



Foundation Models Meet Embodied Agents



Manling Li
Northwestern



Yunzhu Li
Columbia



Jiayuan Mao
MIT



Wenlong Huang
Stanford



**Northwestern
University**



COLUMBIA



**Stanford
University**



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



Part I: Motivation and Overview

Manling Li, Assistant Professor at Northwestern University

AAAI Tutorial: Foundation Models Meet Embodied Agents



Northwestern
University



COLUMBIA



S **Stanford**
University

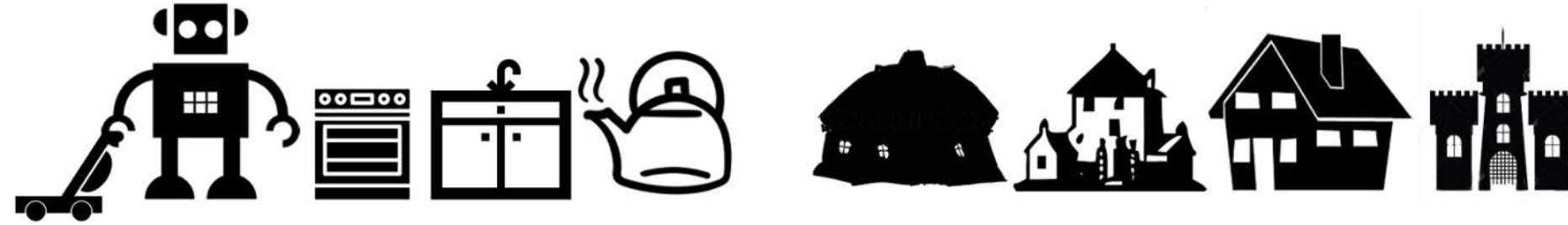
AAAI 2025 Tutorial T04

Time: 2025-02-25 8:30-12:30

Location: 118A Pennsylvania Convention Center

What is a generalist agent?

What is a generalist agent?



Having a robot that can do many tasks, across many environments.

BEHAVIOR-1K

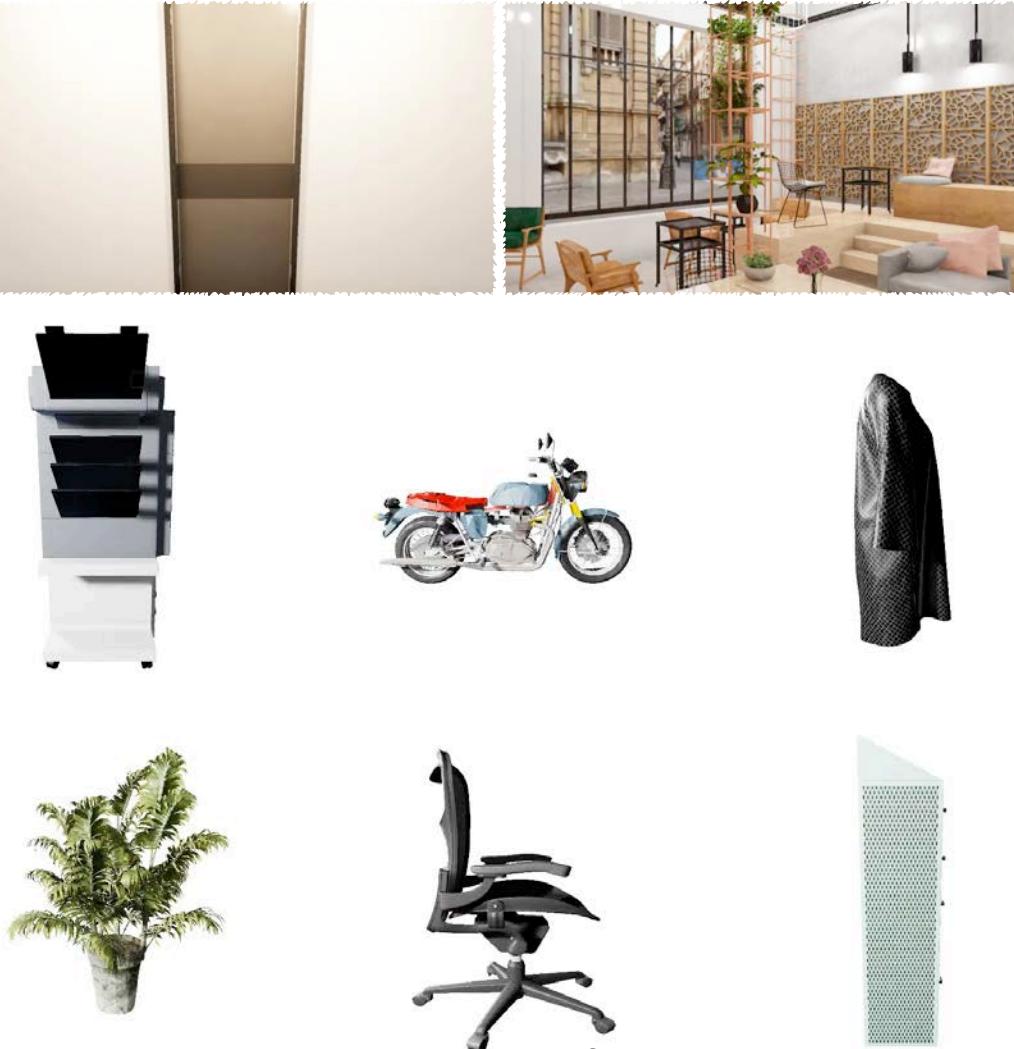
simulating and benchmarking robot tasks that **matter** to humans

<https://behavior.stanford.edu/>

Observation



State: 3D assets & states



Transition Model



tasks that **matter**

What would you like a robot to help you with?



Cleaning the floor?

Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?



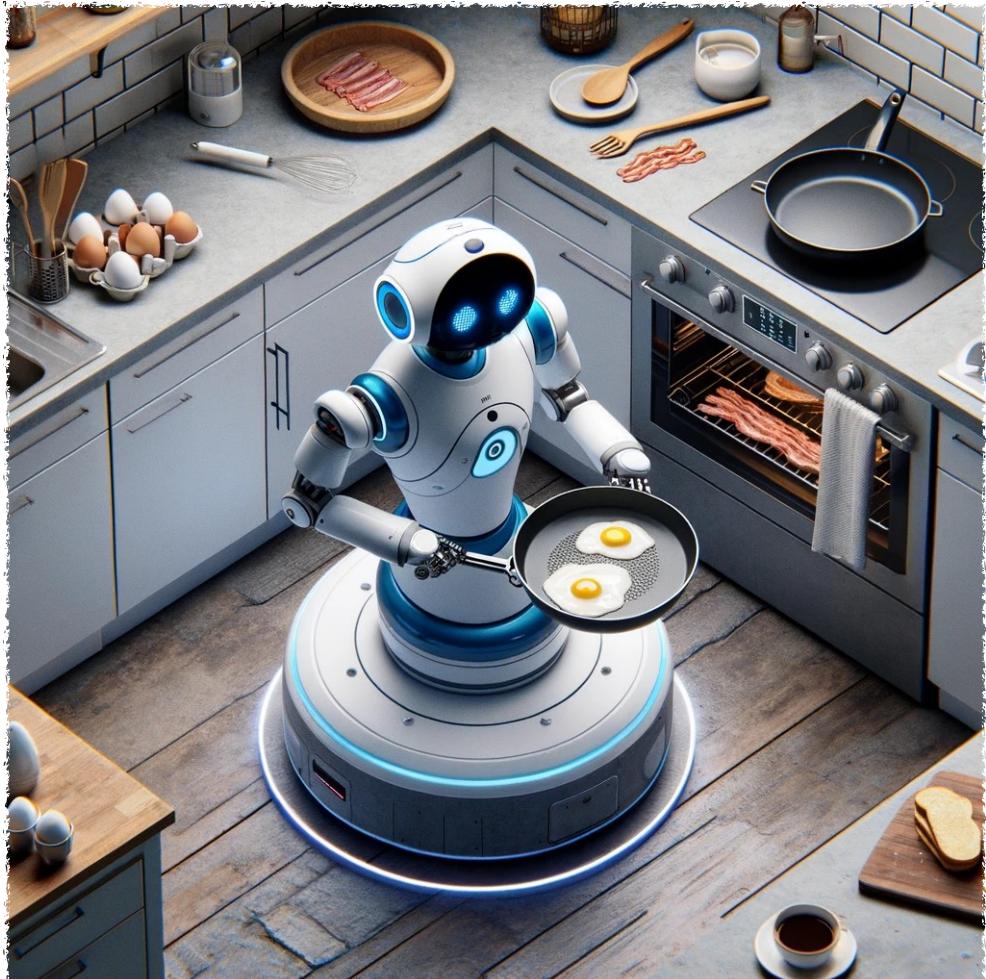
Folding Laundry?



Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?



Cooking Breakfast?



Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?



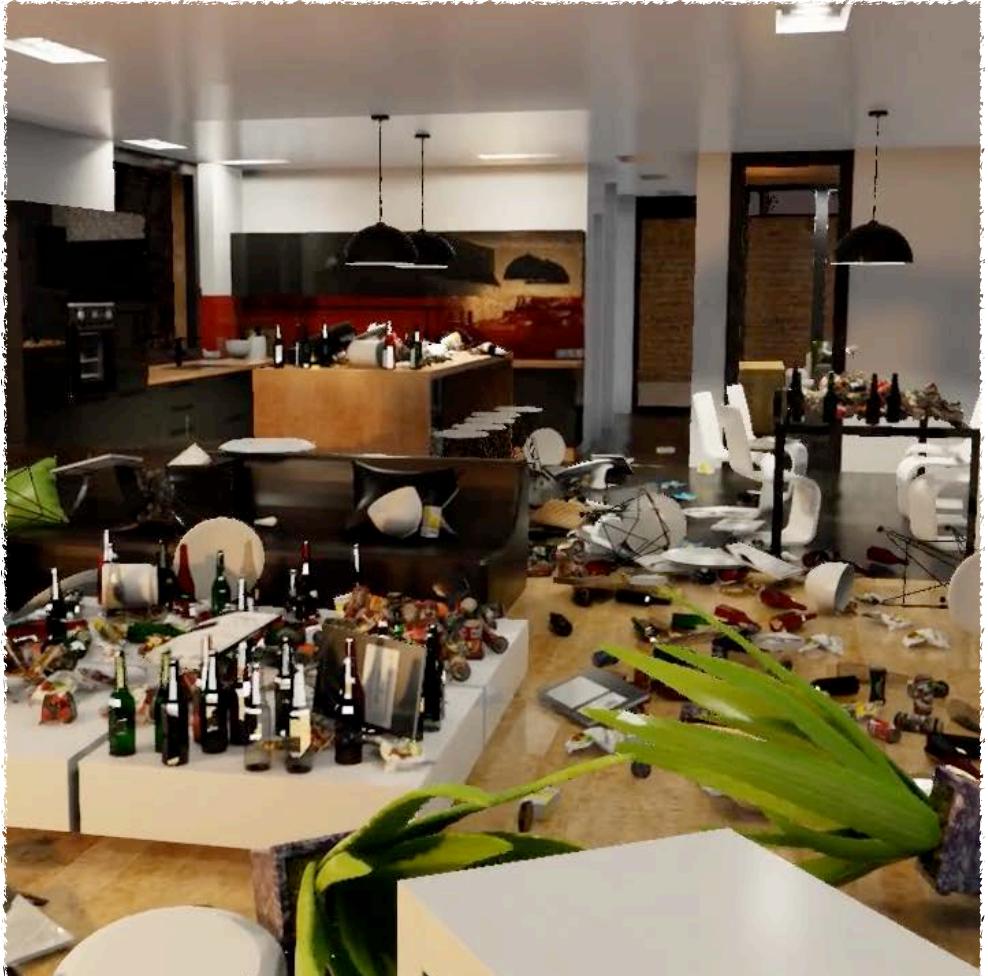
Opening gifts?



Images Generated by DALL-E 3

tasks that **matter**

What would you like a robot to help you with?



Cleaning after a wild party?

Images Generated by DALL-E 3

Robotic Manipulation



Human: pack the yoshi figure in the brown box



LEO:

$$P = [0.1, -0.2, 0]$$

$$R = [0, 0, 0, 1]$$

Task Planning



Human: Organize and tidy up the bedroom.

