

Part II: Foundation Models meet Virtual Agents

Manling Li, Assistant Professor at Northwestern University



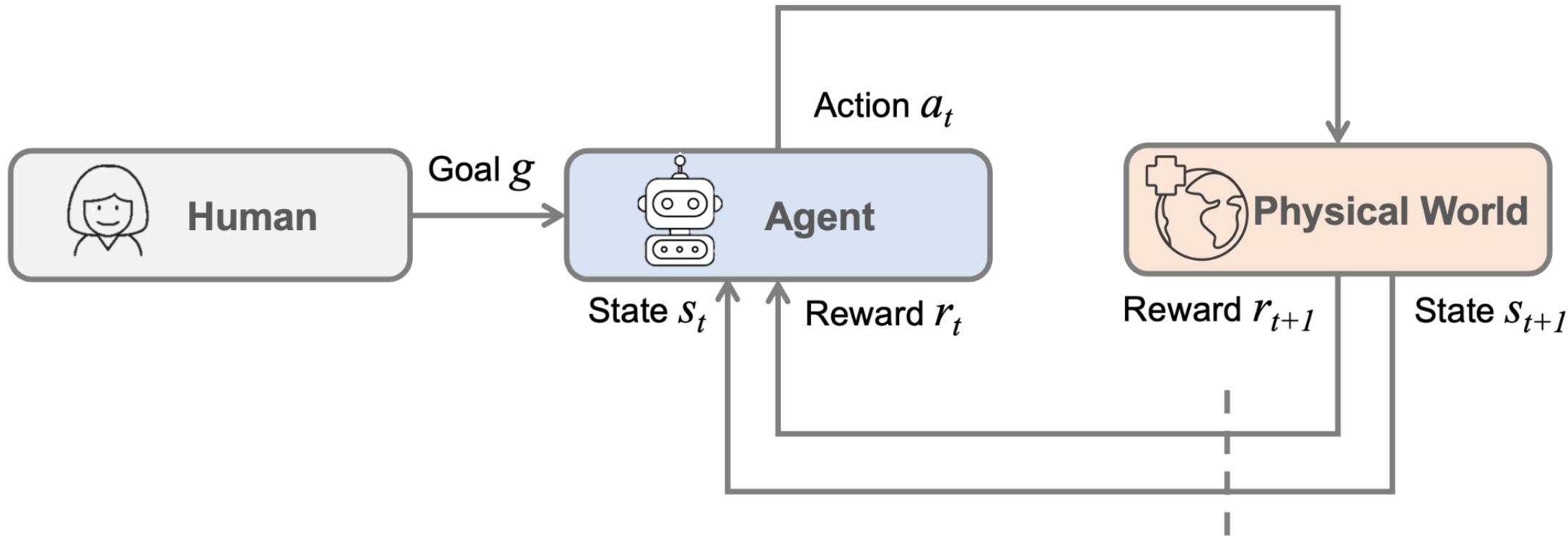
Northwestern
University

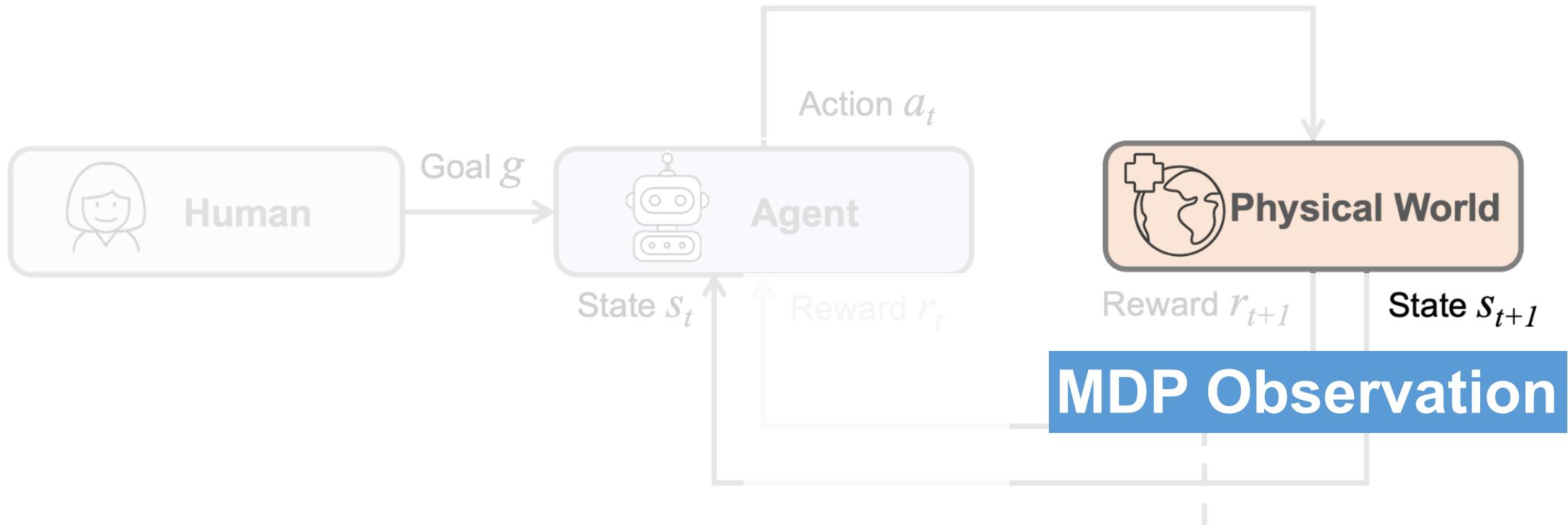


COLUMBIA



Stanford
University







An Envisioned Agent Society



MDP Environment: Embodied Simulator



General Simulators

Simulator	Year	Physics Engine	Applications
Isaac Sim	2023	PhysX	Navigation, Autonomous Driving
Isaac Gym	2019	PhysX	Reinforcement Learning, Large-Scale Parallel Simulation
Unity ML-Agents	2017	Custom	Reinforcement Learning, Robotics Simulation
AirSim	2017	Custom	Drone Simulation, Autonomous Driving, Reinforcement Learning
PyBullet	2017	Bullet	Reinforcement Learning, Robotics Simulation
MORSE	2015	Bullet	Navigation, Multi-Robot
V-REP (CoppeliaSim)	2013	Bullet/ODE/Vortex/Newton	Multi-Robot, Robotics Simulation
MuJoCo	2012	Custom	Reinforcement Learning, Robotics Simulation
Gazebo	2004	ODE/Bullet/Simbody/DART	Navigation, Multi-Robot
Webots	1996	ODE	Robotics Simulation

Real-Scene Based Simulators

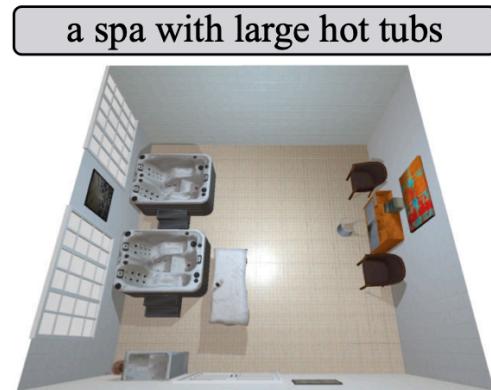
Simulator	Year	Scenes	Modalities
iGibson	2021	15	RGB-D, LiDAR, Learning
SAPIEN	2020	46	RGB-D, Joint Object Interaction
Habitat	2019	1000	RGB-D, Supports Multi-Agent
Matterport 3D	2018	90	RGB-D, Navigation Benchmark
Virtual Home	2018	50	RGB-D, Environment Graph
AI2-THOR ALFRED	2017	120	RGB-D, Supports Multi-Agent

Automated Scene Construction



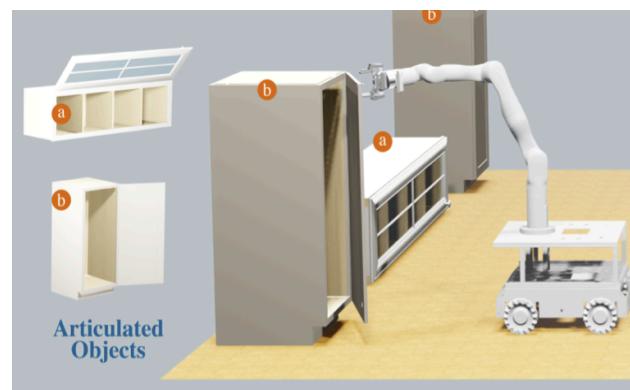
RoboGen

(Luo et al. 2023)



HOLODECK

(Kapelyukh et al. 2018)



PhyScene

(Yang et al. 2024)



ProcTHOR

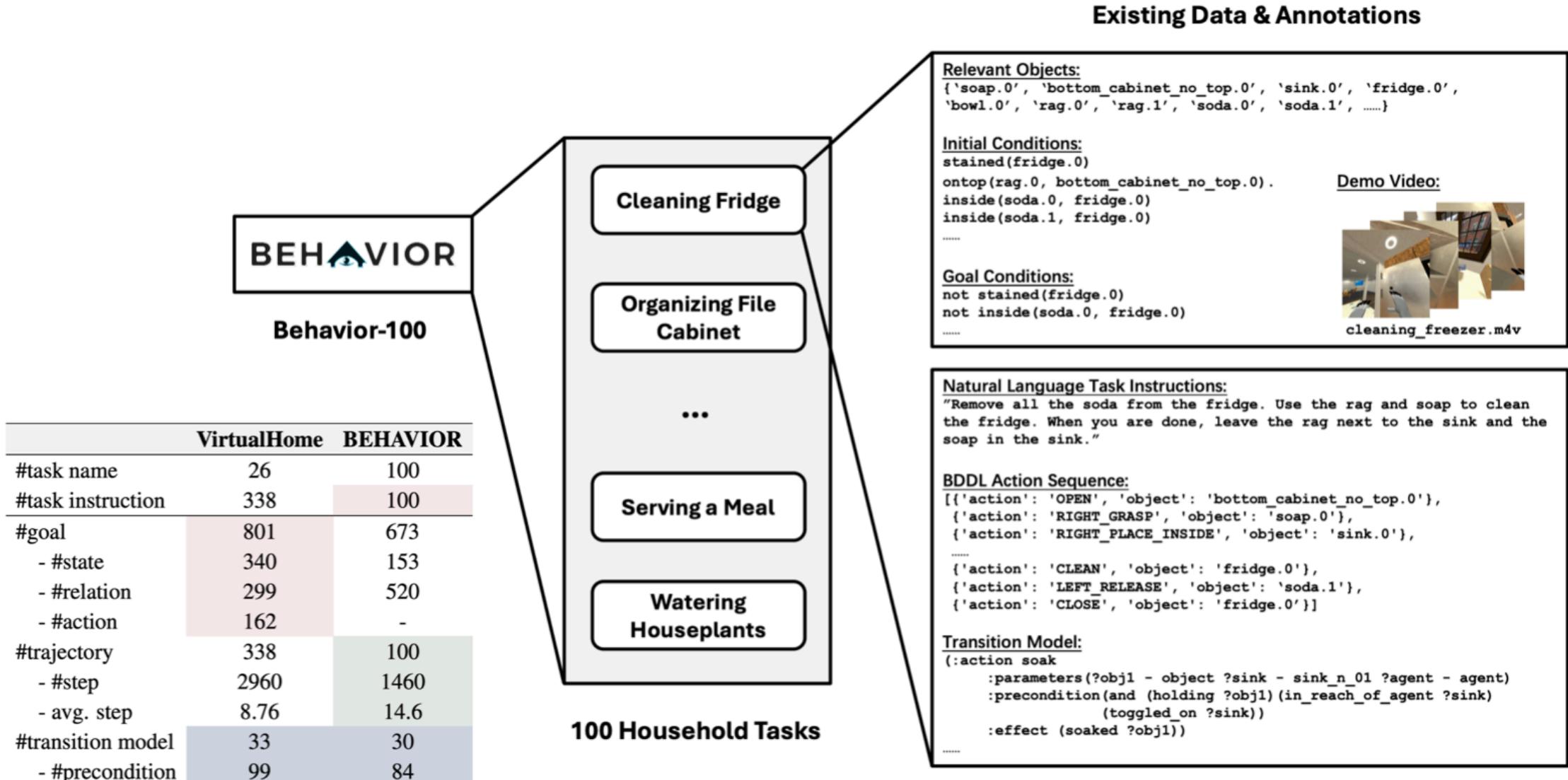
(Yang et al. 2022)

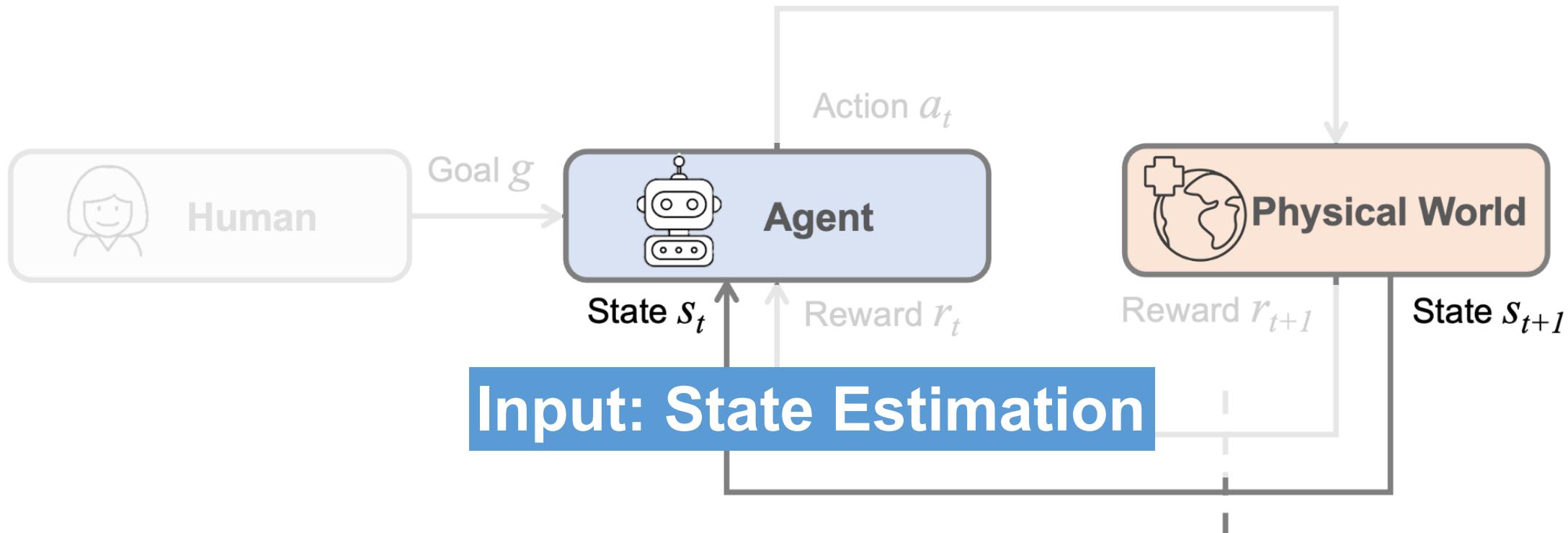
Further enhance these env by generating high-quality 3D scenes, facilitating diverse training scenarios

Real-Scene Based Simulators

Simulator	Year	Scenes	Modalities
iGibson	2021	15	RGB-D, LiDAR, Learning
SAPIEN	2020	46	RGB-D, Joint Object Interaction
Habitat	2019	1000	RGB-D, Supports Multi-Agent
Matterport 3D	2018	90	RGB-D, Navigation Benchmark
Virtual Home	2018	50	RGB-D, Environment Graph
AI2-THOR	2017	120	RGB-D, Supports Multi-Agent

What is a typical dataset?





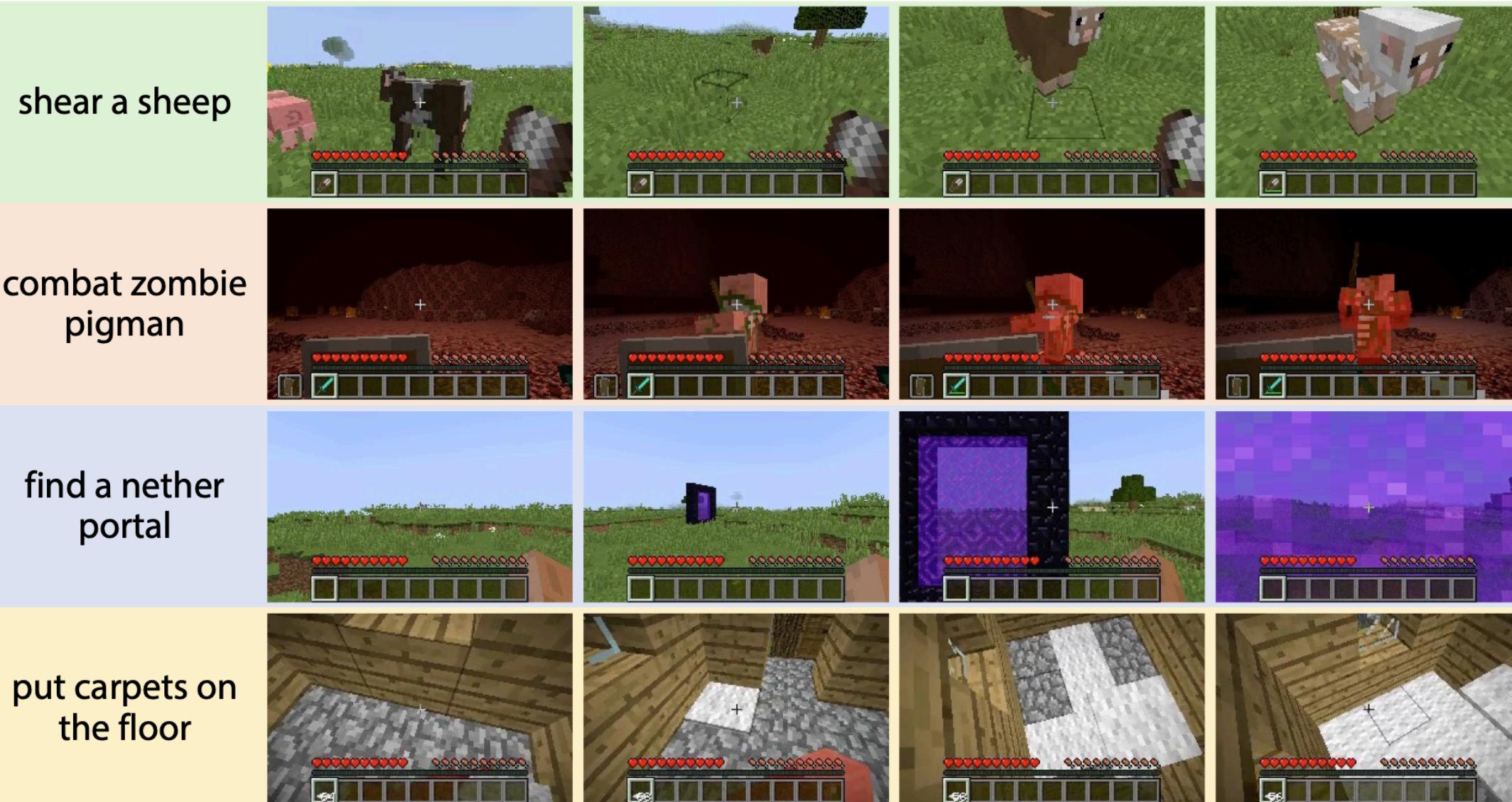
Perception / State Estimation

$$o \rightarrow s$$



Environment : Observation (Rendered 2D Images)

MDP Observation: Rendered 2D Images





Environment : Observation → State (3D Assets & States)

BEHAVIOR Vision Suite

8000+ 3D Objects



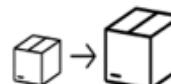
Customizable Data Generator
w/ Controlled generation



Camera pose



Lighting



Object properties

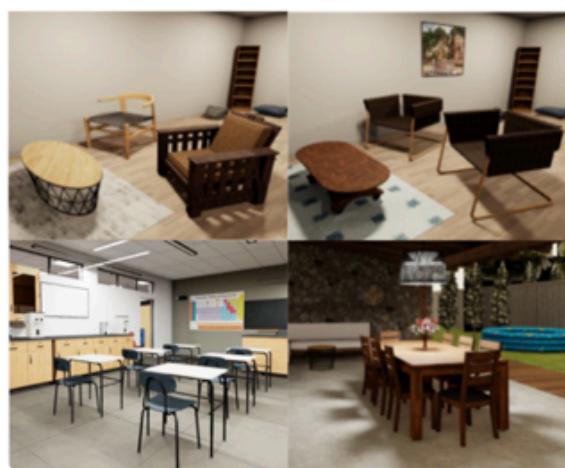


Object states



Spatial relationships

1000 Scene Instances



- ❑ A large-scale RGB-D dataset containing 10,800 panoramic views from 194,400 RGB-D images of 90 building-scale scenes



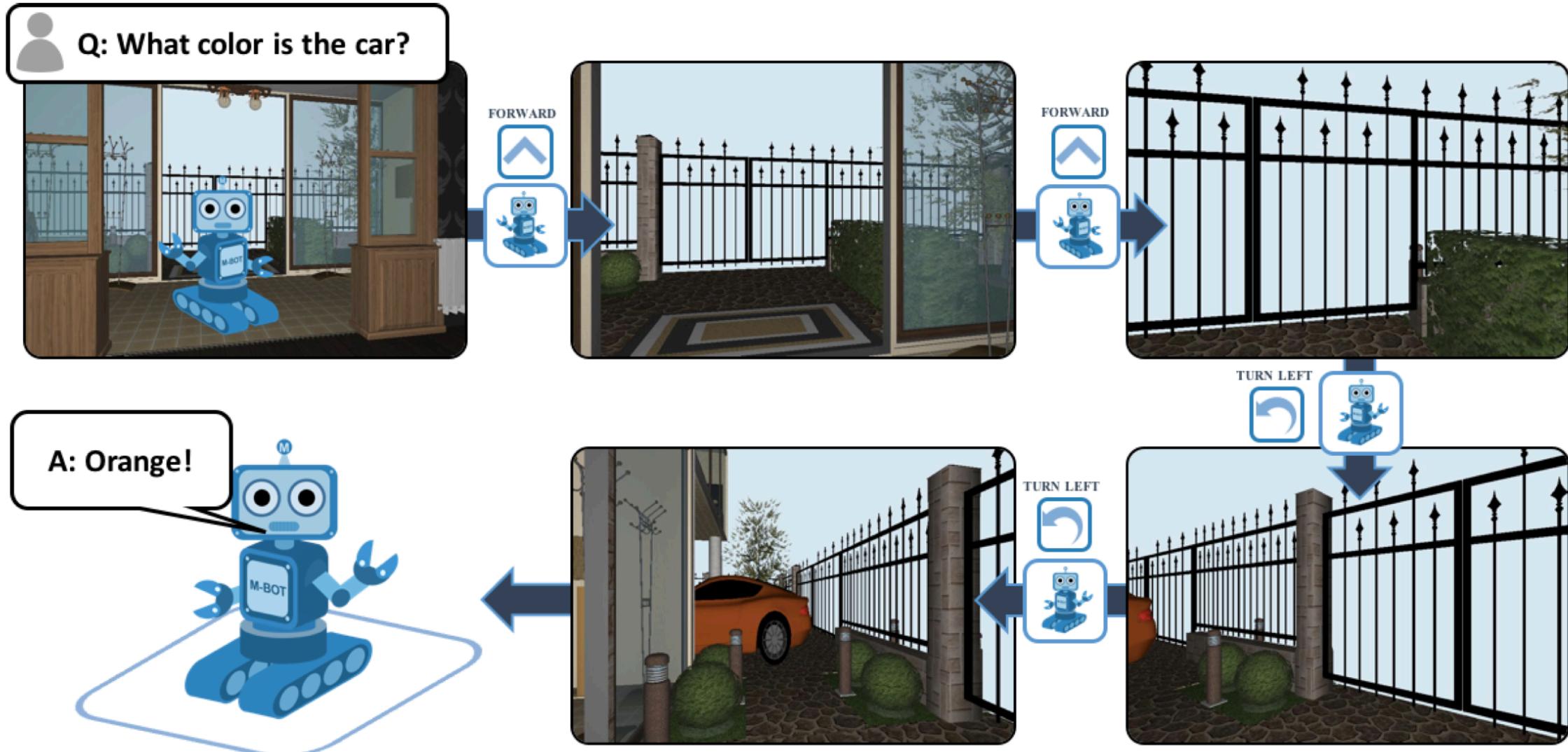
Textured 3D Mesh



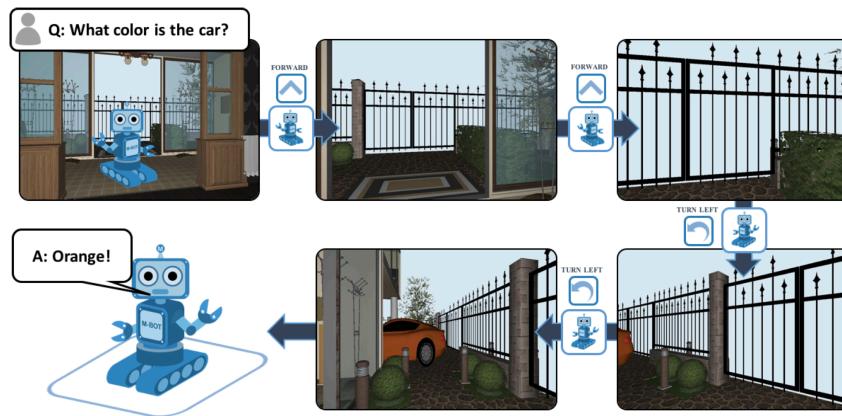
Panoramas



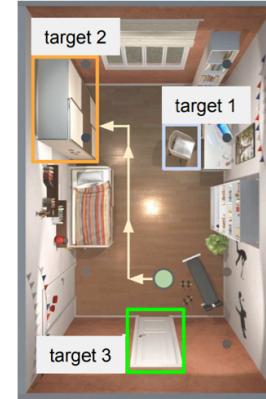
Object Instances



EQA: Embodied Question Answering



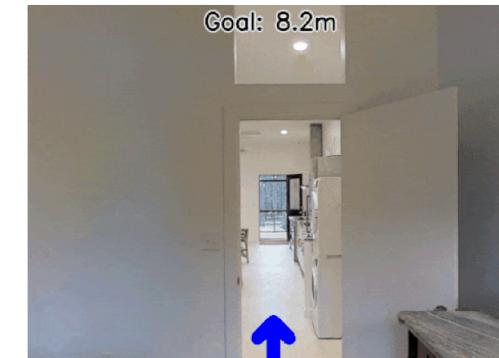
EmbodiedQA
(Das et al., 2018)



Visual Navigation
(Zhu et al., 2017, Gupta et al., 2017)

Question and answer	Initial Image	Scene View
Q: Is there bread in the room? A: No		
Q: How many mugs are in the room? A: 3		
Q: Is there a tomato in the fridge? A: Yes		

Interactive QA
(Gordon et al., 2018)

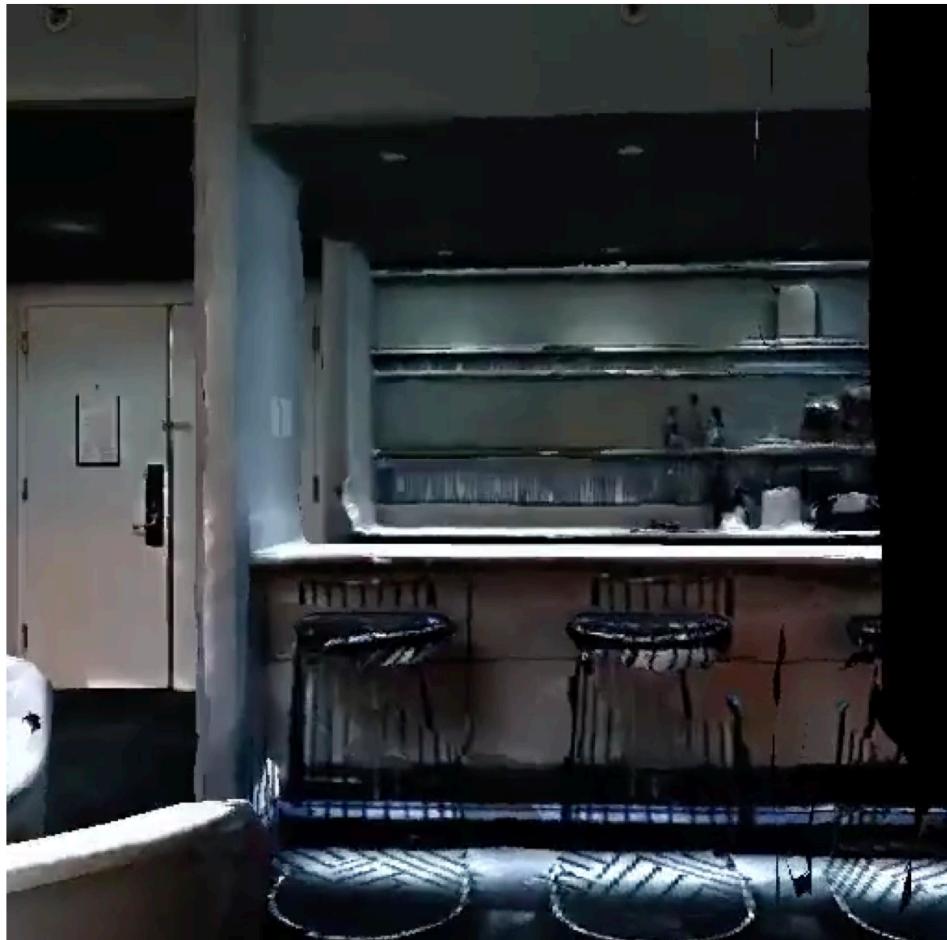


Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

Vision-Language Navigation
(Anderson et al., 2018)

- Two types of perception abilities:

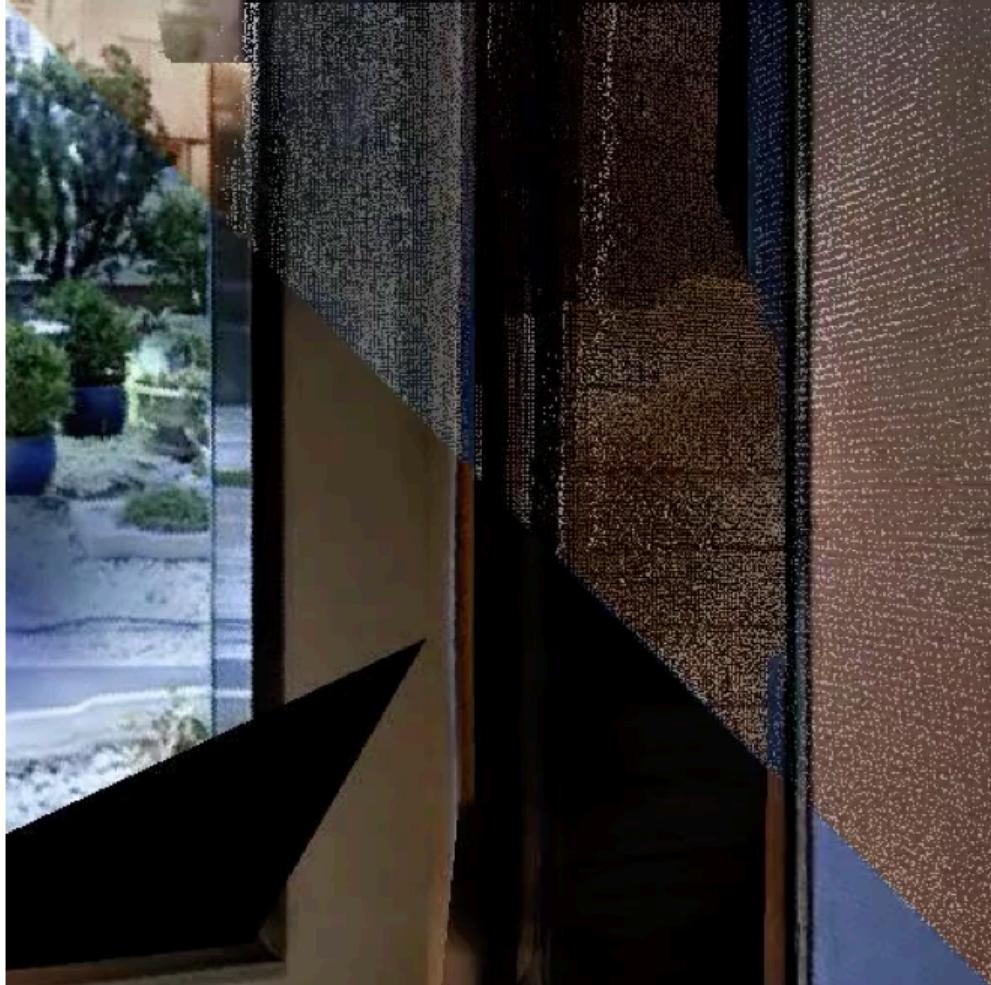
RGB Perception



Point Cloud Perception

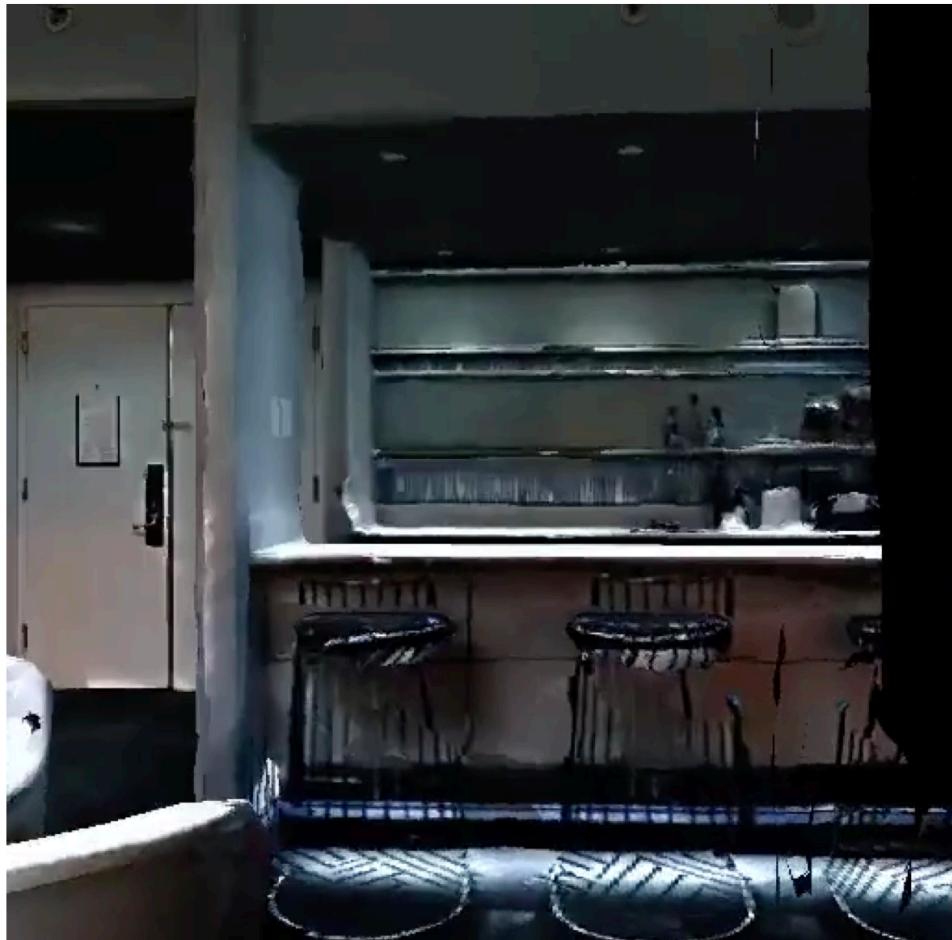


Q: What color is the counter in the hallway?



- Two types of perception abilities:

RGB Perception

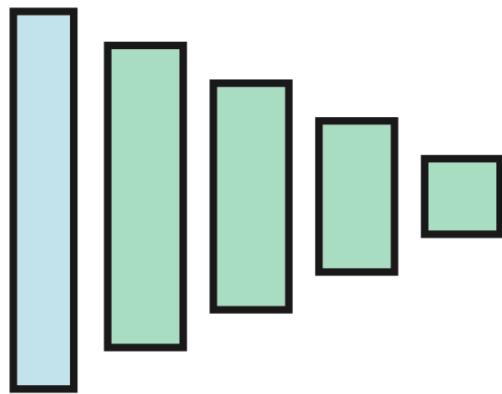


Point Cloud Perception



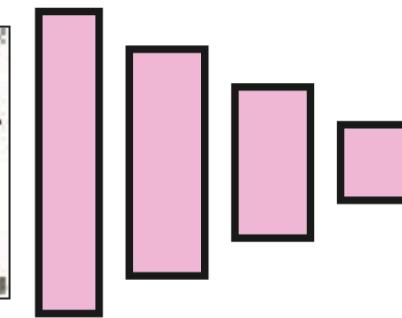
Vision

Perception
for RGB

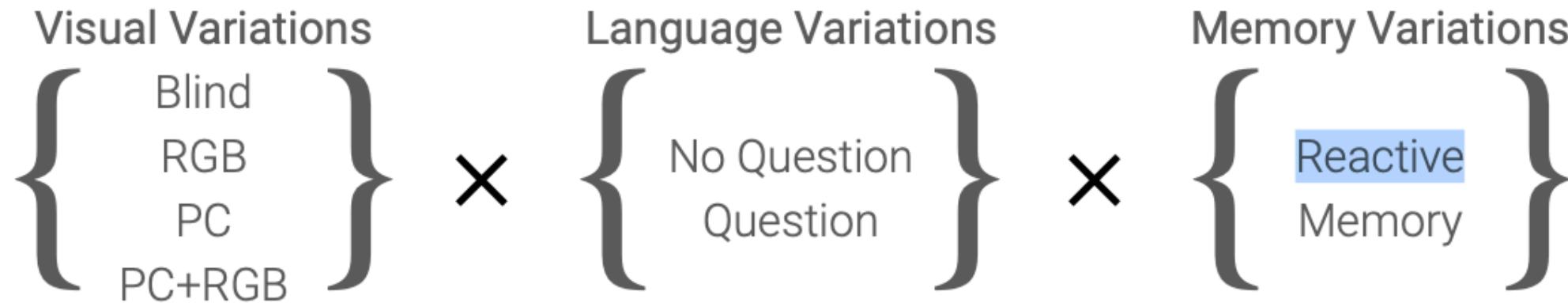


ResNet50 Encoder

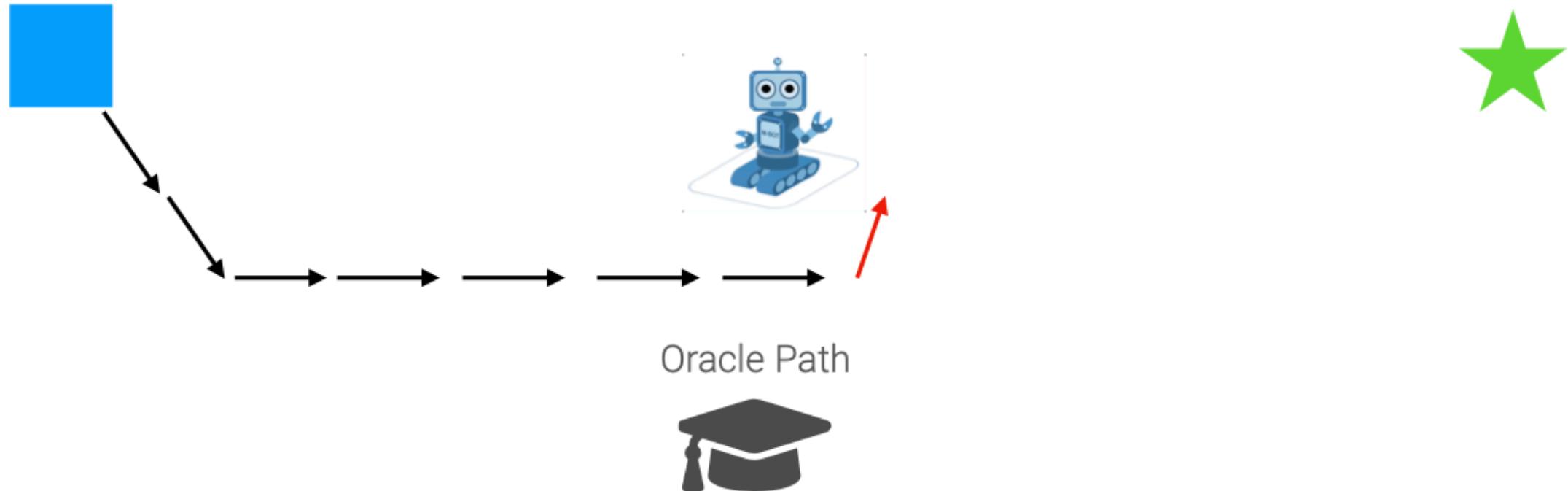
Perception for
Point Clouds



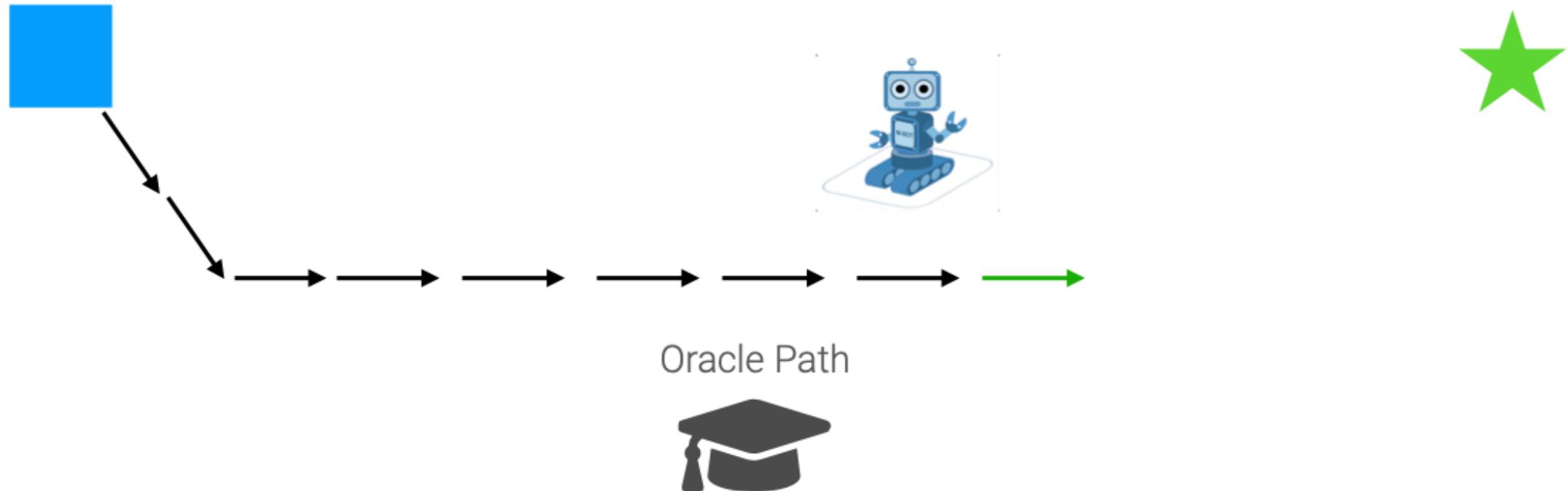
PointNet++ Encoder



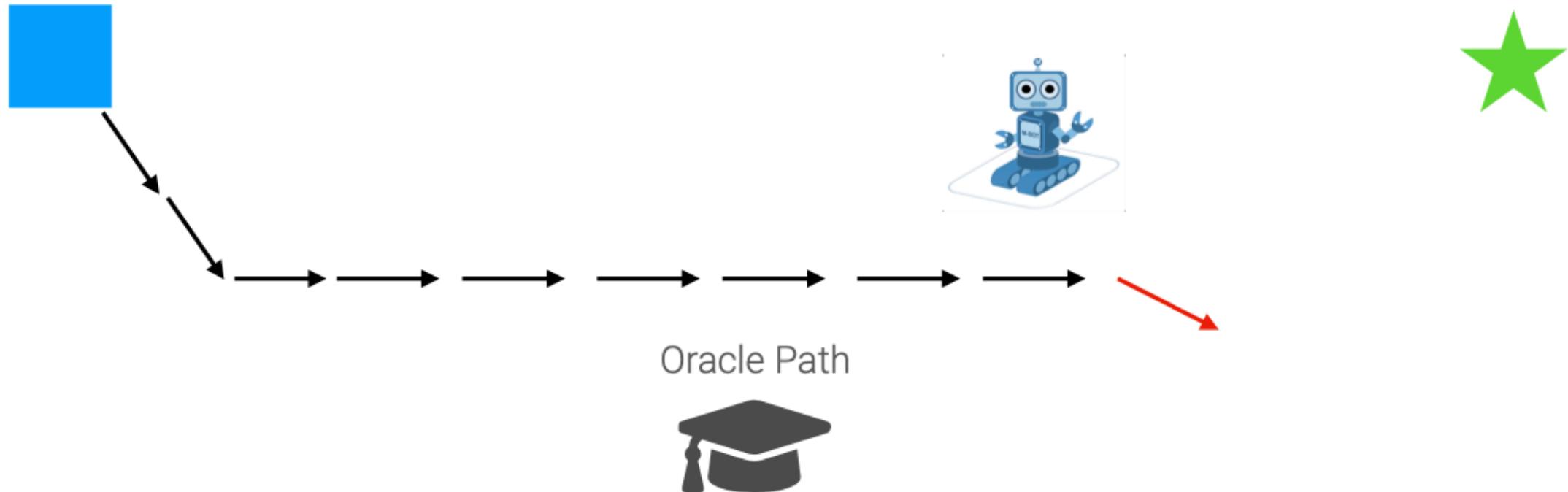
Training Agents for Navigation



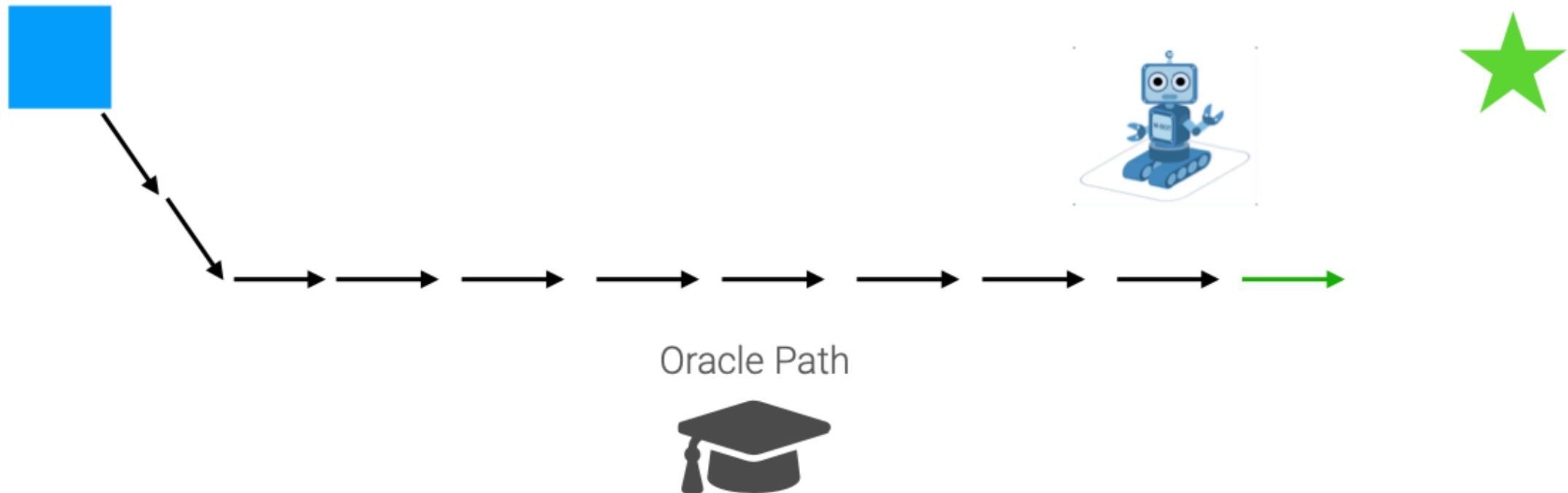
Training Agents for Navigation



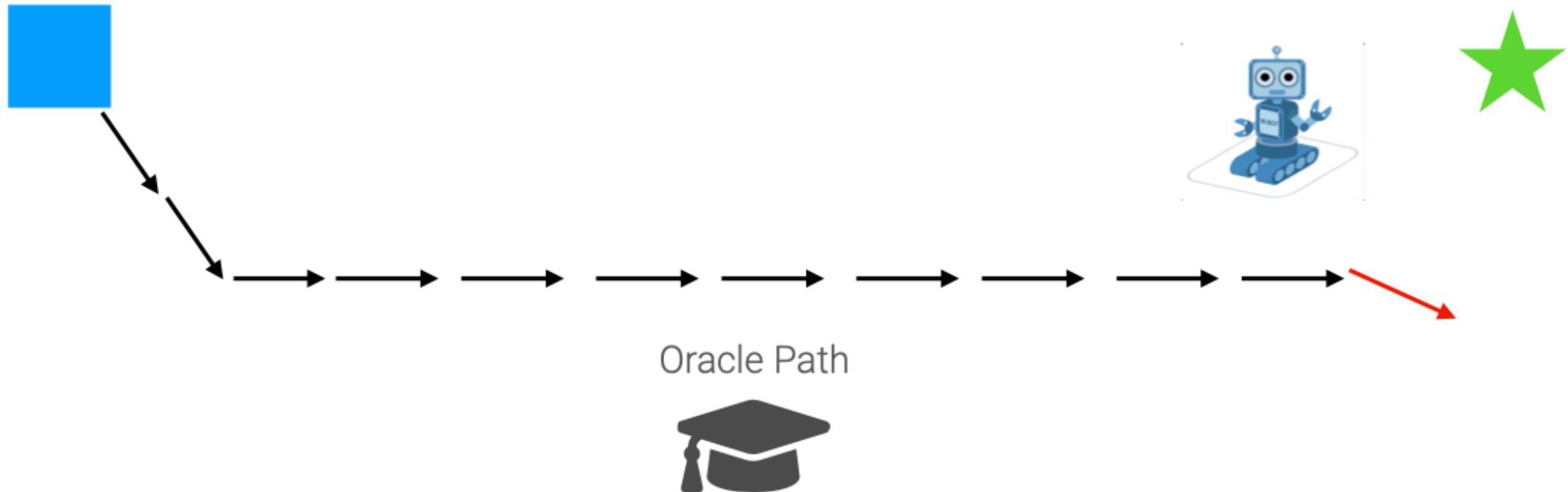
Training Agents for Navigation



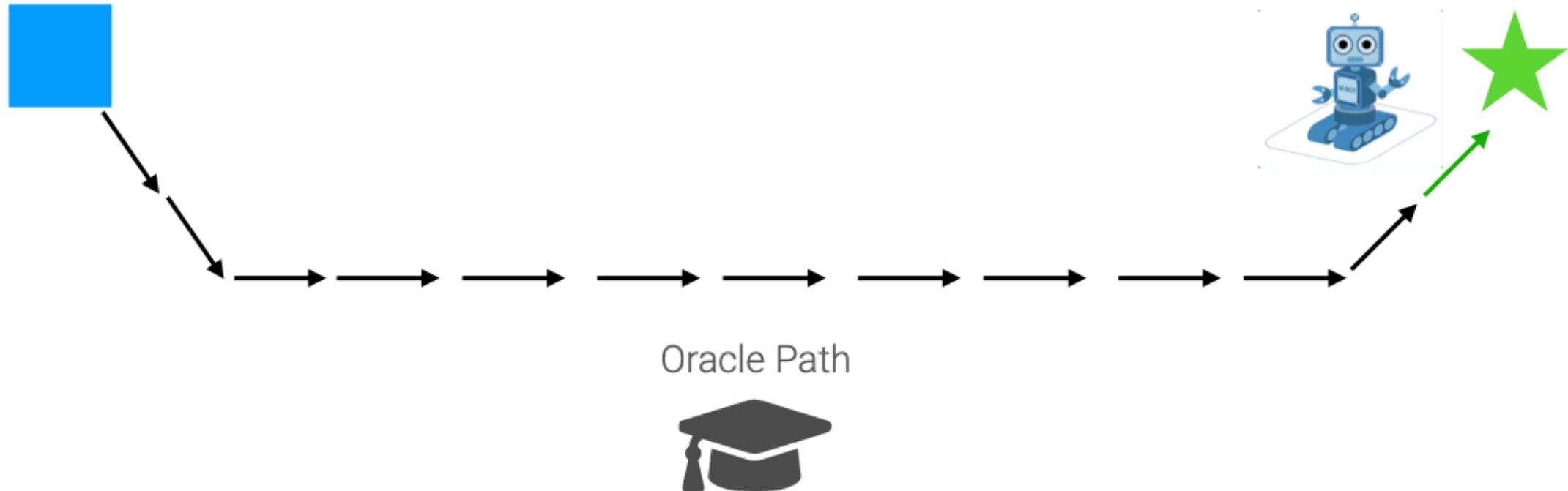
Training Agents for Navigation



Training Agents for Navigation



Training Agents for Navigation



Extension to Real-World: OpenEQA

Episode History H



Question-Answer (Q, A^*) Categories

Object Recognition



Q: What is left of the kitchen pass through?

A*: A bicycle

Attribute Recognition



Q: What colors is the kitchen backsplash?

A*: Black

Object State Recognition



Q: Is the microwave door propped open?

A*: No

Object Localization



Q: Where is the checkers board?

A*: Entryway table

Spatial Reasoning



Q: Can another cookie jar fit on the cookie jar shelf?

A*: Yes

Functional Reasoning



Q: Where can I store the house key?

A*: The lockbox on the door

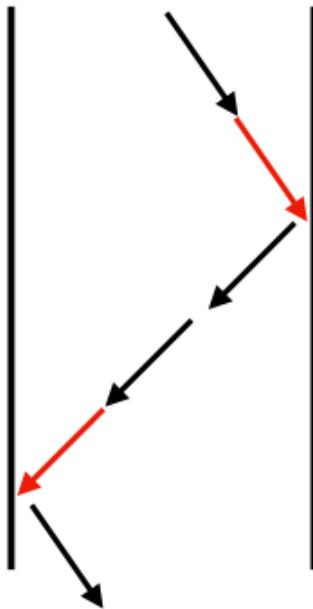
World Knowledge



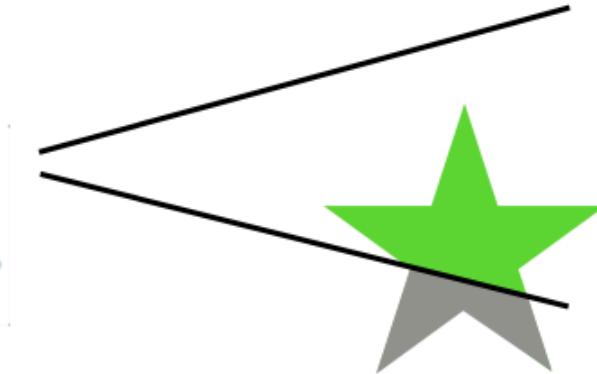
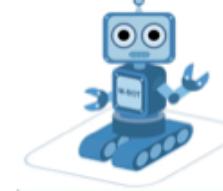
Q: Does this house have forced air heating?

A*: No

Collision Rate (\downarrow better)



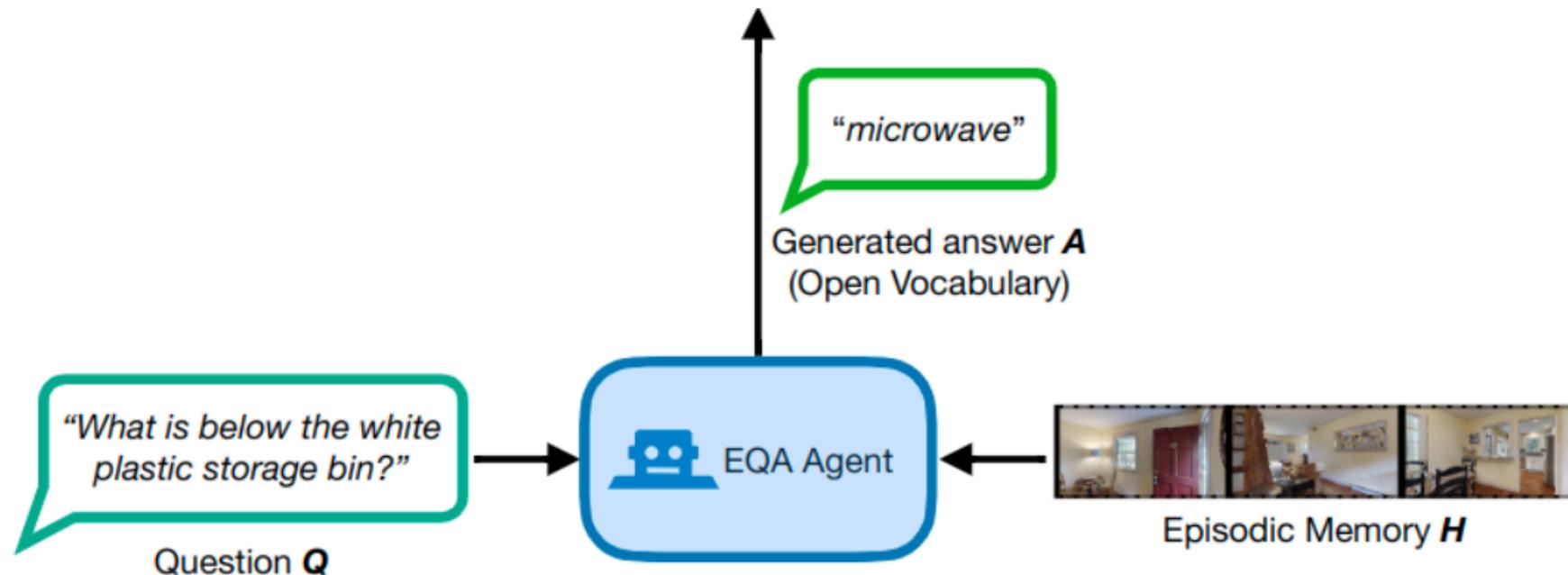
View Quality (\uparrow better)



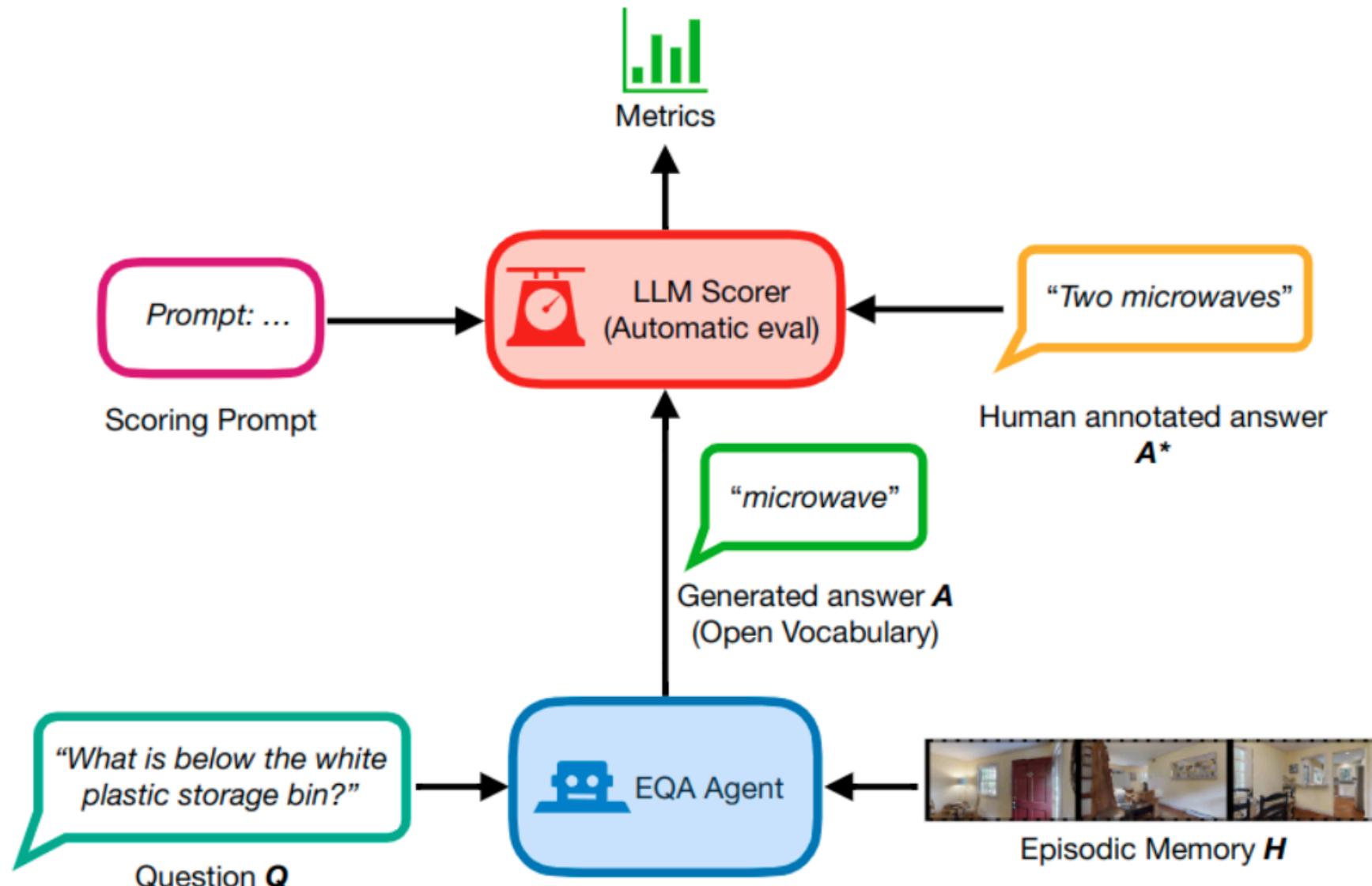
Now Foundation Models are Scorers

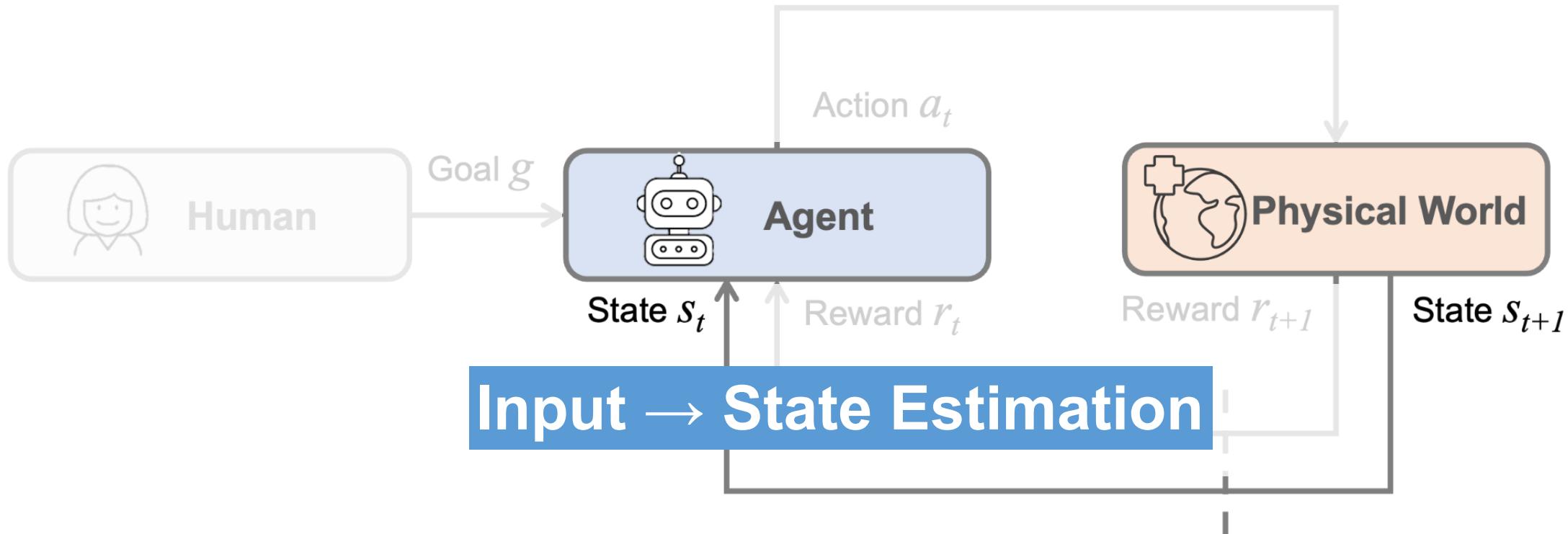


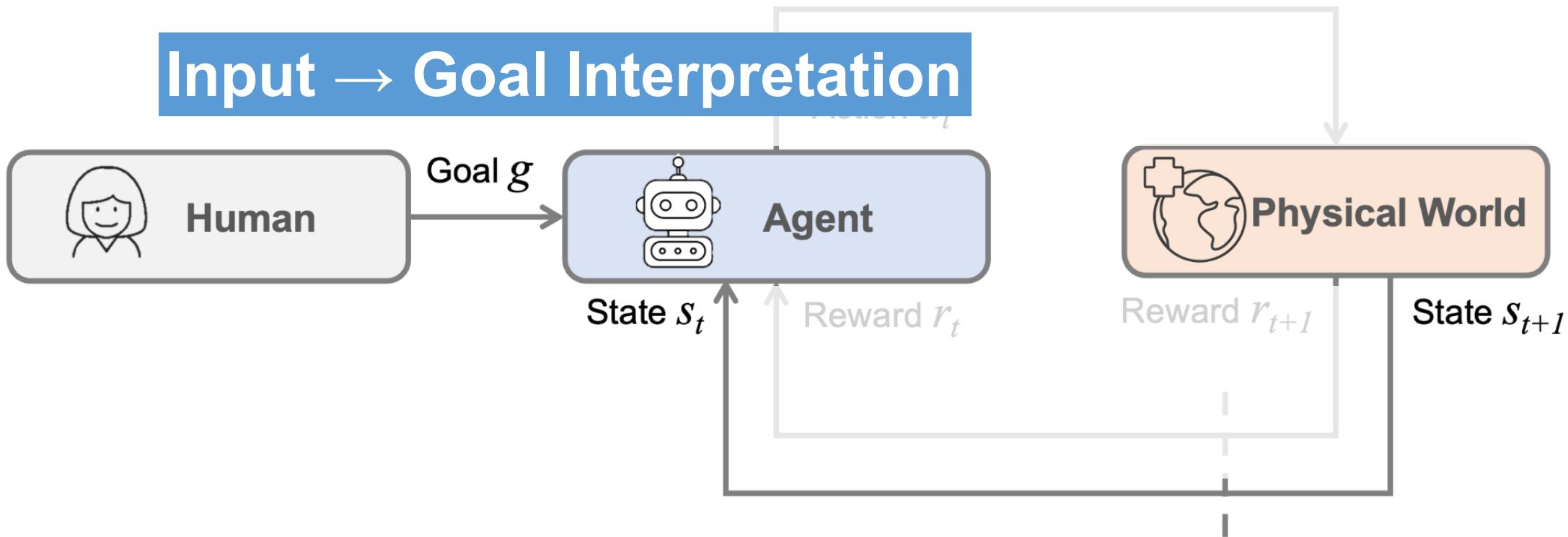
Now Foundation Models are Scorers



New Eval Trend: Now Foundation Models are Scorers







Goal Interpretation

g

Input

Instructions

Put a bowtie into the basket.



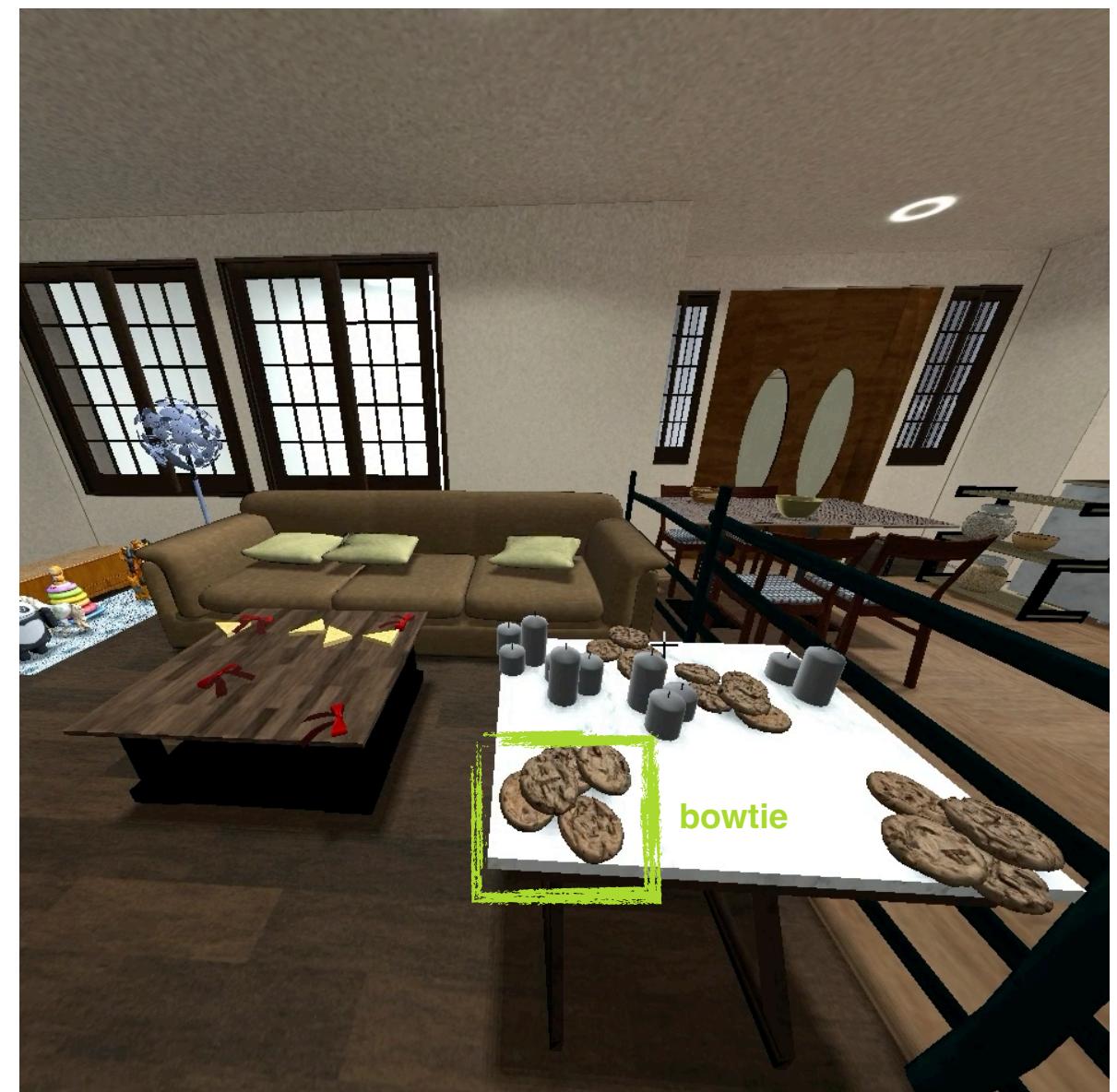
Environment

ontop(bowtie, table)



Symbolic Goal

not ontop(bowtie, table)

LLM Output

Input

Instructions

Put a bowtie into the basket.



Environment

ontop(bowtie, table)

**Symbolic Goal**

not ontop(bowtie, table)

inside(bowtie, basket)

LLM Output

Input

Instructions

Put a bowtie into the basket.



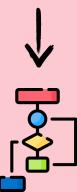
Environment

ontop(bowtie, table)

**Symbolic Goal**

not ontop(bowtie, table)

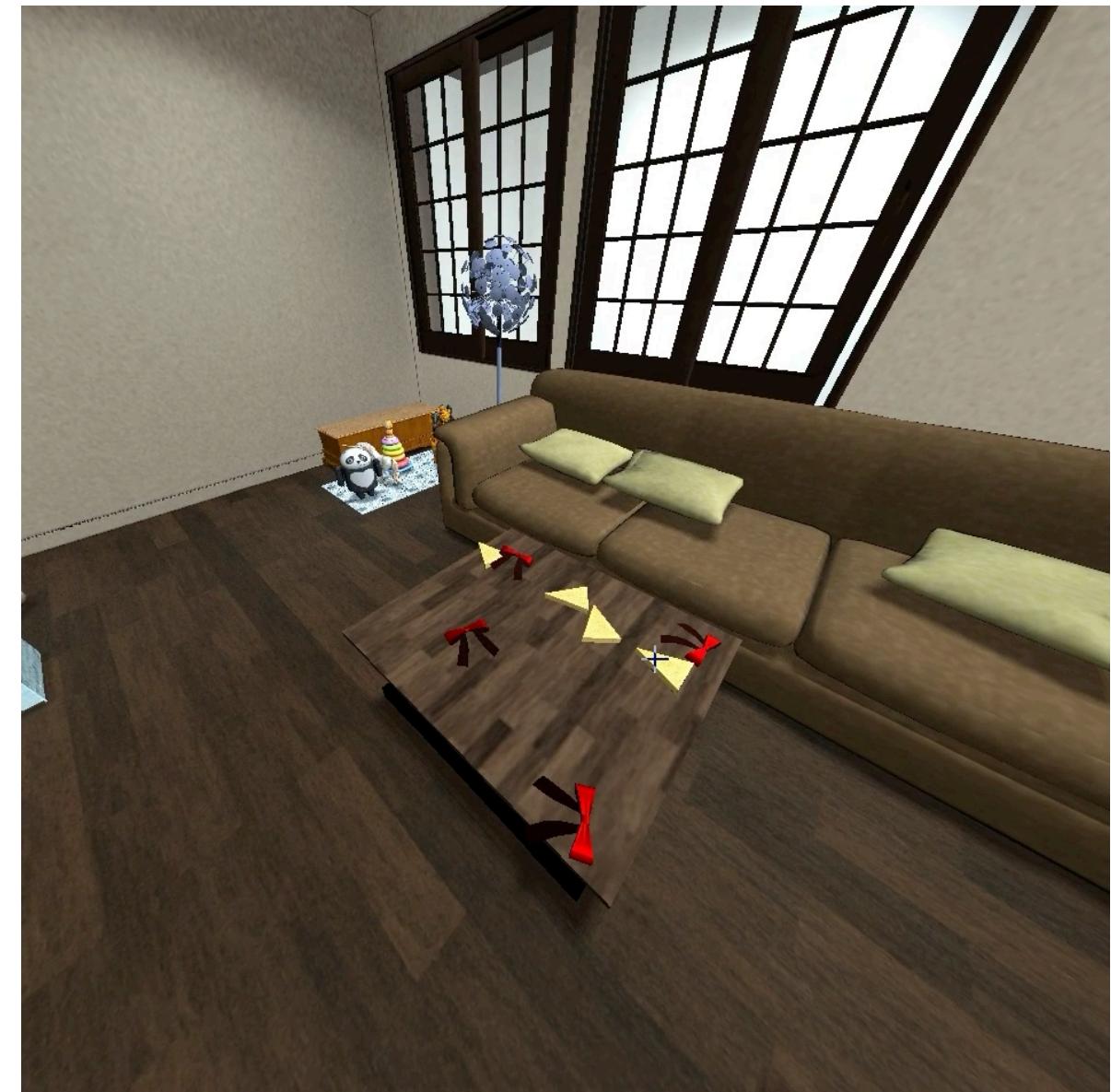
inside(bowtie, basket)



Symbolic Planner



Execute

LLM Output

This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/embodied-agent-interface/embodied-agent-interface>.

VIDEO INPUT



TEXT INPUT

What's inside the apartment: ... The kitchen is equipped with a microwave, eight cabinets, ... Inside the microwave, there is a cupcake. There is a wine glass and an apple on one of the kitchen tables. There are water glasses, a bottle wine, a condiment bottle, and a bag of chips in inside the cabinets. ...

Actions taken by Emily: Emily is initially in the bathroom. She then walks to the kitchen, goes to the sixth cabinet, opens it, subsequently closes it, and then goes towards the fourth cabinet.

QUESTION

Which one of the following statements is more likely to be true?

- (a) Emily has been trying to get a cupcake. ✓ (b) Emily has been trying to get a wine glass. ✗

Multimodal Theory of Mind



Belief

Belief Inference

Type 1.1: True belief, short-term



Scene: ... Inside the bridge, you'll find a bottle of wine...

Actions: ... Finally, she moves towards the fridge, preparing to open it.

Question: If Elizabeth has been trying to get a bottle of wine, which one of the following statements is more likely to be true?

- (a) *Elizabeth thinks that there is a bottle of wine inside the fridge.*
- (b) Elizabeth thinks that there isn't any bottle of wine inside the fridge.

Type 2.1: Goal given true belief



Scene: ... The living room is furnished with a cabinet, ... The cabinet is filled with two apples, ..., and a bottle of wine. ... Inside the fridge, there are two apples.

Actions: James... then opens the fridge, closes it... Finally, he walks towards the living room and approaches the cabinet.

Question: Which one of the following statements is more likely to be true?

- (a) *James has been trying to get a bottle of wine.*
- (b) James has been trying to get an apple.

Type 1.2: False belief, short-term



Scene: ... The living room features a cabinet... The cabinet is filled with a bag of chips, a remote controller, a bottle of wine, and a water glass.

Actions: Jennifer is situated in the living room. She heads towards the cabinet and is about to open it.

Question: If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?

- (a) Jennifer thinks that there isn't a cupcake inside the cabinet.
- (b) *Jennifer thinks that there is a cupcake inside the cabinet.*

Type 1.3: Belief tracking, long-term



Scene: ... The kitchen is equipped with a fridge, sofa, dishwasher, eight cabinets, a stove, a microwave, and a kitchen table...

Actions: ... He walks to the seventh kitchen cabinet, opens and closes it. He repeats the same action with the sixth kitchen cabinet. Subsequently, he moves towards the dishwasher.

Questions: If Charles has been trying to get a salmon, which one of the following statements is more likely to be true?

- (a) Charles thinks that there is a salmon inside the fridge.
- (b) *Charles thinks that there isn't any salmon inside the fridge.*

Goal

Goal Inference

Type 2.1: Goal given true belief



Scene: ... There is a water glass inside the seventh cabinet... The fridge stores two cupcakes...

Actions: Mark... advances towards the seventh kitchen cabinet.

Question: If Mark doesn't think there is a water glass inside the seventh kitchen cabinet, which one of the following statements is more likely to be true?

- (a) Mark has been trying to get a water glass.
- (b) *Mark has been trying to get a condiment bottle.*

Type 2.3: Goal given updated belief



Scene: ... The first cabinet, from left to right, contains a bag of chips.

Actions: Mary... walks towards the first kitchen cabinet, opens it, and then closes it.

Question: Which one of the following statements is more likely to be true?

- (a) Mary has been trying to get a bag of chips.
- (b) *Mark has been trying to get a condiment bottle.*

Type 2.4: Goal given future actions



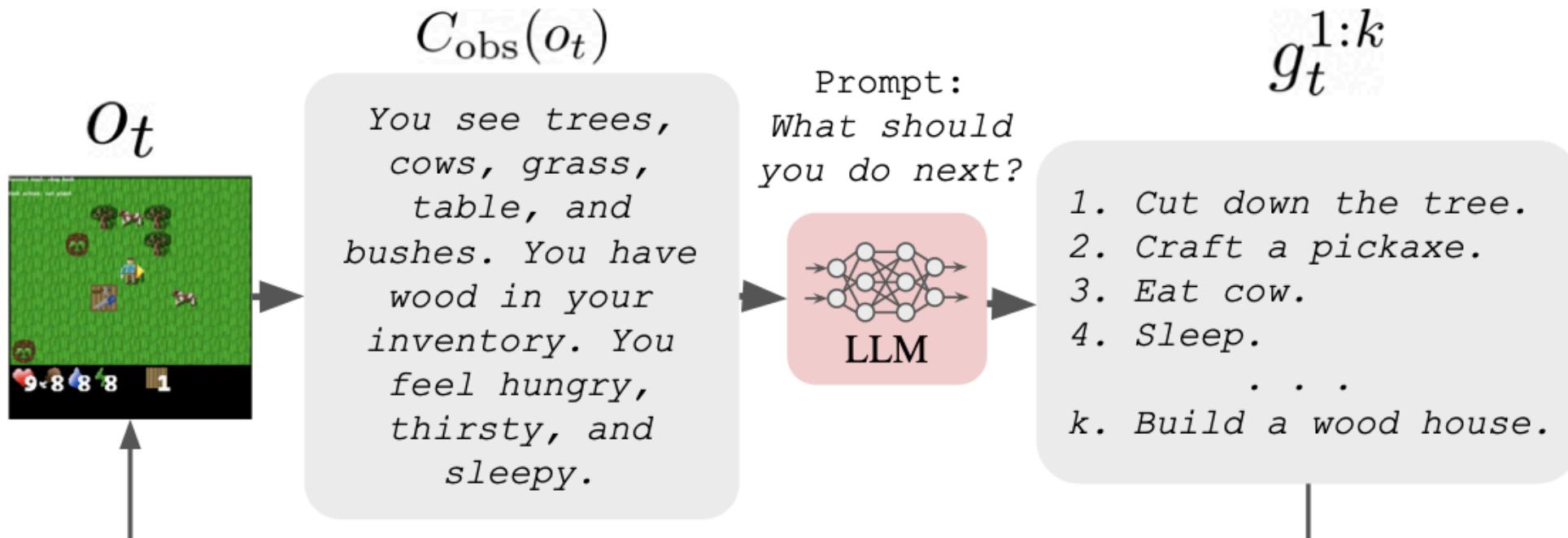
Scene: ... The dishwasher holds a dish bowl... The first cabinet from the left holds a bag of chips and a wine glass... The fifth cabinet has an apple...

Actions: Williams... advances towards the first kitchen cabinet, opens it, and then shuts it. He then moves towards the fifth kitchen cabinet.

Question: Which one of the following statements is more likely to be true?

- (a) William has been trying to get a wine glass.
- (b) *William has been trying to get a dish bowl.*

- Learn to predict plausibly useful goals in a task-agnostic way



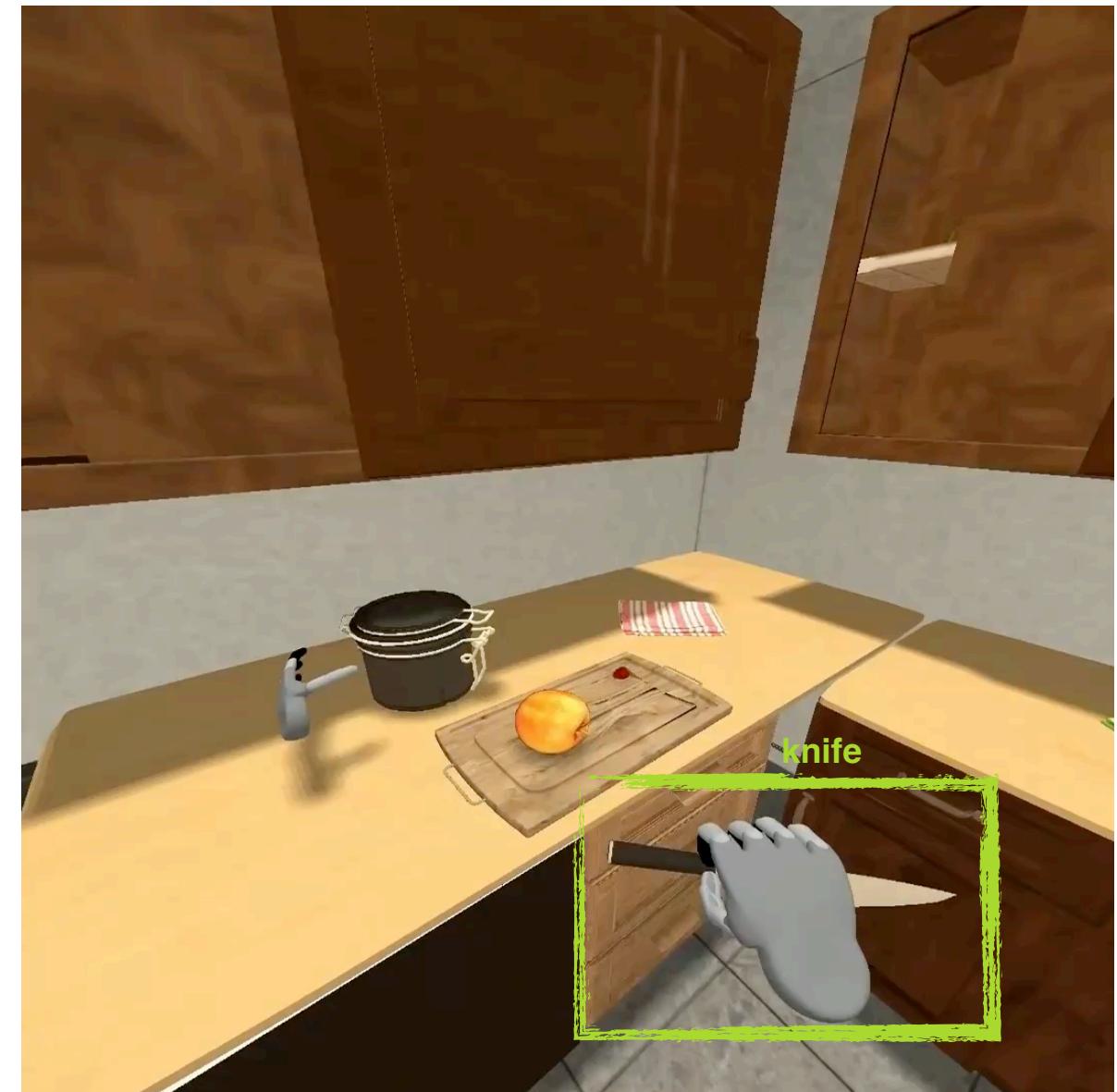
$$\pi(a_t | o_t, g_t^{1:k})$$

$$R_{\text{int}} = \max \left(\Delta(C_{\text{transition}}(o_t, a_t, o_{t+1}), g_t^i) \right), i \in [1 .. k].$$

Input: Bottling Fruit

Environment

right_hand_hold(knife)

**Subgoal Trajectory****LLM Output**

Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)	sliced(strawberry)
inside(peach, jar.0)	inside(strawberry, jar.1)



Subgoal Trajectory

LLM Output



Input: Bottling Fruit

Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)

**Subgoal Trajectory****LLM Output**

Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) **and** sliced(peach)

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) **and** sliced(peach)**S2** inside(peach, jar.0)

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)	sliced(strawberry)
inside(peach, jar.0)	inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) **and** sliced(peach)**S2** inside(peach, jar.0)**S3** inside(strawberry, jar.1)**Actions****A1, A2, A3****LLM Output**

Input: Bottling Fruit



Environment

right_hand_hold(knife)

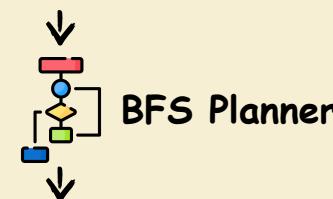


sliced(peach) sliced(strawberry)

inside(peach, jar.0) inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) **and** sliced(peach)**S2** inside(peach, jar.0)**S3** inside(strawberry, jar.1)**Actions****A1, A2, A3****LLM Output**

Input: Bottling Fruit



right_hand_hold(knife)

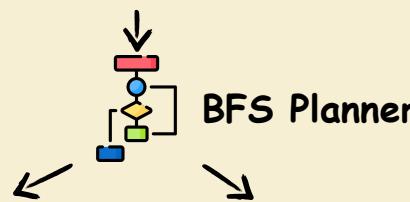


sliced(peach) sliced(strawberry)
inside(peach, jar.0) inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)



GRAB(knife)

SLICE(strawberry)

SLICE(peach)

...

