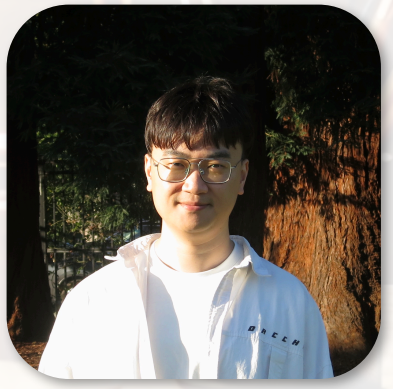
Foundation Models Meet Embodied Agents

Manling Li
Northwestern

Yunzhu Li
Columbia

Jiayuan Mao
MIT



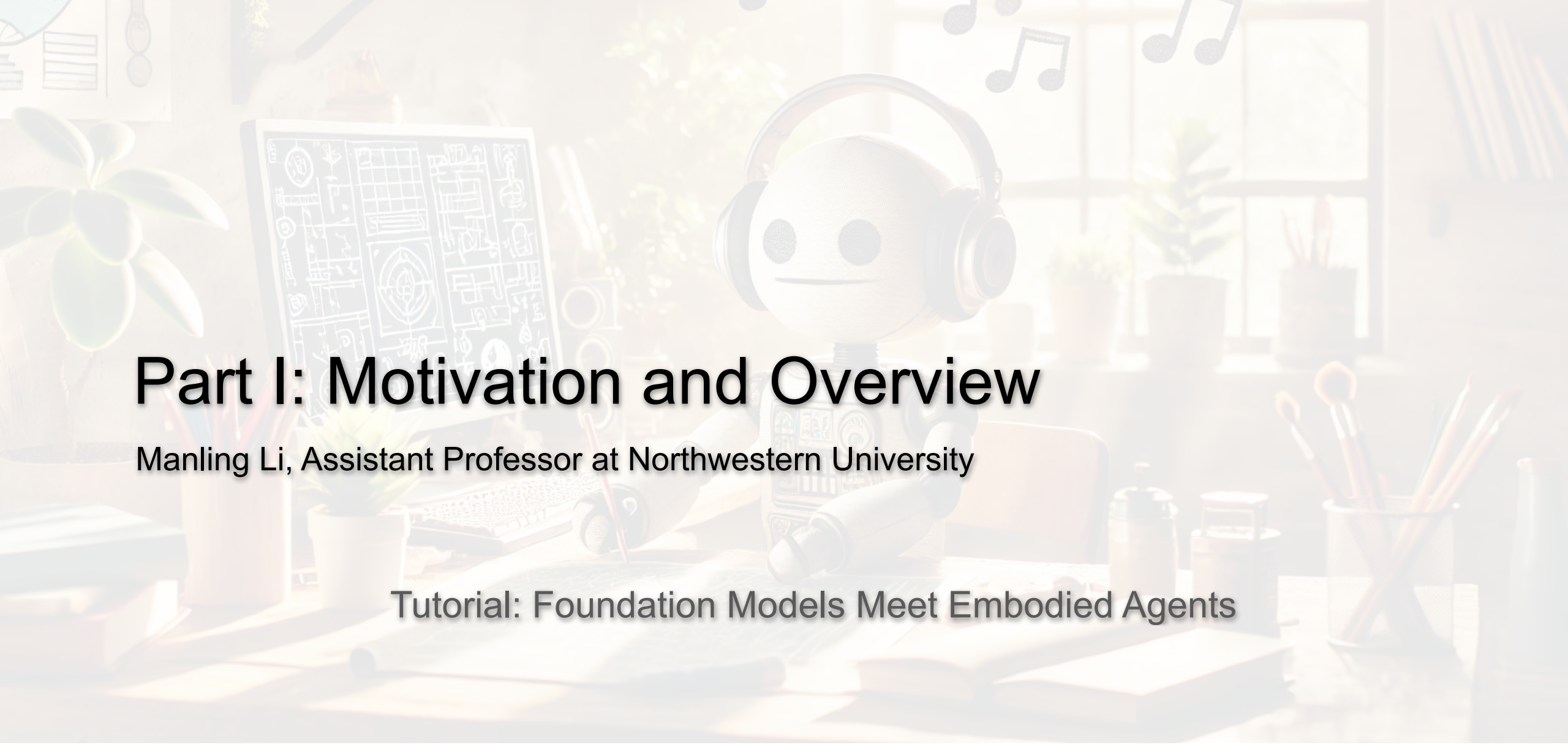
Wenlong Huang
Stanford

Northwestern University

COLUMBIA

MIT

Stanford University



Part I: Motivation and Overview

Manling Li, Assistant Professor at Northwestern University

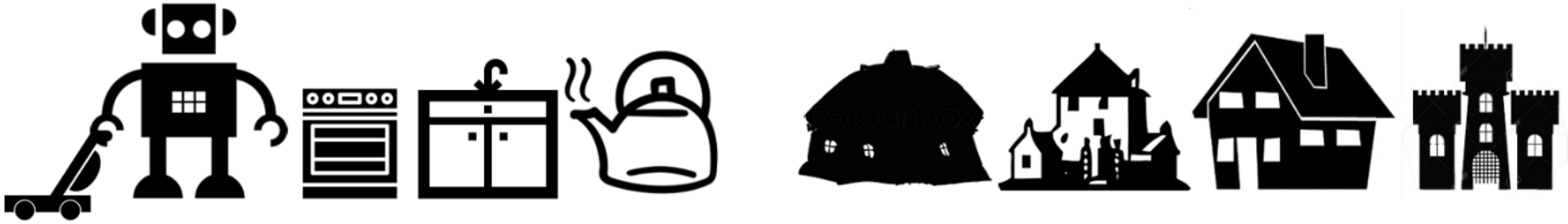
Tutorial: Foundation Models Meet Embodied Agents

Northwestern University

Stanford University

What is a generalist agent?

What is a generalist agent?



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

Figure credits: Jiayuan Mao


BEHAVIOR-1K
