AAAI 2025 Tutorial T04
Time: 2025-02-25 8:30-12:30
Location: 118A Pennsylvania Convention Center

Foundation Models Meet Embodied Agents

Manling Li
Northwestern

Yunzhu Li
Columbia

Jiayuan Mao
MIT

Wenlong Huang
Stanford

Northwestern University · Columbia · MIT · Stanford University

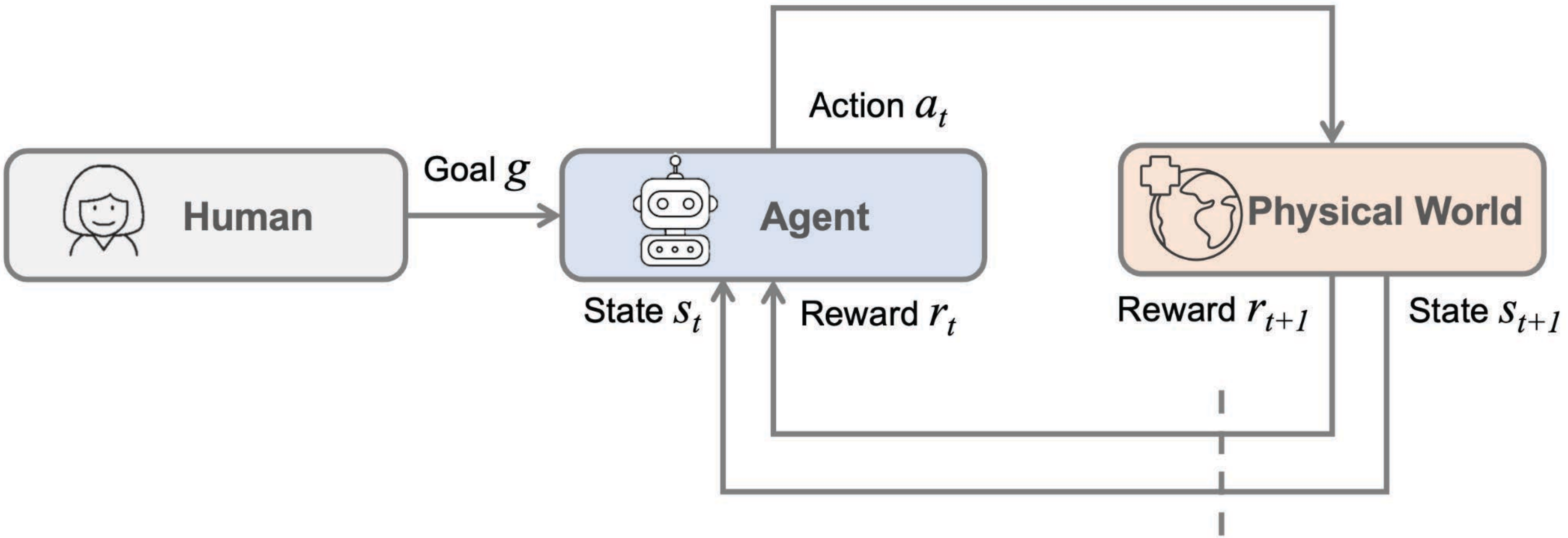
AAAI 2025 Tutorial T04
Time: 2025-02-25 8:30-12:30
Location: 118A Pennsylvania Convention Center

Part II: Foundation Models meet Virtual Agents

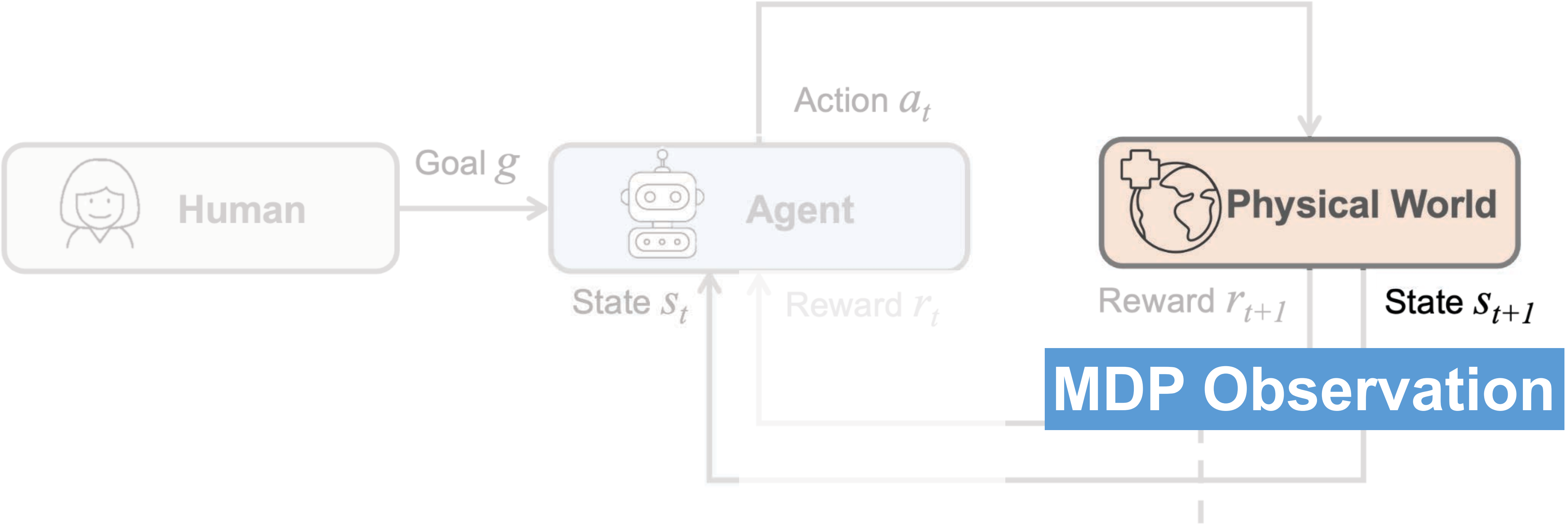
Manling Li, Assistant Professor at Northwestern University

Northwestern University
COLUMBIA
MIT
Stanford University

Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



An Envisioned Agent Society

https://arxiv.org/pdf/2309.07864

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

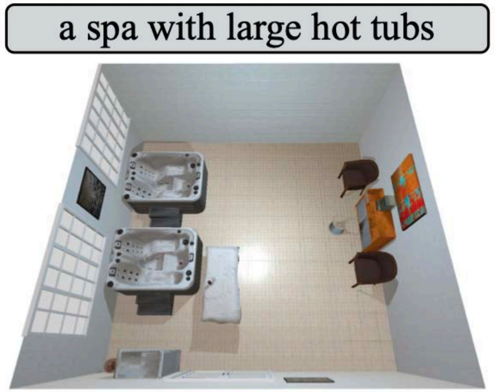
Simulator	Year	Scenes	Modalities
iGibson	2021	15	RGB-D, LiDAR, Learning
SAPIEN	2020	46	RGB-D, Joint Object Interaction
Habitat	2019	1000	RGB-D, Supports Multi-Agent
Matterport 3D	2018	90	RGB-D, Navigation Benchmark
Virtual Home	2018	50	RGB-D, Environment Graph
AI2-THOR ALFRED	2017	120	RGB-D, Supports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

MDP Environment: Embodied Simulator

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?

Existing Data & Annotations

BEHAVIOR

Behavior-100

	VirtualHome	BEHAVIOR
#task name	26	100
#task instruction	338	100
#goal	801	673
- #state	340	153
- #relation	299	520
- #action	162	-
#trajectory	338	100
- #step	2960	1460
- avg. step	8.76	14.6
#transition model	33	30
- #precondition	99	84
- #effect	57	51

Cleaning Fridge

Organizing File Cabinet

...

Serving a Meal

Watering Houseplants

100 Household Tasks

Relevant Objects:
```
{'soap.0', 'bottom_cabinet_no_top.0', 'sink.0', 'fridge.0',
'bowl.0', 'rag.0', 'rag.1', 'soda.0', 'soda.1', ......}
```

Initial Conditions:
```
stained(fridge.0)
ontop(rag.0, bottom_cabinet_no_top.0).
inside(soda.0, fridge.0)
inside(soda.1, fridge.0)
......
```

Goal Conditions:
```
not stained(fridge.0)
not inside(soda.0, fridge.0)
......
```

Demo Video:

cleaning_freezer.m4v

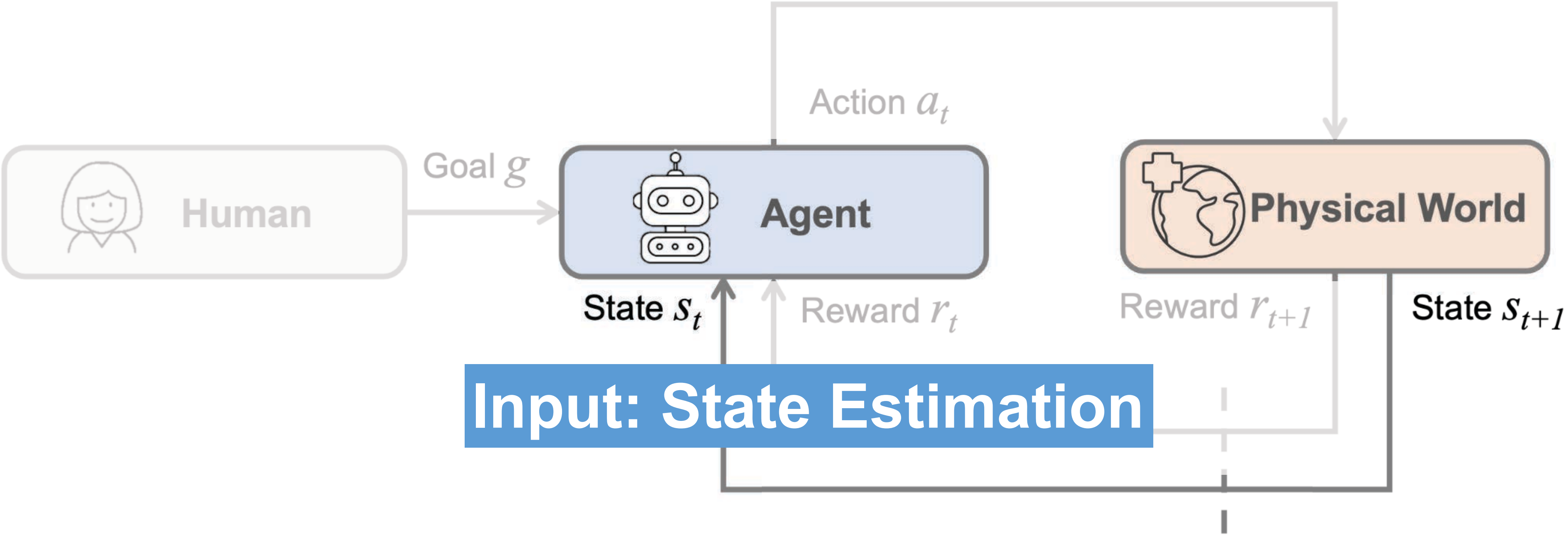
Natural Language Task Instructions:
```
"Remove all the soda from the fridge. Use the rag and soap to clean
the fridge. When you are done, leave the rag next to the sink and the
soap in the sink."
```

BDDL Action Sequence:
```
[{'action': 'OPEN', 'object': 'bottom_cabinet_no_top.0'},
 {'action': 'RIGHT_GRASP', 'object': 'soap.0'},
 {'action': 'RIGHT_PLACE_INSIDE', 'object': 'sink.0'},
 ......
 {'action': 'CLEAN', 'object': 'fridge.0'},
 {'action': 'LEFT_RELEASE', 'object': 'soda.1'},
 {'action': 'CLOSE', 'object': 'fridge.0'}]
```

Transition Model:
```
(:action soak
     :parameters(?obj1 - object ?sink - sink_n_01 ?agent - agent)
     :precondition(and (holding ?obj1)(in_reach_of_agent ?sink)
            (toggled_on ?sink))
     :effect (soaked ?obj1))
......
```

Additional Human Annotations

Let us go back to MDPs (Markov Decision Processes)

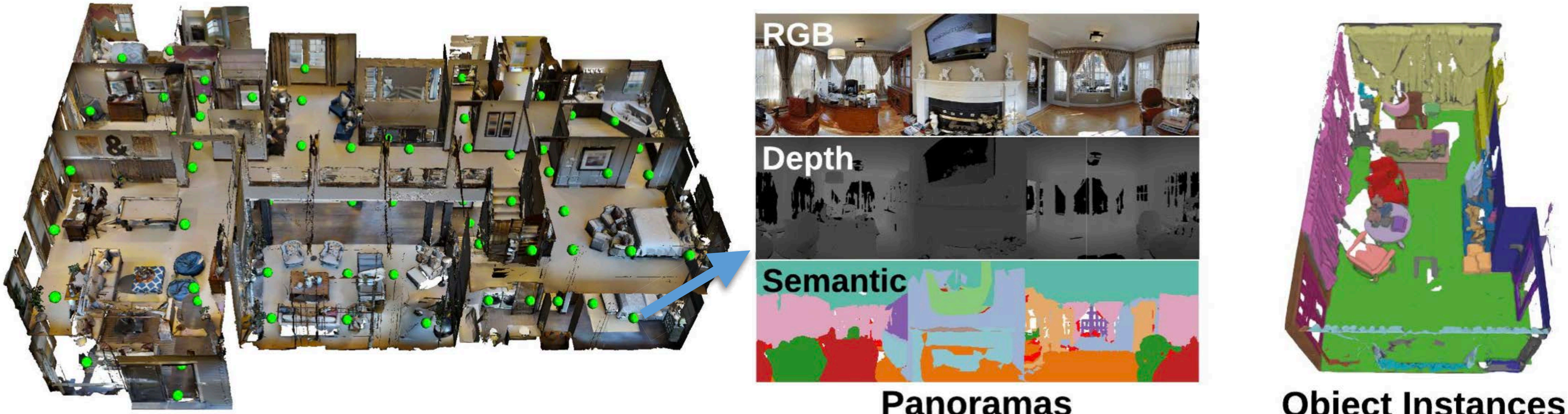


Perception / State Estimation

$$o \rightarrow s$$

MDP Environment: Matterport3D

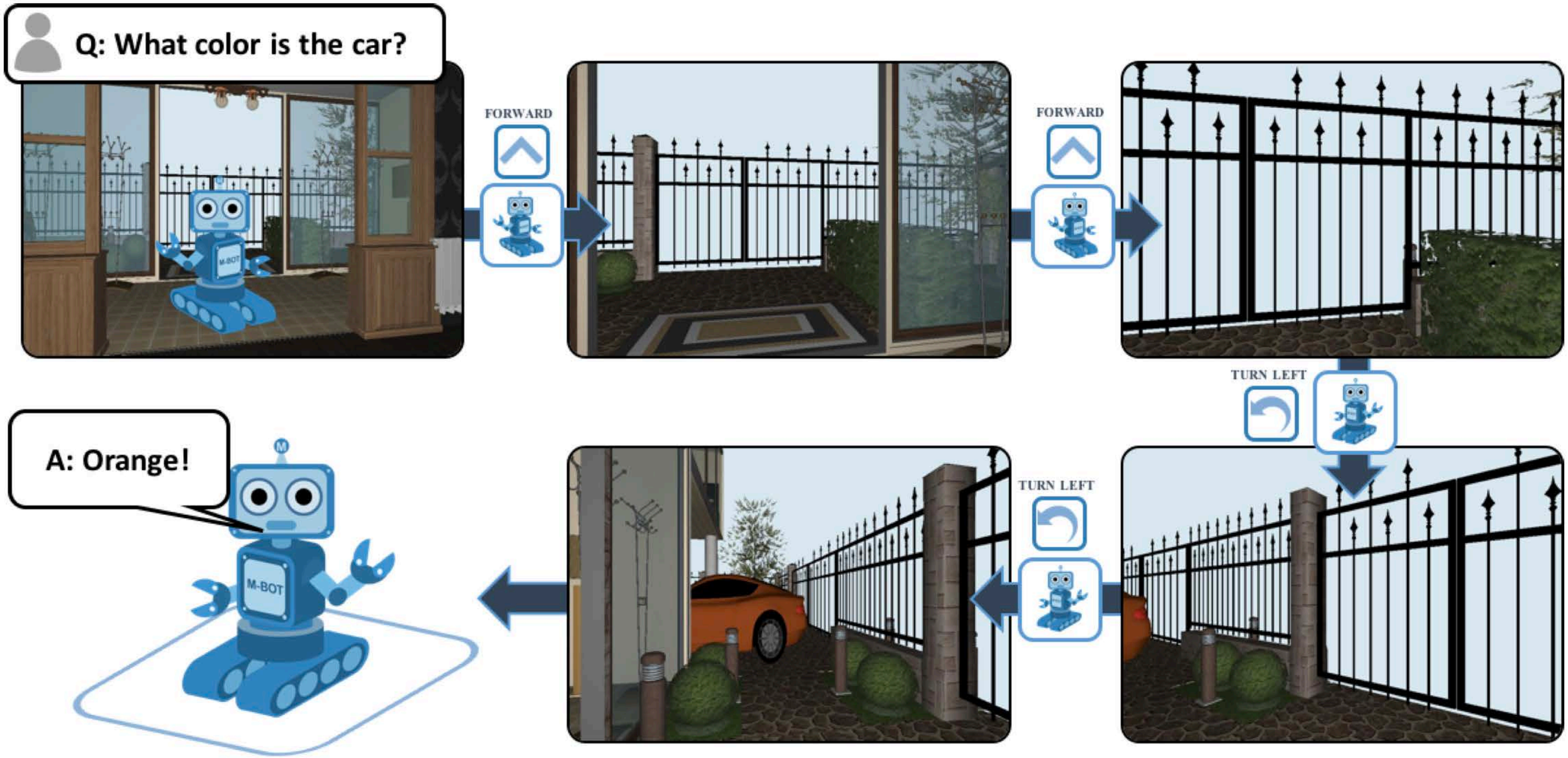
- A large-scale RGB-D dataset containing 10,800 panoramic views from 194,400 RGB-D images of 90 building-scale scenes



Textured 3D Mesh **Panoramas** **Object Instances**

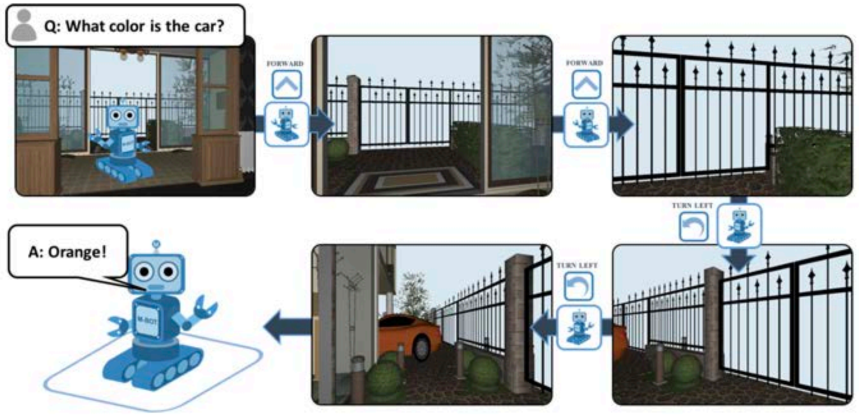
https://arxiv.org/pdf/1709.06158

Embodied QA



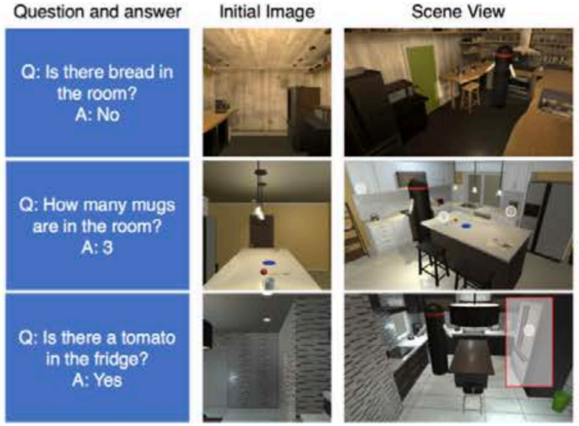
EmbodiedQA: Das et. al., 2018

EQA: Embodied Question Answering



EmbodiedQA
(Das et al., 2018)

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



Interactive QA
(Gordon et al., 2018)

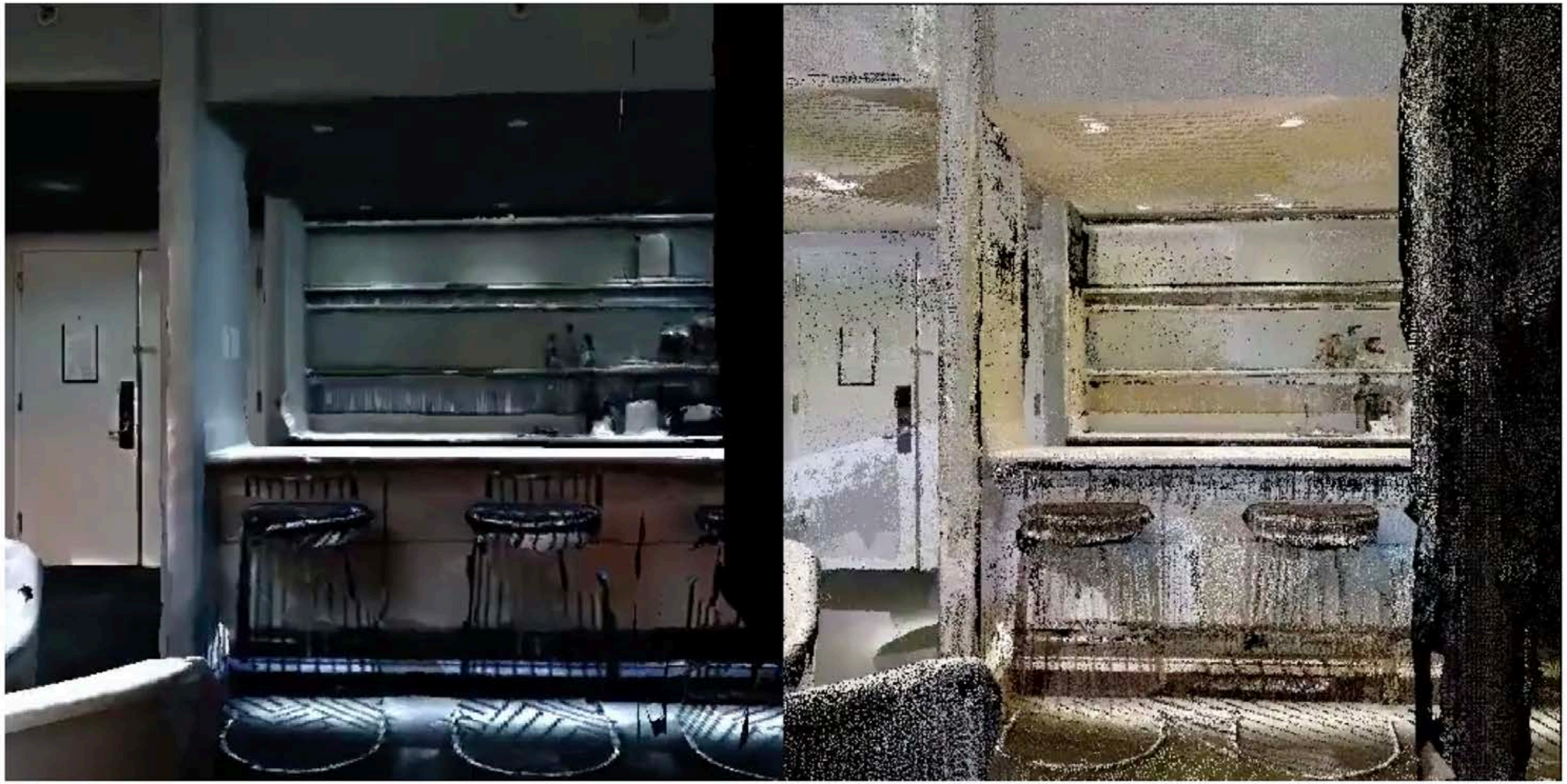
Vision-Language Navigation
(Anderson et al., 2018)

https://embodiedqa.org/slides/eqa_matterport.slides.pdf

MDP Observation: 3D Integration

- Two types of perception abilities:

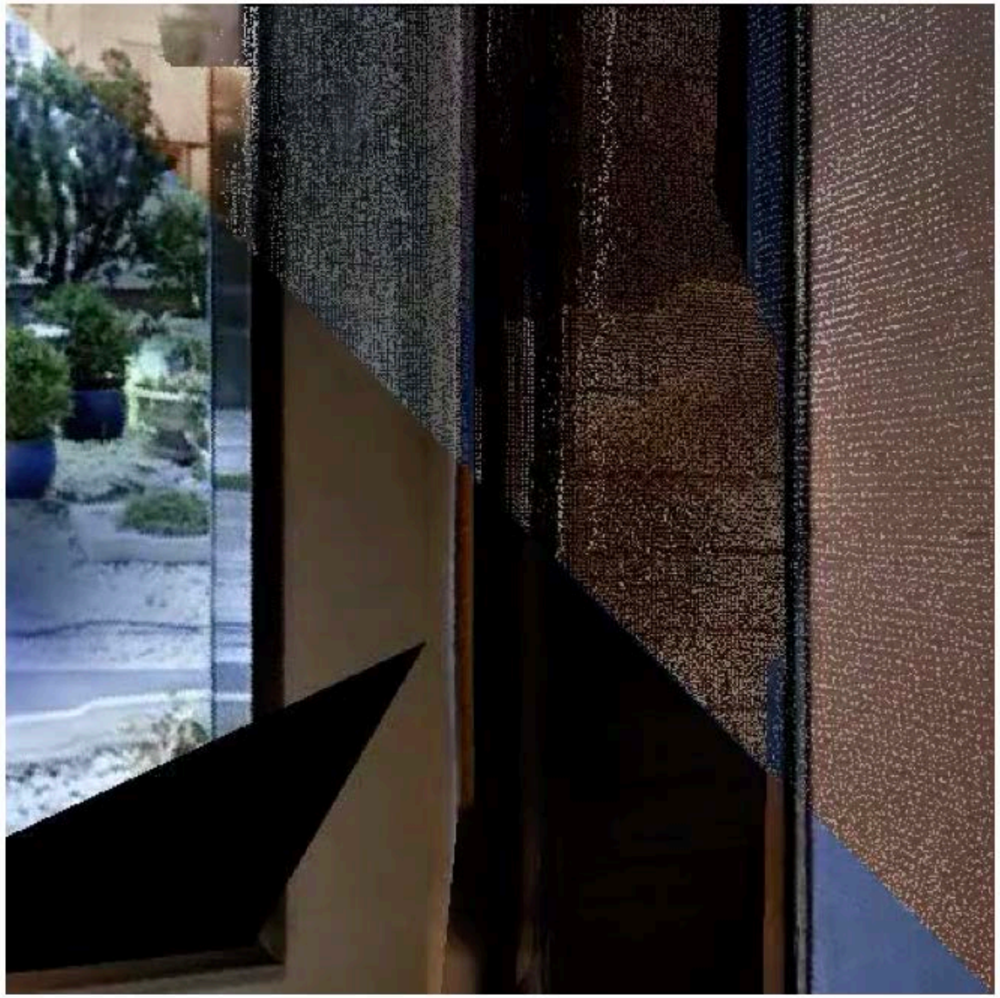
RGB Perception | Point Cloud Perception



https://embodiedqa.org/slides/eqa_matterport.slides.pdf

MDP Observation: 3D Integration

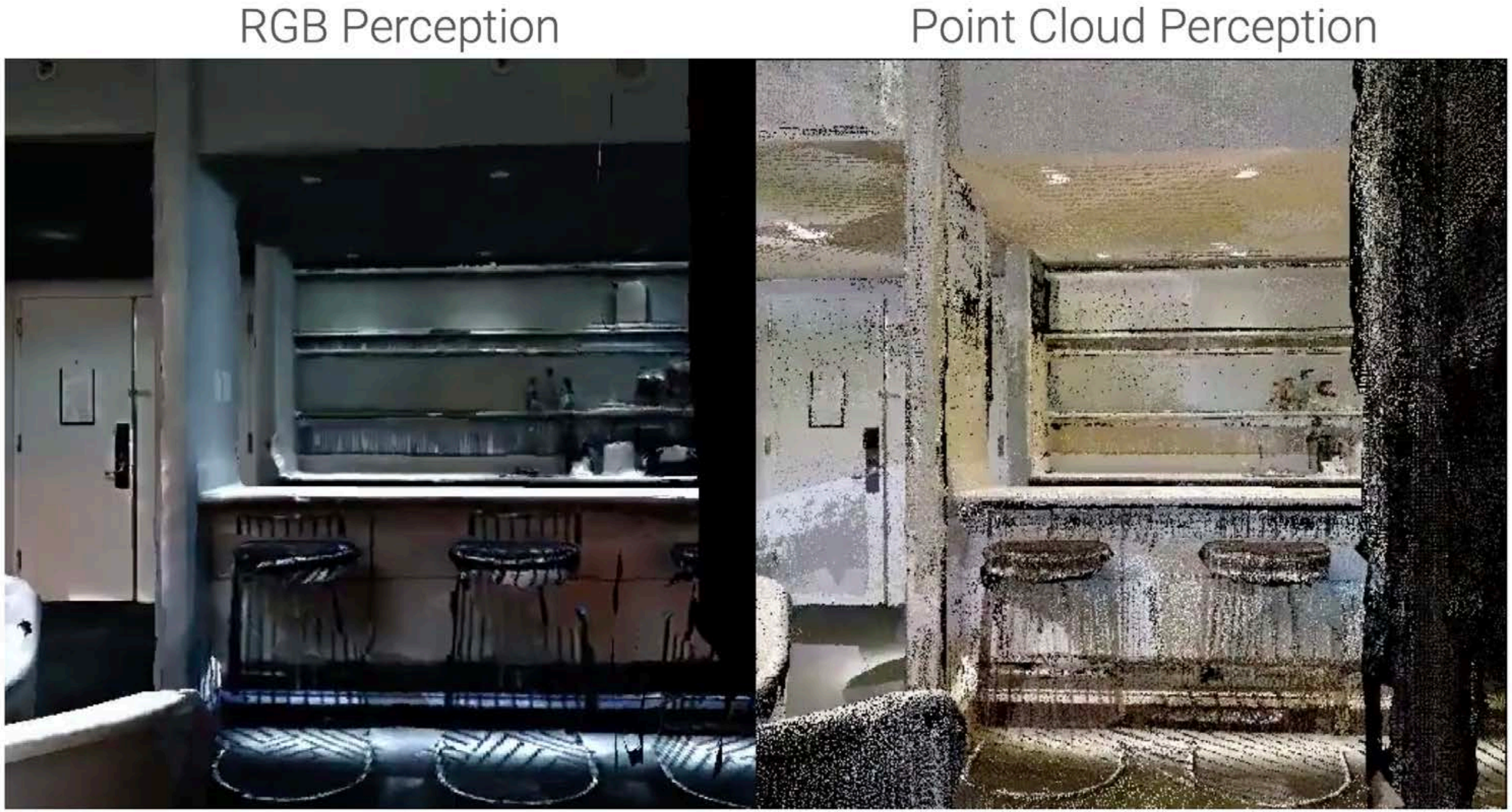
Q: What color is the counter in the hallway?



https://embodiedqa.org/slides/eqa_matterport.slides.pdf

MDP Observation: 3D Integration

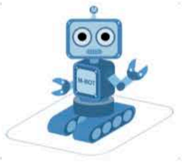
- Two types of perception abilities:

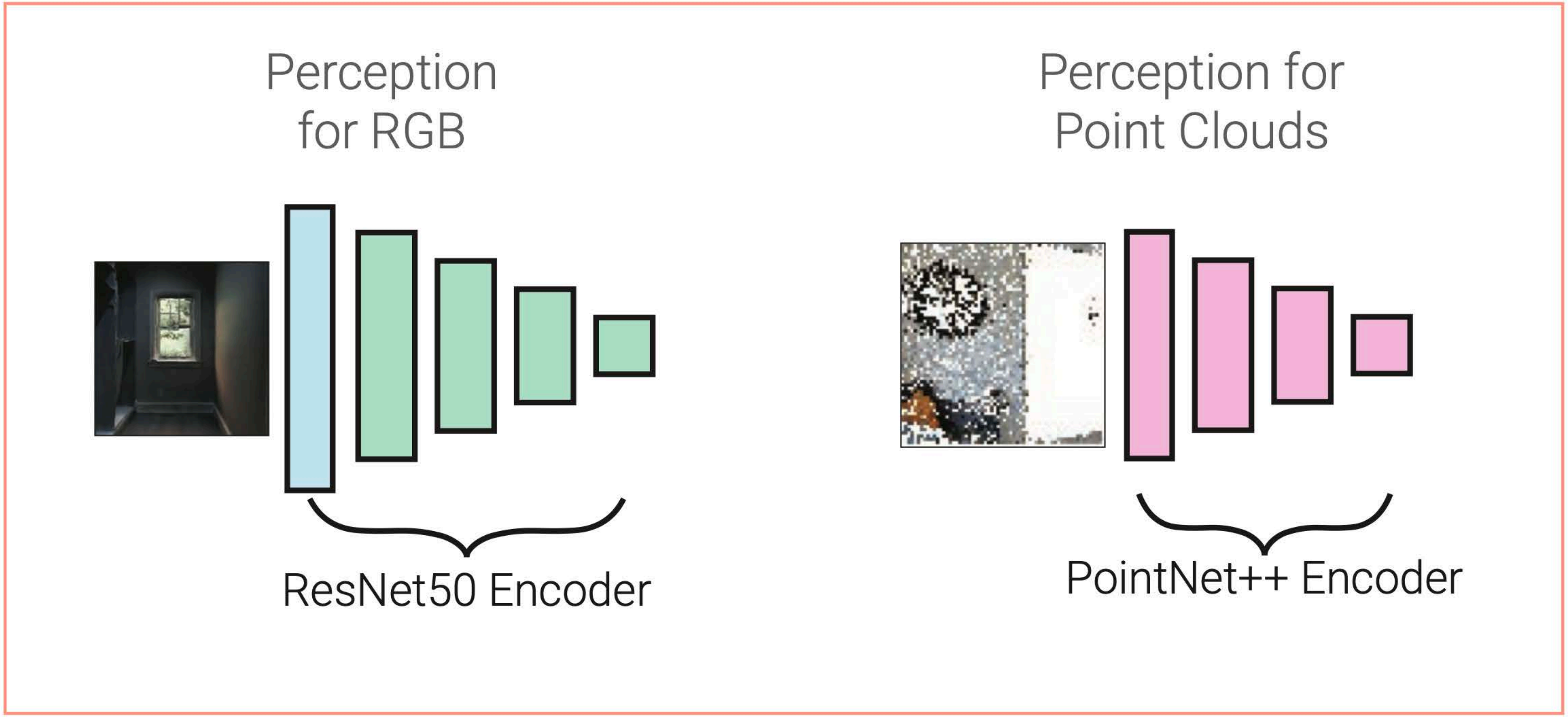


https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Agent Perception / State Estimation