LEO:

1. *Clean the floor by sweeping to remove ...*
2. *Make the bed ...*
3. *Fold and organize ...*

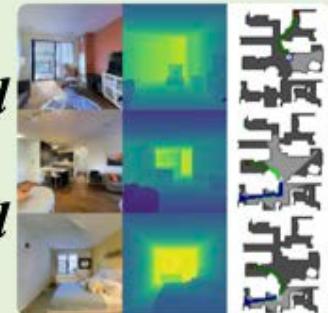
Embodied Navigation



Human: Find the pillow

LEO:

1. *Move forward*
2. *Turn right*
3. *Move forward*
4. ...



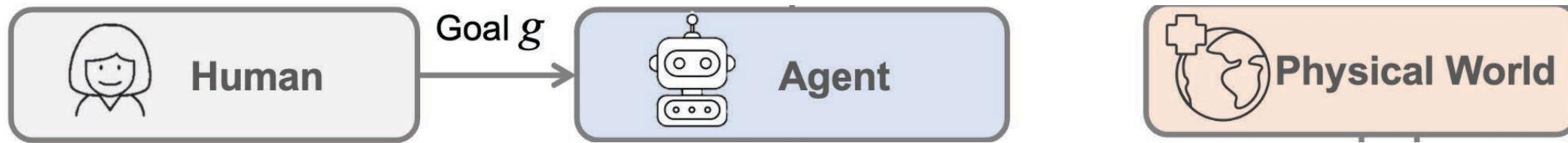
What is “embodied decision making”?

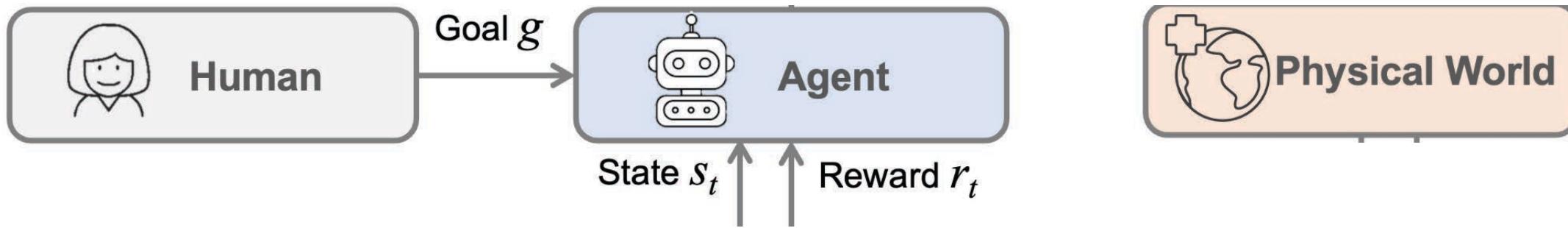
Can we leverage MDP as a guiding principle to categorize “foundation models”?

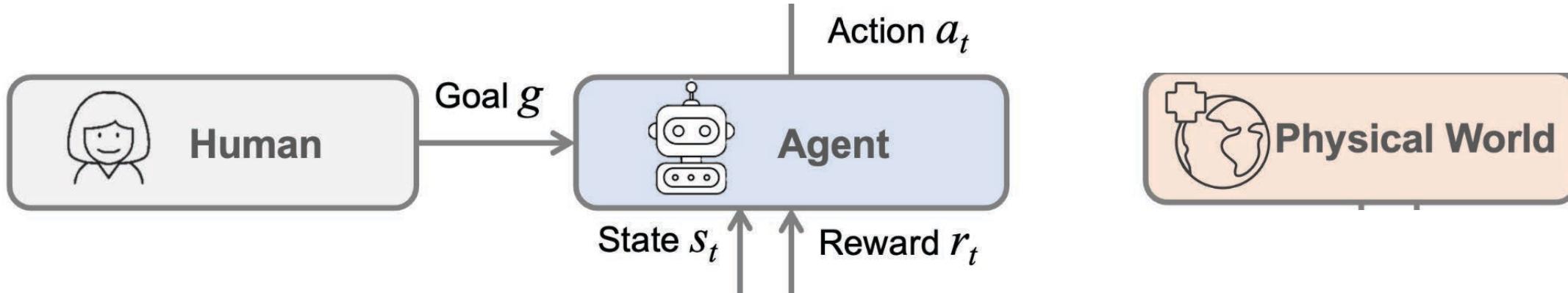
Let us go back to MDPs (Markov Decision Processes)

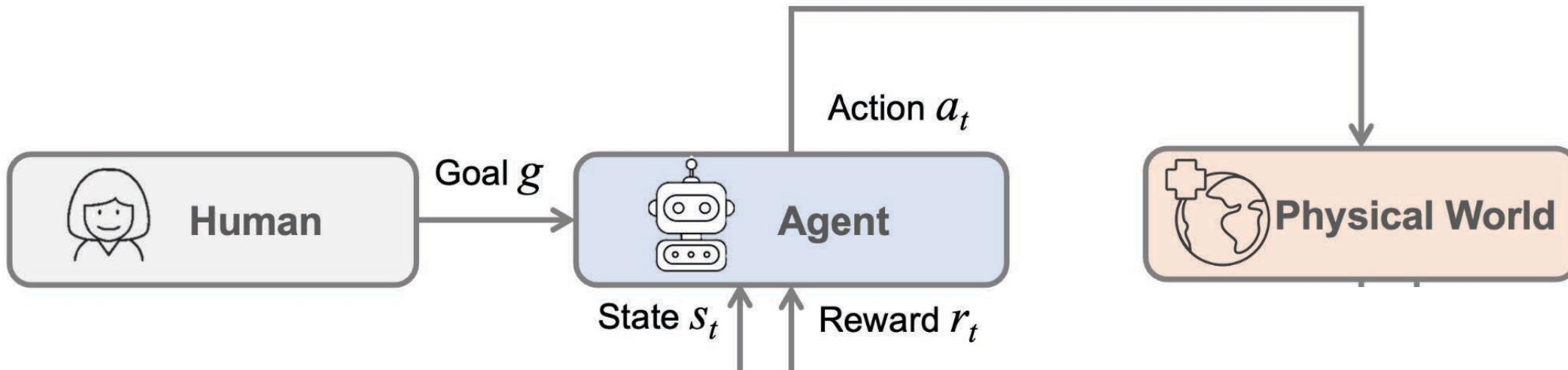


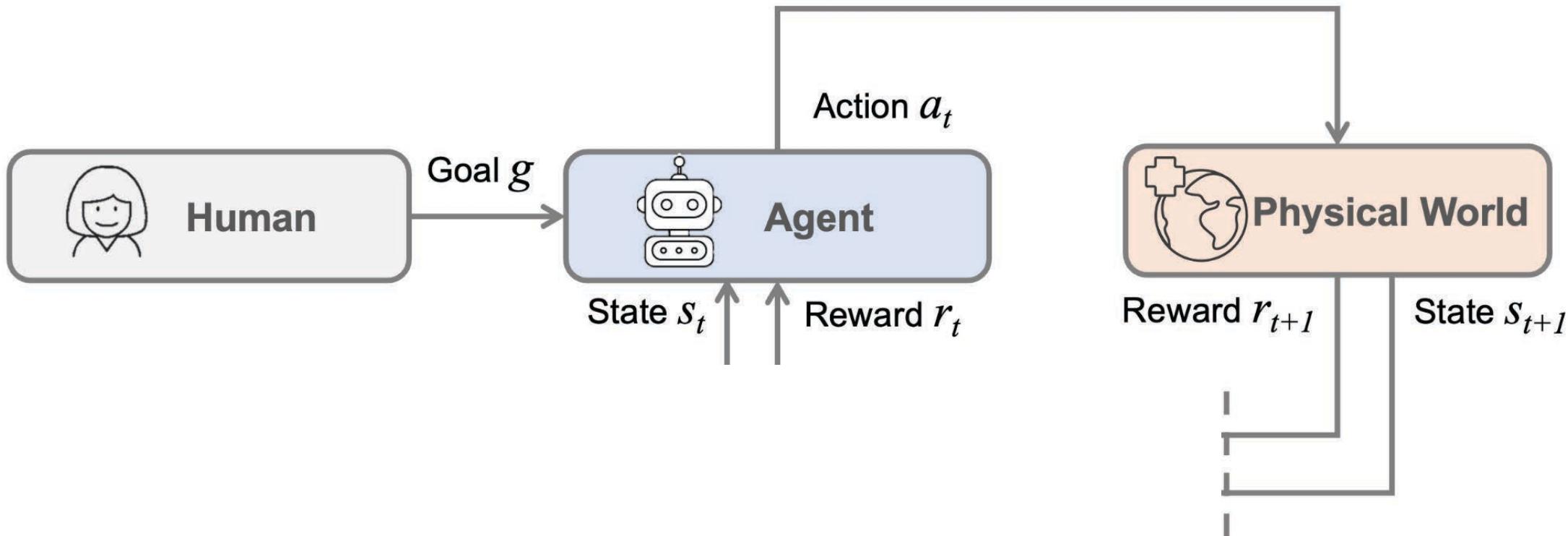
Let us go back to MDPs (Markov Decision Processes)