simulating and benchmarking robot tasks that matter to humans
https://behavior.stanford.edu/

What is a generalist agent?

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

Figure credits: Jiayuan Mao

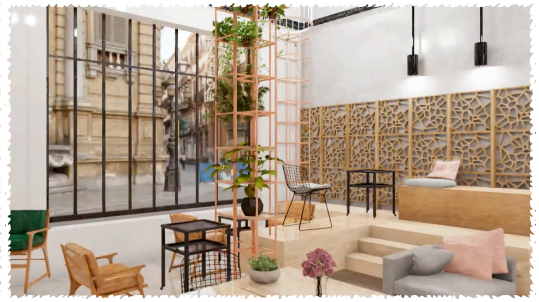
BEHAVIOR-1K

Environment

Observation



State: 3D assets & states

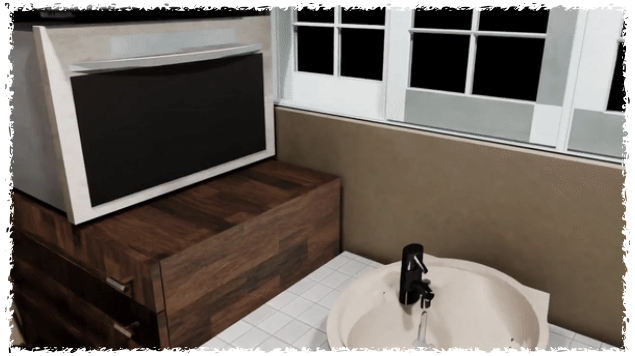







Transition Model





https://behavior.stanford.edu/

What is a generalist agent?

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

Figure credits: Jiayuan Mao

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

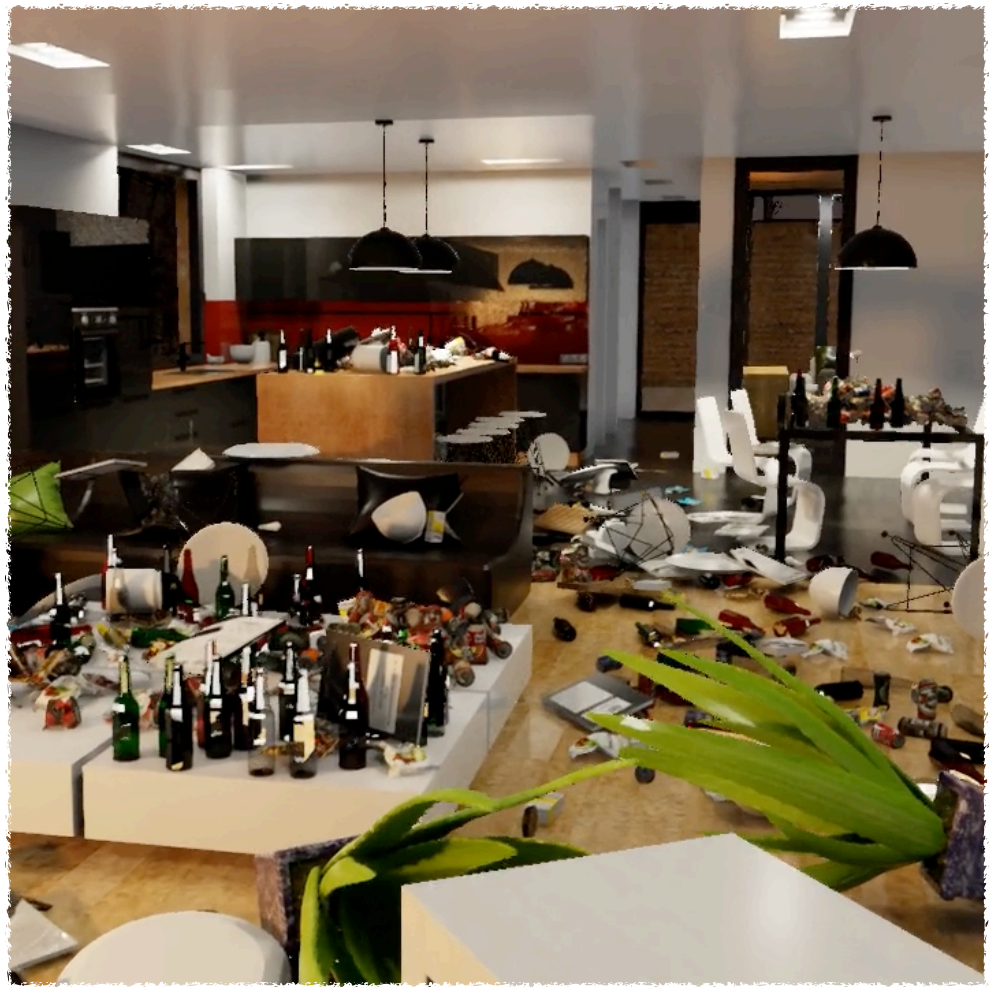
What would you like a robot to help you with?