Vision





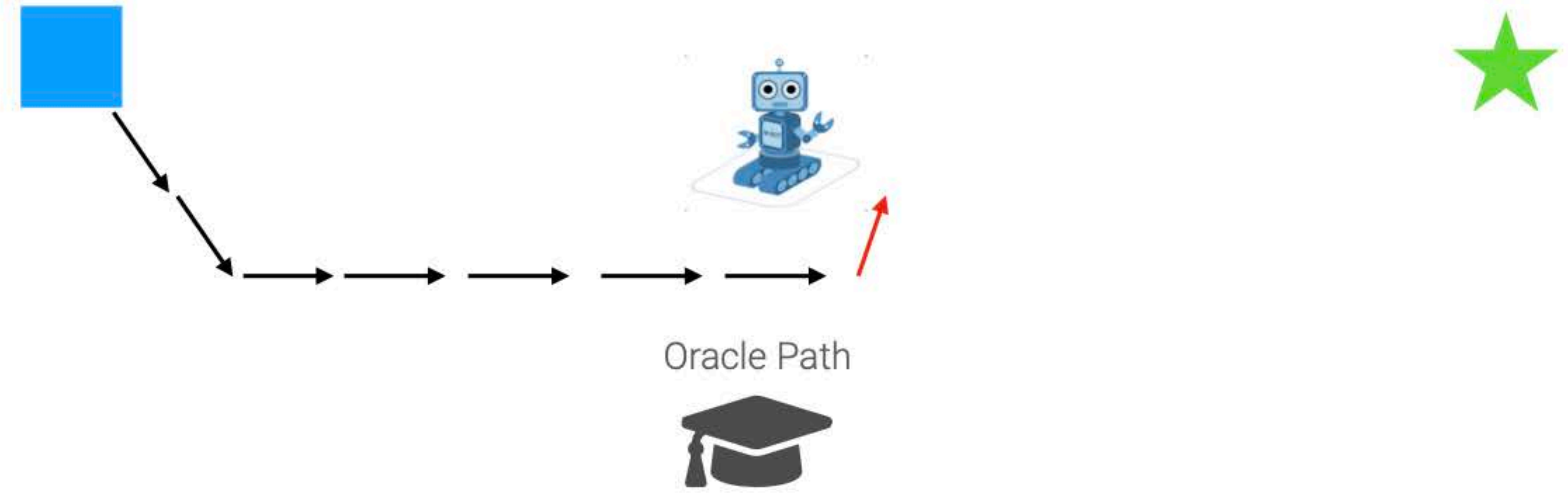
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Agent Navigation for Embodied QA

Visual Variations { Blind, RGB, PC, PC+RGB } × Language Variations { No Question, Question } × Memory Variations { Reactive, Memory }

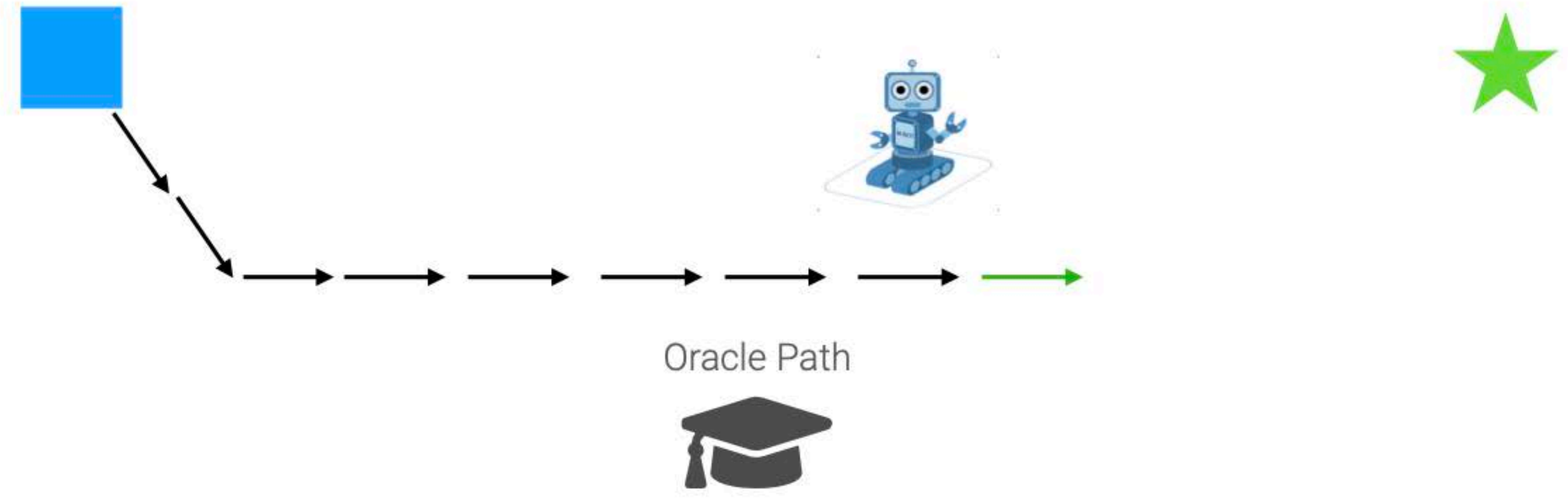
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Training Agents for Navigation



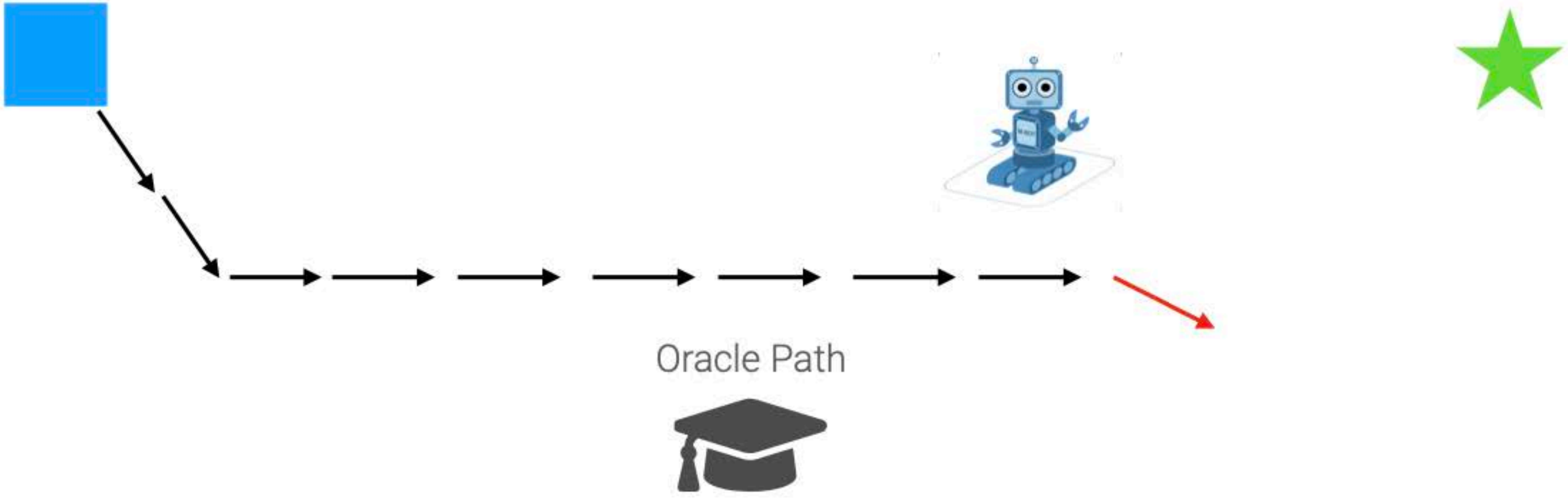
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Training Agents for Navigation



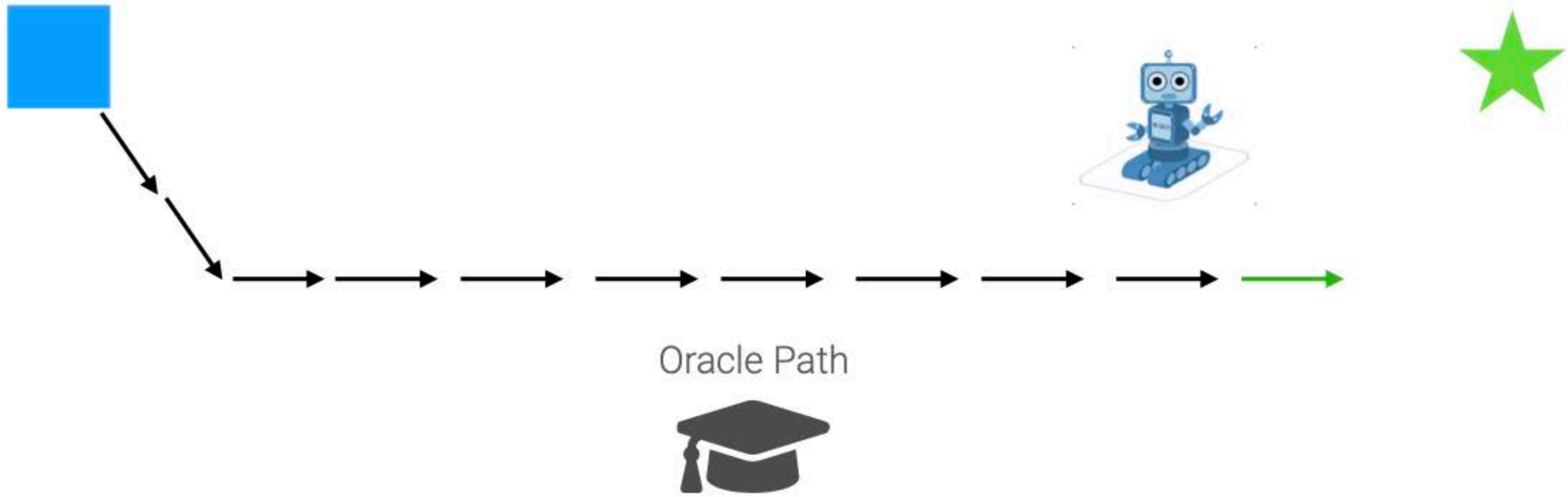
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Training Agents for Navigation



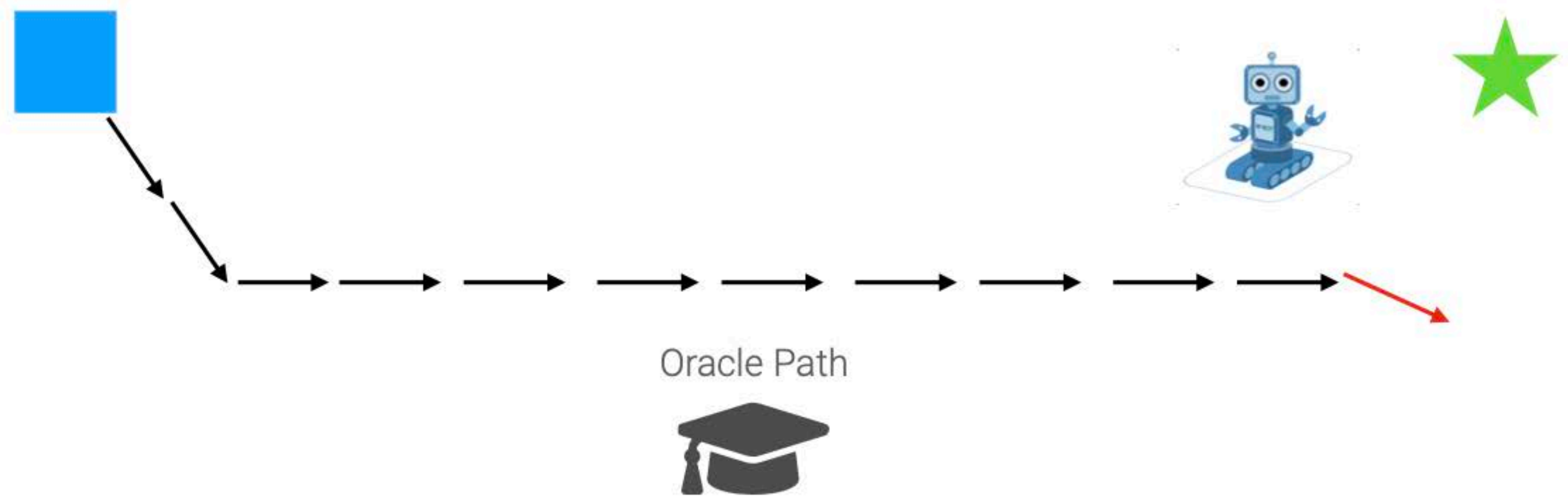
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Training Agents for Navigation



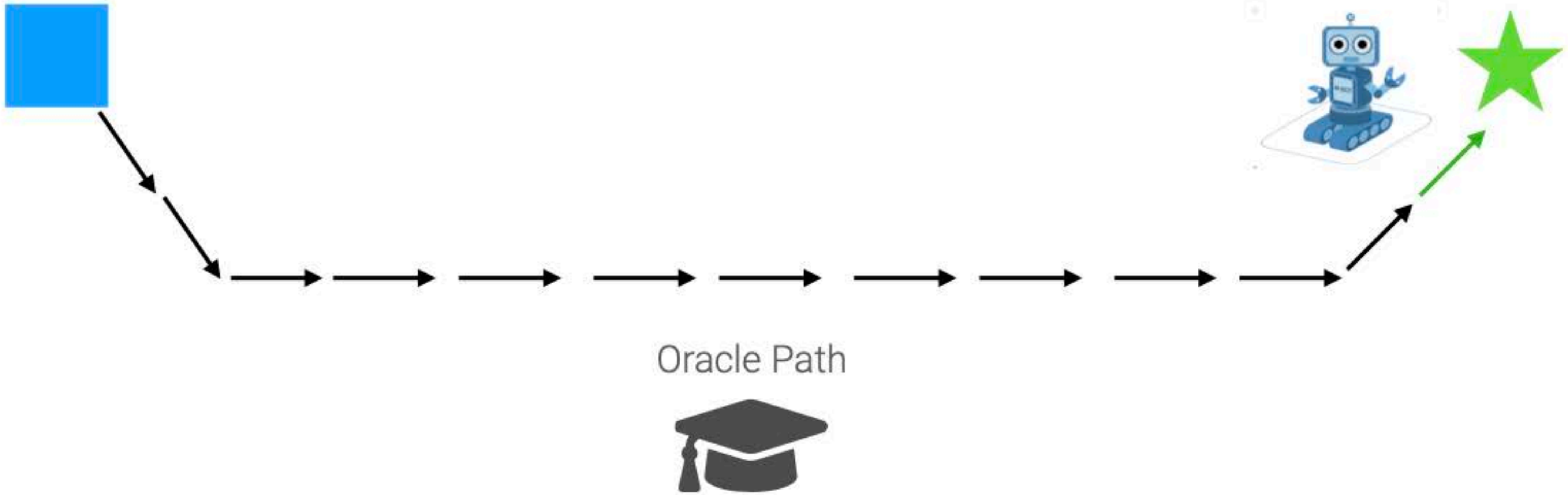
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Training Agents for Navigation



https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Training Agents for Navigation



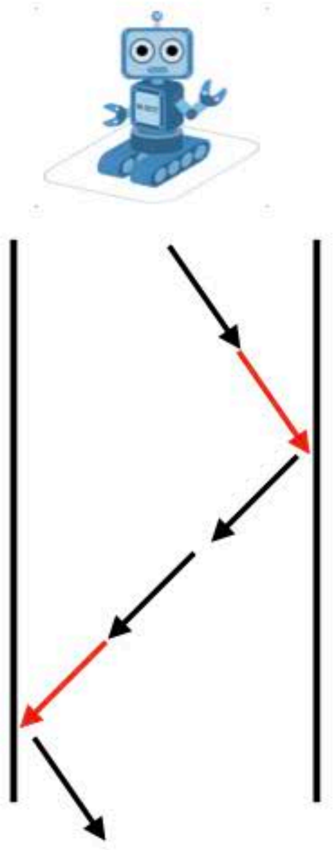
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Extension to Real-World: OpenEQA

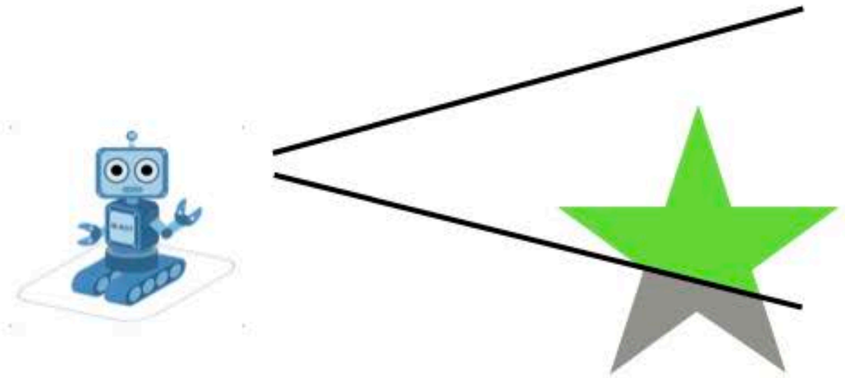
https://open-eqa.github.io/

Previous Metrics

Collision Rate (↓ better)



View Quality (↑ better)

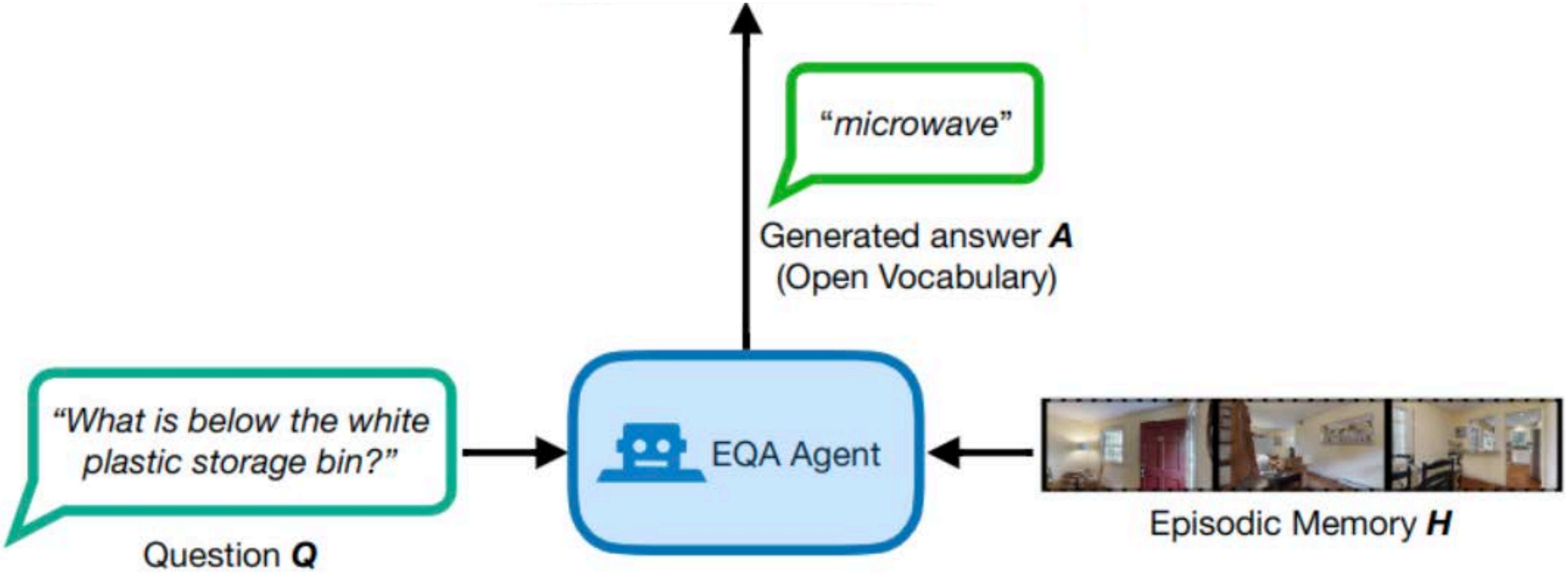


https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Now Foundation Models are Scorers

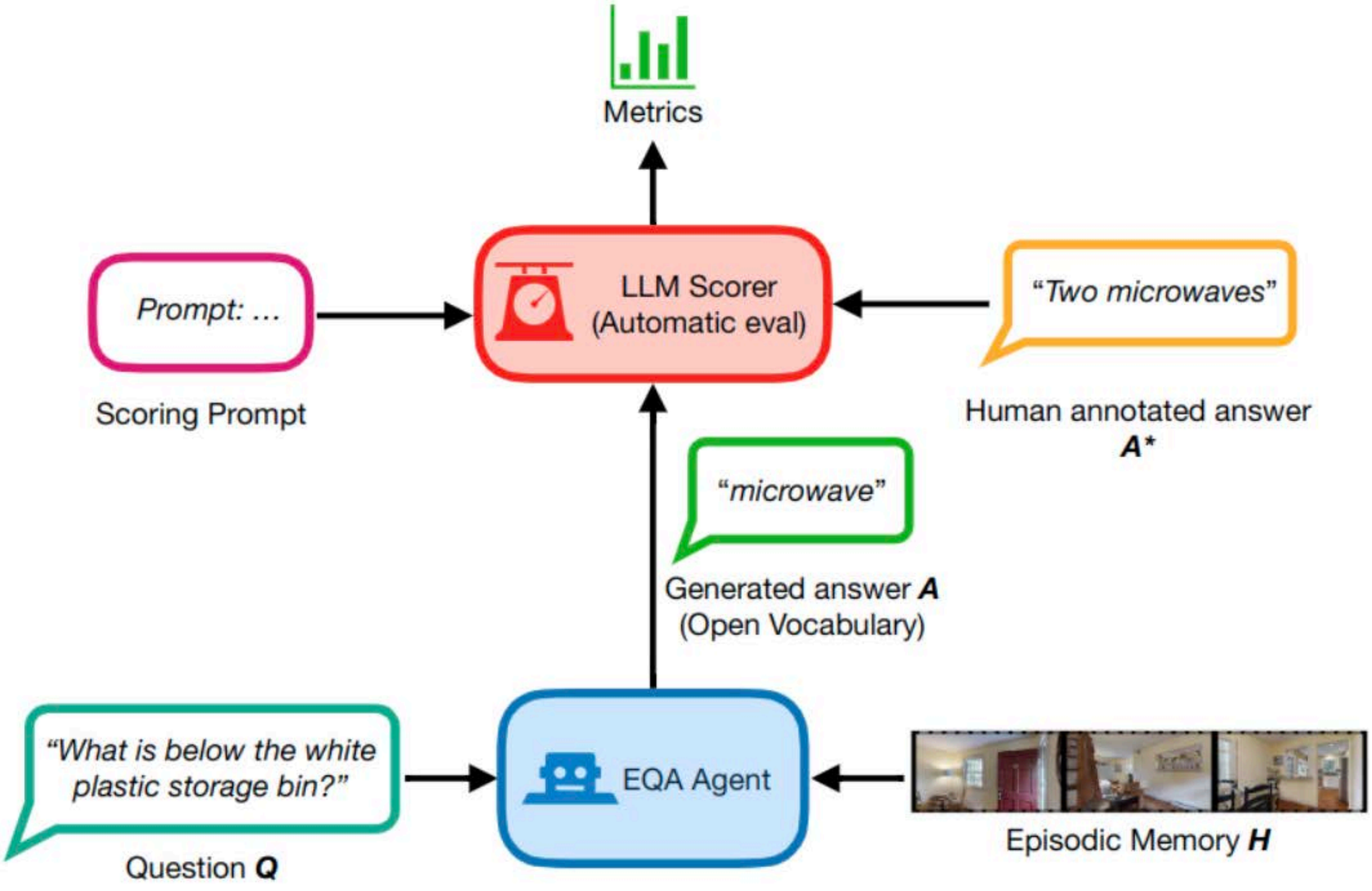
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Now Foundation Models are Scorers



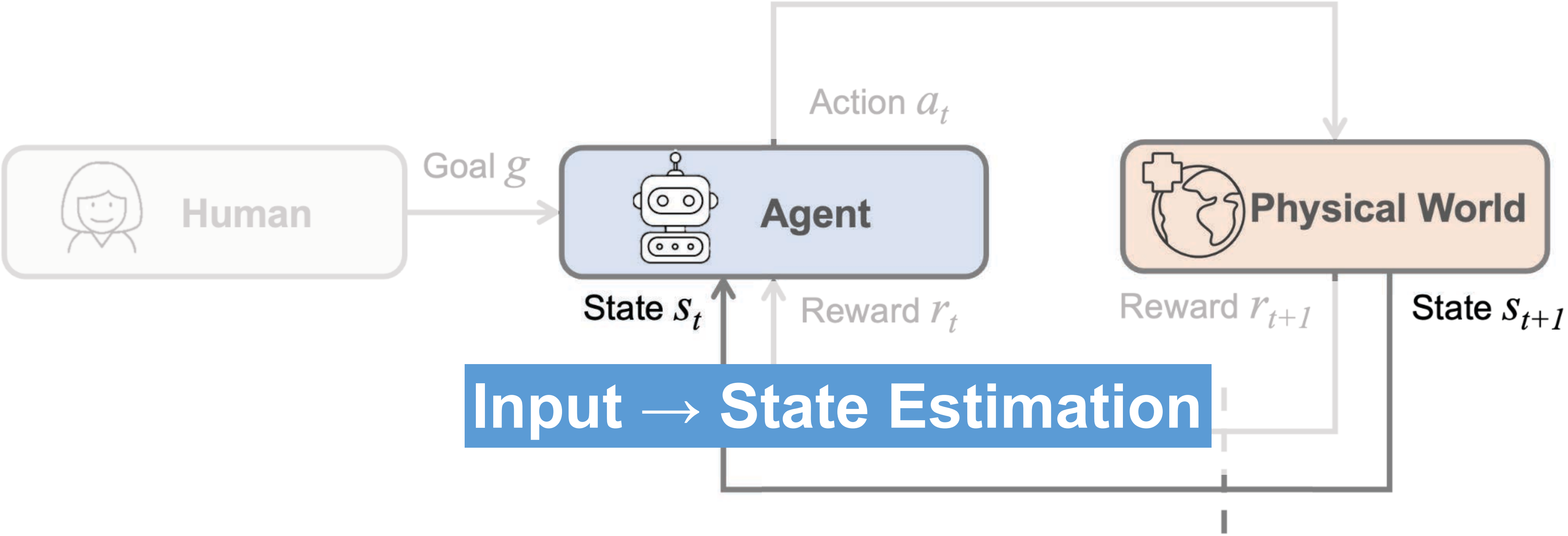
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

New Eval Trend: Now Foundation Models are Scorers



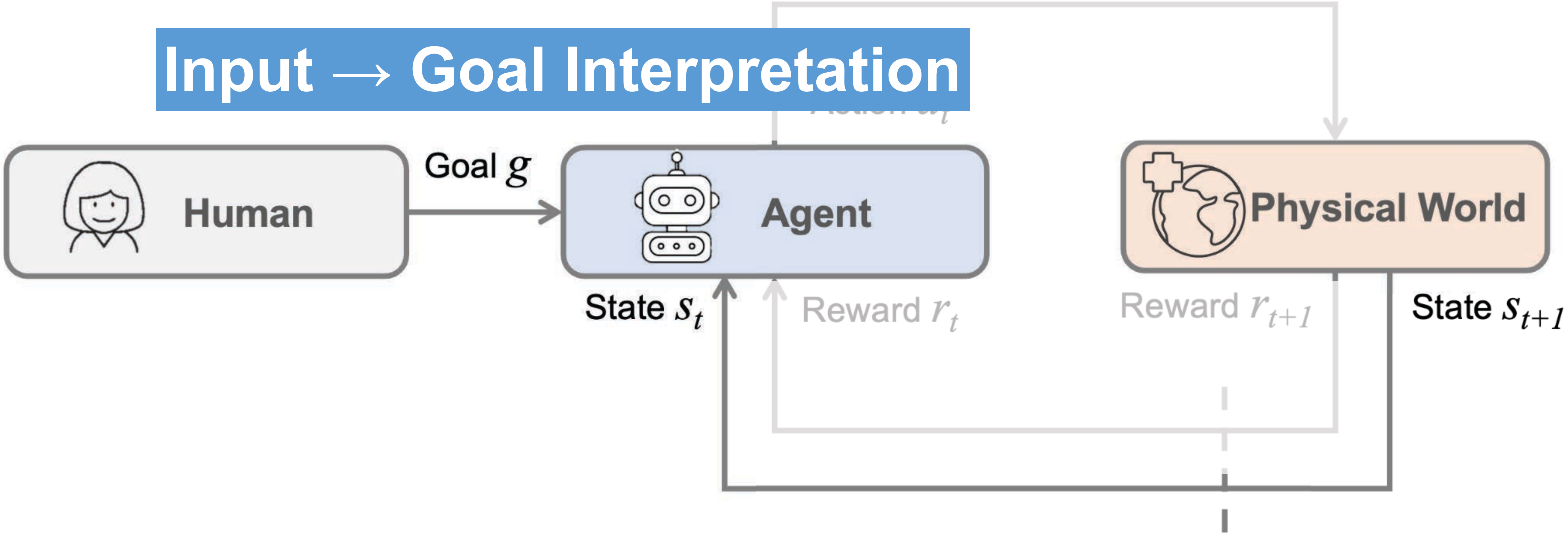
https://embodiedqa.org/slides/eqa_matterport.slides.pdf

Let us go back to MDPs (Markov Decision Processes)



Input → State Estimation

Input → Goal Interpretation

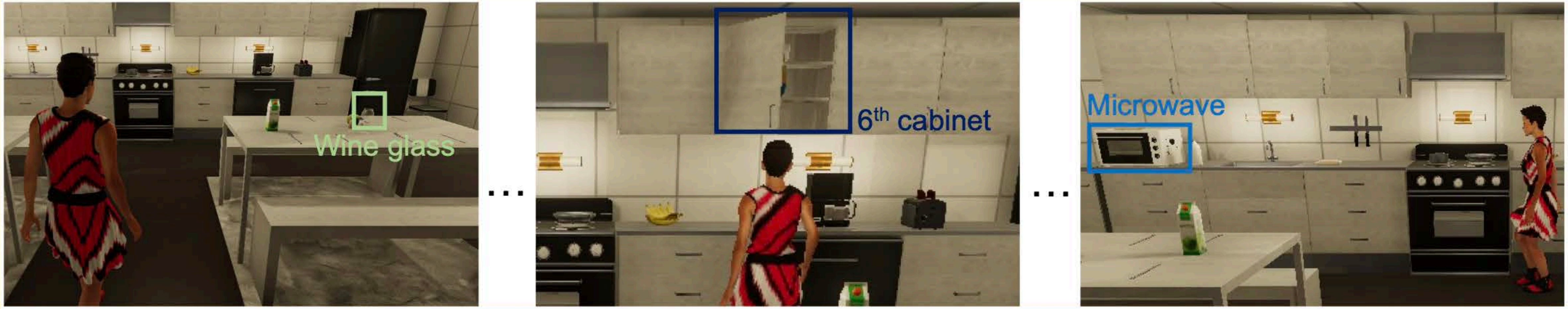


Goal Interpretation

g

Multimodal Theory of Mind

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. ✖

https://arxiv.org/pdf/2401.08743

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.

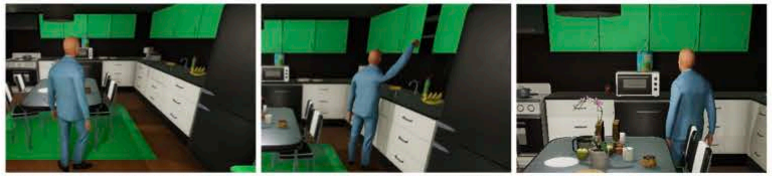
Question: If Elizabeth has been trying to get a bottle of wine, which one of the following statements is more likely to be true?
(a) Elizabeth thinks that there is a bottle of wine inside the fridge.
(b) Elizabeth thinks that there isn't any bottle of wine inside the fridge.

Type 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: … 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 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 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

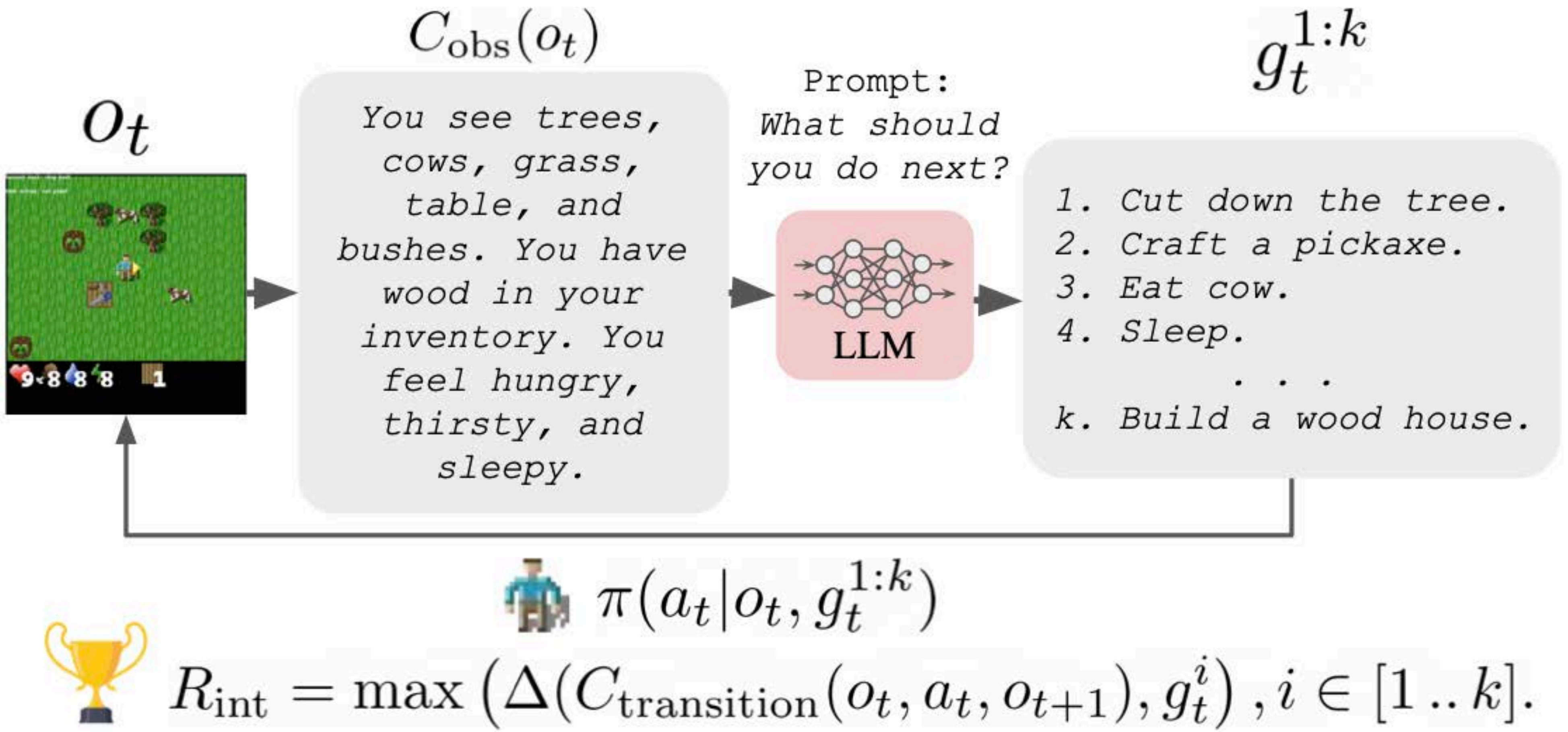
Scene: … The dishwasher holds a dish bowl… The first cabinet from the left holds a bag of chips and a wine glass… The fifth cabinet has an apple…
Actions: Williams… advances towards the first kitchen cabinet, opens it, and then shuts it. He then moves towards the fifth kitchen cabinet.