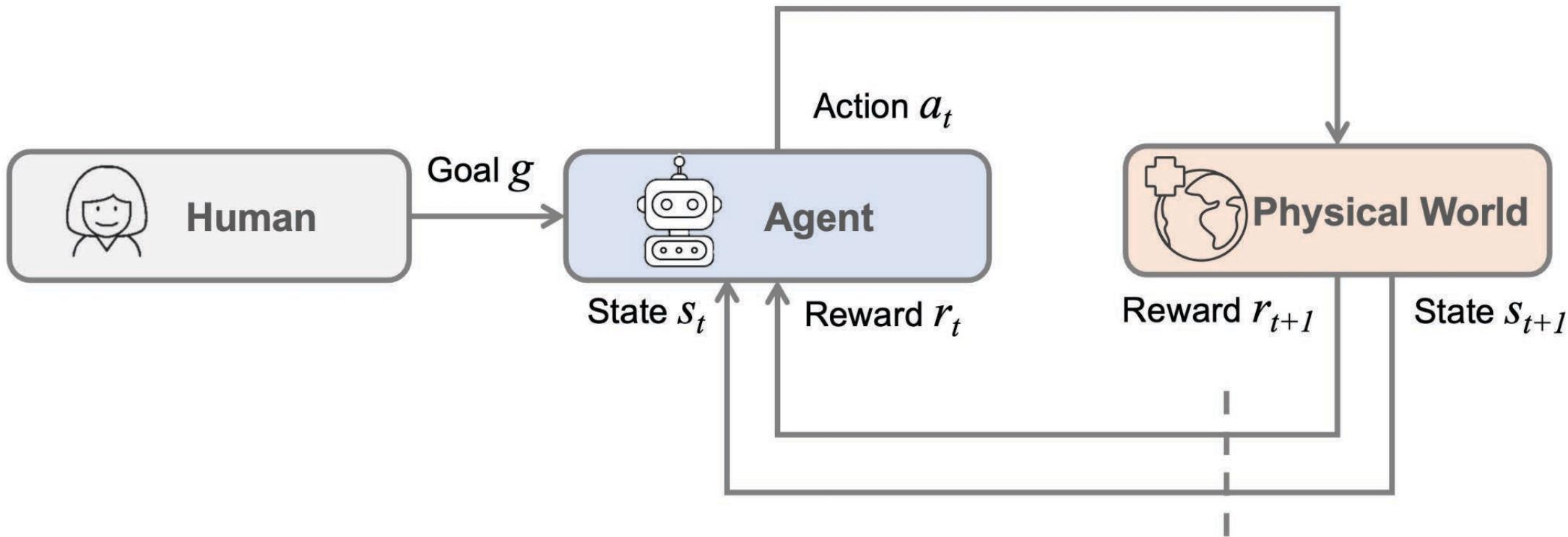


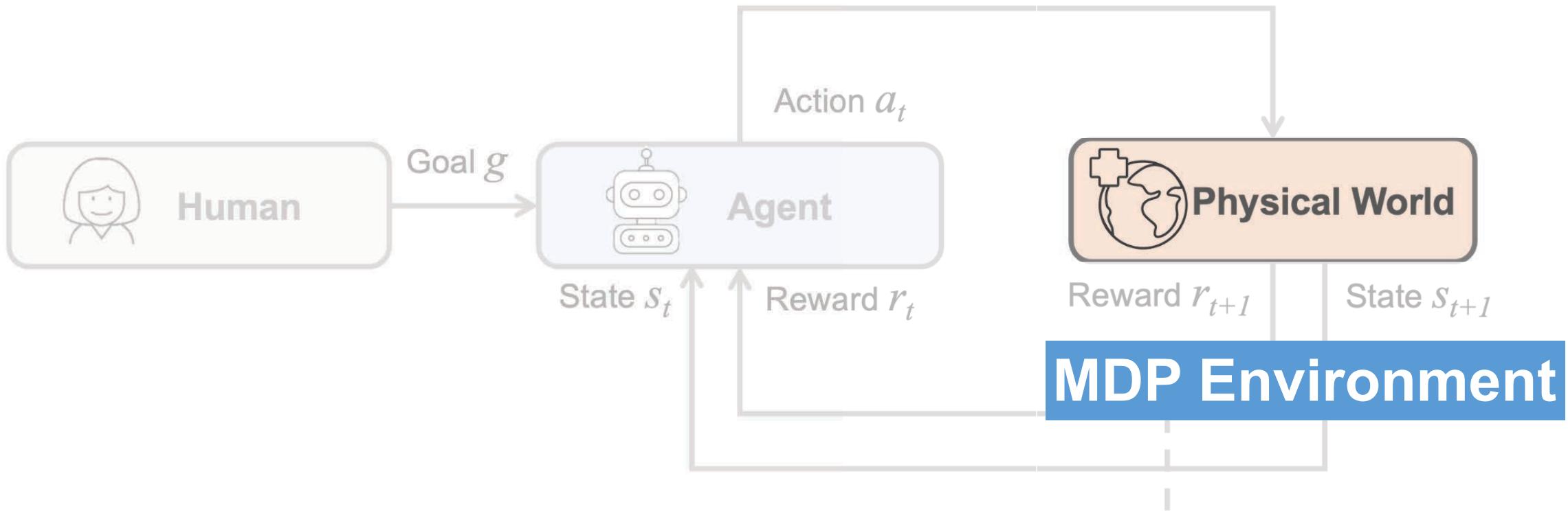












Open-ended Environments

Craft Glass Bridge



Build Oak House



Make Ice Igloo



Combat Zombie



Fish Squid



Farm Sugar Cane



Find Ocean Monument



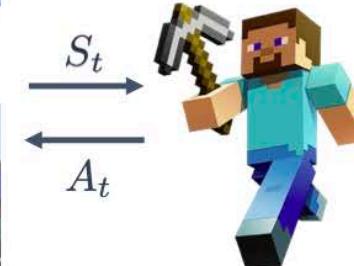
Explore Desert Temple



Treasure Hunt in End City



Generalist Agent



Internet-scale Knowledge Base

	Features	Description	Screenshot
River	Water, Sand, Clay, Sugar Cane, Seagrass, Salmon, Squid, Drowned	Temperature: 0.5. Rainfall: 0.5. A biome that consists of water blocks in an elongated, curving shape similar to a real river. Rivers are a reliable source of clay. They are good for fishing, but drowned can spawn at night.	

MineDojo



Heading OCR

Question: Tell me the heading text of this screenshot of webpage.

Answer: Discover, Appreciate, & Understand the Animal World!

Captioning

Question: What is the meta description of this website?

Answer: The world's largest & most trusted collection of animal facts, pictures and more!

WebQA

Question: What additional platform is mentioned for following the website's content?

Answer: YouTube Channel

VisualWebBench

- Website-wise Task
- Element-wise Task
- Action-wise Task

Element OCR

Question: Tell me the text content in the red bounding box

Answer: We believe that if people know about the world's creatures they will better care for them. That's why we add new animals for you to discover ...

A: Logo and navigation bar (All Animals, Animals, Articles, Reviews, Pets, Places, Quizzes, About, Search)

B: Search bar

C: Animal of the Day (Lion) with text: "Lion Lives in small groups called prides!"

D: Main title "Discover, Appreciate, & Understand the Animal World!" with a red bounding box around it.

E: Latest Product Review (Dog and Cat)

F: Trending on A-Z Animals (Puppy)

G: Search bar at the bottom left

H: Bottom navigation bar (Search, Home, About, Contact, Privacy Policy, Terms of Service, Advertise, Sitemap, RSS Feed, XML Sitemap, Help, Feedback, Log In, Sign Up, Logout, Language: English, Currency: USD, Cart: 0 items, Total: \$0.00)

Action Grounding

Question: I have labeled bright IDs for some HTML elements in this website screenshot. Tell me which one I should click to complete the instruction: learn about the animal of the day

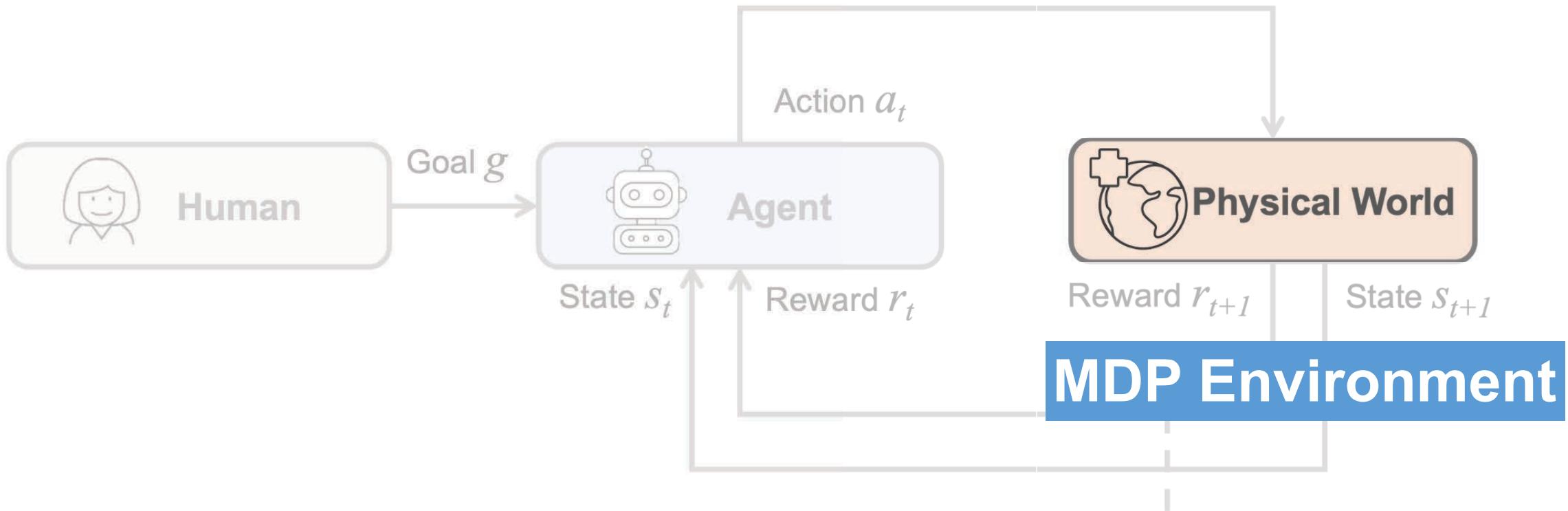
Answer: C

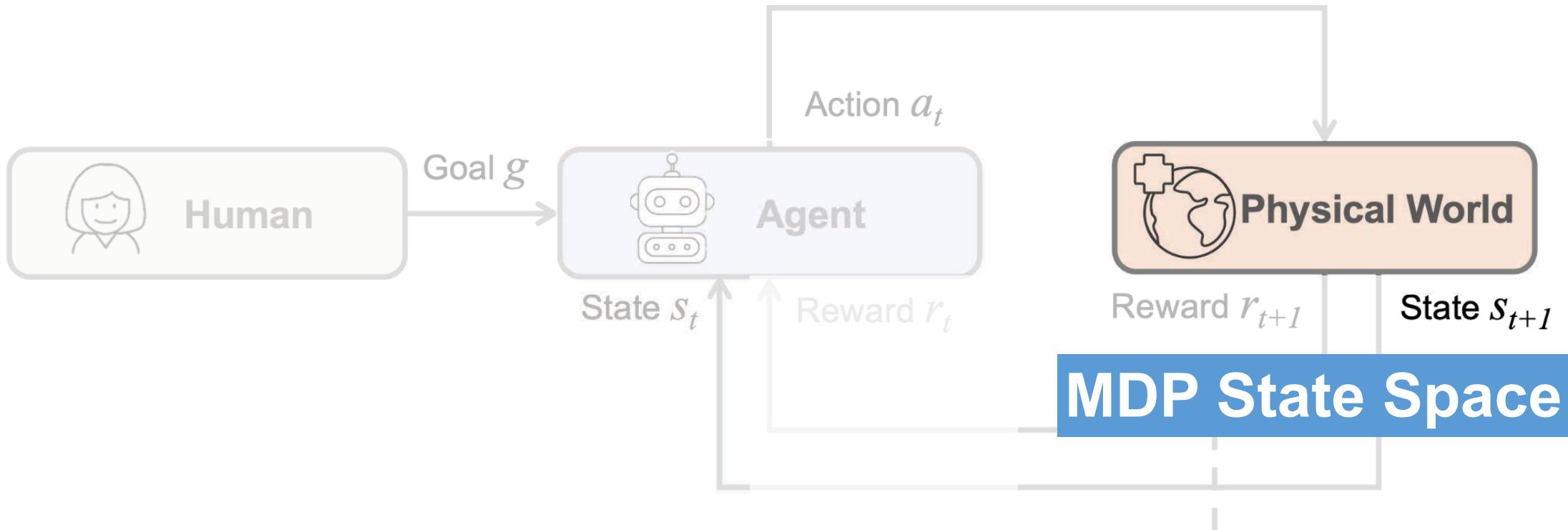
Action Prediction

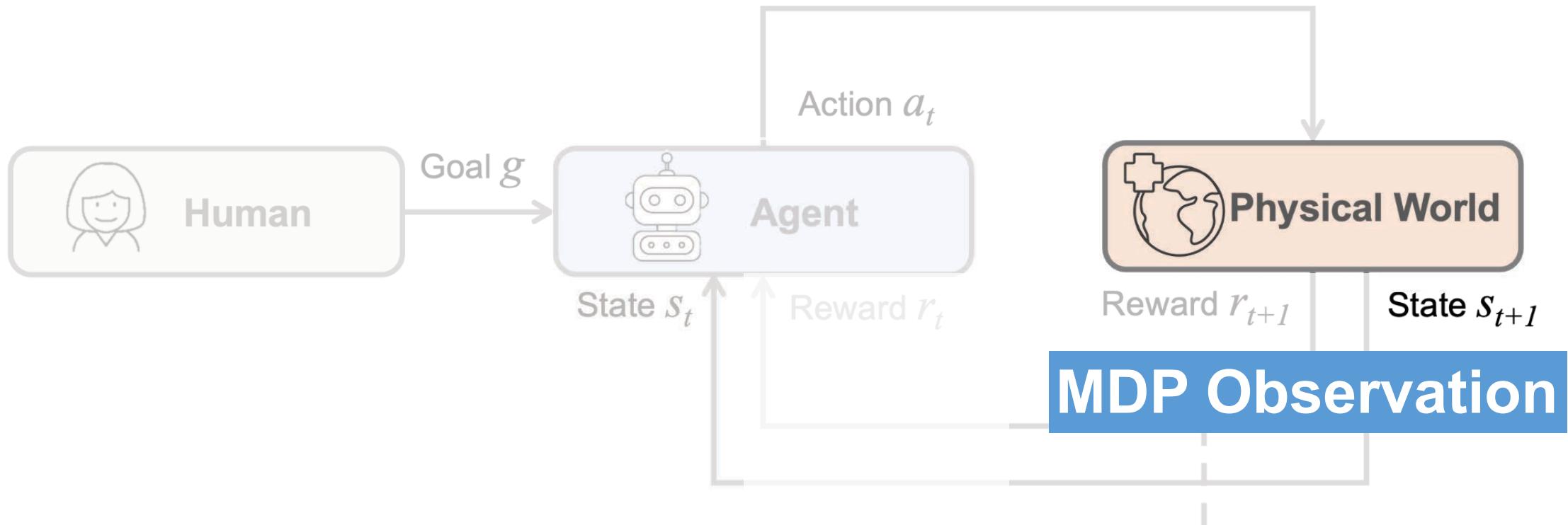
Question: After clicking the element in the bounding box, which one is the best description of the new webpage?

