT then RL can boost the performance

SFT+RL

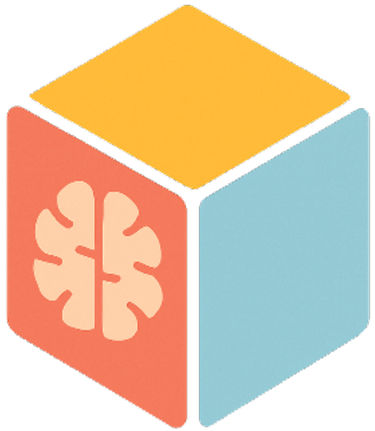
Frozen VLM

RL

SFT







MindCube

https://mll-lab-nu.github.io/mind-cube

ActionEQA: Action Interface for Embodied Question Answering


Tianwei Bao*


Qineng Wang*


Kangrui Wang


Mingkai Deng


Guangyi Liu


Jiayuan Mao


Larry Birnbaum


Zhiting Hu


Eric P. Xing


Zhaoran Wang


Manling Li

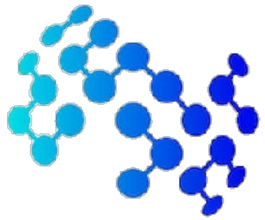












MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

What will happen next?



Current state

“Put the lemon on the plate.”



?

Which next state matches the action?

Current state



“Put the lemon on the plate.”



A



B



C



D

Predict the outcome



“Close the microwave.”



Current state

Next state

State + action → **Next state**

What action caused this change?







Before

After

Action = ?

What action caused this change?







Before

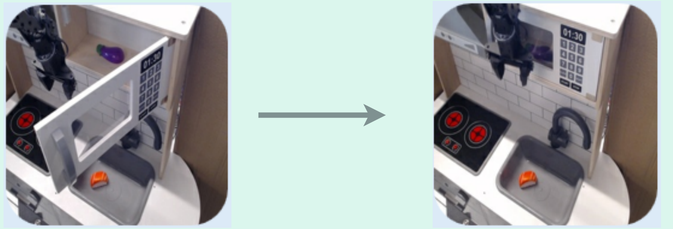
After

“Put the candy bar on the left side of the first shelf.”

At what level do we understand an action?

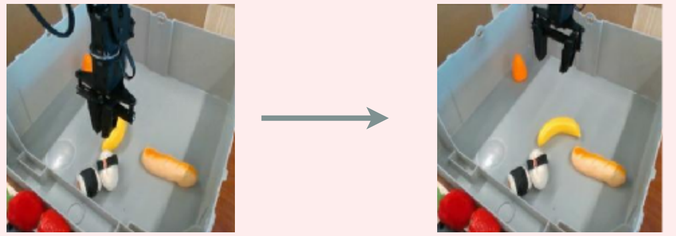
Task

“Close the microwave.”



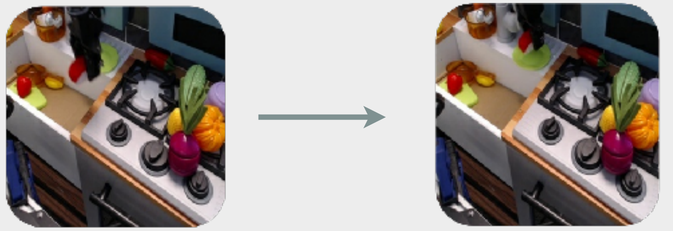
Motion

Move along +X. Open the gripper.



Control

[Δx, Δy, Δz, Δroll, Δpitch, Δyaw, Δgrip]



Two directions × three action levels

Action level	Predict the next state	Infer the action
Task	Goal-conditioned outcome	Task goal from state change
Motion	Motion-conditioned outcome	Motion from state change
Control	Control-conditioned outcome	Control from state change

Six ways to test action–state understanding.

Start from real robot experience



BridgeData V2



DROID



RT-1

Distractors must test the action





Observed transition

Opposite direction

Wrong axis

Dominant axis removed

Plausible alternatives reveal specific errors.

ActionEQA

8,795

questions about actions and their outcomes

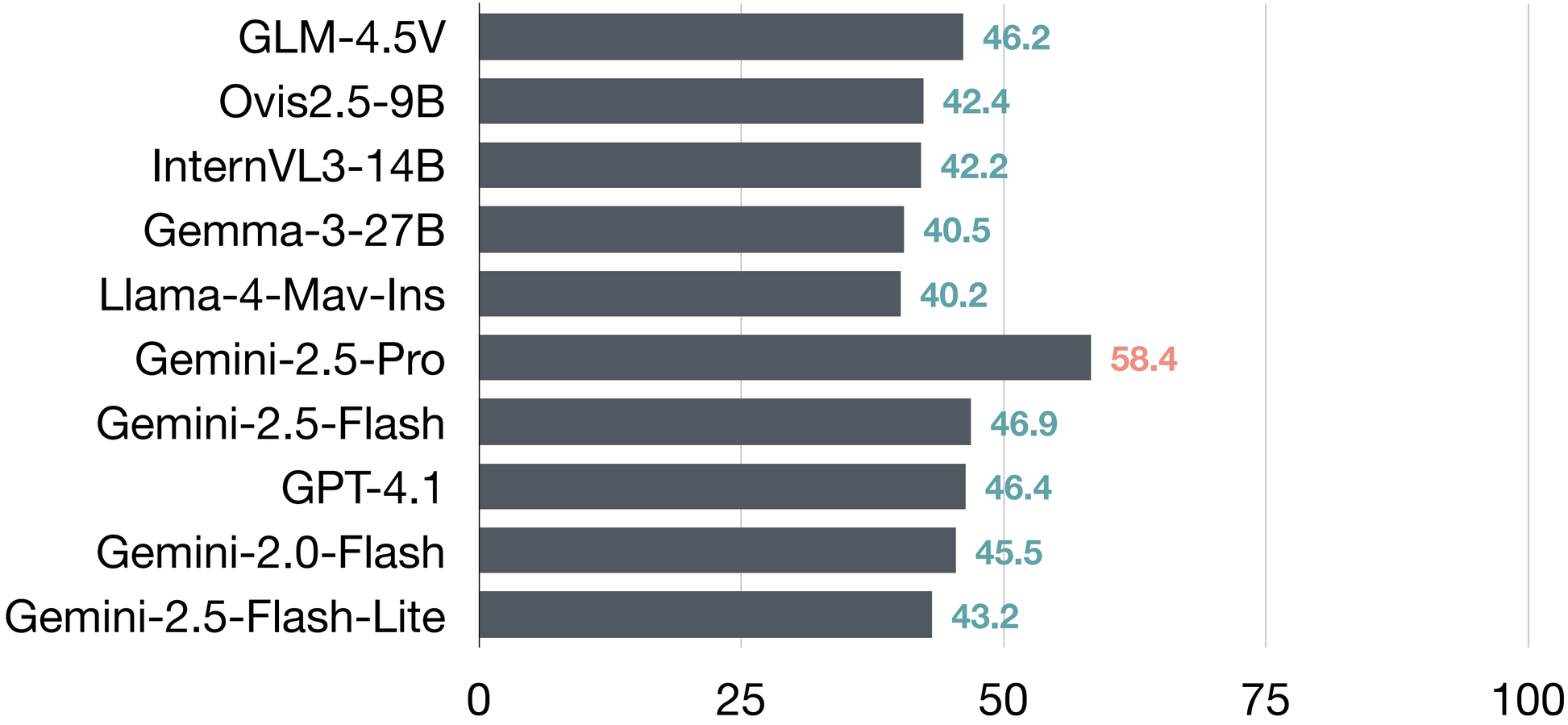
6 tasks

3 real-robot datasets





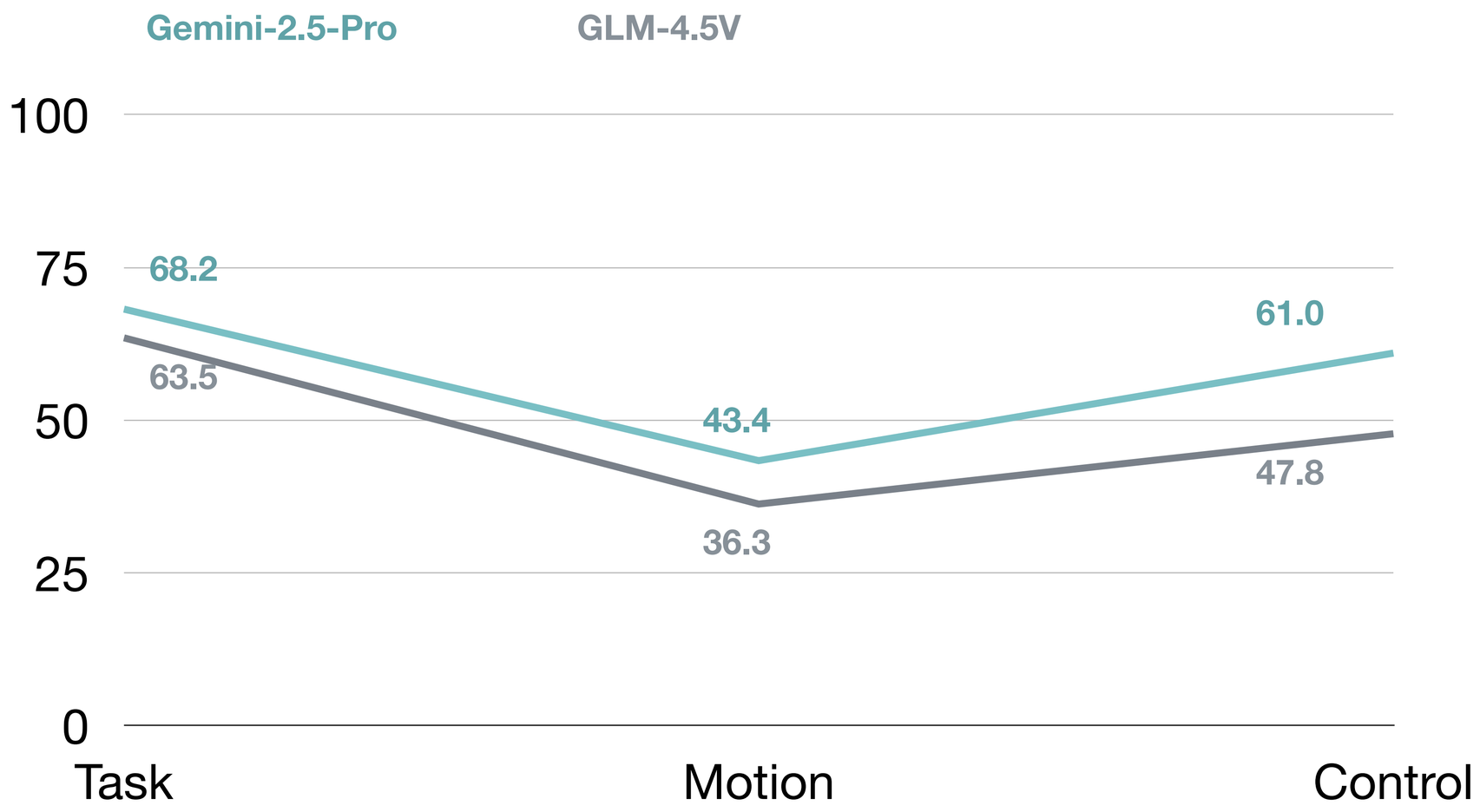
The strongest tested model reaches 58.4%.



What does the overall score hide?

ActionEQA · overall accuracy (%) · 10 evaluated models

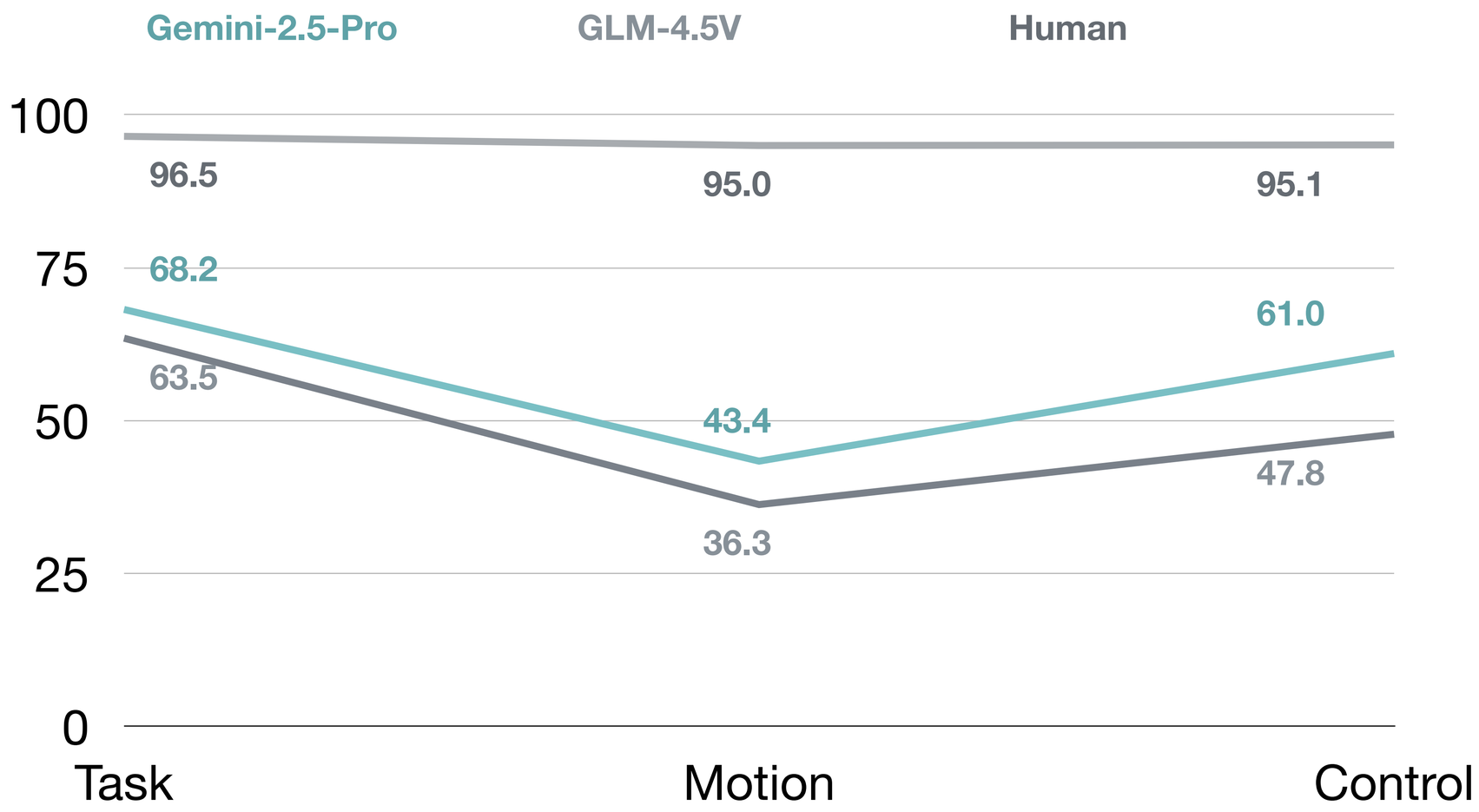
Motion is harder than goals or controls.