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.

https://arxiv.org/pdf/2401.08743

Goal Prediction

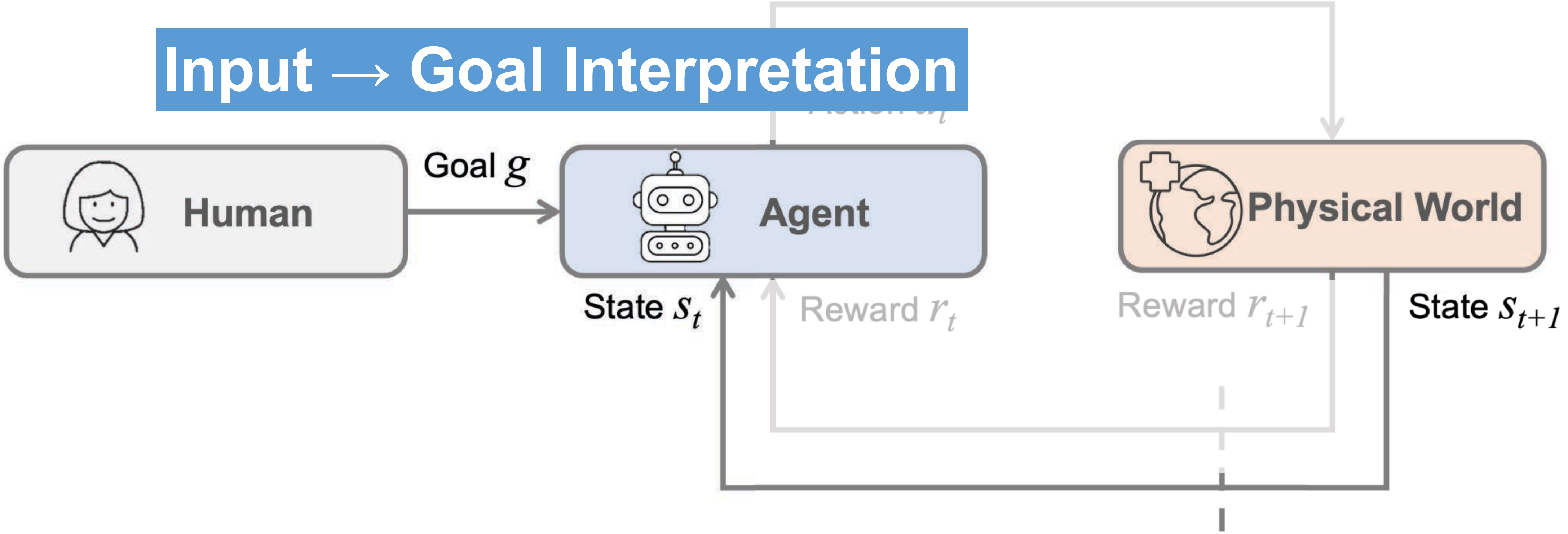
- Learn to predict plausibly useful goals in a task-agnostic way



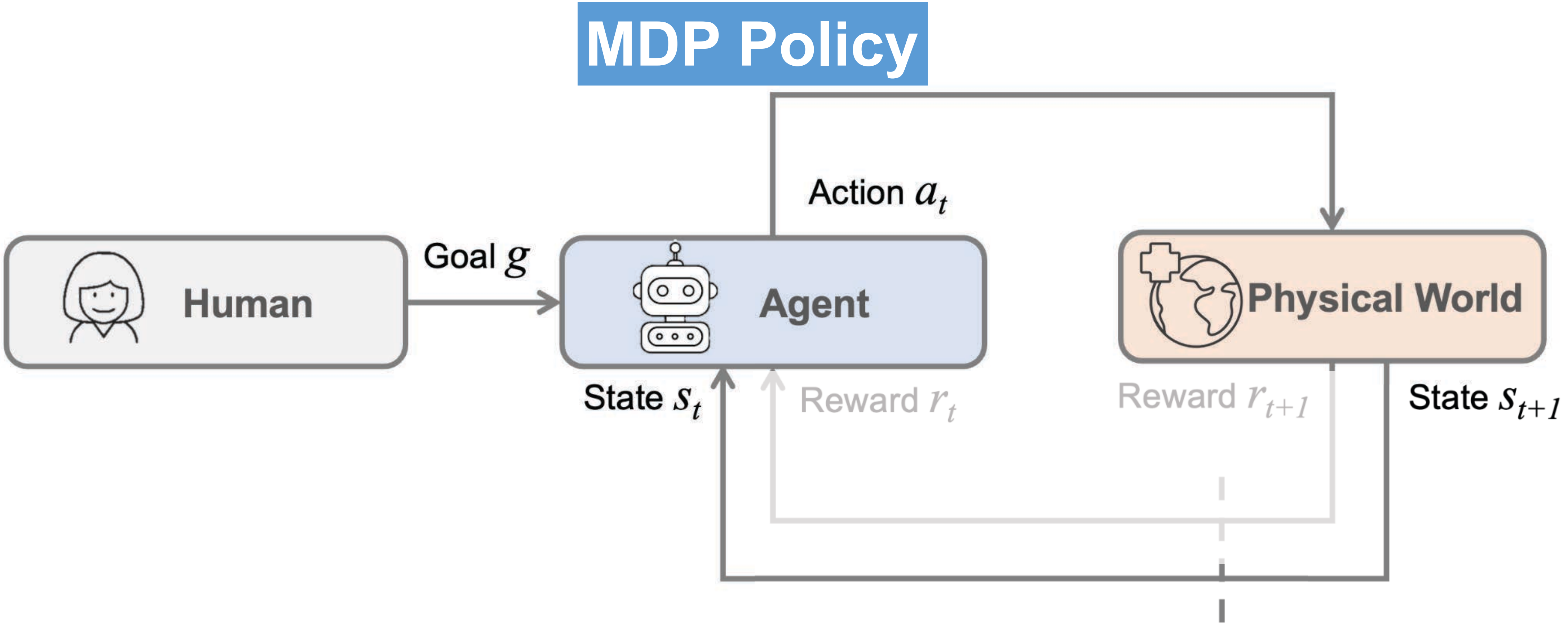
https://proceedings.mlr.press/v202/du23f/du23f.pdf

Let us go back to MDPs (Markov Decision Processes)

Input → Goal Interpretation



Let us go back to MDPs (Markov Decision Processes)

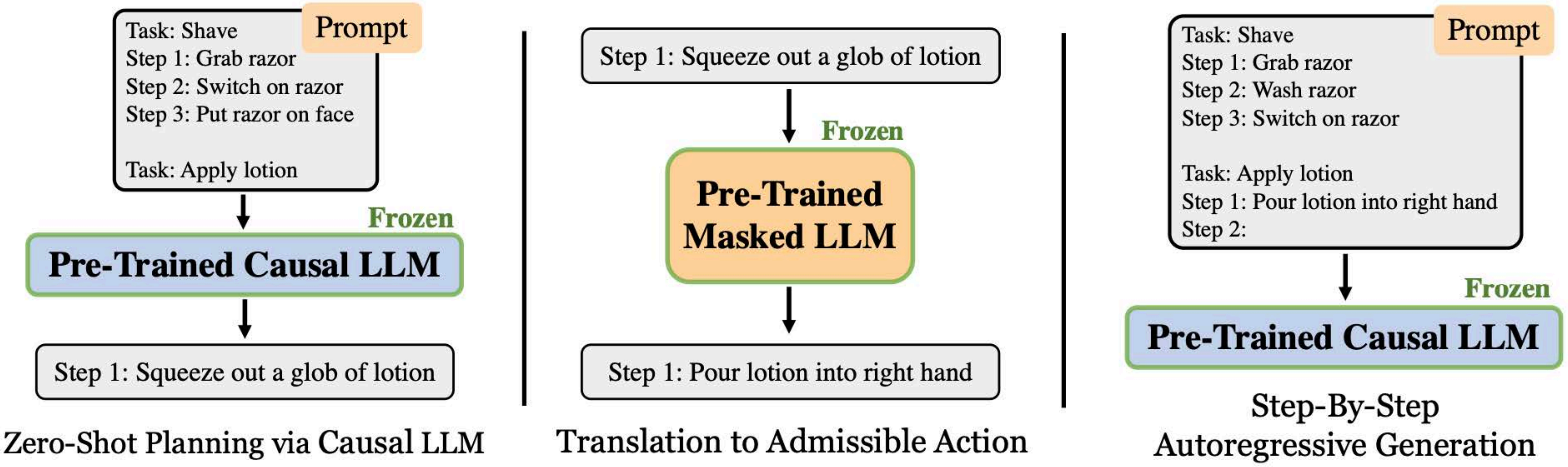


Policy

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

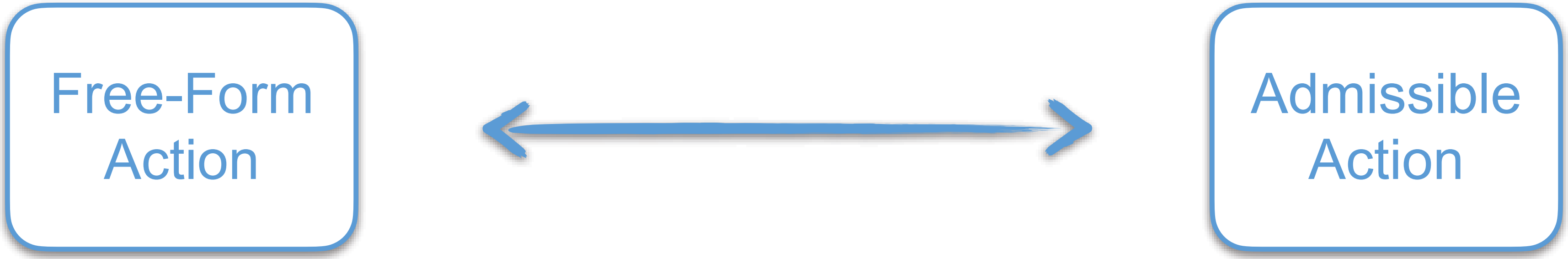
First Work: Language Models as Zero-Shot Planners

- LLMs as the Planner



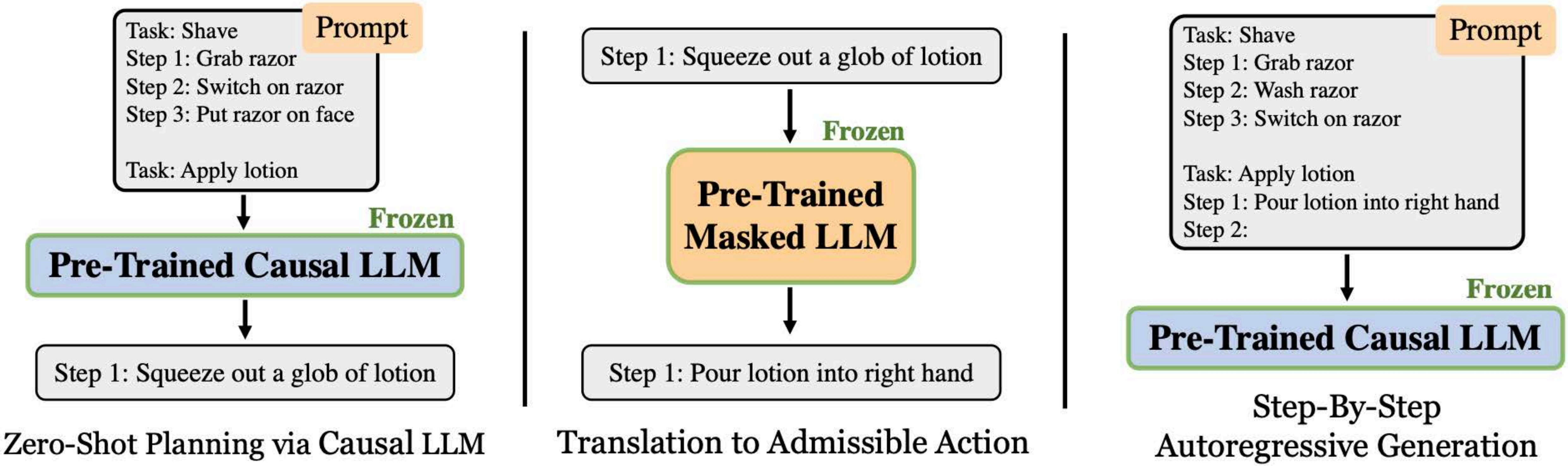
https://wenlong.page/language-planner/

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



First Work: Language Models as Zero-Shot Planners

- LLMs as the Planner

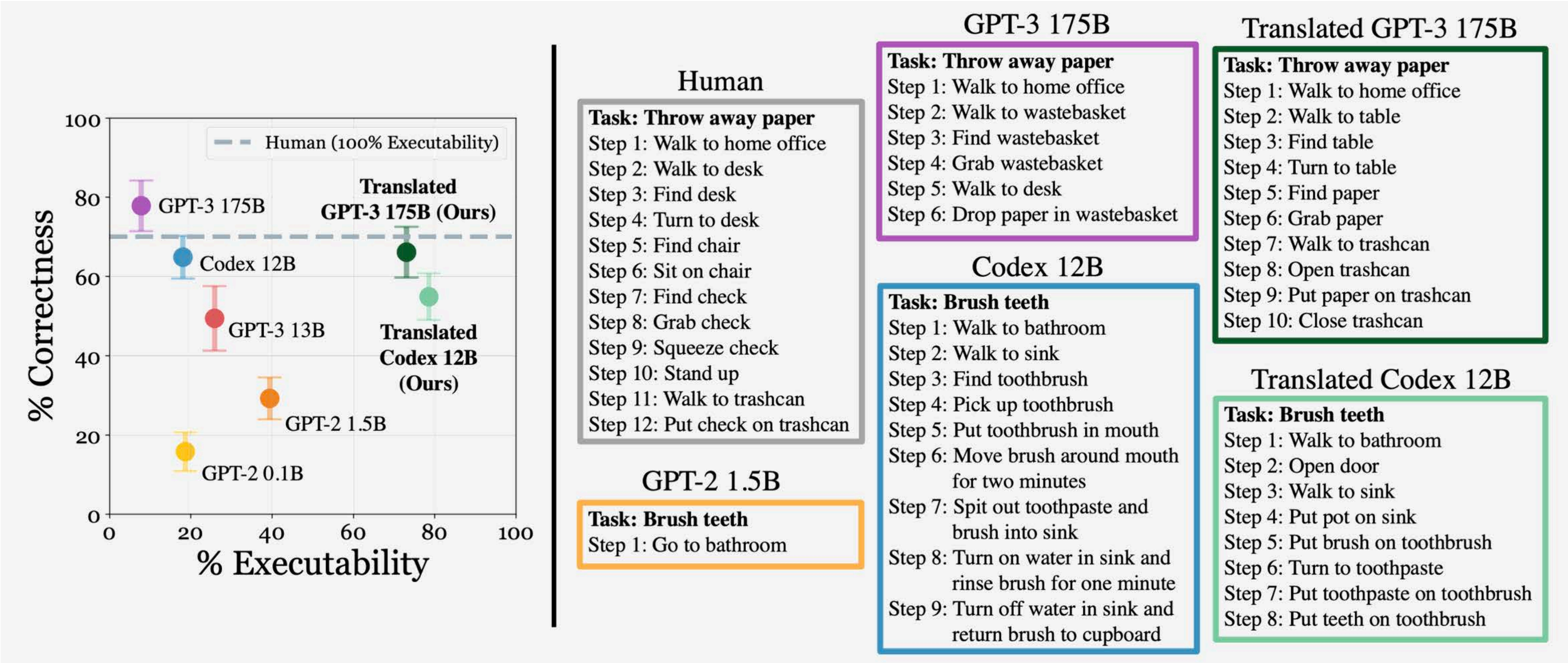


https://wenlong.page/language-planner/

First Work: Language Models as Zero-Shot Planners

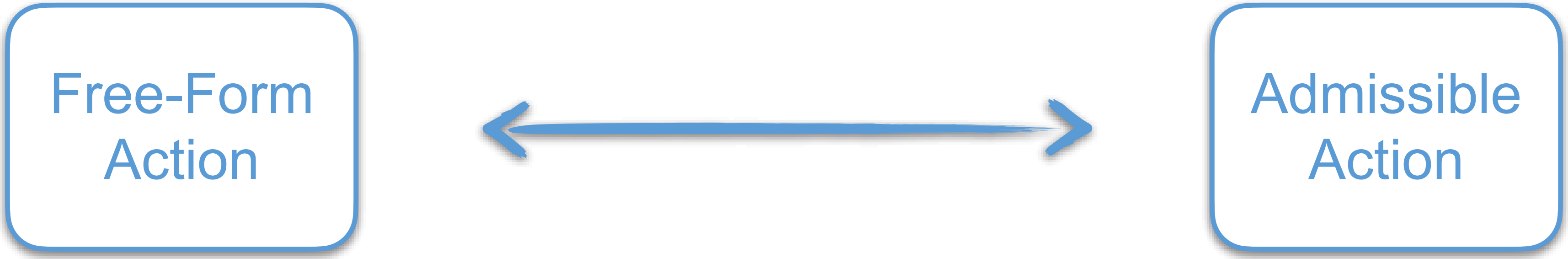
https://wenlong.page/language-planner/

First Work: Language Models as Zero-Shot Planners

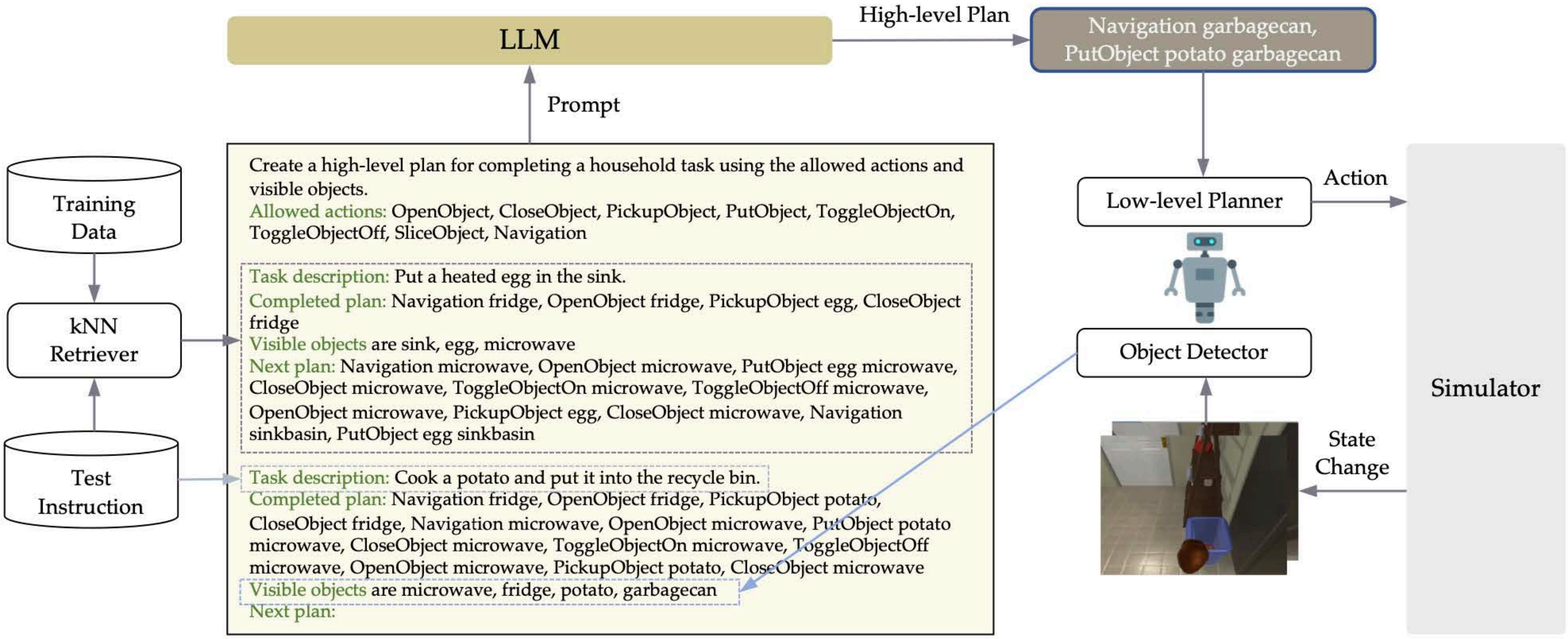


https://wenlong.page/language-planner/

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



Another way: Retrieval Based

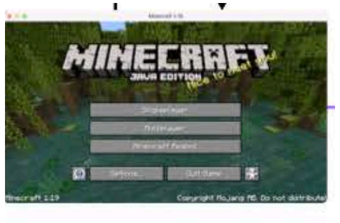


https://arxiv.org/pdf/2310.09971

Learn the Skill Library (Admissible Action)

Voyager: Policy & Skill Acquisition

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

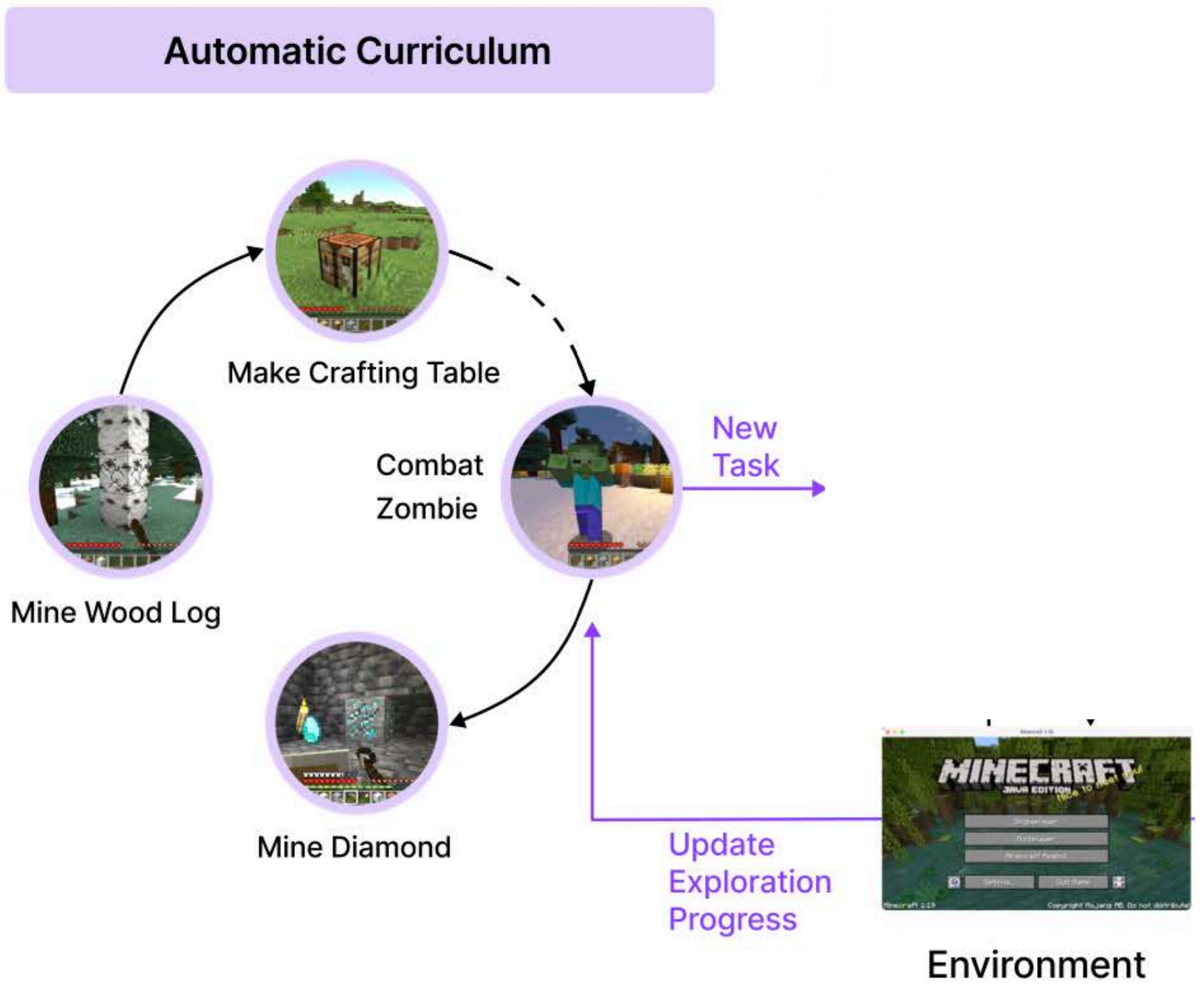


Environment

https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

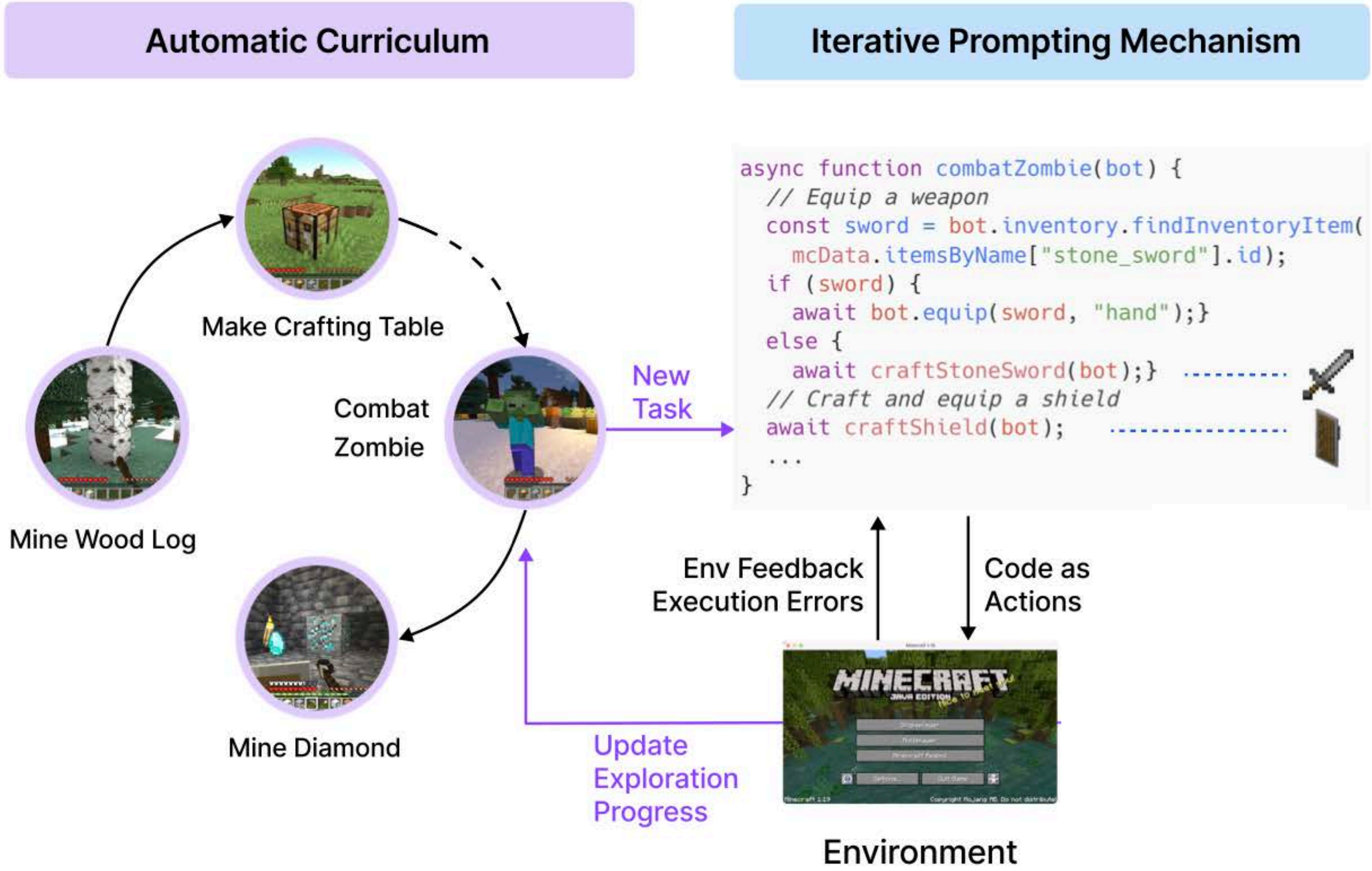
- Focuses on agent policies for exploration and skill acquisition in Minecraft



https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

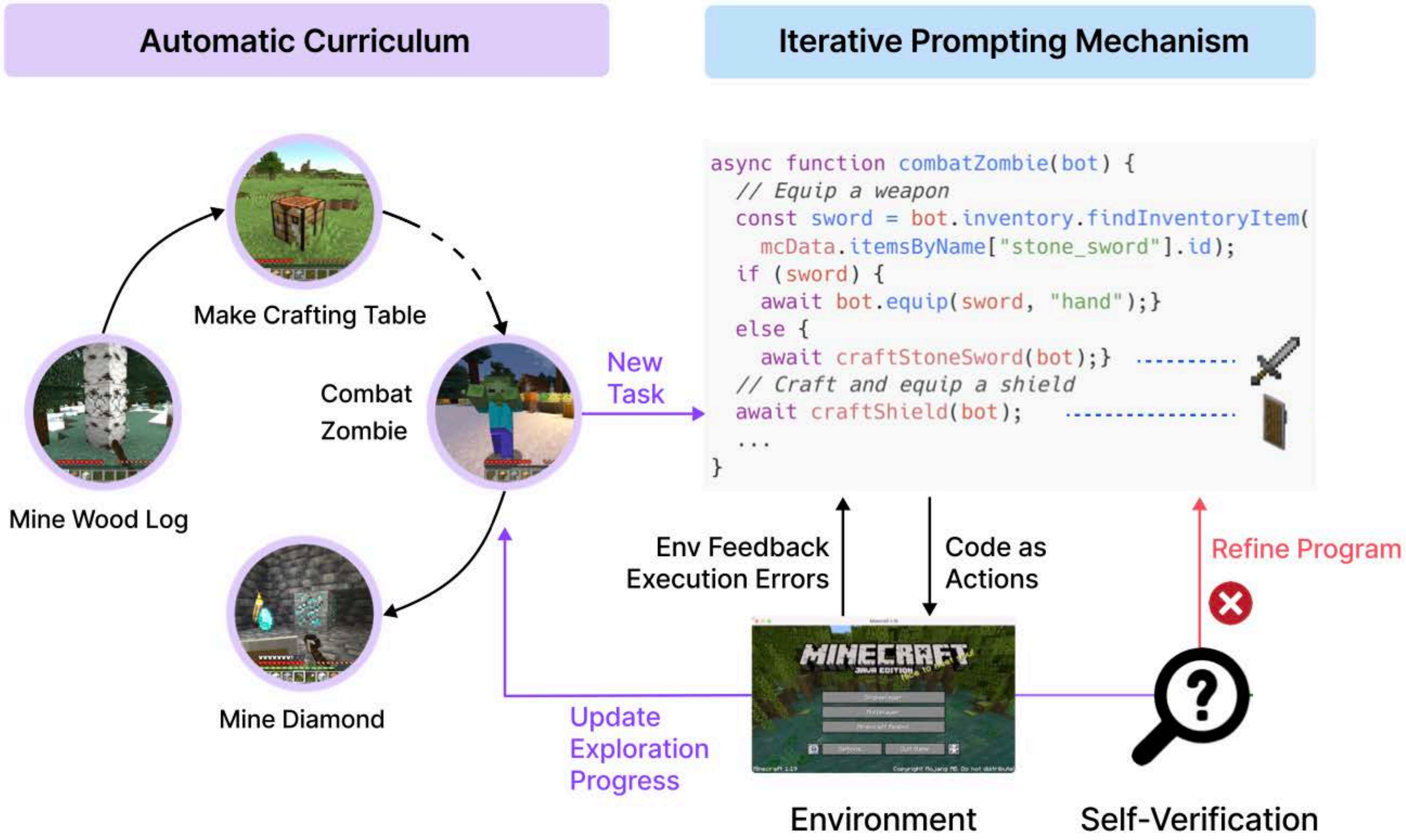
- Focuses on agent policies for exploration and skill acquisition in Minecraft



https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

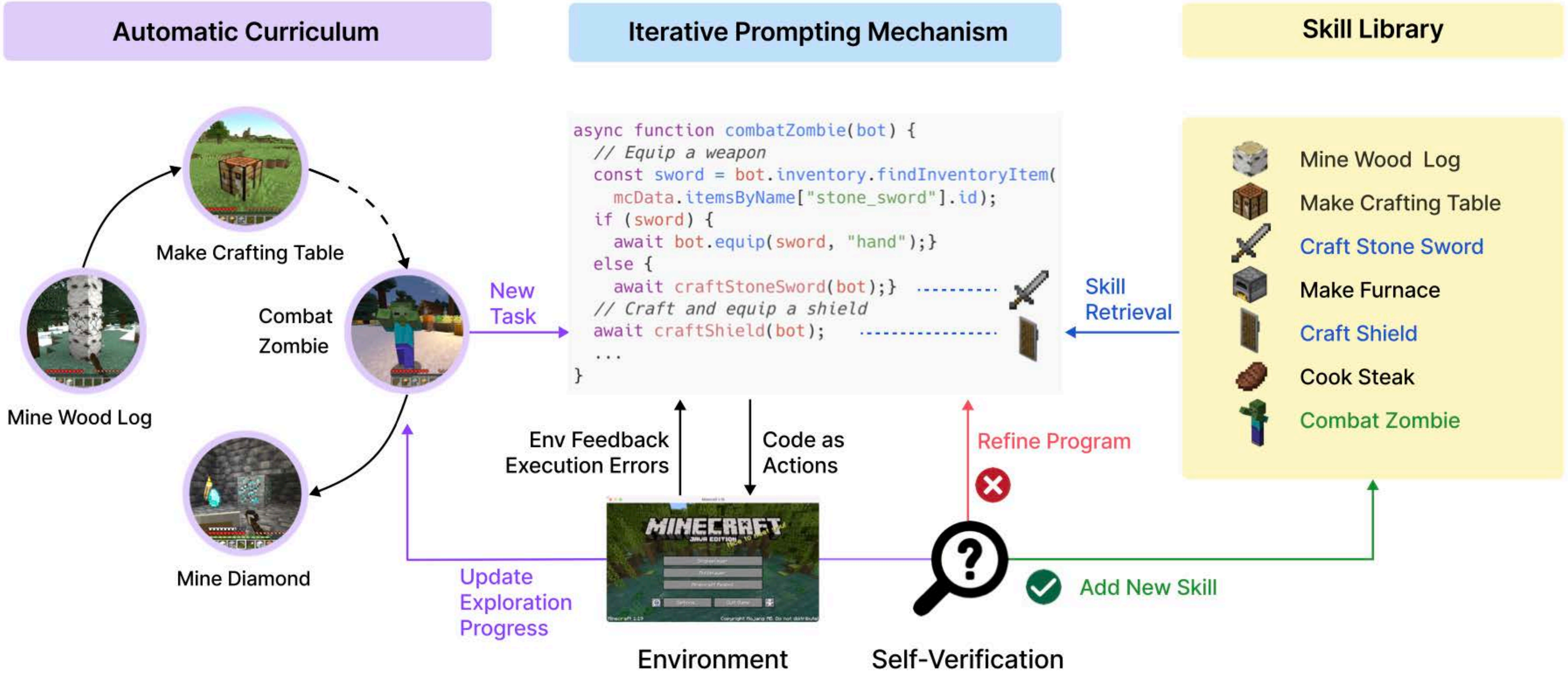
- Focuses on agent policies for exploration and skill acquisition in Minecraft