(A) Animal news, facts, ...
 (B) All animals A-Z List
 (C) The 7 best pet ...
 (D) Search any animals!

Answer: C









Environment : Observation (Rendered 2D Images)

Enabled by large dataset of realistic interactive **scenes** and objects



50 Scenes





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

50 Scenes

Enabled by large dataset of realistic interactive scenes and objects



10000 Objects

Semantic

Properties: cookable, sliceable, freezable, burnable, deformable

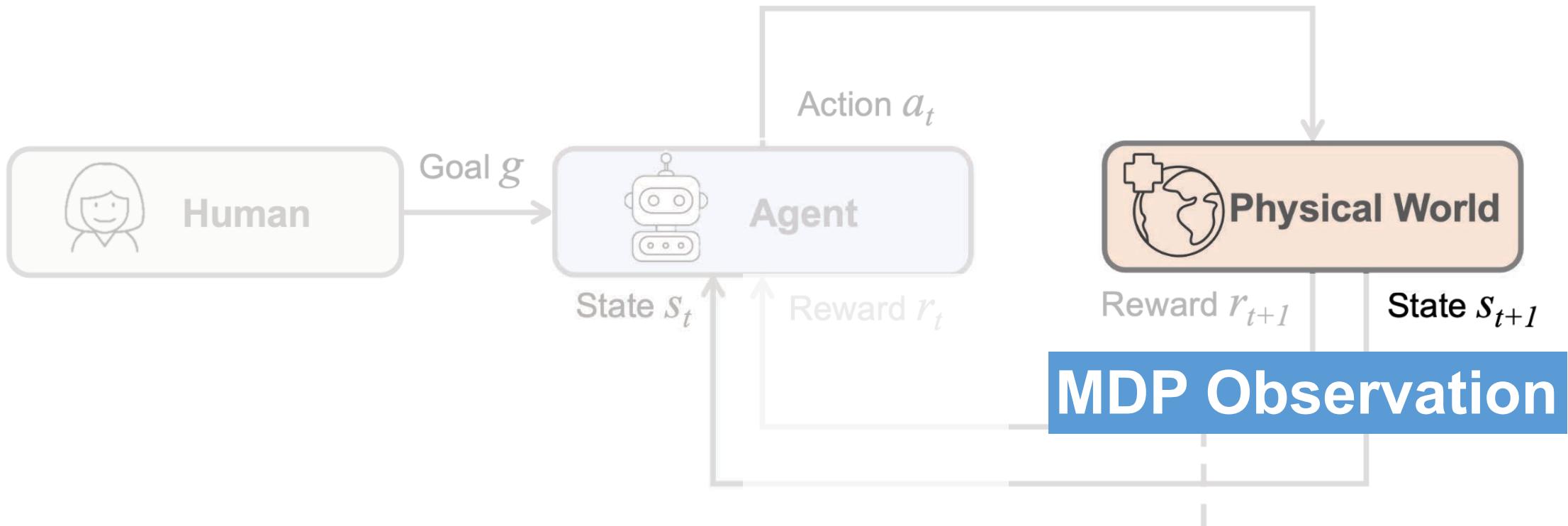
... Cooking temperature: 58°C

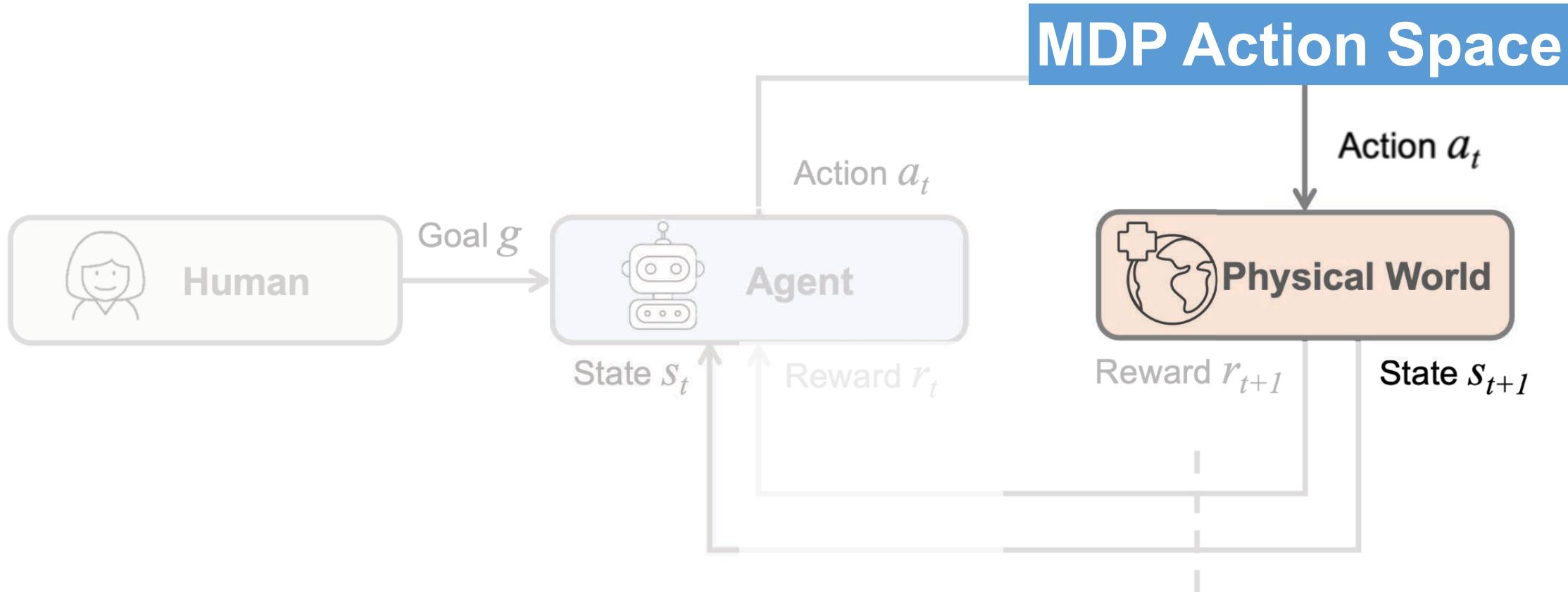


Physical

Articulation annotation (joint type, origin, axis, limit)
Mass, friction, CoM, ...
Canonical size and orientation