Middle-layer bottleneck

ActionEQA · accuracy by action level (%)

The human baseline stays near 95%.



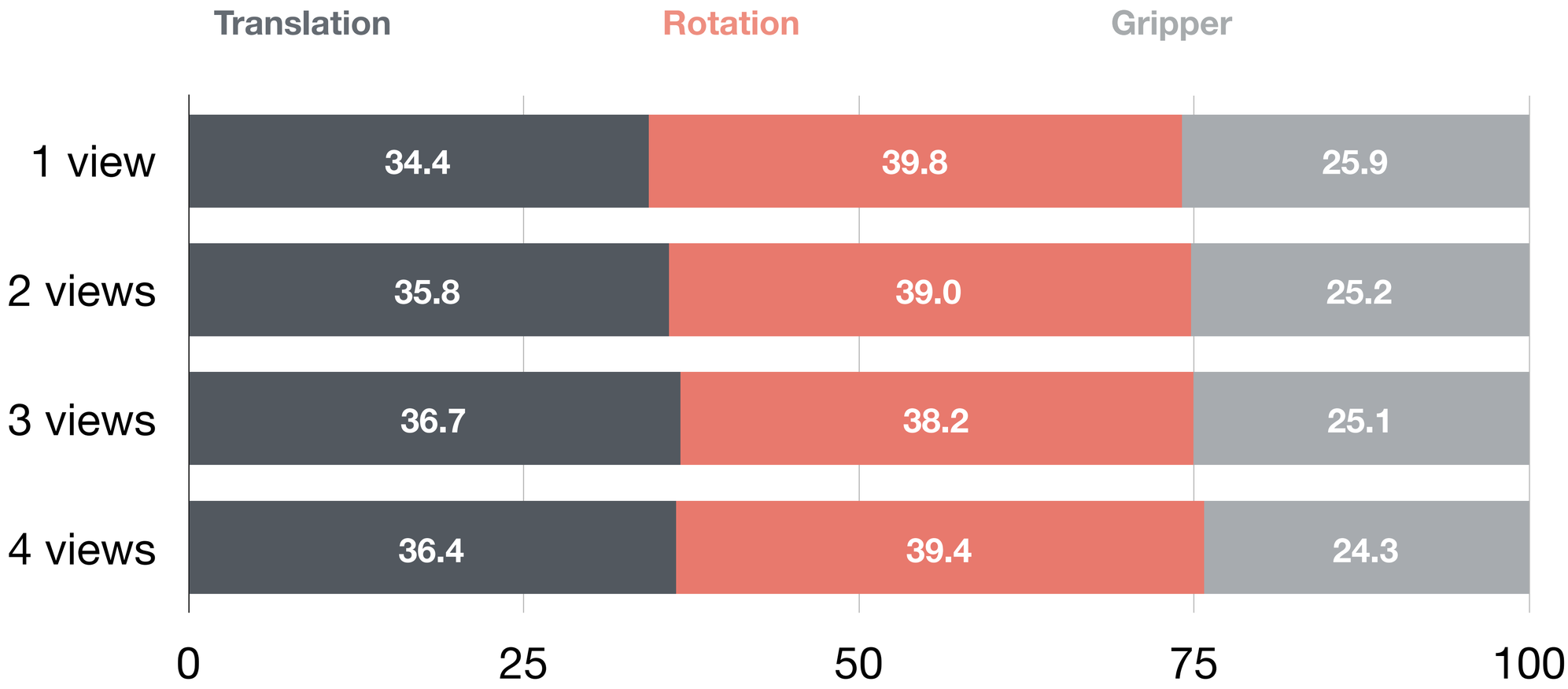
51.6 pp

Human – Gemini

at the motion level

ActionEQA · accuracy by action level (%)

Rotation is the largest error category.



Rotation accounts for 38–40% of categorized errors.

ActionEQA · share of categorized motion errors (%) · 1–4 views

Combining views can weaken the answer.

Task	Level	Ego	Exo	Both	Δ best
State prediction	High	55.4	43.2	48.6	-6.8
	Mid	35.9	34.0	35.5	-0.4
	Low	35.5	34.8	33.7	-1.8
Action inference	High	80.3	56.5	77.0	-3.3
	Mid	37.2	31.6	33.7	-3.5
	Low	51.6	48.6	51.8	+0.2 pp

5 of 6 settings fall below their best single view.

ActionEQA · accuracy (%) · Δ best = combined minus best single view (pp)

A convincing explanation can invent an outcome.

“Pick up the sponge and wipe the plate.”



Current state



Model chose A



Correct answer: B

“most likely a wet mark from the sponge”

Return to our opening question

Current state



“Put the lemon on the plate.”



A ✓



B



C



D

From executable actions to grounded outcomes

Spatial state



Action meaning

“Put the lemon on the plate.”

State transition



What changes when actions form a sequence?

ENACT: Evaluating Embodied Cognition with World Modeling of Egocentric Interaction


Qineng Wang*


Wenlong Huang*


Yu Zhou


Hang Yin


Tianwei Bao


Jianwen Lyu


Weiyu Liu


Ruohan Zhang†

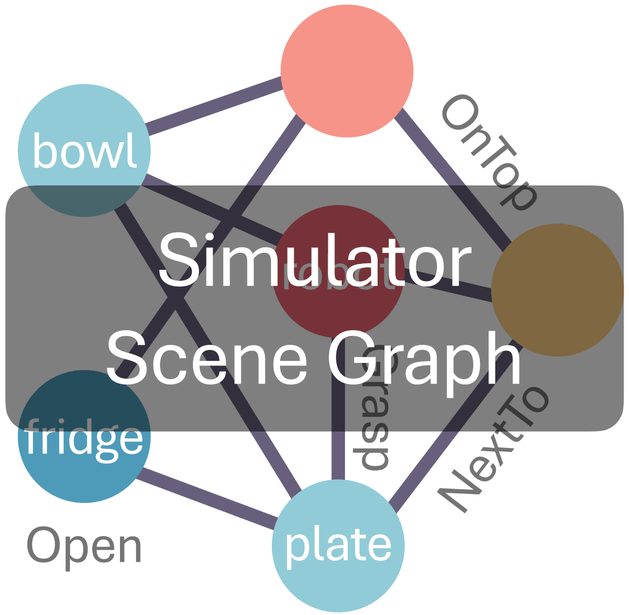

Jiajun Wu†


Li Fei-Fei†


Manling Li†

Casting embodied cognition as POMDP world modeling

ENACT evaluates embodied cognition through POMDP world modeling via egocentric interaction.





diff(s_1 , s_0)