https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

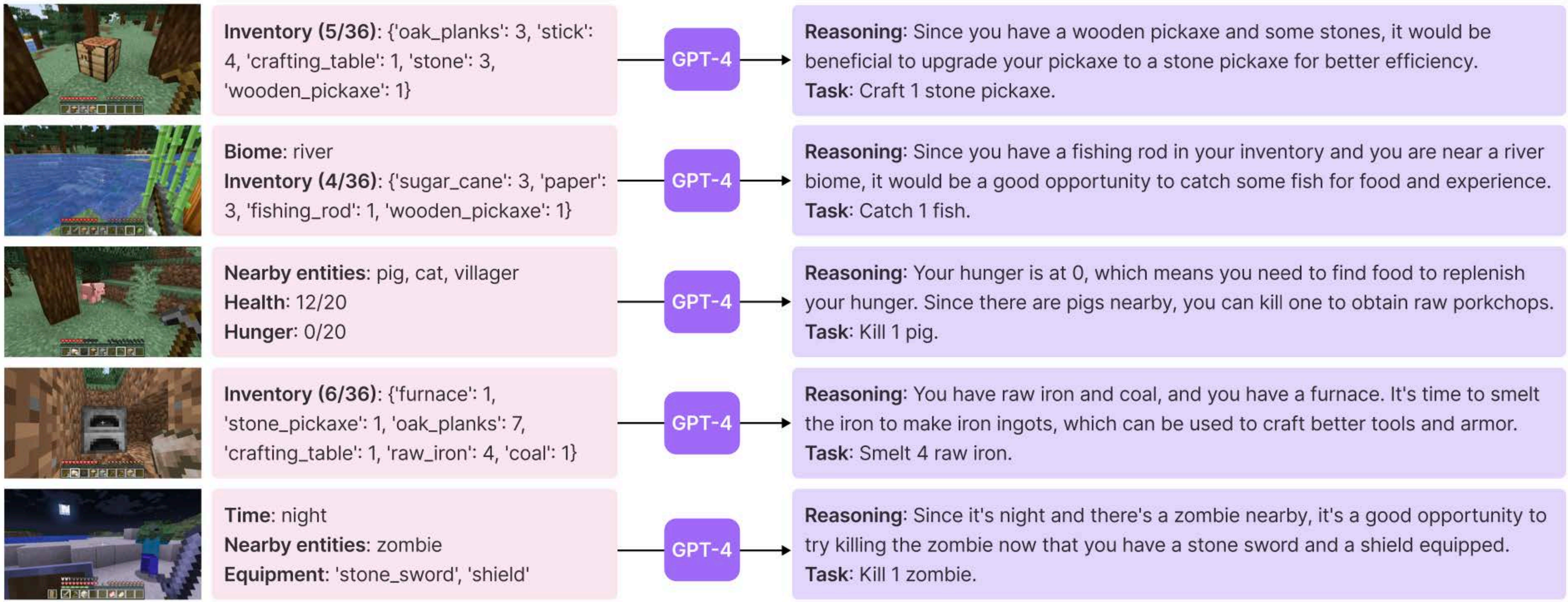
- Focuses on agent policies for exploration and skill acquisition in Minecraft



https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

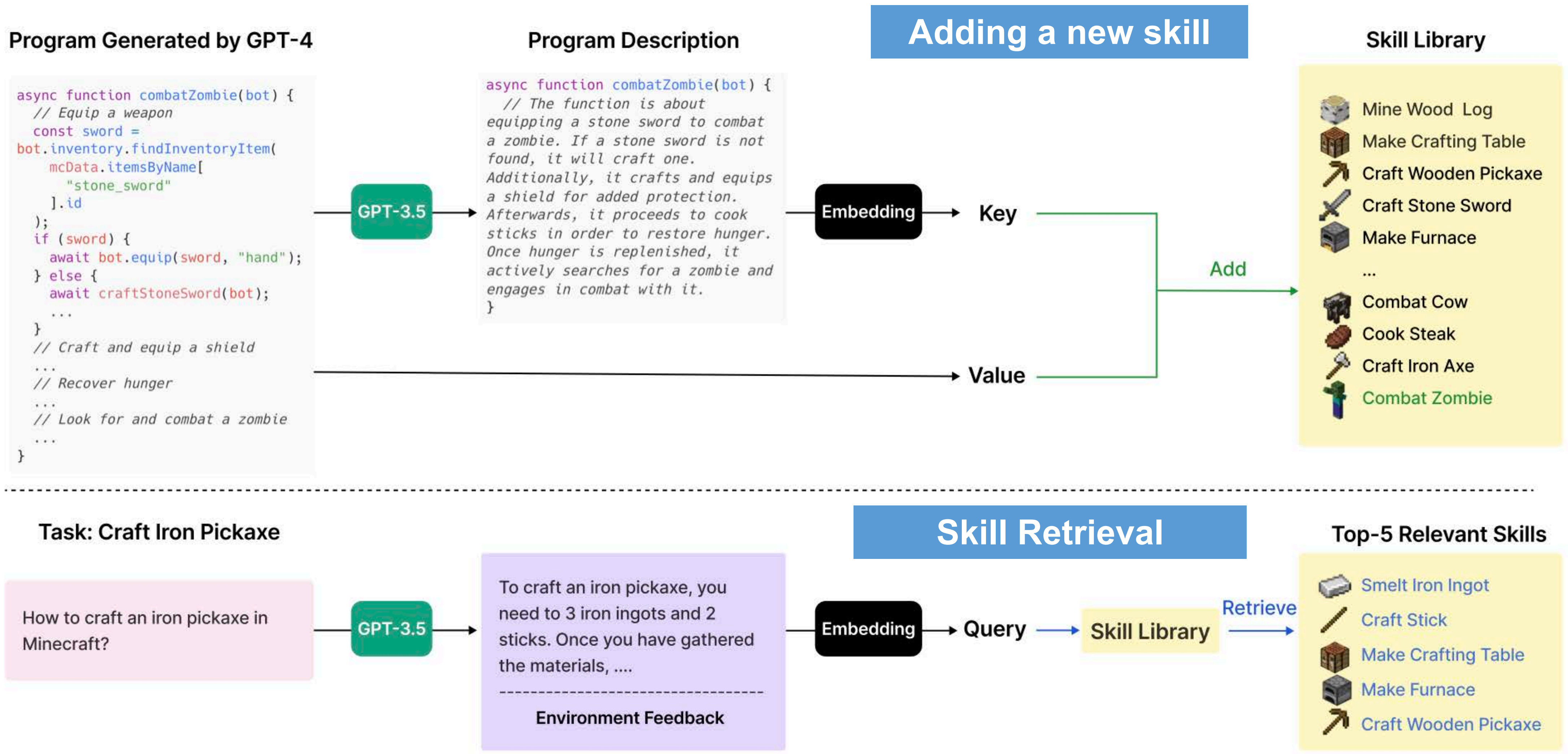
- Step 1: Propose tasks by the automatic curriculum



https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

- Step 2: Skill Library



https://arxiv.org/pdf/2305.16291

Voyager: Policy & Skill Acquisition

- Code as Actions, refining with environment feedback

Environment Feedback

I cannot make stick because I need: 2 more planks

I cannot make stone_shovel because I need: 2 more stick

GPT-4

```
async function craftStoneShovelWithTable(bot) {
   // If not enough cobblestone, mine cobblestone
   ...
+  // If not enough sticks, check if there are
+  // enough planks in the inventory
   if (sticksCount < 2) {
+    const planksCount = bot.inventory.count(
+      mcData.itemsByName.acacia_planks.id);
+    if (planksCount < 2) {
+      // Collect or craft planks
+      await mineBlock(bot, "acacia_log", 1);
+      await craftItem(bot, "acacia_planks", 1);
+    }
+    // Craft sticks using planks
     await craftItem(bot, "stick", 1);}
   await craftItem(bot, "stone_shovel", 1);
}
```

Execution Error

throw new Error(`No item named ${name}`);

No item named acacia_axe

at line 18:await craftItem(bot, "acacia_axe", 1);

GPT-4

```
-async function craftAcaciaAxe(bot) {
+async function craftWoodenAxe(bot) {
   // Craft anacia planks and sticks
   ...

   // Place the crafting table near the bot
   ...

-  // Craft an acacia axe using 3 acacia planks
-  // and 2 sticks
-  await craftItem(bot, "acacia_axe", 1);
-  bot.chat("Acacia axe crafted.");
+  // Craft a wooden axe using 3 acacia planks
+  // and 2 sticks
+  await craftItem(bot, "wooden_axe", 1);
+  bot.chat("Wooden axe crafted.");
}
```

https://arxiv.org/pdf/2305.16291

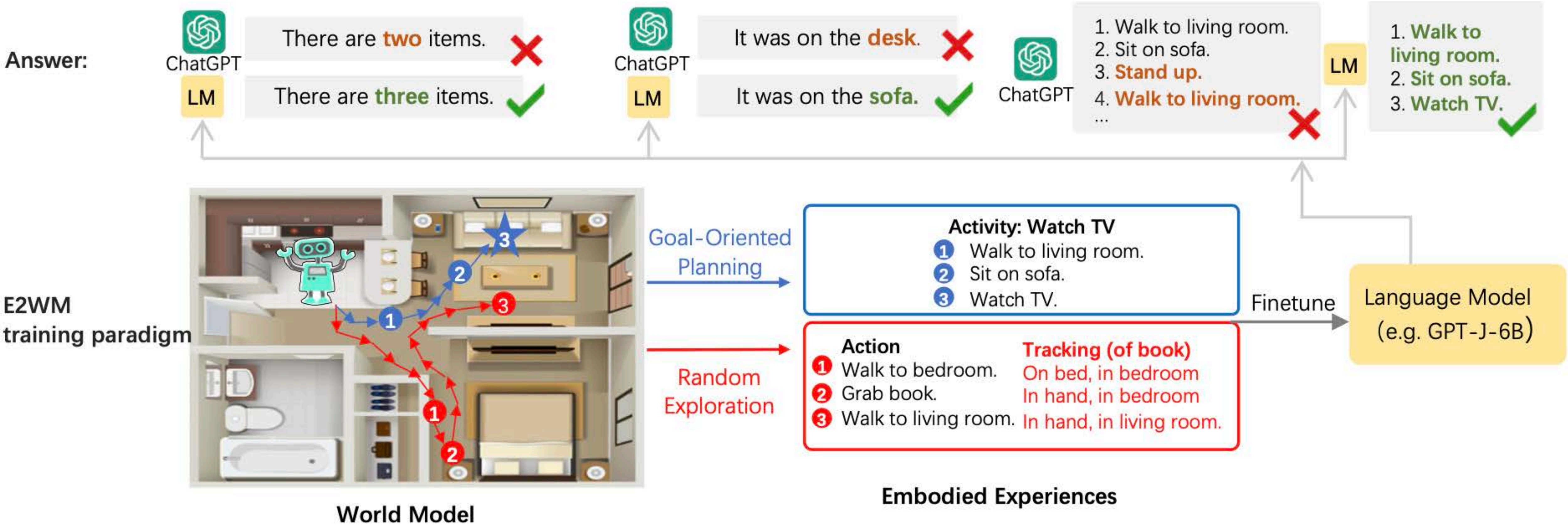

LLMs
Embodied
Agents

LLMs can also be finetuned / trained


LLMs
Embodied
Agents

Training LLMs for better planning

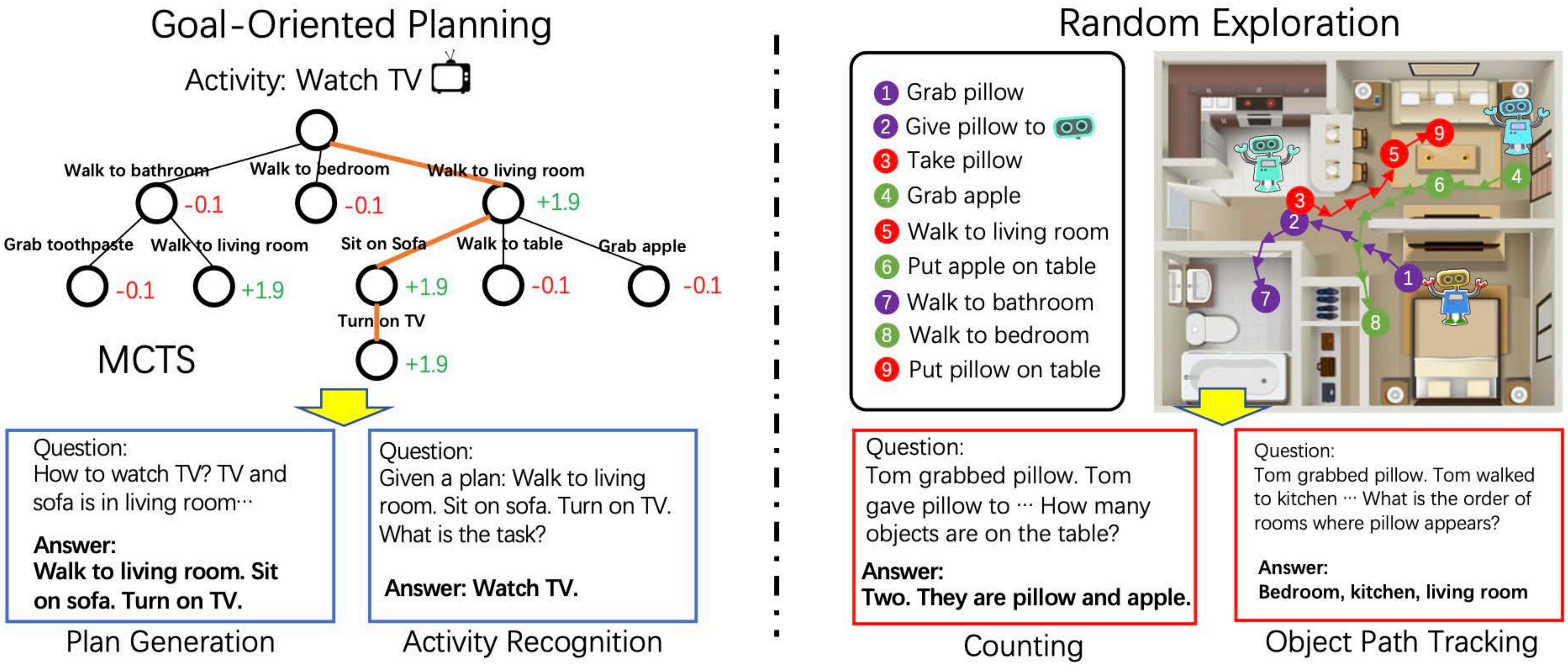
- Embodied Experiences from World Models: Synthetic Data Creation



https://arxiv.org/pdf/2305.10626

Training LLMs for better planning

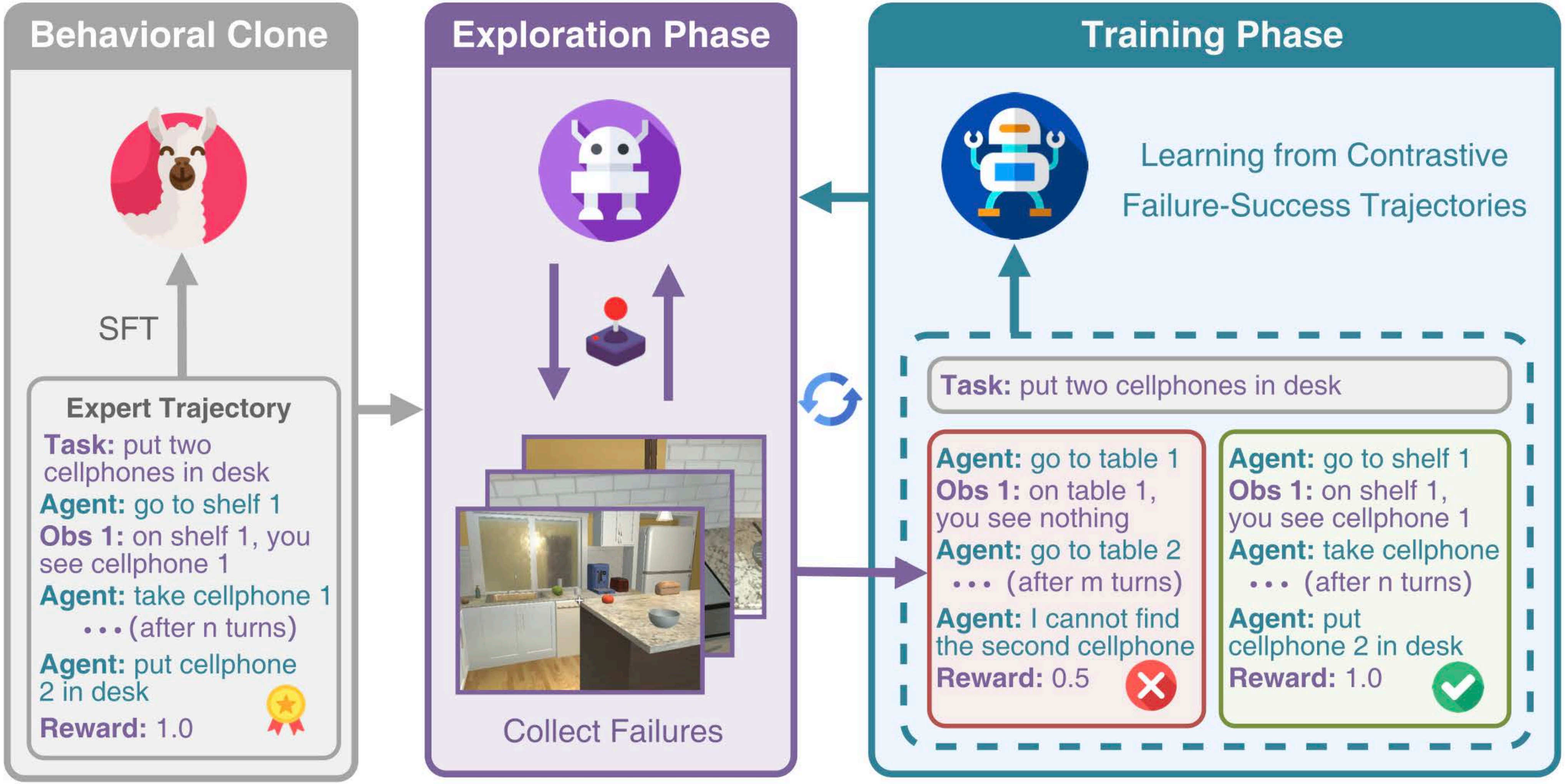
- Embodied Experiences from World Models: Synthetic Data Creation



https://arxiv.org/pdf/2305.10626

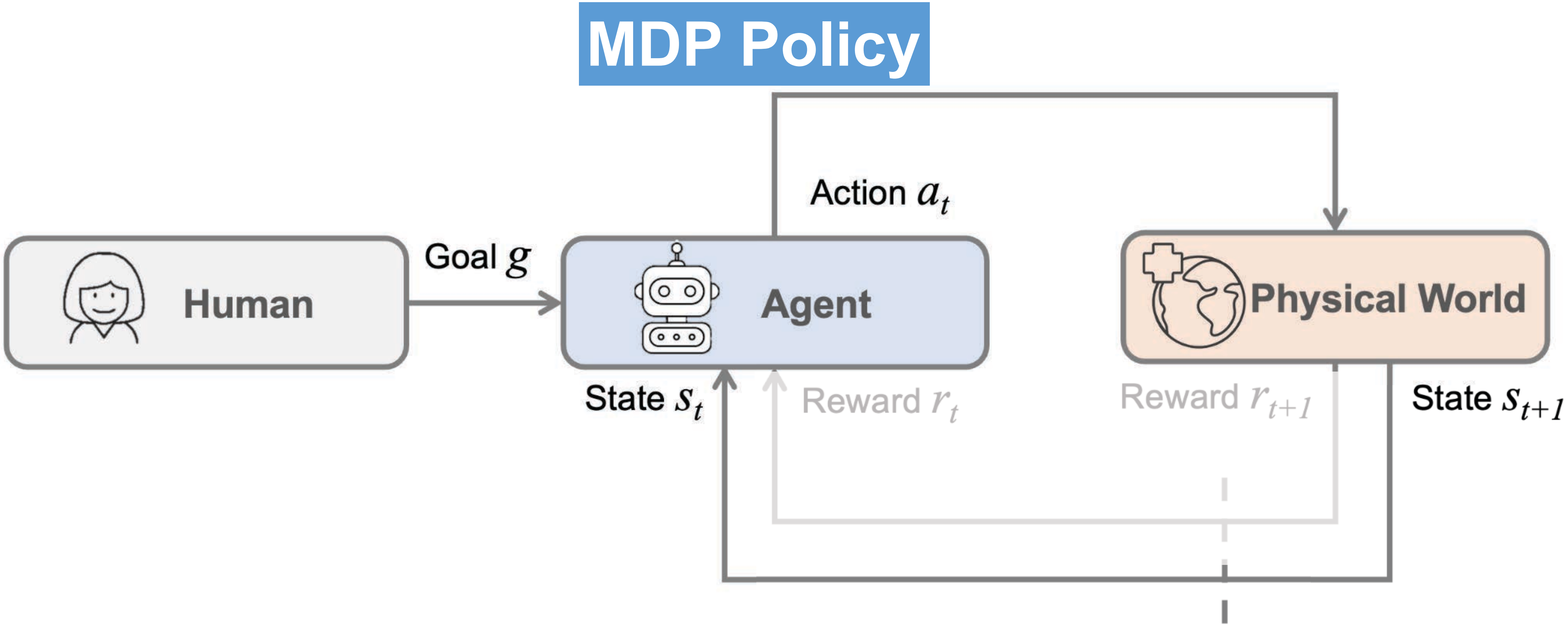
Training LLMs for better planning

- Transferring DPO to Trajectory: Exploration-based Trajectory Optimization

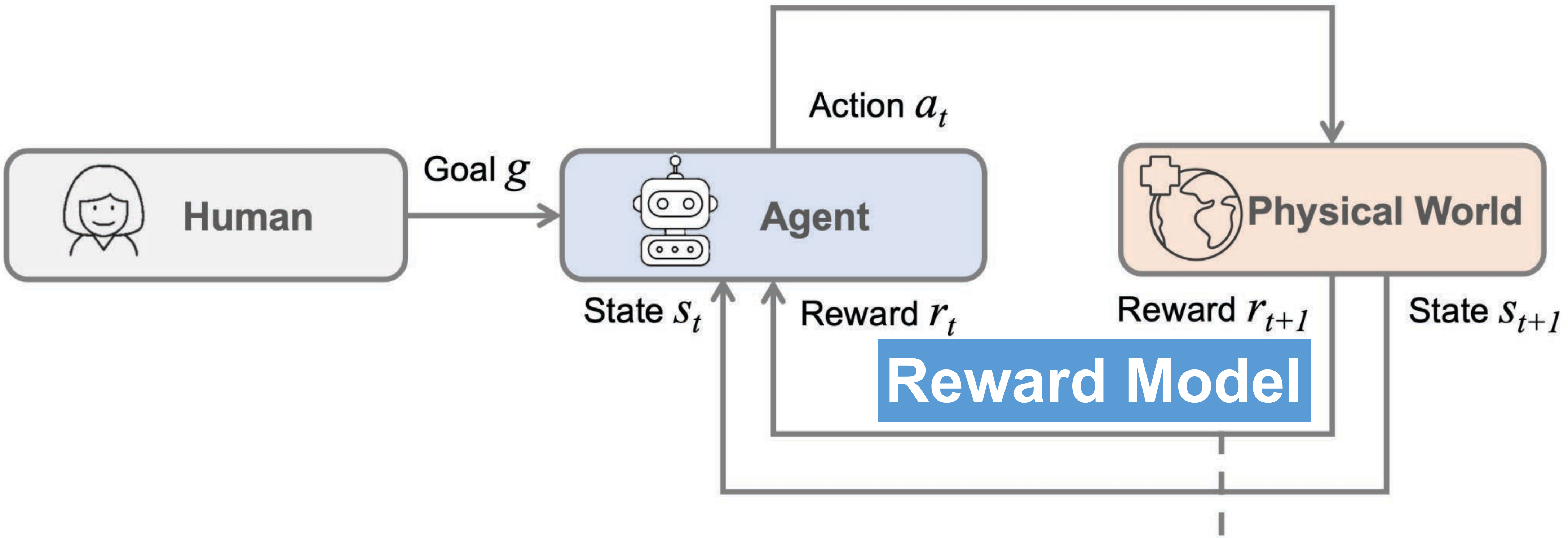


https://arxiv.org/pdf/2403.02502

Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)

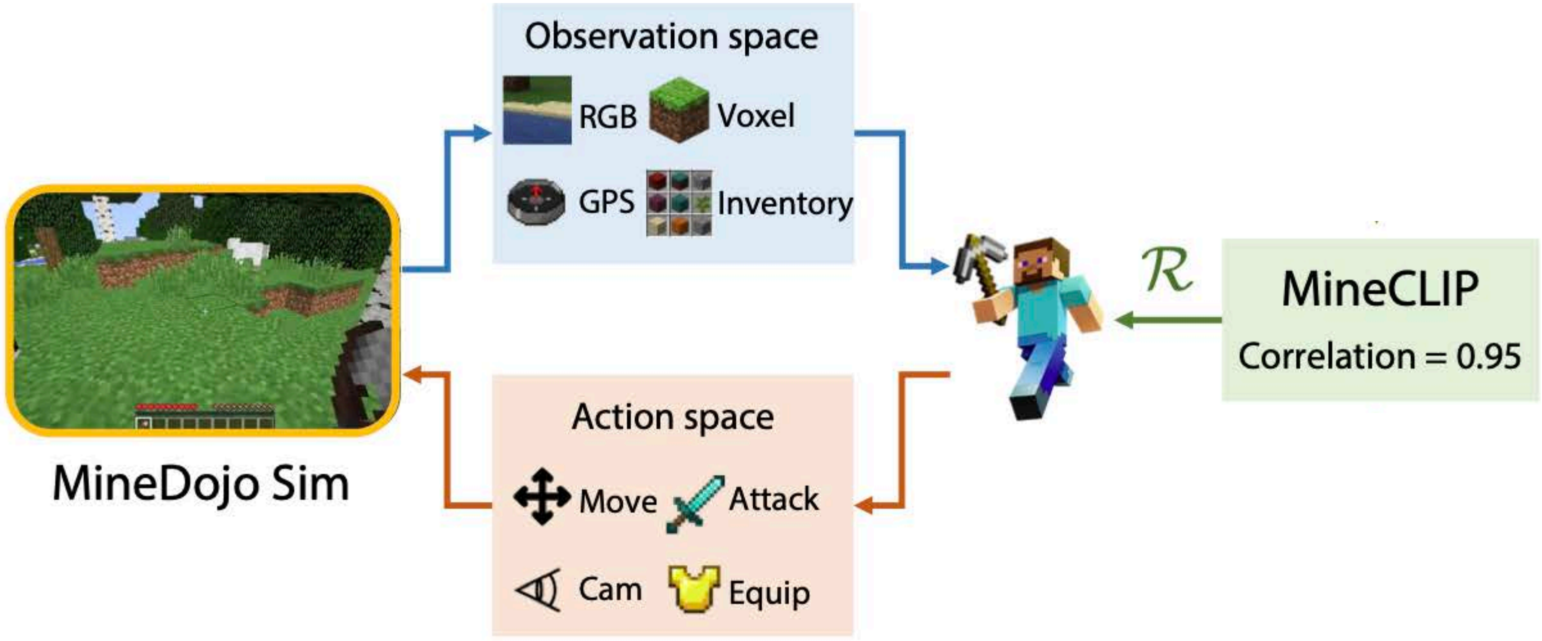


Reward Model

$$o, a \rightarrow r$$

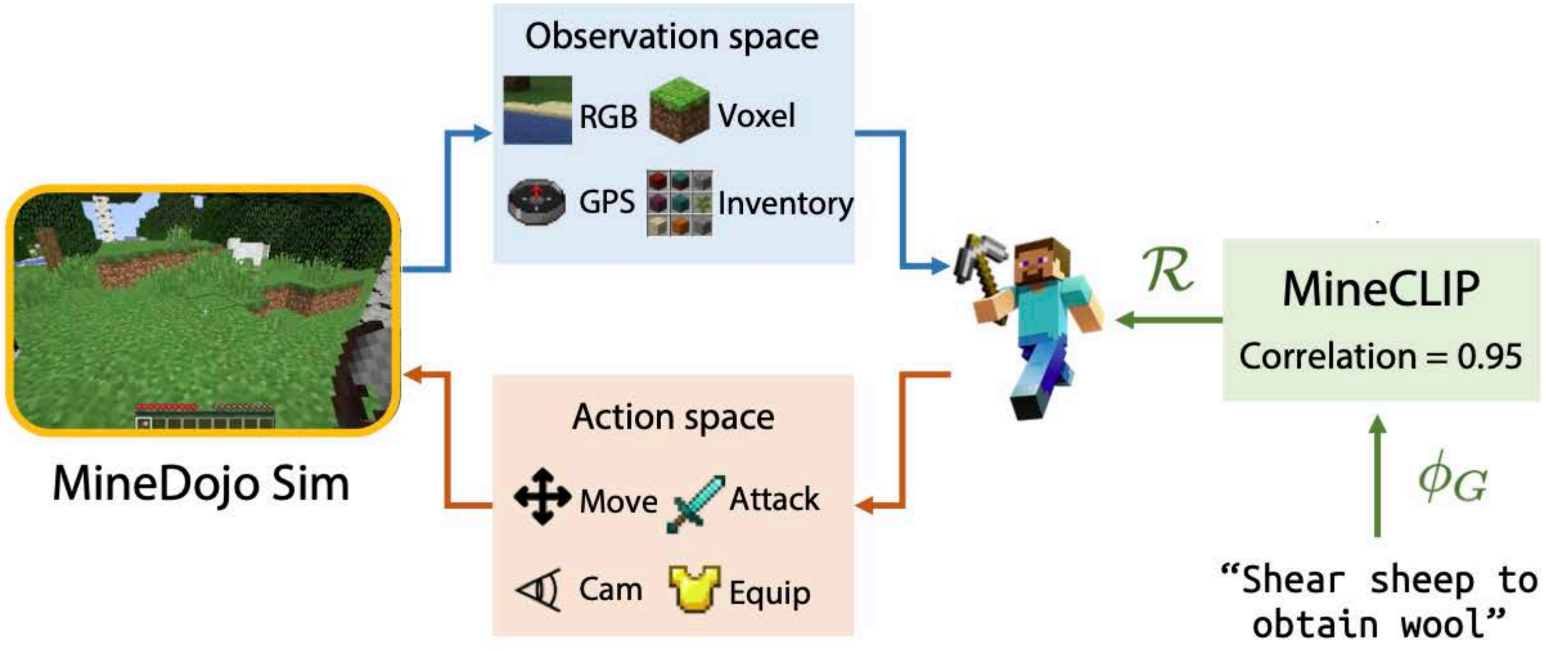
MineCLIP: Contrastive Video-Language Model

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



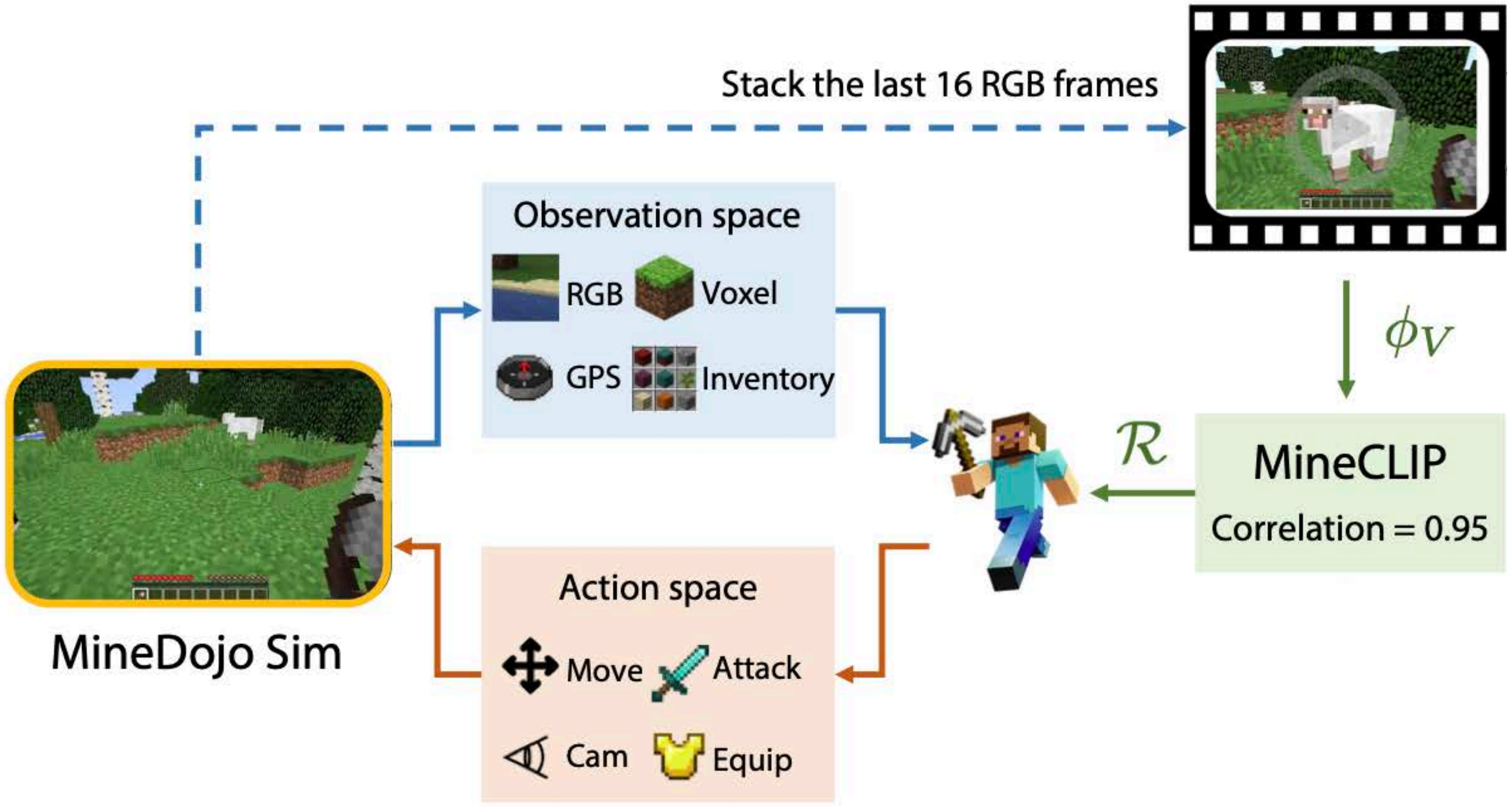
MineCLIP: Contrastive Video-Language Model