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

Typical Embodied Agent Tasks

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. …

https://arxiv.org/pdf/2311.12871

To achieve these tasks in these environments

“embodied decision making”

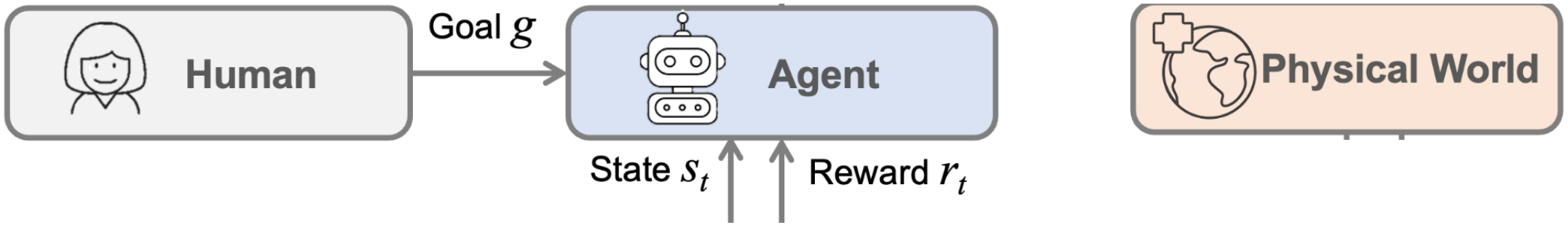
What is “embodied decision making” ?

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

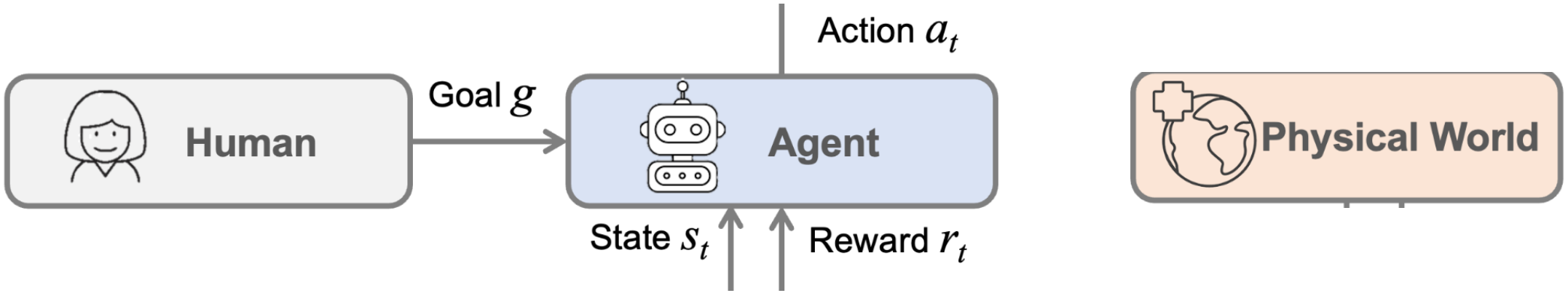
Let us go back to MDPs (Markov Decision Processes)