Scene Graph Difference

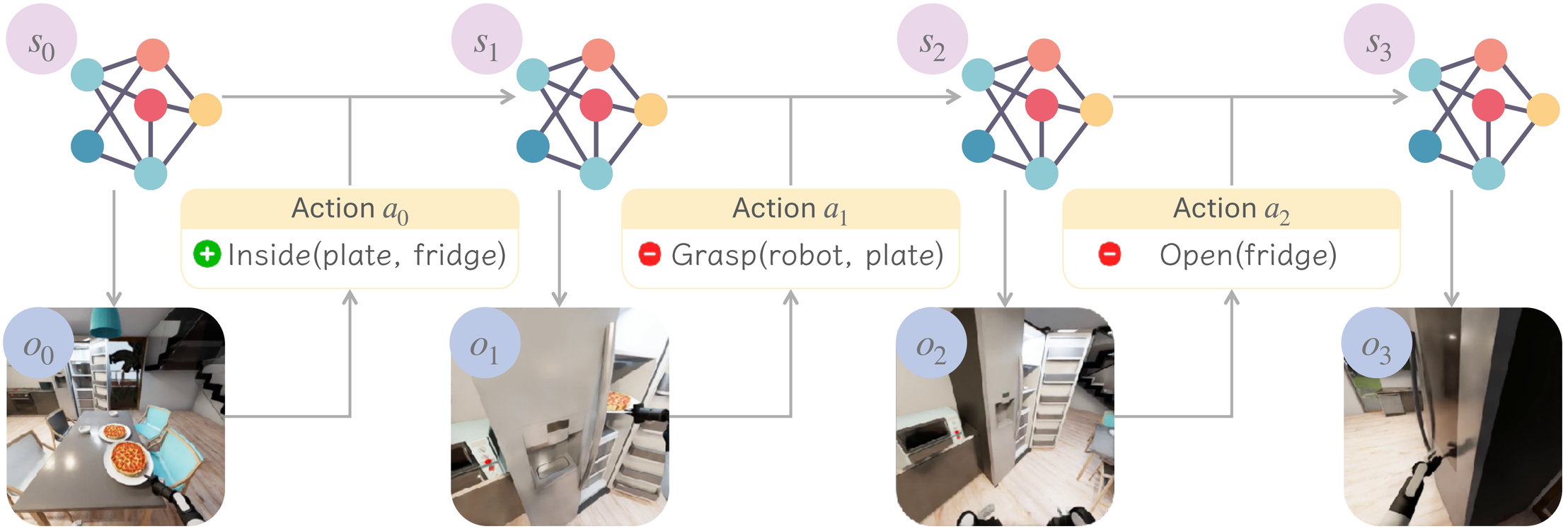
Tomato→Inside of fridge

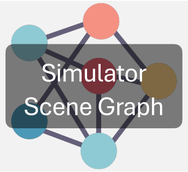
Pizza→On top of plate

Robot→Not holding plate

State Space s Observation Space o Action Space a

Casting embodied cognition as POMDP world modeling

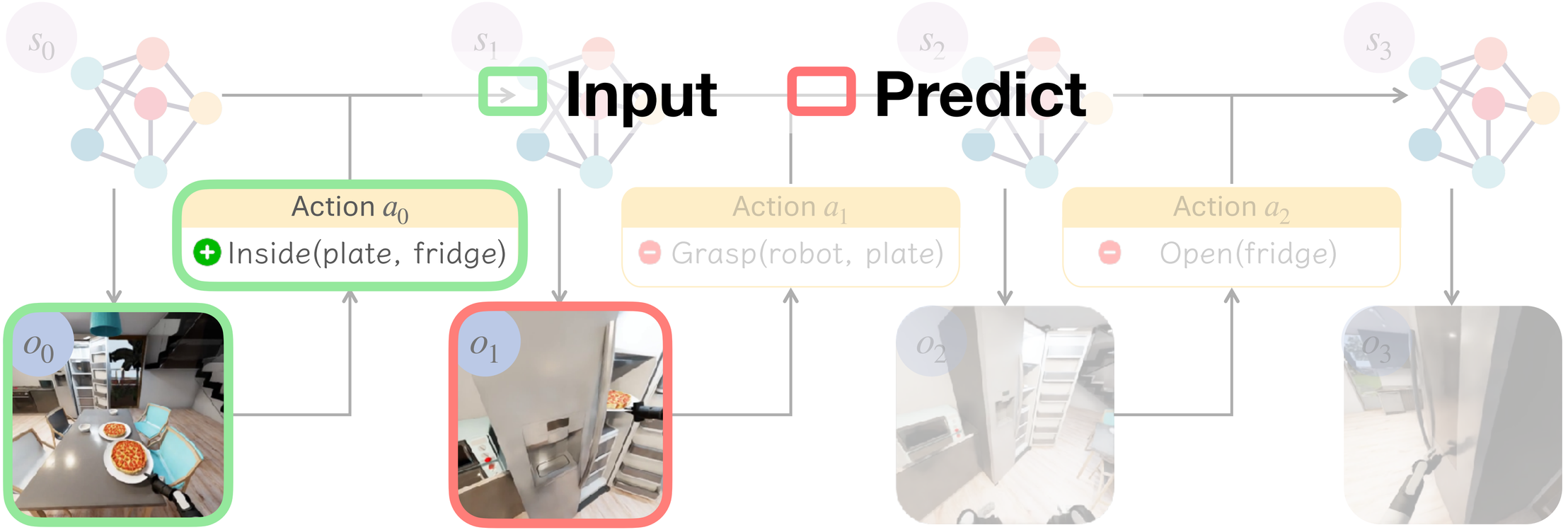


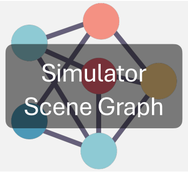
State s 

Observation o 

Action a s_1 - s_0

Forward World Modeling

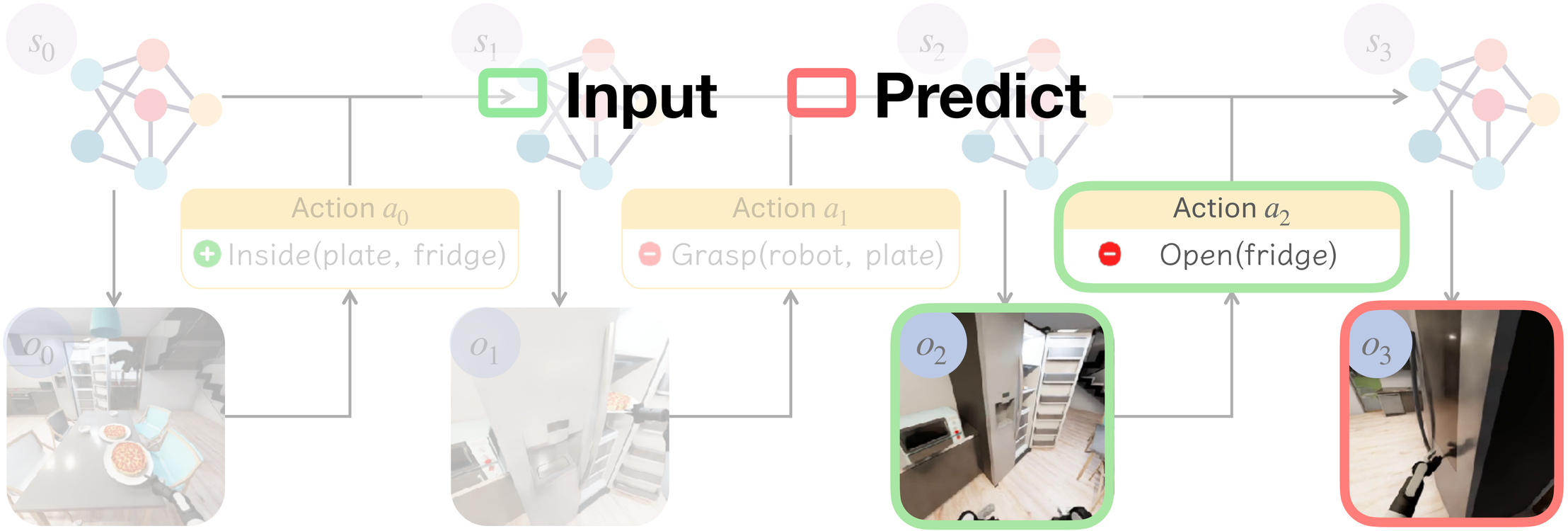


State s 

Observation o 

Action a s_1 - s_0

Forward World Modeling

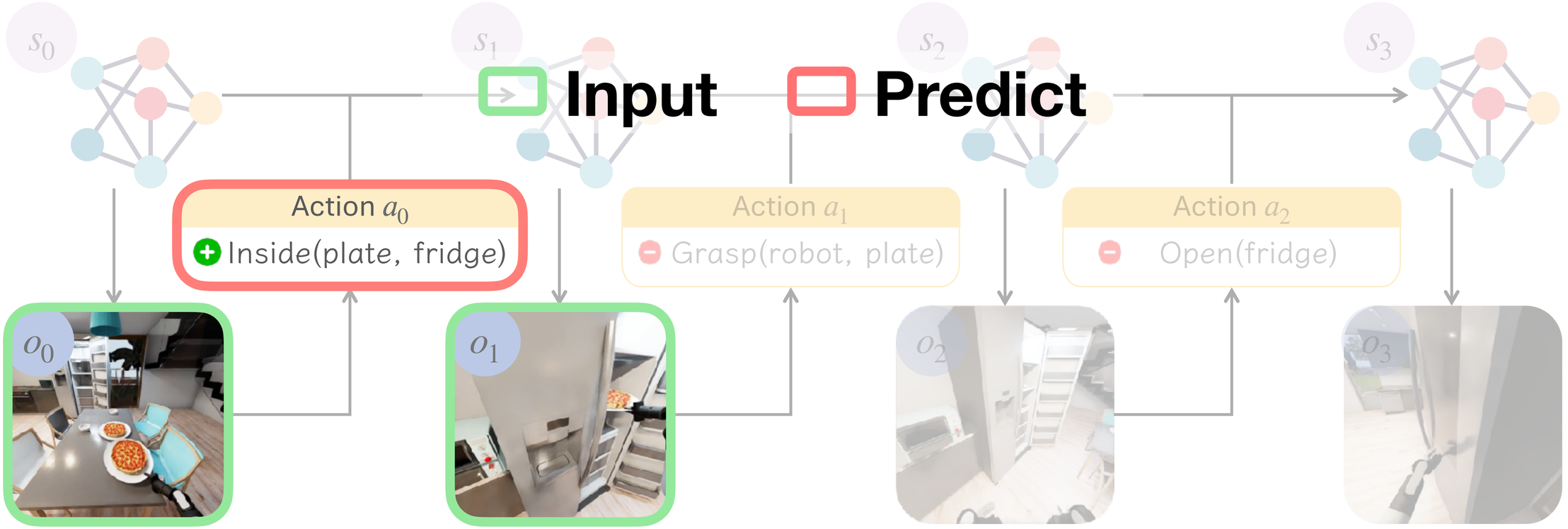


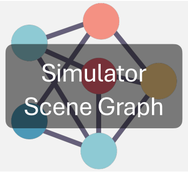
State s — Simulator Scene Graph

Observation o — RGB Image

Action a — s_1 - s_0

Inverse World Modeling



State s 

Observation o 

Action a s_1 - s_0

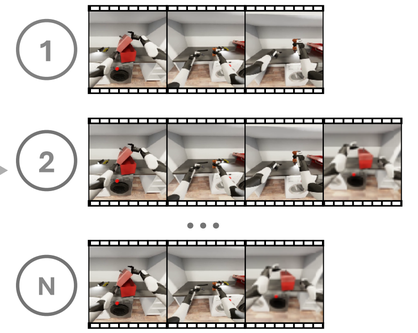
How we collect ENACT dataset?



Mobile Manipulation Demonstrations (30Hz)



Segmented Frames with Abstract State Changes



Filtering and Augmentation

Forward World Modeling QAs

Inverse World Modeling QAs

QA Curation

World Modeling as Reordering Frames Based on Actions

Shuffling

Current State





− OnTop(drill, table)



+ Grasping(driver)



− OnTop(driver, table)



+ Inside(driver, toolbox)

World Modeling as Reordering Frames Based on Actions

Current State





− OnTop(drill, table)



+ Grasping(driver)



− OnTop(driver, table)



+ Inside(driver, toolbox)

VLM Output









And Its Inverse …

Current State





− OnTop(drill, table)



− OnTop(driver, table)



+ Grasping(driver)



+ Inside(driver, toolbox)

VLM Output

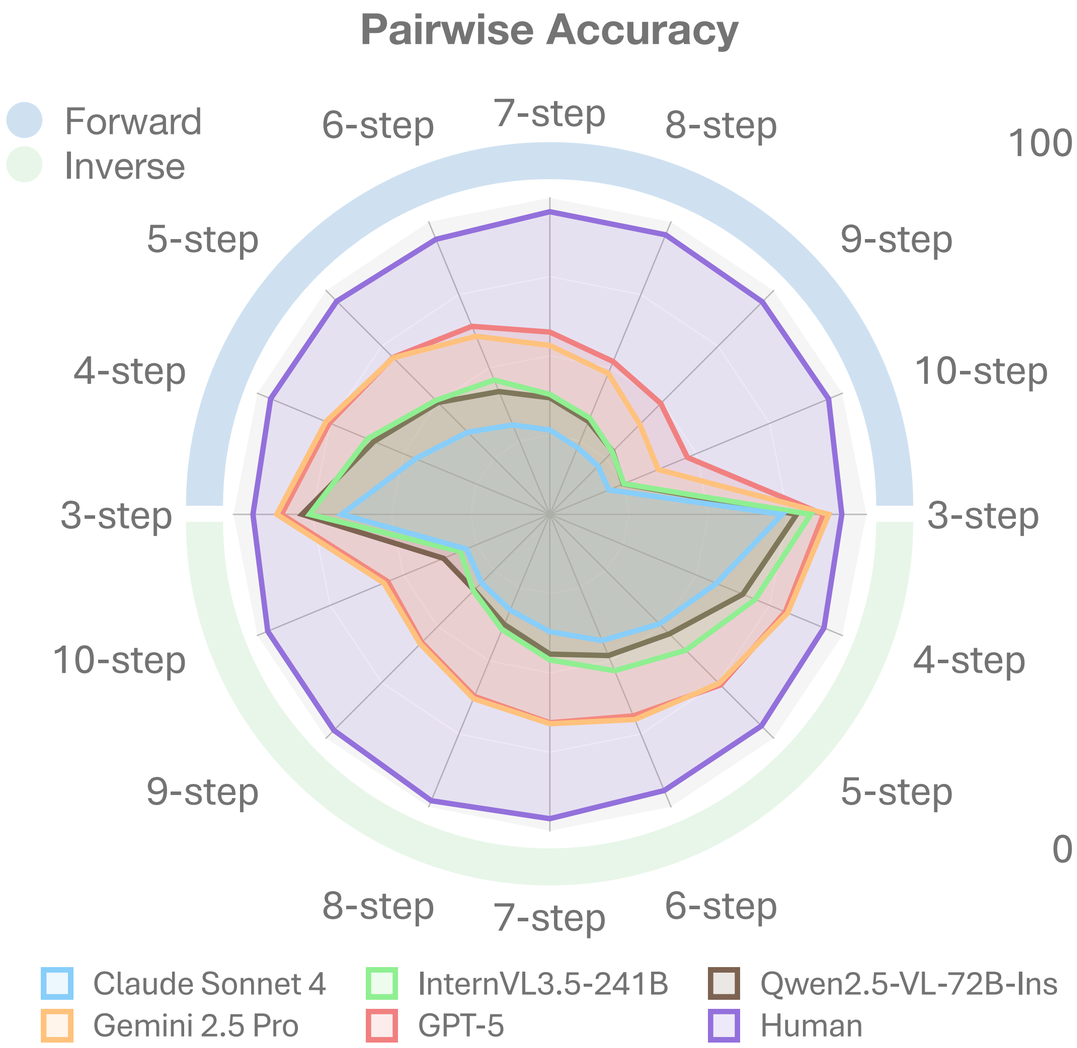
− OnTop(drill, table)

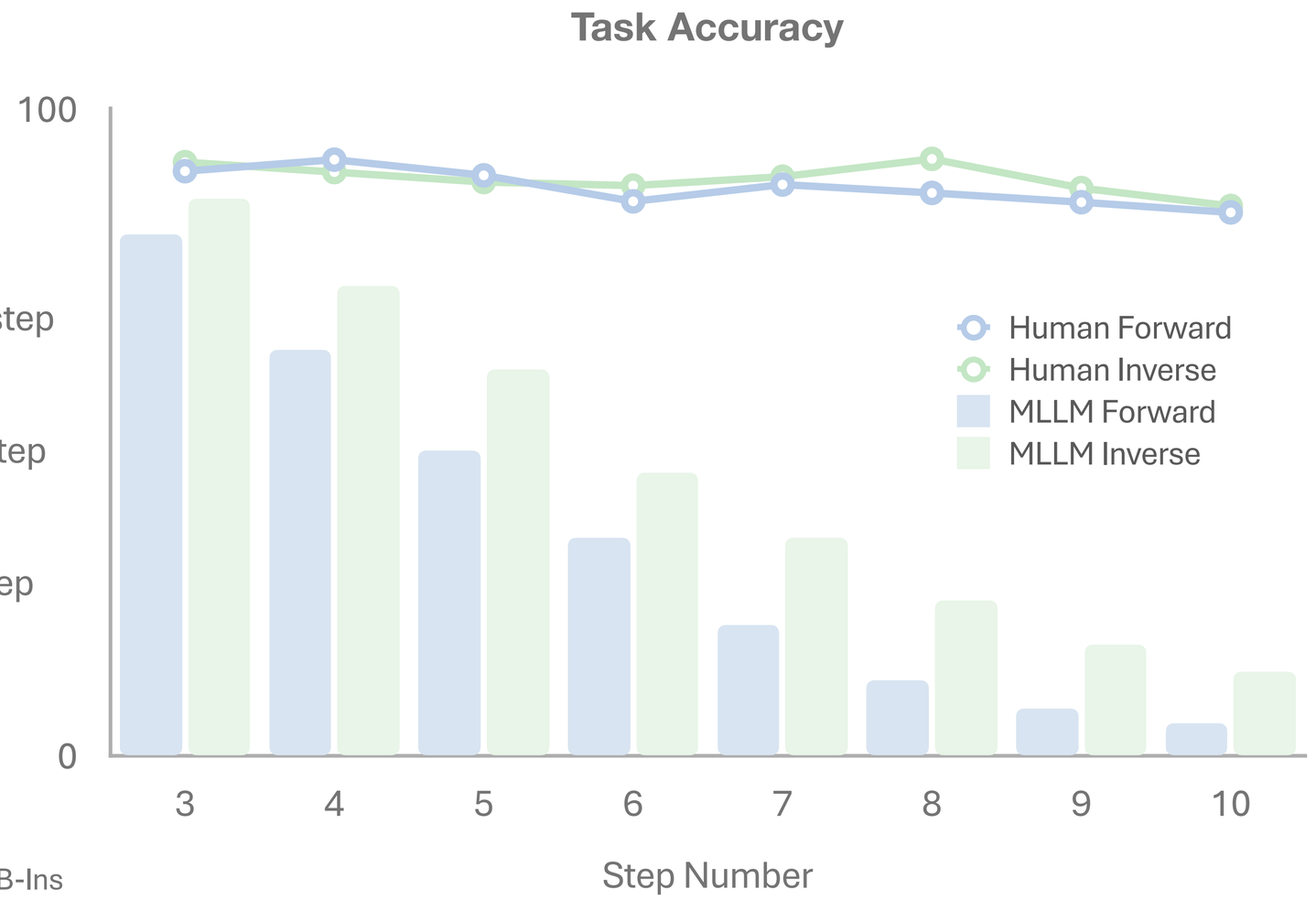
+ Grasping(driver)