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



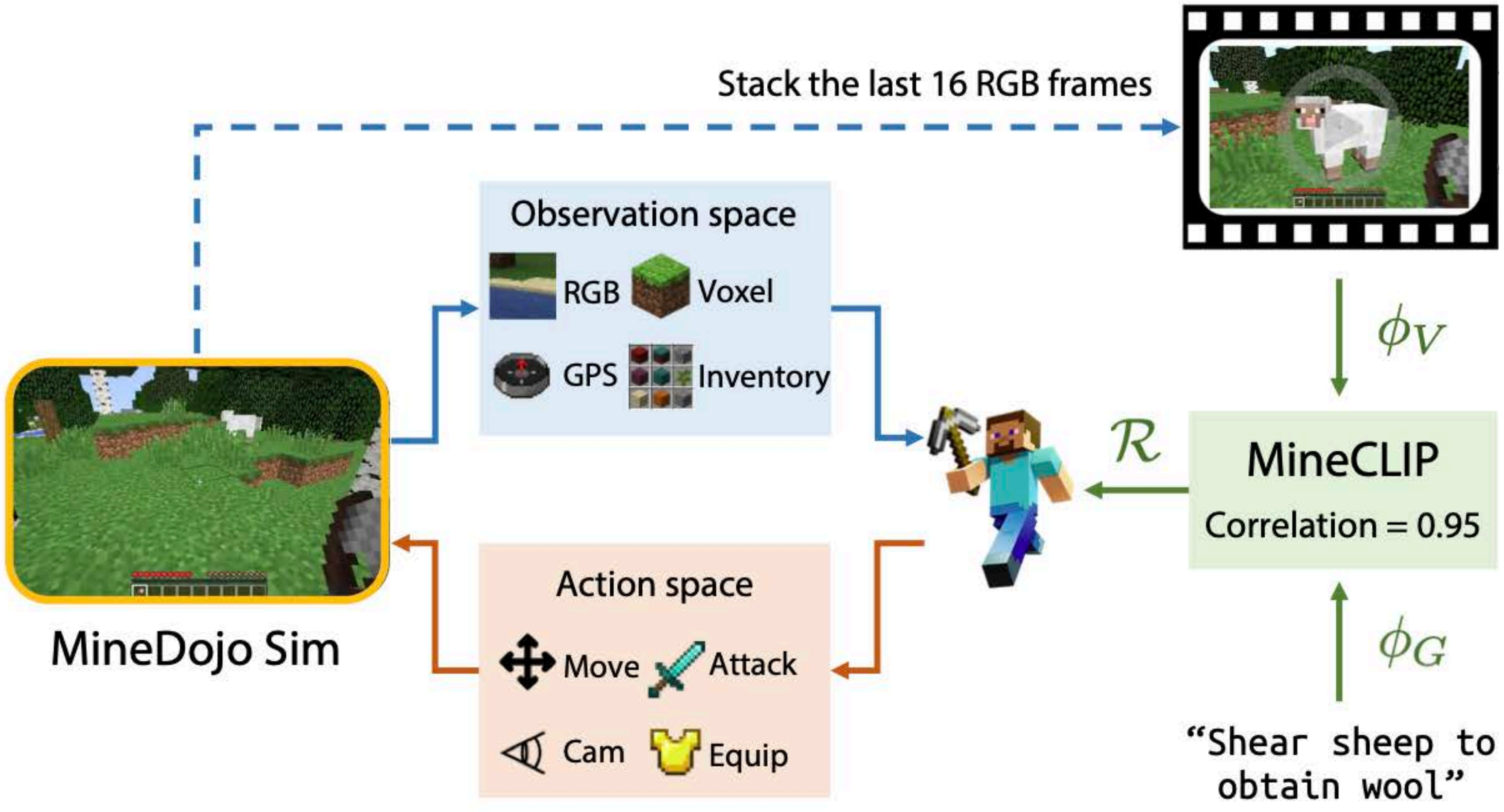
MineCLIP: Contrastive Video-Language Model

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



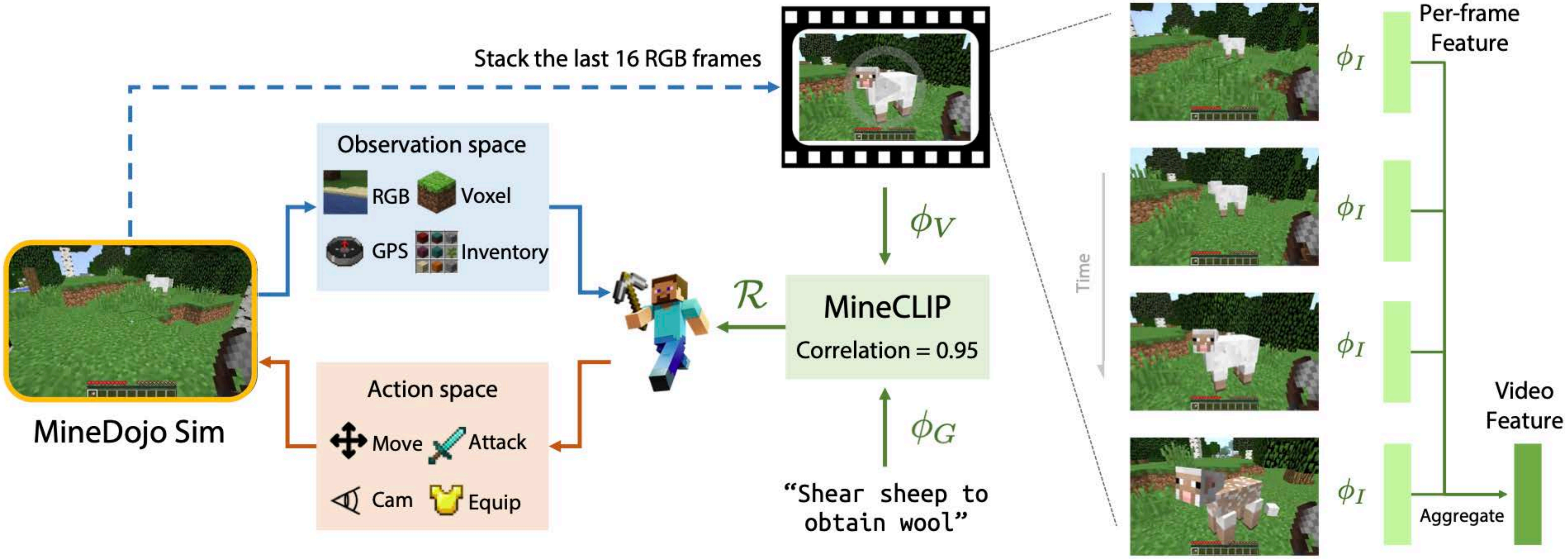
MineCLIP: Contrastive Video-Language Model

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

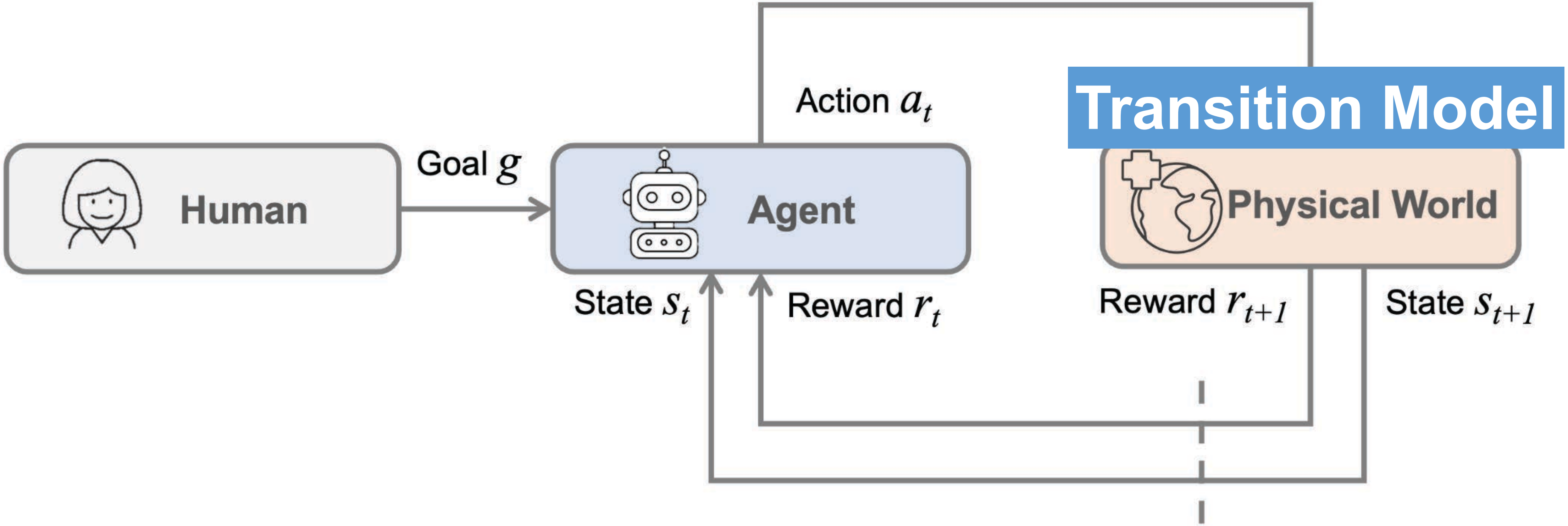


MineCLIP: Contrastive Video-Language Model

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



Let us go back to MDPs (Markov Decision Processes)



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

Goal Interpretation | Subgoal Decomposition | Action Sequencing | **Transition Modeling**

Input: Washing Dishes

Environment: stained(bowl)

Goal: not stained(bowl)

Operator Name: CLEAN_WITH_BRUSH

Transition Modeling

transitioning

Preconditions

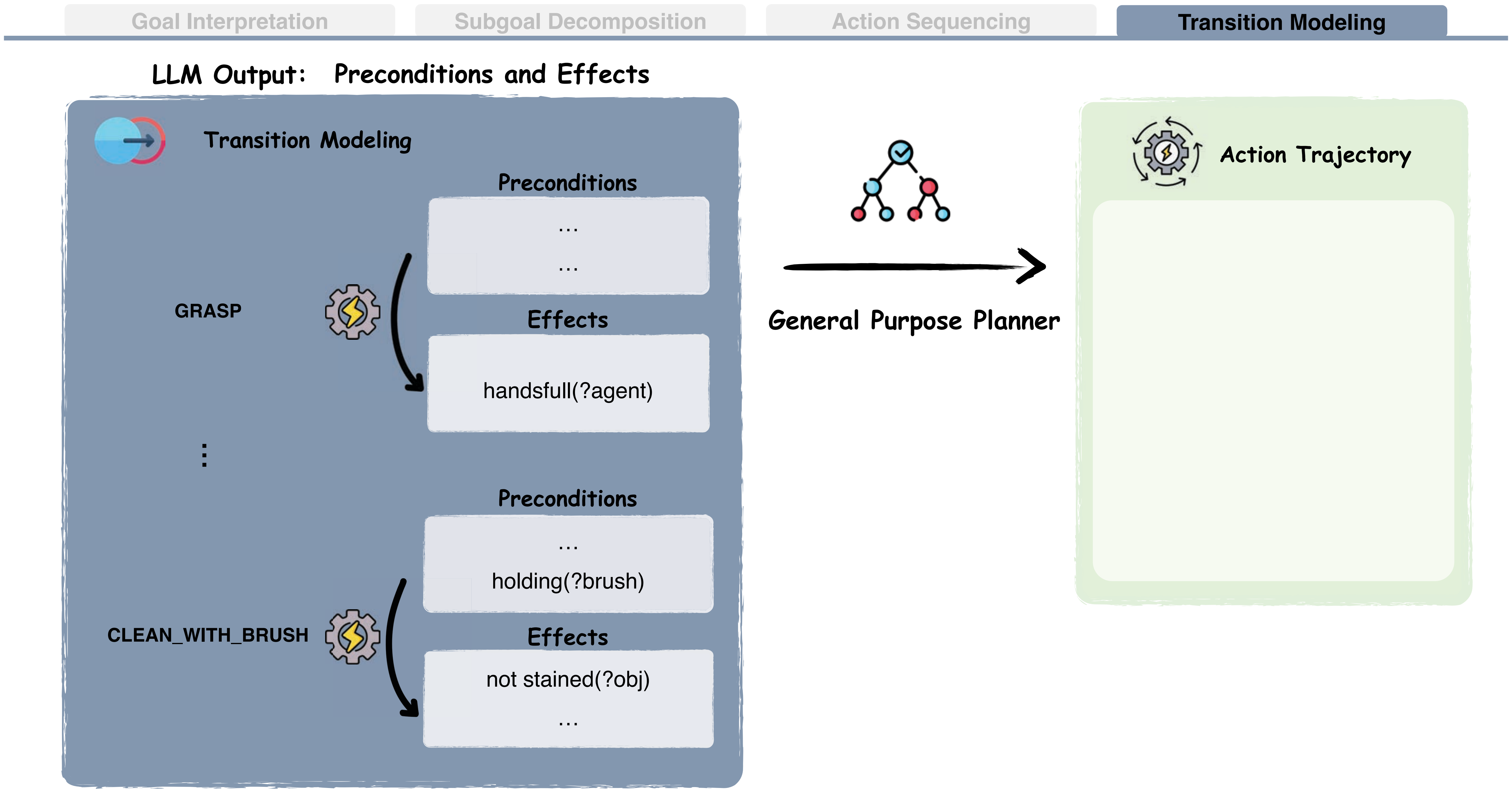
- stained(bowl) ✓
- soaked(scrub brush) ✓
- holding(scrub brush) ✓

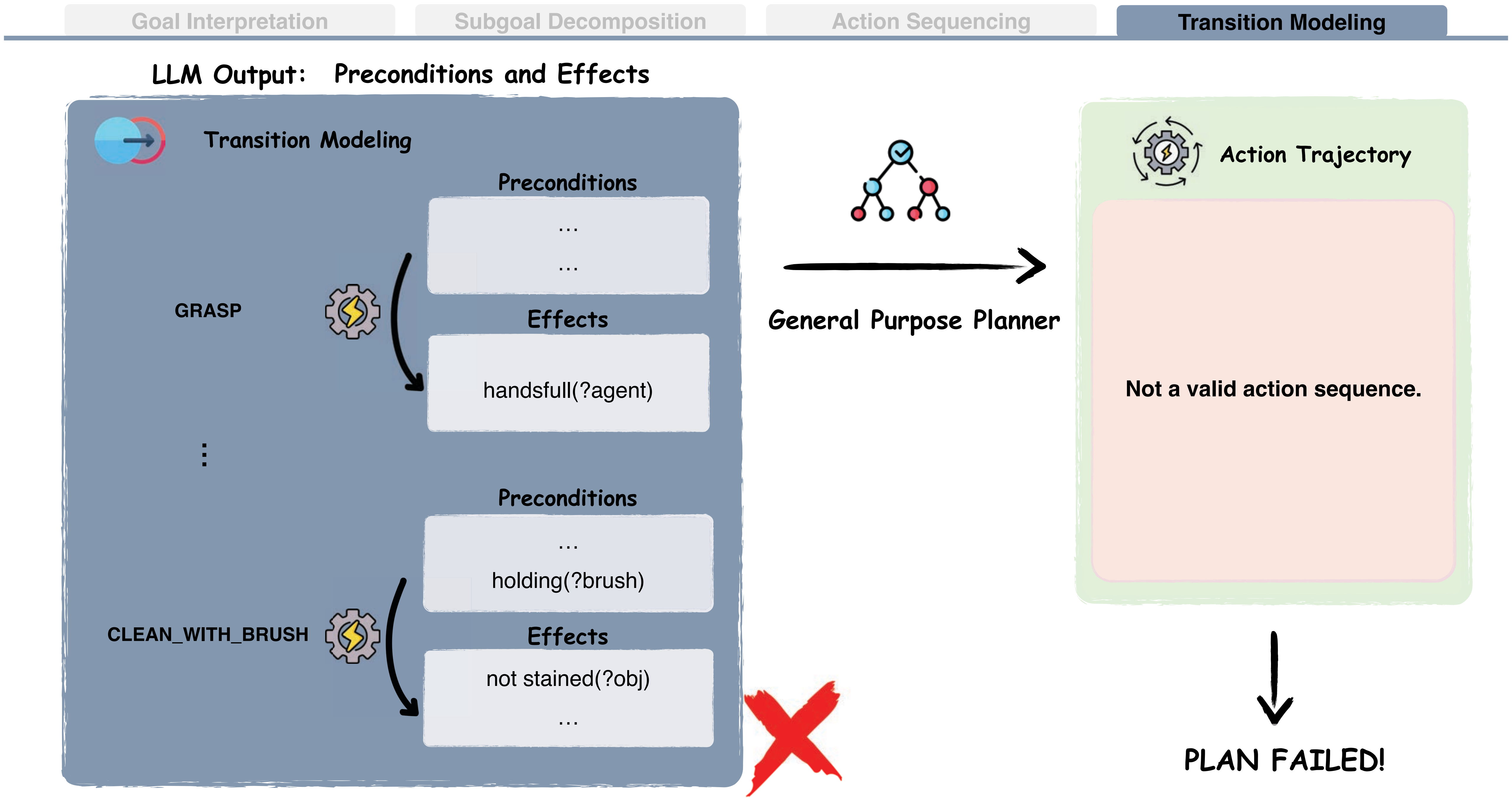
Effects

- not stained(bowl) ✓

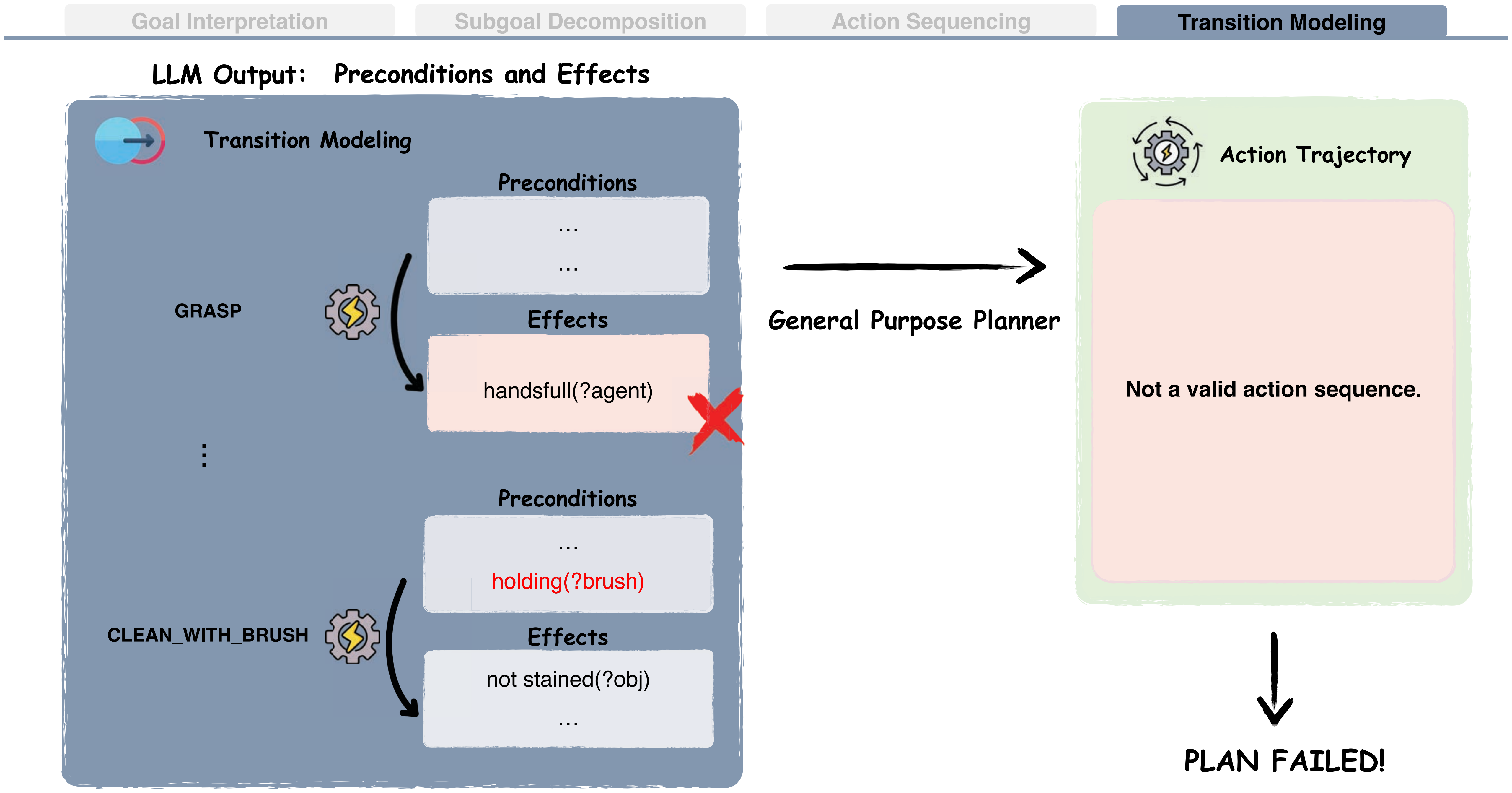
LLM Output

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.

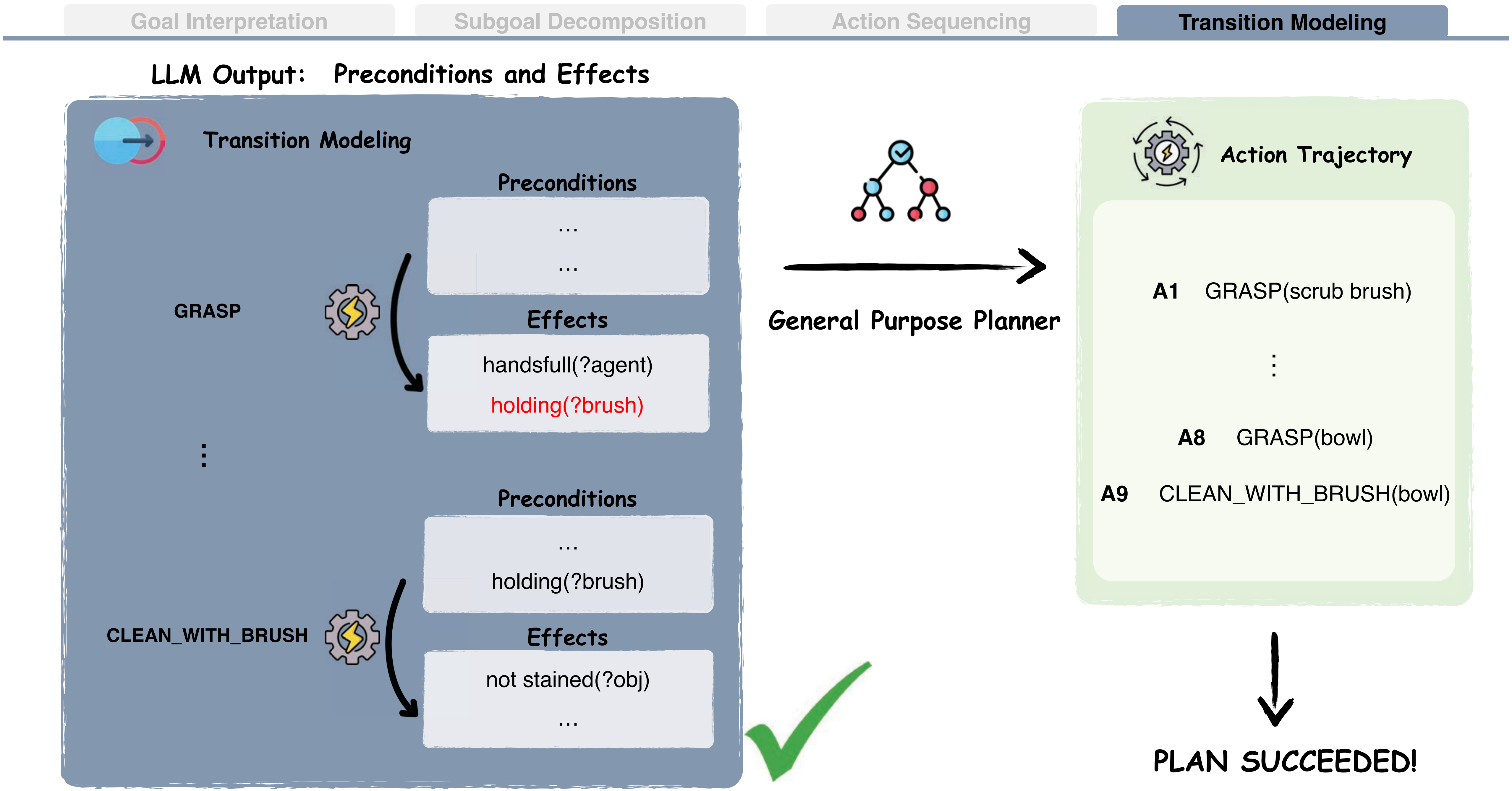

Goal Interpretation
Subgoal Decomposition
Action Sequencing
Transition Modeling
LLM Output: Preconditions and Effects
Transition Modeling
Preconditions
…
…
GRASP
Effects
handsfull(?agent)
⋮
Preconditions
…
holding(?brush)
CLEAN_WITH_BRUSH
Effects
not stained(?obj)
…
General Purpose Planner
Action Trajectory



Goal Interpretation
Subgoal Decomposition
Action Sequencing
Transition Modeling
LLM Output: Preconditions and Effects
Transition Modeling
GRASP
Preconditions
…
…
Effects
handsfull(?agent)
⋮
CLEAN_WITH_BRUSH
Preconditions
…
holding(?brush)
Effects
not stained(?obj)
…
General Purpose Planner
Action Trajectory
Not a valid action sequence.
PLAN FAILED!



Goal Interpretation
Subgoal Decomposition
Action Sequencing
Transition Modeling
LLM Output: Preconditions and Effects
Transition Modeling
Preconditions
...
...
GRASP
Effects
handsfull(?agent)
⋮
Preconditions
...
holding(?brush)
CLEAN_WITH_BRUSH
Effects
not stained(?obj)
...
General Purpose Planner
Action Trajectory
Not a valid action sequence.
PLAN FAILED!



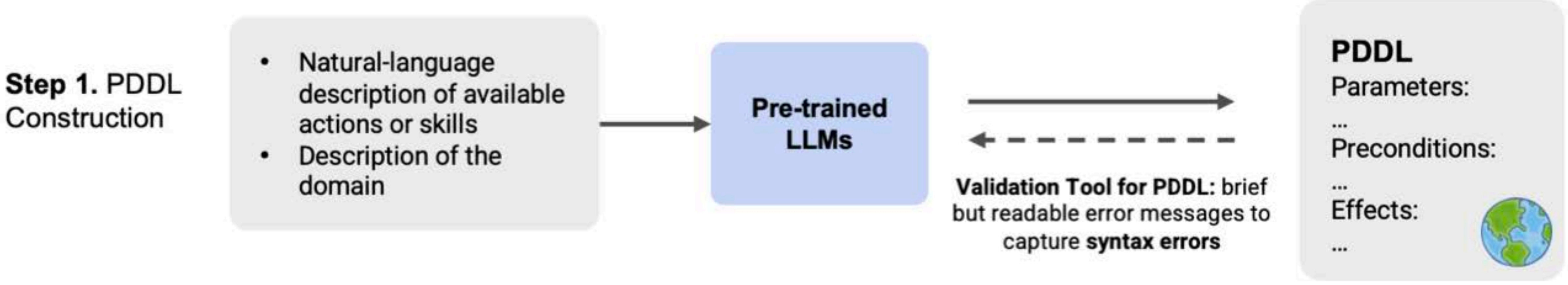
Goal Interpretation
Subgoal Decomposition
Action Sequencing
Transition Modeling
LLM Output: Preconditions and Effects
Transition Modeling
GRASP
Preconditions
...
...
Effects
handsfull(?agent)
holding(?brush)
CLEAN_WITH_BRUSH
Preconditions
...
holding(?brush)
Effects
not stained(?obj)
...
General Purpose Planner
Action Trajectory
A1 GRASP(scrub brush)
A8 GRASP(bowl)
A9 CLEAN_WITH_BRUSH(bowl)
PLAN SUCCEEDED!

PDDL as World Model

```
...
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
```

https://guansuns.github.io/pages/llm-dm/

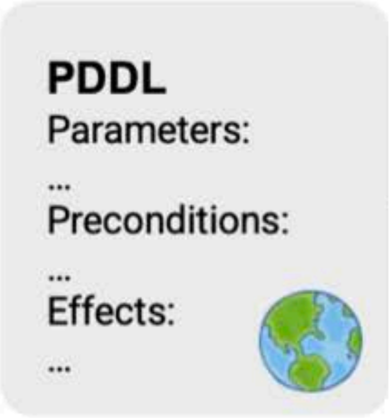
PDDL as World Model



https://guansuns.github.io/pages/llm-dm/

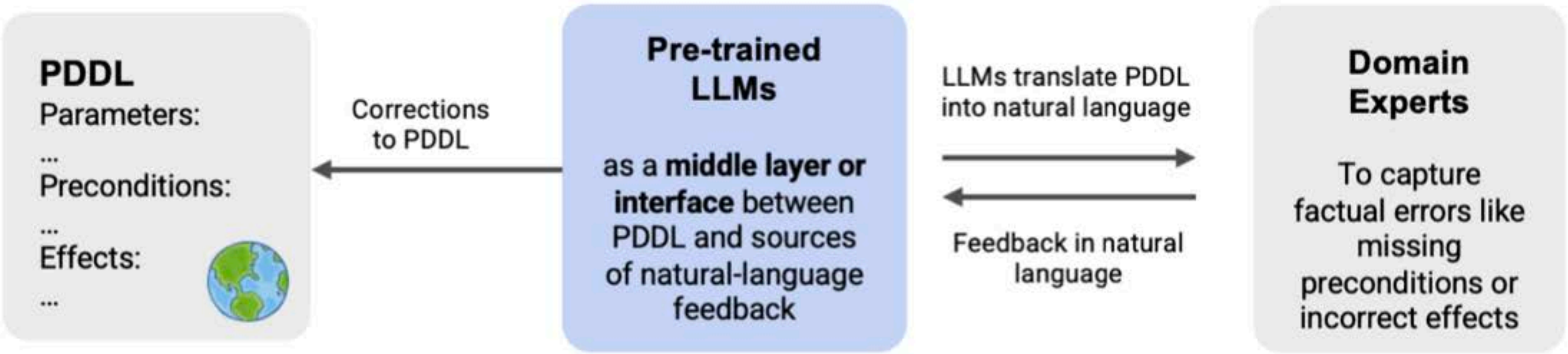
PDDL as World Model

Step 2. Human Correction

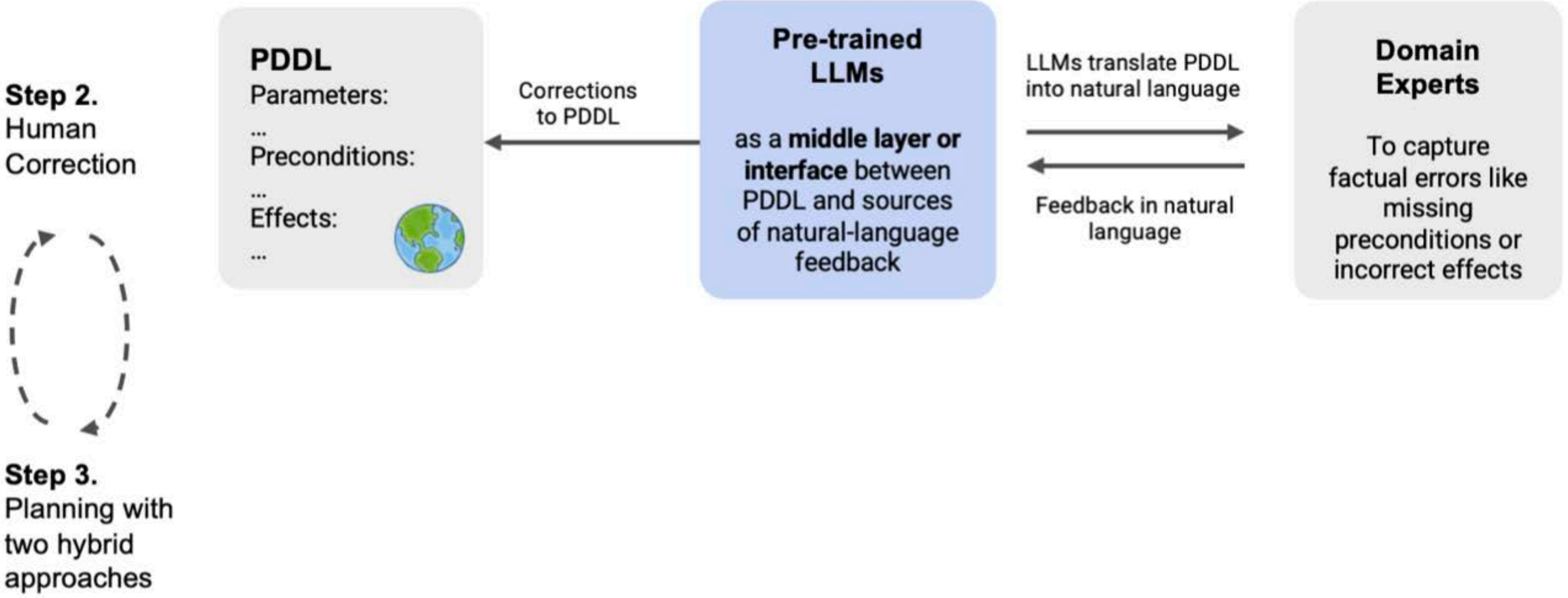


PDDL as World Model

Step 2.
Human
Correction

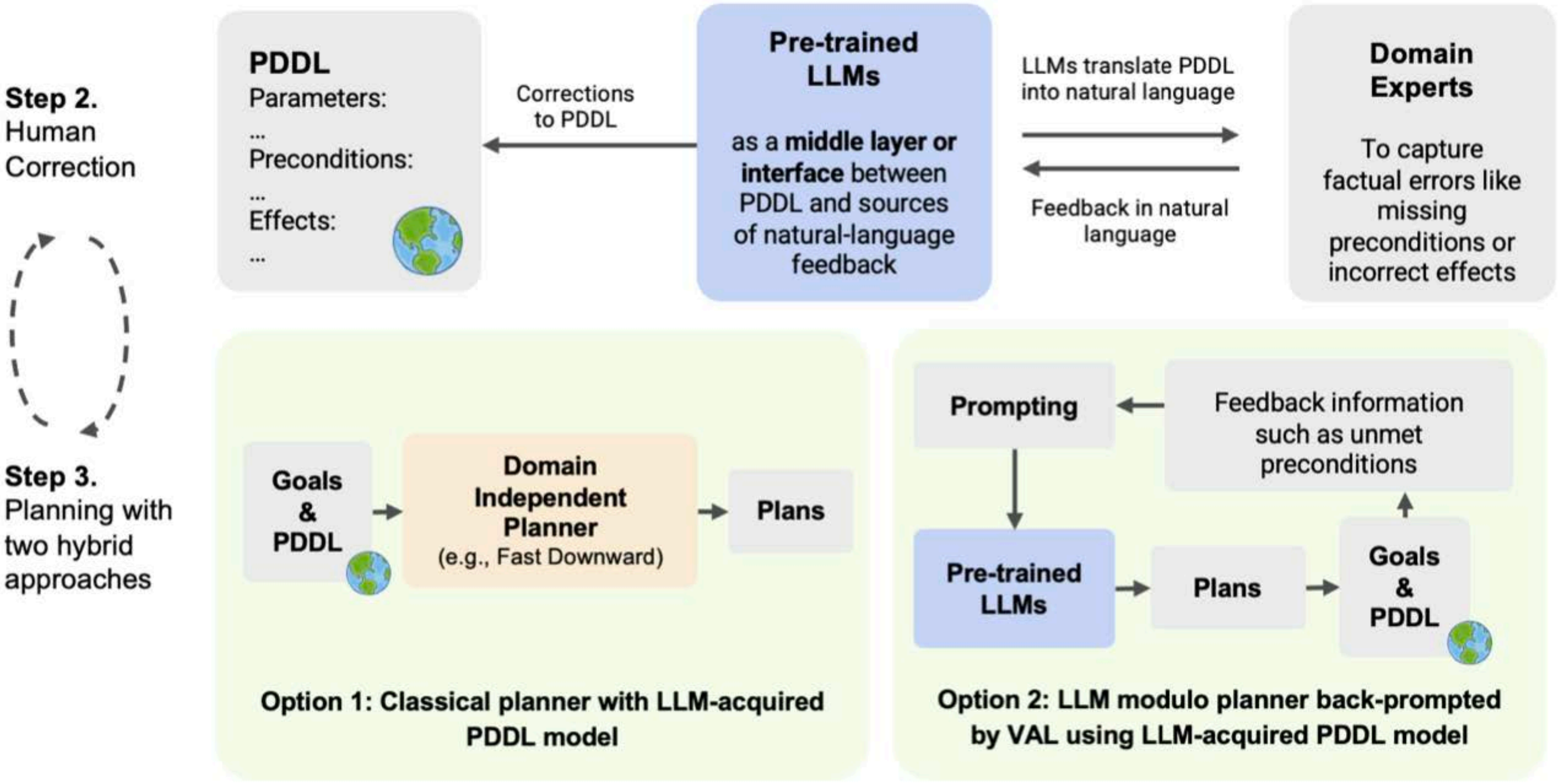


PDDL as World Model



https://guansuns.github.io/pages/llm-dm/

PDDL as World Model

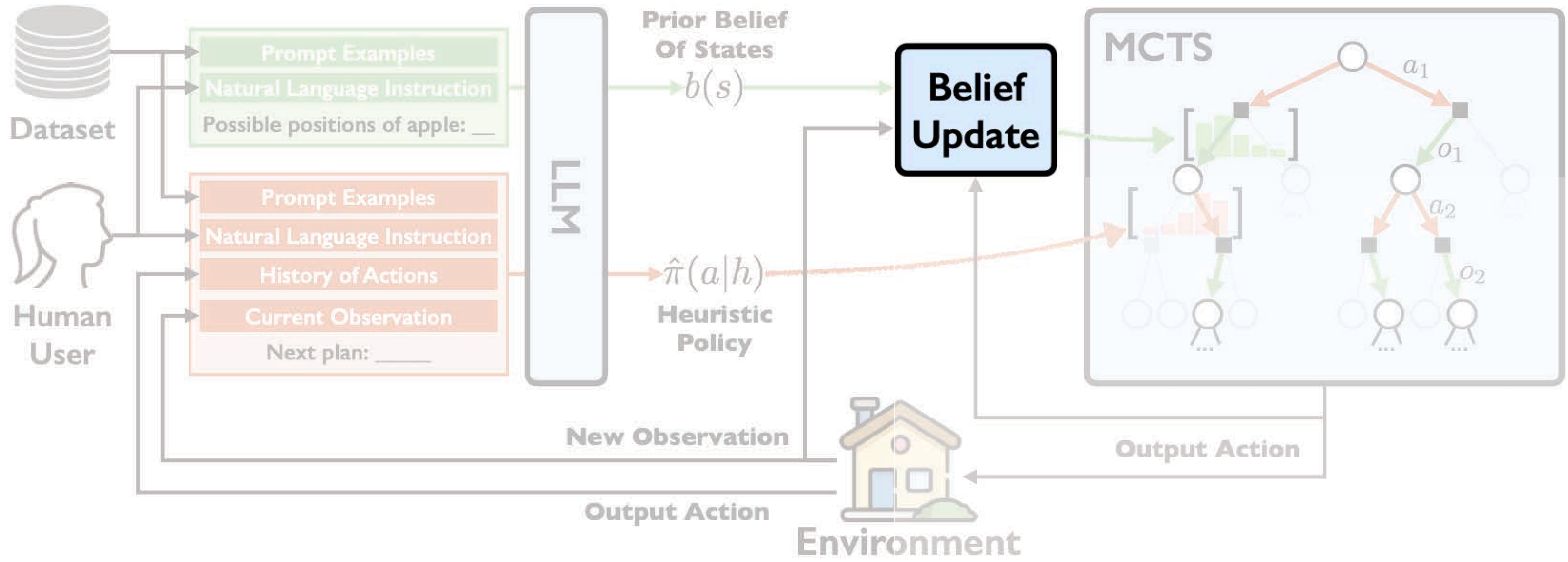


https://guansuns.github.io/pages/llm-dm/

LLMs as World Models

LLM-MCTS: LLMs as Commonsense Knowledge

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



https://arxiv.org/pdf/2305.14078

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.