Let us go back to MDPs (Markov Decision Processes)

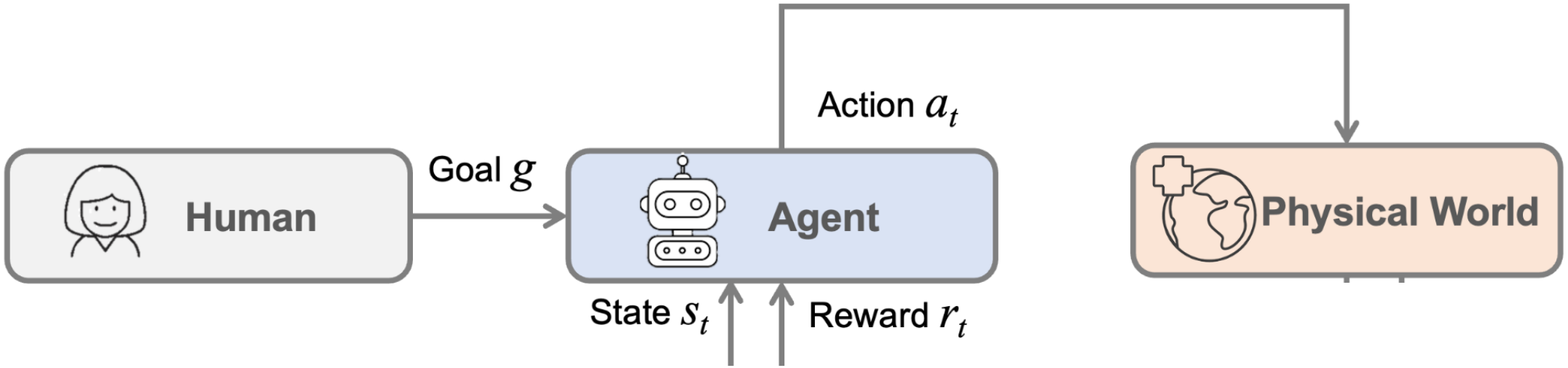
Let us go back to MDPs (Markov Decision Processes)



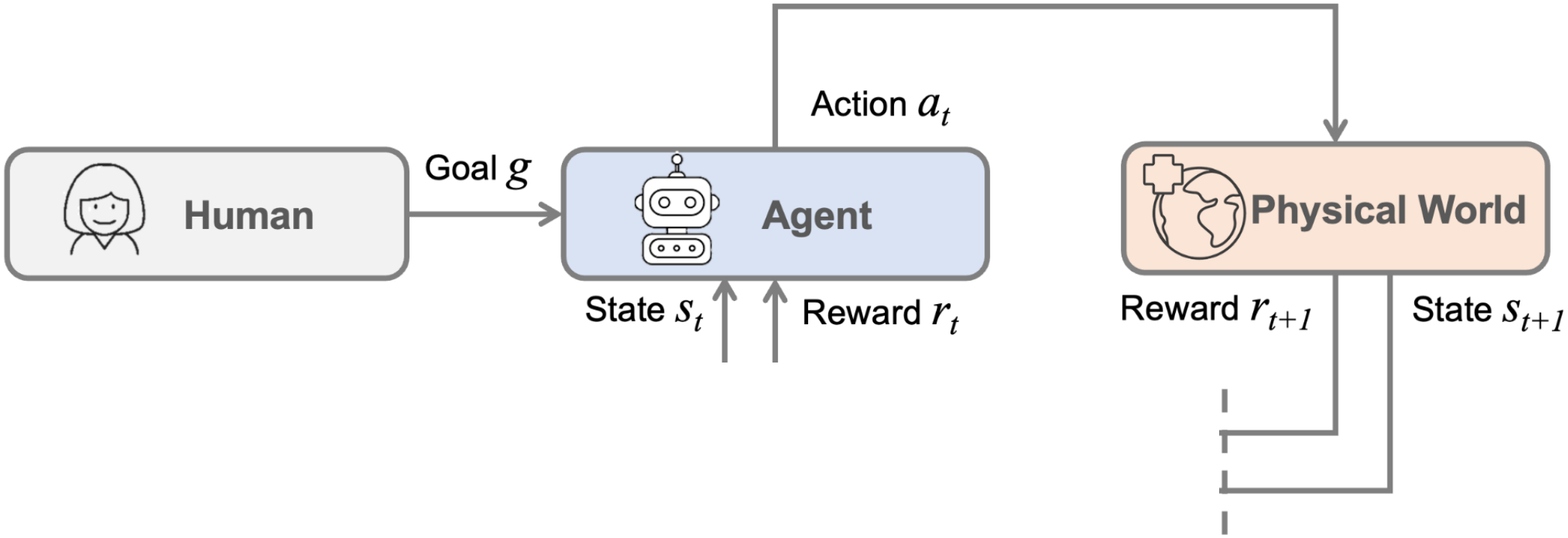
Let us go back to MDPs (Markov Decision Processes)