− OnTop(driver, table)

+ Inside(driver, toolbox)

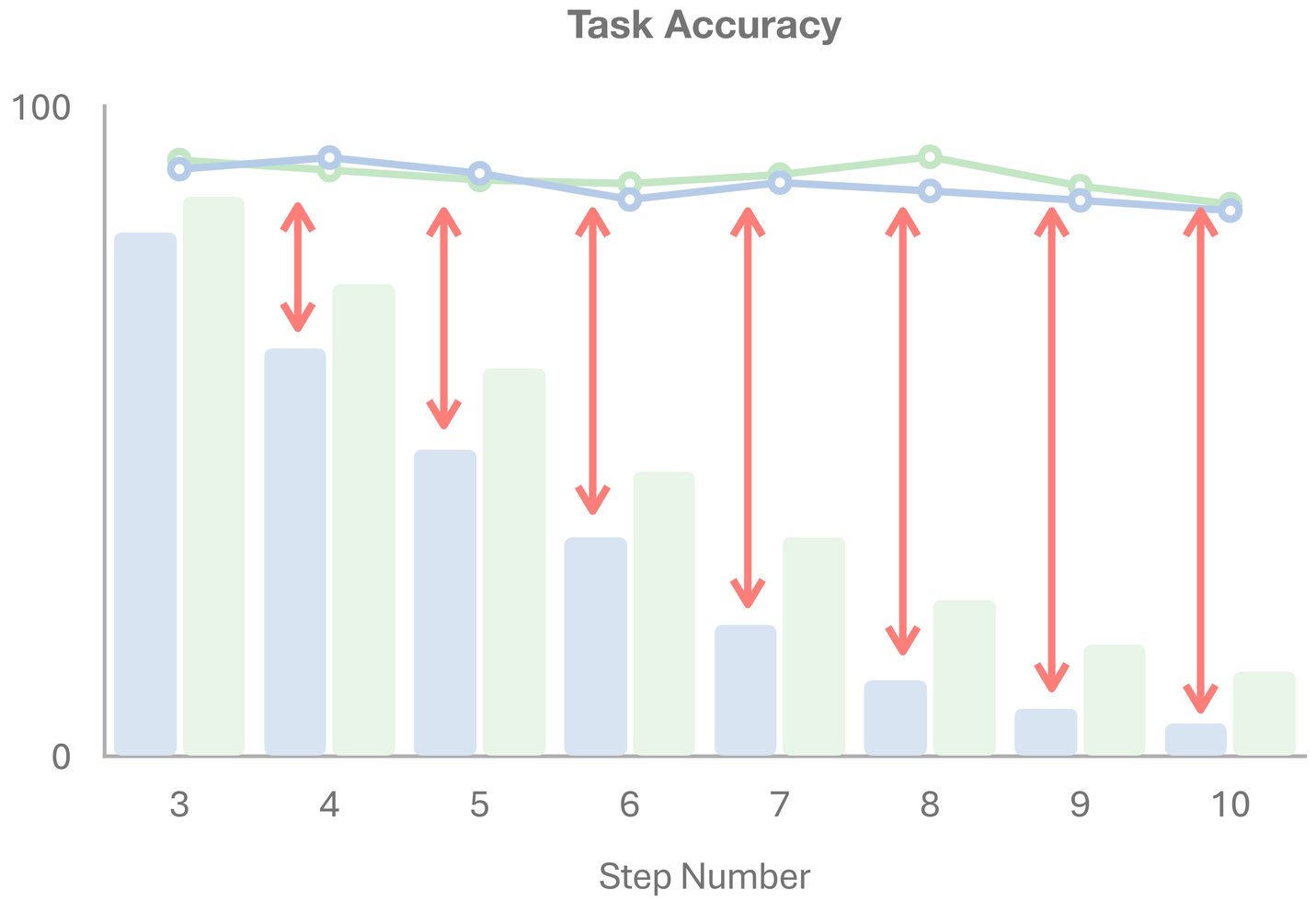
MLLMs performance on ENACT across varied step numbers



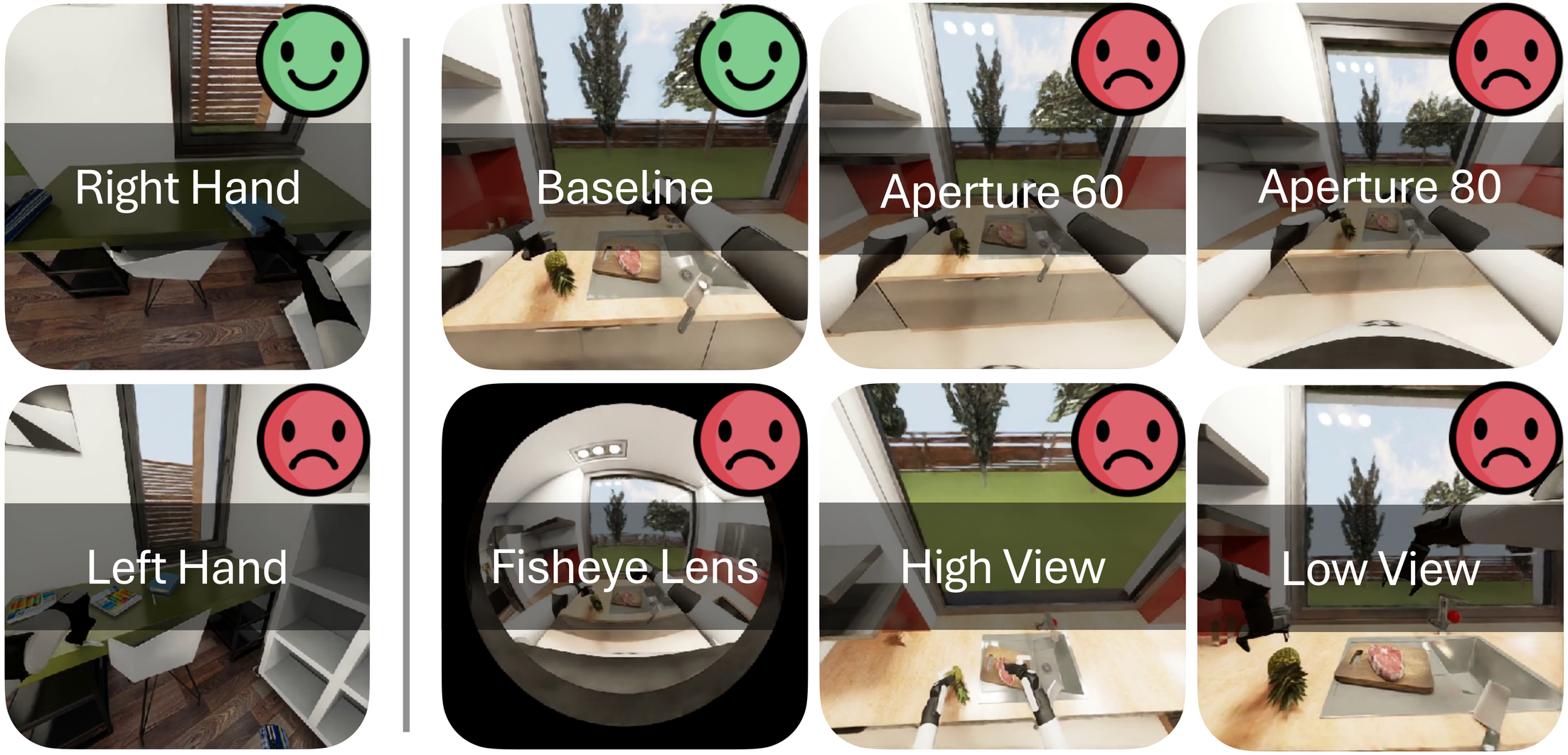


MLLMs performance on ENACT across varied step numbers

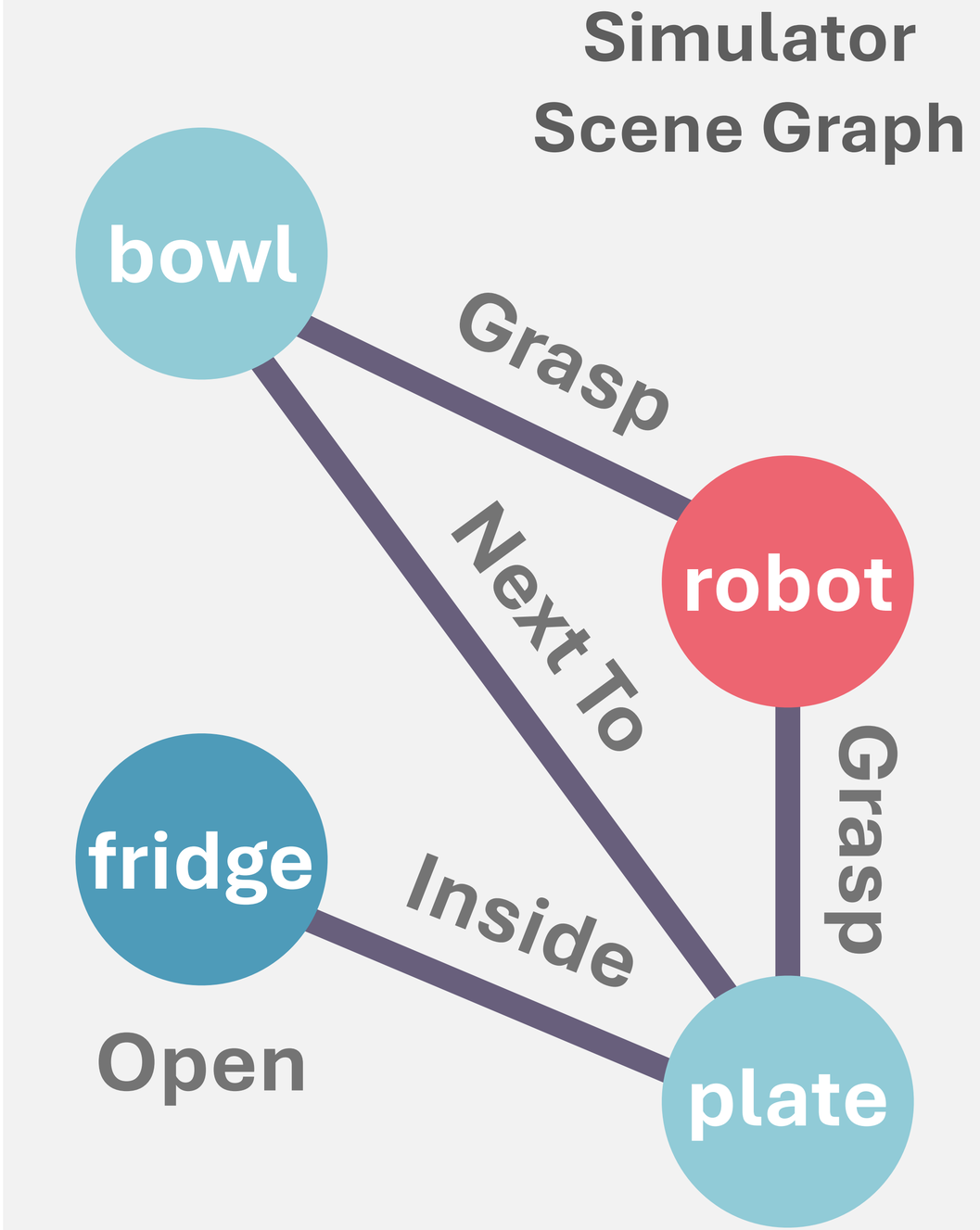
- MLLMs' performance decreases as step number increases
- MLLMs perform better on inverse setting than forward setting
- Humans achieve near-ceiling performance
- Remains a large gap between MLLMs and humans, especially at longer steps