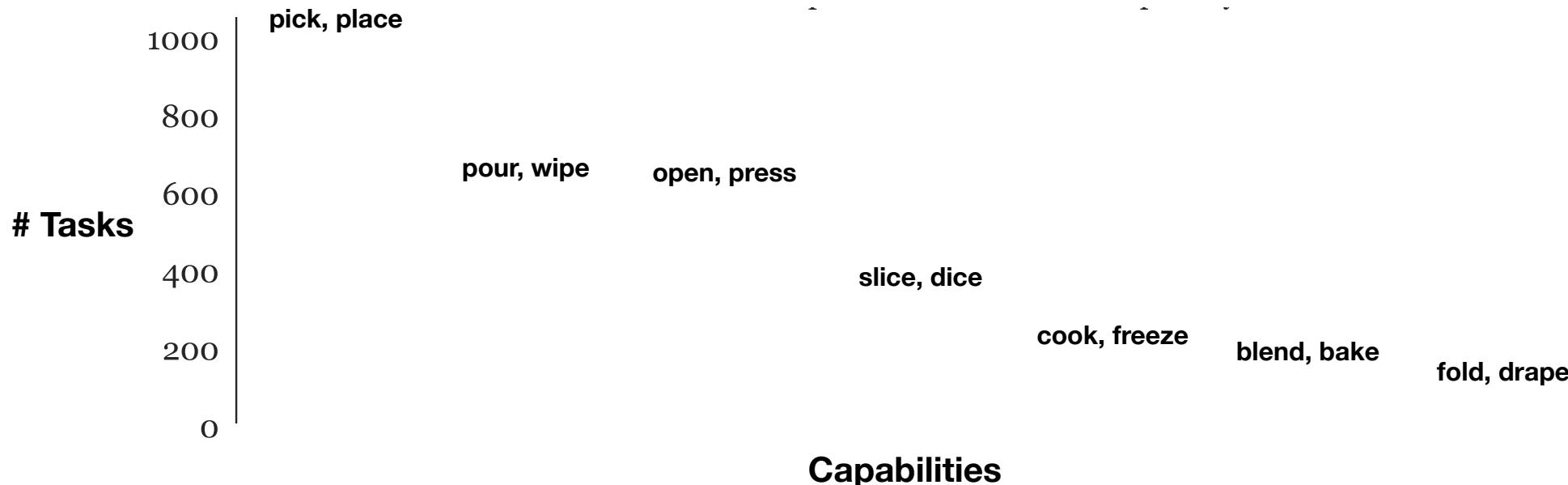


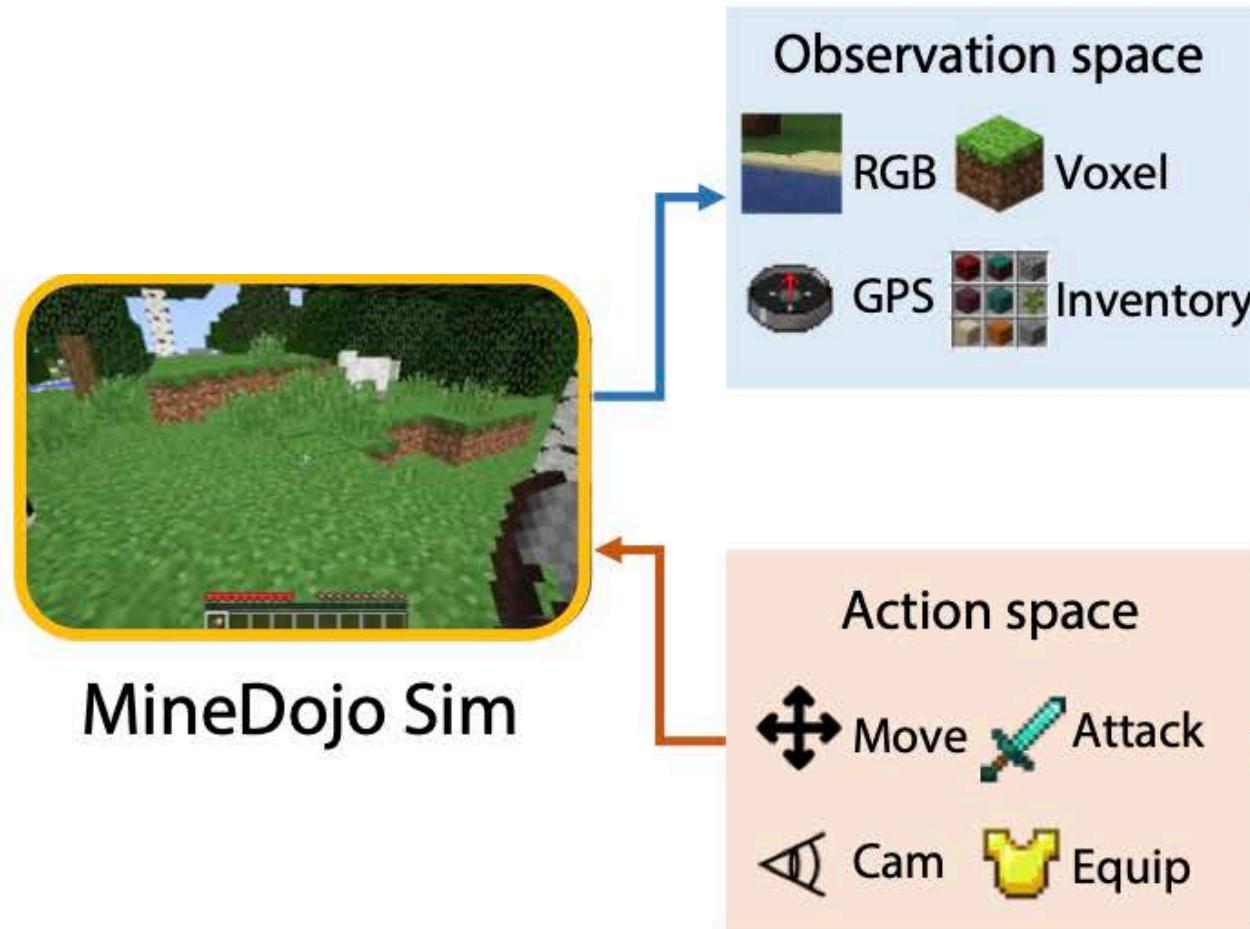




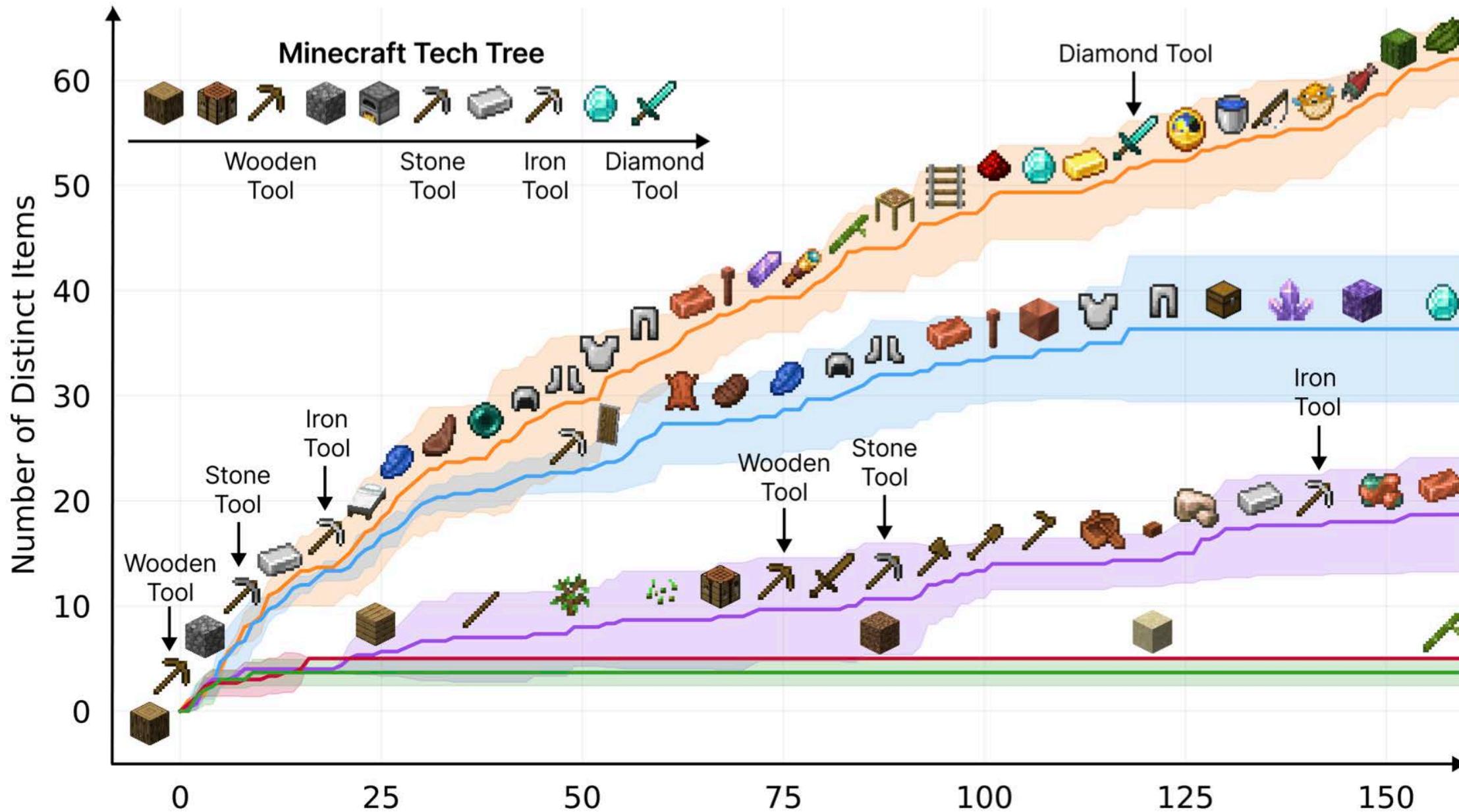
Action : can robots learn to **solve** these tasks?

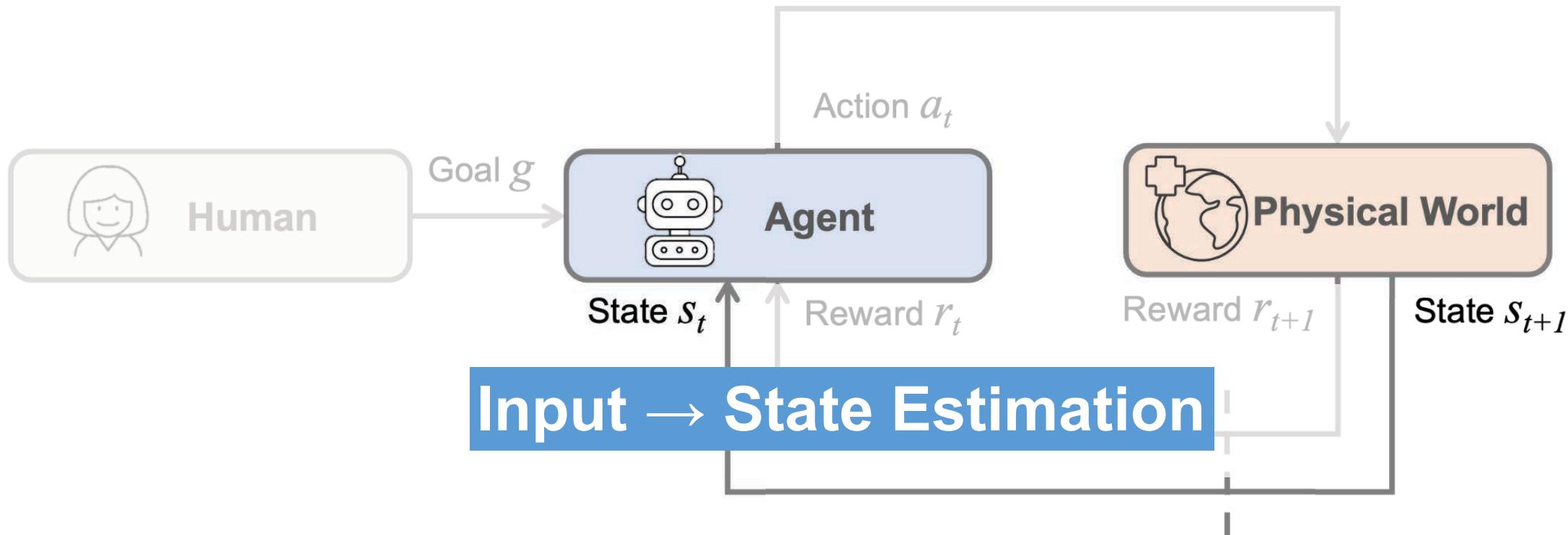
→ What **capabilities** are needed?





MDP Action Space: Skills





Perception / State Estimation

$$o \rightarrow s$$

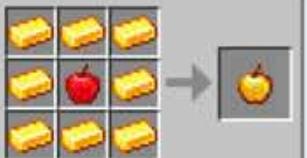
Observation (2D rendered scenes)

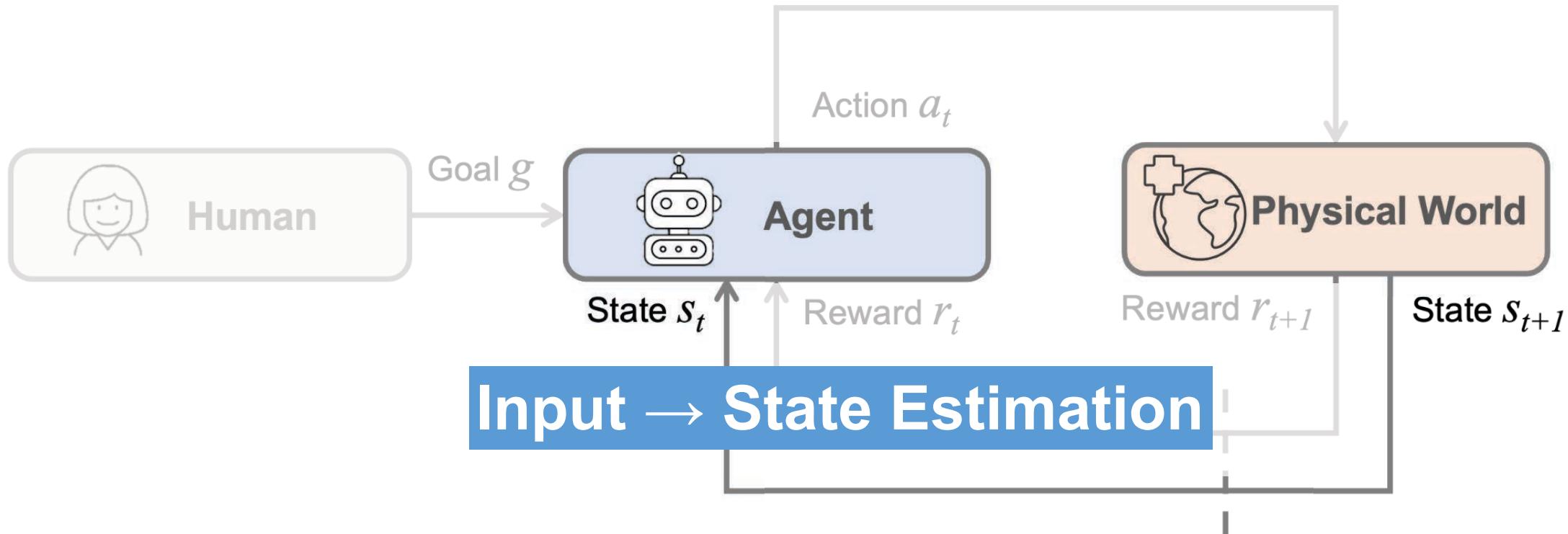


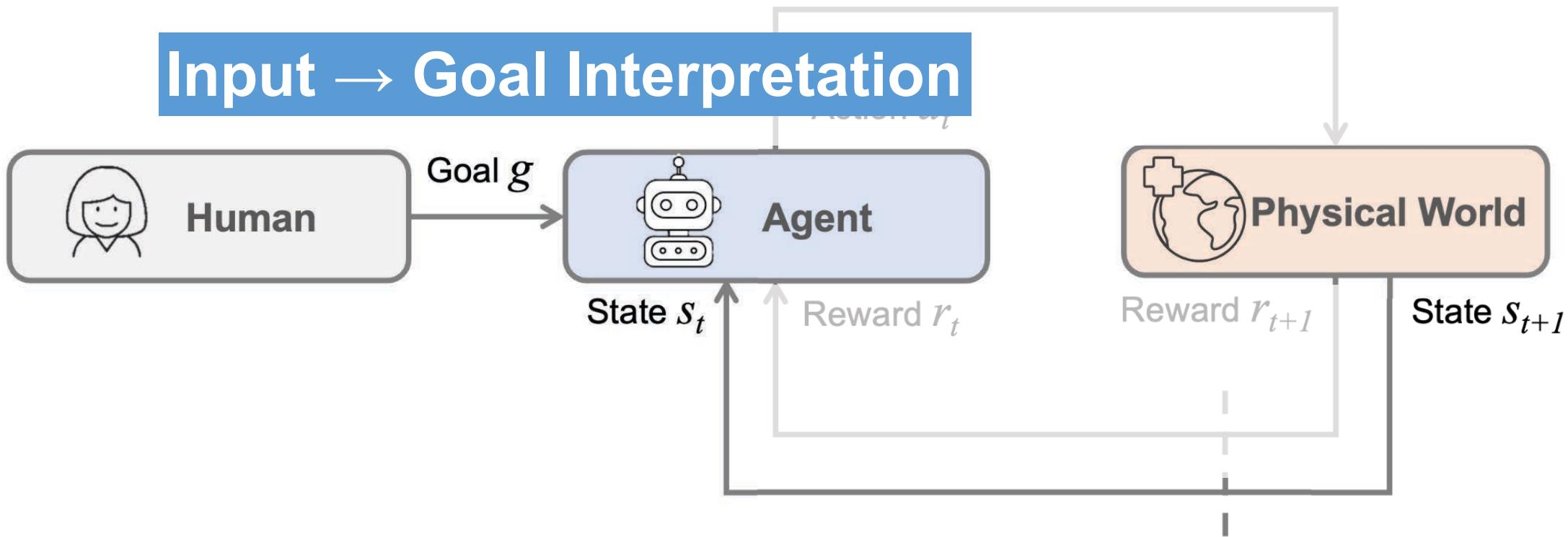
State (3D assets)