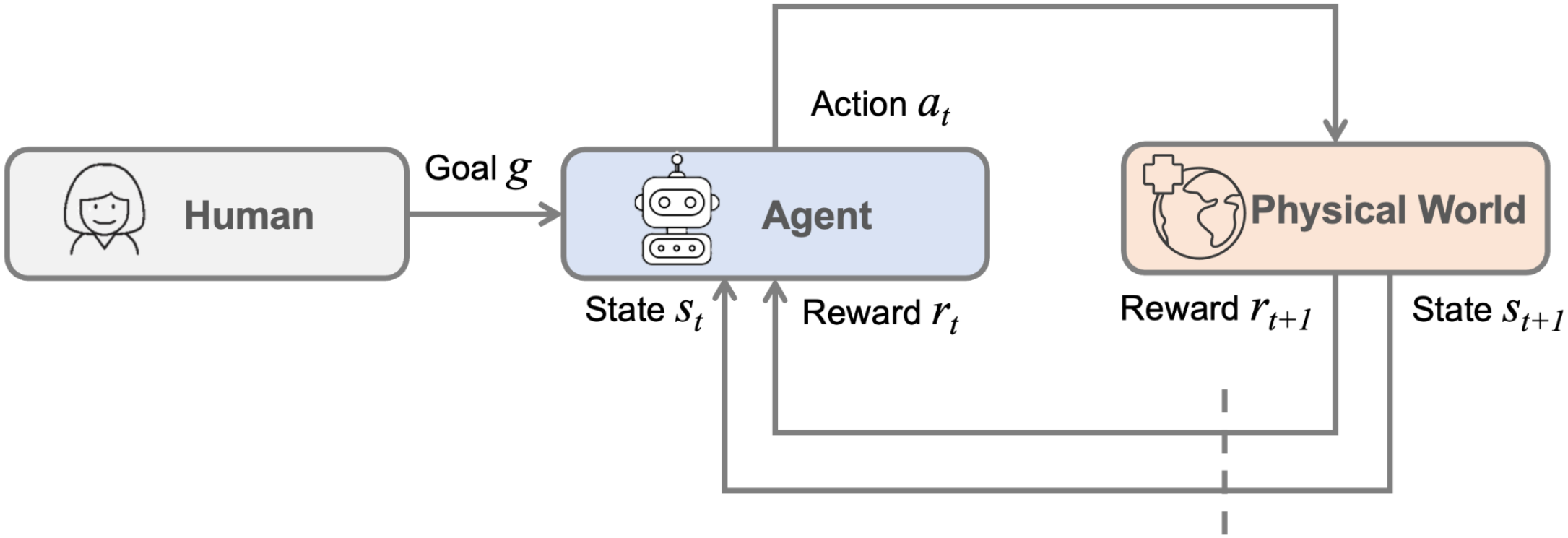
Let us go back to MDPs (Markov Decision Processes)



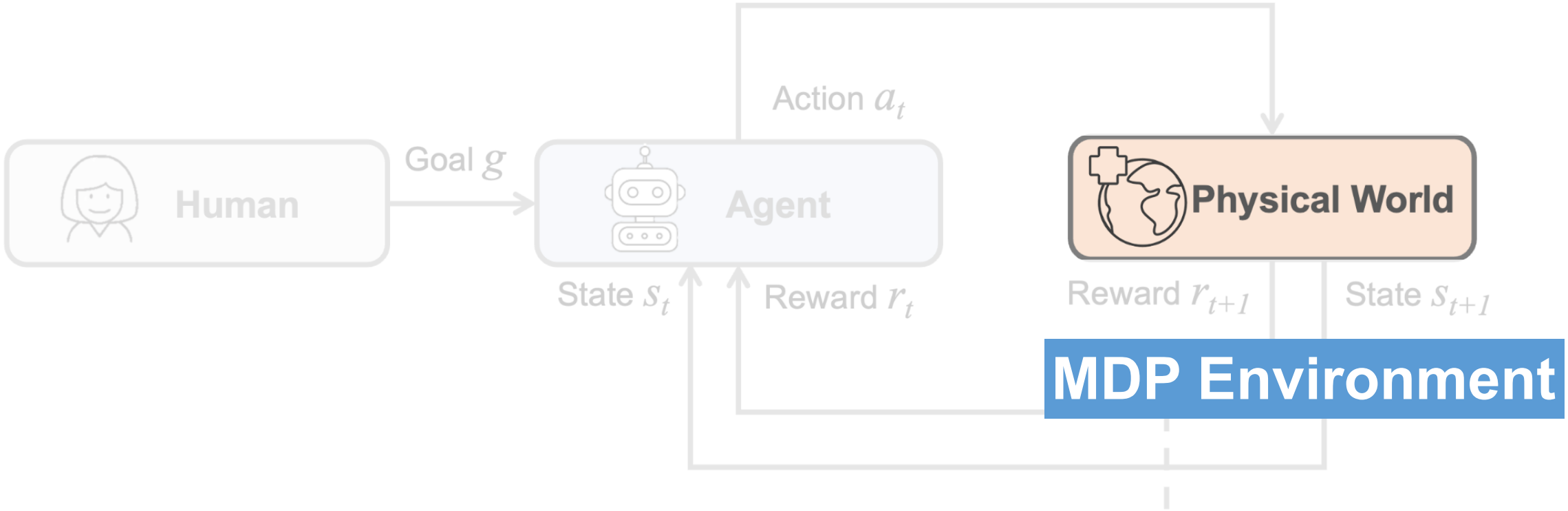
Let us go back to MDPs (Markov Decision Processes)