SLICE(strawberry)

SLICE(peach)

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



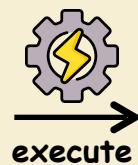
Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)

GRAB(knife)

SLICE(strawberry)

SLICE(peach)



Action Trajectory 1

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)

GRAB(knife)

SLICE(strawberry)

SLICE(peach)



execute

Action Trajectory 1

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



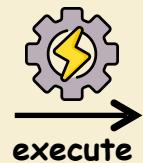
Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)

GRAB(knife)

SLICE(strawberry)

SLICE(peach)



knife is already held

Action Trajectory 1

Failed

LLM Output



Input: Bottling Fruit



right_hand_hold(knife)



sliced(peach)

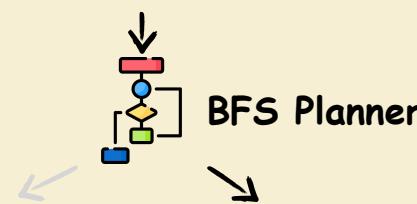
sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)

GRAB(knife)

SLICE(strawberry)

SLICE(peach)

...

SLICE(strawberry)

SLICE(peach)

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)

execute

SLICE(strawberry)

SLICE(peach)

Action Trajectory n

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry)

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) and sliced(peach)



SLICE(strawberry)

SLICE(peach)

Action Trajectory n

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry) ✓

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) ✓ sliced(peach)



SLICE(strawberry)

SLICE(peach)

Action Trajectory n

LLM Output



Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach)

sliced(strawberry) ✓

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) ✓ sliced(peach)



SLICE(strawberry)

SLICE(peach)

Action Trajectory n

LLM Output



Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Bottling Fruit



Environment

right_hand_hold(knife)



sliced(peach) ✓

sliced(strawberry) ✓

inside(peach, jar.0)

inside(strawberry, jar.1)



Subgoal Trajectory

S1 sliced(strawberry) ✓ sliced(peach) ✓

SLICE(strawberry)

SLICE(peach)



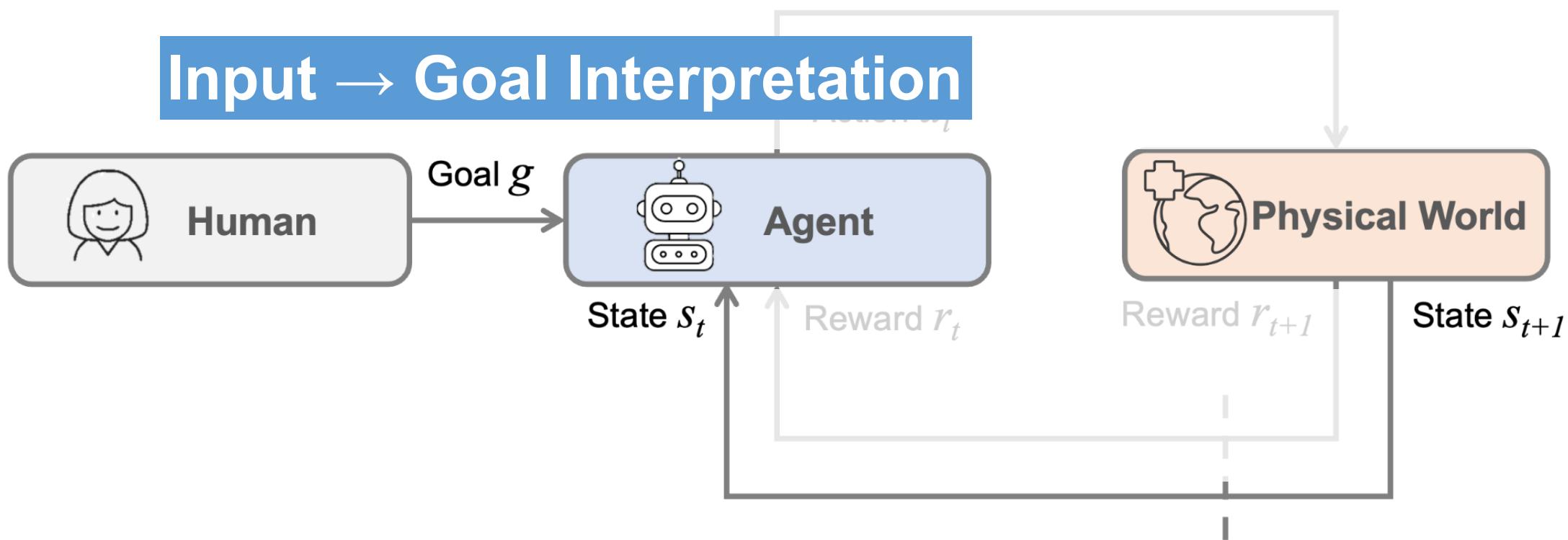
S1 is satisfied

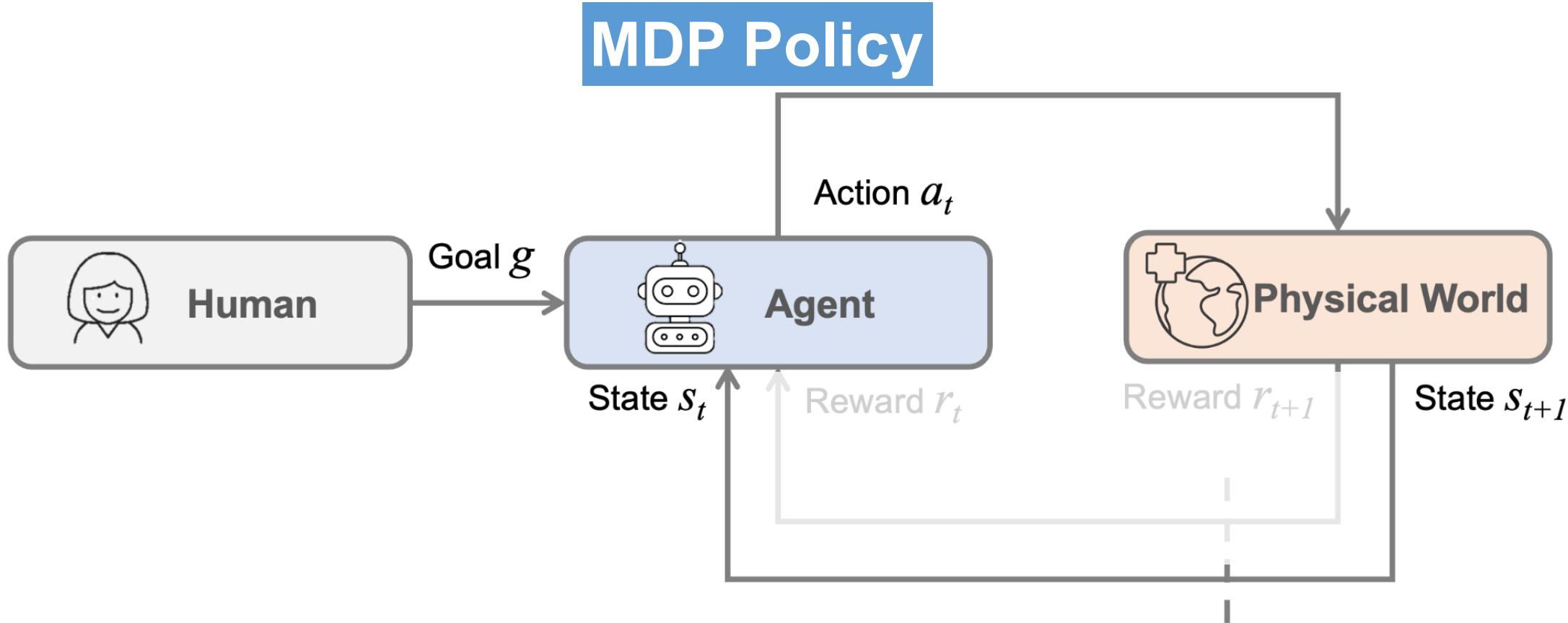
Success ✓

Action Trajectory n

LLM Output







Policy

$$\pi(o, g) \rightarrow a$$

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

...

A7 OPEN (oven)

A8 COOK (strawberry)

...

LLM Output

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

...

A7 OPEN (oven)

A8 COOK (strawberry)

...

LLM Output

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

... ✓

A7 OPEN (oven) ✓

A8 COOK (strawberry)

...

LLM Output

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

... ✓

A7 OPEN (oven) ✓

A8 COOK (strawberry) ✘

...

LLM Output

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

...



A7 OPEN (oven)

**Replanning
Activated**

A8 COOK (strawberry)



...

LLM Output

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

- ... ✓
- A7 OPEN(oven) ✓
- A8 RIGHT_GRASP(pan) ✓
- A9 RIGHT_PLACE_INSIDE(oven)
- A10 CLOSE(oven)
- A11 COOK(strawberry)
- ...

LLM Output

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

- ... ✓
- A7 OPEN (oven) ✓
- A8 RIGHT_GRASP(pan) ✓
- A9 RIGHT_PLACE_INSIDE (oven) ✓
- A10 CLOSE(oven)
- A11 COOK(strawberry)
- ...

LLM Output

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



Operator

BEHAVIOR**Action Trajectory**

- ... ✓
- A7 OPEN(oven) ✓
- A8 RIGHT_GRASP(pan) ✓
- A9 RIGHT_PLACE_INSIDE(oven) ✓
- A10 CLOSE(oven) ✓
- A11 COOK(strawberry)
- ...

LLM Output

Embodied Agent Interface: Benchmarking LLMs for Embodied Decision Making.
Li et al. NeurIPS 2024. <https://embodied-agent-interface.github.io/>

Input: Preserving food

Environment

... inside (strawberry, pan) ...



Goal

... cooked (strawberry) ...



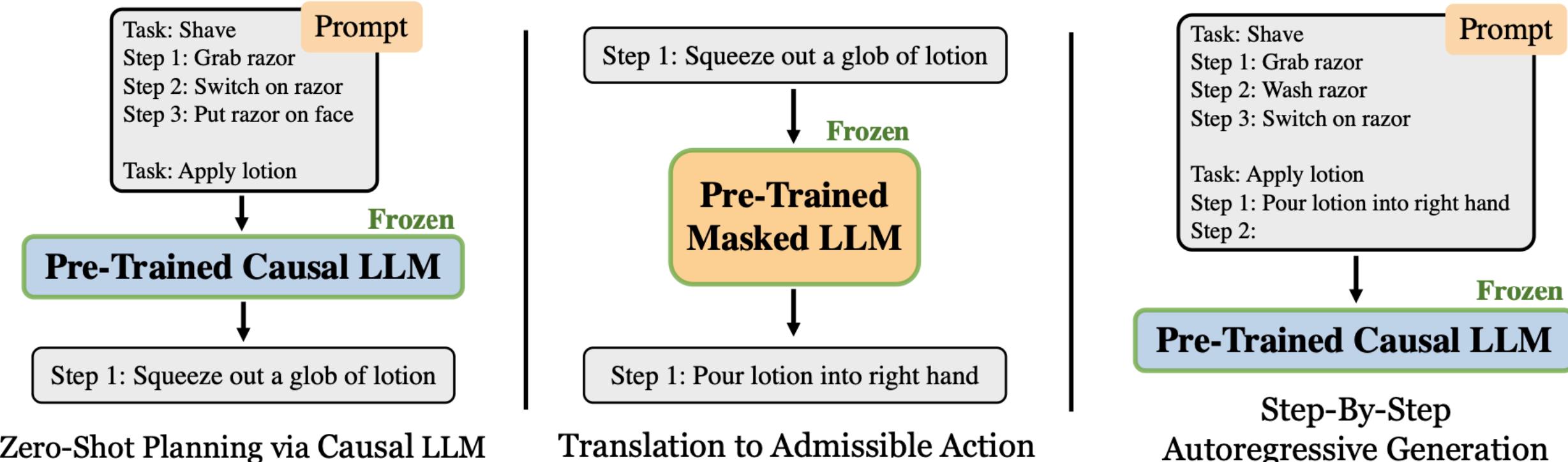
Operator

BEHAVIOR**Action Trajectory**

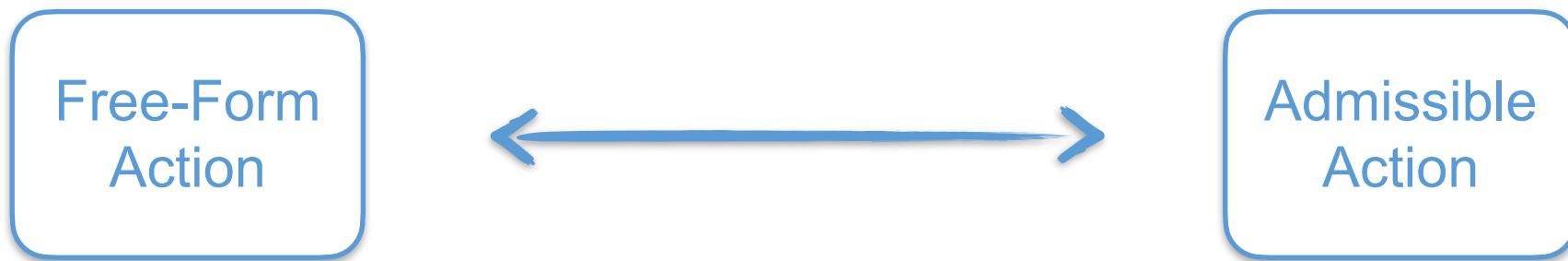
- ... ✓
- A7 OPEN (oven) ✓
- A8 RIGHT_GRASP(pan) ✓
- A9 RIGHT_PLACE_INSIDE (oven) ✓
- A10 CLOSE(oven) ✓
- A11 COOK(strawberry) ✓
- ...

LLM Output

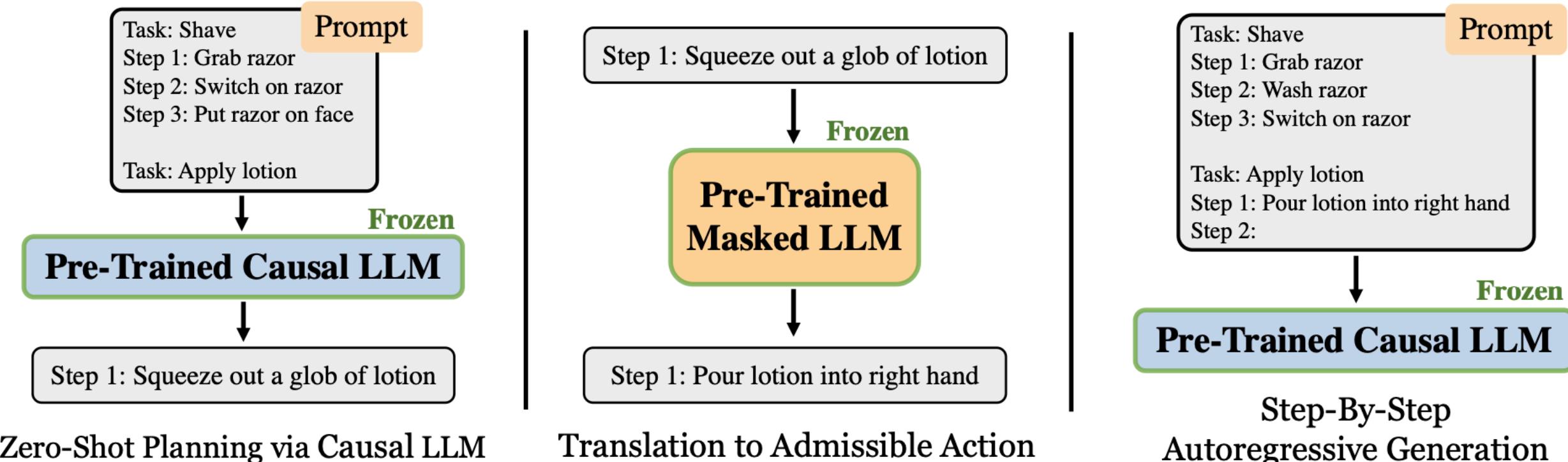
❑ LLMs as the Planner



The key is <Free-Form Language Action, Admissible Action>



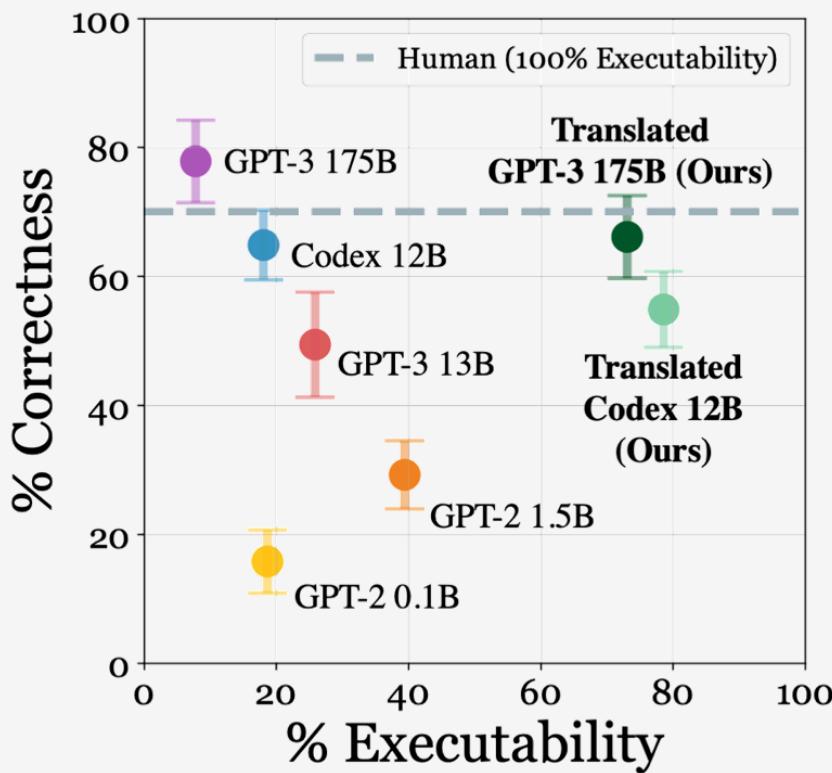
❑ LLMs as the Planner



First Work: Language Models as Zero-Shot Planners



First Work: Language Models as Zero-Shot Planners



Human

Task: Throw away paper

Step 1: Walk to home office
Step 2: Walk to desk
Step 3: Find desk
Step 4: Turn to desk
Step 5: Find chair
Step 6: Sit on chair
Step 7: Find check
Step 8: Grab check
Step 9: Squeeze check
Step 10: Stand up
Step 11: Walk to trashcan
Step 12: Put check on trashcan

GPT-2 1.5B

Task: Brush teeth

Step 1: Go to bathroom

GPT-3 175B

Task: Throw away paper

Step 1: Walk to home office
Step 2: Walk to wastebasket
Step 3: Find wastebasket
Step 4: Grab wastebasket
Step 5: Walk to desk
Step 6: Drop paper in wastebasket

Codex 12B

Task: Brush teeth

Step 1: Walk to bathroom
Step 2: Walk to sink
Step 3: Find toothbrush
Step 4: Pick up toothbrush
Step 5: Put toothbrush in mouth
Step 6: Move brush around mouth for two minutes
Step 7: Spit out toothpaste and brush into sink
Step 8: Turn on water in sink and rinse brush for one minute
Step 9: Turn off water in sink and return brush to cupboard

Translated GPT-3 175B

Task: Throw away paper

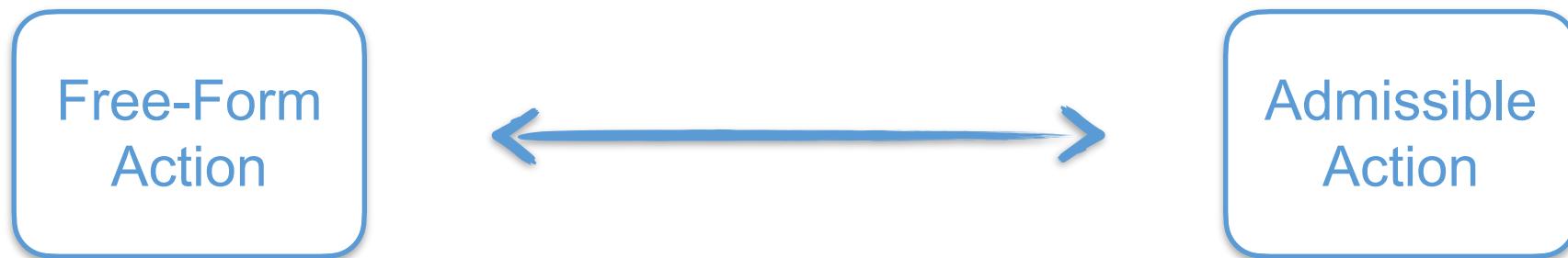
Step 1: Walk to home office
Step 2: Walk to table
Step 3: Find table
Step 4: Turn to table
Step 5: Find paper
Step 6: Grab paper
Step 7: Walk to trashcan
Step 8: Open trashcan
Step 9: Put paper on trashcan
Step 10: Close trashcan

Translated Codex 12B

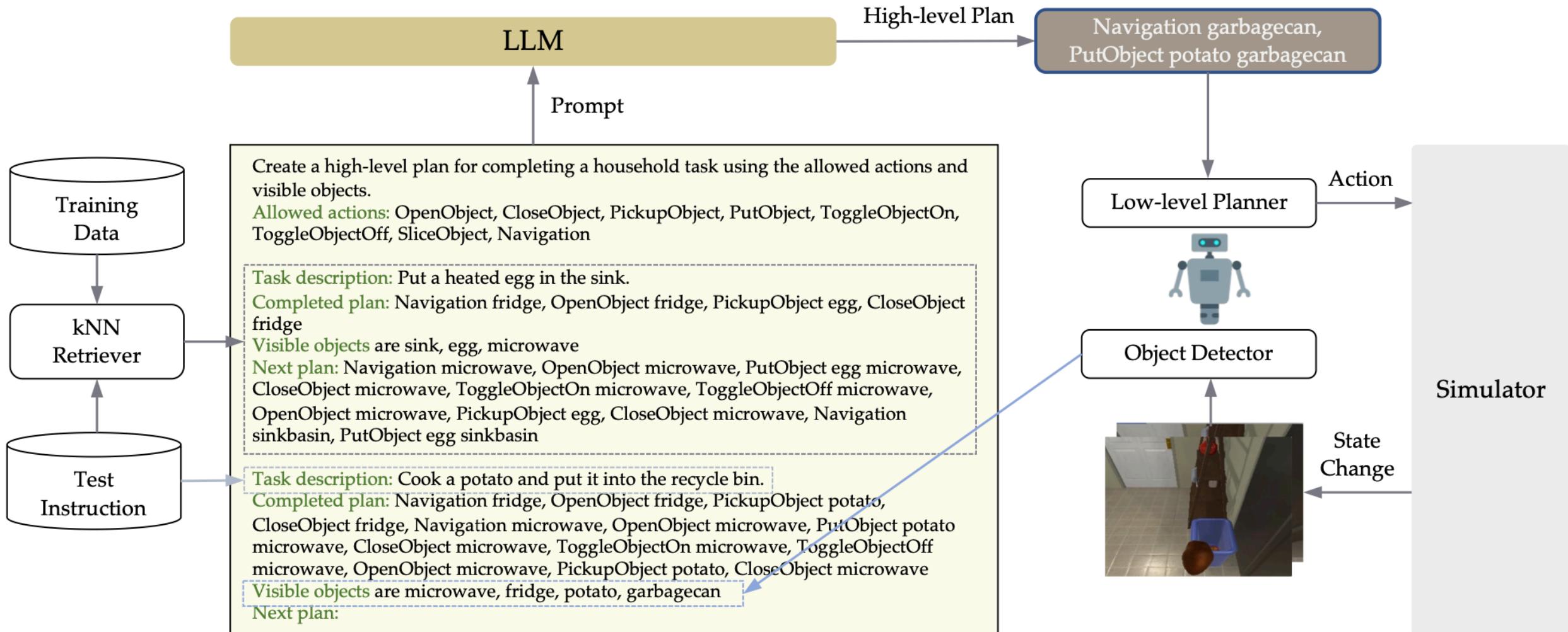
Task: Brush teeth

Step 1: Walk to bathroom
Step 2: Open door
Step 3: Walk to sink
Step 4: Put pot on sink
Step 5: Put brush on toothbrush
Step 6: Turn to toothpaste
Step 7: Put toothpaste on toothbrush
Step 8: Put teeth on toothbrush

Another way of mapping <Free-Form Language Action, Admissible Action>



Another way: Retrieval Based



Learn the Skill Library (Admissible Action)

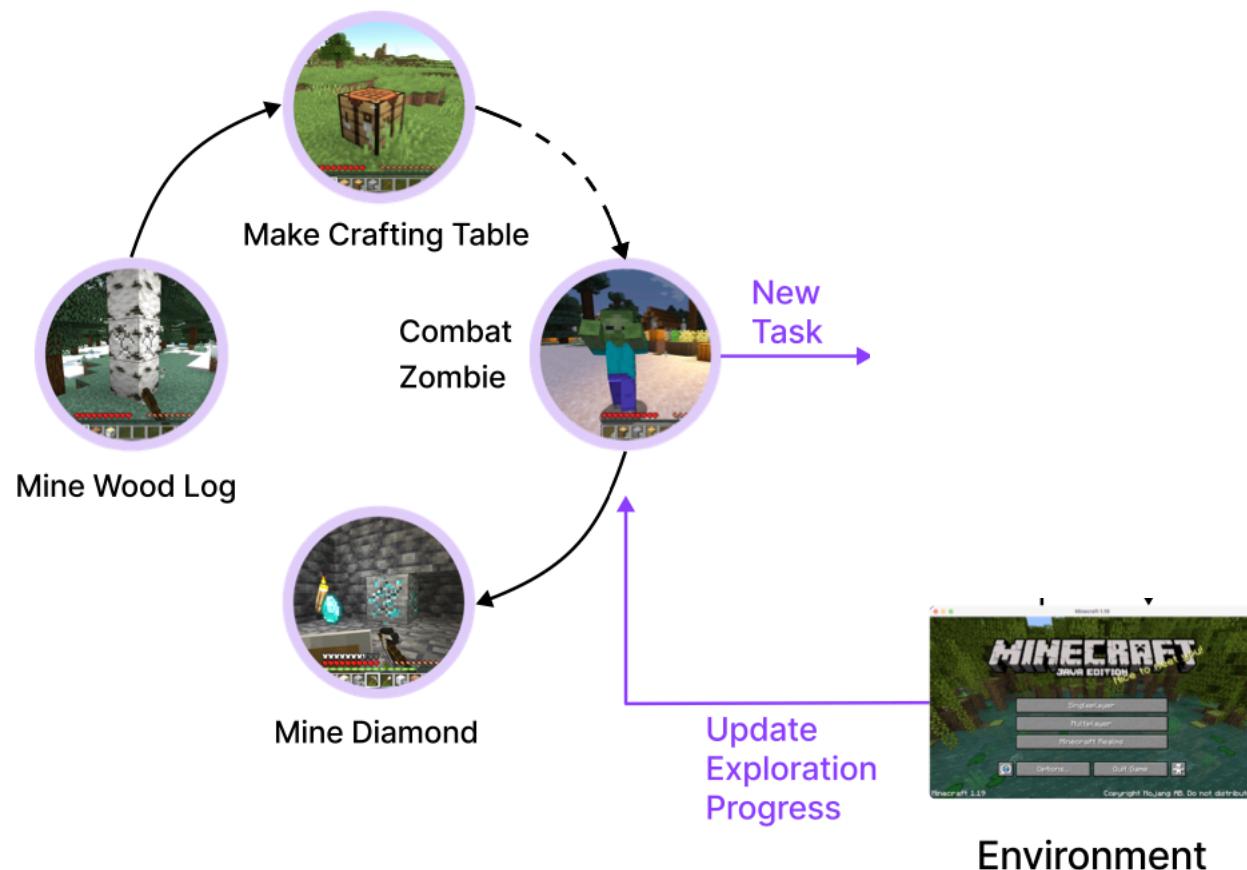
- Focuses on agent policies for exploration and skill acquisition in Minecraft



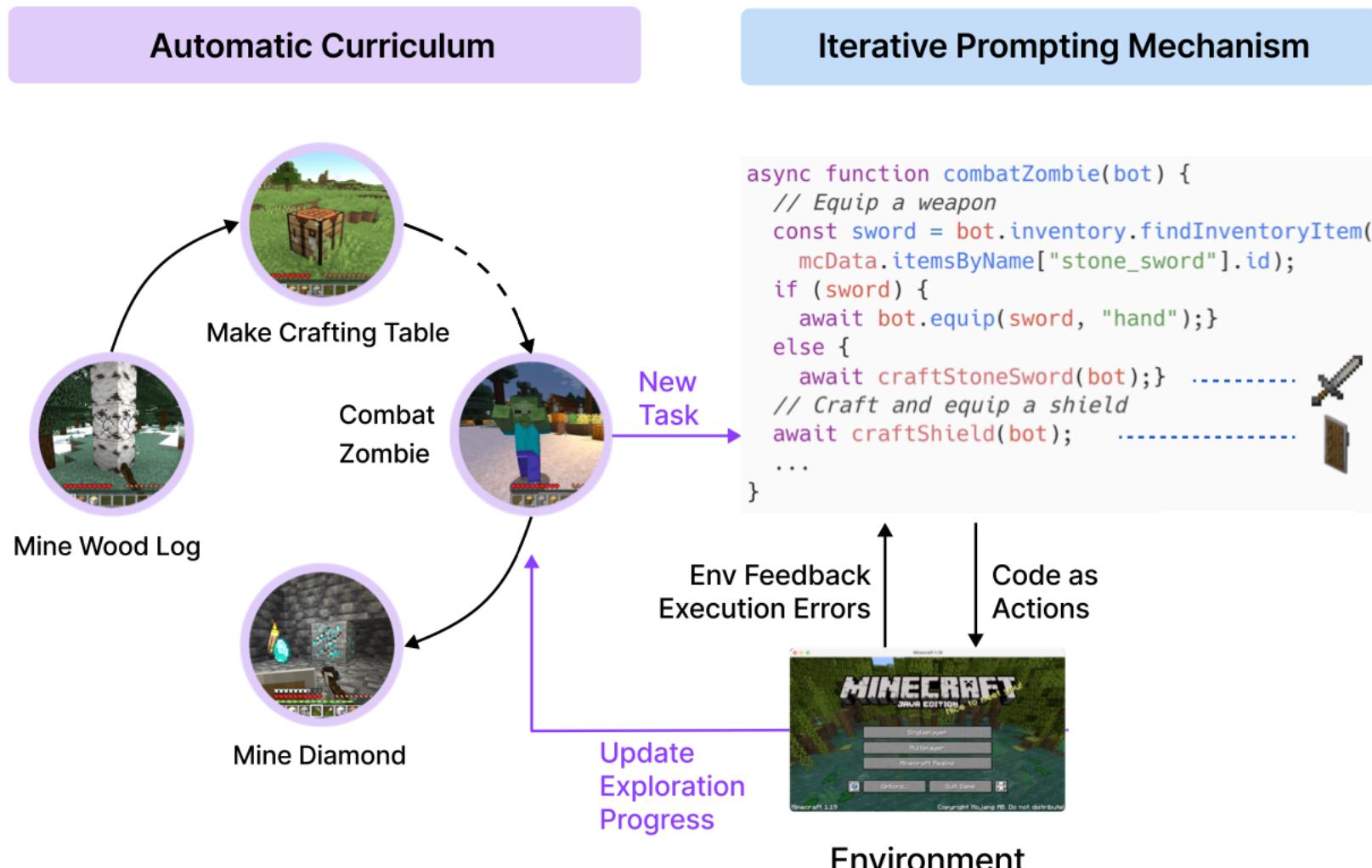
Environment

- Focuses on agent policies for exploration and skill acquisition in Minecraft

Automatic Curriculum

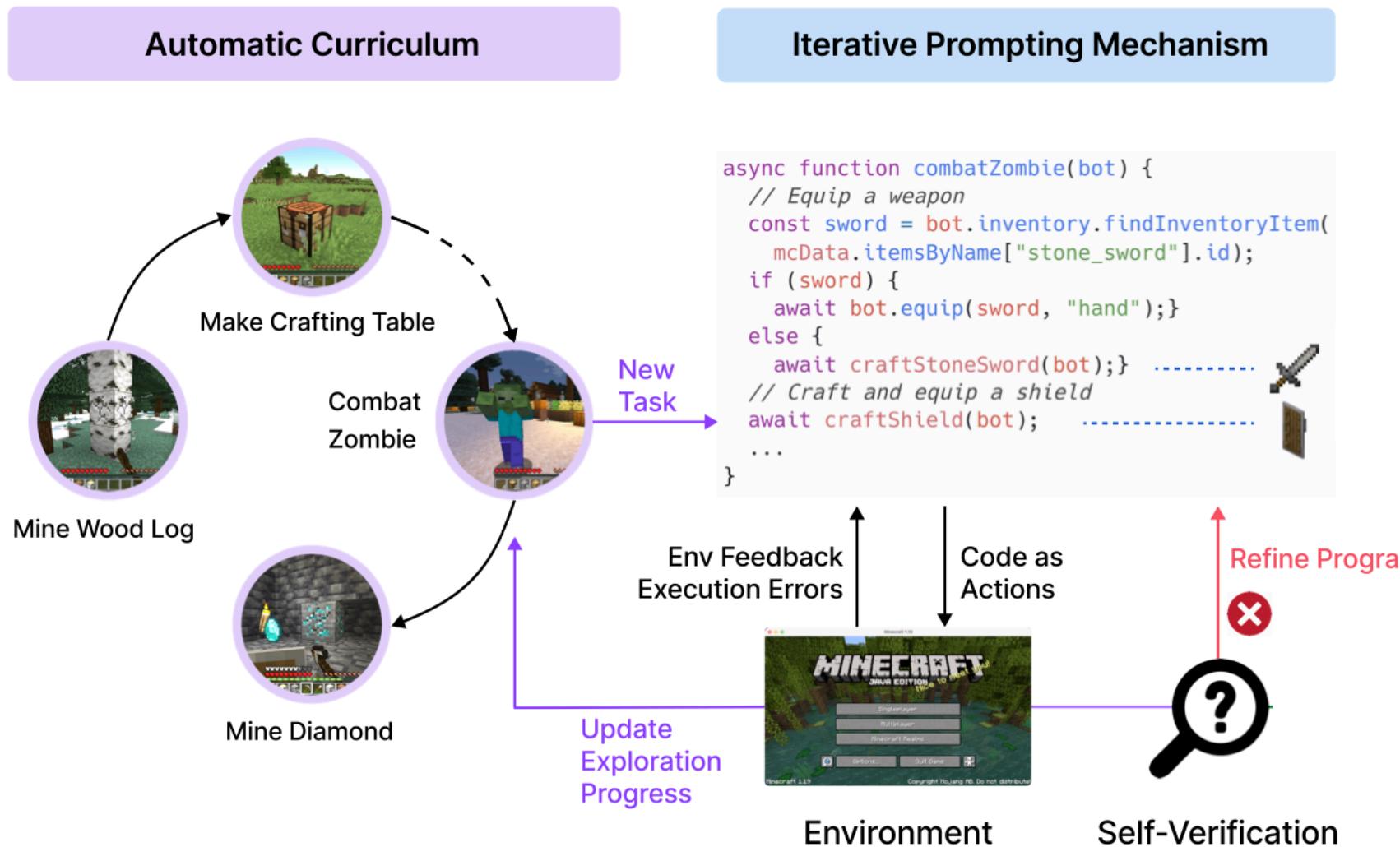


- Focuses on agent policies for exploration and skill acquisition in Minecraft



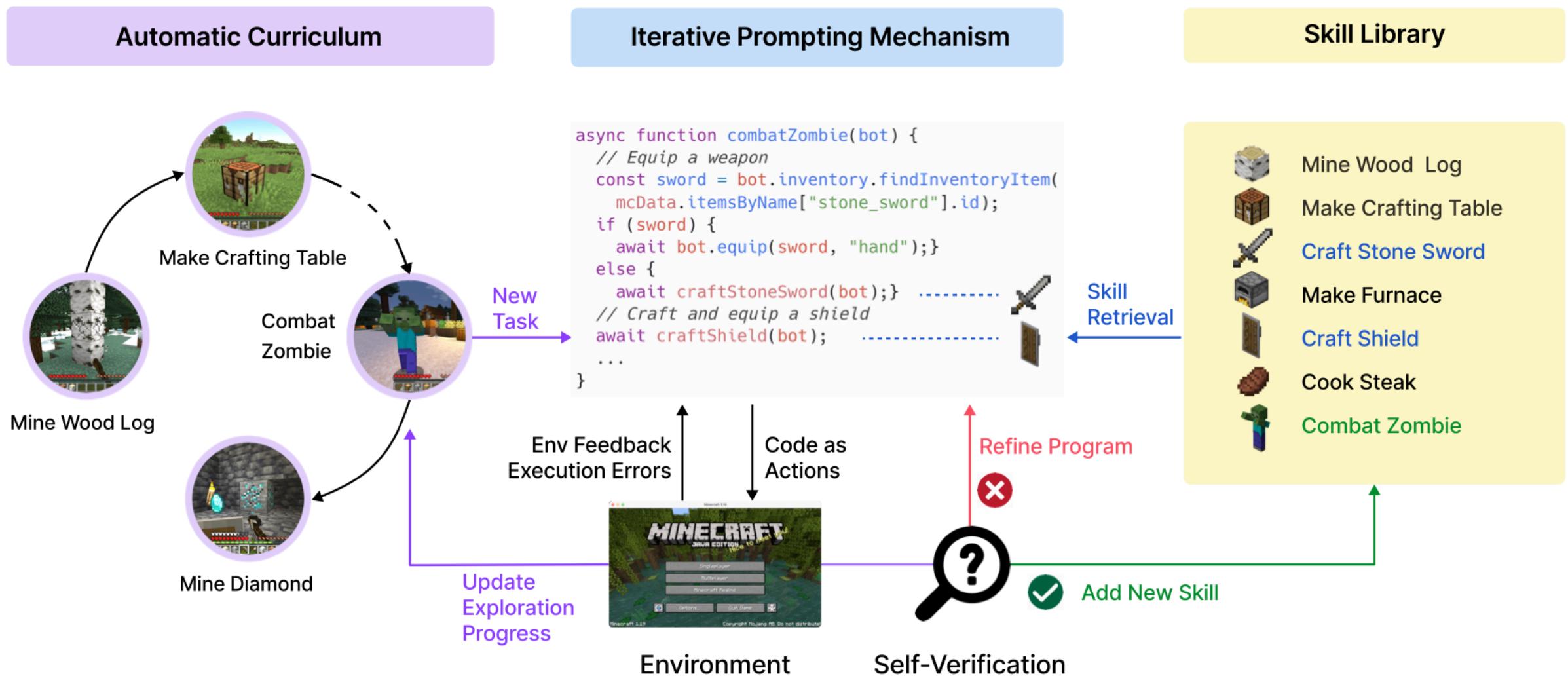
Voyager: Policy & Skill Acquisition

- Focuses on agent policies for exploration and skill acquisition in Minecraft



Voyager: Policy & Skill Acquisition

- Focuses on agent policies for exploration and skill acquisition in Minecraft



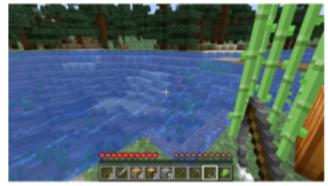
□ Step 1: Propose tasks by the automatic curriculum



Inventory (5/36): {'oak_planks': 3, 'stick': 4, 'crafting_table': 1, 'stone': 3, 'wooden_pickaxe': 1}

GPT-4

Reasoning: Since you have a wooden pickaxe and some stones, it would be beneficial to upgrade your pickaxe to a stone pickaxe for better efficiency.
Task: Craft 1 stone pickaxe.



Biome: river

Inventory (4/36): {'sugar_cane': 3, 'paper': 3, 'fishing_rod': 1, 'wooden_pickaxe': 1}

GPT-4

Reasoning: Since you have a fishing rod in your inventory and you are near a river biome, it would be a good opportunity to catch some fish for food and experience.
Task: Catch 1 fish.



Nearby entities: pig, cat, villager

Health: 12/20

Hunger: 0/20

GPT-4

Reasoning: Your hunger is at 0, which means you need to find food to replenish your hunger. Since there are pigs nearby, you can kill one to obtain raw porkchops.
Task: Kill 1 pig.



Inventory (6/36): {'furnace': 1, 'stone_pickaxe': 1, 'oak_planks': 7, 'crafting_table': 1, 'raw_iron': 4, 'coal': 1}

GPT-4

Reasoning: You have raw iron and coal, and you have a furnace. It's time to smelt the iron to make iron ingots, which can be used to craft better tools and armor.
Task: Smelt 4 raw iron.



Time: night

Nearby entities: zombie

Equipment: 'stone_sword', 'shield'

GPT-4

Reasoning: Since it's night and there's a zombie nearby, it's a good opportunity to try killing the zombie now that you have a stone sword and a shield equipped.
Task: Kill 1 zombie.

Voyager: Policy & Skill Acquisition



□ Step 2: Skill Library

Program Generated by GPT-4

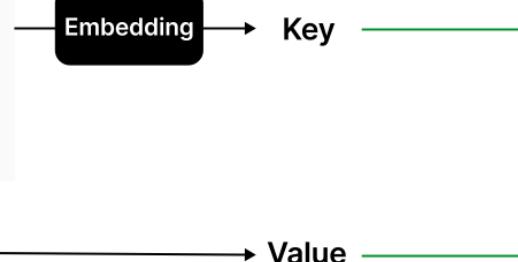
```
async function combatZombie(bot) {  
    // Equip a weapon  
    const sword =  
    bot.inventory.findInventoryItem(  
        mcData.itemsByName[  
            "stone_sword"  
        ].id  
    );  
    if (sword) {  
        await bot.equip(sword, "hand");  
    } else {  
        await craftStoneSword(bot);  
    }  
    // Craft and equip a shield  
    ...  
    // Recover hunger  
    ...  
    // Look for and combat a zombie  
    ...  
}
```

Program Description

```
async function combatZombie(bot) {  
    // The function is about  
    equipping a stone sword to combat  
    a zombie. If a stone sword is not  
    found, it will craft one.  
    Additionally, it crafts and equips  
    a shield for added protection.  
    Afterwards, it proceeds to cook  
    sticks in order to restore hunger.  
    Once hunger is replenished, it  
    actively searches for a zombie and  
    engages in combat with it.  
}
```

GPT-3.5

Adding a new skill



Skill Library

	Mine Wood Log
	Make Crafting Table
	Craft Wooden Pickaxe
	Craft Stone Sword
	Make Furnace
...	
	Combat Cow
	Cook Steak
	Craft Iron Axe
	Combat Zombie

Task: Craft Iron Pickaxe

How to craft an iron pickaxe in
Minecraft?