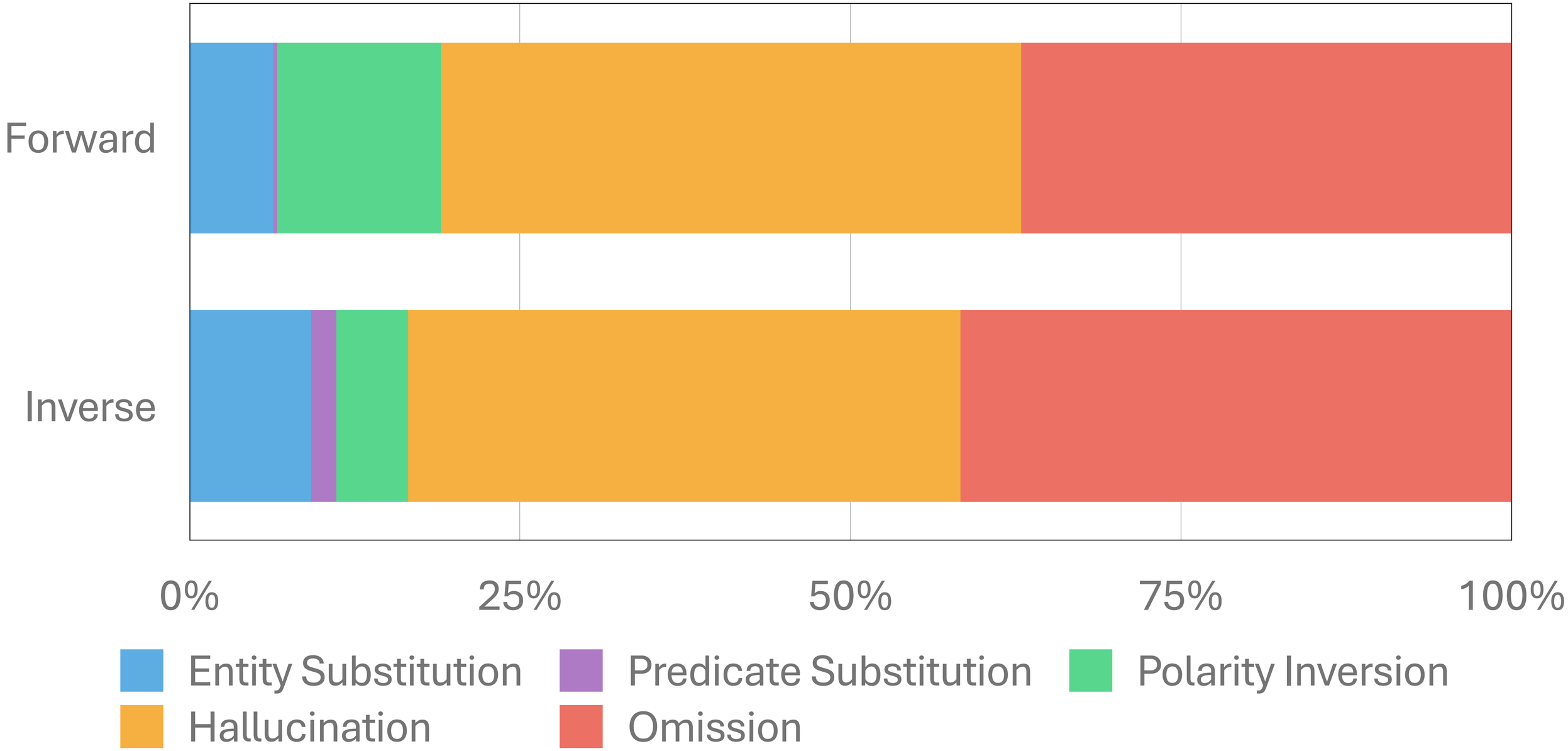
Performance depends on the viewpoint and acting hand



Ground-Truth	Inside(plate, fridge)
Entity Substitution	Inside(bowl, fridge)
Predicate Substitution	OnTop(plate, fridge)
Polarity Inversion	~Inside(plate, fridge)
Hallucination	Under(bowl, chair)
Omission	Not reported



MLLMs error distribution



ICLR



ENACT

https://enact-embodied-cognition.github.io/

https://github.com/mll-lab-nu/ENACT

MindTopo

Can Foundation Models Reason in Topological Space?

Northwestern University Microsoft Research Stanford University


Yunfei Ge*


Anbang Liu*


Qineng Wang*†


Johnalbert Garnica*


Jianwen Lyu


Zihan Wang


Reuben Tan


Jianfeng Gao


Ruohan Zhang


Yining Hong


Jiajun Wu


Manling Li

* Equal contribution. † Project lead.

Which metal ring can be removed?

Without cutting either ring.

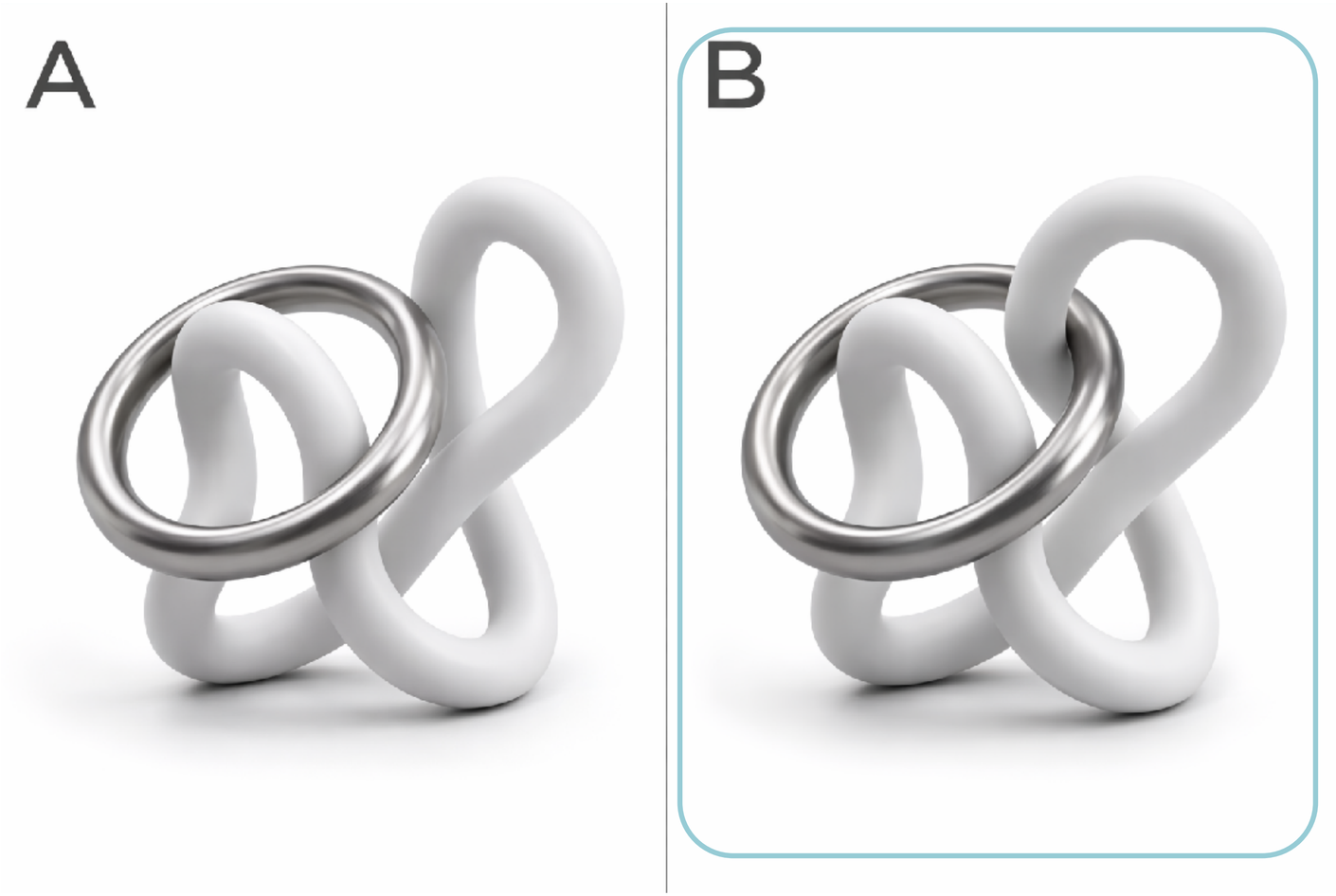
A

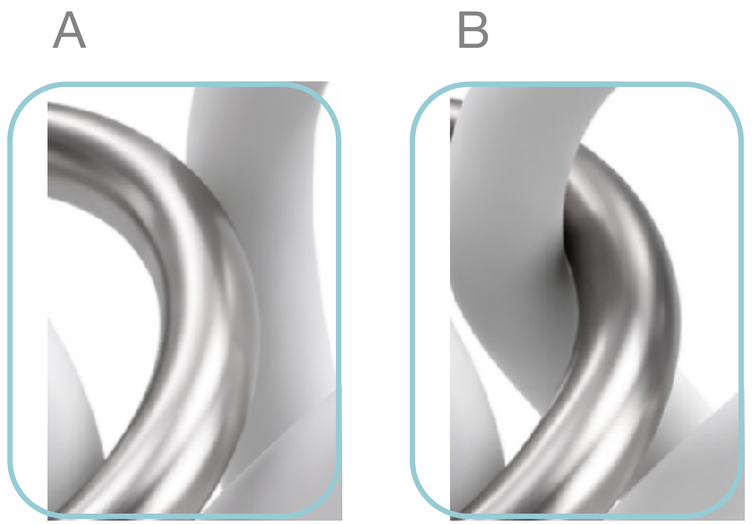


B



B can be removed.





Trace what passes in front.

B — removable

Similar shapes. Different structure.

A



B



A: interlocked

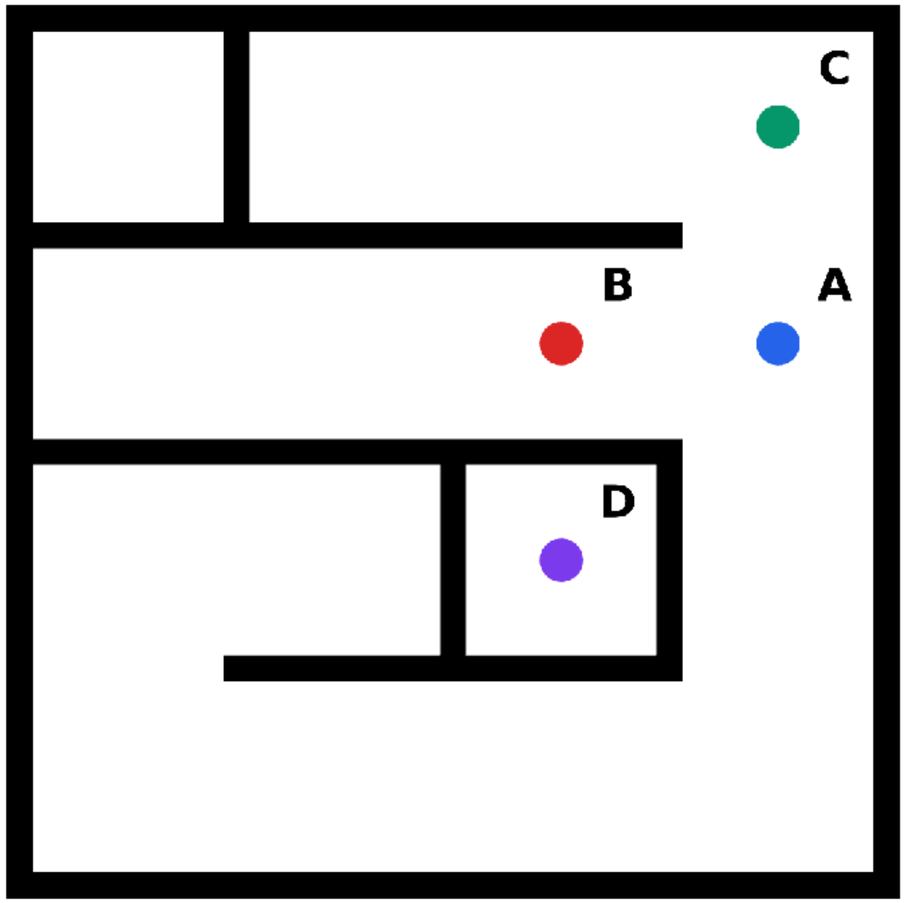
B: separable

The question is

which parts are linked.

Topological structure

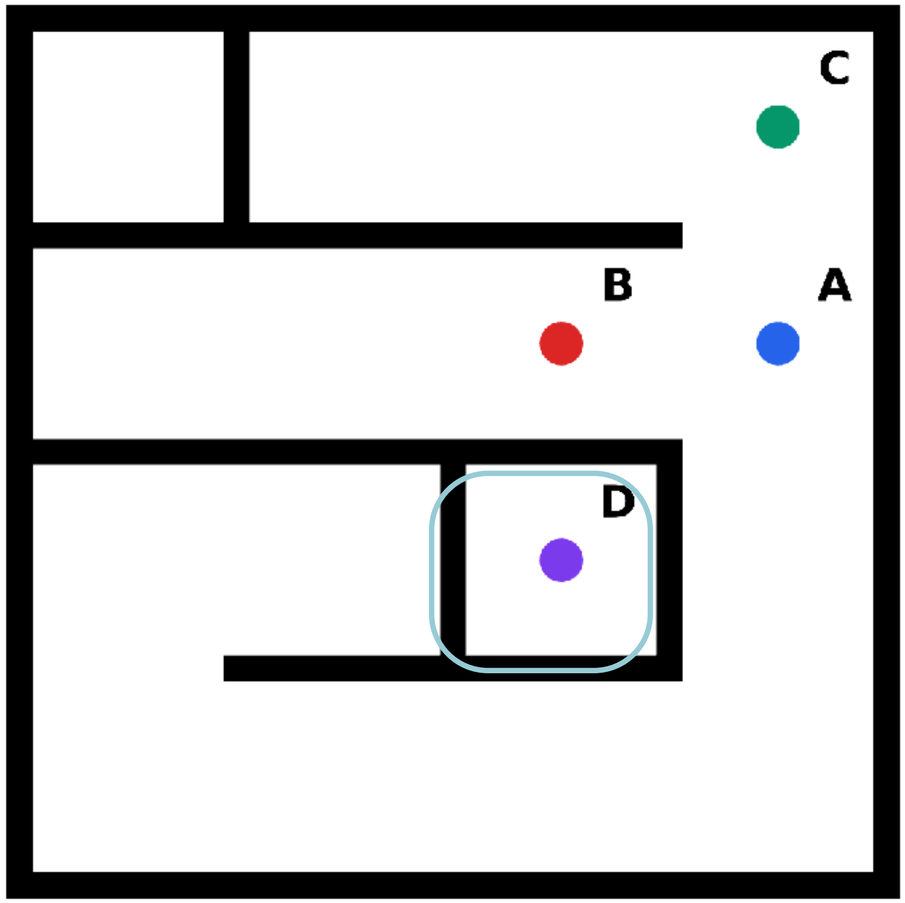
Try a simpler spatial relation.



Which points connect to A?

B C D

Connectivity is about reachable regions.