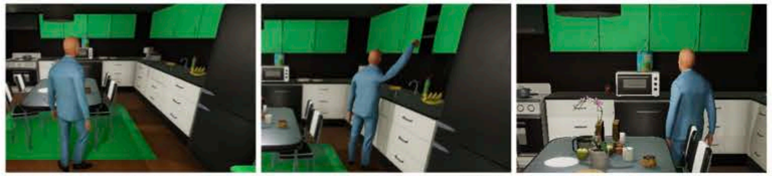
Question: If Elizabeth has been trying to get a bottle of wine, which one of the following statements is more likely to be true?
(a) Elizabeth thinks that there is a bottle of wine inside the fridge.
(b) Elizabeth thinks that there isn't any bottle of wine inside the fridge.

Type 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: … 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 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 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

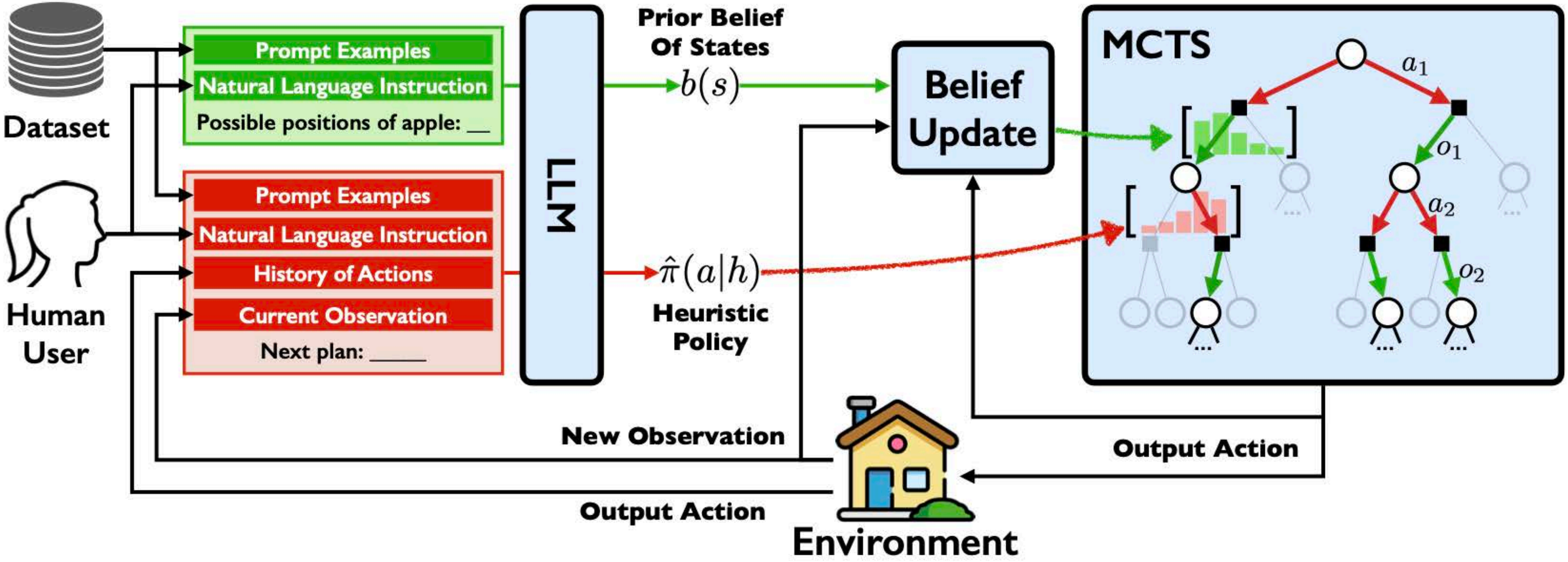
Scene: … The dishwasher holds a dish bowl… The first cabinet from the left holds a bag of chips and a wine glass… The fifth cabinet has an apple…
Actions: Williams… advances towards the first kitchen cabinet, opens it, and then shuts it. He then moves towards the fifth kitchen cabinet.

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.

https://arxiv.org/pdf/2401.08743

LLM-MCTS: LLMs as Commonsense Knowledge

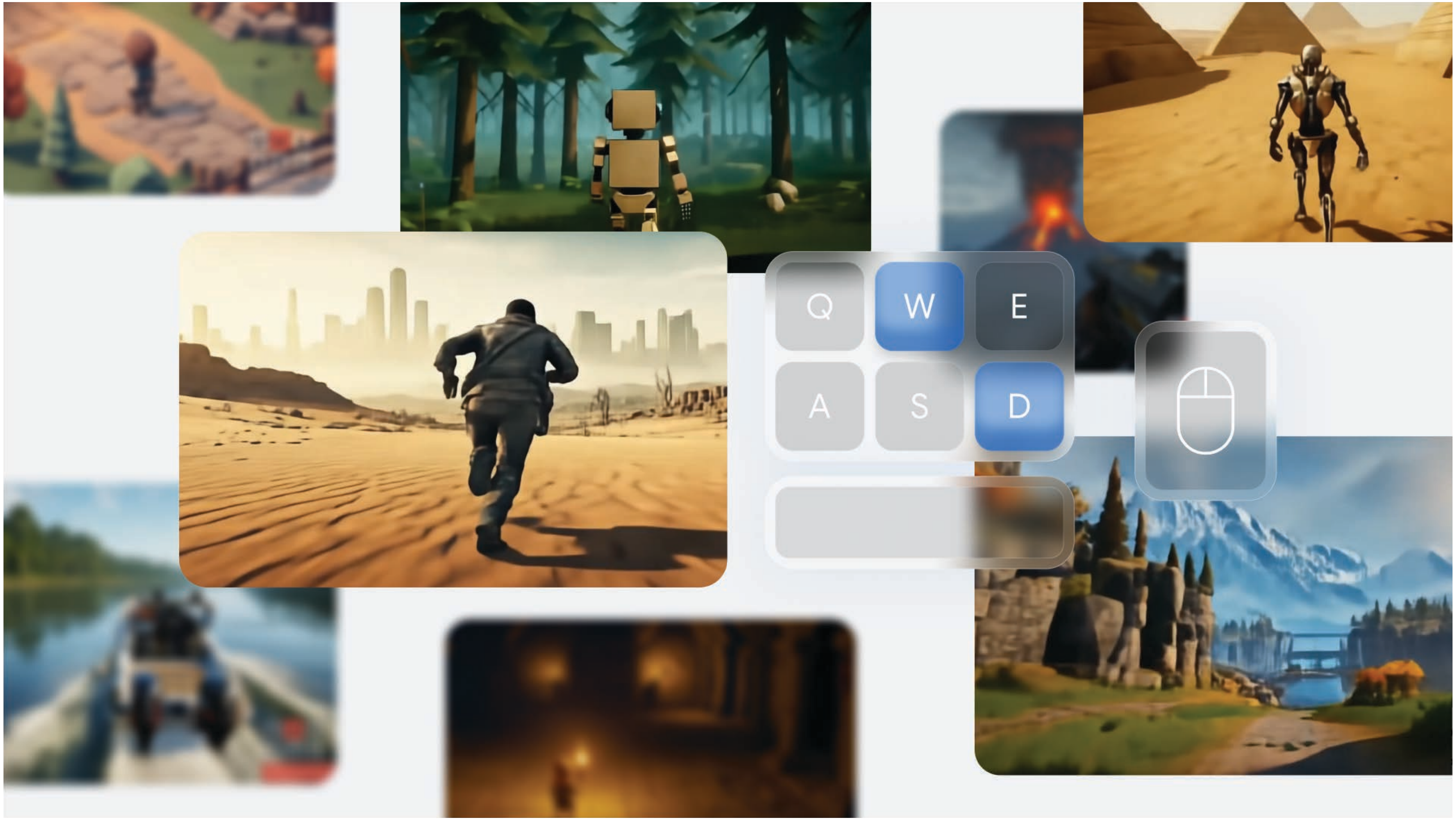
- Sample from the commonsense belief to obtain an initial state of the world



https://arxiv.org/pdf/2305.14078

Video Generation as World Model (World Foundation Model)

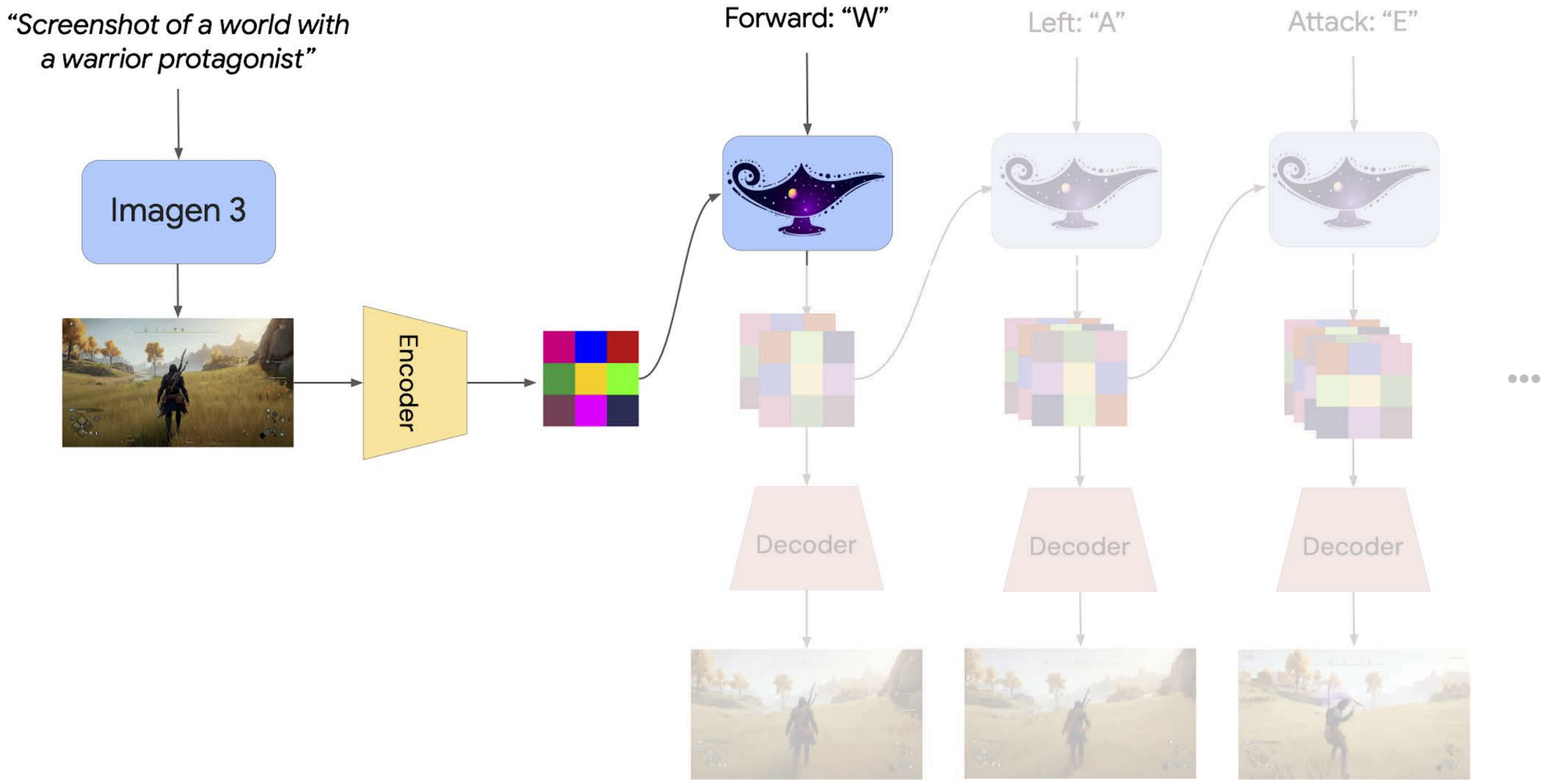
Genie 2: World Foundation Model



https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

Genie 2: World Foundation Model

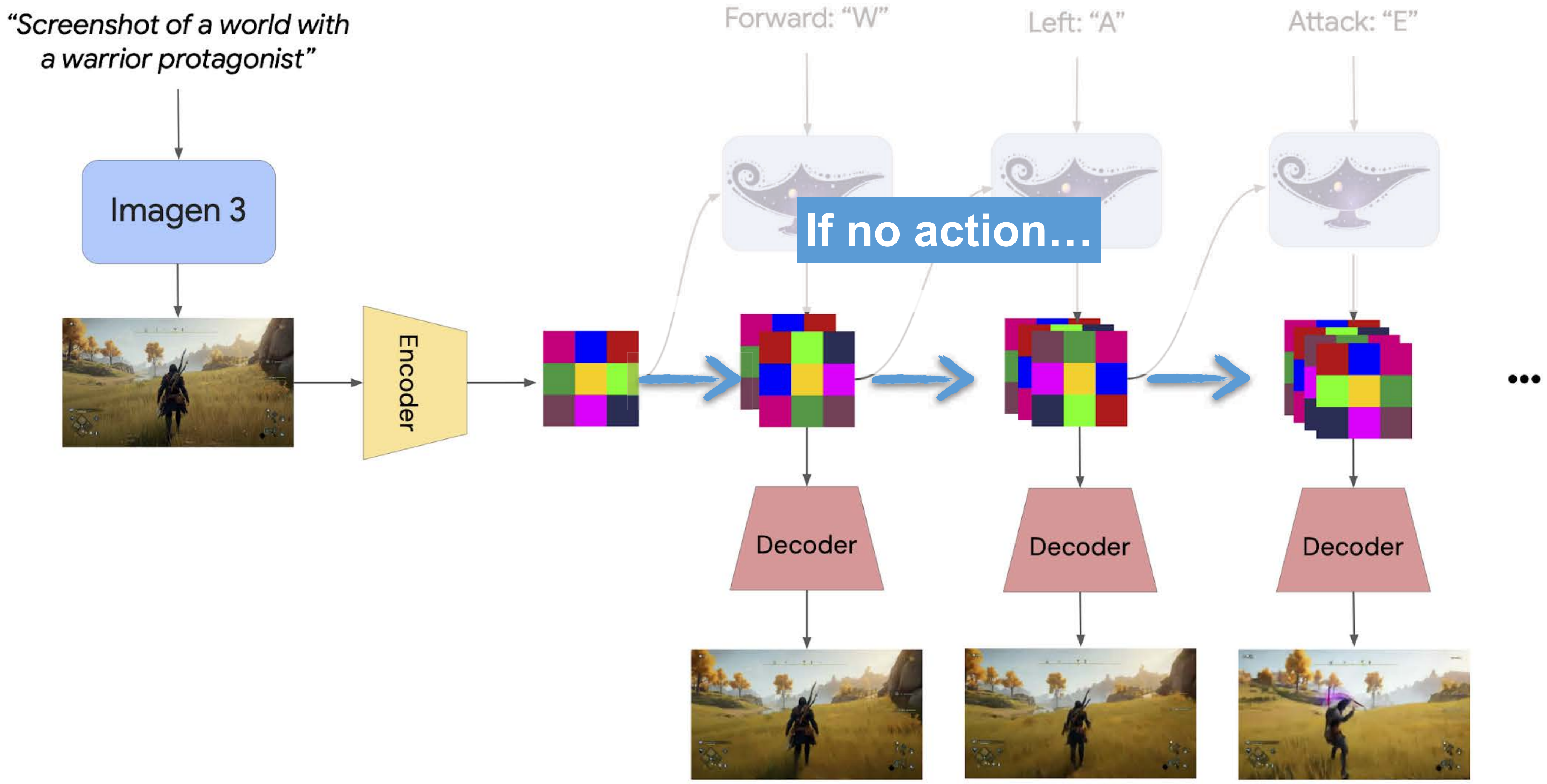
- Diffusion world model



https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

Genie 2: World Foundation Model

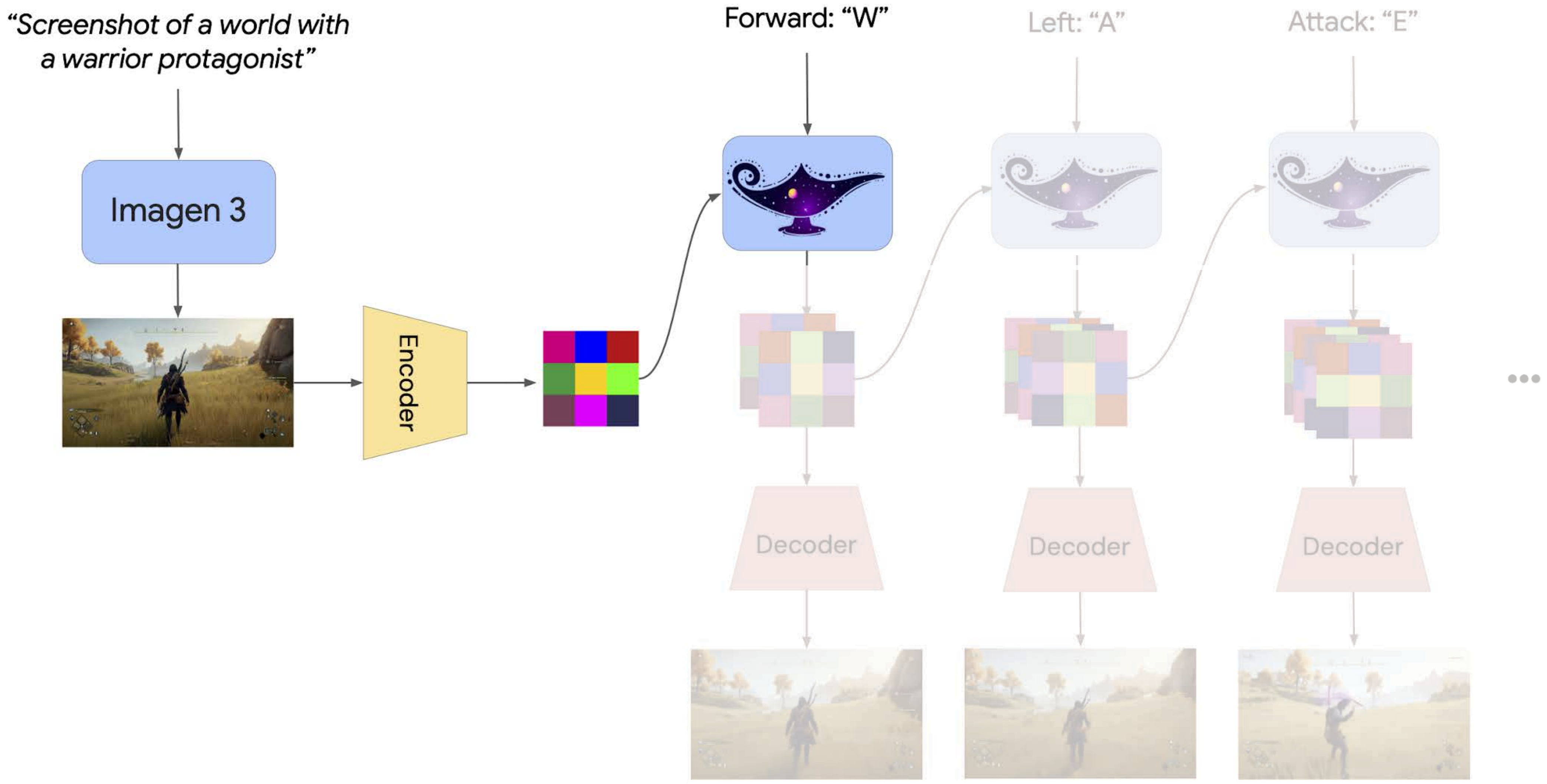
- Diffusion world model



https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

Genie 2: World Foundation Model

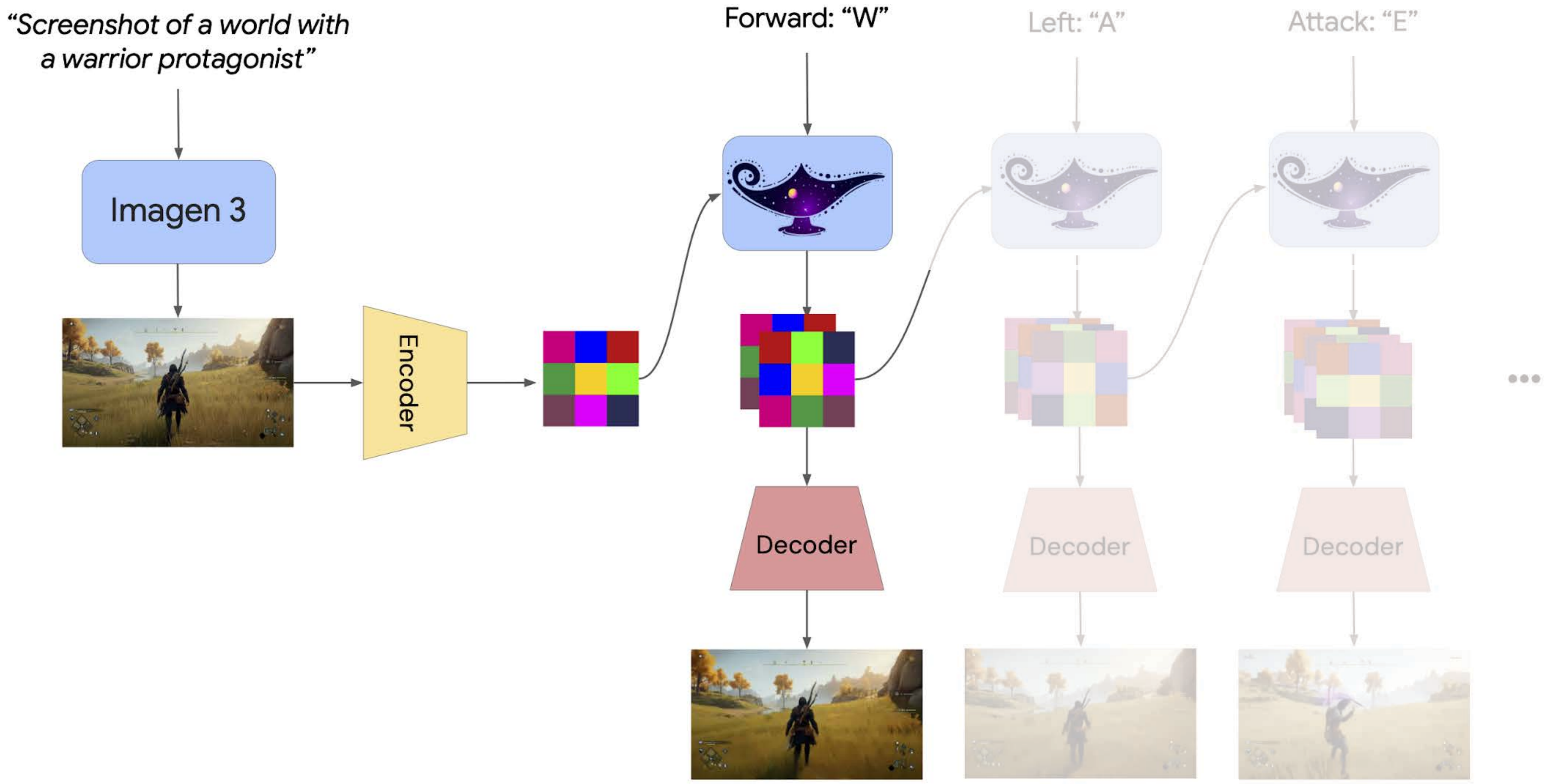
- Diffusion world model



https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

Genie 2: World Foundation Model

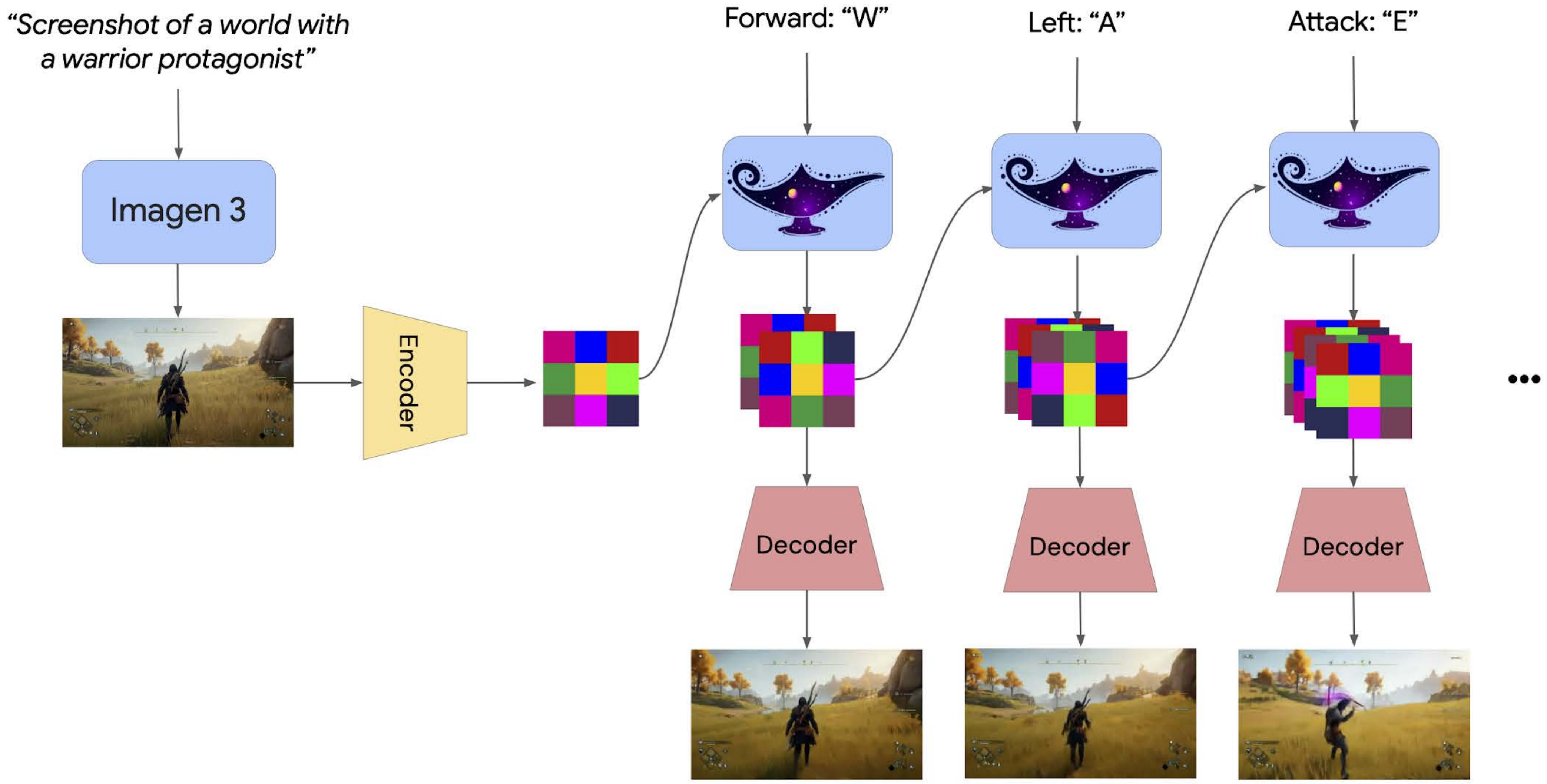
- Diffusion world model



https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

Genie 2: World Foundation Model

- Diffusion world model



https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/

NVIDIA Cosmos World Model

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

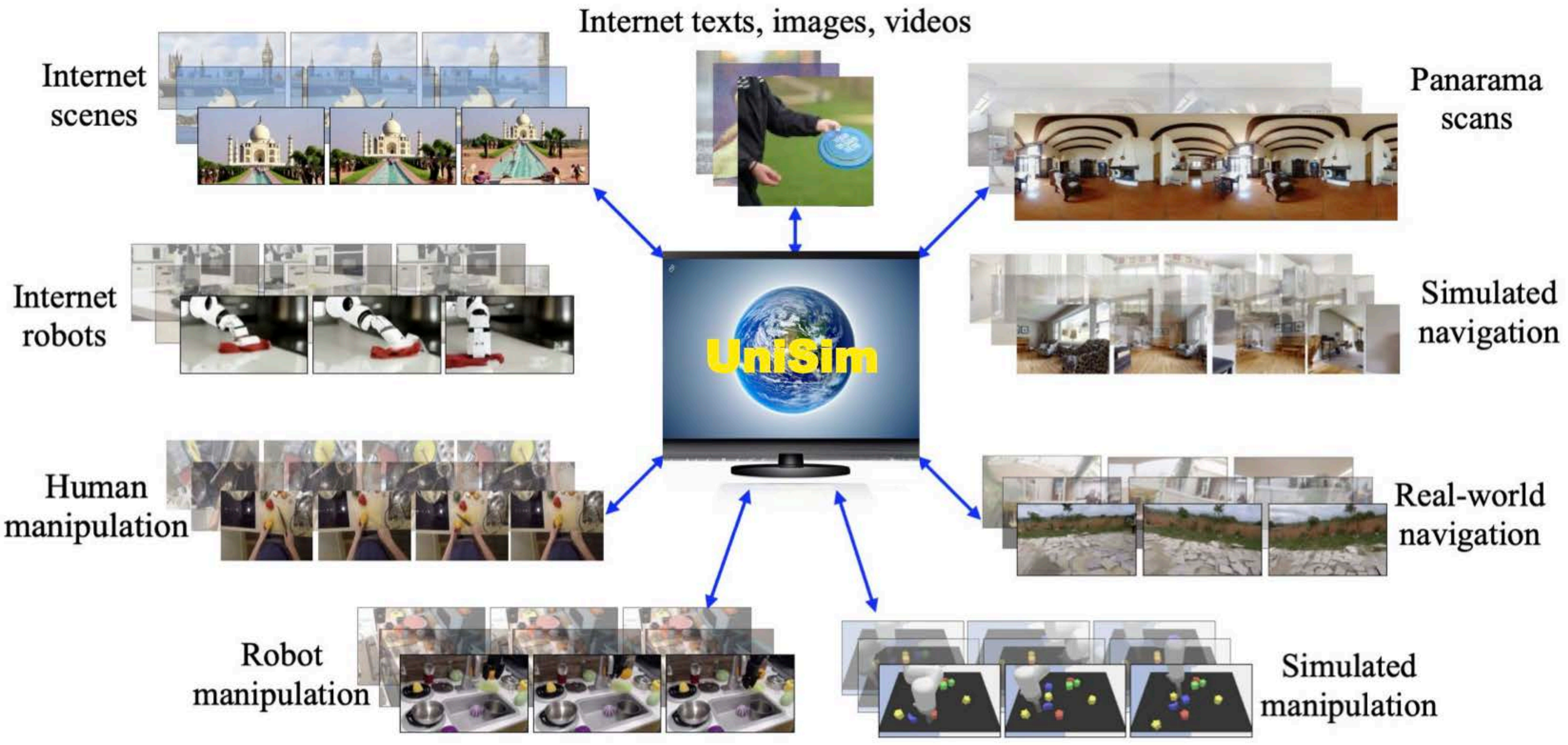
Cosmos Super

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

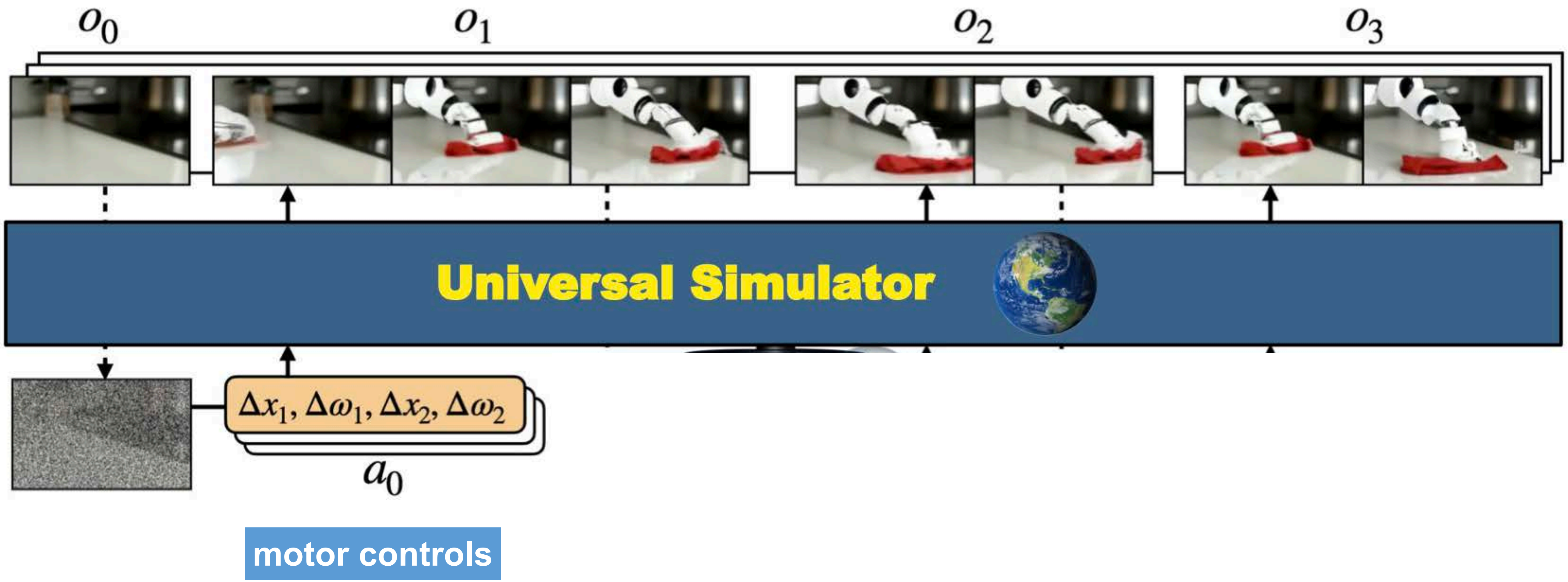
UniSim: Learning Interactive Real-World Simulators