GPT-3.5

To craft an iron pickaxe, you
need to 3 iron ingots and 2
sticks. Once you have gathered
the materials,

Environment Feedback

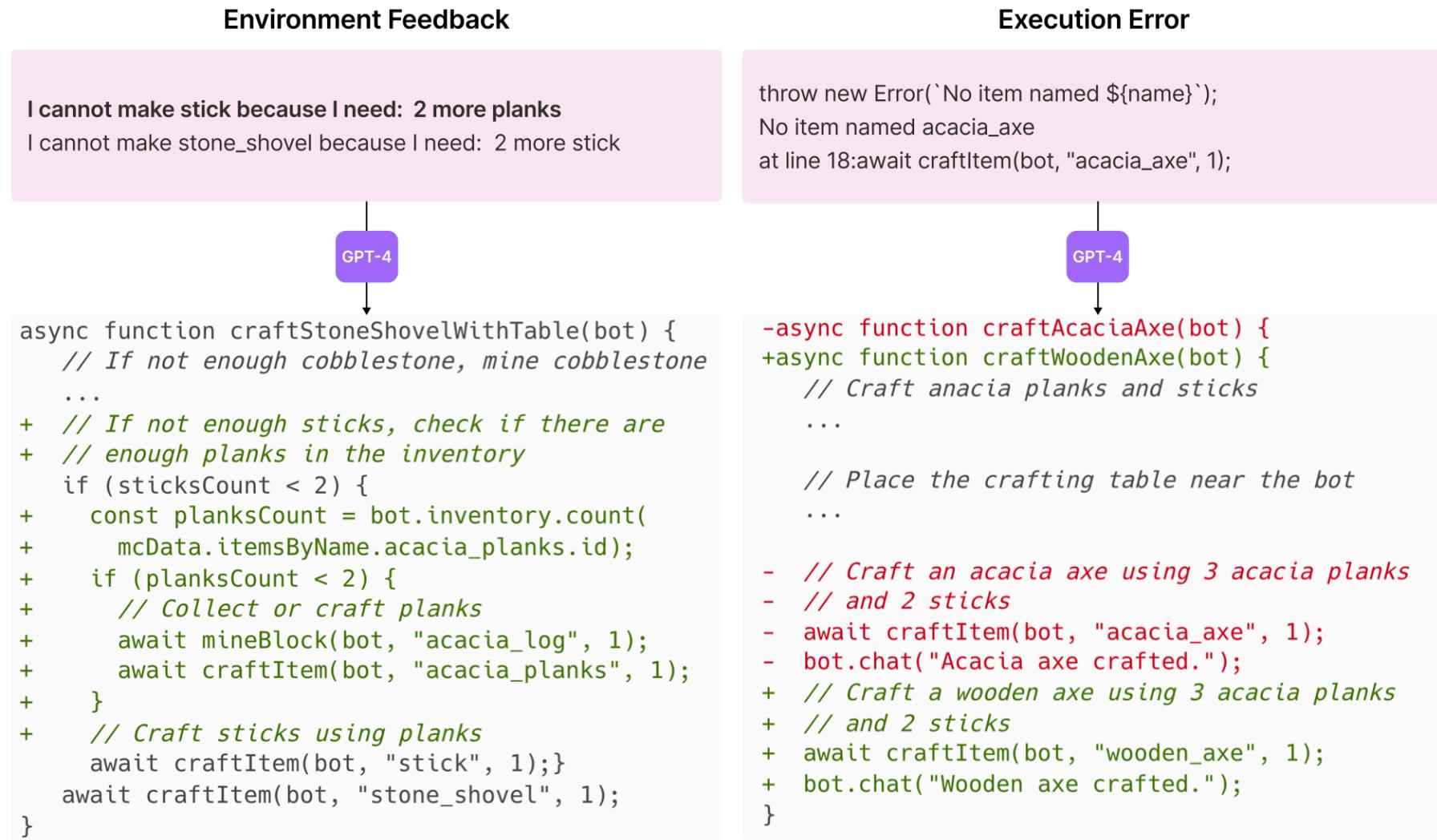
Skill Retrieval



Top-5 Relevant Skills

	Smelt Iron Ingot
	Craft Stick
	Make Crafting Table
	Make Furnace
	Craft Wooden Pickaxe

Code as Actions, refining with environment feedback



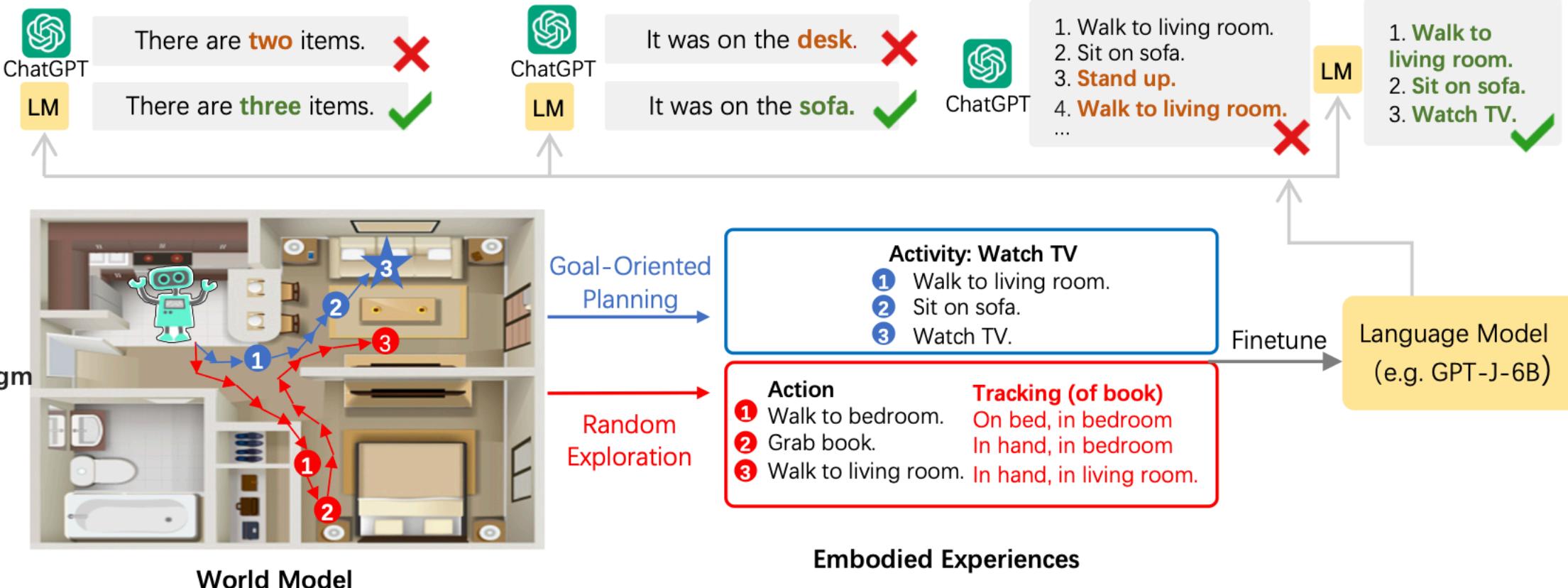


LLMs can also be finetuned / trained



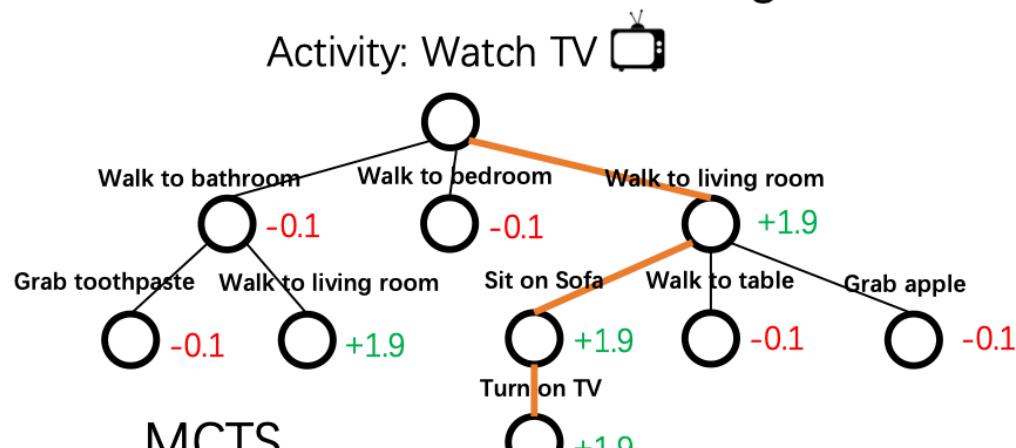
□ Embodied Experiences from World Models: Synthetic Data Creation

Answer:



□ Embodied Experiences from World Models: Synthetic Data Creation

Goal-Oriented Planning



Question:
How to watch TV? TV and sofa is in living room...

Answer:
Walk to living room. Sit on sofa. Turn on TV.

Plan Generation

Question:
Given a plan: Walk to living room. Sit on sofa. Turn on TV.
What is the task?

Answer: Watch TV.

Activity Recognition

Random Exploration

- 1 Grab pillow
- 2 Give pillow to 
- 3 Take pillow
- 4 Grab apple
- 5 Walk to living room
- 6 Put apple on table
- 7 Walk to bathroom
- 8 Walk to bedroom
- 9 Put pillow on table



Question:
Tom grabbed pillow. Tom gave pillow to ... How many objects are on the table?

Answer:
Two. They are pillow and apple.

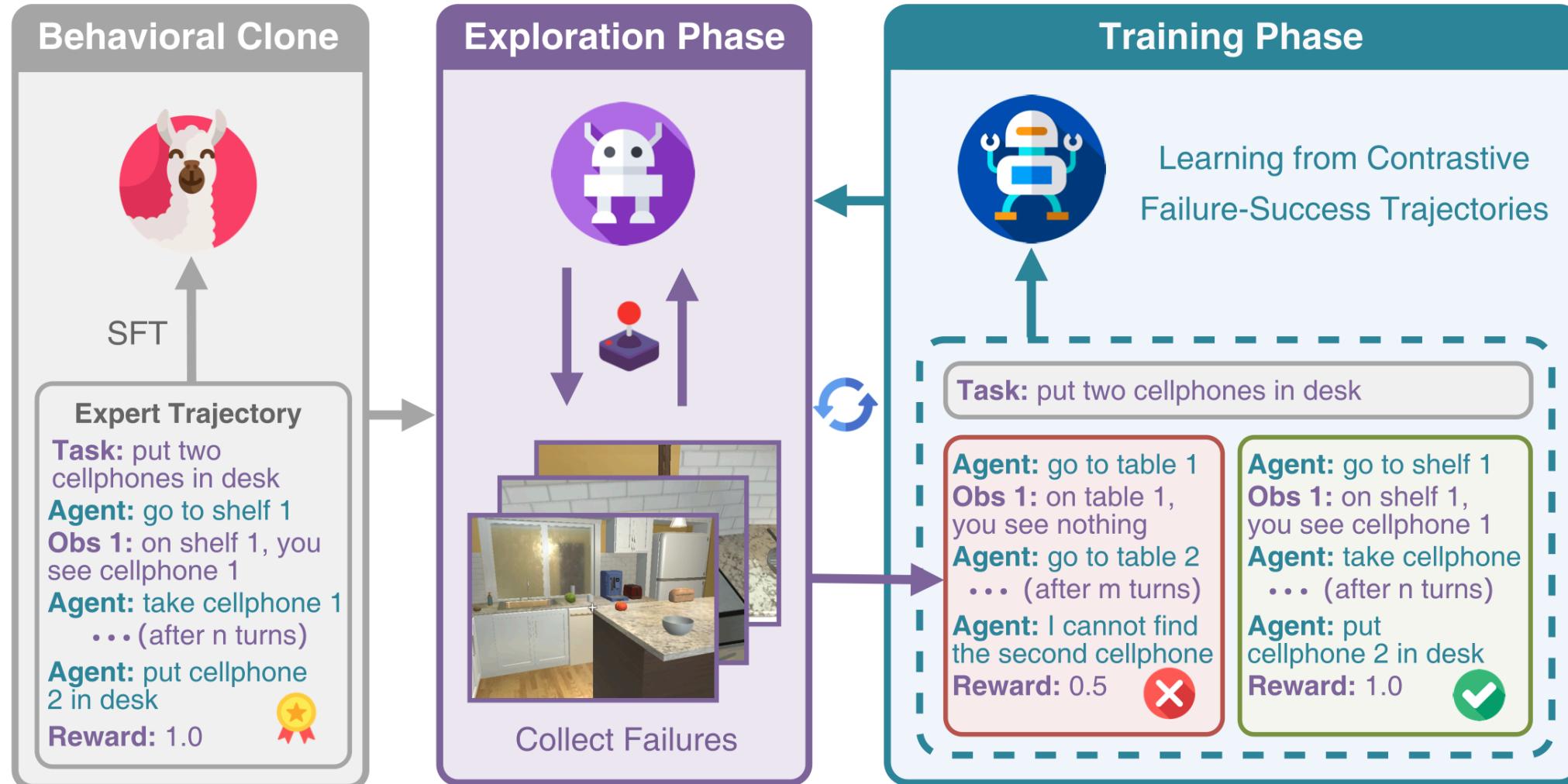
Counting

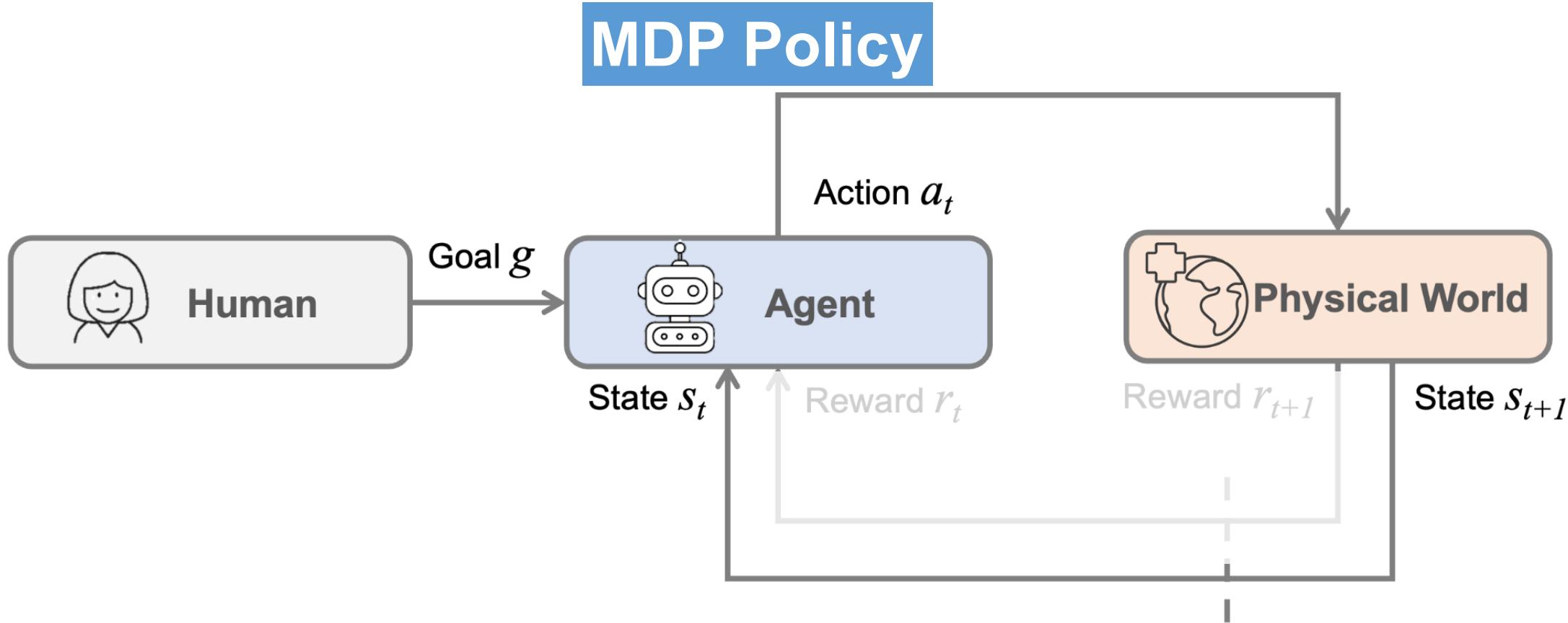
Question:
Tom grabbed pillow. Tom walked to kitchen ... What is the order of rooms where pillow appears?

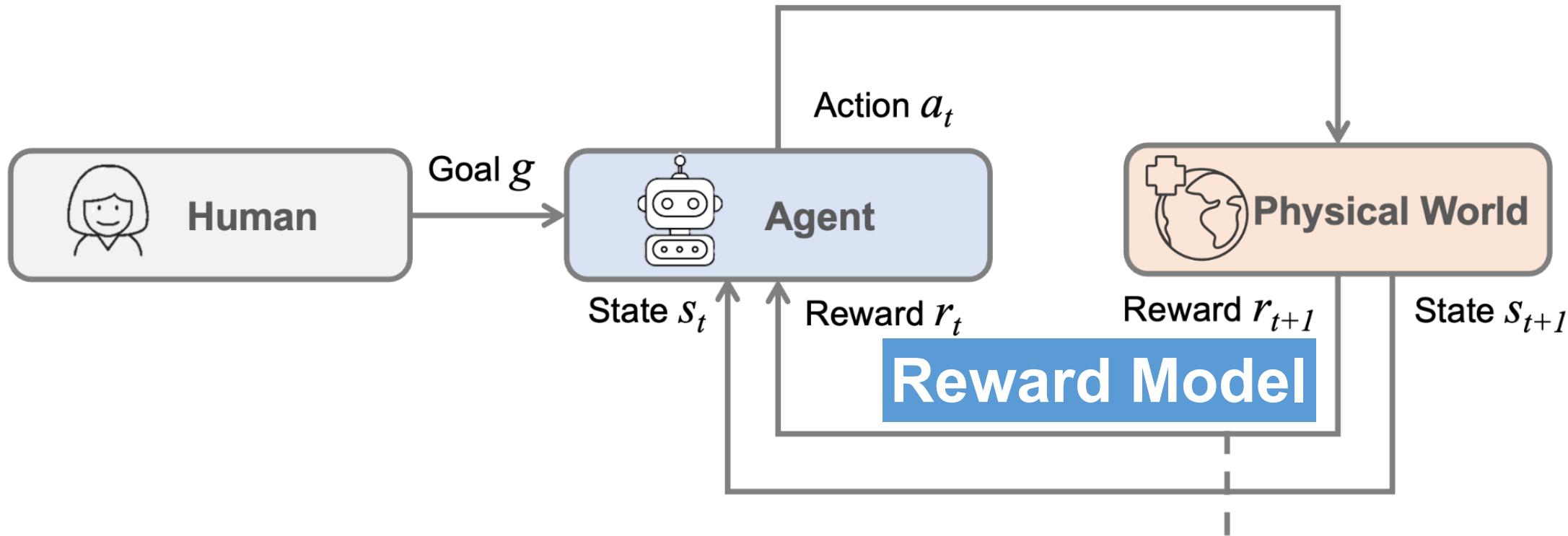
Answer:
Bedroom, kitchen, living room

Object Path Tracking

❑ Transferring DPO to Trajectory: Exploration-based Trajectory Optimization





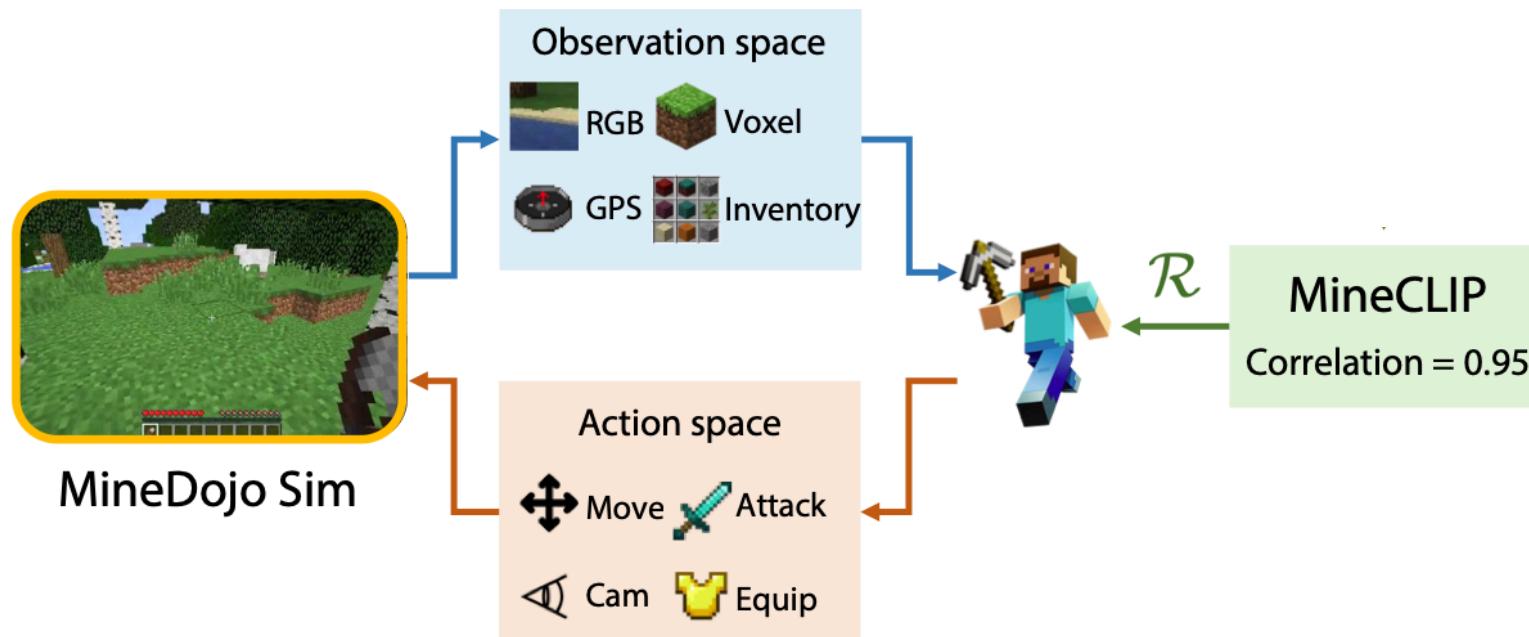


Reward Model

$$o, a \rightarrow r$$

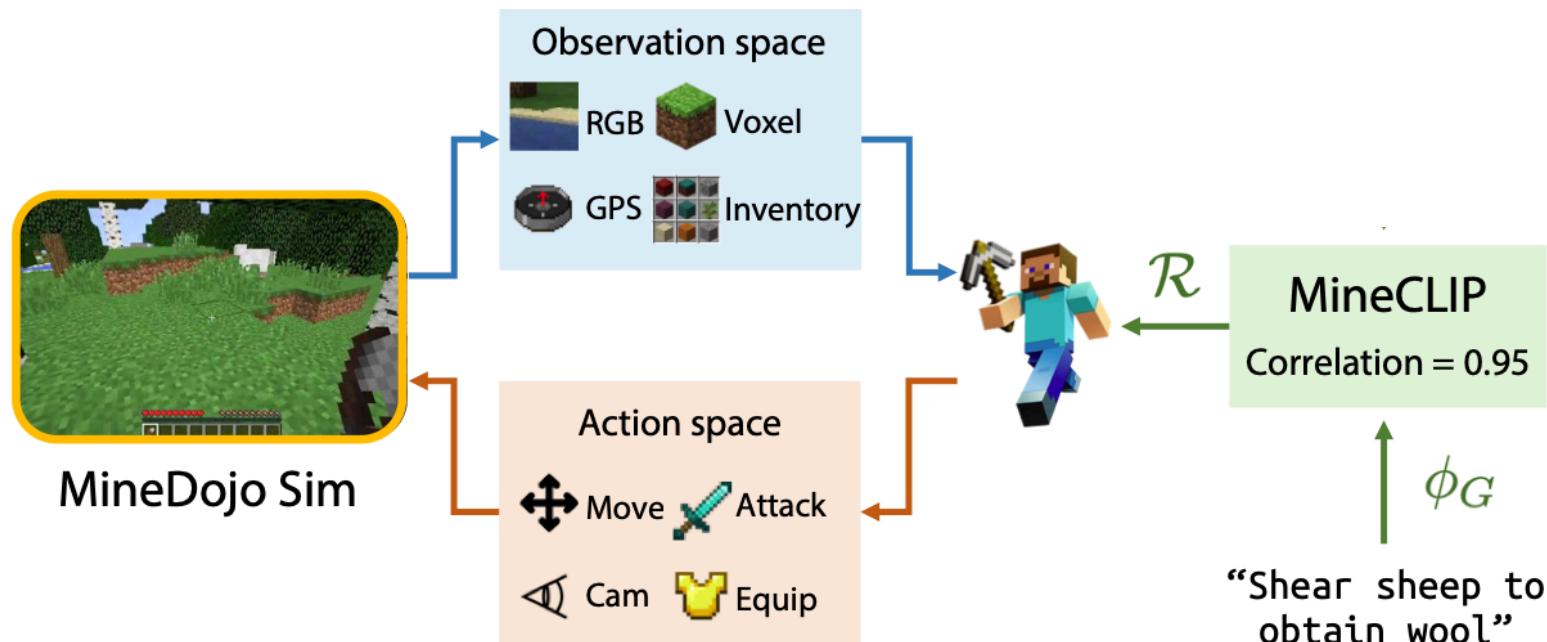
MineCLIP: Contrastive Video-Language Model

- The correlation score can be used as a learned dense reward function to train a strong multi-task RL agent

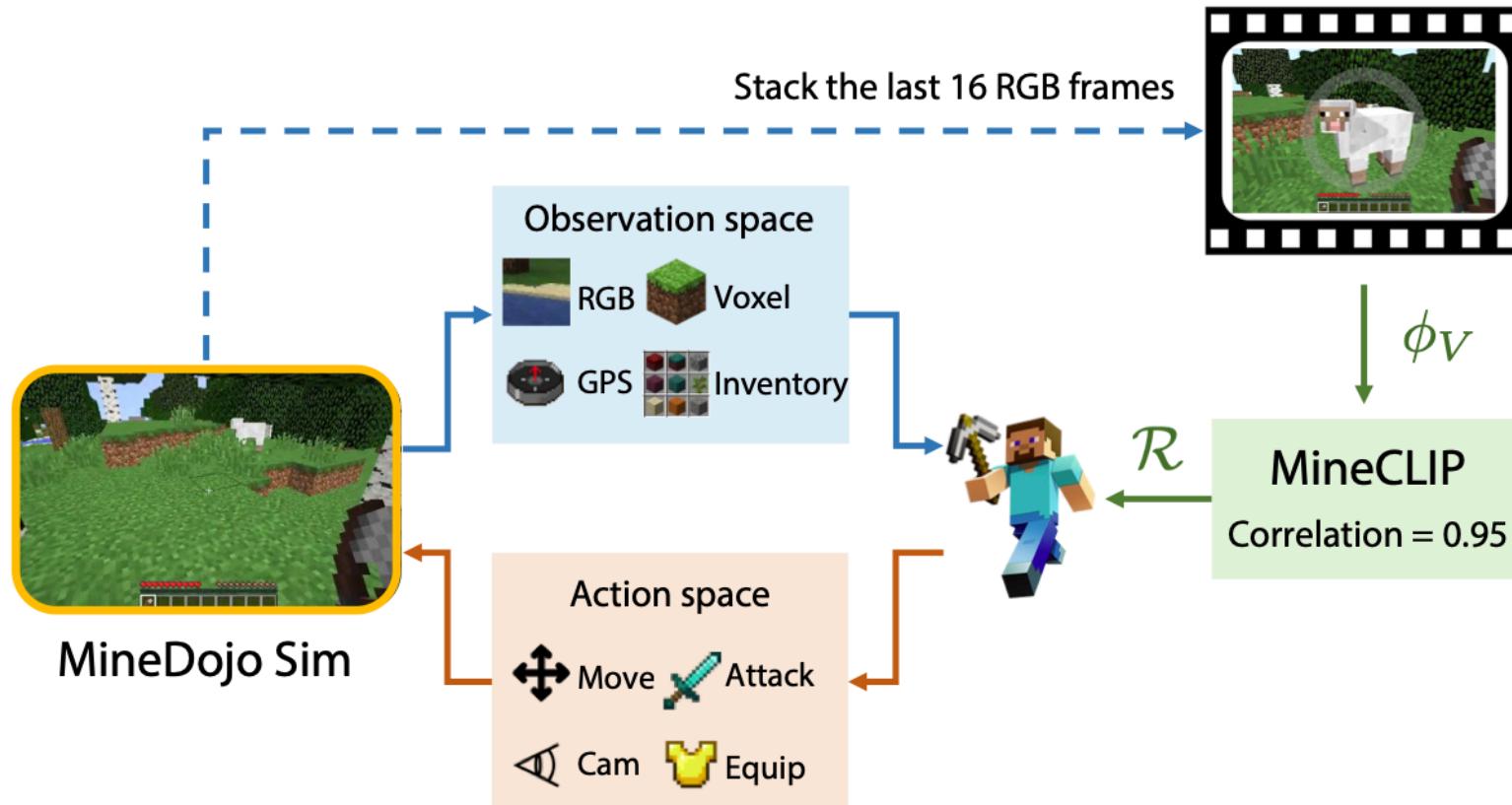


MineCLIP: Contrastive Video-Language Model

- The correlation score can be used as a learned dense reward function to train a strong multi-task RL agent

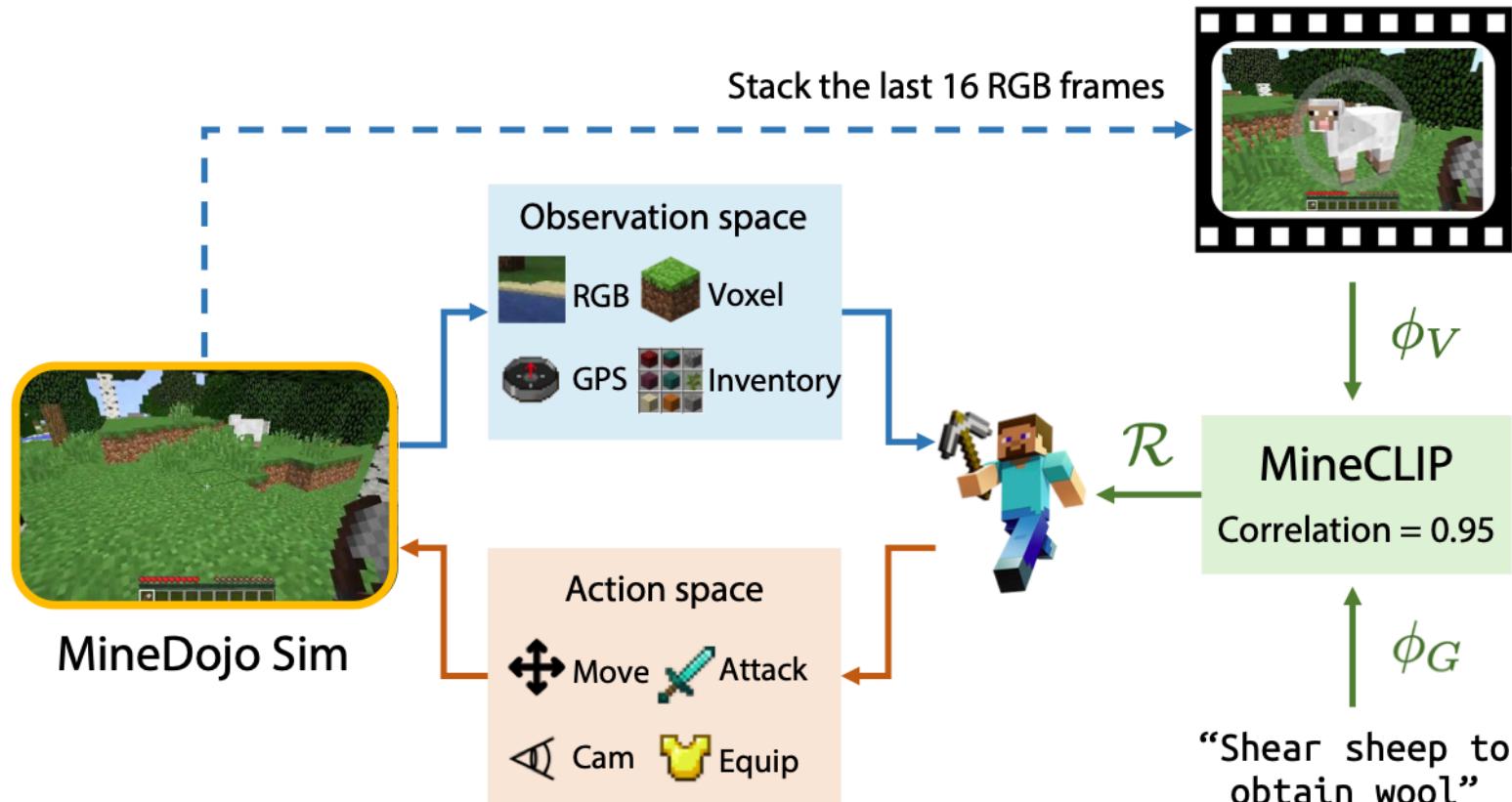


- The correlation score can be used as a learned dense reward function to train a strong multi-task RL agent



MineCLIP: Contrastive Video-Language Model

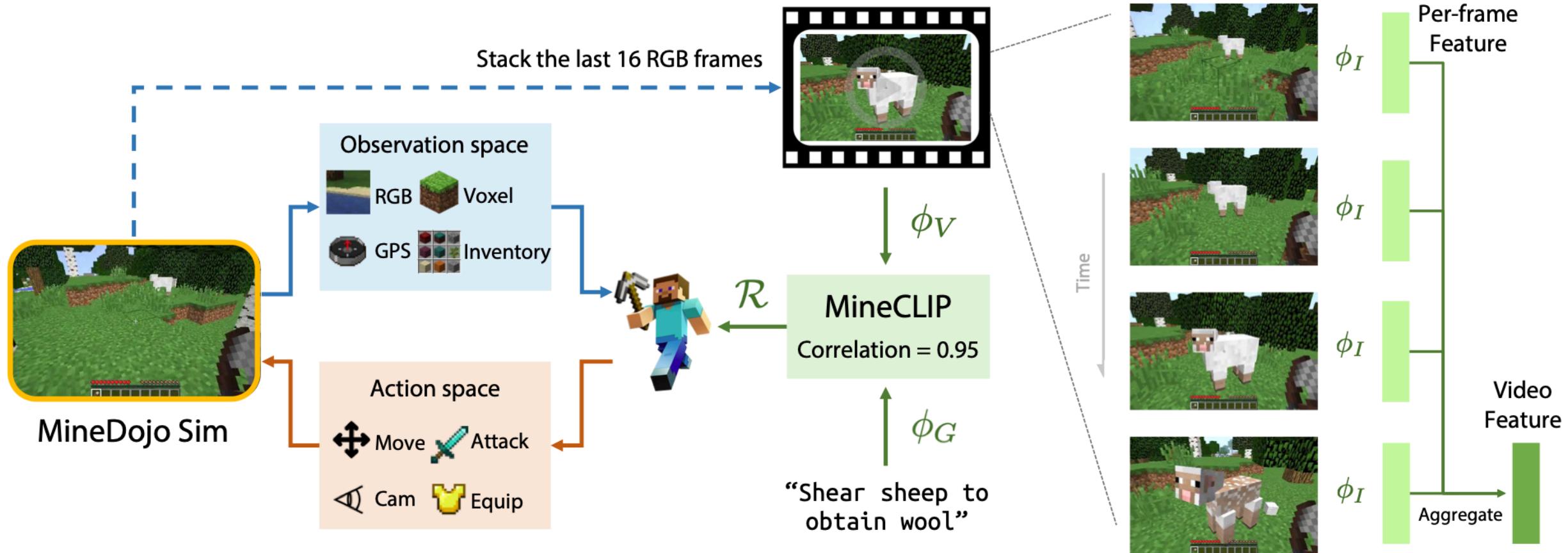
- The correlation score can be used as a learned dense reward function to train a strong multi-task RL agent

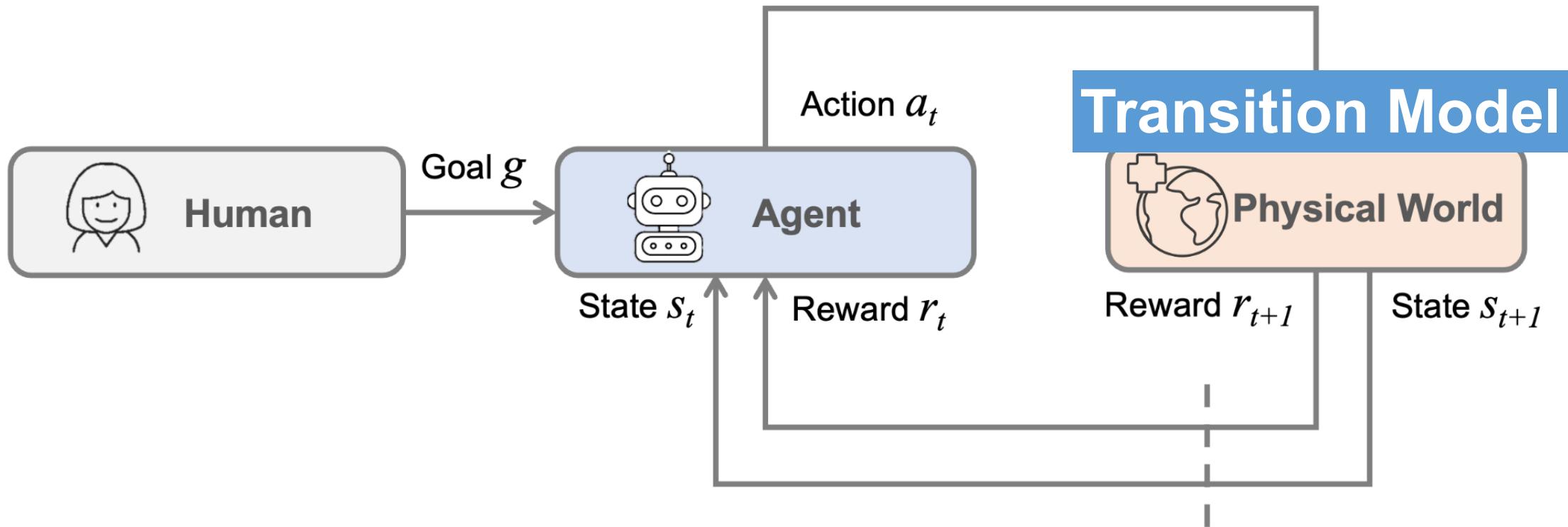


MineCLIP: Contrastive Video-Language Model



- The correlation score can be used as a learned dense reward function to train a strong multi-task RL agent





Transition Model

$$o_t, a \rightarrow o_{t+1}$$

"World Modeling"

Different Instantiation of World Models



- PDDL as world models
- LLMs as world models
- Video generation as world models "Large World Model"

PDDL as World Models

Input: Washing Dishes

Environment



Goal



Operator Name

stained(bowl)

not stained(bowl)

CLEAN_WITH_BRUSH

Transition Modeling**Preconditions**

stained(bowl)

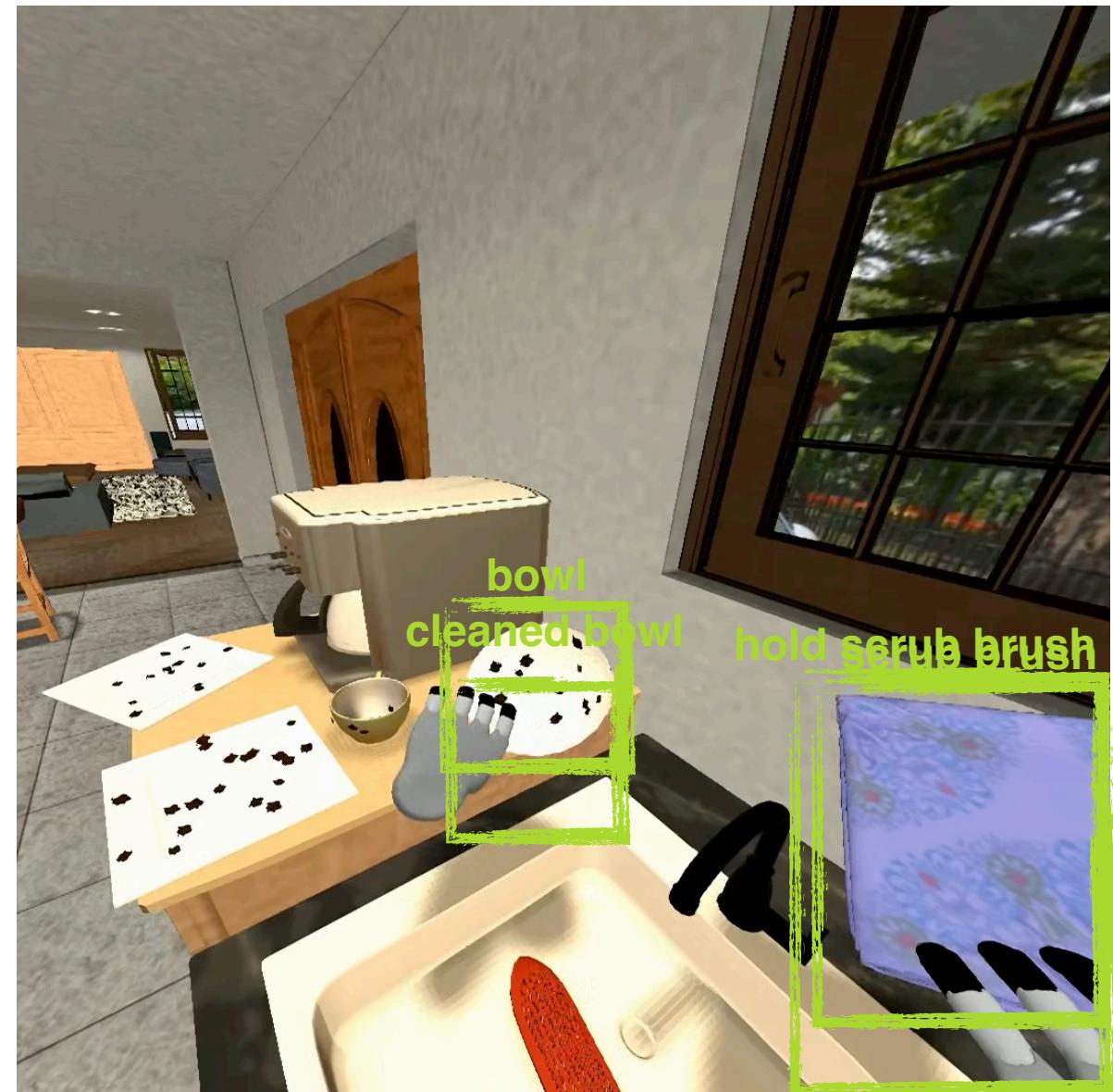
soaked(scrub brush)

holding(scrub brush)