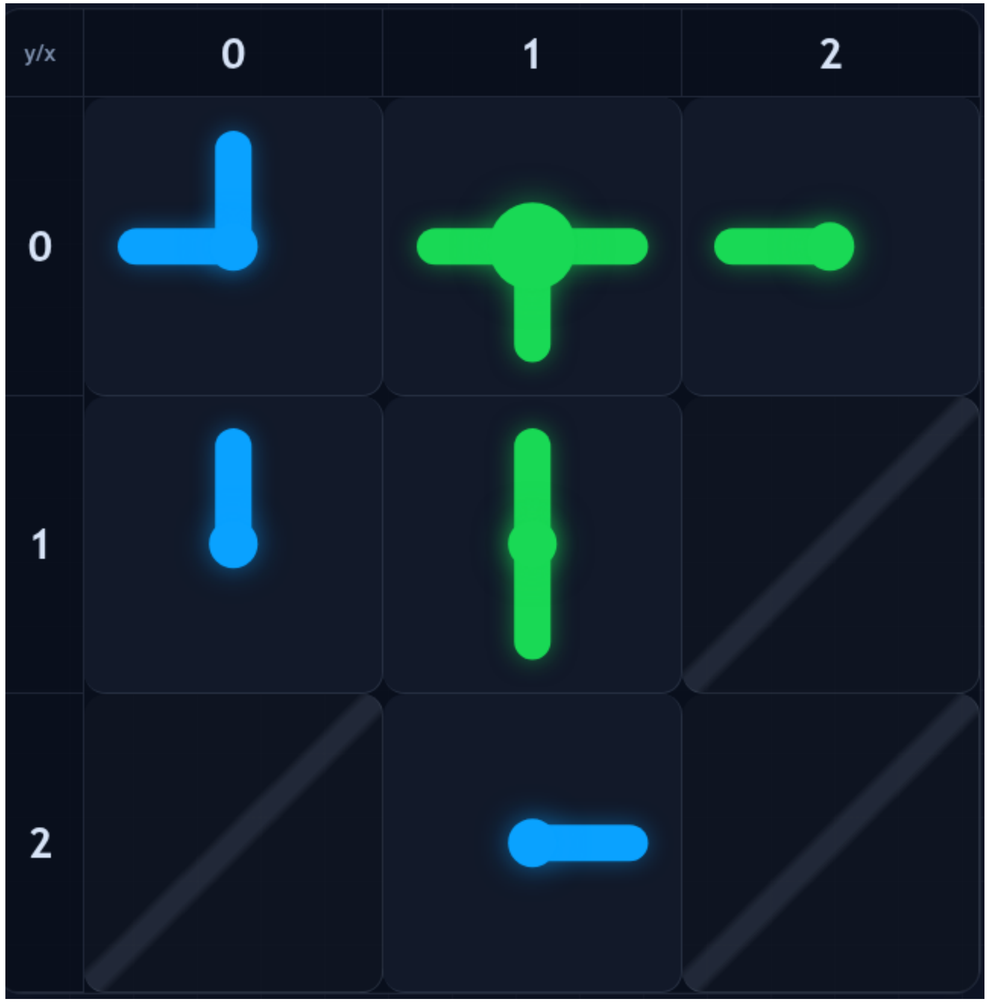
Which points connect to A?

B C D

B and C

D lies in a separate region.

Now create connectivity.



Connect every pipe

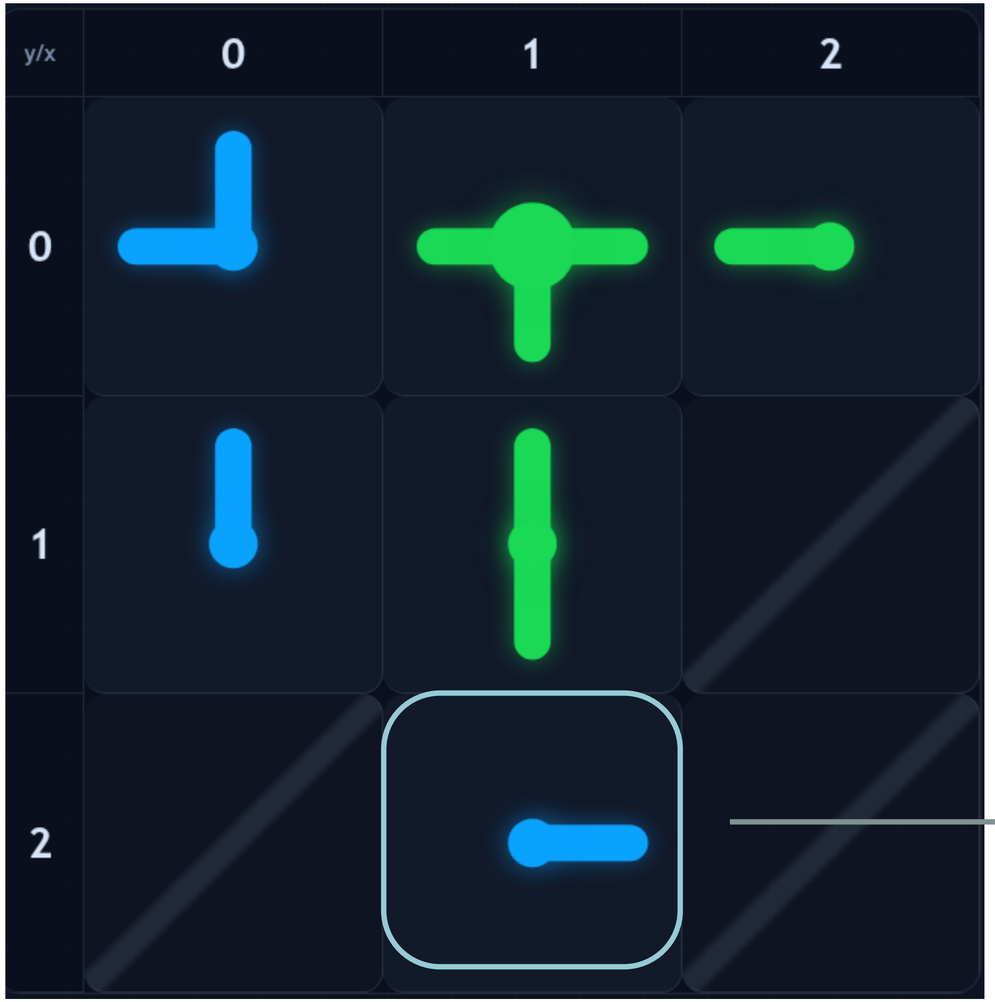
to the green source.

Allowed action:

rotate one pipe by 90°.

What would you rotate first?

Turn the bottom endpoint toward the source.



Connect every pipe

to the green source.

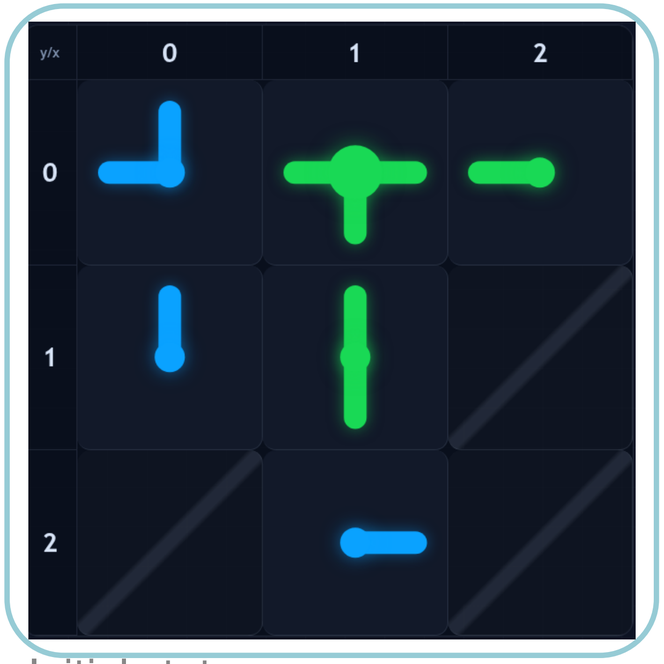
Allowed action:

rotate one pipe by 90°.

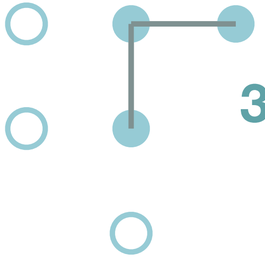
Bottom row, middle column

Rotate 90° counterclockwise.

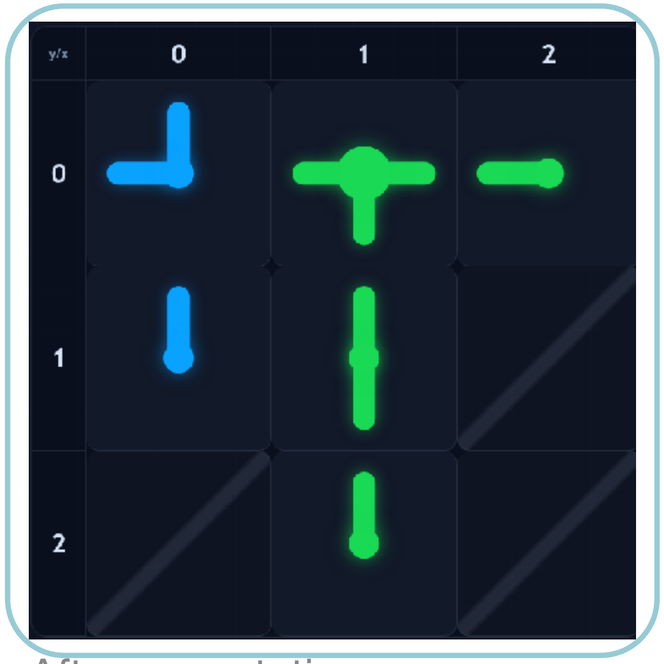
One legal action changes connectivity.



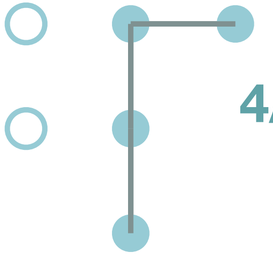
Initial state



3/6 connected



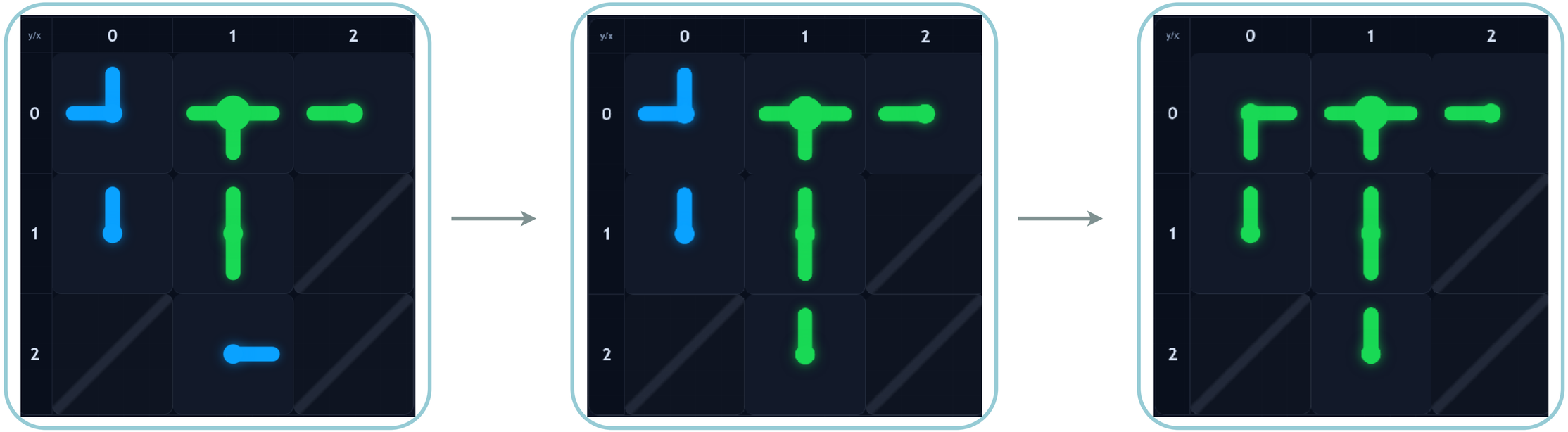
After one rotation



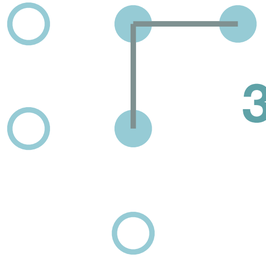
4/6 connected

Pipe · connectivity to the source · tile identities stay fixed

The goal requires a sequence of valid changes.