Passive mobs

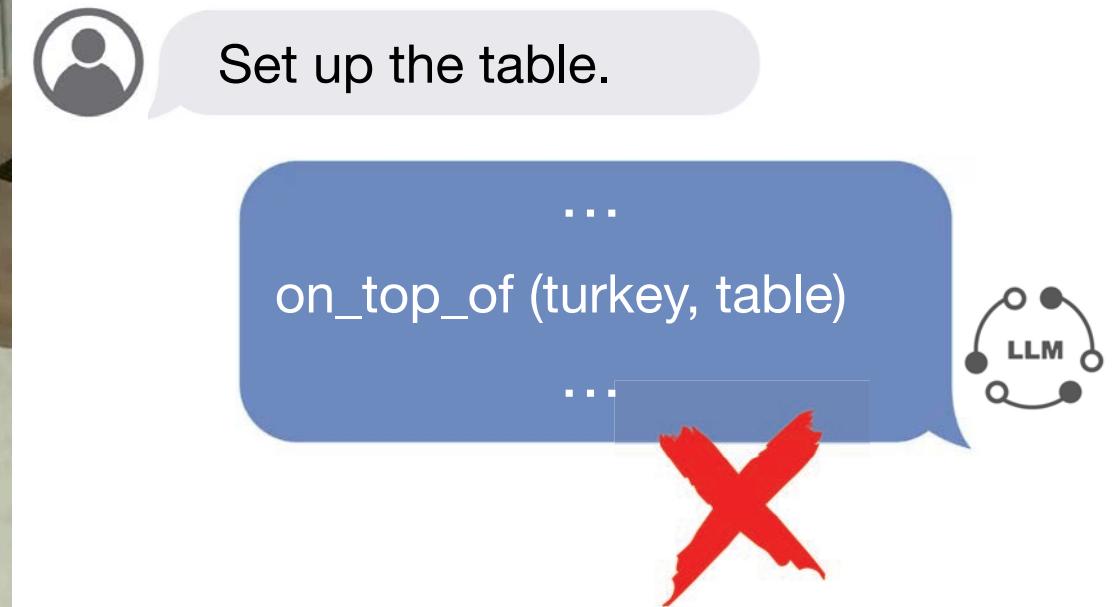


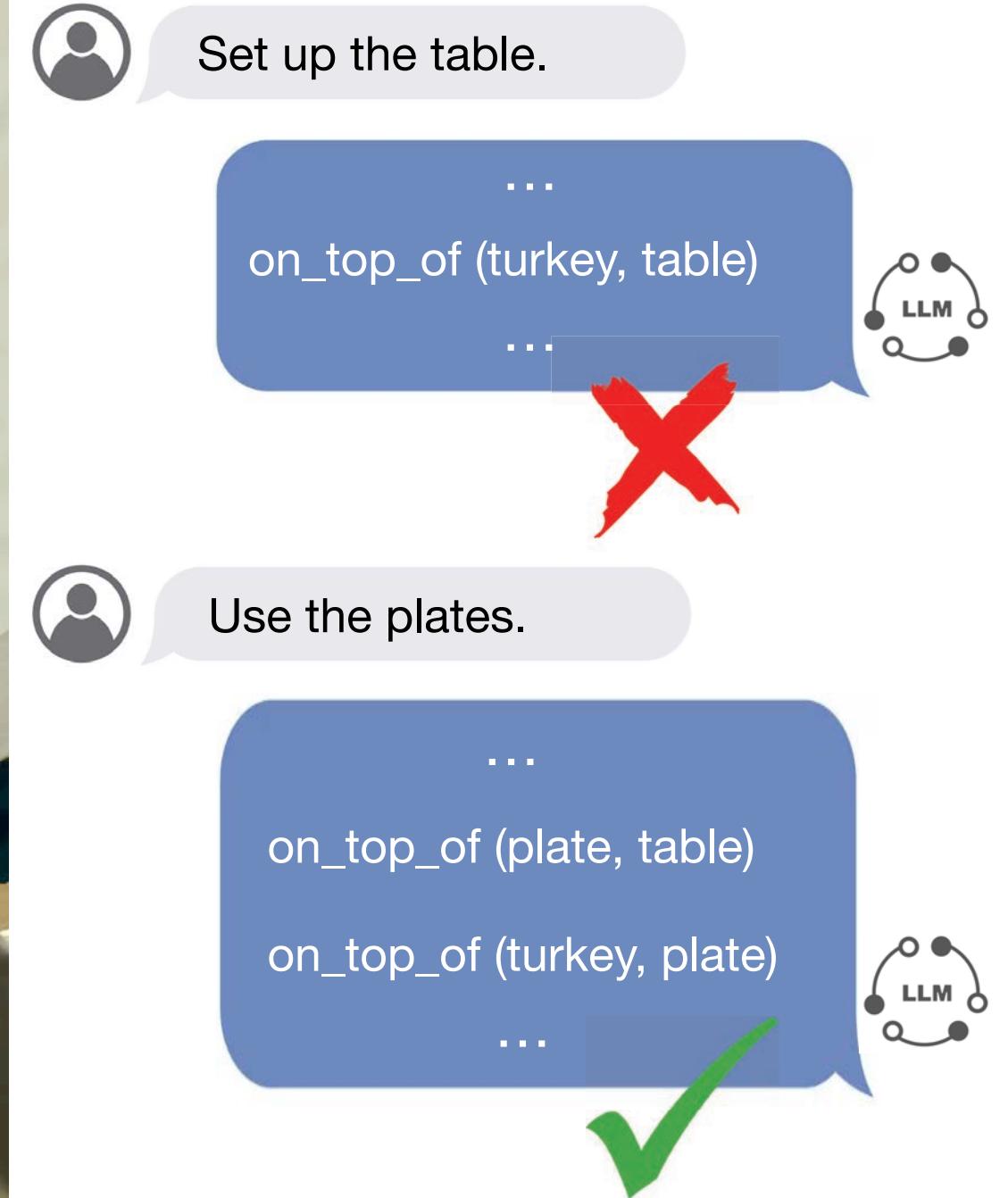
Name	Ingredients	Crafting recipe
Cake	Milk Bucket + Sugar + Egg + Wheat	 
Golden Apple	Gold Ingot + Apple	 





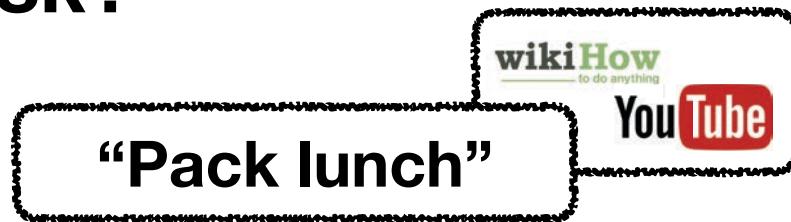
Goal Interpretation g







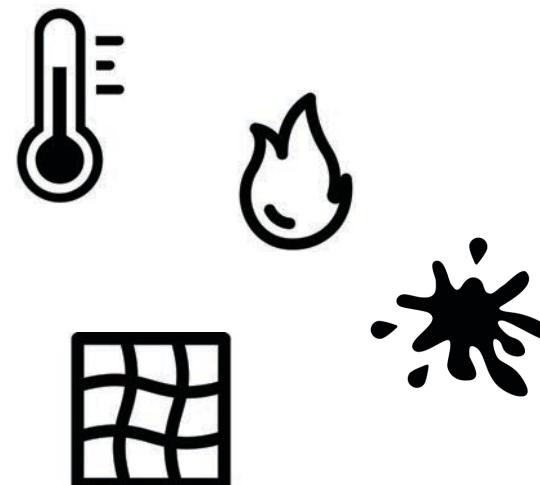
Goal : defines a task?



What objects?



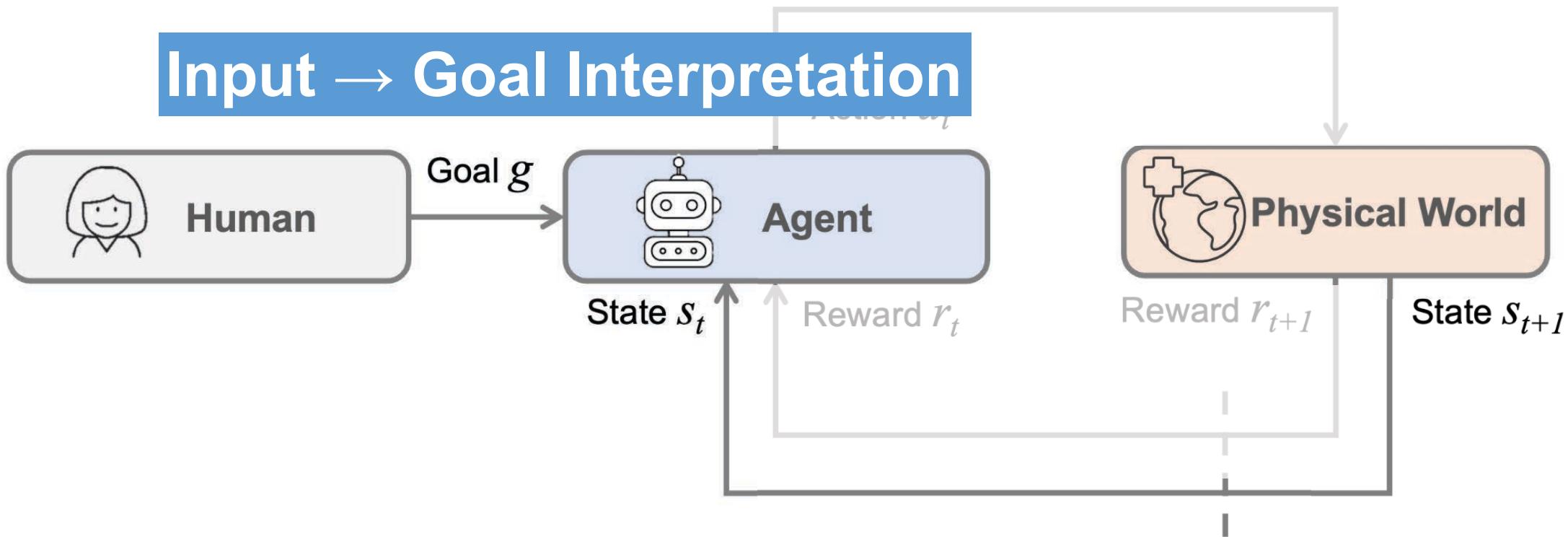
What properties?