Effects

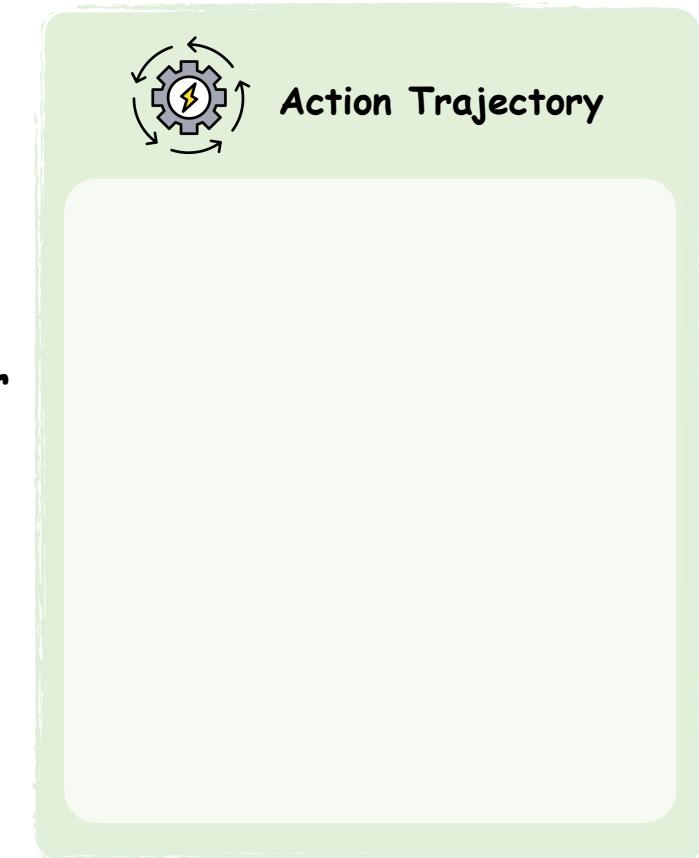
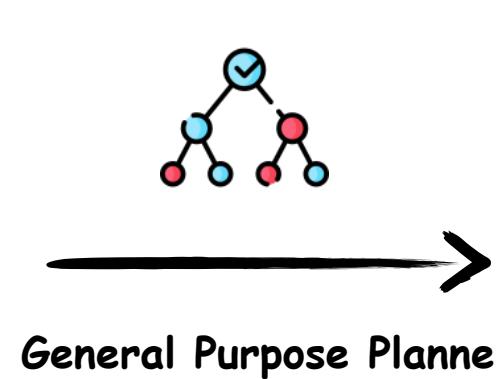
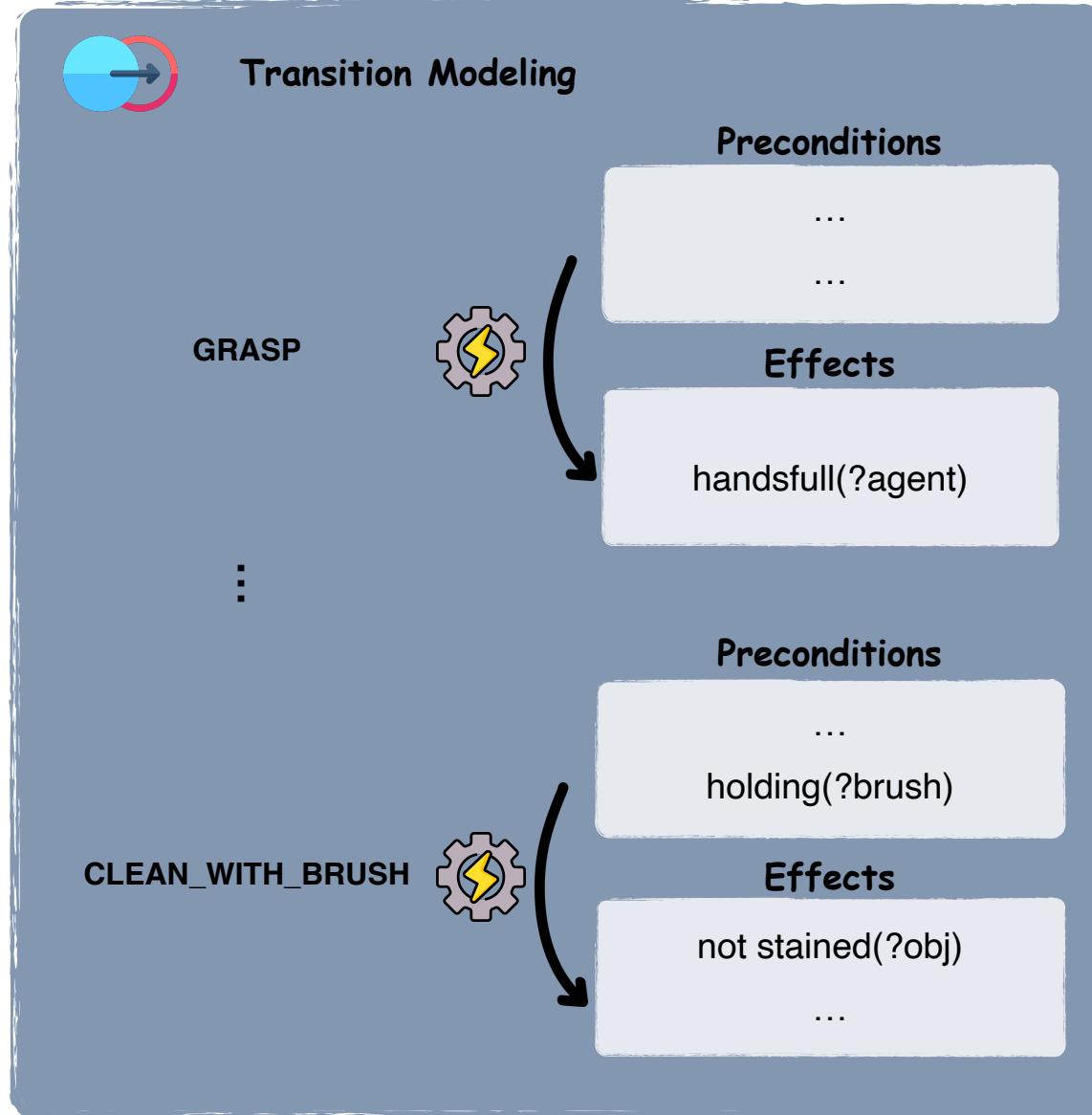
not stained(bowl)

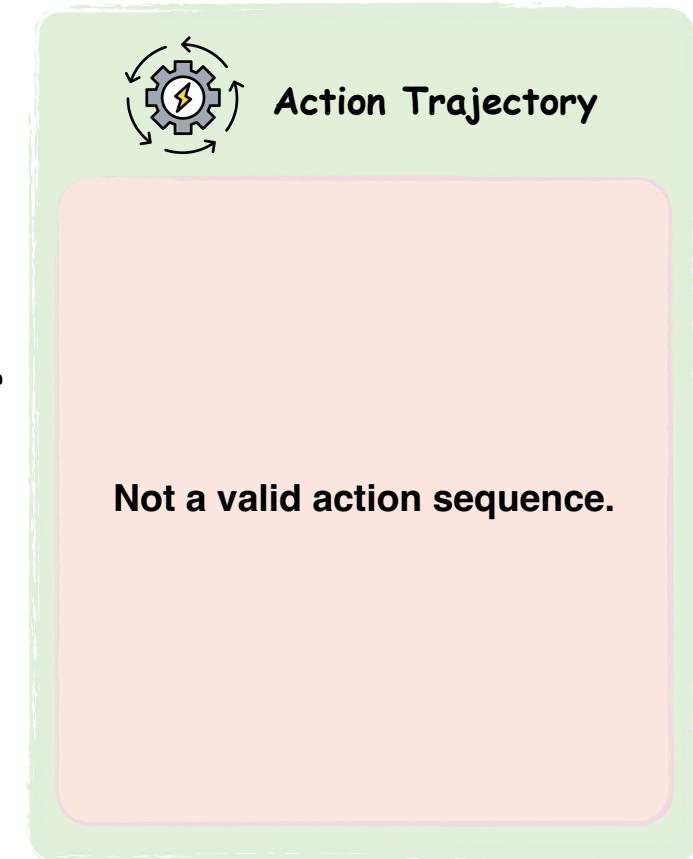
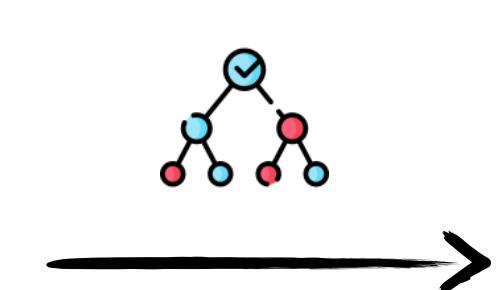
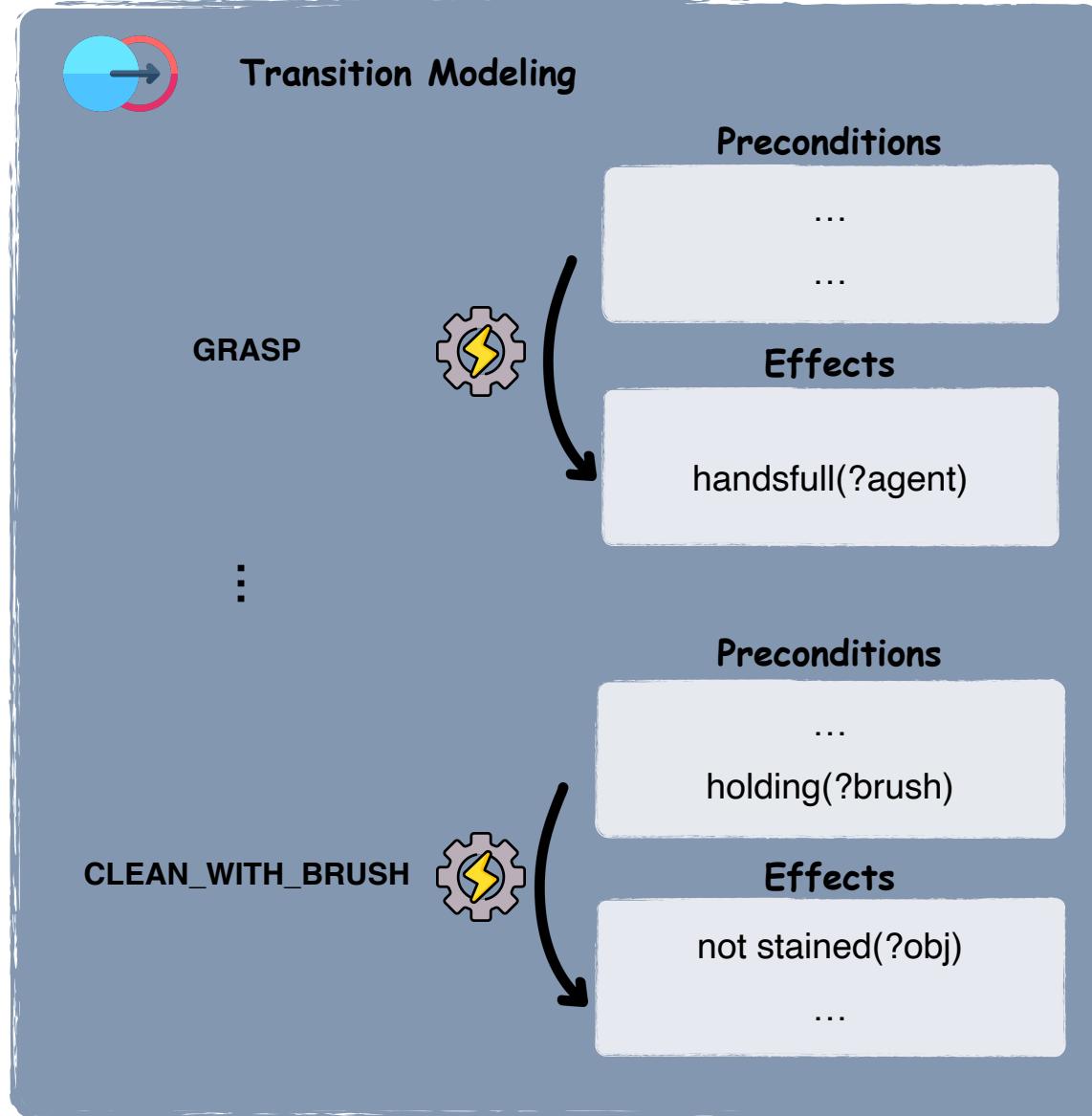
transitioning

LLM Output

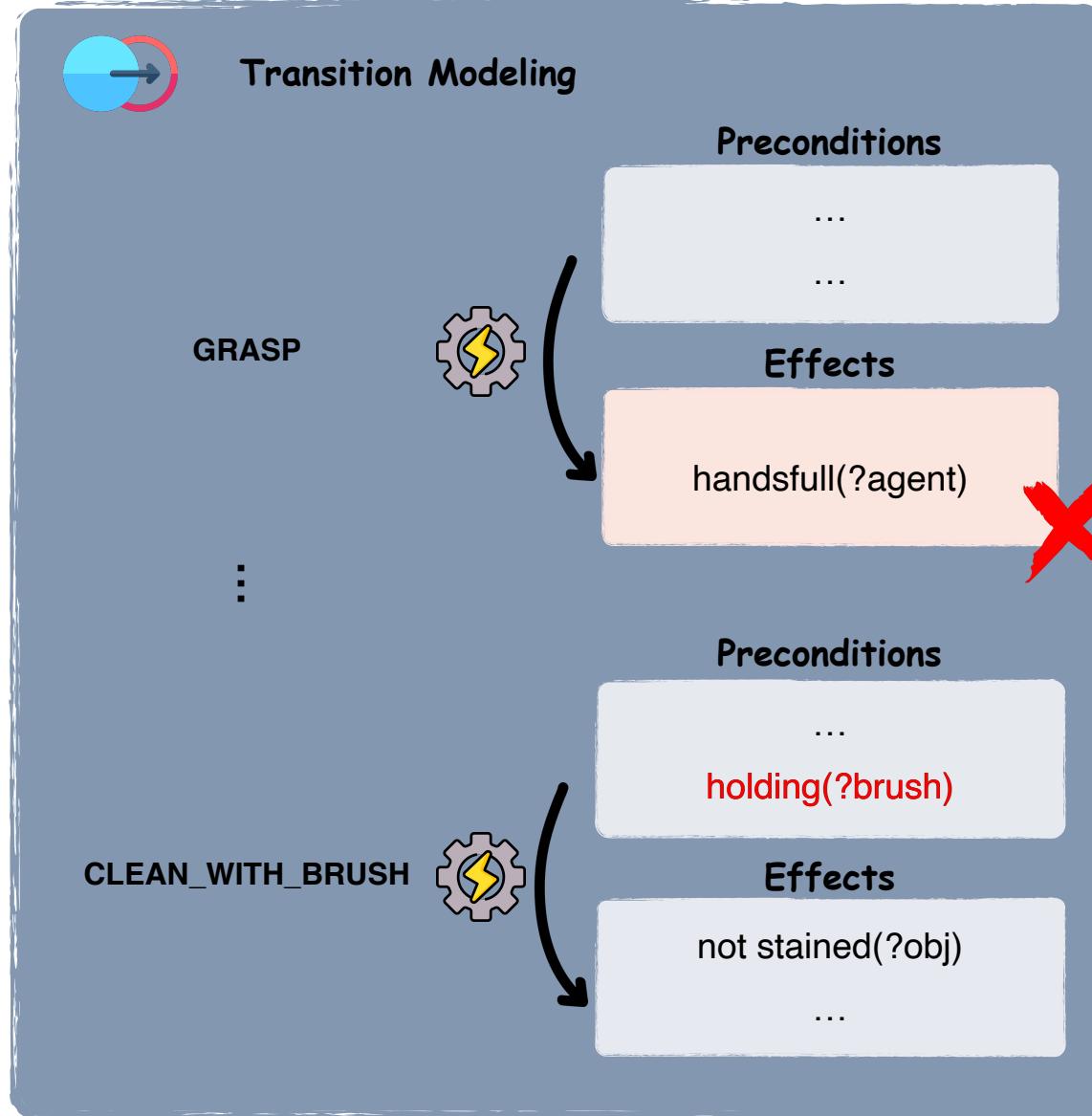
This video is for demonstration only. There're no actual controller-level actions. For action execution examples, visit our repository: <https://github.com/embodied-agent-interface/embodied-agent-interface>.

LLM Output: Preconditions and Effects

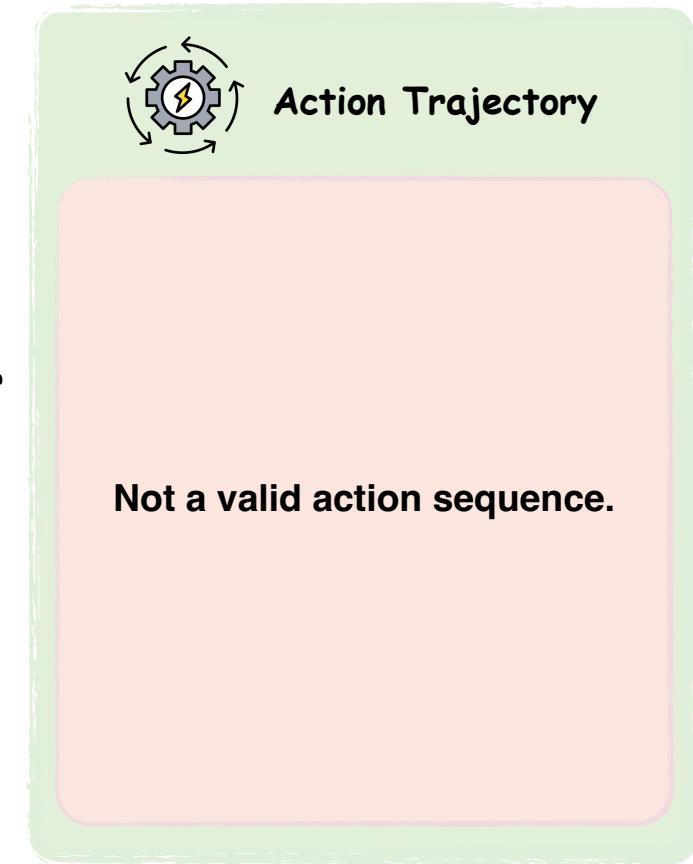


LLM Output: Preconditions and Effects

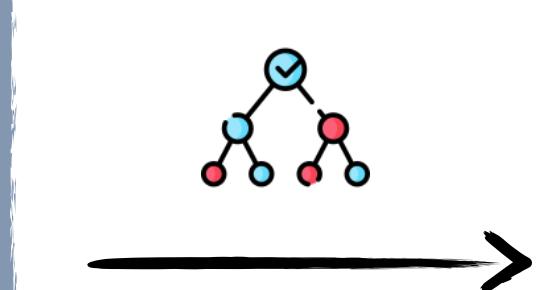
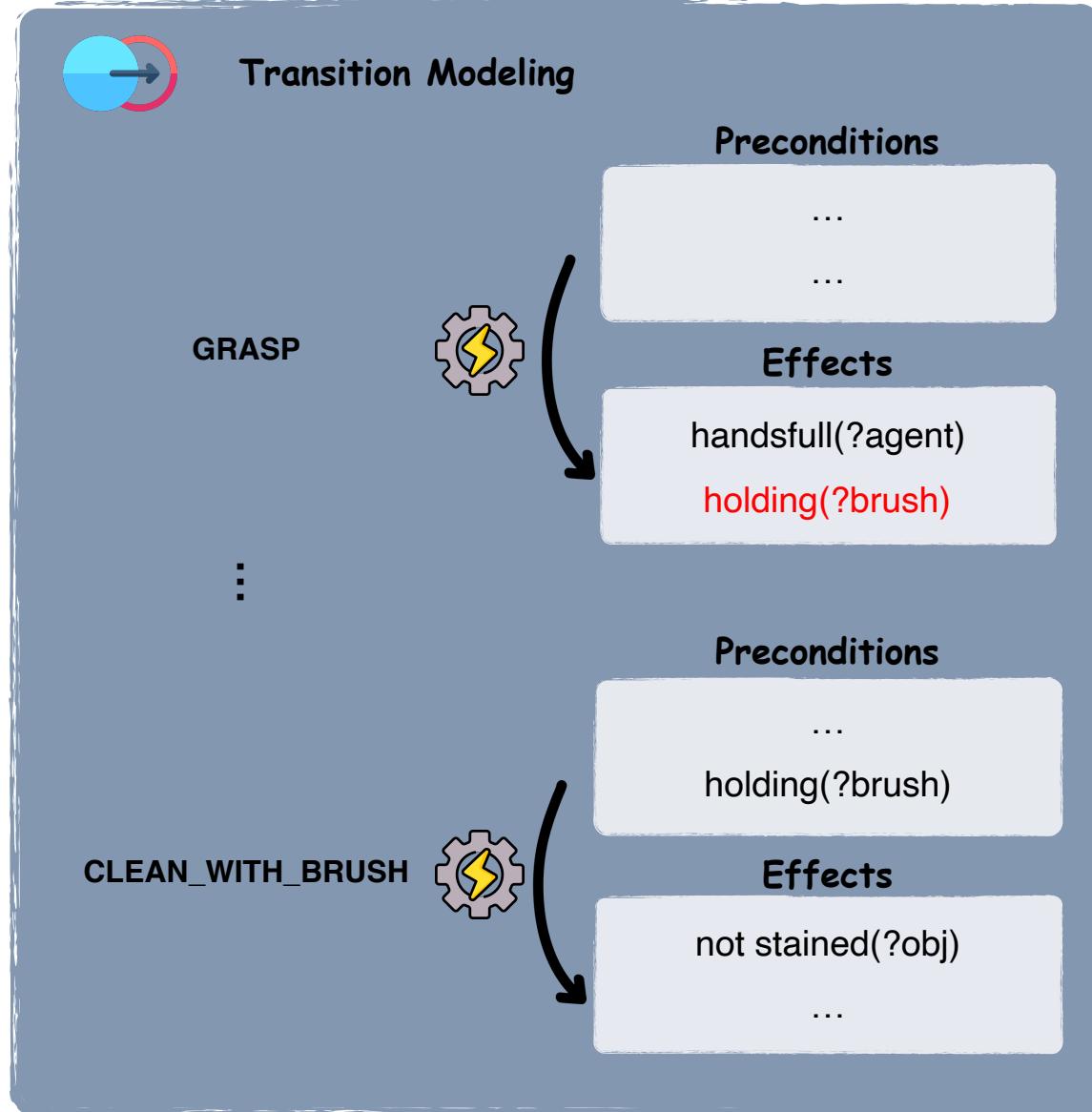
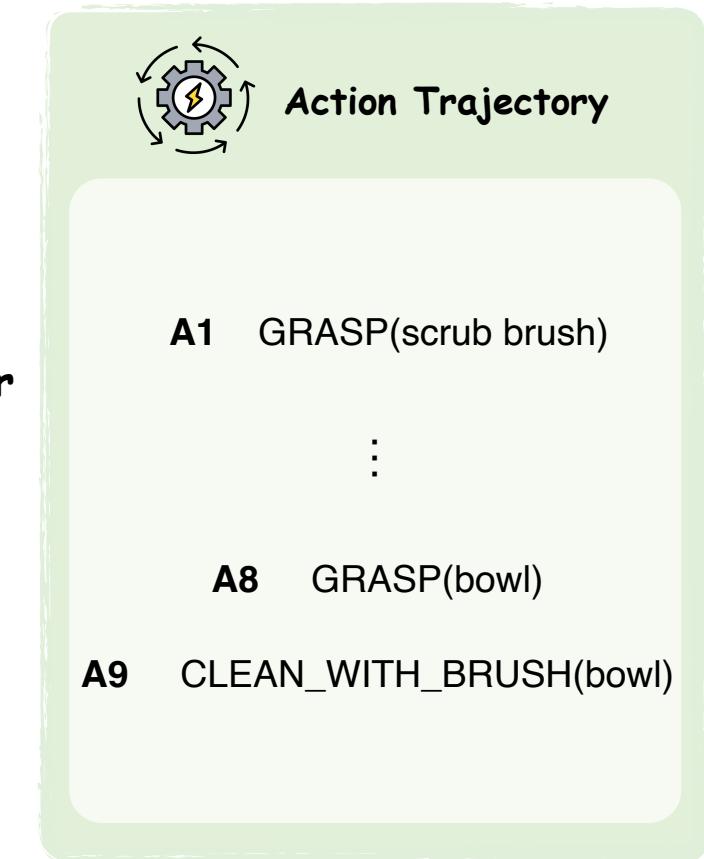
PLAN FAILED!

LLM Output: Preconditions and Effects

General Purpose Planner



PLAN FAILED!

LLM Output: Preconditions and Effects**General Purpose Planner****PLAN SUCCEEDED!**

```
...
2. ?o - householdObject: the small appliance to be toggled on
```

...

Preconditions:

```
(and
```

...

```
  (not (appliance-on ?o))
```

```
)
```

Effects:

```
(and
```

```
  (appliance-on ?o)
```

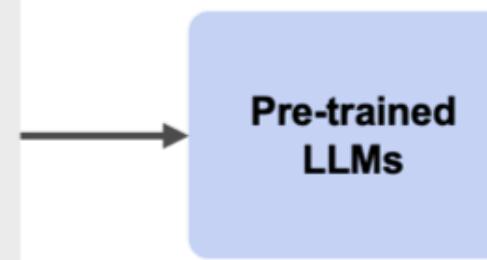
```
)
```

New Predicates:

```
1. (appliance-on ?o - householdObject): true if the small appliance ?o is switched on
```

Step 1. PDDL Construction

- Natural-language description of available actions or skills
- Description of the domain



Validation Tool for PDDL: brief but readable error messages to capture **syntax errors**

PDDL

Parameters:
...
Preconditions:
...
Effects:
...



Step 2. Human Correction

PDDL

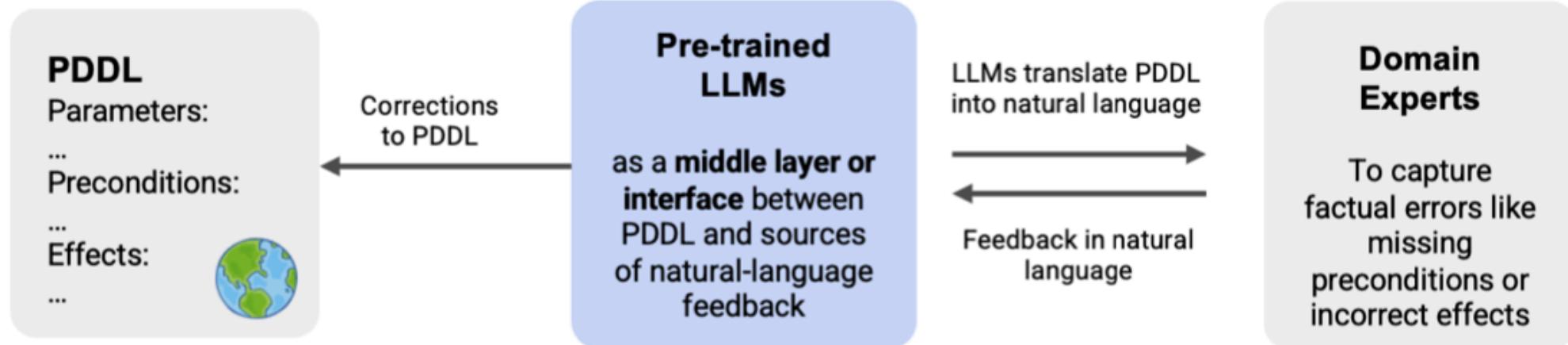
Parameters:
...

Preconditions:
...

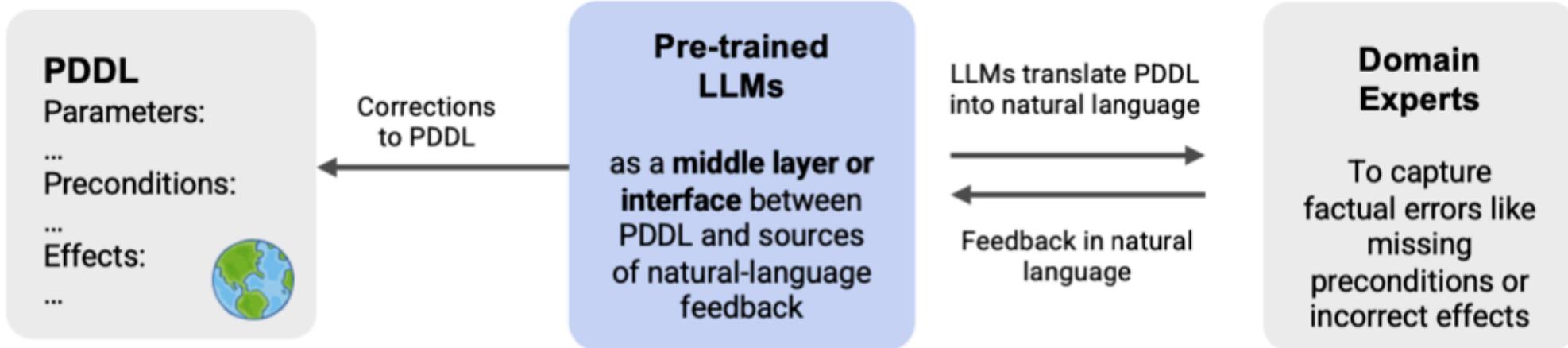
Effects:
...



Step 2. Human Correction

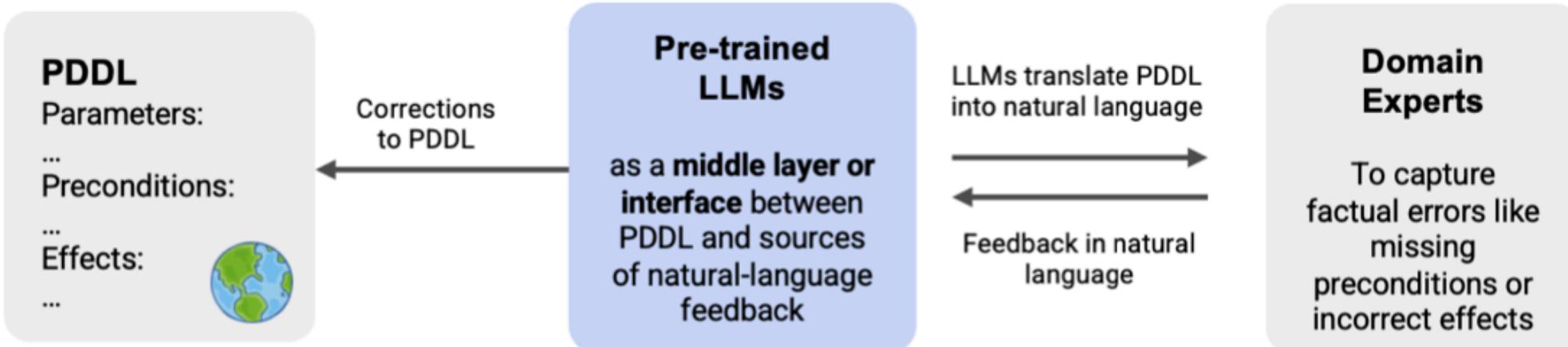


Step 2. Human Correction

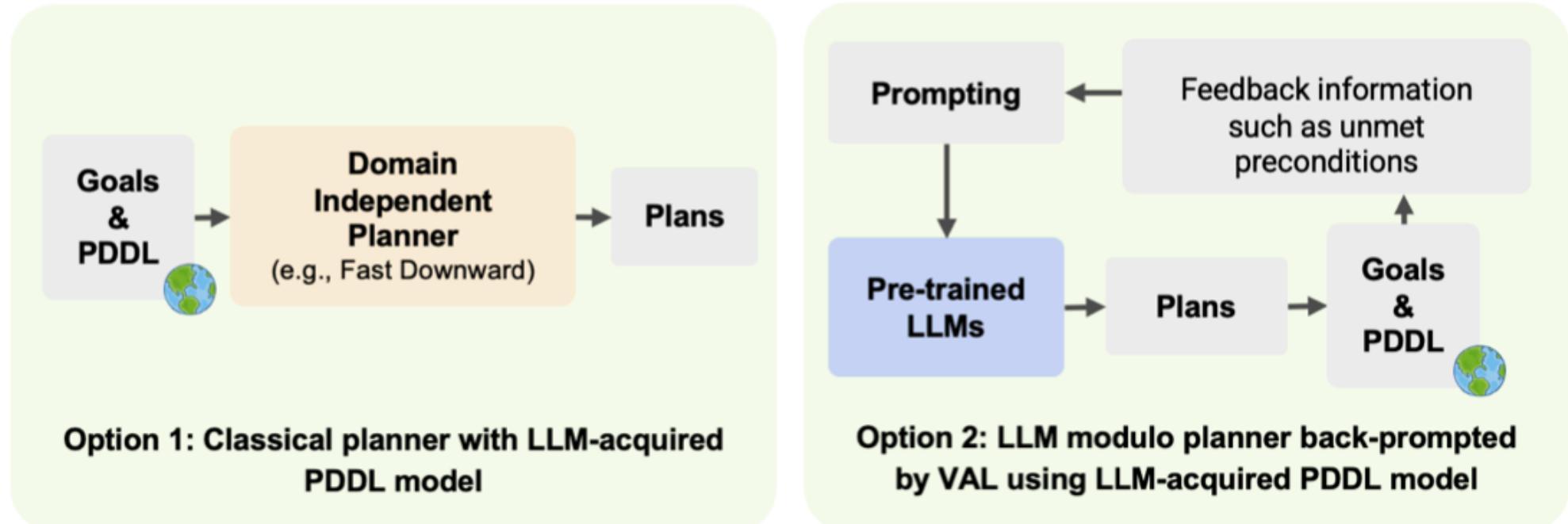


Step 3. Planning with two hybrid approaches

Step 2.
Human
Correction

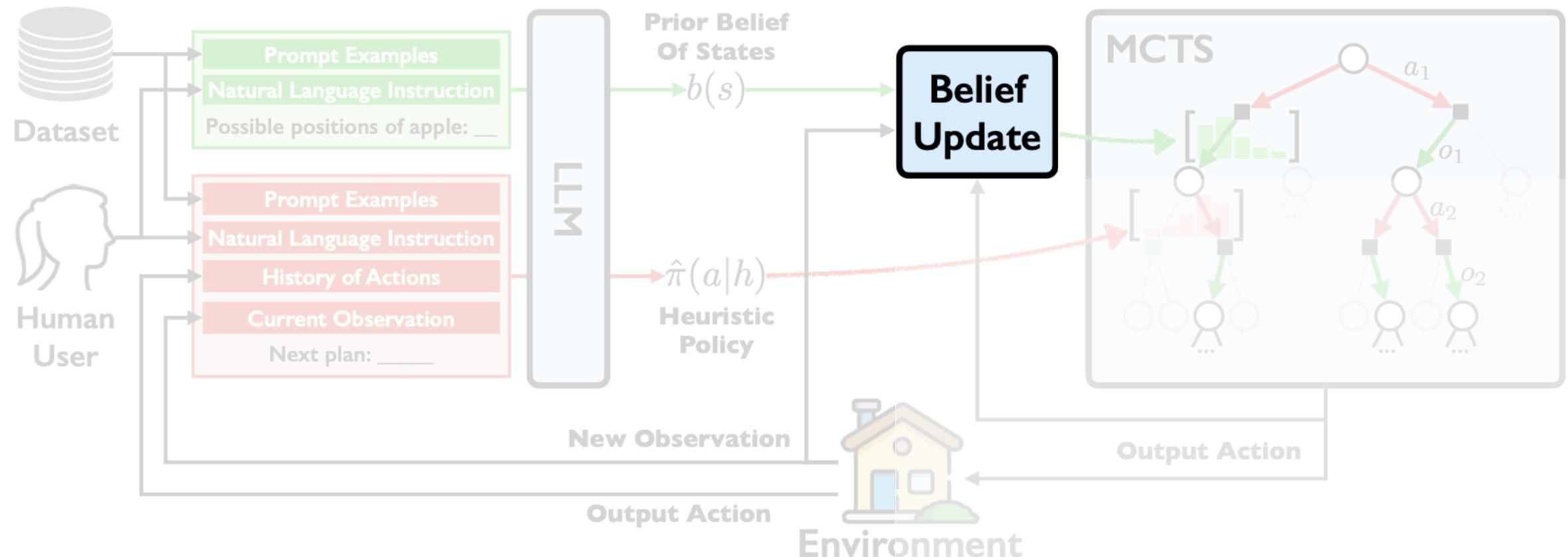


Step 3.
Planning with two hybrid approaches



LLMs as World Models

- Sample from the commonsense belief to obtain an initial state of the world



Multimodal Theory of Mind

Belief

Belief Inference

Type 1.1: True belief, short-term



Scene: ... Inside the bridge, you'll find a bottle of wine...

Actions: ... Finally, she moves towards the fridge, preparing to open it.

Question: If Elizabeth has been trying to get a bottle of wine, which one of the following statements is more likely to be true?

- (a) *Elizabeth thinks that there is a bottle of wine inside the fridge.*
- (b) Elizabeth thinks that there isn't any bottle of wine inside the fridge.

Type 2.1: Goal given true belief



Scene: ... The living room is furnished with a cabinet, ... The cabinet is filled with two apples, ..., and a bottle of wine. ... Inside the fridge, there are two apples.

Actions: James... then opens the fridge, closes it... Finally, he walks towards the living room and approaches the cabinet.

Question: Which one of the following statements is more likely to be true?

- (a) *James has been trying to get a bottle of wine.*
- (b) James has been trying to get an apple.

Goal Inference

Type 1.2: False belief, short-term



Scene: ... The living room features a cabinet... The cabinet is filled with a bag of chips, a remote controller, a bottle of wine, and a water glass.

Actions: Jennifer is situated in the living room. She heads towards the cabinet and is about to open it.

Question: If Jennifer has been trying to get a cupcake, which one of the following statements is more likely to be true?

- (a) Jennifer thinks that there isn't a cupcake inside the cabinet.
- (b) *Jennifer thinks that there is a cupcake inside the cabinet.*

Type 2.2: Goal given false belief



Scene: ... There is a water glass inside the seventh cabinet... The fridge stores two cupcakes...

Actions: Mark... advances towards the seventh kitchen cabinet.

Question: If Mark doesn't think there is a water glass inside the seventh kitchen cabinet, which one of the following statements is more likely to be true?

(a) Mark has been trying to get a water glass.

(b) *Mark has been trying to get a cupcake.*

Type 1.3: Belief tracking, long-term



Scene: ... The kitchen is equipped with a fridge, sofa, dishwasher, eight cabinets, a stove, a microwave, and a kitchen table...

Actions: ... He walks to the seventh kitchen cabinet, opens and closes it. He repeats the same action with the sixth kitchen cabinet. Subsequently, he moves towards the dishwasher.

Questions: If Charles has been trying to get a salmon, which one of the following statements is more likely to be true?

- (a) Charles thinks that there is a salmon inside the fridge.
- (b) *Charles thinks that there isn't any salmon inside the fridge.*

Type 2.3: Goal given updated belief



Scene: ... The first cabinet, from left to right, contains a bag of chips.

Actions: Mary... walks towards the first kitchen cabinet, opens it, and then closes it.

Question: Which one of the following statements is more likely to be true?

(a) Mary has been trying to get a bag of chips.

(b) *Mark has been trying to get a condiment bottle.*

Type 2.4: Goal given future actions



Scene: ... The dishwasher holds a dish bowl... The first cabinet from the left holds a bag of chips and a wine glass... The fifth cabinet has an apple...

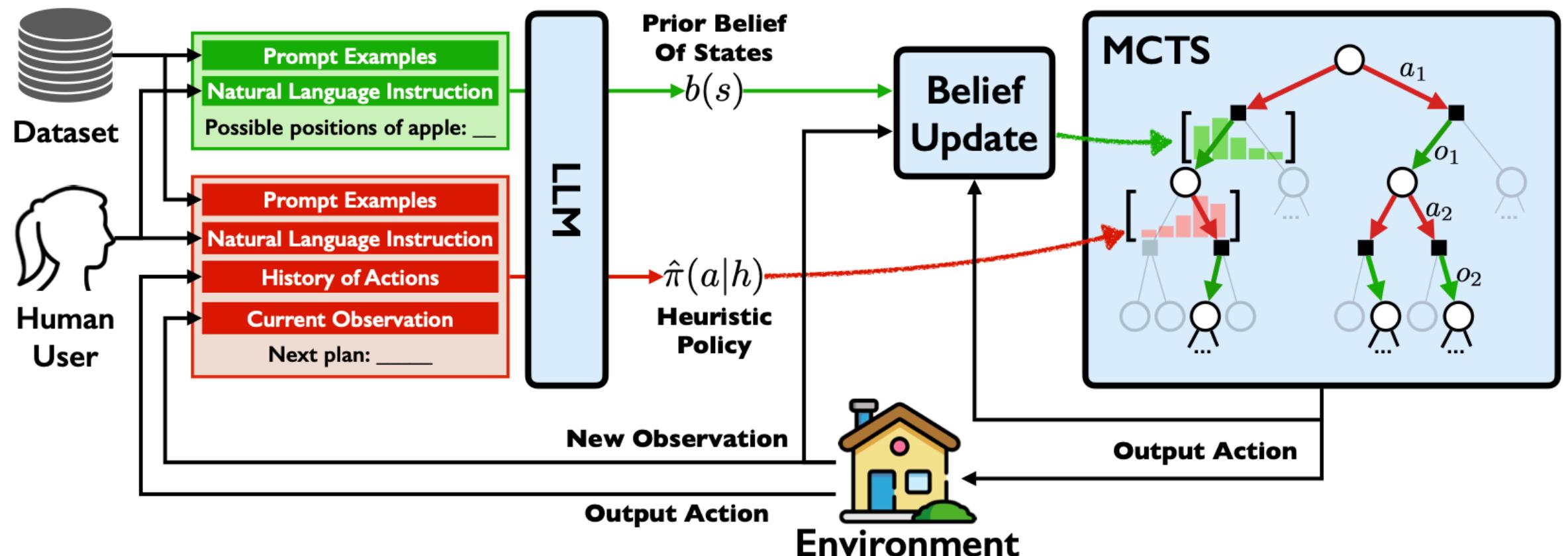
Actions: Williams... advances towards the first kitchen cabinet, opens it, and then shuts it. He then moves towards the fifth kitchen cabinet.

Question: Which one of the following statements is more likely to be true?

(a) William has been trying to get a wine glass.