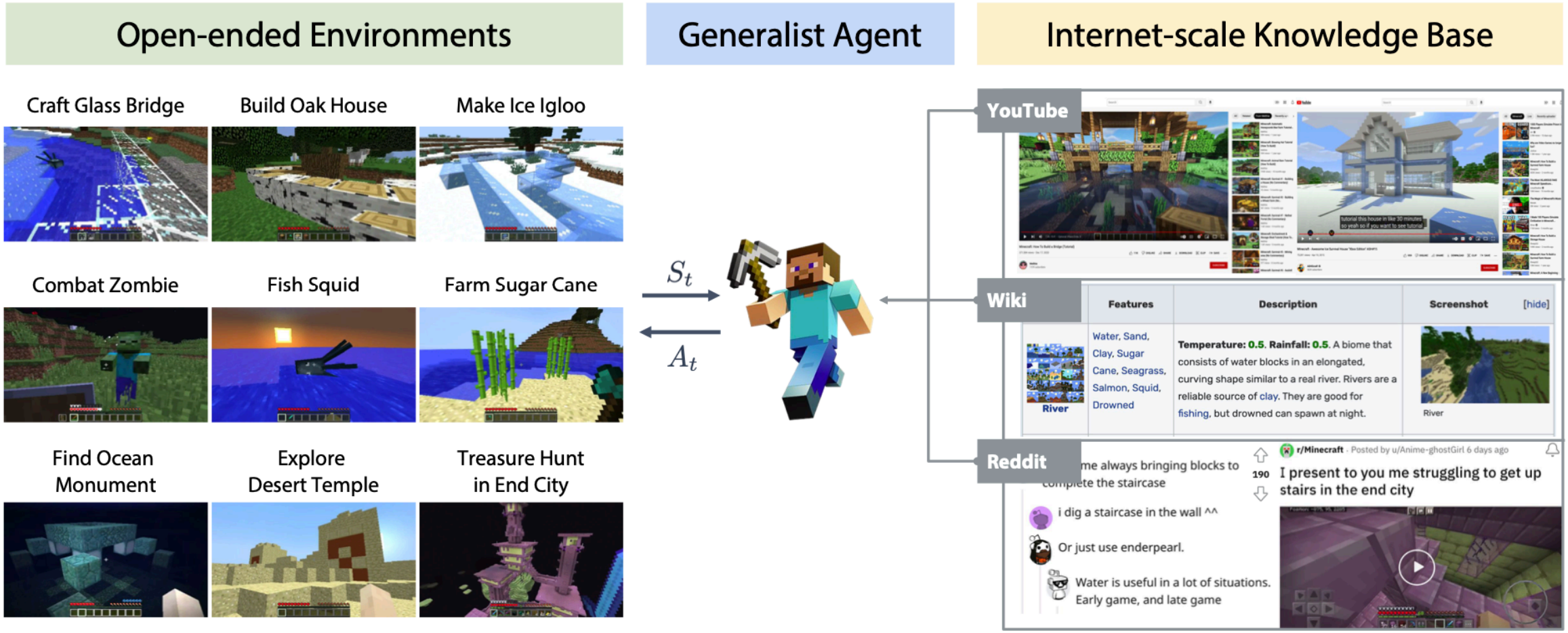
Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