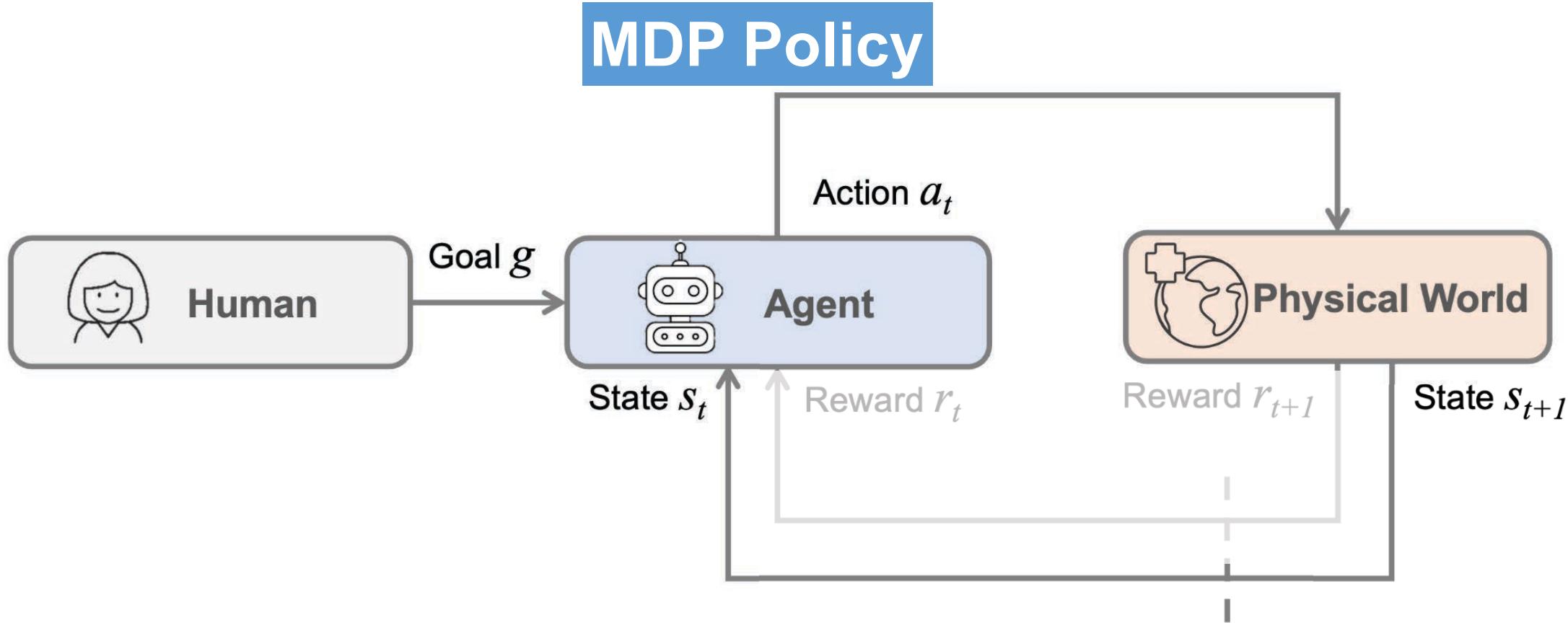


Start & Goal?

- apple in fridge
- burger in fridge
- water bottle in fridge
- paper bag on counter







Policy

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

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

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>.

Input: Preserving food

Environment

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

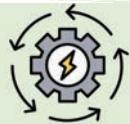


Goal

... 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

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>.

Input: Preserving food

Environment

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

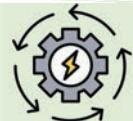


Goal

... 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

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>.

Input: Preserving food

Environment

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



Goal

... cooked (strawberry) ...



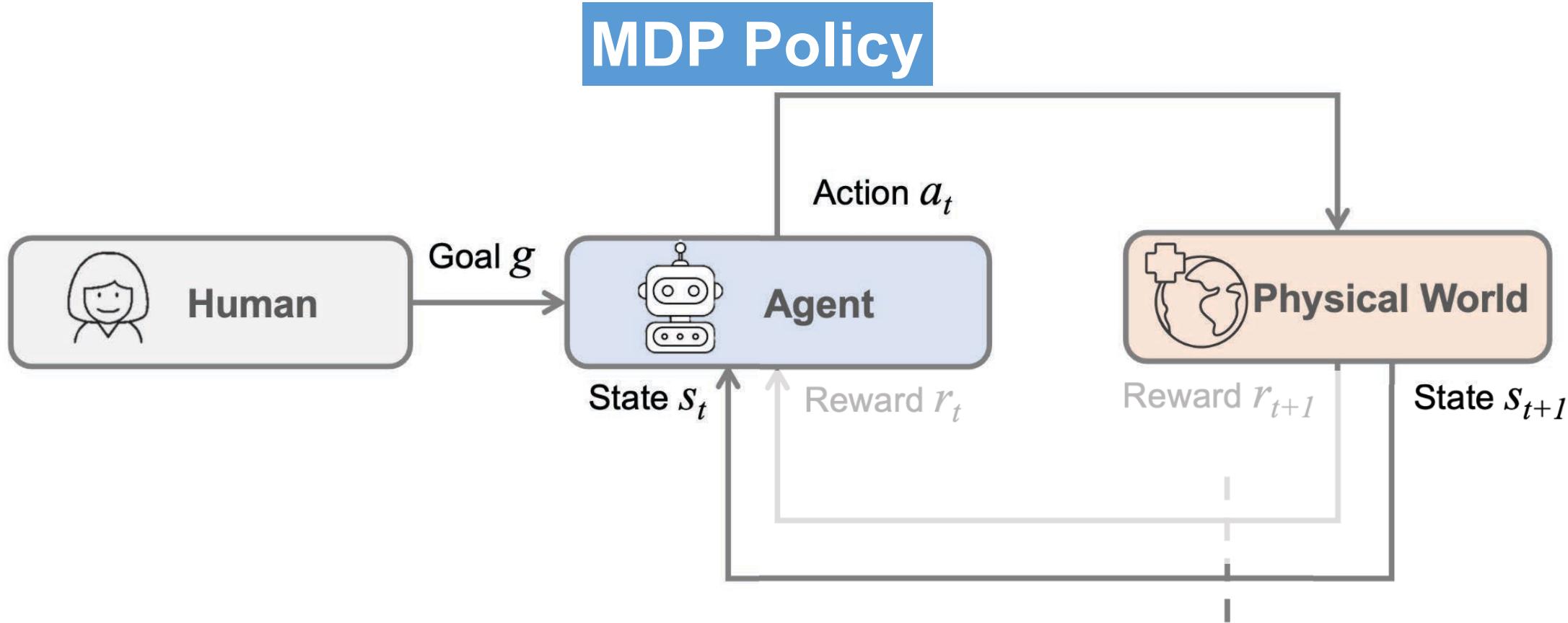
Operator

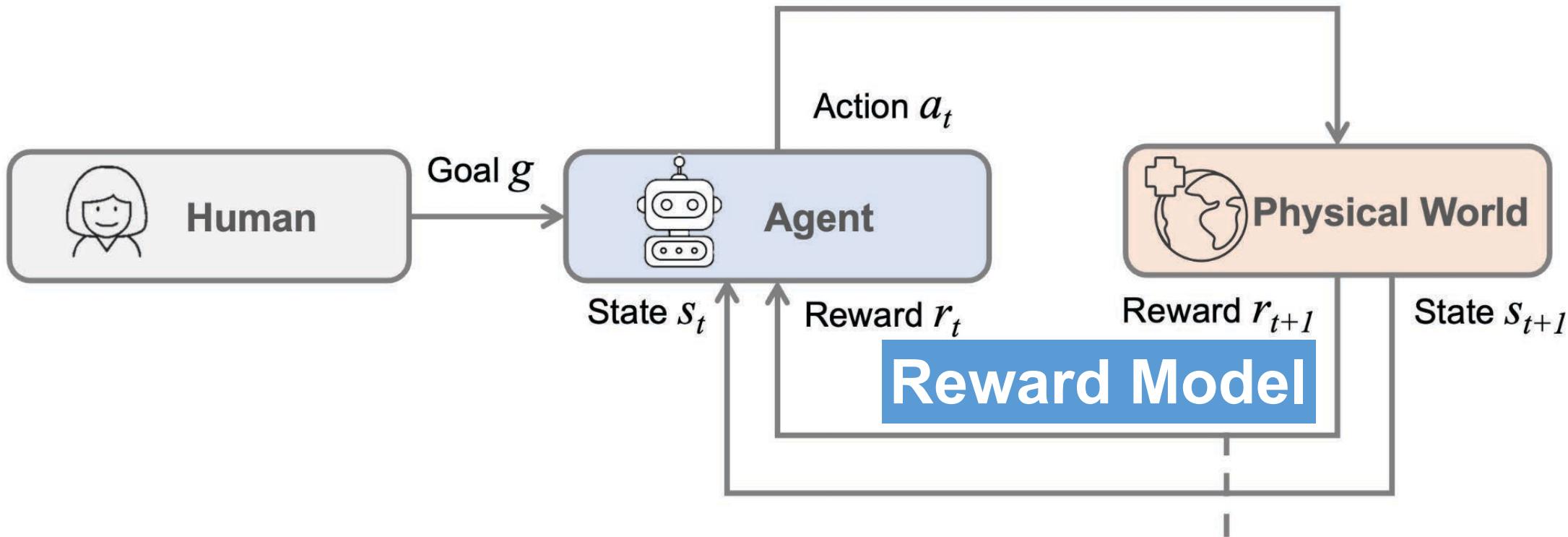
BEHAVIOR**Action Trajectory**

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

LLM Output

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>.