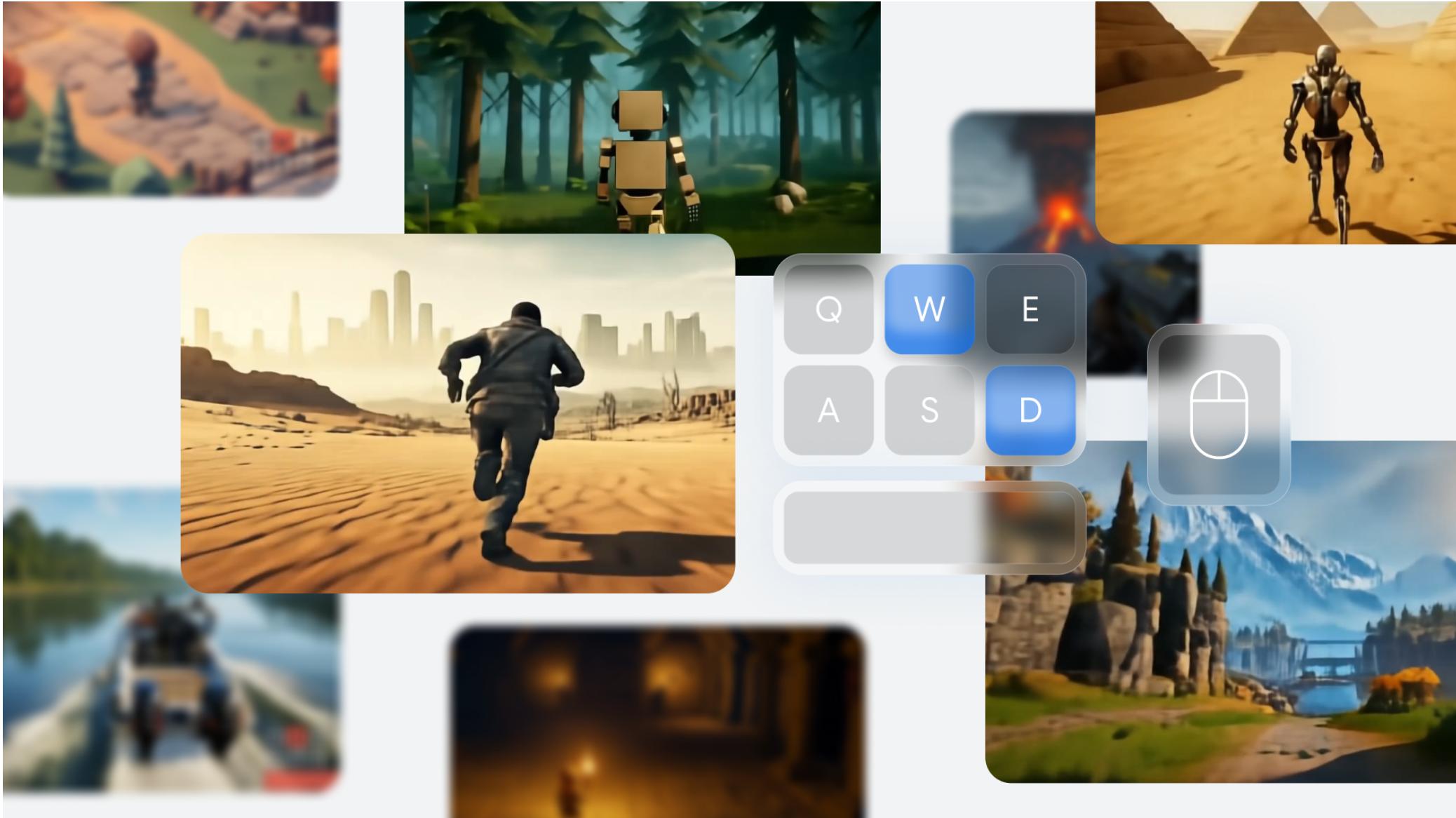
(b) *William has been trying to get a dish bowl.*

- Sample from the commonsense belief to obtain an initial state of the world



Video Generation as World Model (World Foundation Model)

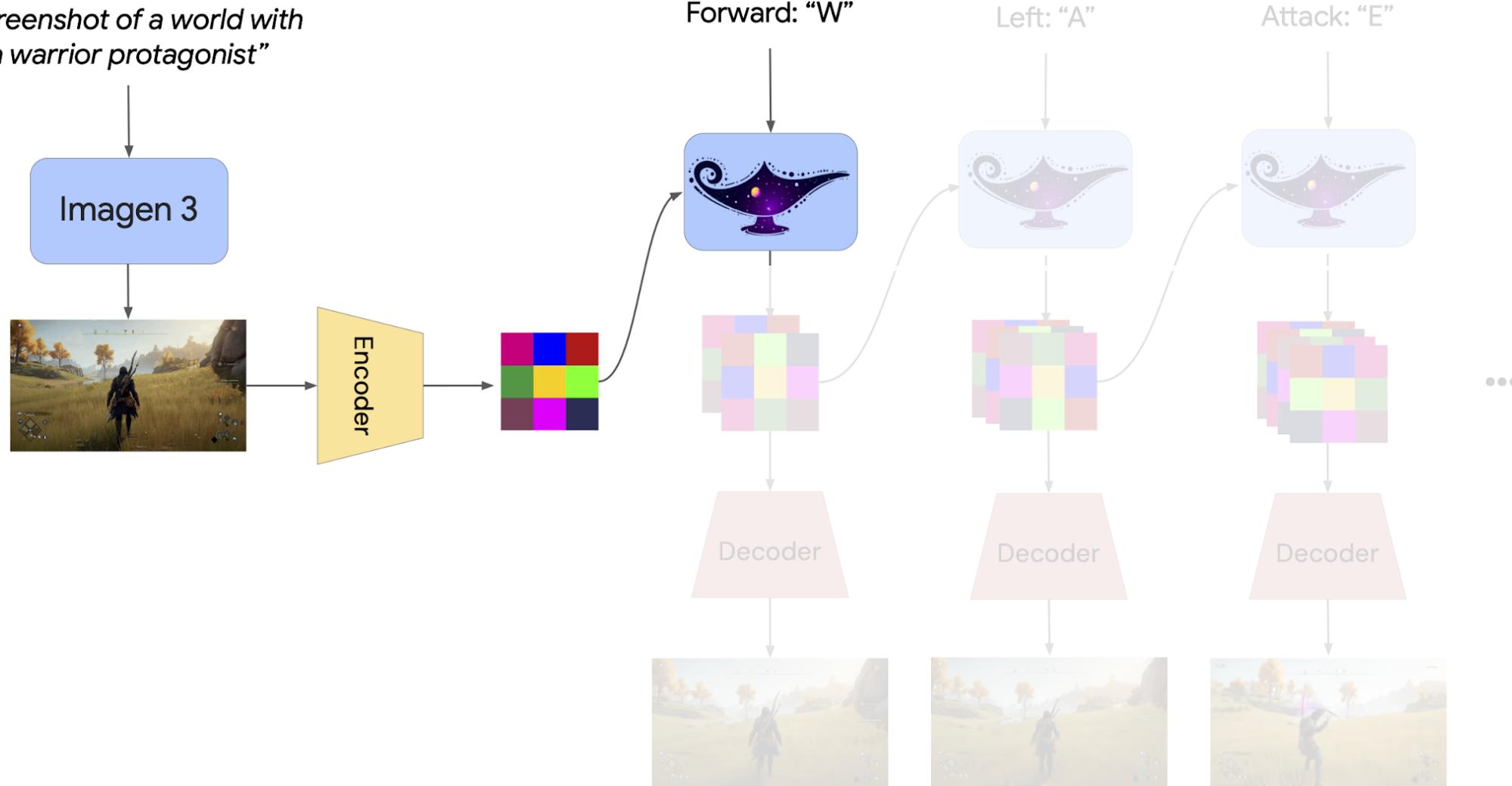
Genie 2: World Foundation Model



Genie 2: World Foundation Model

Diffusion world model

“Screenshot of a world with a warrior protagonist”

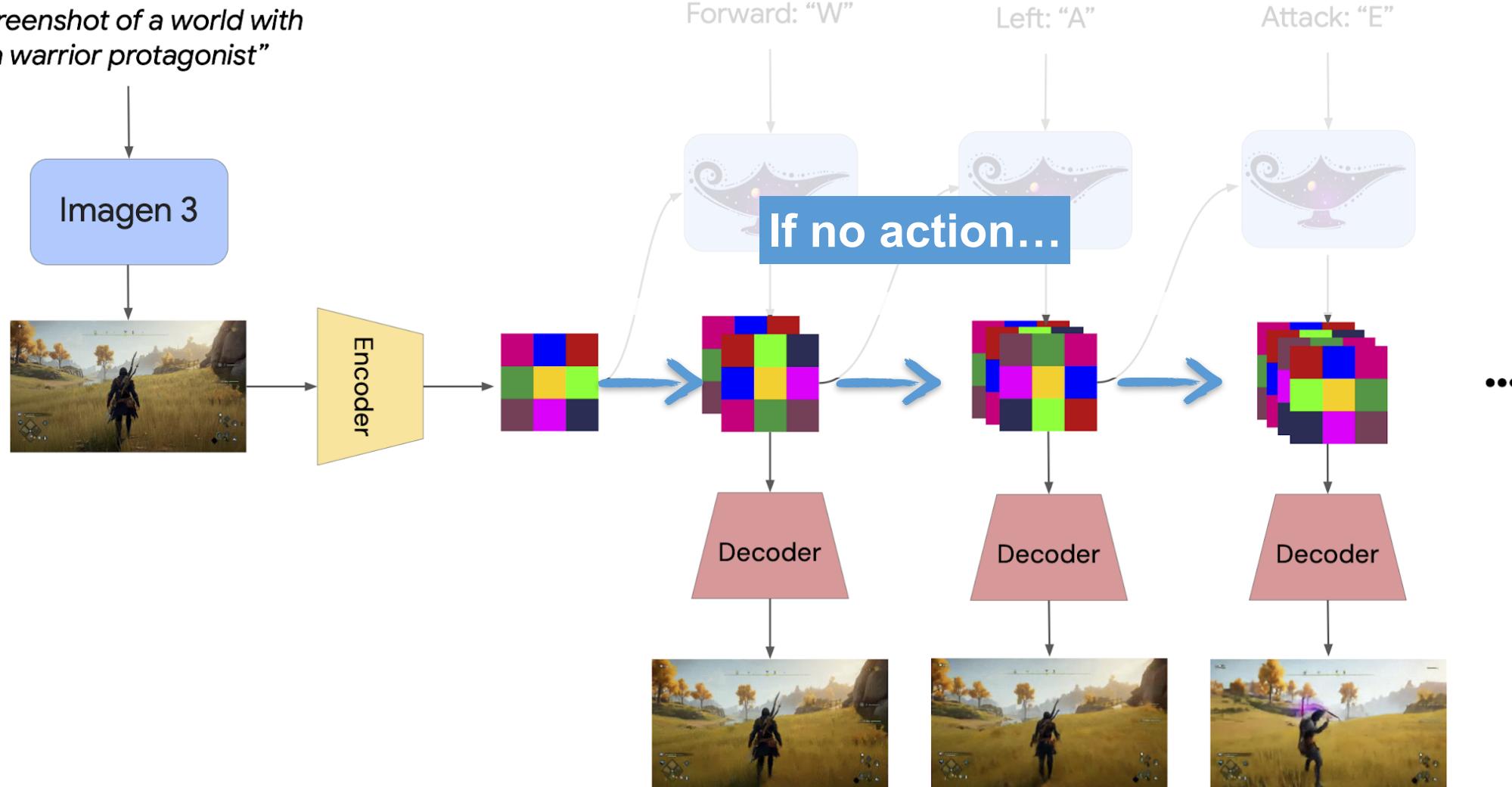


Genie 2: World Foundation Model



Diffusion world model

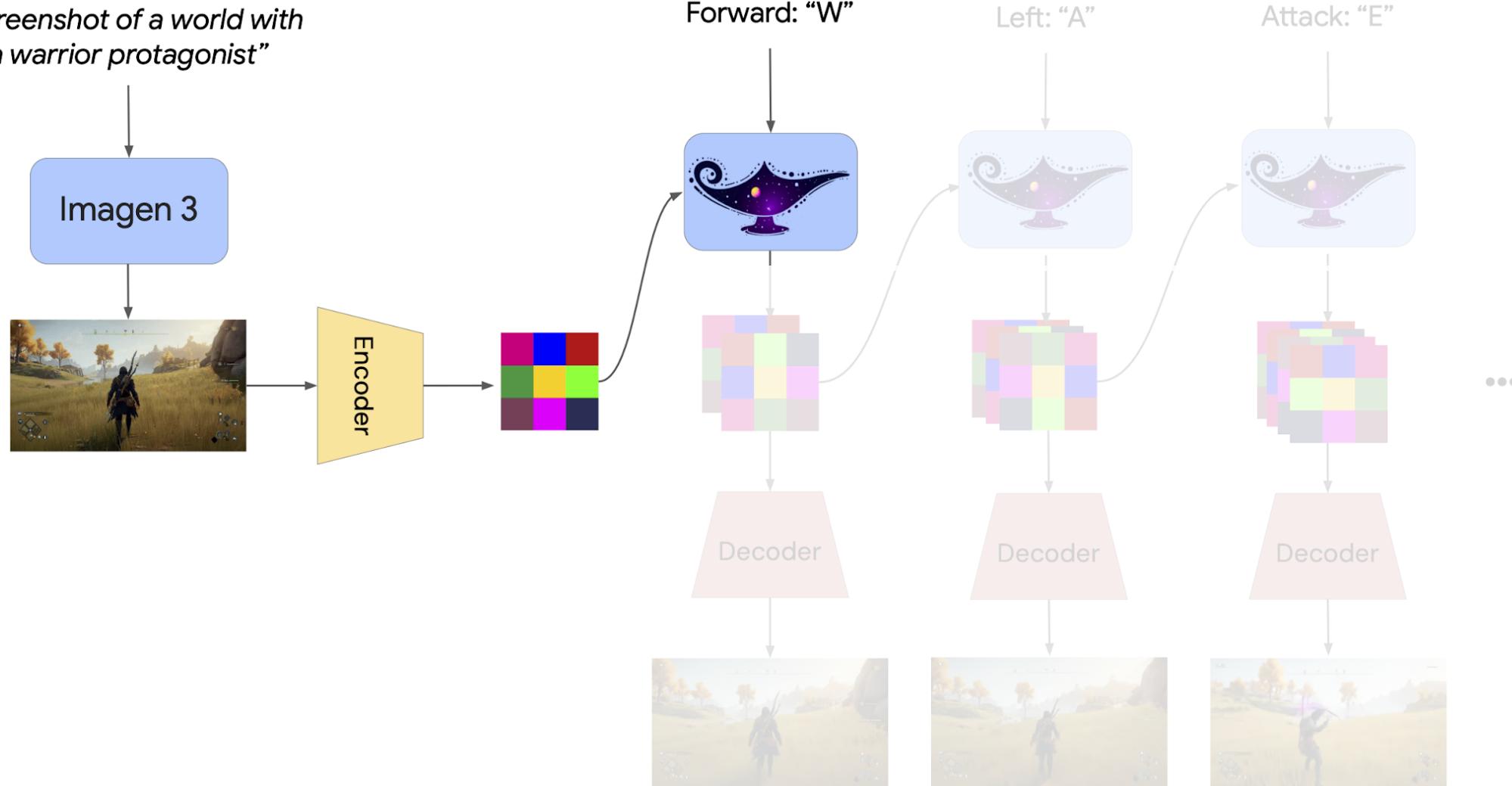
"Screenshot of a world with a warrior protagonist"



Genie 2: World Foundation Model

Diffusion world model

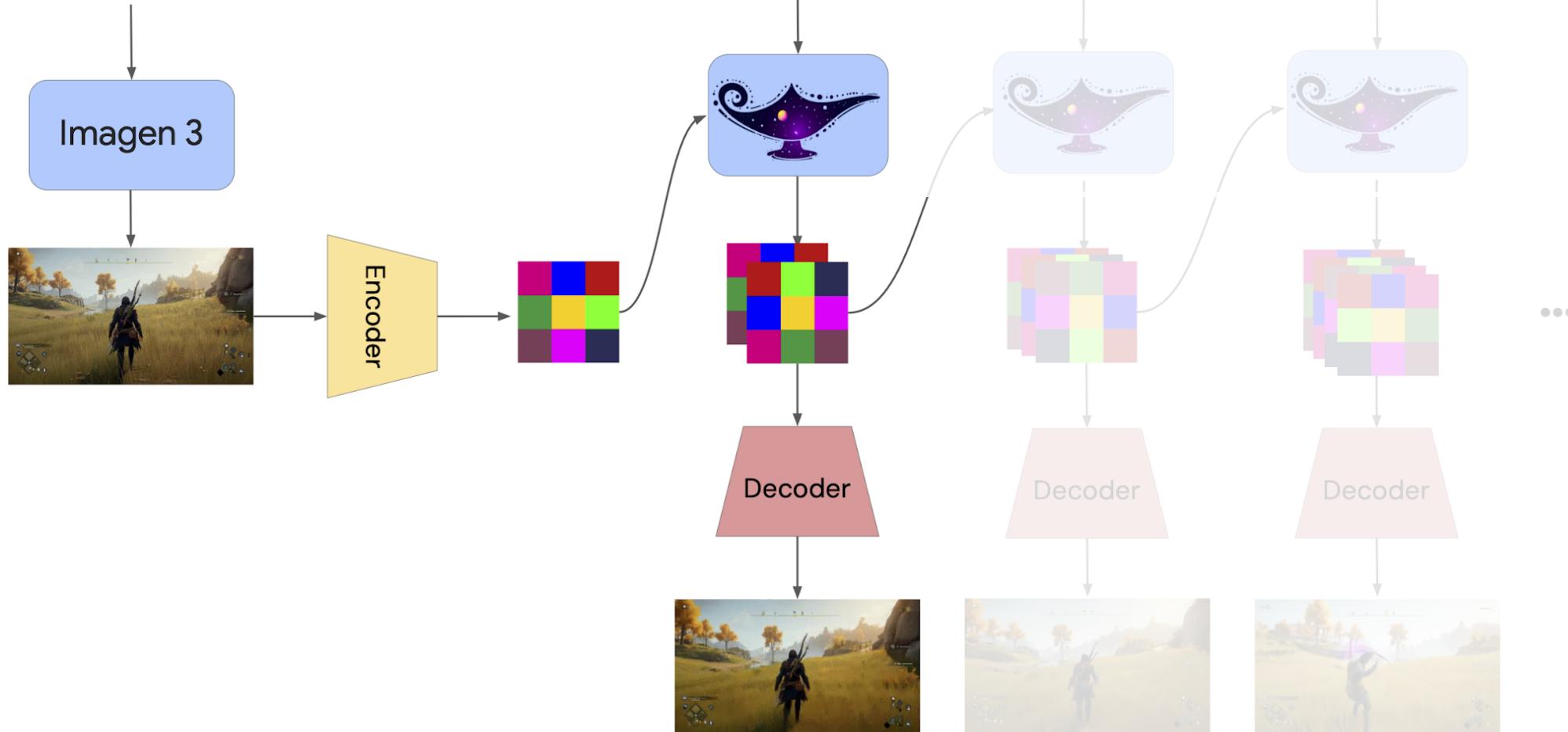
“Screenshot of a world with a warrior protagonist”



Genie 2: World Foundation Model

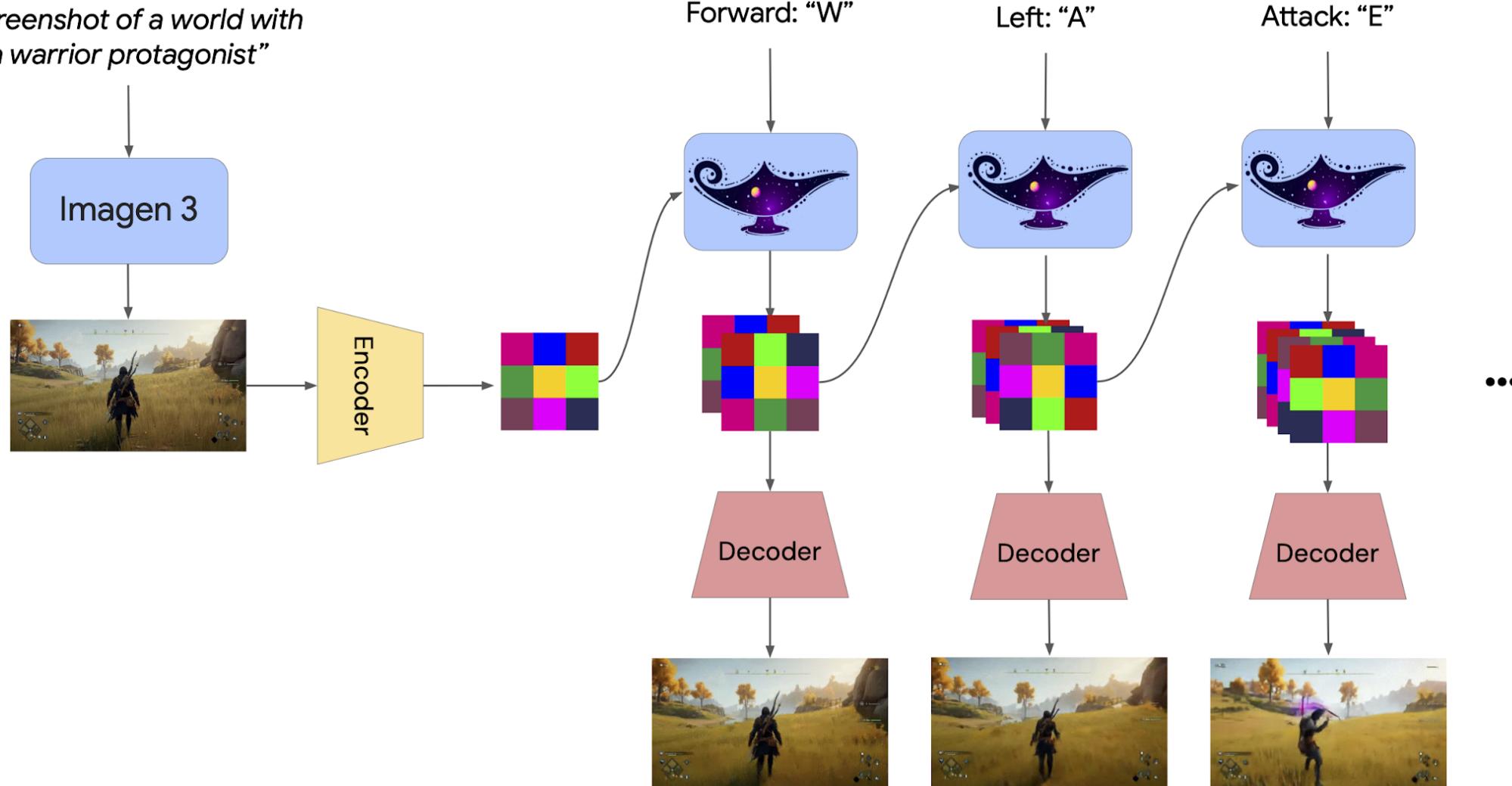
Diffusion world model

“Screenshot of a world with a warrior protagonist”



Diffusion world model

“Screenshot of a world with a warrior protagonist”



NVIDIA Cosmos World Foundation Models

A family of pre-trained models purpose-built for generating physics-aware videos and world states for physical AI development.

Learn more about model architectures, development resources, and availability [here](#).

Cosmos **Nano**

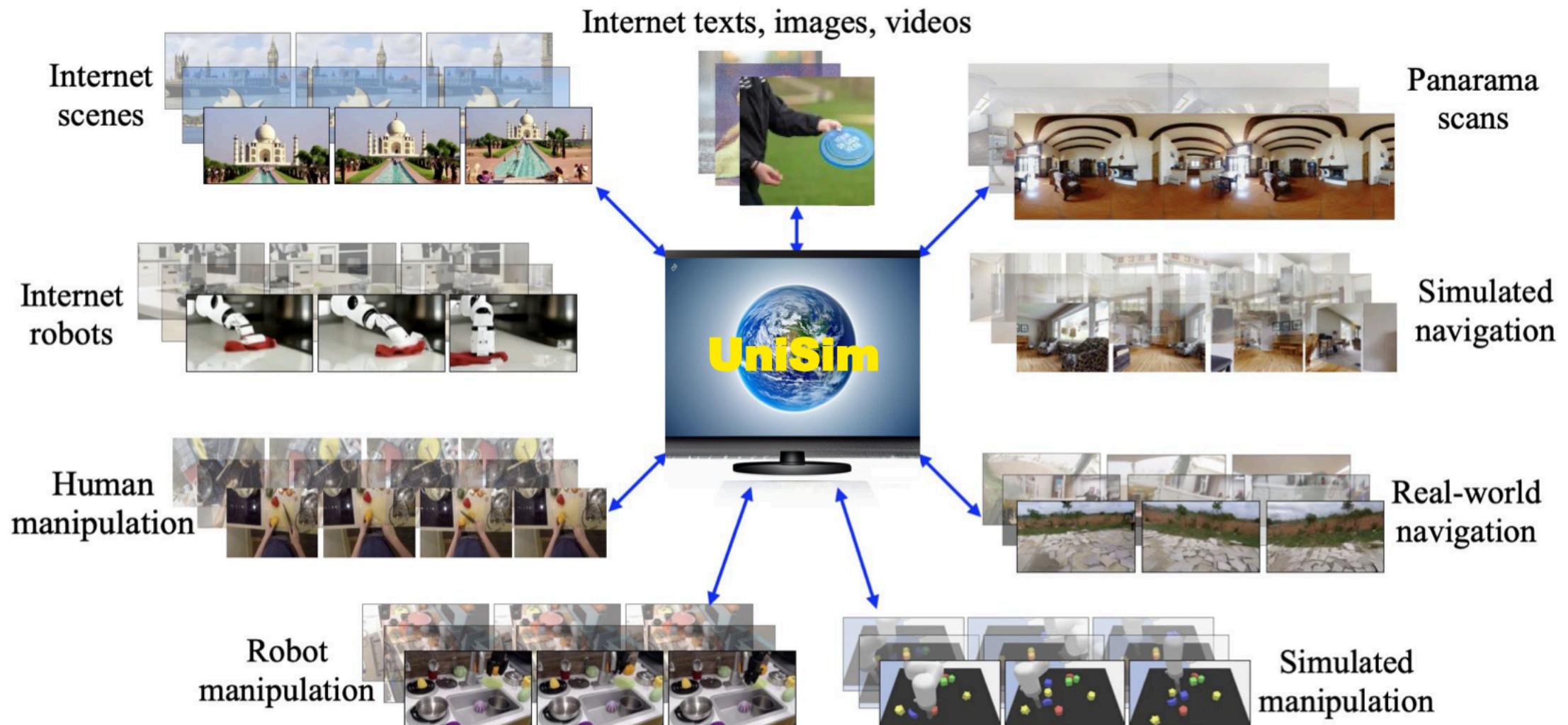
Super low-latency, real-time
models optimized for deploying at
the edge

Cosmos **Super**

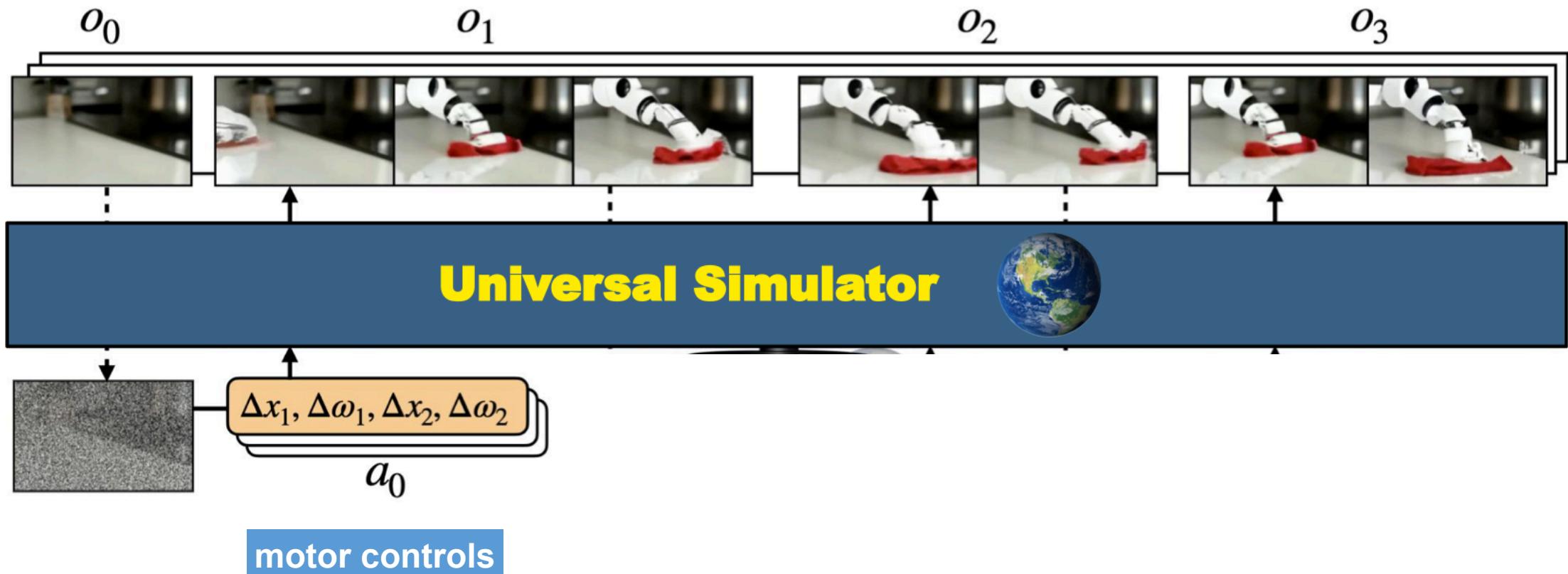
Highly performant baseline
models for out-of-the-box fine-
tuning and deployment

Cosmos **Ultra**

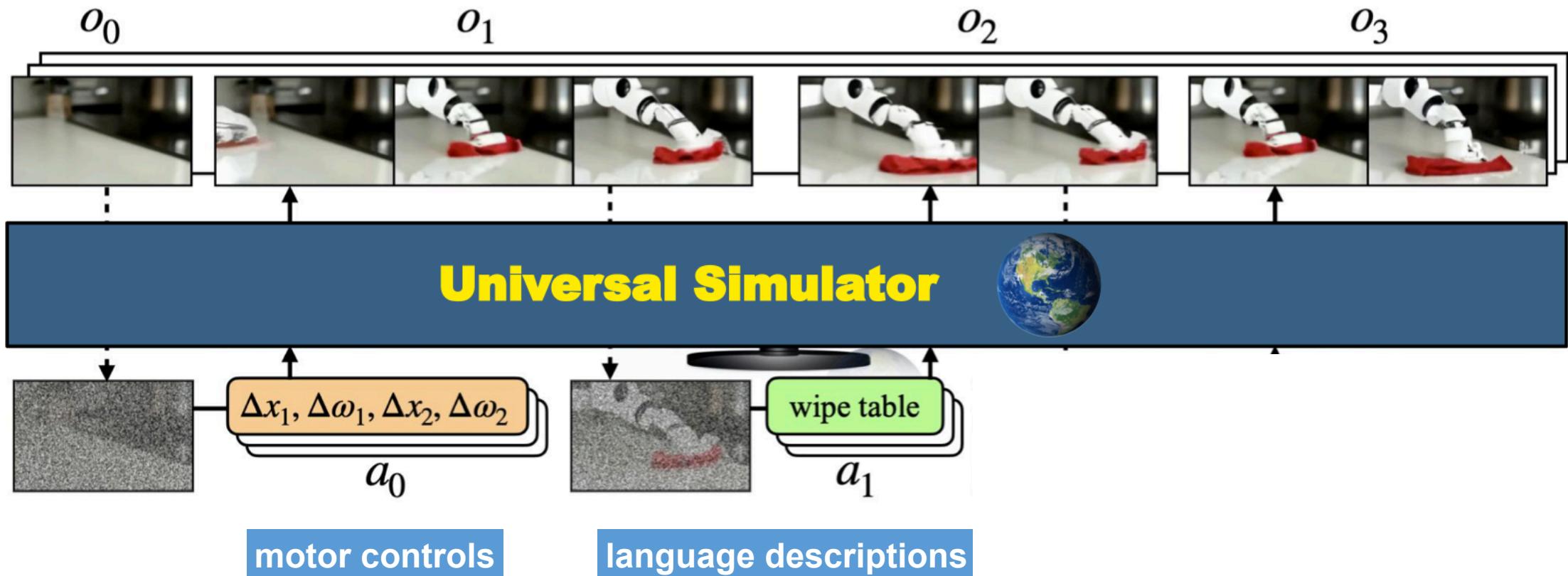
Maximum-accuracy and quality,
provides best-fidelity knowledge
transfer for distilling custom
models



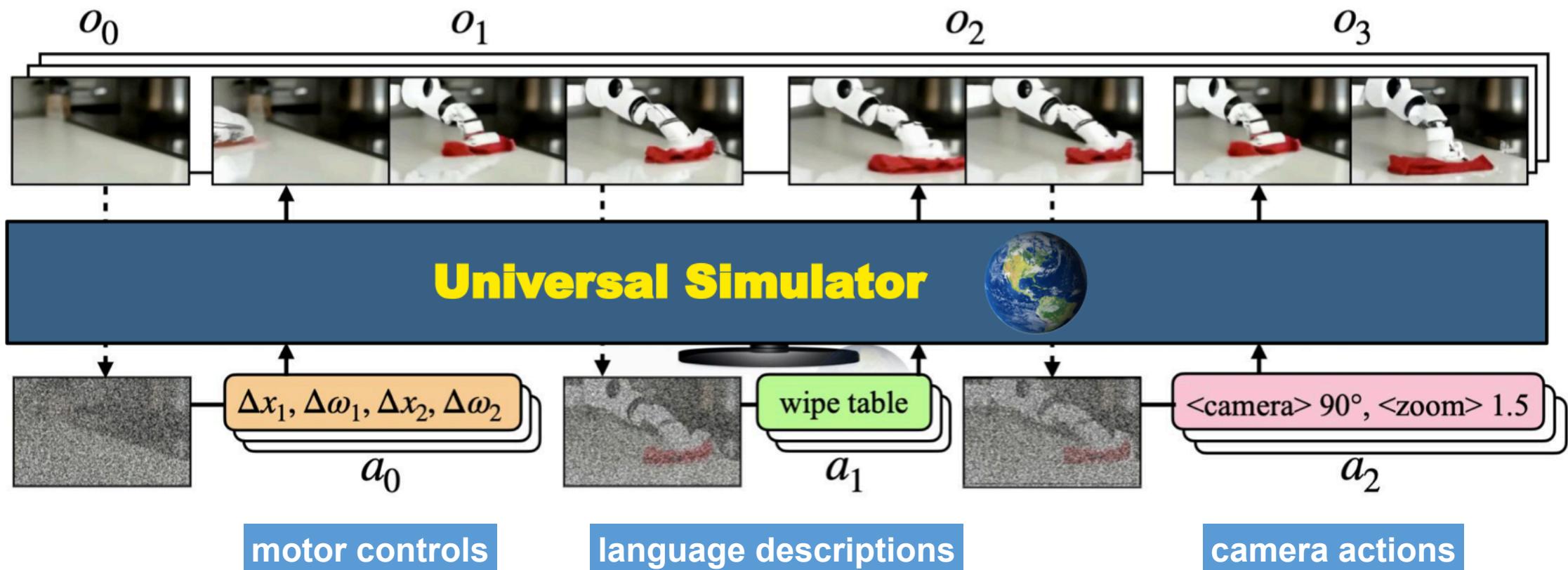
- ❑ A video diffusion model trained to predict the next (variable length) set of observation frames



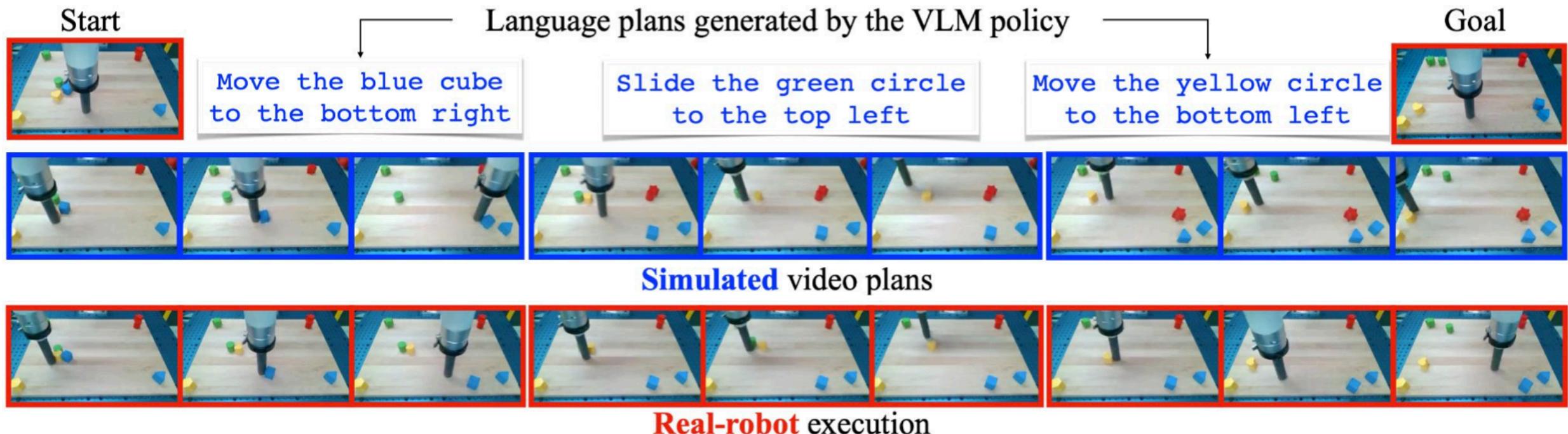
- ❑ A video diffusion model trained to predict the next (variable length) set of observation frames

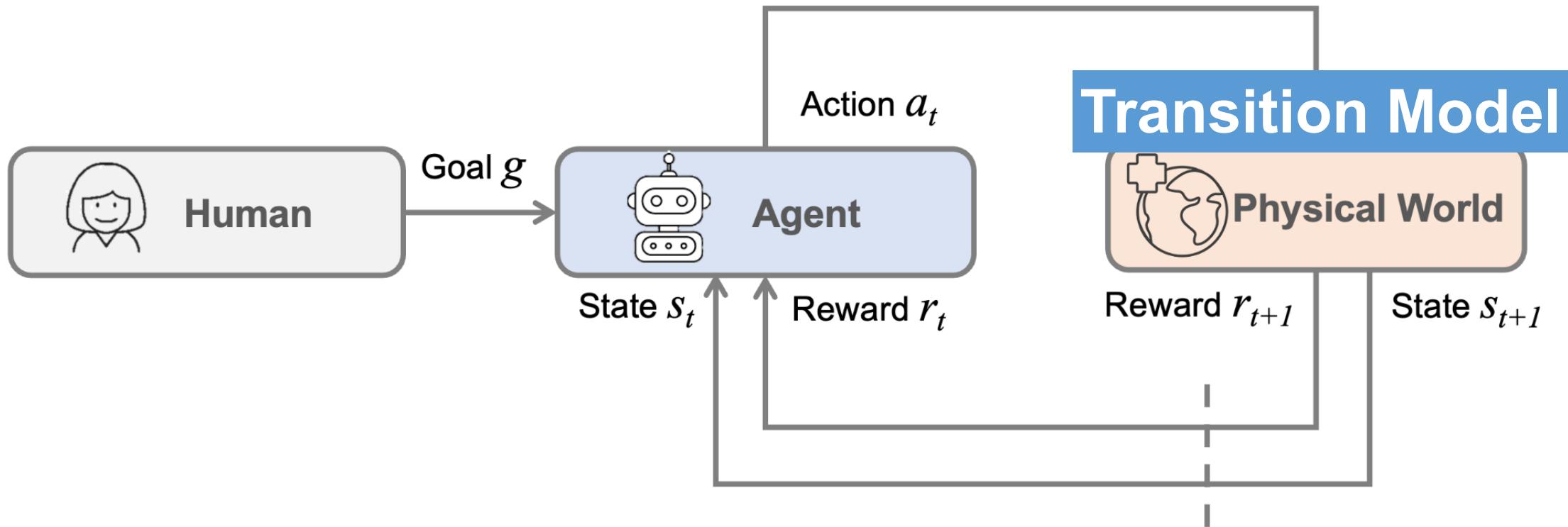


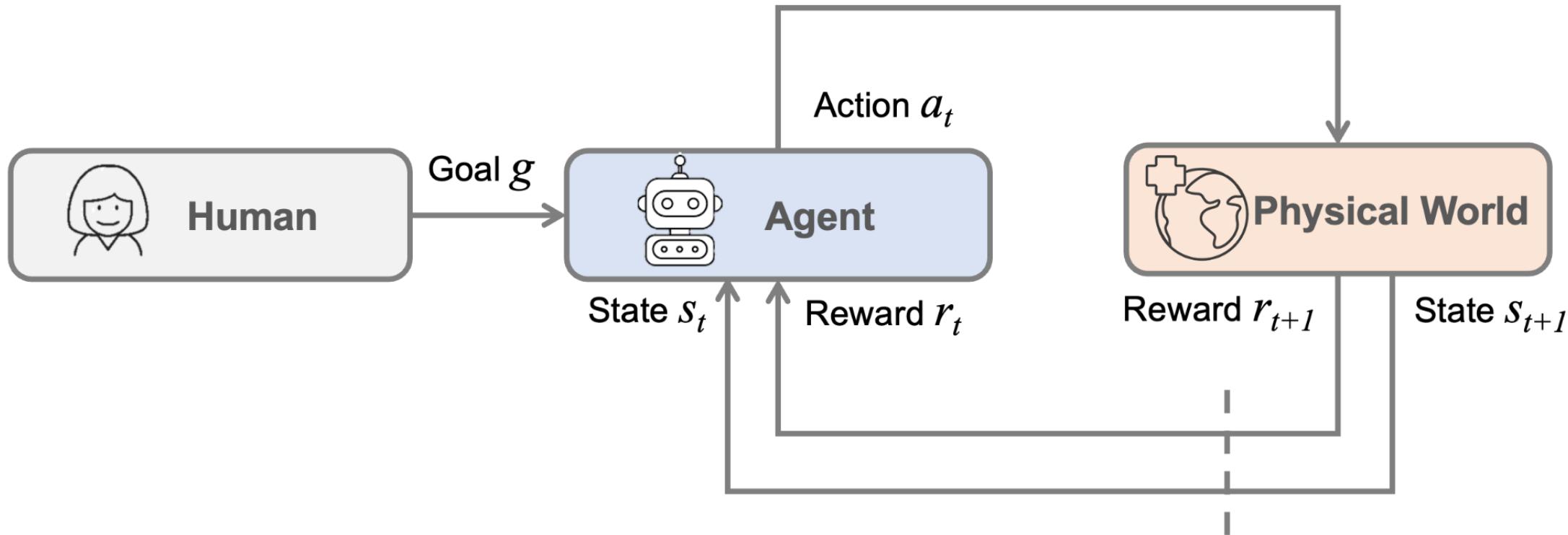
- ❑ A video diffusion model trained to predict the next (variable length) set of observation frames