https://universal-simulator.github.io/unisim/

UniSim: Learning Interactive Real-World Simulators

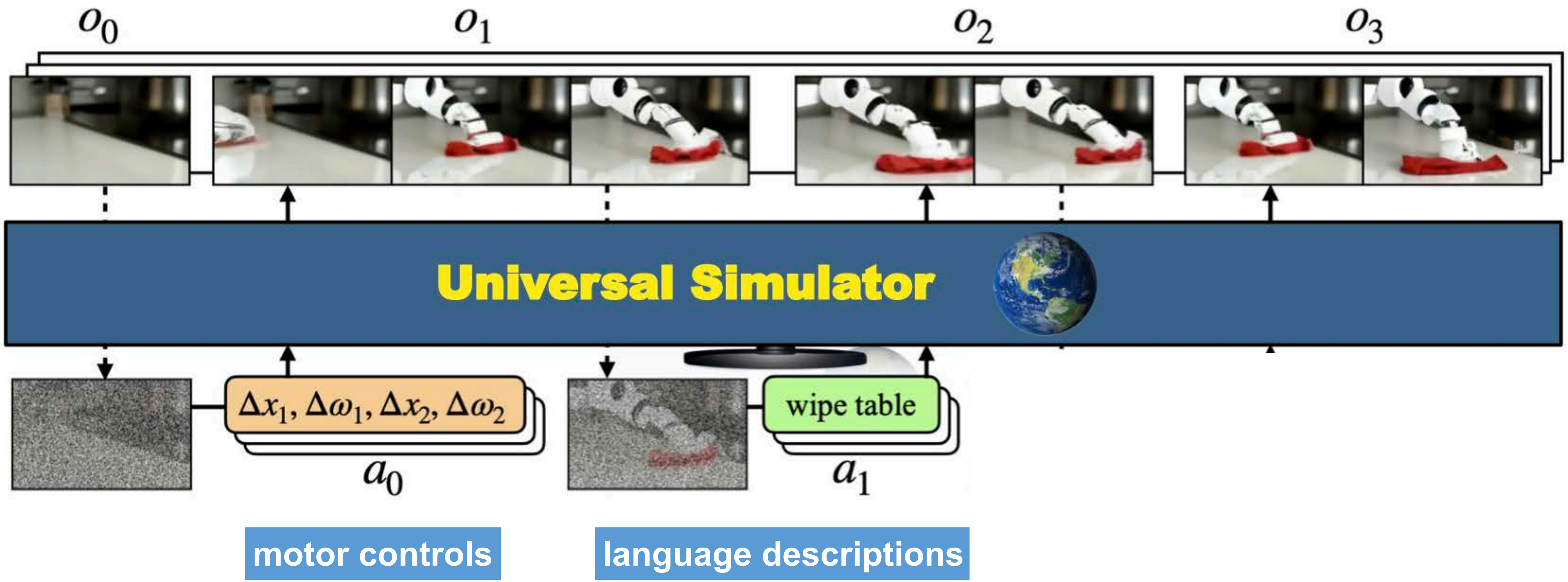
- A video diffusion model trained to predict the next (variable length) set of observation frames



https://universal-simulator.github.io/unisim/

UniSim: Learning Interactive Real-World Simulators

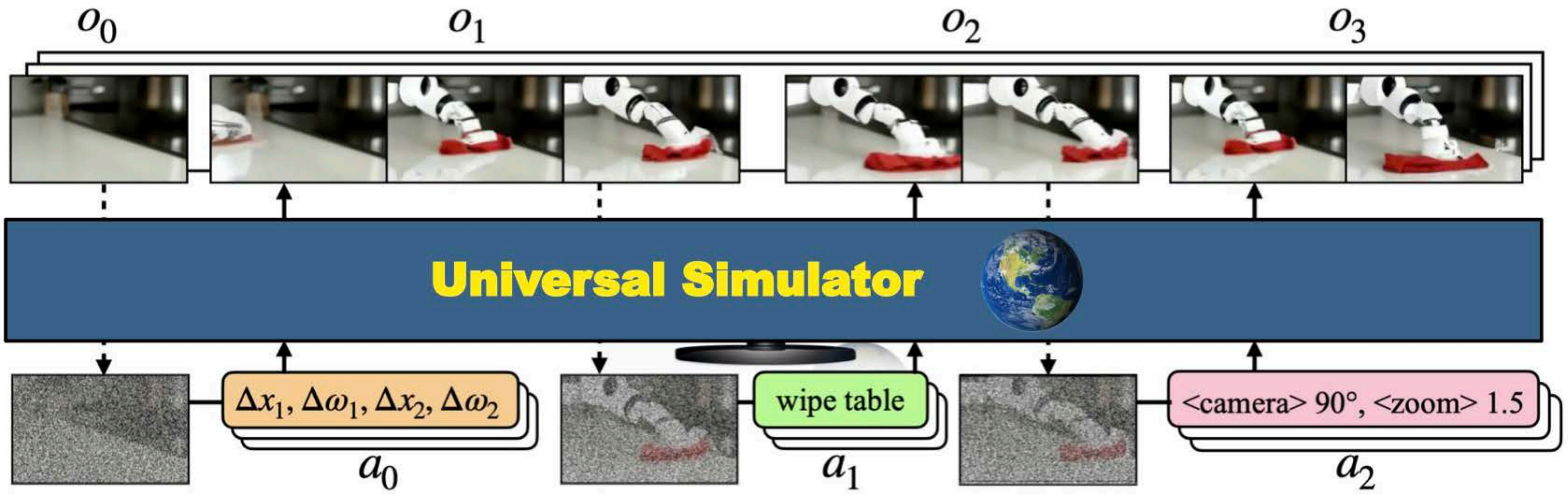
- A video diffusion model trained to predict the next (variable length) set of observation frames



https://universal-simulator.github.io/unisim/

UniSim: Learning Interactive Real-World Simulators

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

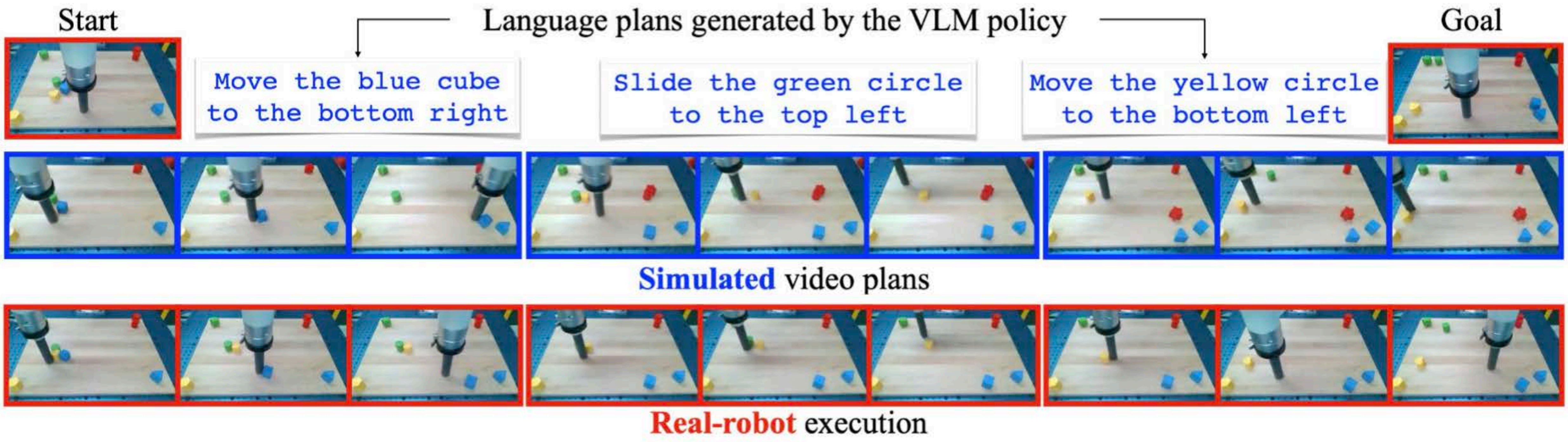


motor controls | language descriptions | camera actions

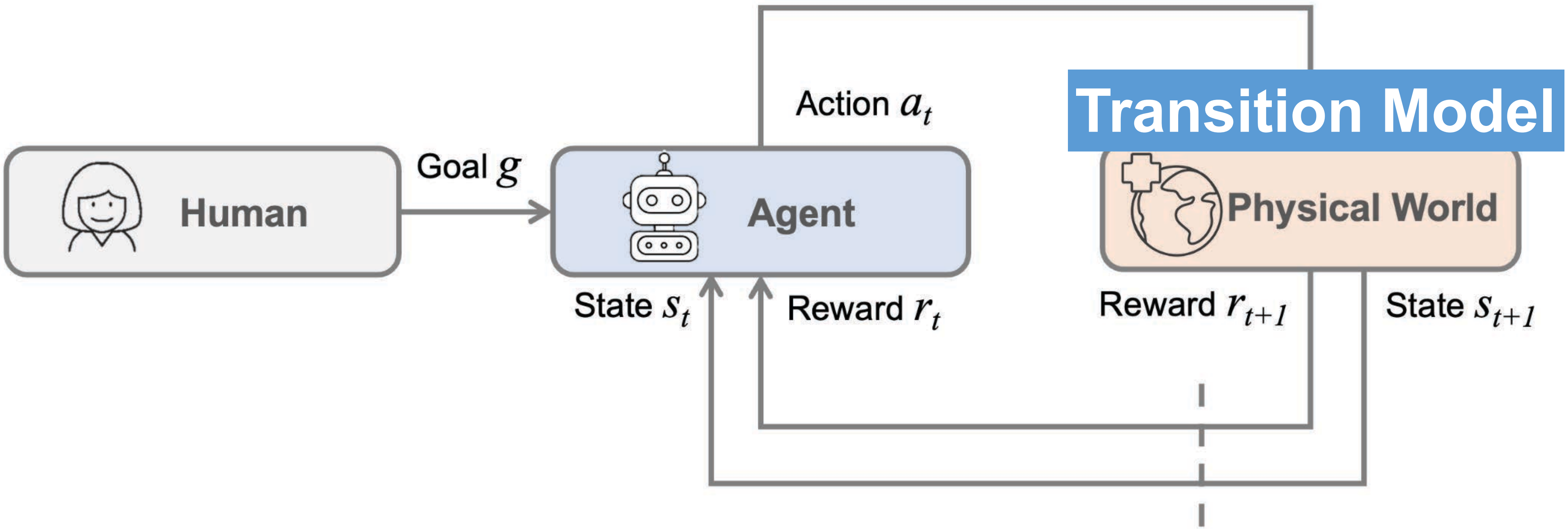
https://universal-simulator.github.io/unisim/

UniSim: Learning Interactive Real-World Simulators

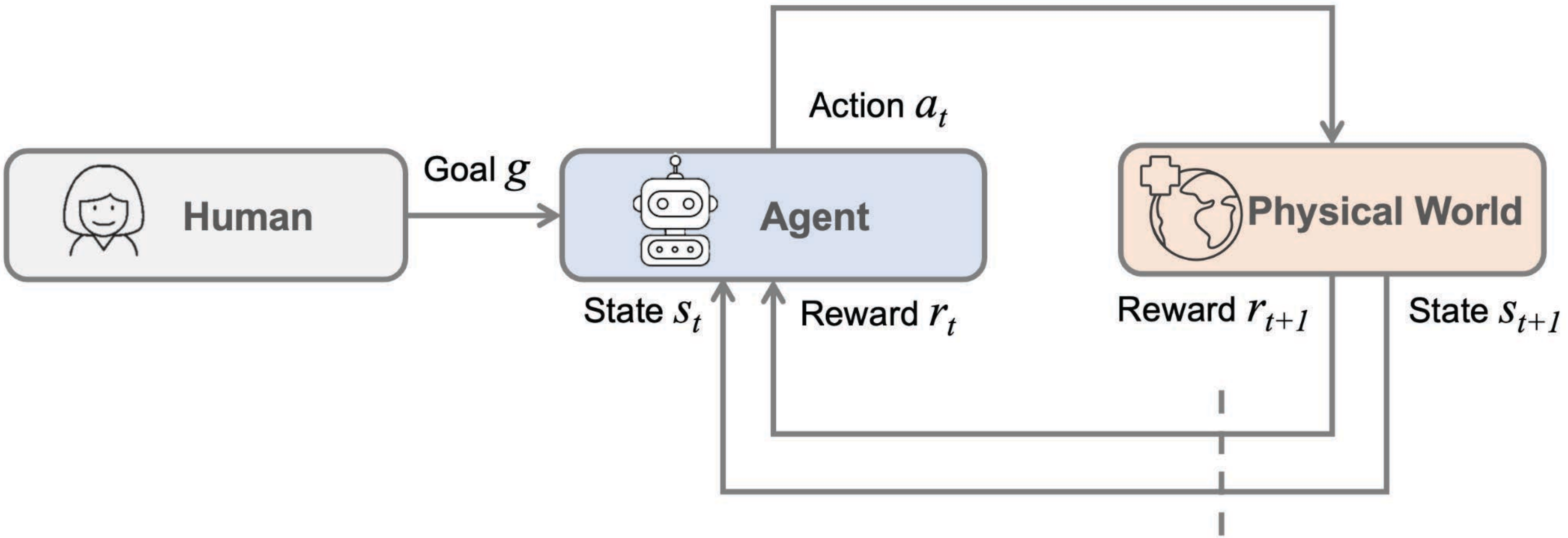
- Support Long-horizon generation



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Unified Virtual Agent

Key idea: MDP as sequence prediction

Goal

State

Action

Reward

Slides Credit: Wenlong

Key idea: MDP as sequence prediction

goal s_1 a_1 r_1 s_2 a_2 r_2 ... s_n a_n r_n

Goal

State

Action

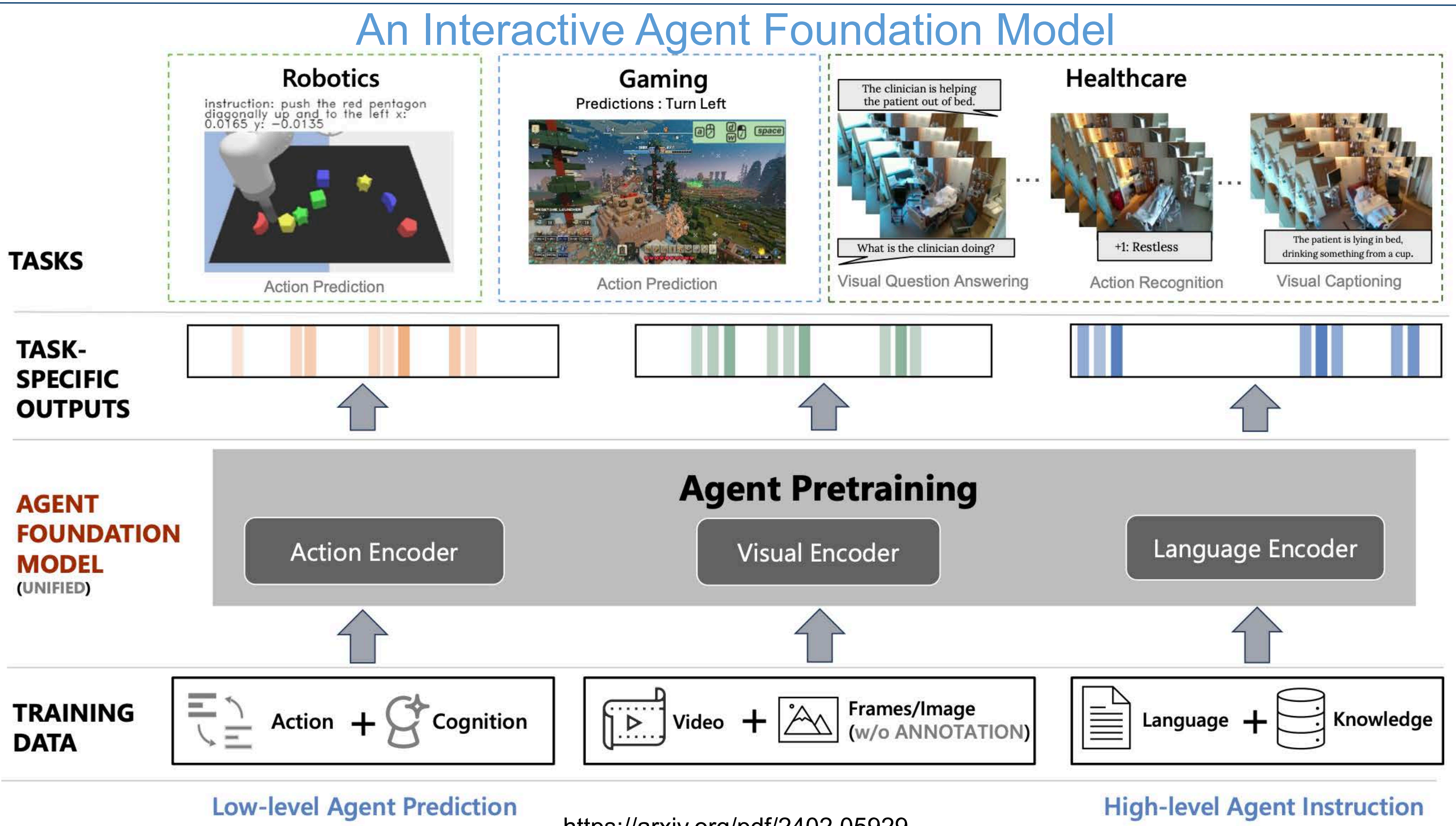
Reward

Option 1: Flat Concatenation

Option 2: Hierarchical Encoding

Slides Credit: Wenlong

Option 1: Flat Concatenation

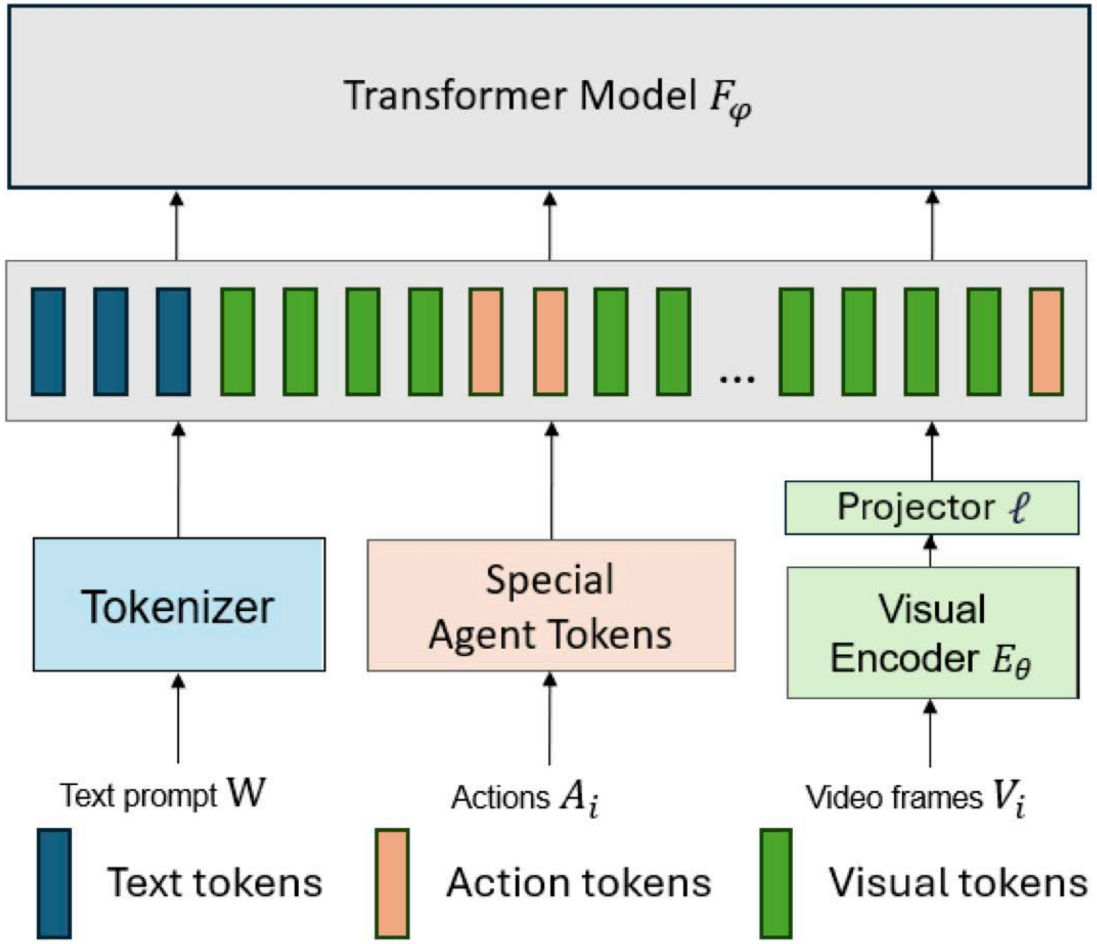


https://arxiv.org/pdf/2402.05929

Option 1: Flat Concatenation

An Interactive Agent Foundation Model

- Unified Tokenization

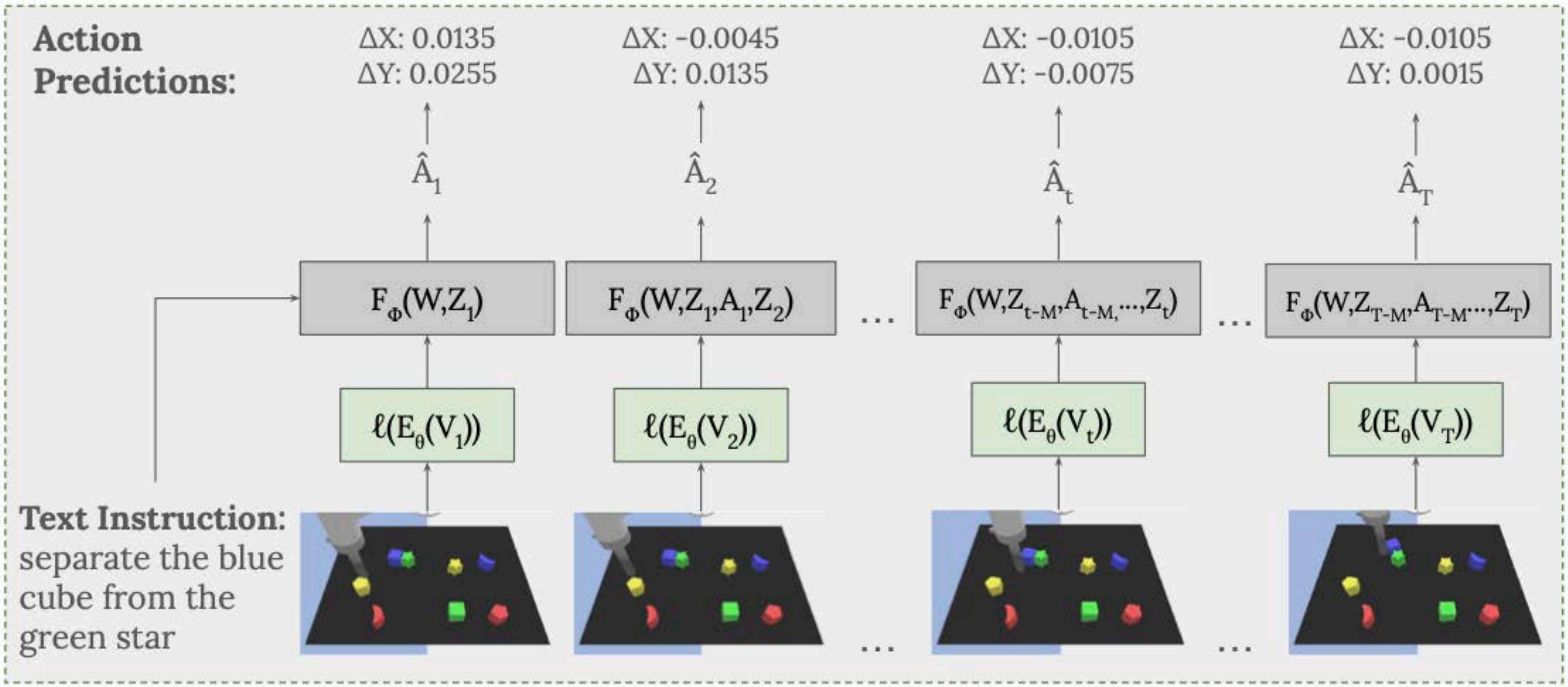


https://arxiv.org/pdf/2402.05929

Option 1: Flat Concatenation

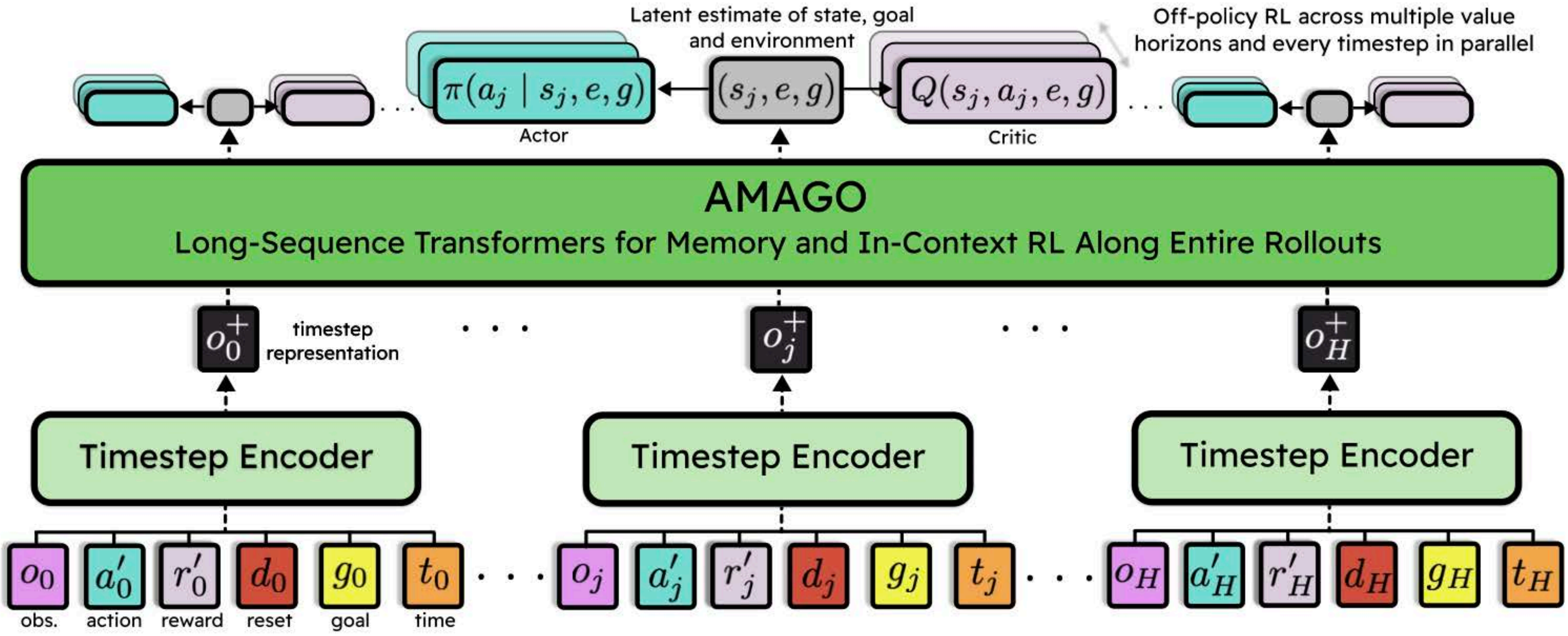
An Interactive Agent Foundation Model

- Pretraining Pipeline:



https://arxiv.org/pdf/2402.05929

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



https://arxiv.org/pdf/2310.09971

Multi-Agent Collaboration

An Envisioned Agent Society
Special dishes?
Planning to cook... I need: 1.Fish, 2.Sauce ...
Kitchen
Task planning and solving
Ordering dishes and cooking
Making lanterns... We need: *10= *100+ *50...
Calculating costs ...
Acting with tools
Discussing decoration
Outdoors
Sure! I play the guitar!
Let's play a song!
I play ...
Concert
Cooperation
Band performing
User
Let me experience the festival in this world...