MDP Environment: MineCraft

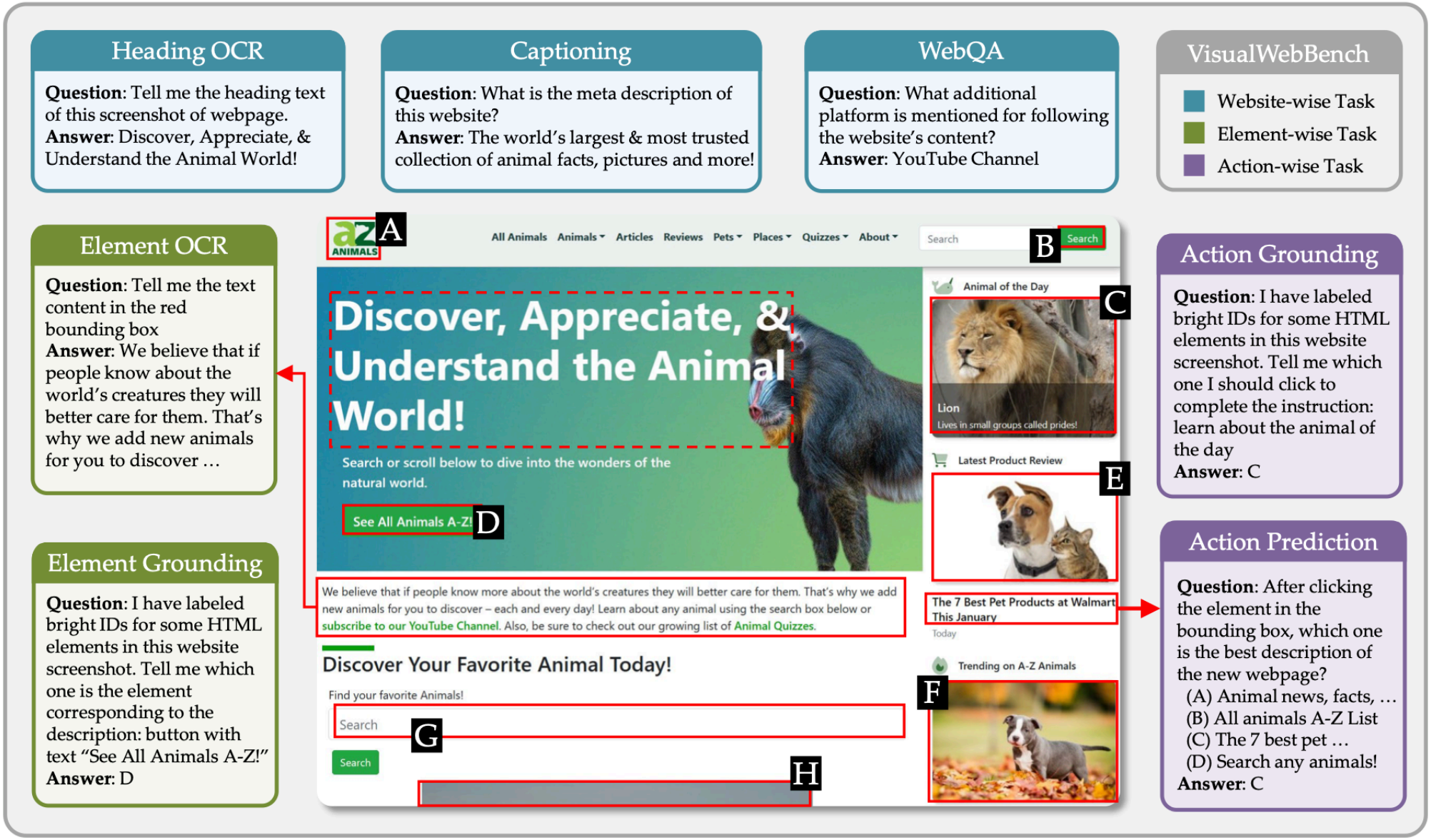


MineDojo

MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge https://arxiv.org/abs/2206.08853