❑ Support Long-horizon generation







Unified Virtual Agent



Goal

State

Action

Reward



Goal

State

Action

Reward

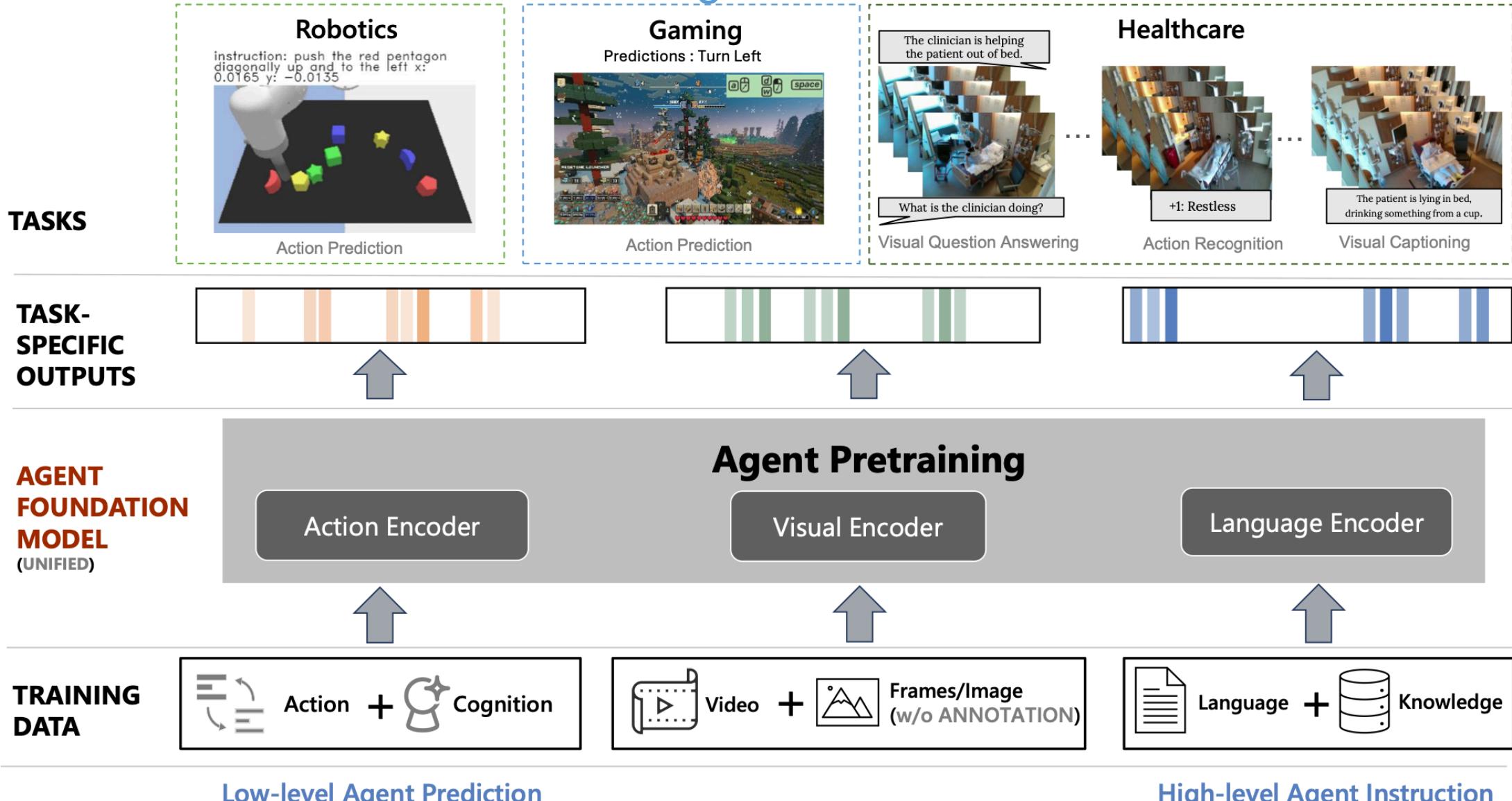
Option 1: Flat Concatenation

Option 2: Hierarchical Encoding

Option 1: Flat Concatenation



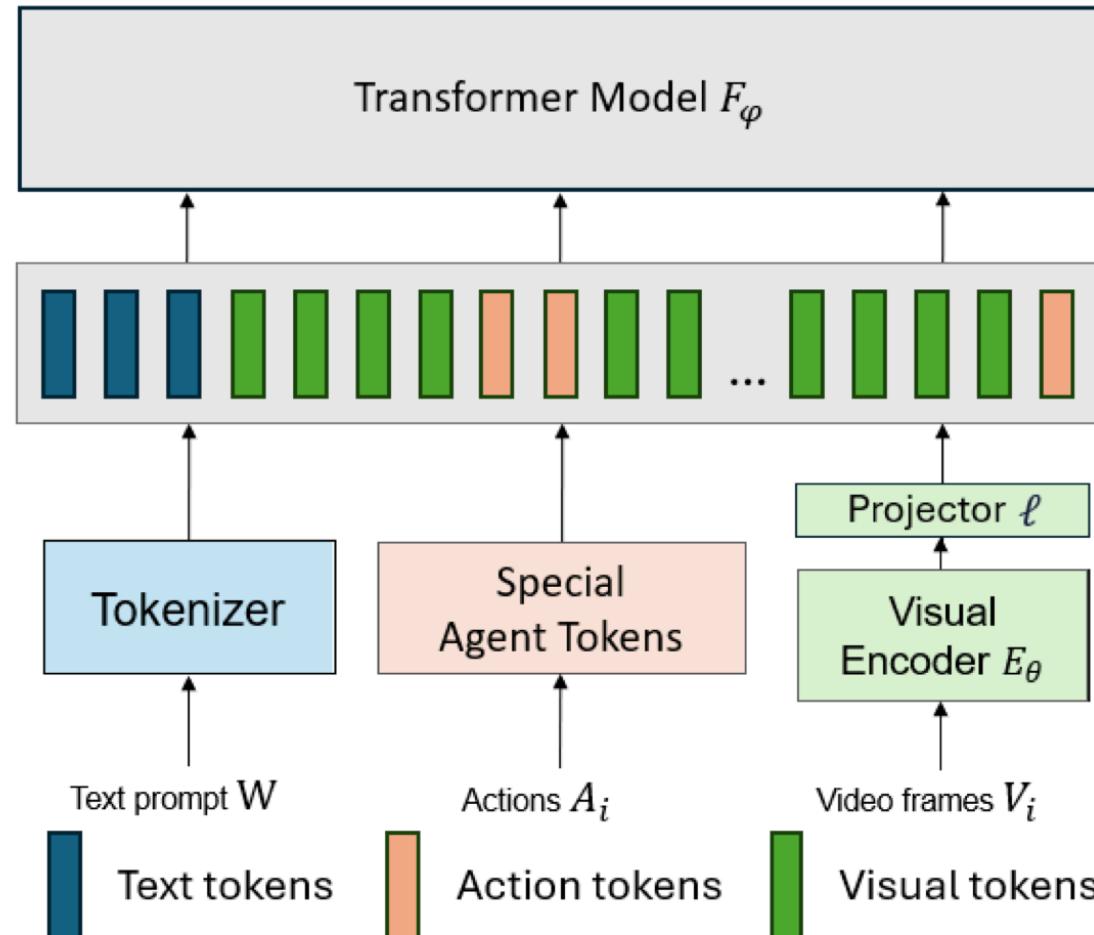
An Interactive Agent Foundation Model



Option 1: Flat Concatenation

An Interactive Agent Foundation Model

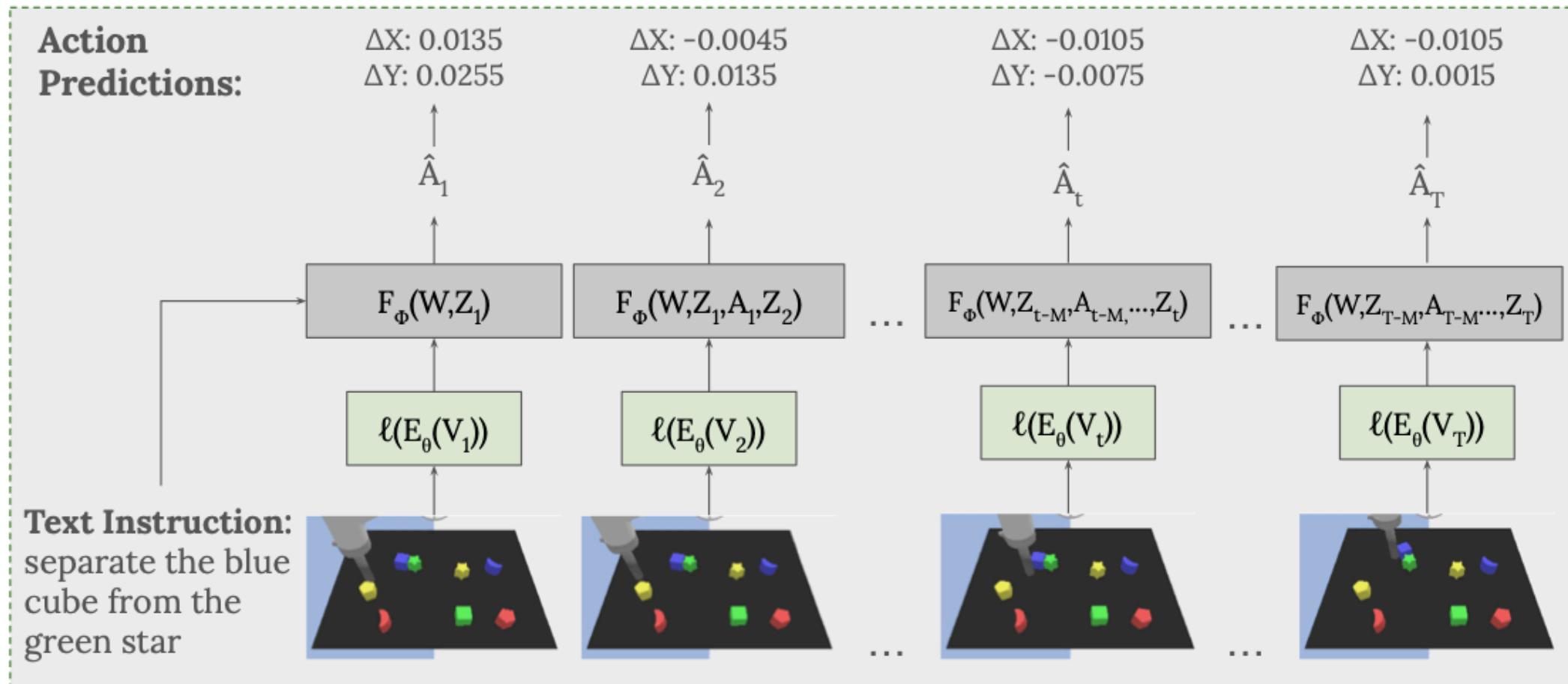
Unified Tokenization



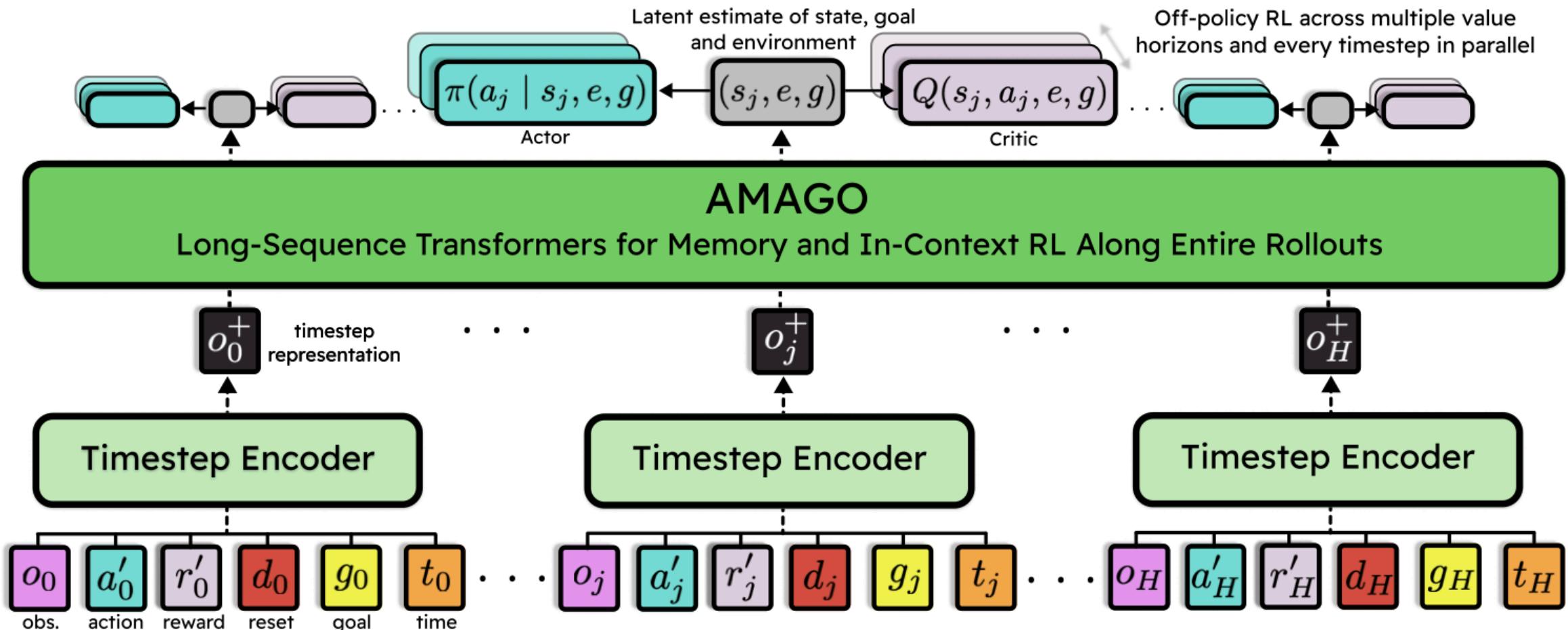
Option 1: Flat Concatenation

An Interactive Agent Foundation Model

❑ Pretraining Pipeline:



Option 2: Aggregating o, a, r to one vector



Multi-Agent Collaboration



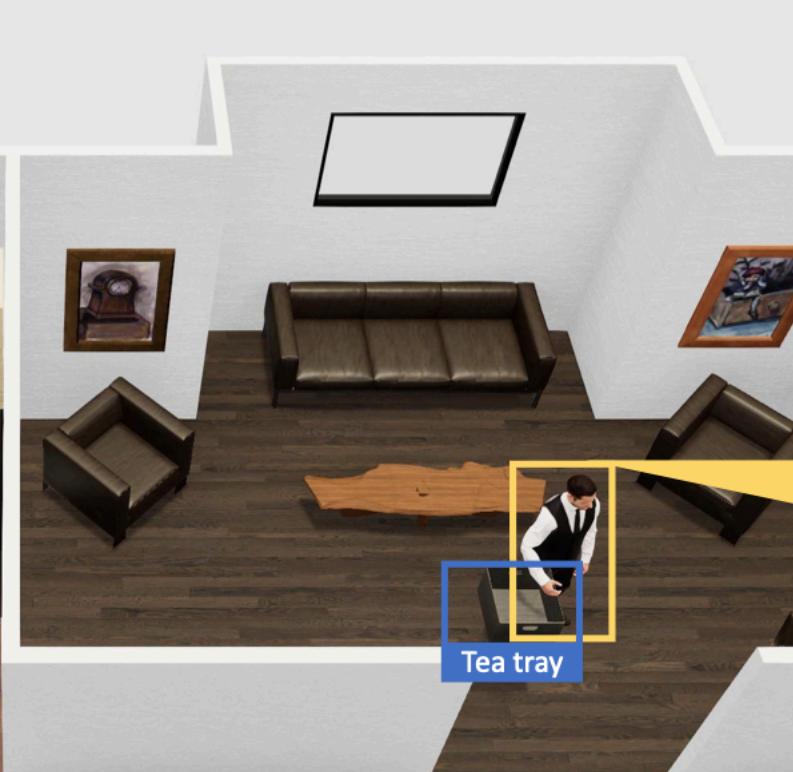
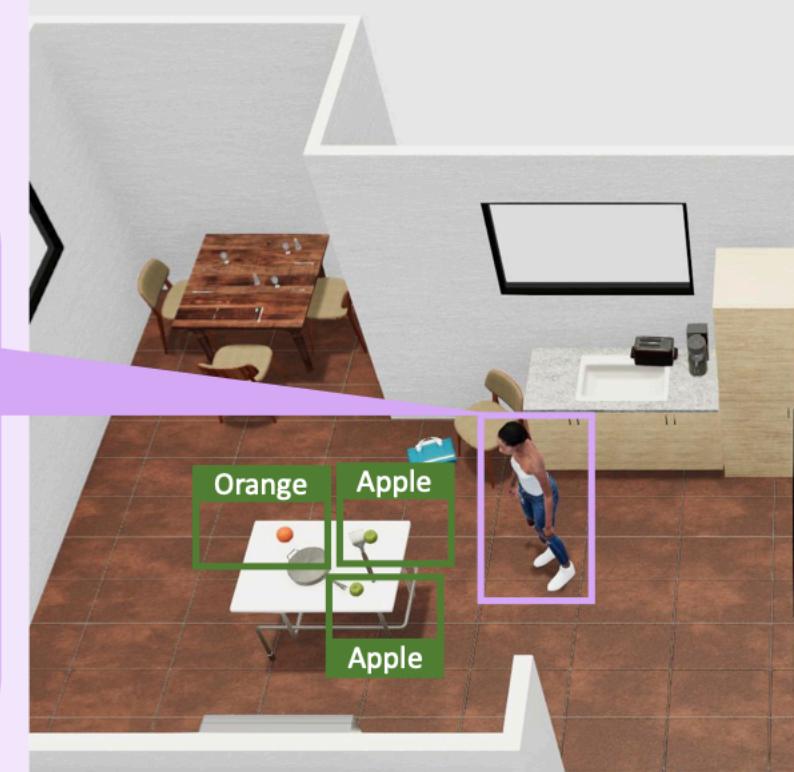
An Envisioned Agent Society





Alice
(LLM)

"Hi, Bob. I found 3 target objects in the kitchen, I remember you were holding an empty container, can you come here to pick them up while I go to explore other rooms?"



Bob
(Human/AI)

"OK, thanks for your information, I'll go to transport them."

Benchmarking

So Many Different Ways of Using LLMs

Existing Work	Goal Interpretation	Action Sequencing	Subgoal Decomposition	Transition Modeling
SayCan	LLMs	LLMs		
Ada	LLMs			LLMs
LLP+P	LLMs			
AutoTAMP		LLMs		LLMs
Code as Policies	LLMs	LLMs	LLMs	
Voyager	LLMs	LLMs		
Demo2Code	LLMs		LLMs	LLMs
LM as ZeroShot Planner		LLMs	LLMs	
SayPlan	LLMs	LLMs		LLMs
Text2Motion		LLMs		
LLMGROP	LLMs	LLMs		
REFLECT	LLMs	LLMs		
Generating Consistent PDDL Domains with LLMs	LLMs			LLMs
PlanSeqLearn		LLMs		
COWP	LLMs	LLMs		LLMs

So Many Different Ways of Using LLMs

Existing Work	Goal Interpretation	Action Sequencing	Subgoal Decomposition	Transition Modeling
CAPE	LLMs	LLMs		
HERACLEs		LLMs		
RoboTool		LLMs		LLMs
PROMST		LLMs		
LLM3	LLMs	LLMs		
Ghost in the Minecraft		LLMs		
PlanBench	LLMs	LLMs		
TaPA	LLMs	LLMs	LLMs	
ChatGPT Robot Control		LLMs		
LLM World Models for Planning	LLMs	LLMs		
DEPS	LLMs	LLMs		
Grounded Decoding		LLMs		
ProgPrompt	LLMs	LLMs		
DROC		LLMs		LLMs
LMPC	LLMs	LLMs		
GPTPDDL		LLMs		

Different Purposes

Different Input/Output

So we need **Standardization!**

Embodied Agent Interface

Embodied Agent Interface

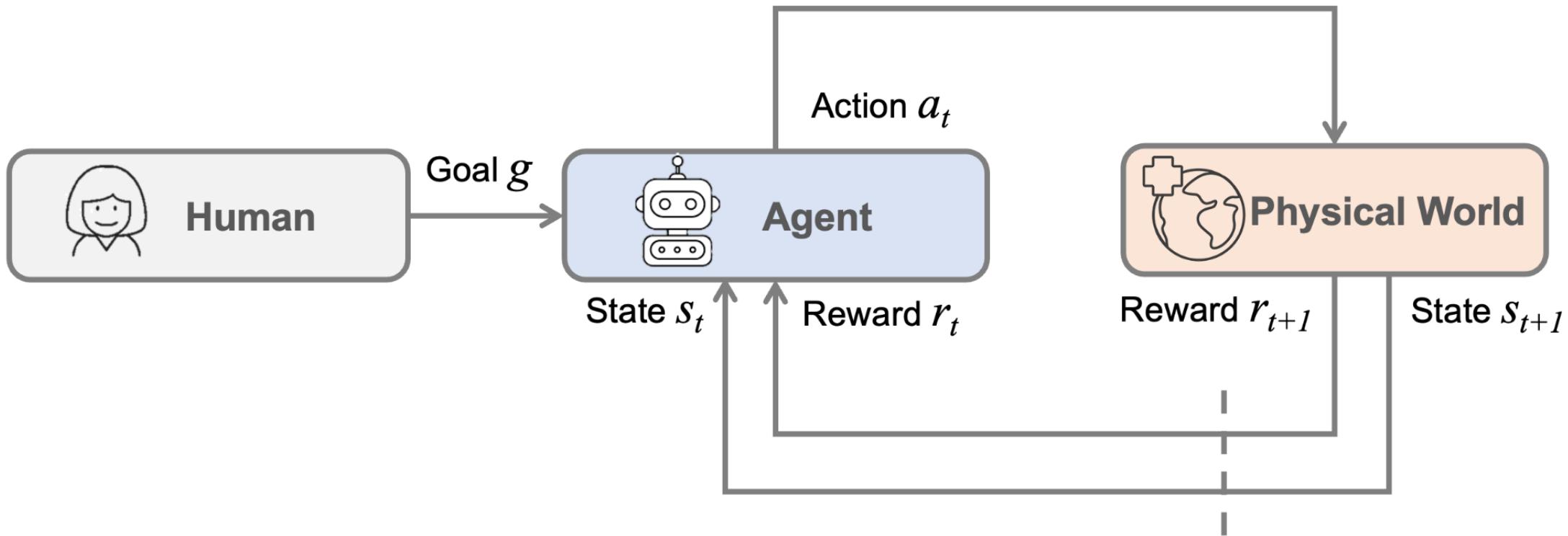
to benchmark LLMs for Embodied Decision Making

LLMs

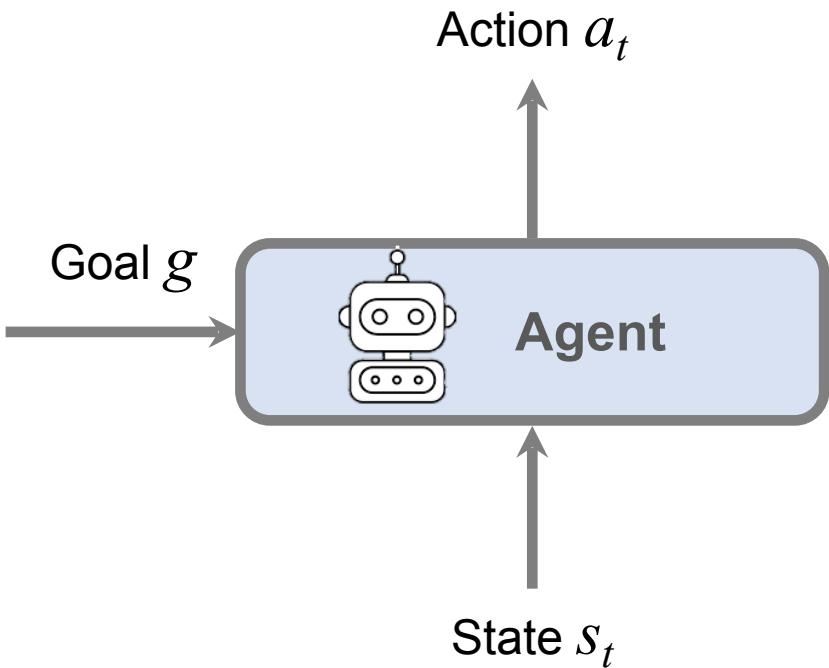
Embodied Agent Interface

Embodied Decision Making

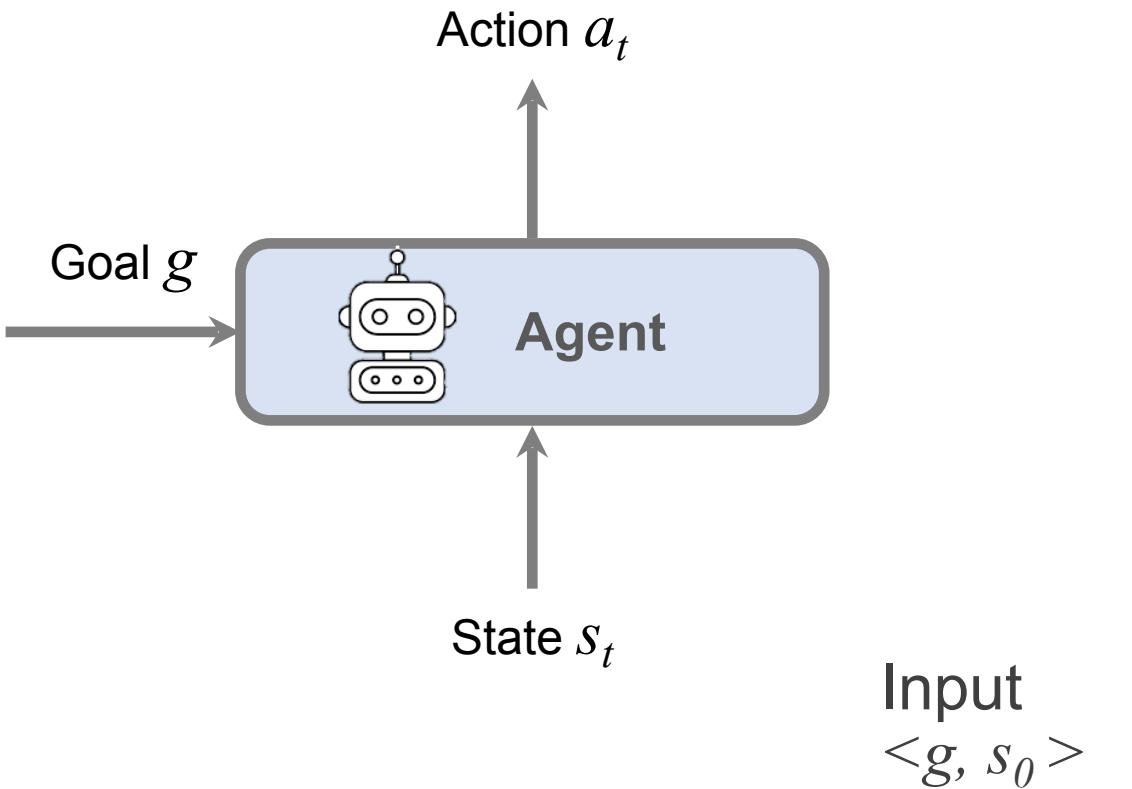
Let us go back to MDPs (Markov Decision Processes)



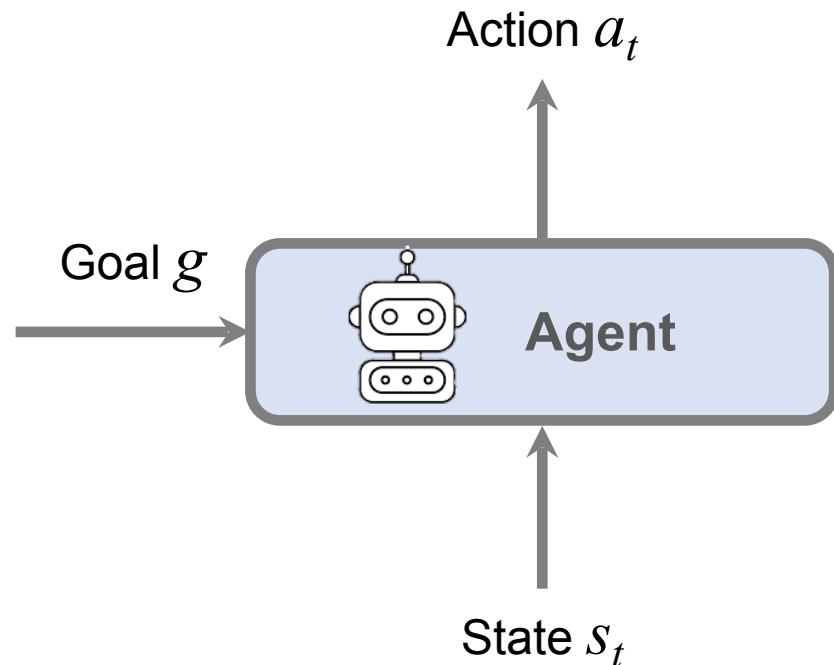
Markov Decision Process



Markov Decision Process



Markov Decision Process



Output

$\langle a_1, a_2, \dots, a_t \rangle$

Action
Sequencing

$\langle s_1, s_2, \dots, s_t \rangle$

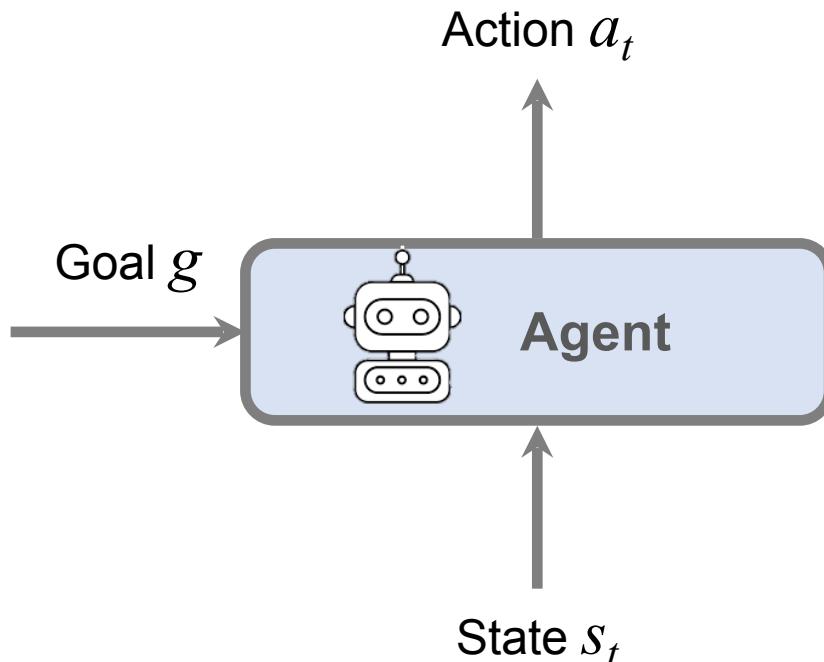
Subgoal
Decomposition

Input

$\langle g, s_0 \rangle$

Goal
Interpretation

Markov Decision Process



Output

$\langle a_1, a_2, \dots, a_t \rangle$

Action Sequencing

$\langle s_1, s_2, \dots, s_t \rangle$

Subgoal Decomposition

Learn

$(s_t, a_t) \rightarrow s_{t+1}$

Transition Modeling

Input

$\langle g, s_0 \rangle$

Goal Interpretation

Embodied Agent Interface

Action
Sequencing

Subgoal
Decomposition

Transition
Modeling

Goal
Interpretation

Embodied Agent Interface

Goal
Interpretation

Subgoal
Decomposition

Action
Sequencing

Transition
Modeling

LLMs

Embodied Agent Interface

Goal
Interpretation

Subgoal
Decomposition

Action
Sequencing

Transition
Modeling

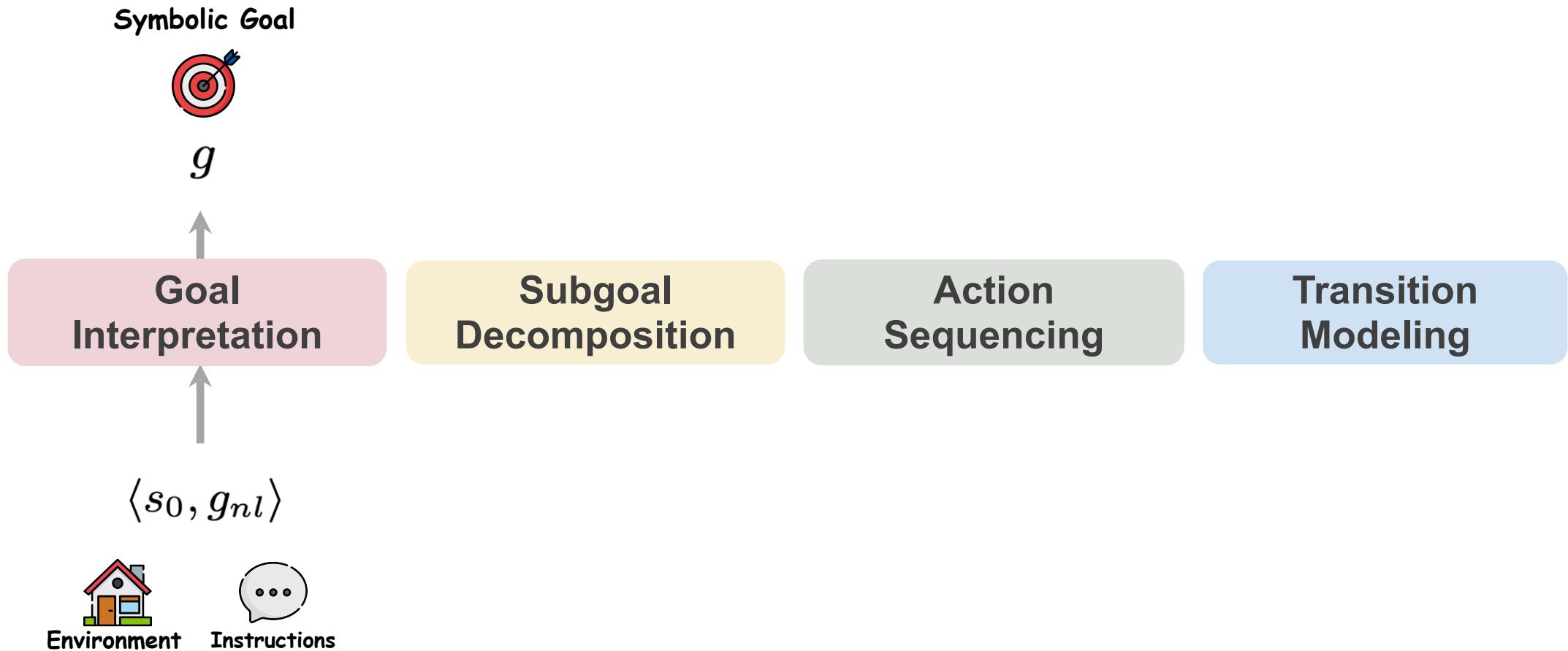
Embodied Environments:

BEHAVIOR

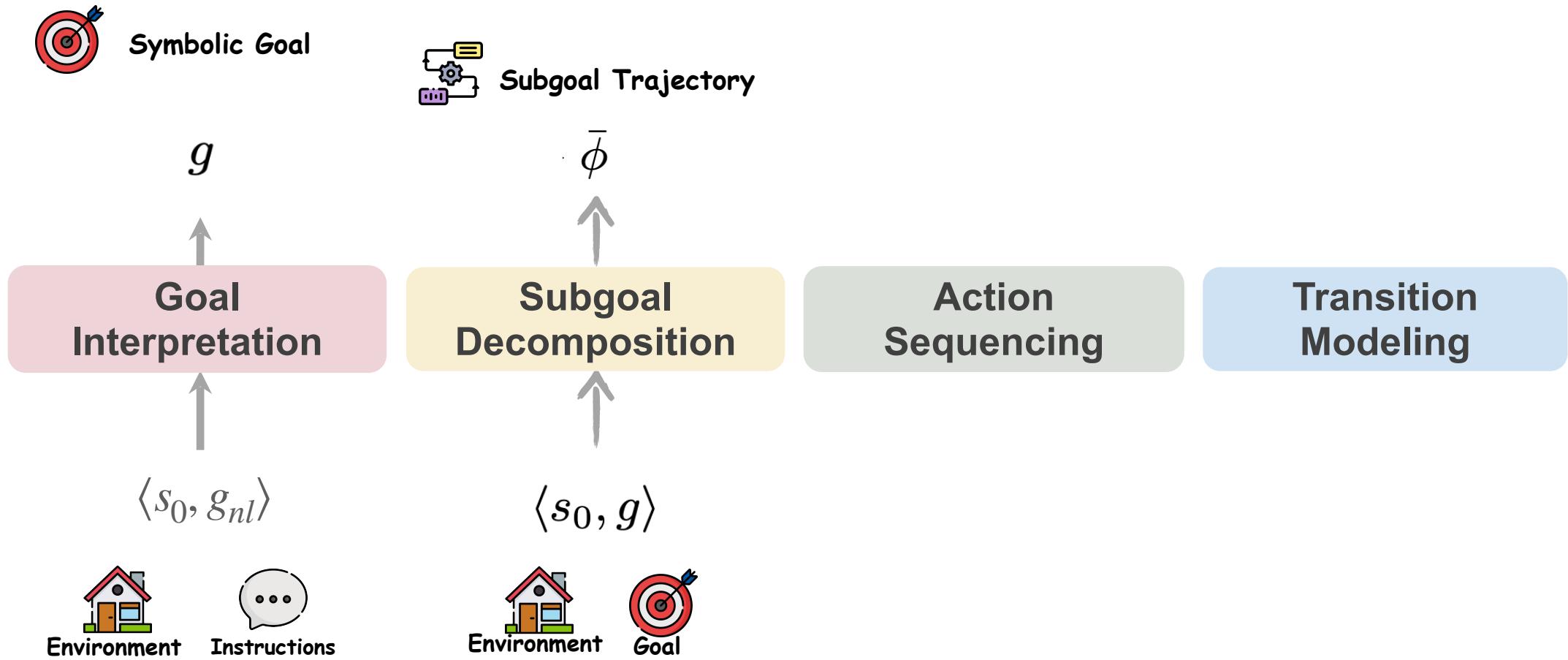
VirtualHome

...

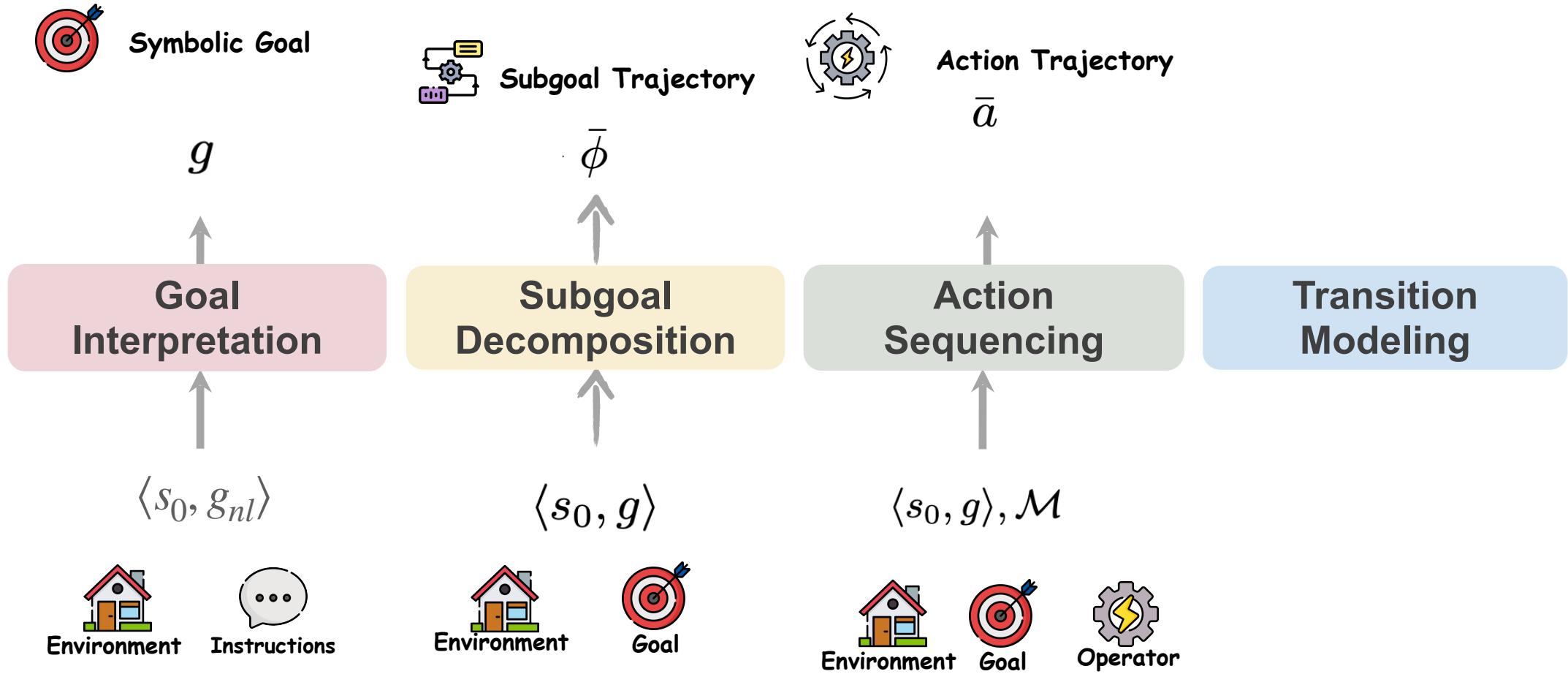
Standardization of modules and interfaces



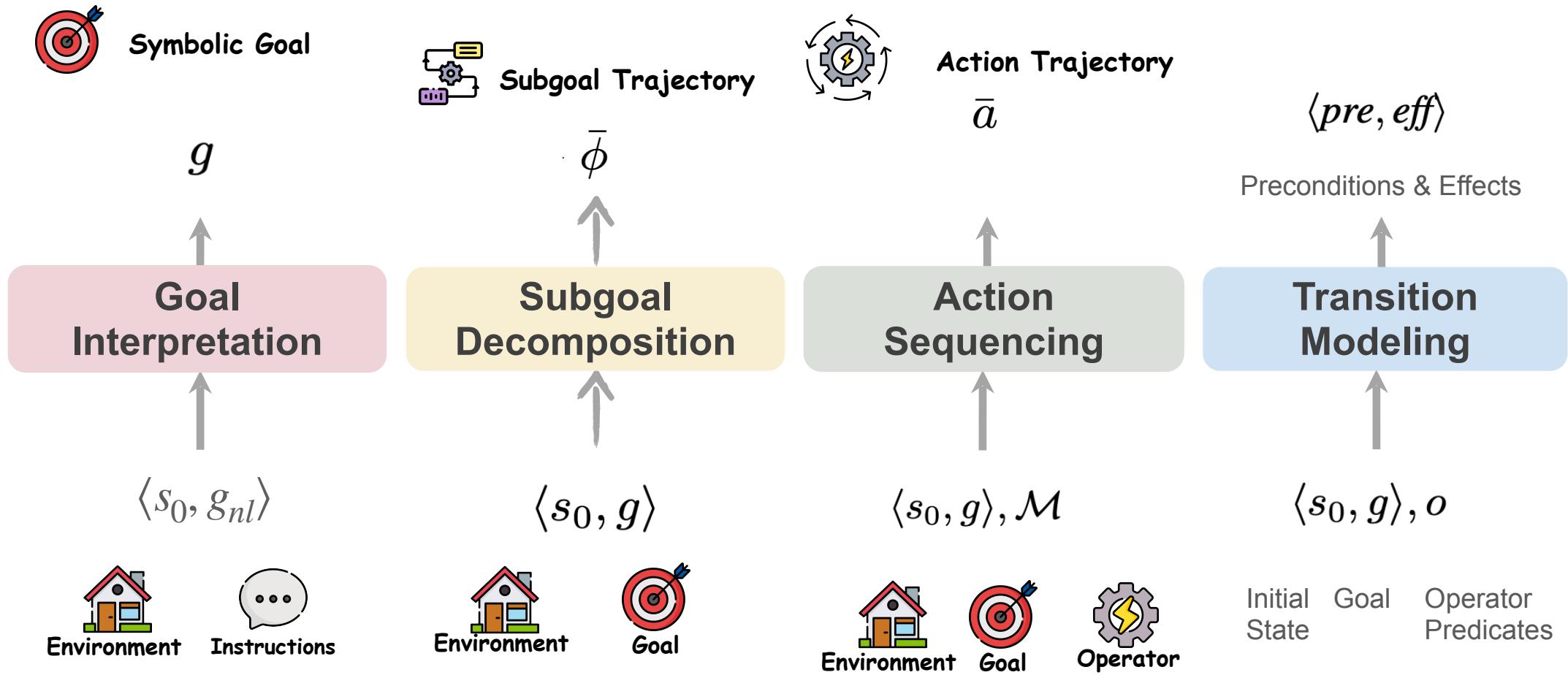
Standardization of modules and interfaces

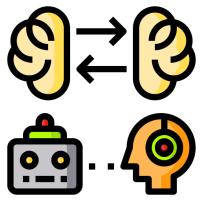


Standardization of modules and interfaces



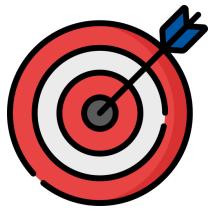
Standardization of modules and interfaces





Standardize Modules and Interfaces

4 modules
438 tasks
1475 goals



Standardize Goal Specifications

Linear
Temporal
Logic



Standardize Goal Specifications

Objects



Task Goal

exist a that: inside **and**
 not inside

forall that: **not** open

Task: Bottling Fruits

open open

strawberry inside

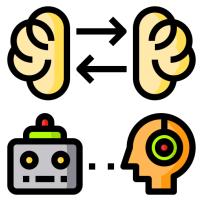
strawberry inside

orange inside

orange inside

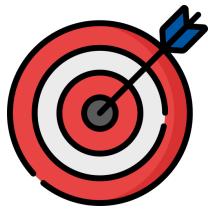
not open **not** open

Linear Temporal Logic



Standardize Modules and Interfaces

4 modules
438 tasks
1475 goals



Standardize Goal Specifications

Linear
Temporal
Logic

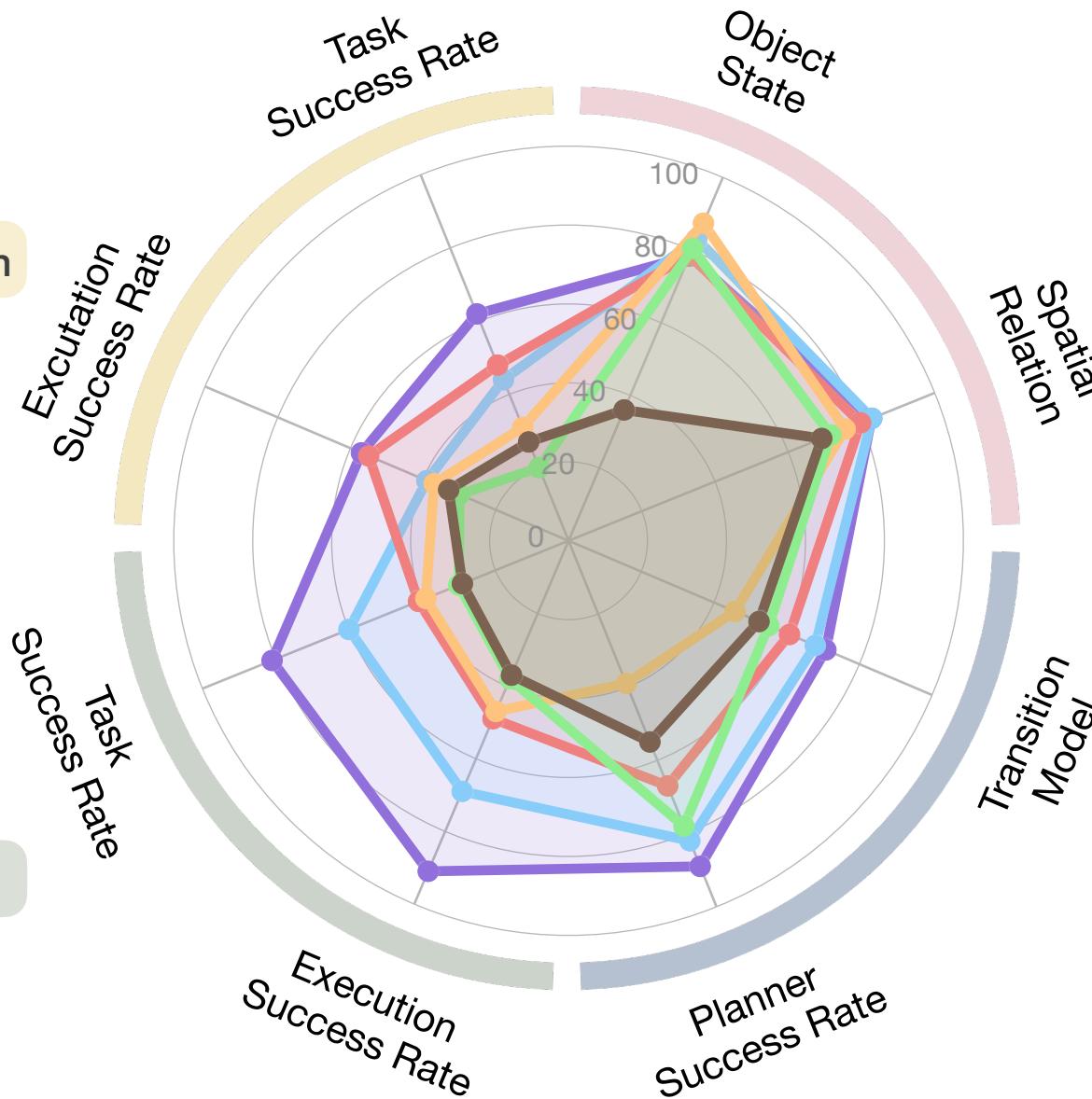


Standardize Fine-grained Metrics

18 models
42 metrics
100+ page analysis

How well do LLMs work?