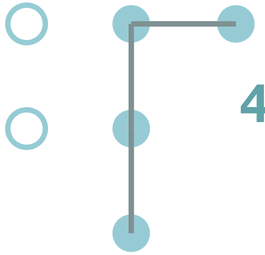


Initial state



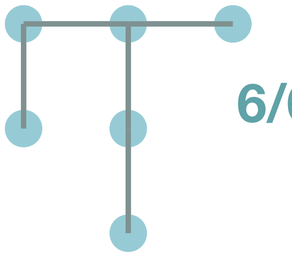
3/6 connected

After one rotation



4/6 connected

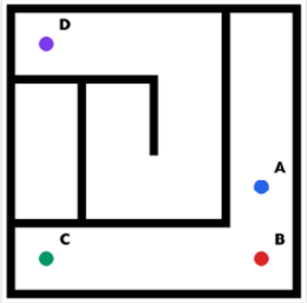
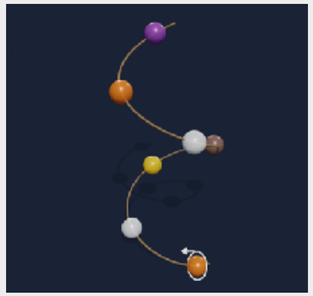
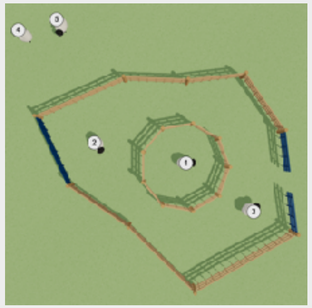
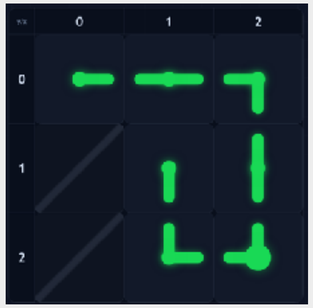
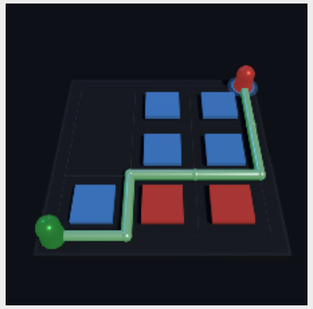
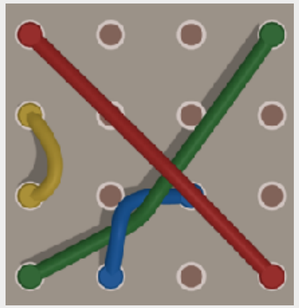
After further rotations



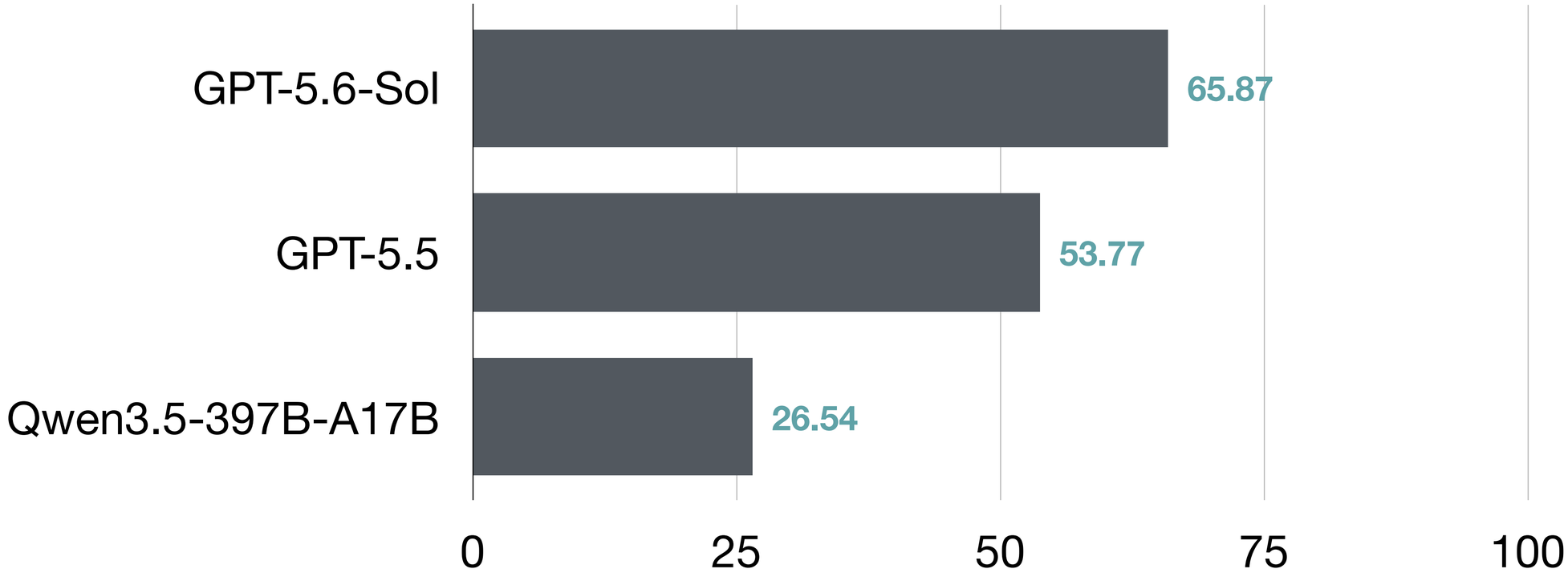
6/6 connected

Pipe · connectivity to the source · tile identities stay fixed

MindTopo: five properties, two demands.

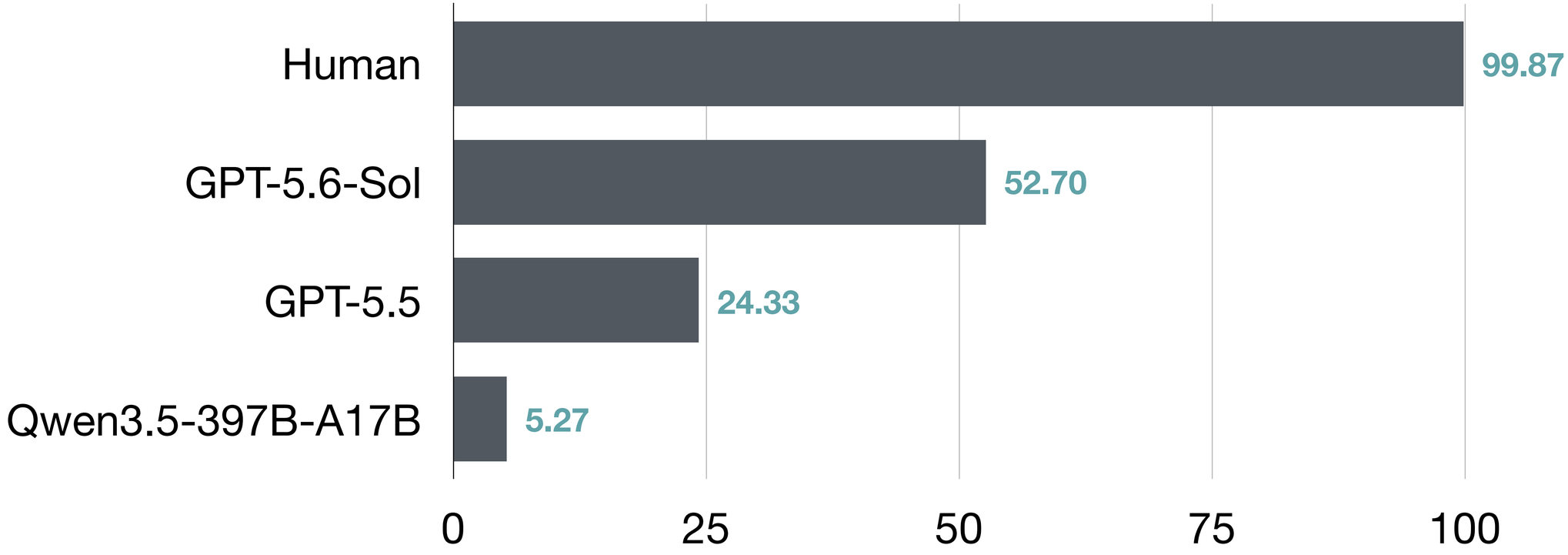
	Continuity	Separation	Order	Enclosure	Knots
Reasoning					
	Connected?	Separate parts?	In what order?	Inside or outside?	Linked or unlinked?
Planning					
	Connect the pipes.	Draw a partition.	Reorder the tiles.	Trap the cat.	Remove crossings.

Does answering imply acting?



MindTopo · task-macro reasoning accuracy (%) · 8 tasks

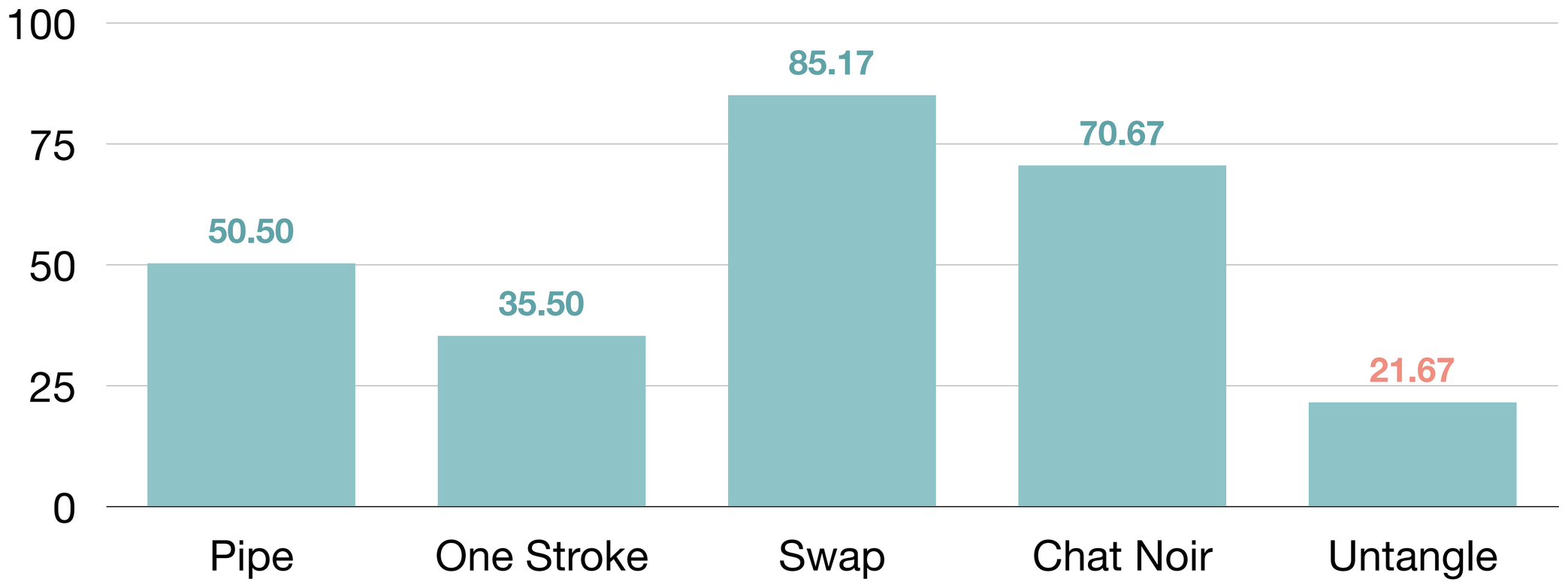
Human planning is near ceiling. Models are not.



47.17 percentage-point gap to the best model

MindTopo · task-macro planning success (%) · 5 environments

The average hides very different planning skills.

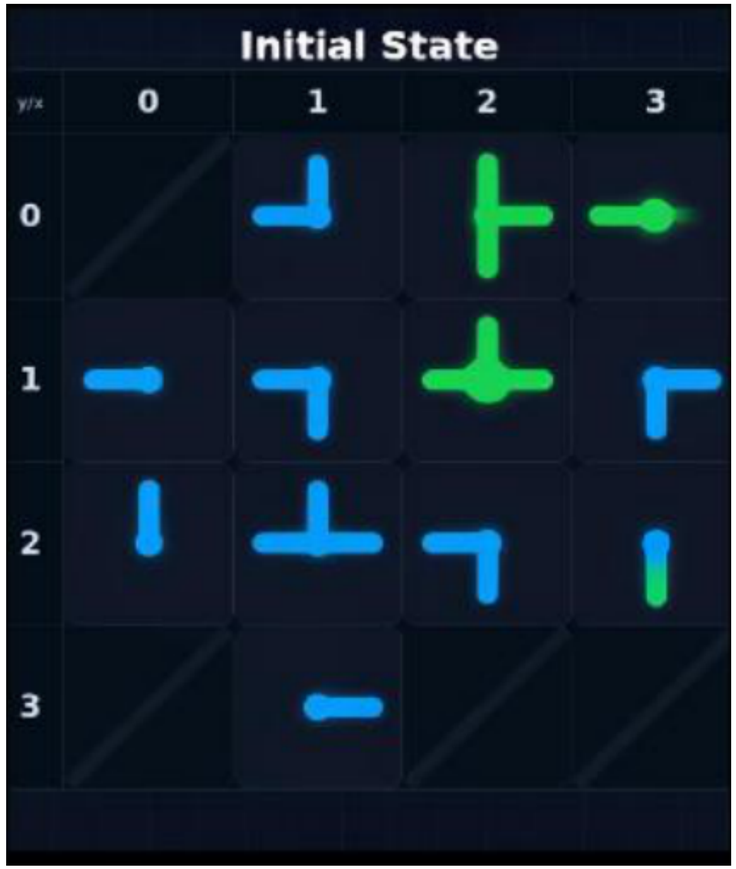


85.17% on Swap; 21.67% on Untangle.

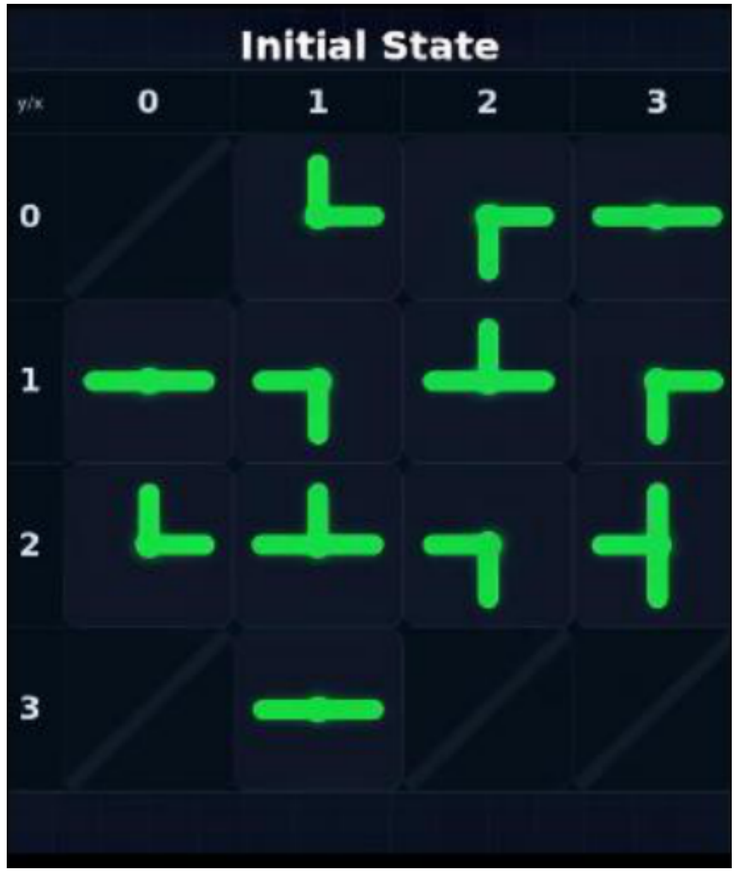
MindTopo · GPT-5.6-Sol · episode success (%)

Could imagined frames help the agent plan?