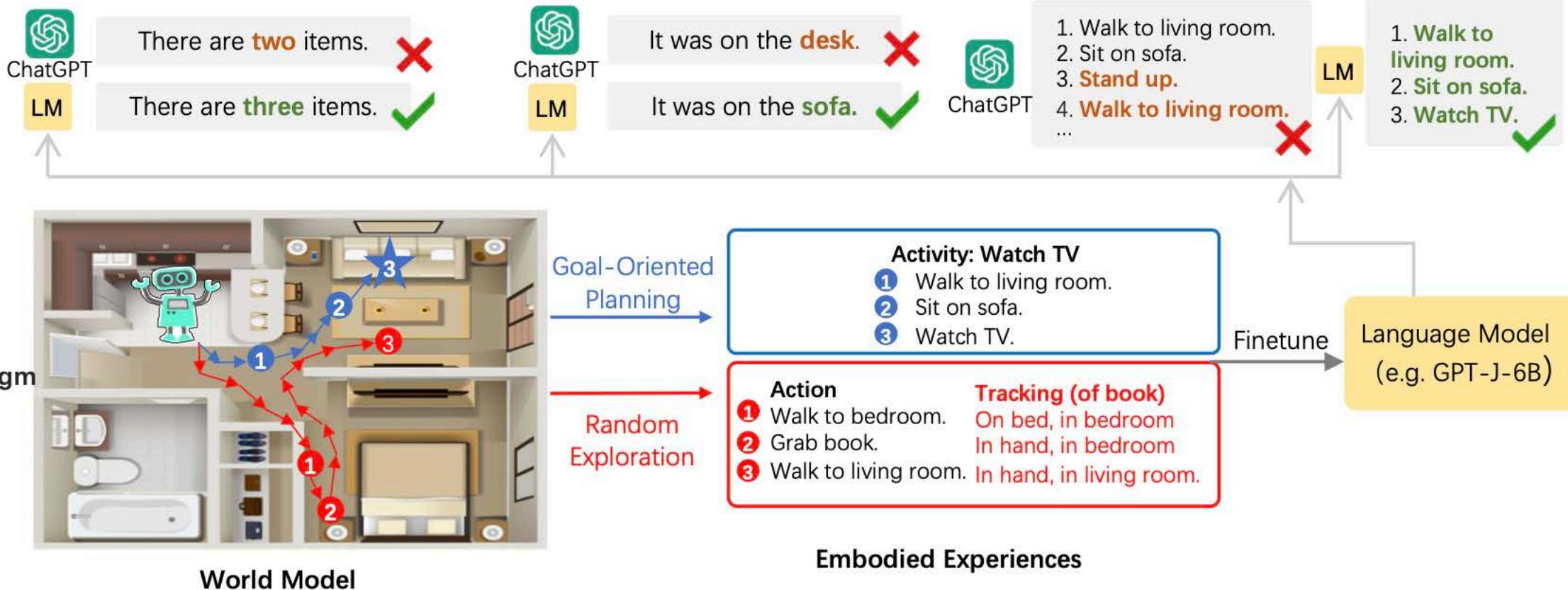


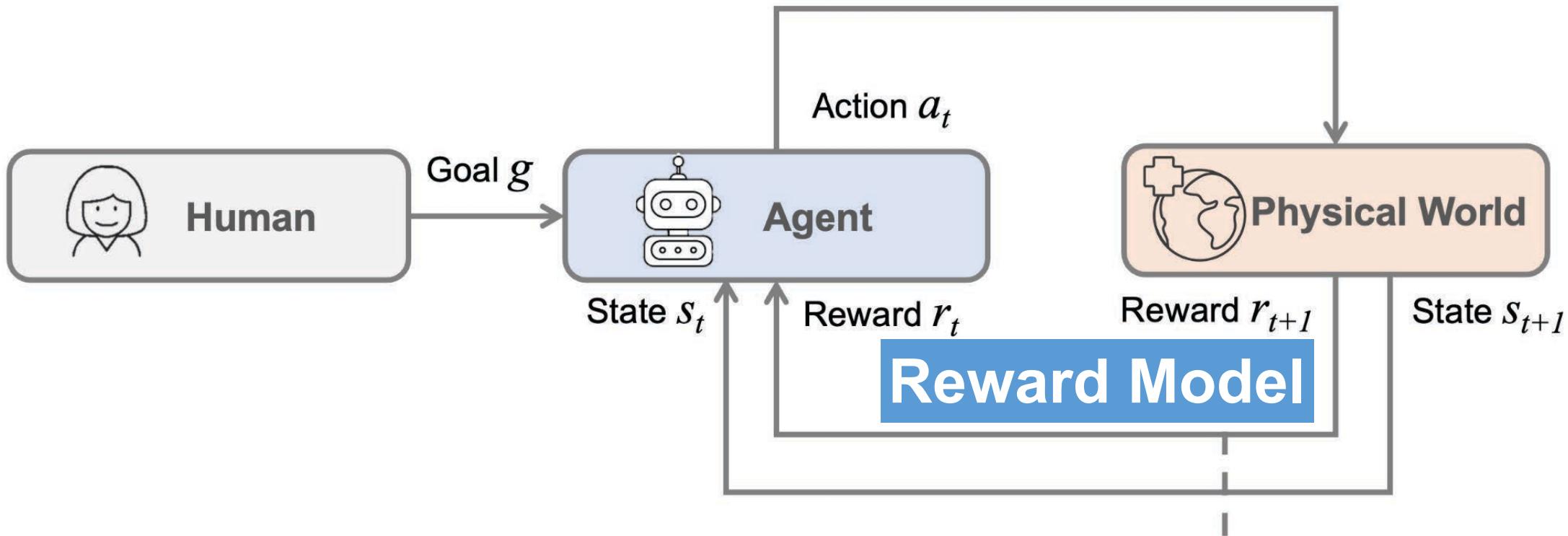
Reward Model

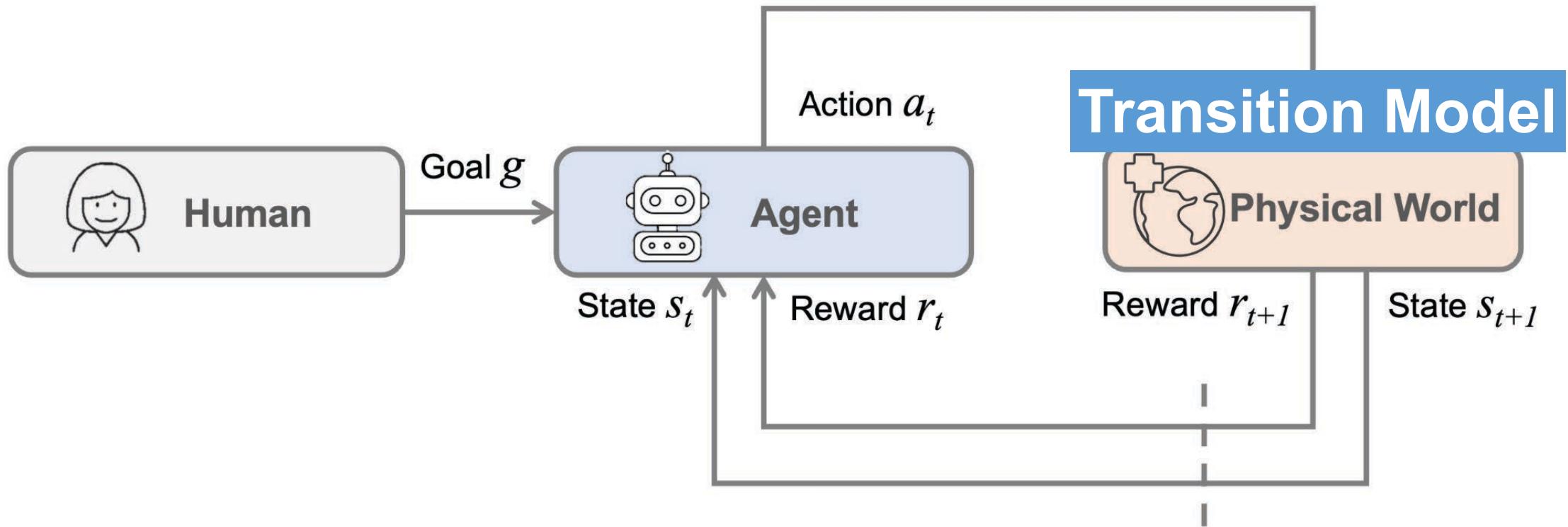
$$o, a \rightarrow r$$

Reward Model

Answer:







Transition Model

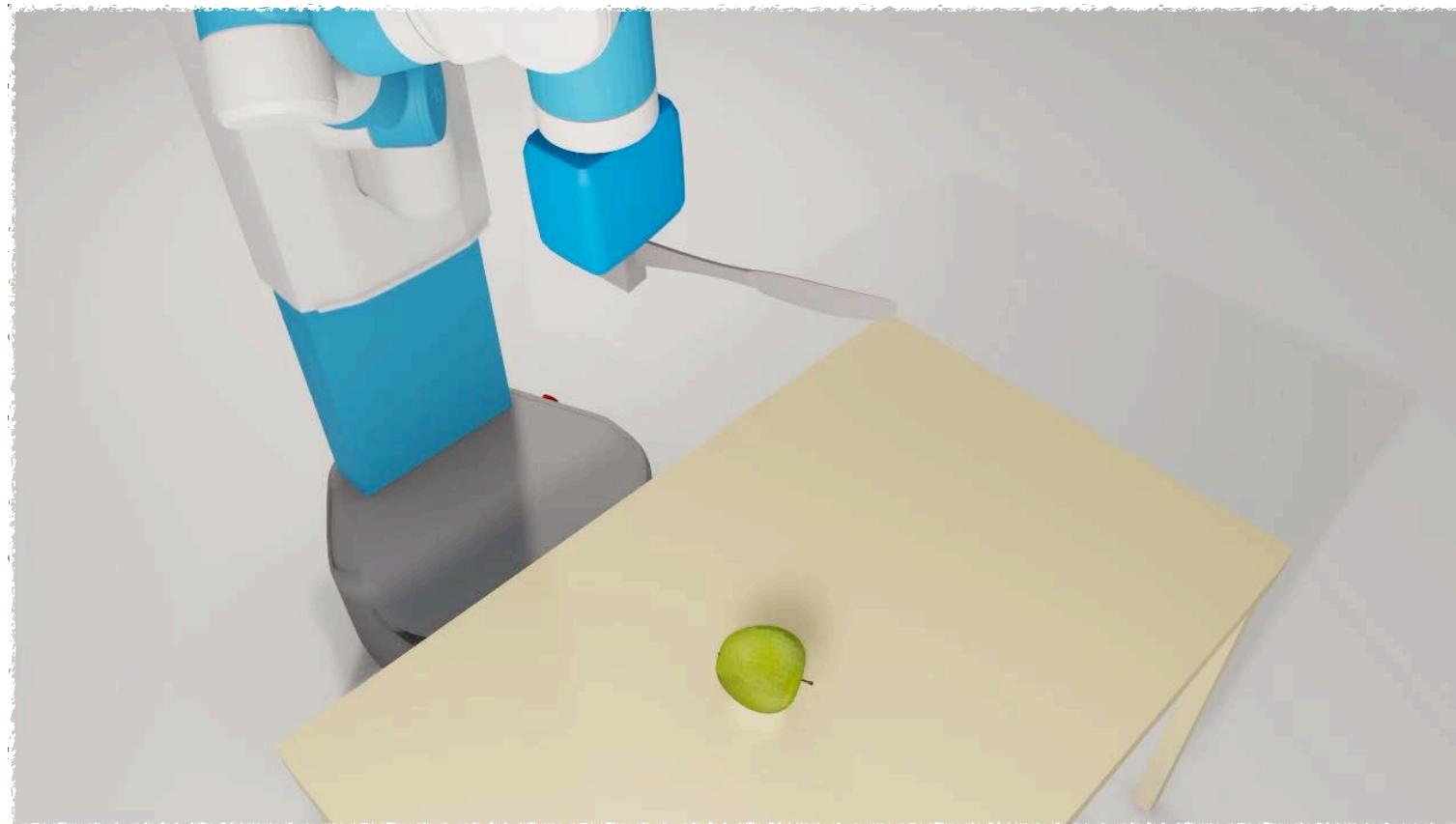
$$o_t, a \rightarrow o_{t+1}$$

"World Modeling"



Physics Transition Examples

```
class SlicingRule:
```

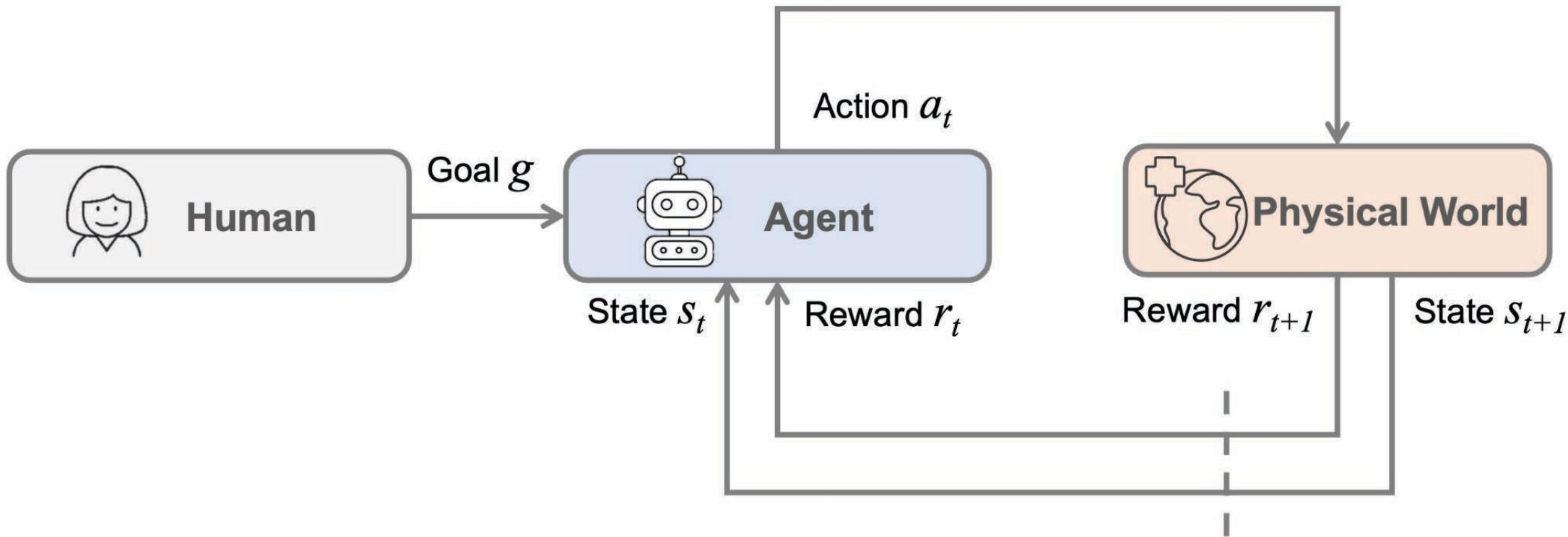


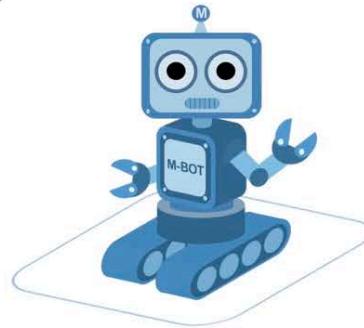


```
class TransitionRule:  
    def condition(self, *args) -> bool  
    def transition(self, *args)
```

Two red arrows point from annotations to specific parts of the code. One arrow points from the text "Determines whether a transition should occur" to the `condition` method. Another arrow points from the text "What should happen when a transition is triggered" to the `transition` method.

Allows us to capture arbitrarily complex **physical phenomena!**





Different Instantiations of MDP



Perfect actuations

Noisy actuations

Perfect odometry

Noisy odometry

RGB only perception

Multiple perceptual modalities

-
-
-

-
-
-

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 Li
5. Remaining Challenges	15min	Yunzhu Li
QA	30min	