https://openreview.net/pdf?id=EnXJfQqy0K

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

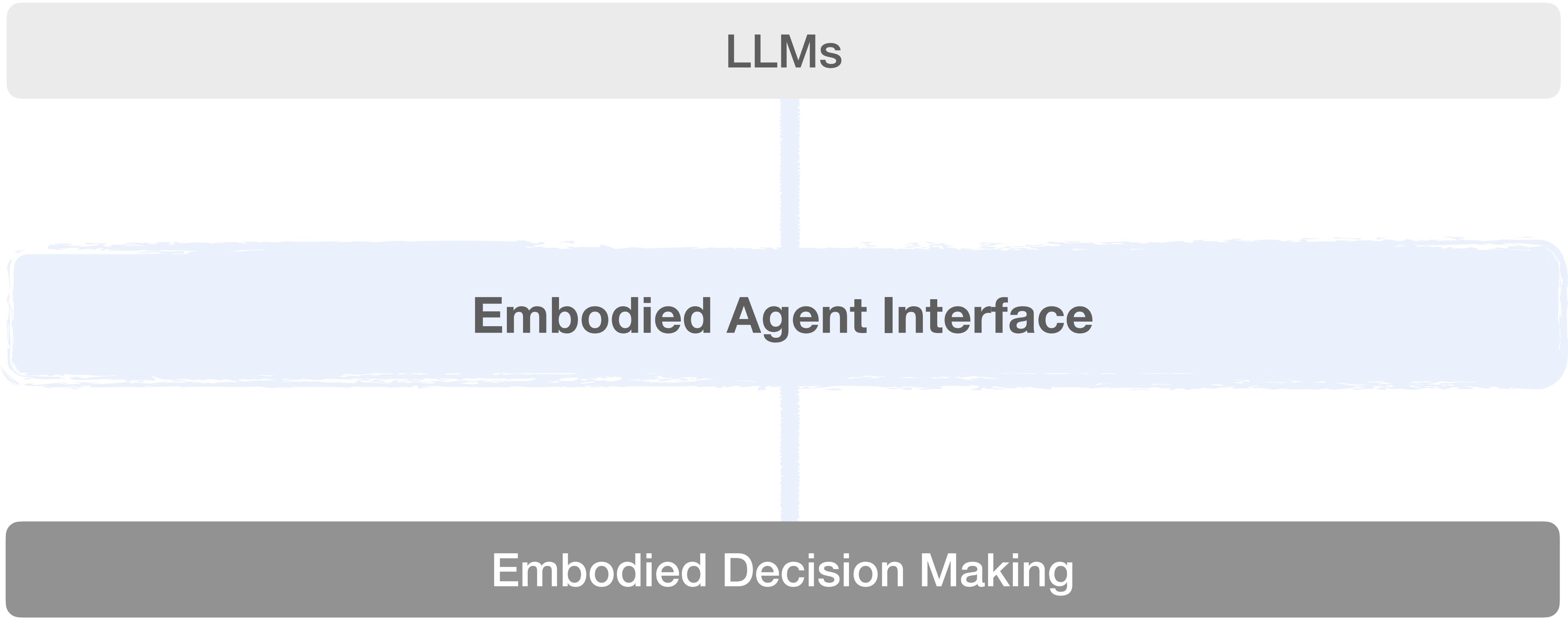
Different Input/Output

So we need **Standardization!**

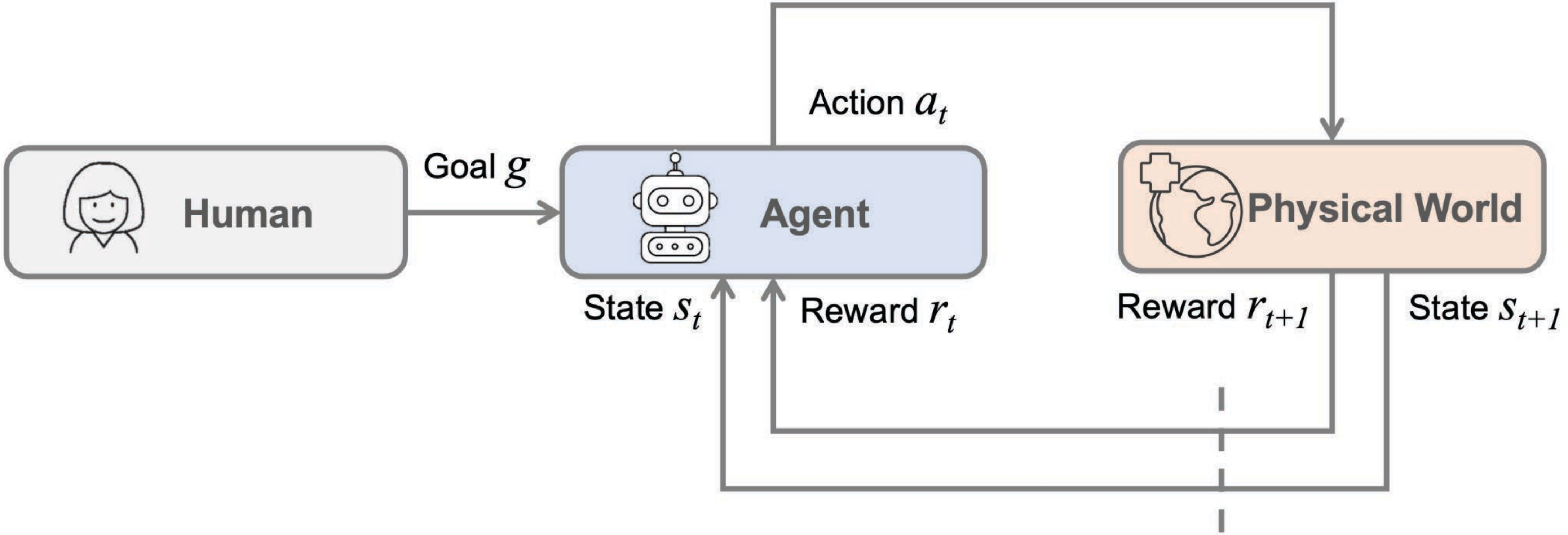
Embodied Agent Interface

Embodied Agent Interface

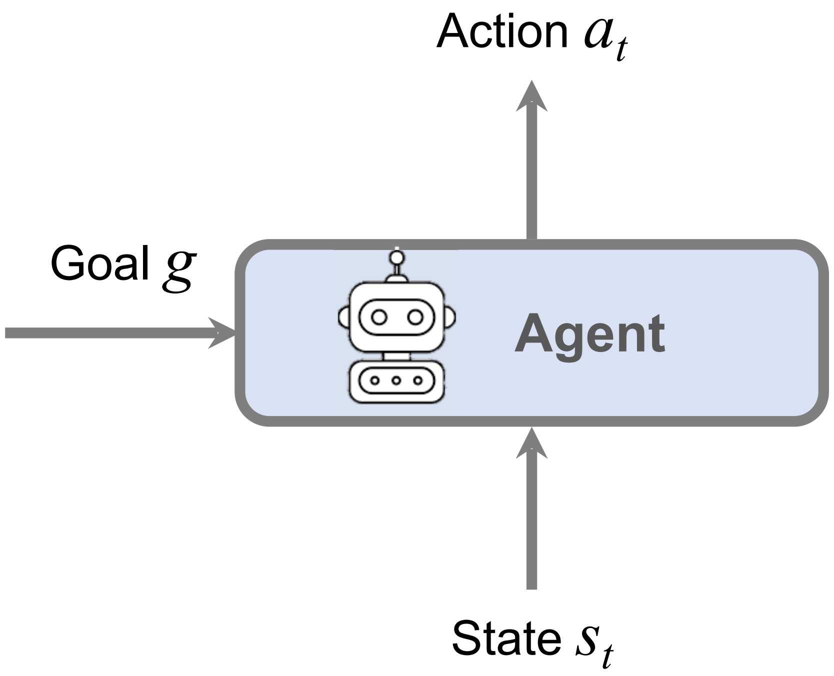
to benchmark **LLMs** for Embodied Decision Making


LLMs
Embodied Agent Interface
Embodied Decision Making

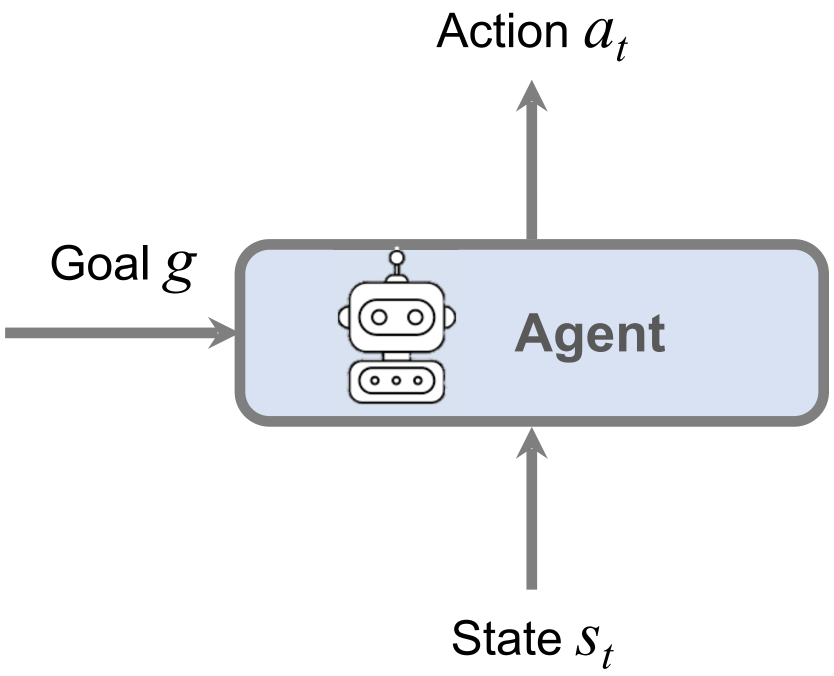
Let us go back to MDPs (Markov Decision Processes)



Markov Decision Process



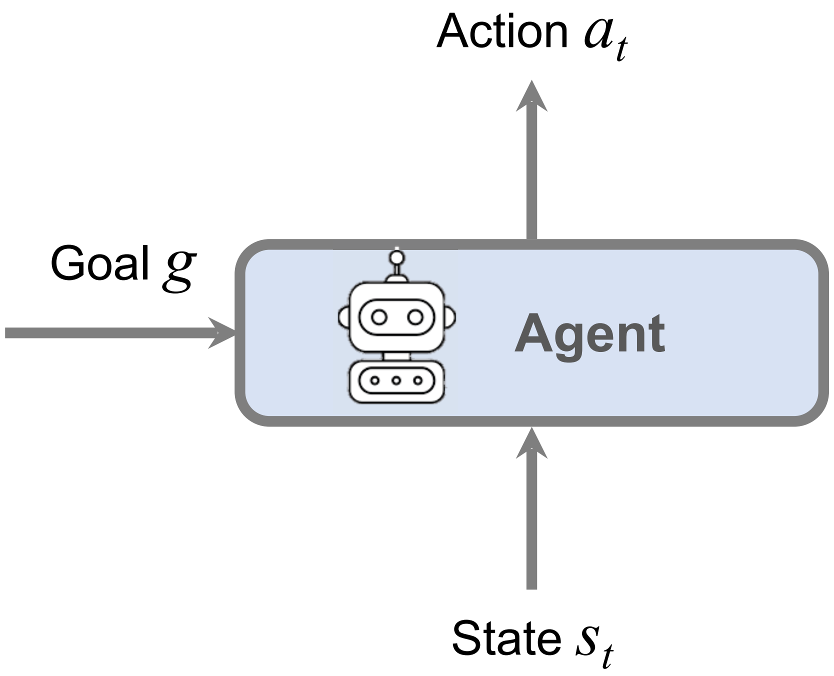
Markov Decision Process



Input
$<g, s_0>$

Goal Interpretation

Markov Decision Process



Output

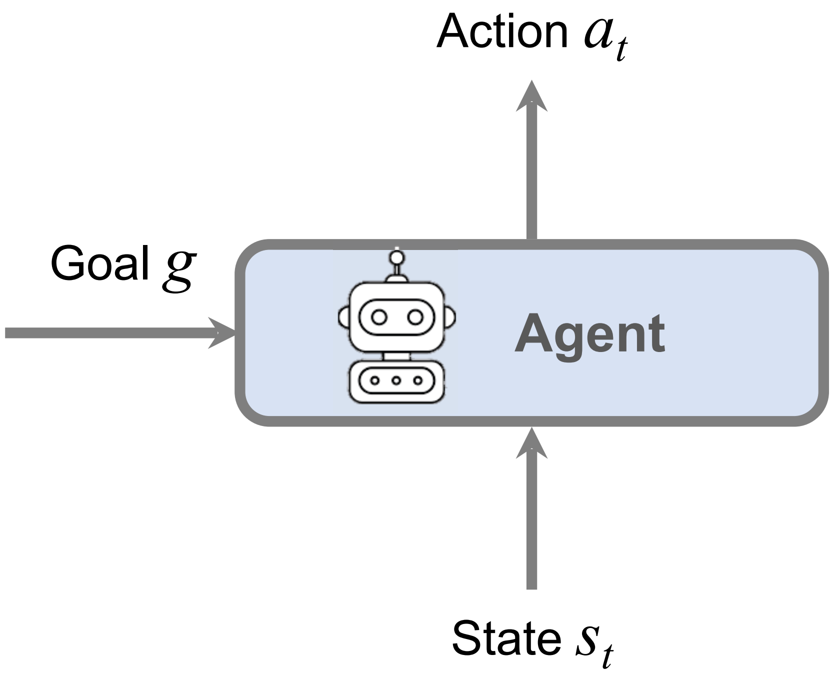
$<a_1, a_2, \ldots, a_t>$ **Action Sequencing**

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

Input

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

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**

Embodied Agent Interface

Action Sequencing

Subgoal Decomposition

Transition Modeling

Goal Interpretation

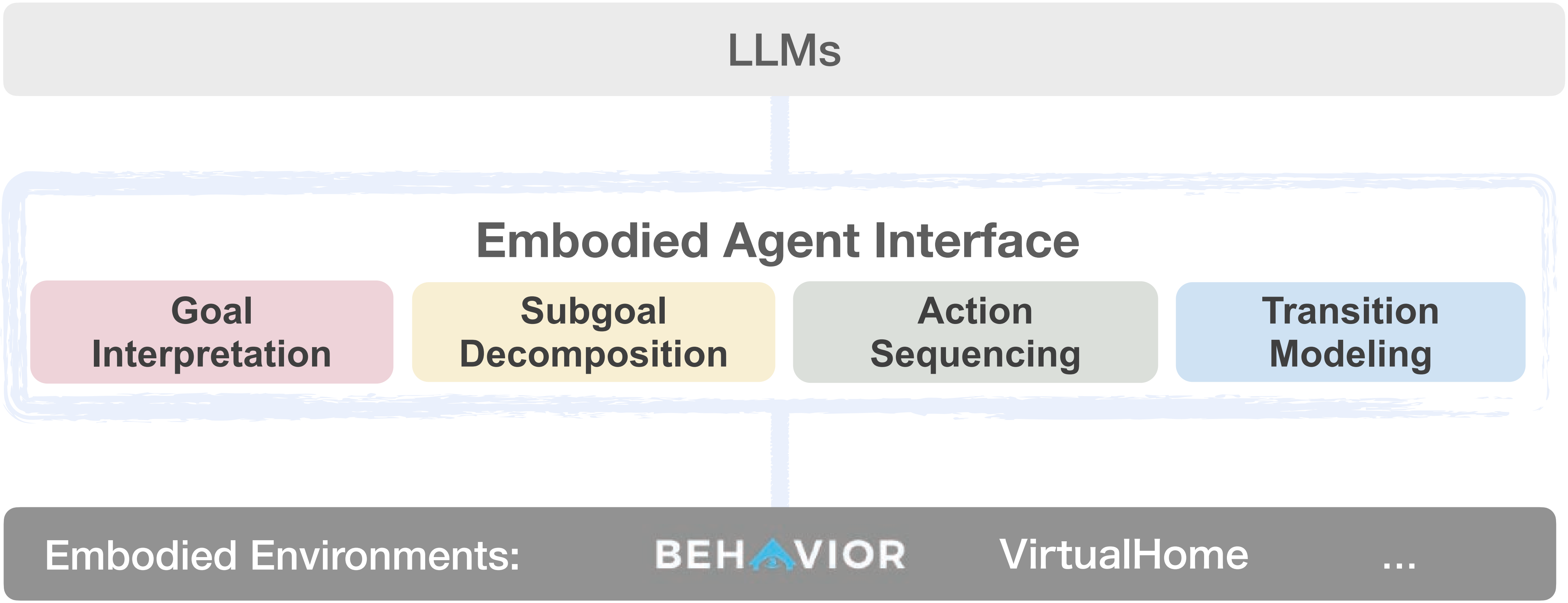
Embodied Agent Interface

Goal Interpretation

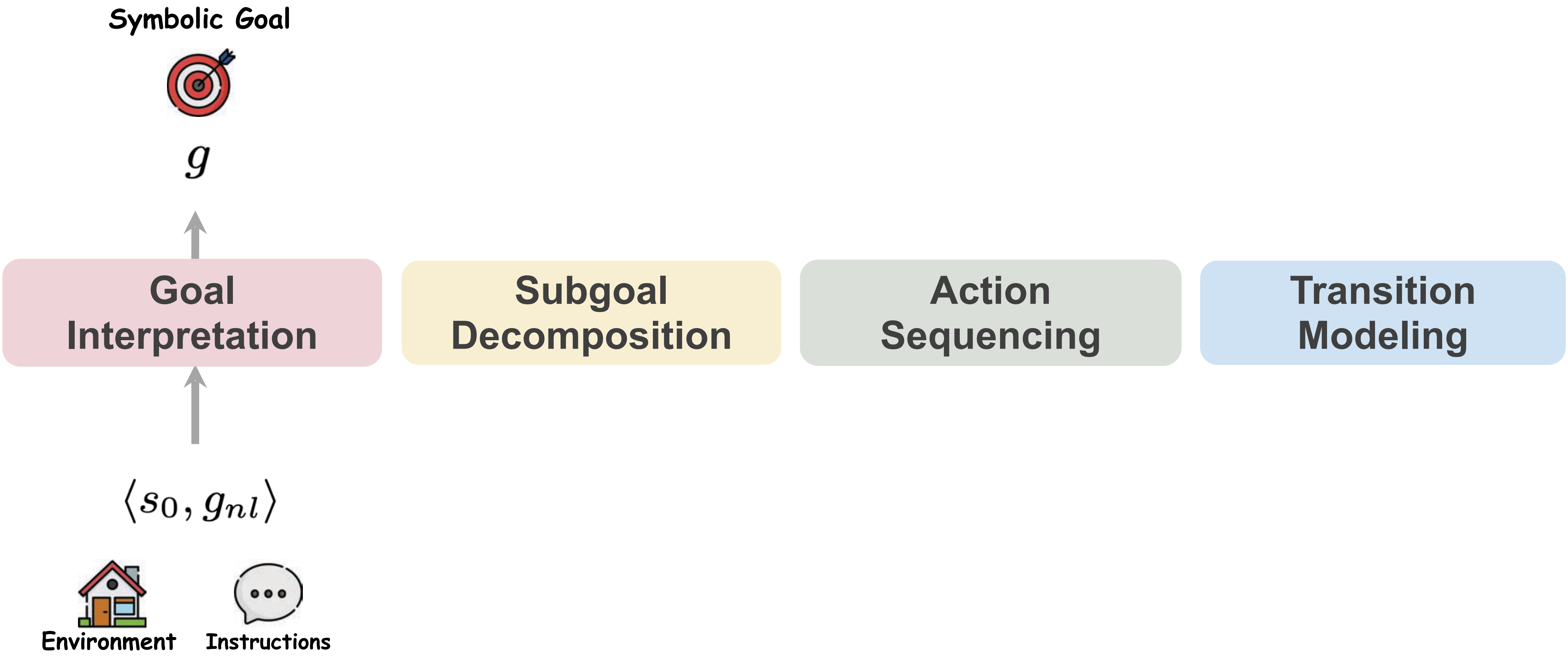
Subgoal Decomposition

Action Sequencing

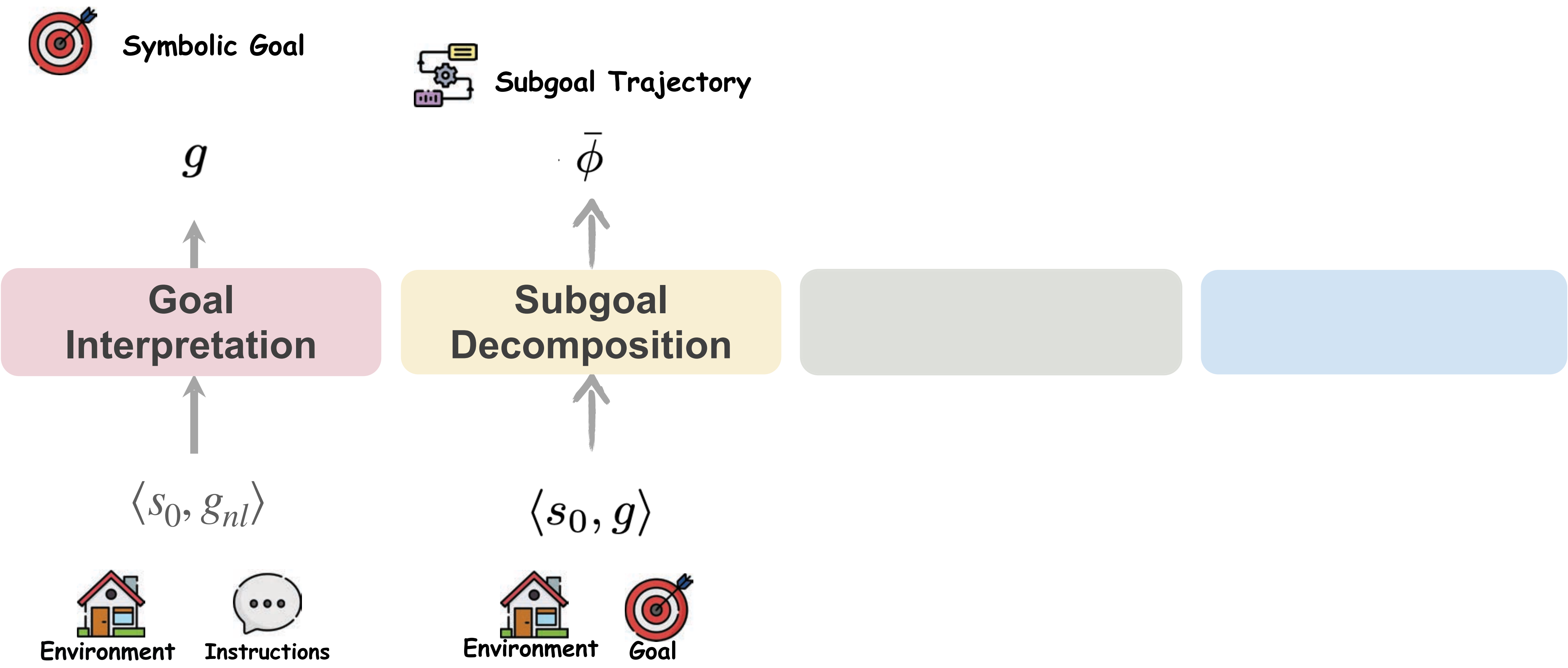
Transition Modeling


LLMs
Embodied Agent Interface
Goal Interpretation
Subgoal Decomposition
Action Sequencing
Transition Modeling
Embodied Environments:
BEHAVIOR
VirtualHome
...

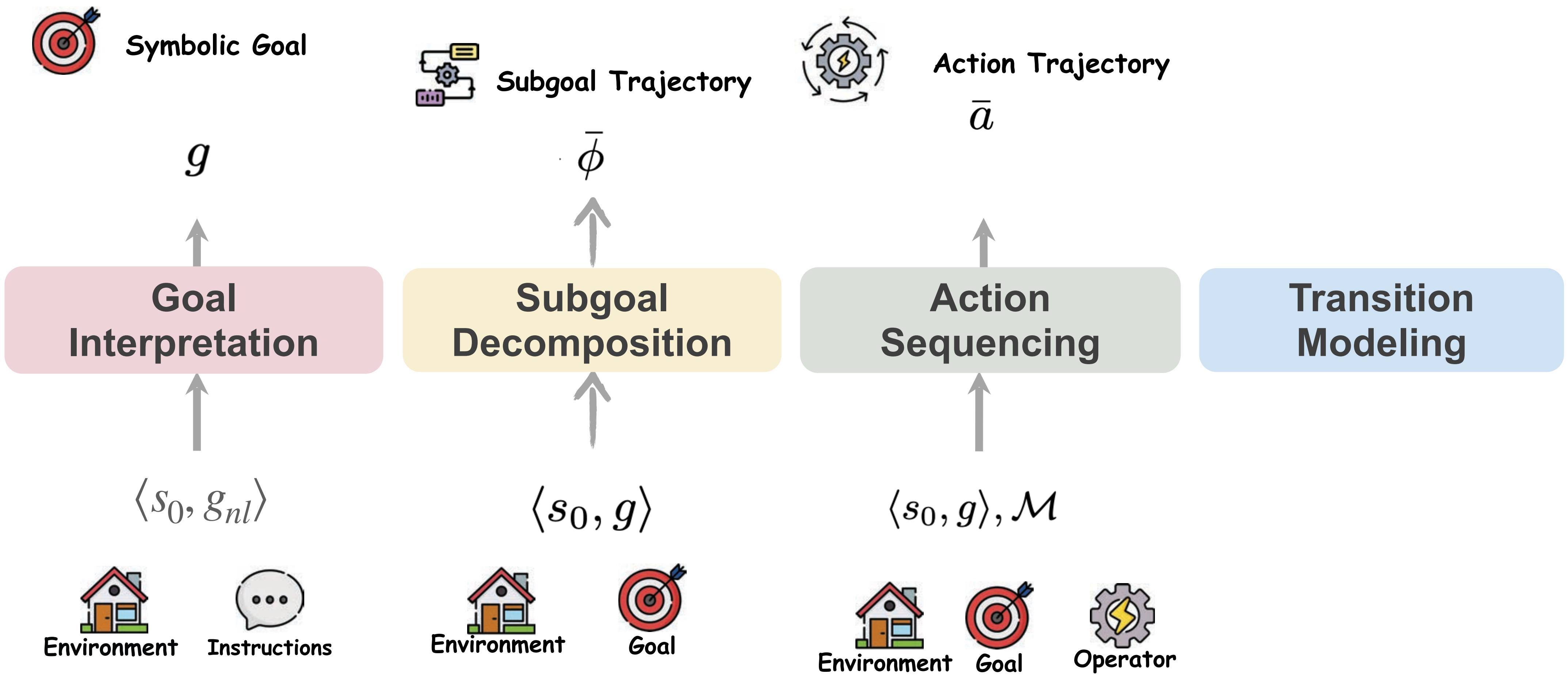
Standardization of modules and interfaces



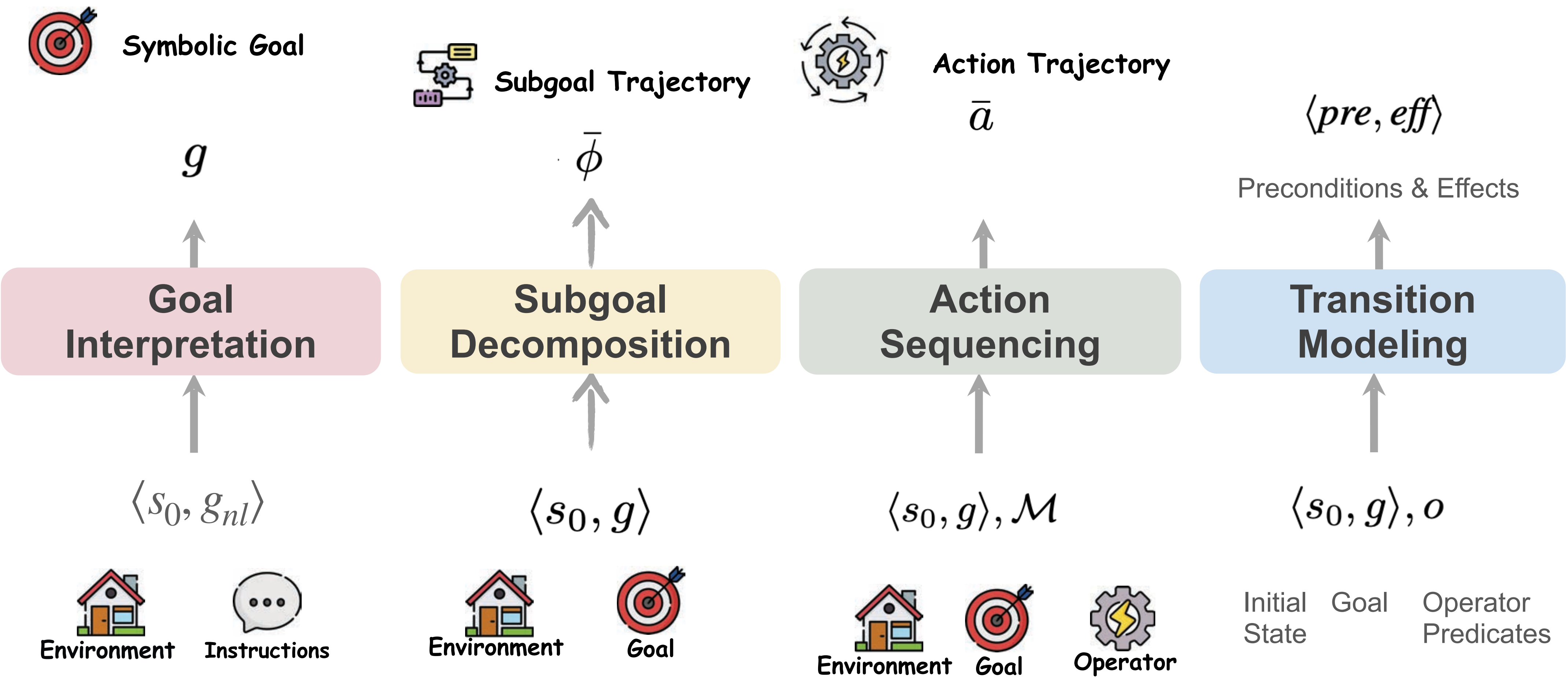
Standardization of modules and interfaces

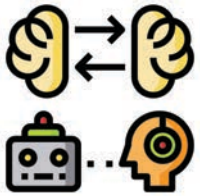


Standardization of modules and interfaces



Standardization of modules and interfaces





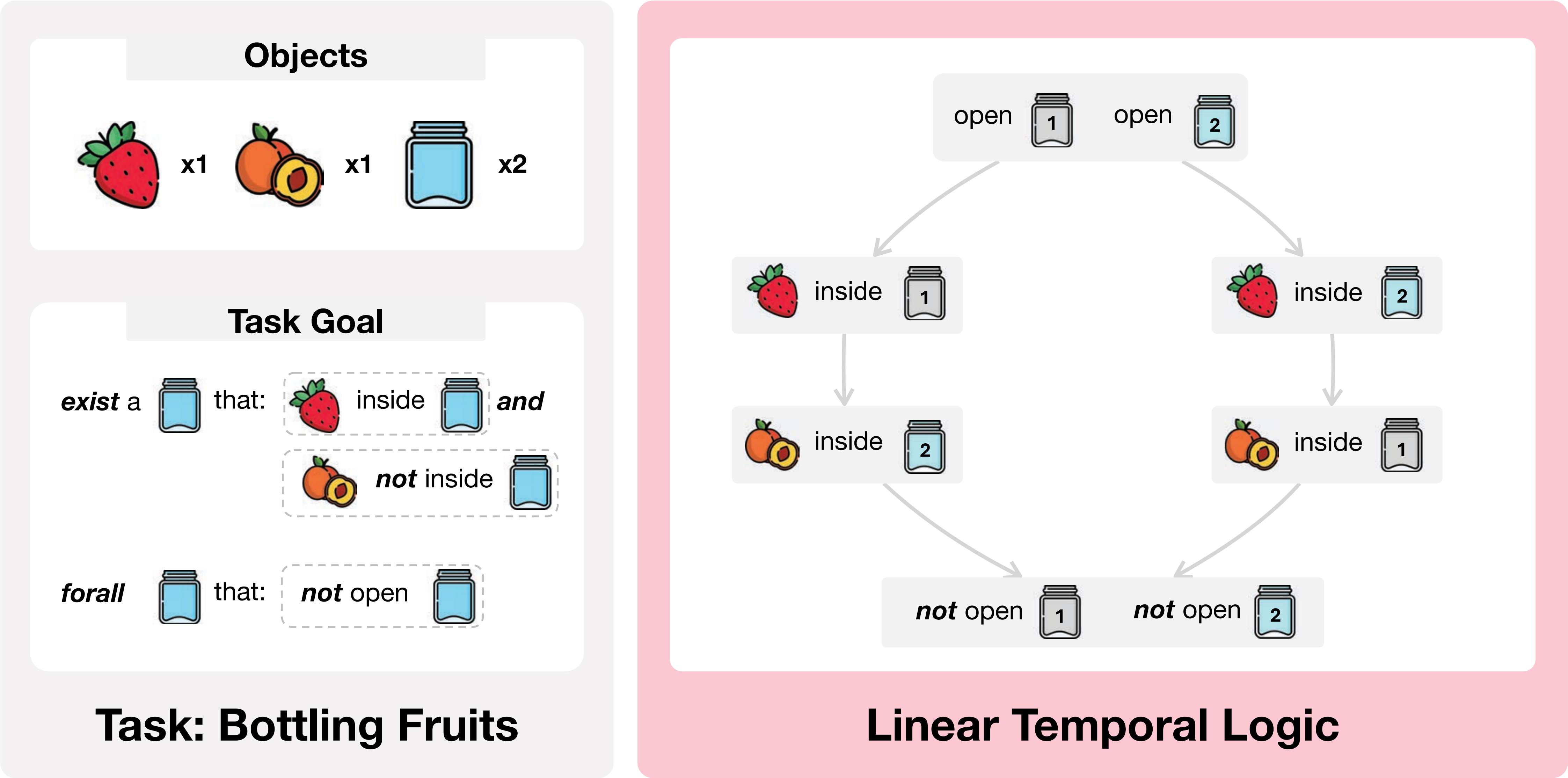
Standardize
Modules and Interfaces

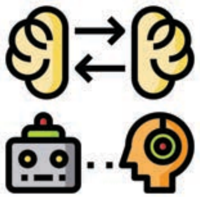
4 modules
438 tasks
1475 goals

Standardize
Goal Specifications

Linear
Temporal
Logic

Standardize Goal Specifications





Standardize Modules and Interfaces

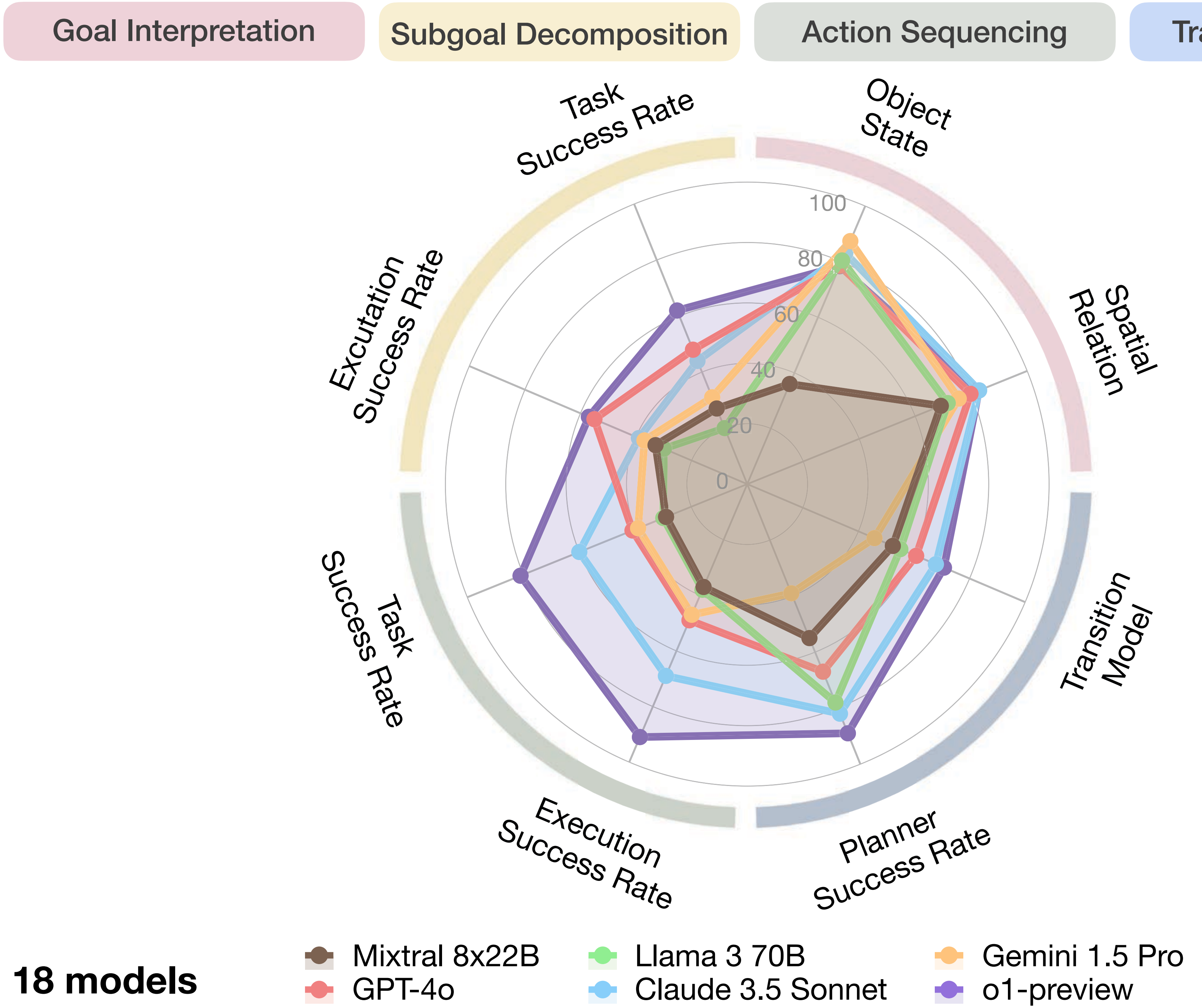
4 modules
438 tasks
1475 goals

Standardize Goal Specifications

Linear
Temporal
Logic

Standardize Fine-grained Metrics

18 models
42 metrics
100+ page analysis



Goal Interpretation
Subgoal Decomposition
Action Sequencing
Transition Modeling
Object State
Spatial Relation
Transition Model
Planner Success Rate
Execution Success Rate
Task Success Rate
Excutation Success Rate
Task Success Rate
100
80
60
40
20
0
18 models
Mixtral 8x22B
GPT-4o
Llama 3 70B
Claude 3.5 Sonnet
Gemini 1.5 Pro
o1-preview

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	25min	Jiayuan Mao
High-level & Low-level Decision Making	50min	Wenlong Huang
Break		
4. Robotic Foundation Models	30min	Yunzhu
5. Remaining Challenges	15min	Yunzhu
QA	30min	