18 models on 4 core abilities:



Subgoal Decomposition

Goal Interpretation

Action Sequencing

Transition Modeling

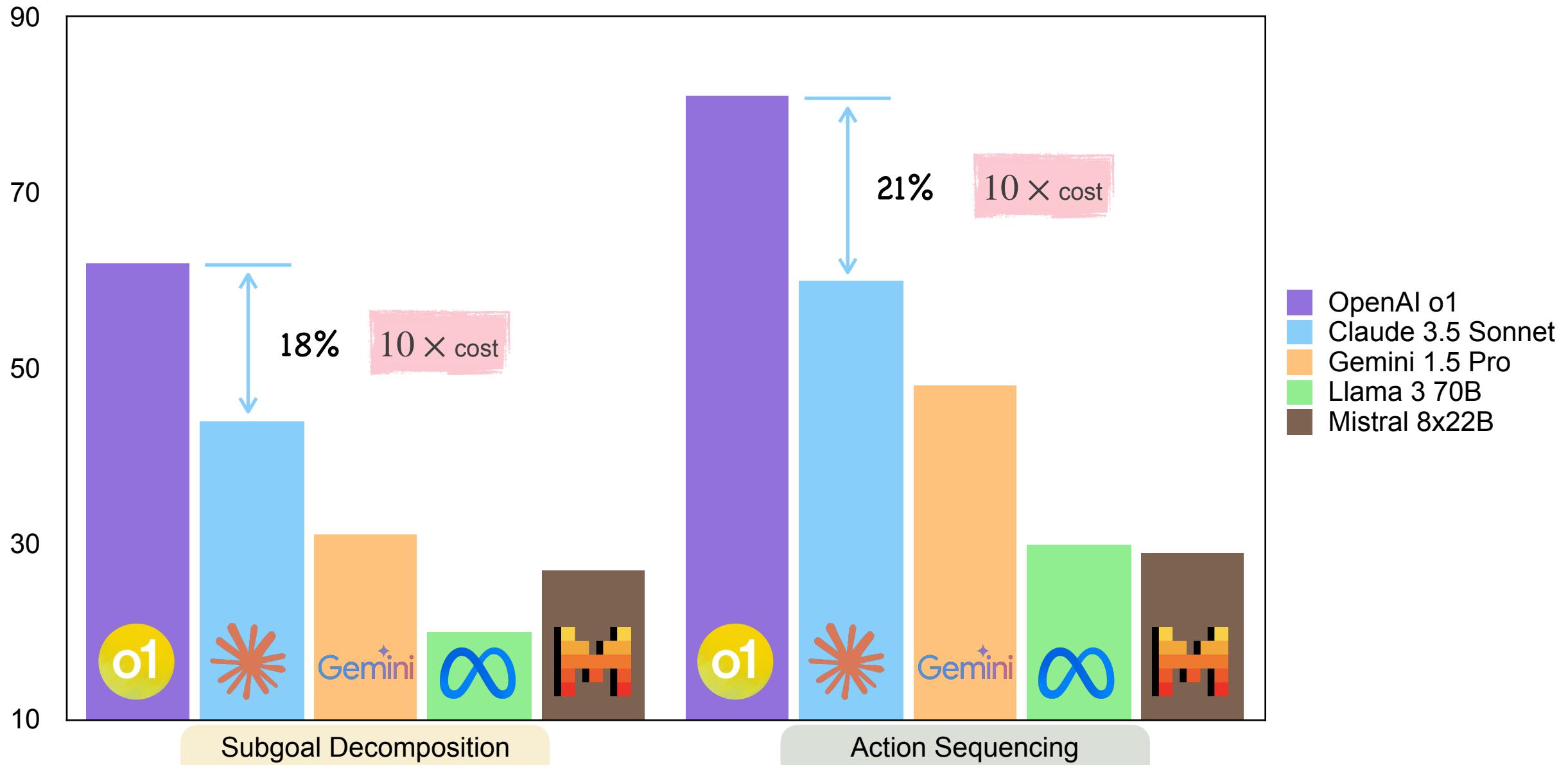
■ Mixtral 8x22B
● GPT-4o

● Llama 3 70B
● Claude 3.5 Sonnet

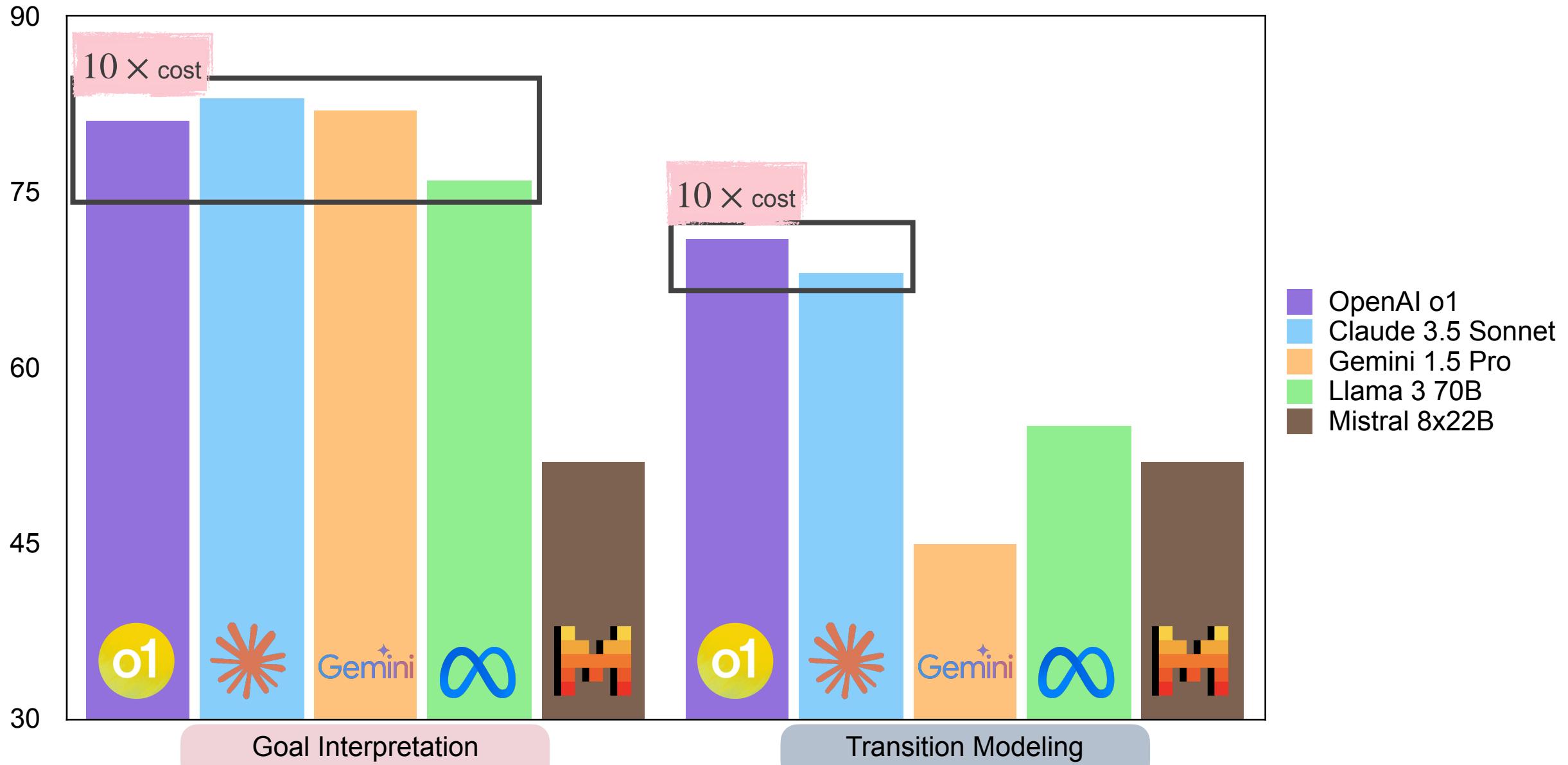
● Gemini 1.5 Pro
● o1-preview

Which model is the best for my task?

Subgoal & Action Sequencing



State Understanding



Where do they fail and why?

How can we improve them?

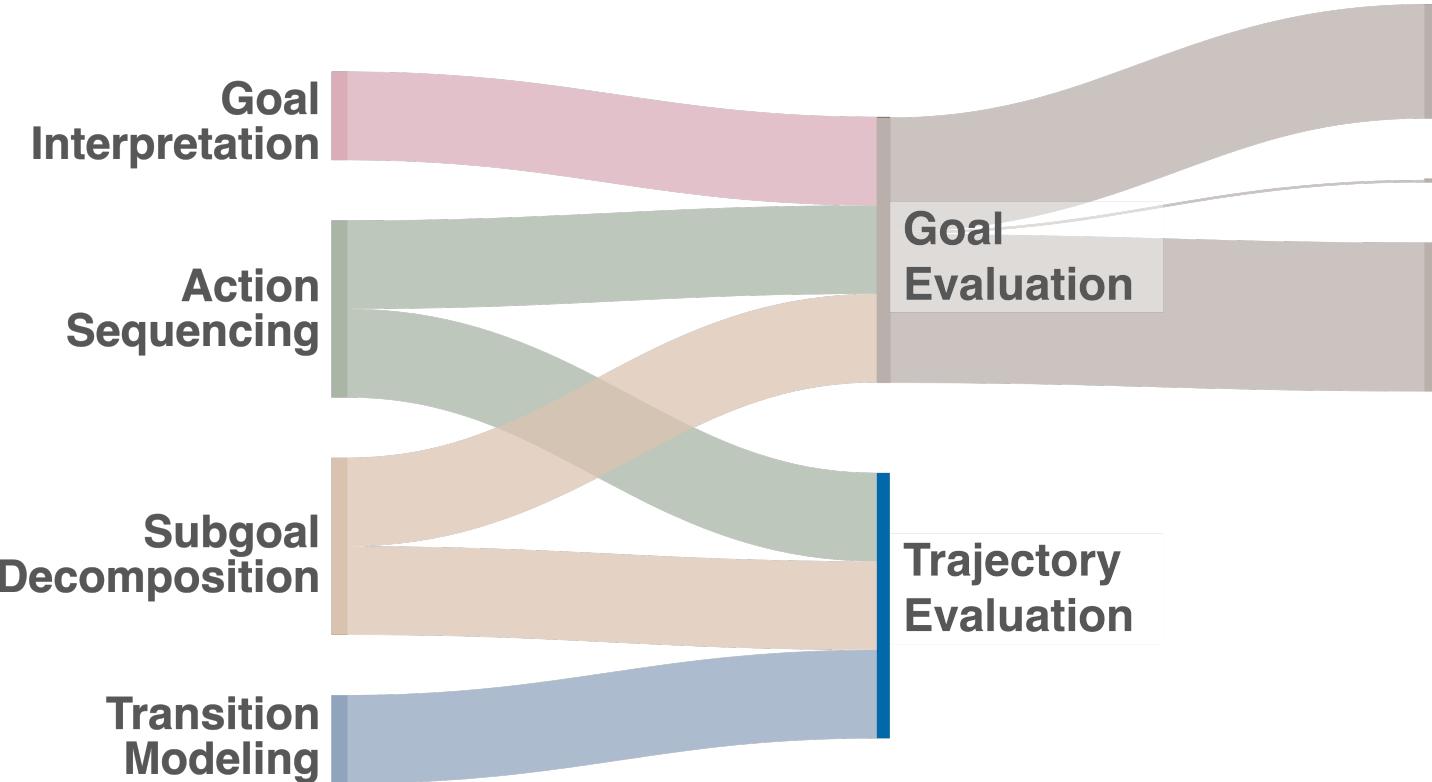
How about Goal Understanding?

43.1% Goal-related Errors



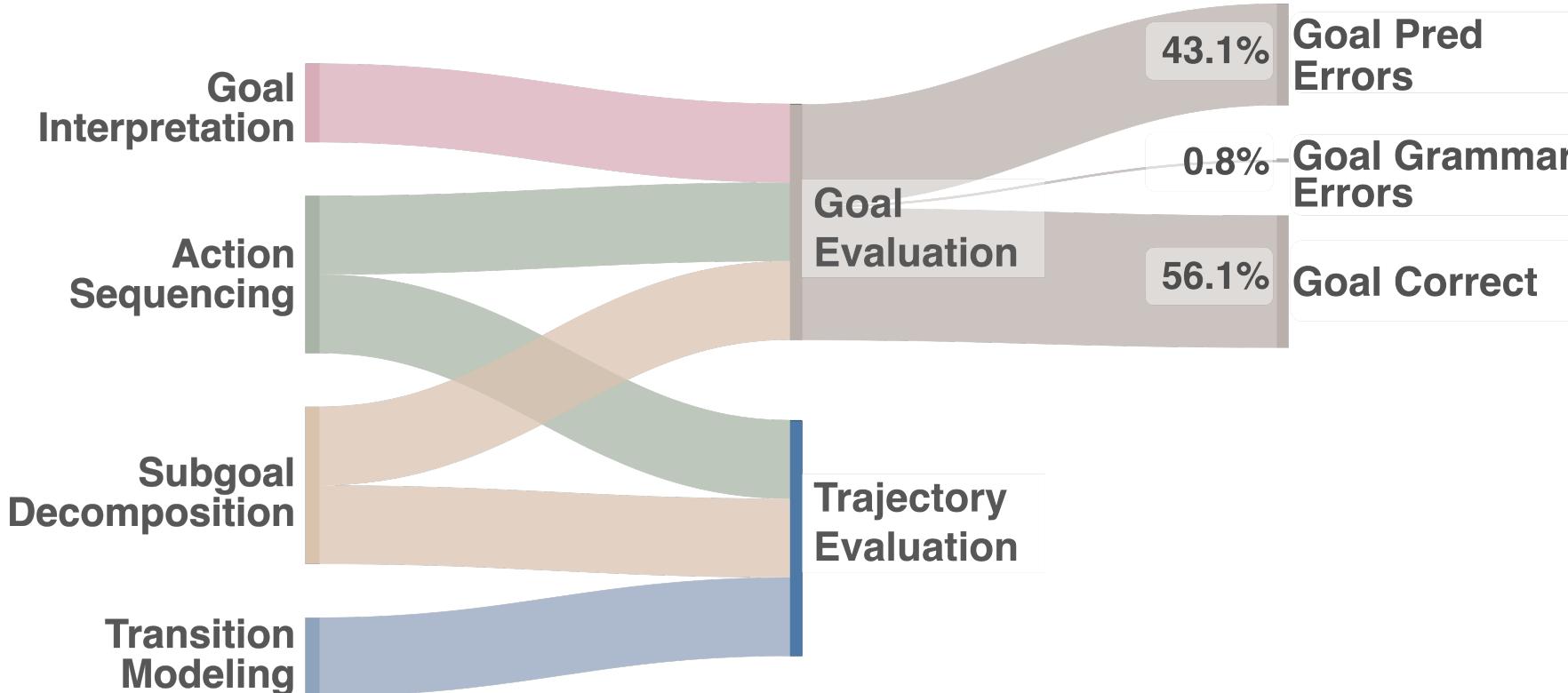
How about Goal Understanding?

43.1% Goal-related Errors



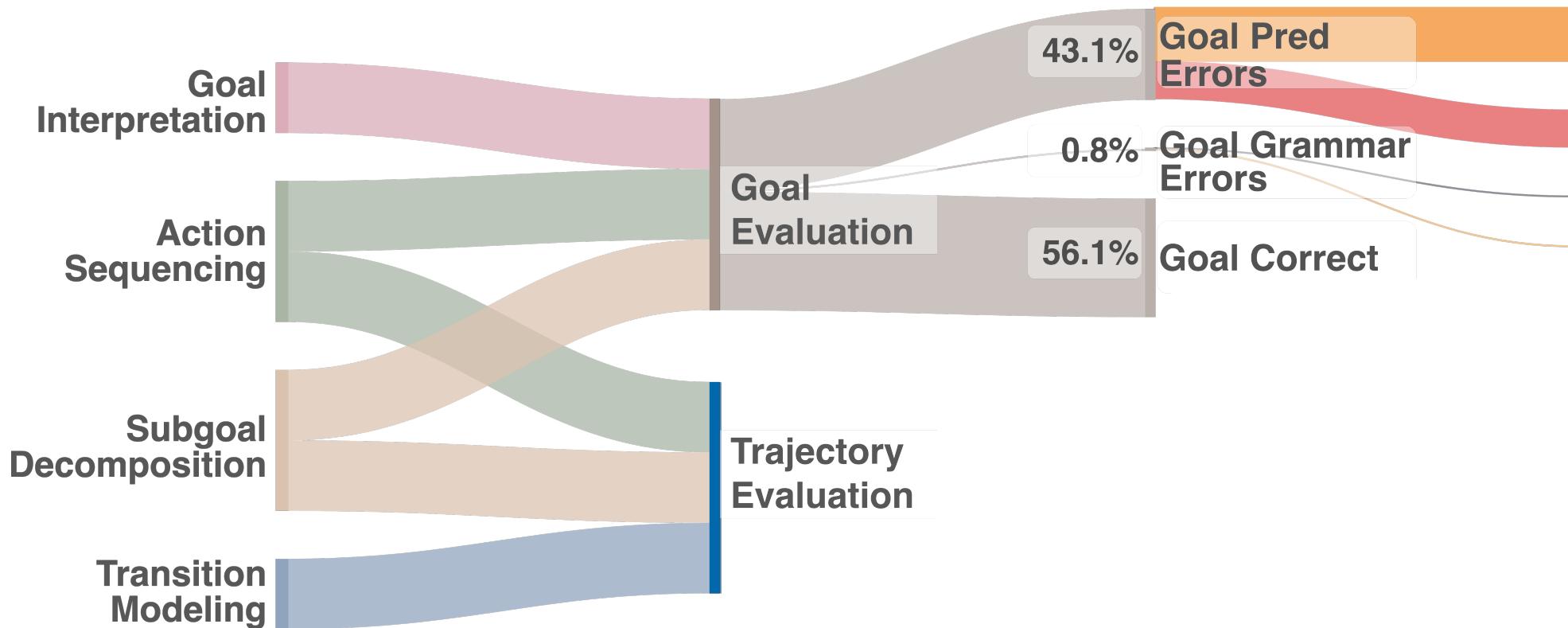
How about Goal Understanding?

43.1% Goal-related Errors



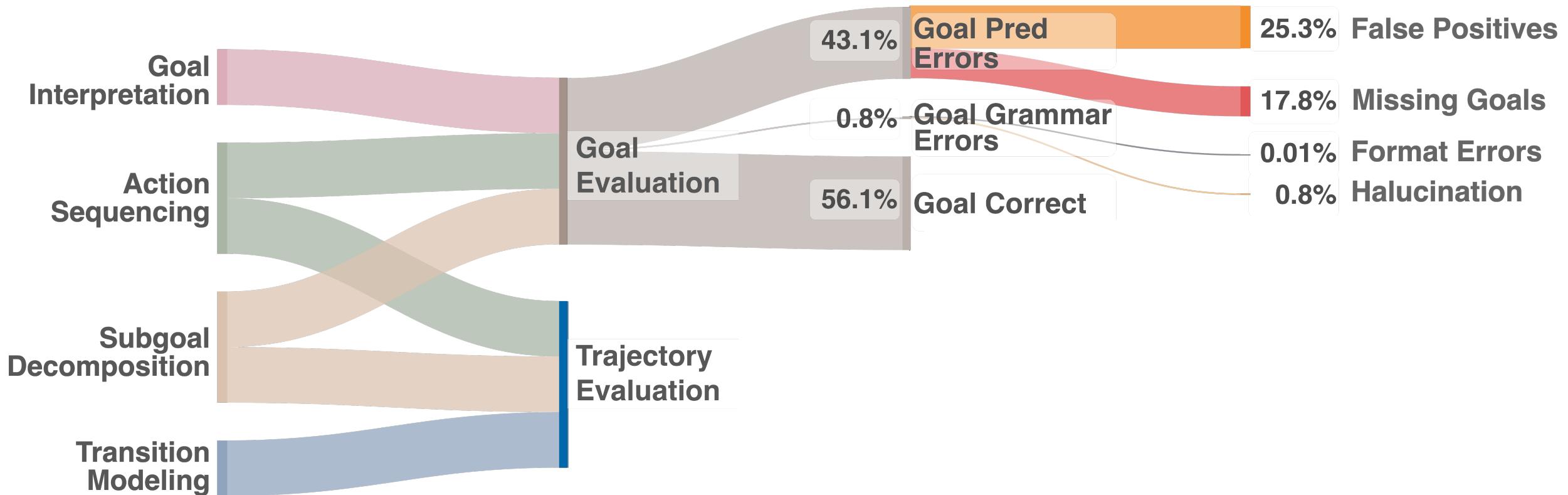
How about Goal Understanding?

43.1% Goal-related Errors



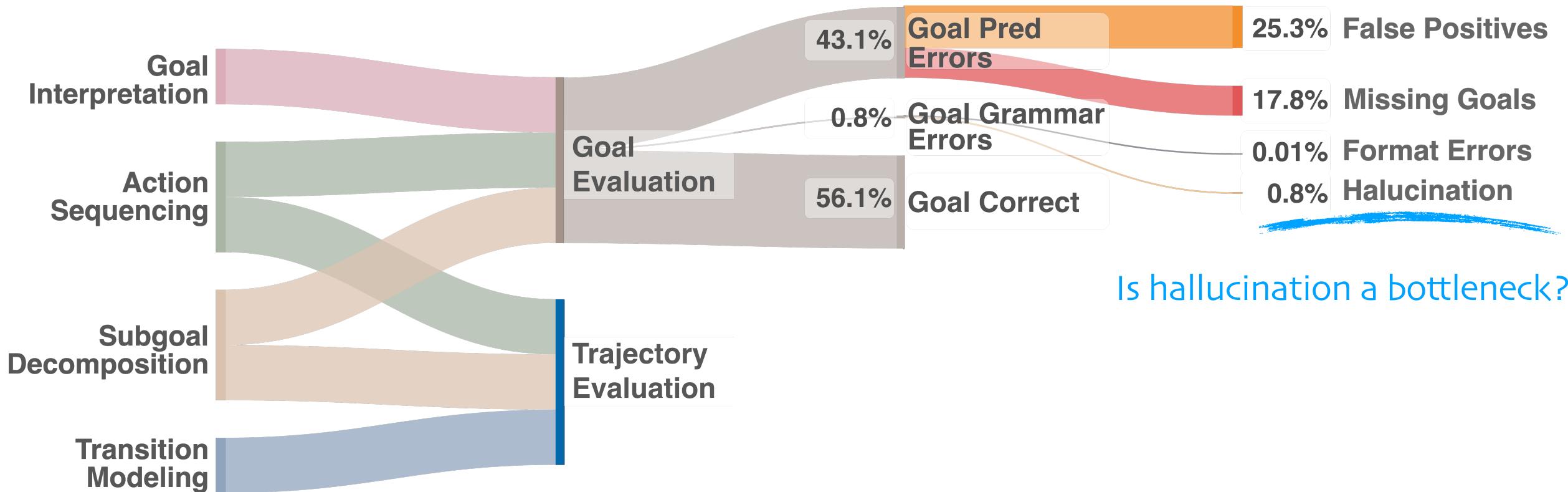
How about Goal Understanding?

43.1% Goal-related Errors



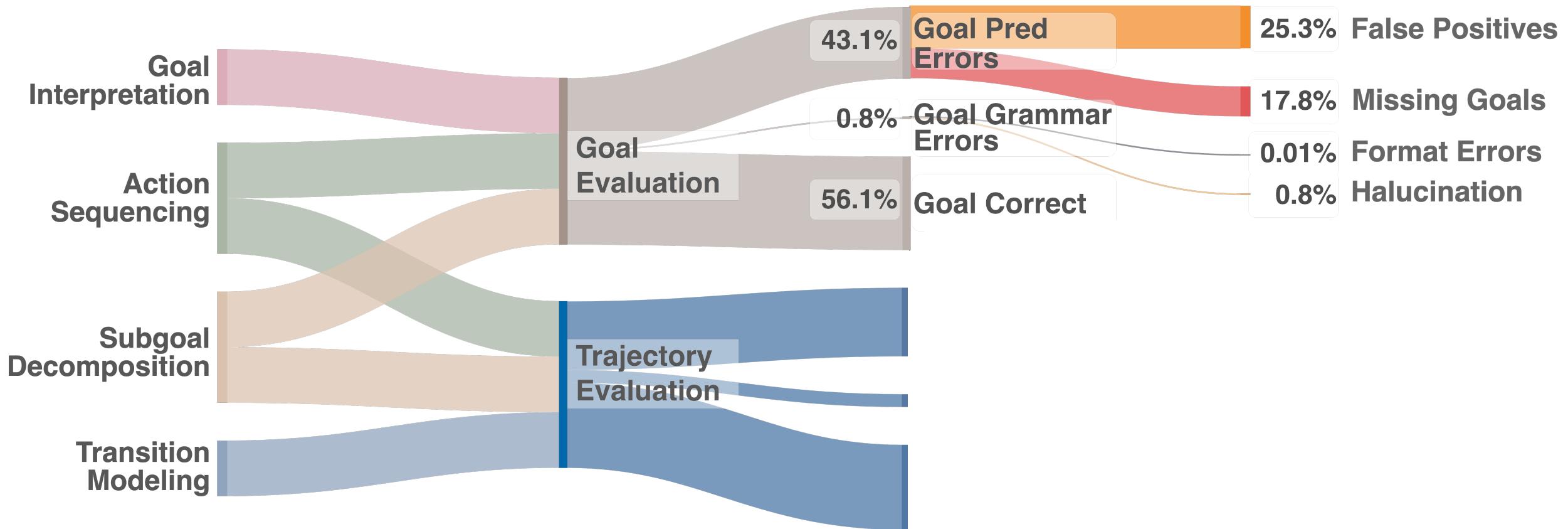
How about Goal Understanding?

Is hallucination a bottleneck?



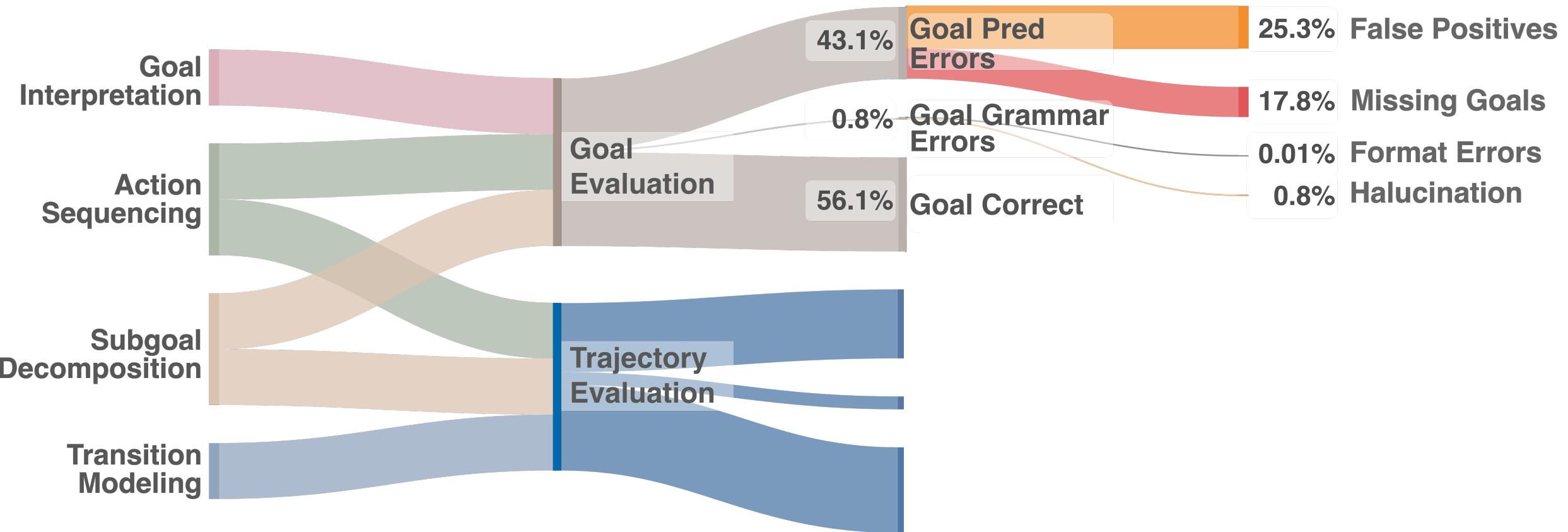
How about Goal Understanding?

Is hallucination a bottleneck?



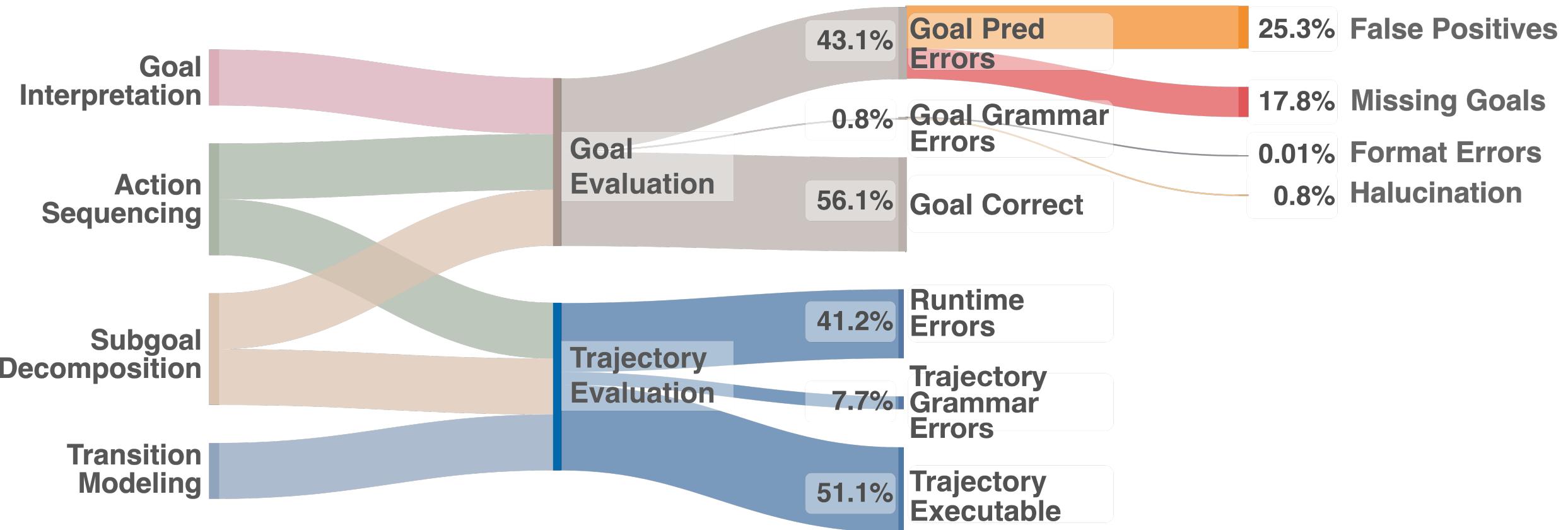
How about Trajectory Prediction?

43.2% Runtime Errors



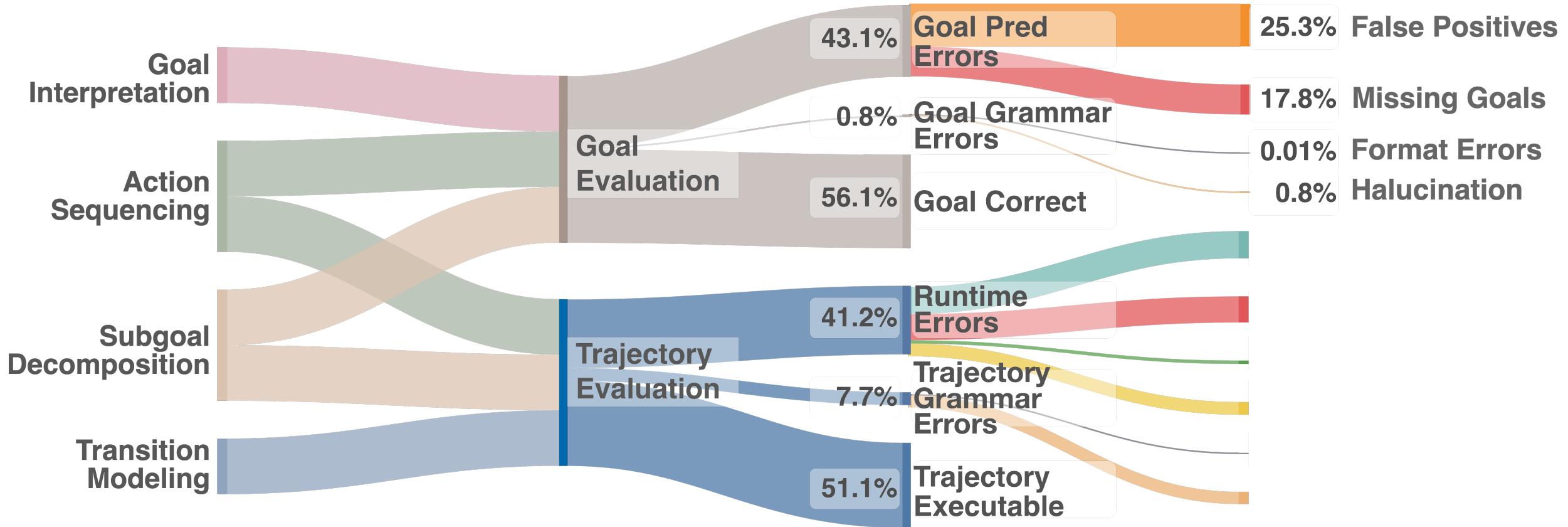
How about Trajectory Prediction?

43.2% Runtime Errors



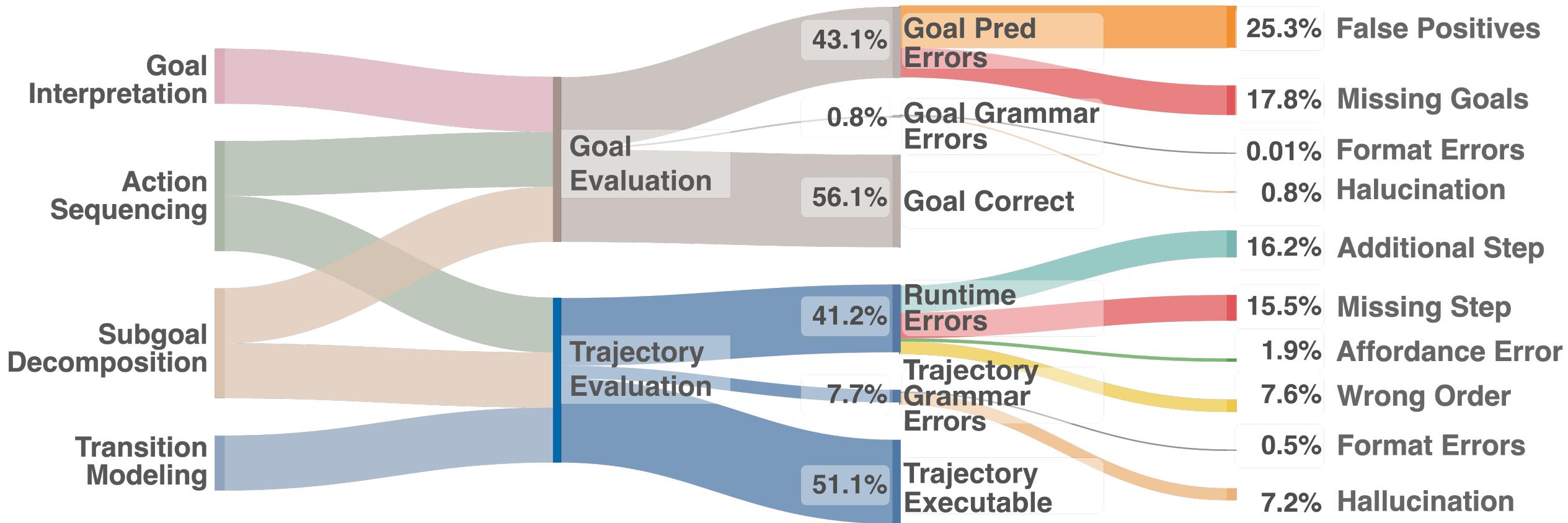
How about Trajectory Prediction?

43.2% Runtime Errors



How about Trajectory Prediction?

43.2% Runtime Errors



How about VLMs?

	Goal Interpretation (F1, %)	Action Sequencing (Success Rate, %)
Scene Graph HOLDS_LH(turkey) ON_TOP(bowl, table) ...	31.5	11.1
	9.1	2.5
 Scene Graph HOLDS_LH(turkey) ON_TOP(bowl, table) ...	25.8	11.0

Vision info are currently providing noise for Decision Making



EMBODIEDBENCH

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

Rui Yang^{1*}, Hanyang Chen^{1*}, Junyu Zhang^{1*}, Mark Zhao^{3*},
Cheng Qian¹, Kangrui Wang², Qineng Wang², Teja Venkat Koripella¹, Marziyeh Movahedi⁴,
Manling Li², Heng Ji¹, Huan Zhang¹, Tong Zhang¹

¹University of Illinois Urbana-Champaign, ²Northwestern University, ³University of Toronto, ⁴Toyota Technological Institute at Chicago

*Equal contribution

Hierarchical Action Representation

High-level: [{"action": "find a HandTowel"}, {"action": "pick up the HandTowel"}, ...]
 Low-level: [X, Y, Z, Roll, Pitch, Yaw, Gripper]

Tasks with various action levels

High-level Household

EB-ALFRED



Instruction: Put the books on the desk.

EB-Habitat



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

Low-level Navigation

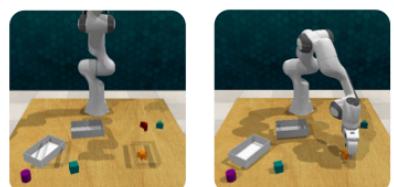
EB-Navigation



Instruction: Navigate to the laptop and stay close.

Low-level Manipulation

EB-Manipulation



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

EmbodiedBench

Vision-Driven Embodied Agents



Gemini



Qwen

Capability-oriented fine-grained evaluation



Base Capability



Place a plate with a spoon on a counter.



Common Sense



I'm *feeling thirsty and need a small container to hold water* ... Please navigate to that object and stay near it. (*cup*)



Complex Instruction



While you're tidying up, *find a spot on the back of the toilet to put two rolls of toilet paper*. It's a great place for convenient access ...



Spatial Awareness



Stack the *right* cylinder on top of the *left* moon.



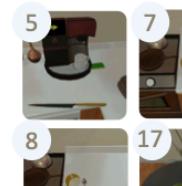
Visual Appearance



Obtain a *round red* fruit and put it in the *right* counter.



Long Horizon



Place a *cold apple slice* on the *table*.

Thank You

Outline



Content	Time	Presenter
1. Motivation and Overview	15min	Manling Li
2. Foundation Models meet Virtual Agents	45min	Manling Li
3. Foundation Models meet Physical Agents Overview & Perception High-level & Low-level Decision Making	25min 50min	Jiayuan Mao Wenlong Huang
Break		
4. Robotic Foundation Models	30min	Yunzhu
5. Remaining Challenges	15min	Yunzhu
QA	30min	