Generated rollout: first frame

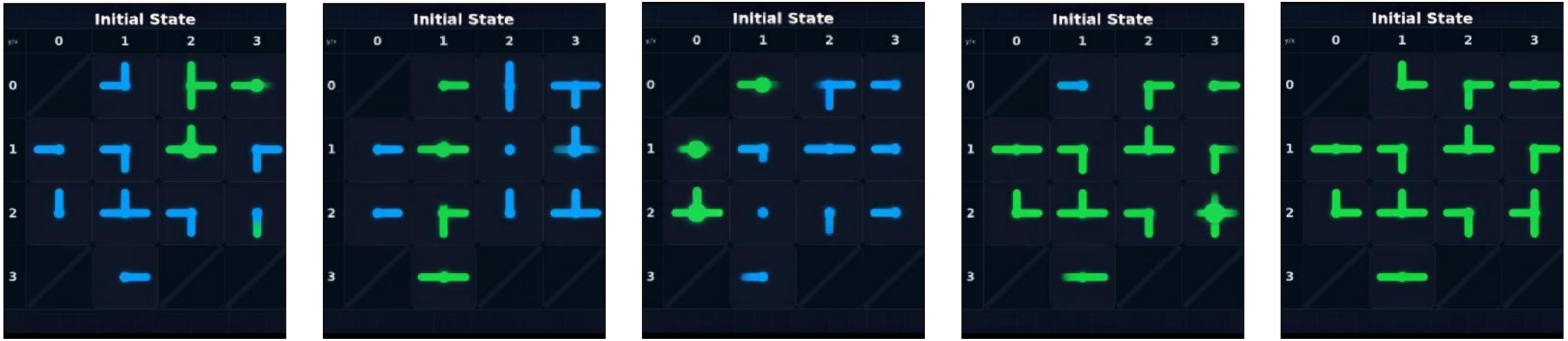


Generated rollout: last frame



Everything looks green. Is that enough?

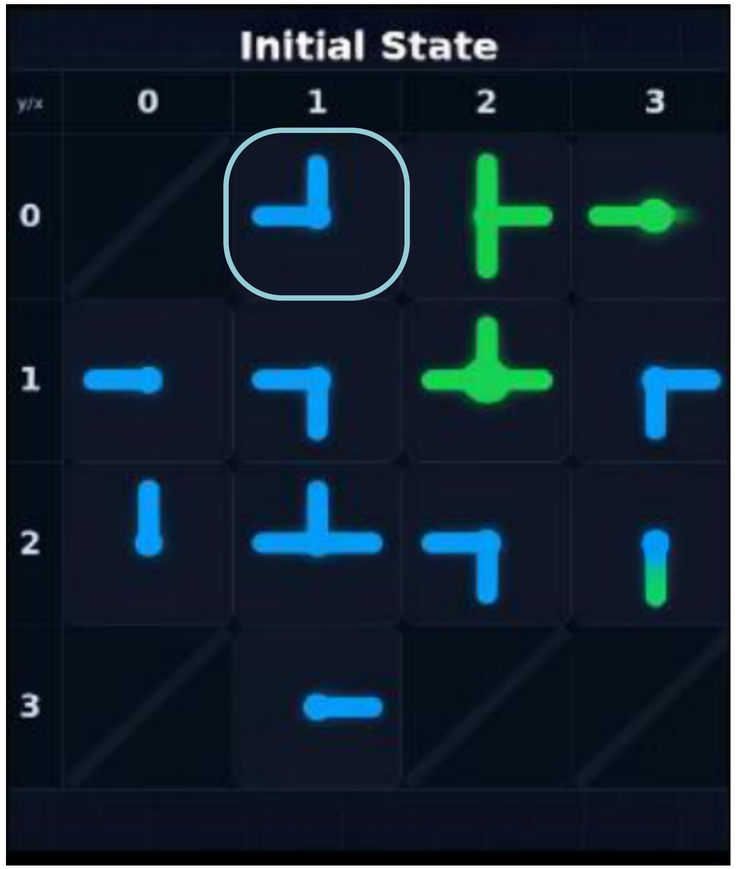
Look between the endpoints.



What changed besides orientation?

Generated Pipe rollout · five sampled frames

Rotation cannot turn an elbow into an endpoint.



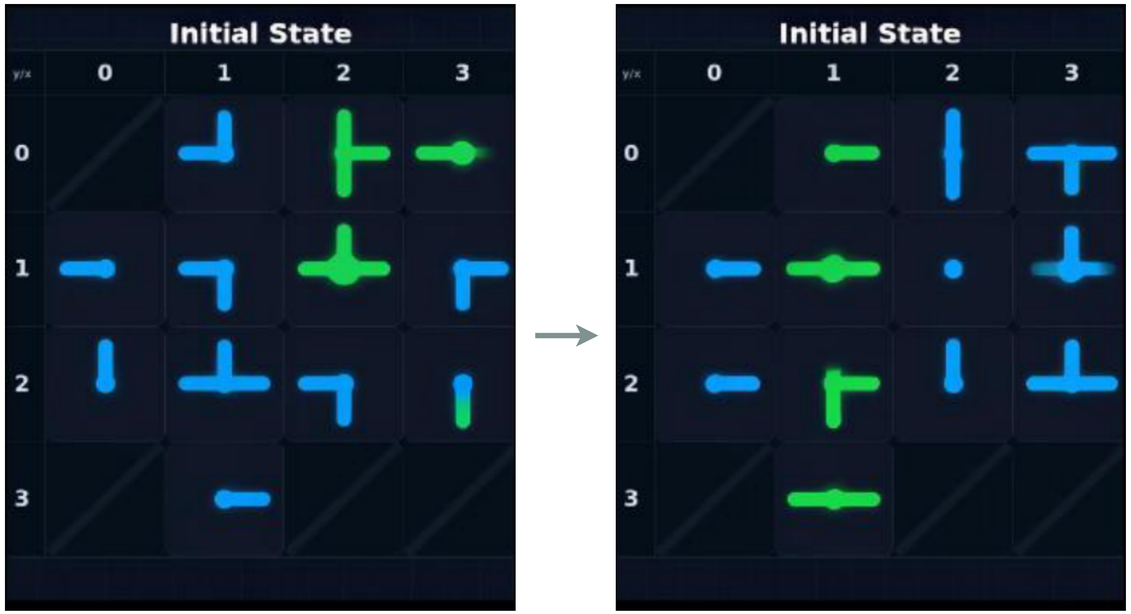




Elbow: two openings

Endpoint: one opening

A plausible image is not a valid next state.



What must stay consistent?

Same occupied cells

Same pipe types

Legal rotations

Predict change without inventing the state.

The video failure is systematic.

116 / 119

audited videos violated environment rules

Failure type	Videos
Dynamics violation	116 / 119
Invariance error	93 / 119
Consistency error	90 / 119

A world model must preserve the rules through time.

MindTopo · Wan2.2 · One Stroke and Untangle · error categories overlap

Final-frame success can hide an invalid process.

Generator / task	Final-frame pass	Full-rollout pass
LTX-2.3 / Untangle	32.4%	0%
LTX-2.5 / Pipe	35.2%	0%

Check the whole transition sequence.

Could a coding agent check this transition?



same_cells(before, after)

same_pipe_types(before, after)

legal_rotation(before, after)

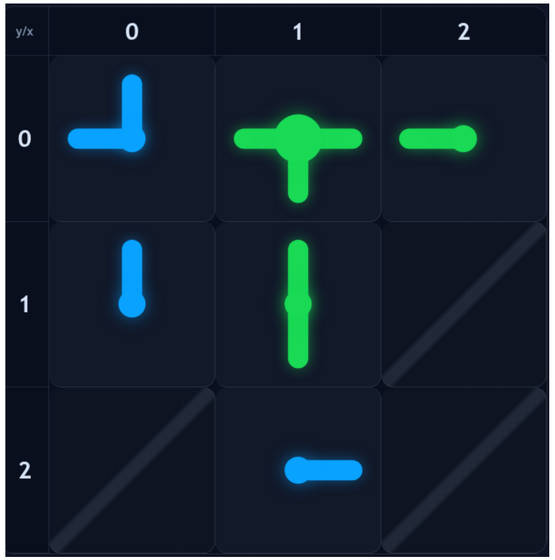
connected_to_source(after)

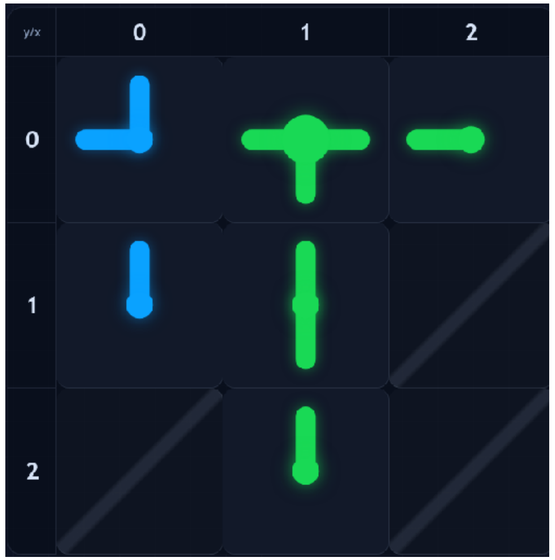
Parsed state

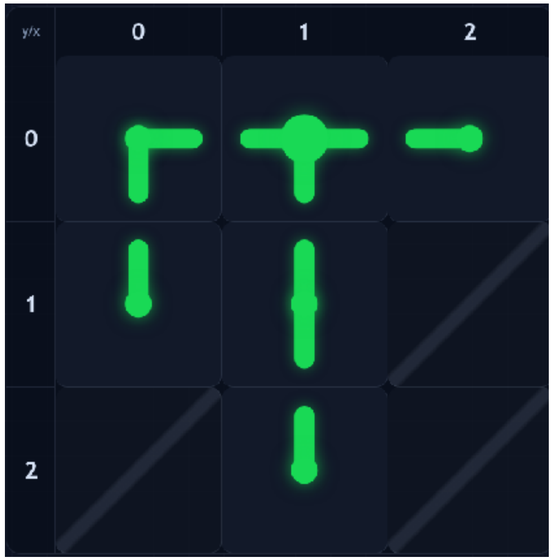
The check is only as good as the grounded state.

Proposed verification step for a coding agent

What did the pipe example require?







Represent the pieces.

Keep the tile types and openings.

Apply legal rotations.

Change orientation, preserve type.

Verify connectivity.

Does the path reach the goal?

What the benchmarks teach us

Representation

Keep objects and relations.

Action consequences

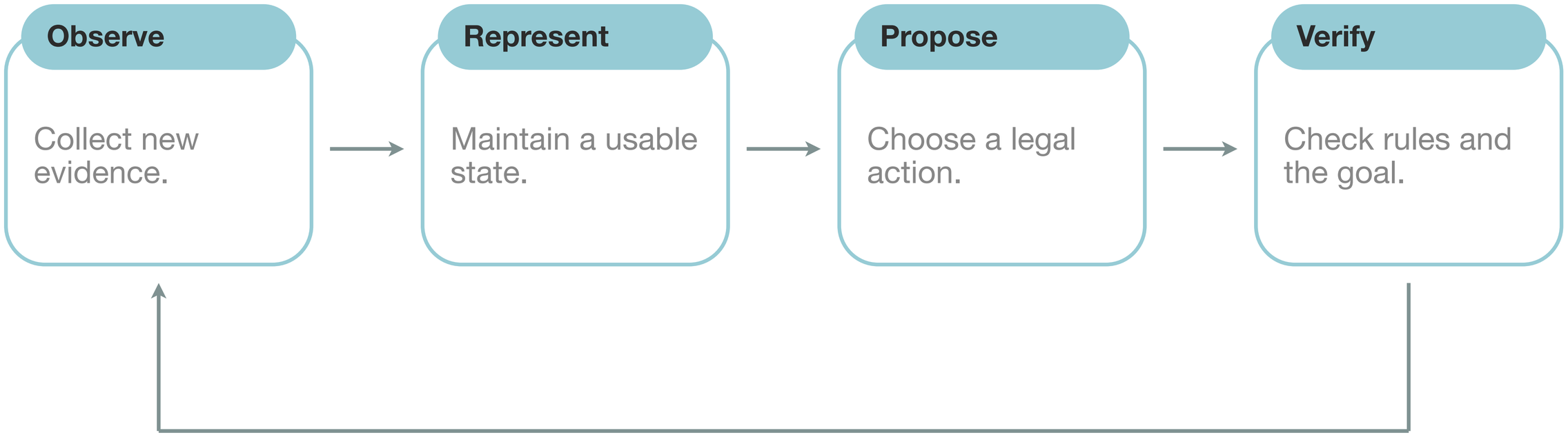
Track changes across time.

Verification

Check rules and success.

Evaluate the entire agent–environment loop.

A coding agent can make the loop explicit.



Execute, inspect feedback, revise the state or plan.

From demonstrations to dependable agents

Can a coding agent build a spatial model,

predict the effects of its actions,

and revise its model from feedback?



Spatial intelligence

World modeling

Embodied intelligence