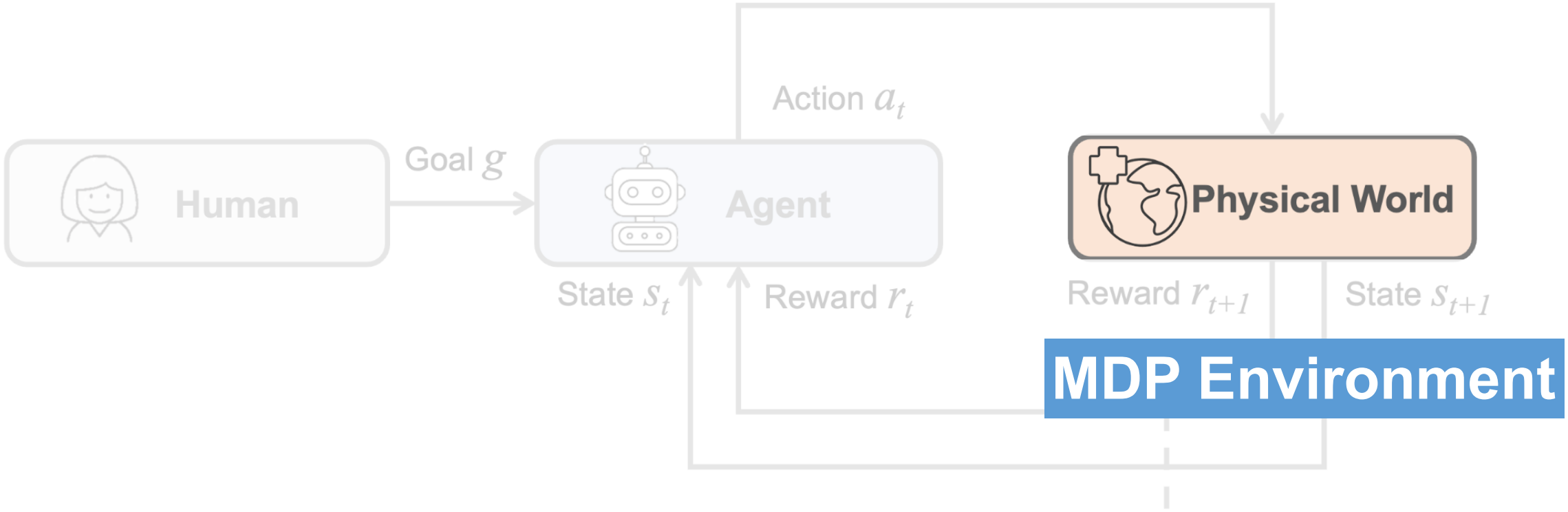
MDP Environment: GTA

MDP Environment: Web Agent Env

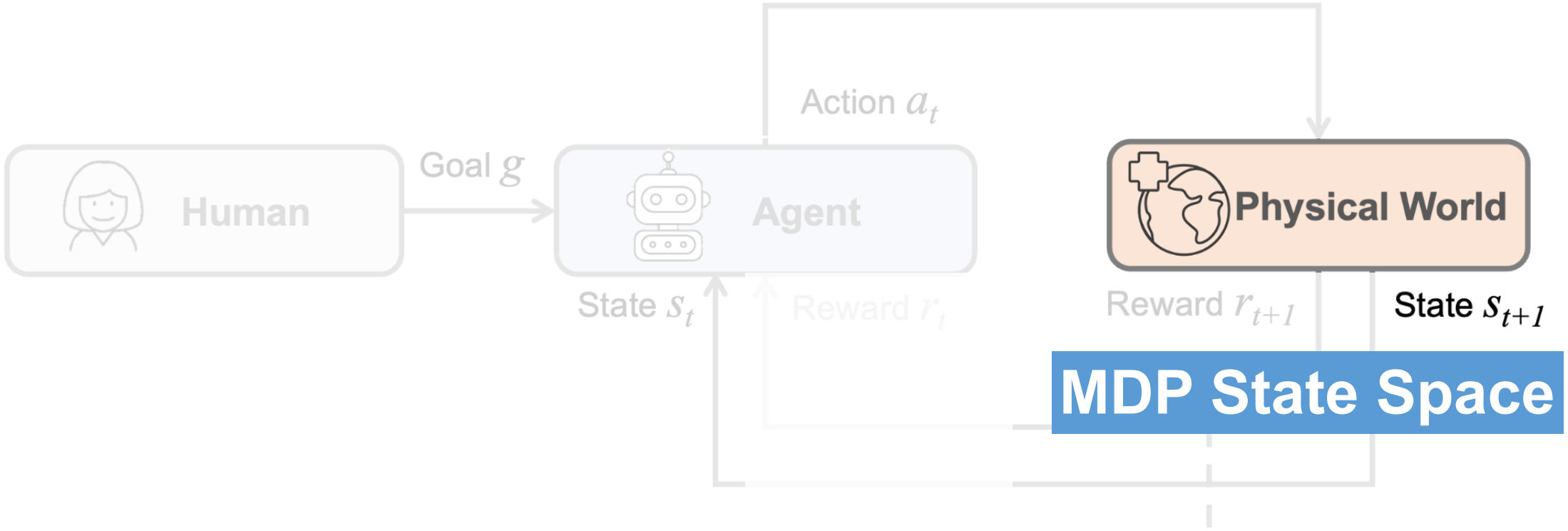


VisualWebBench: How Far Have Multimodal LLMs Evolved in Web Page Understanding and Grounding? https://arxiv.org/pdf/2404.05955

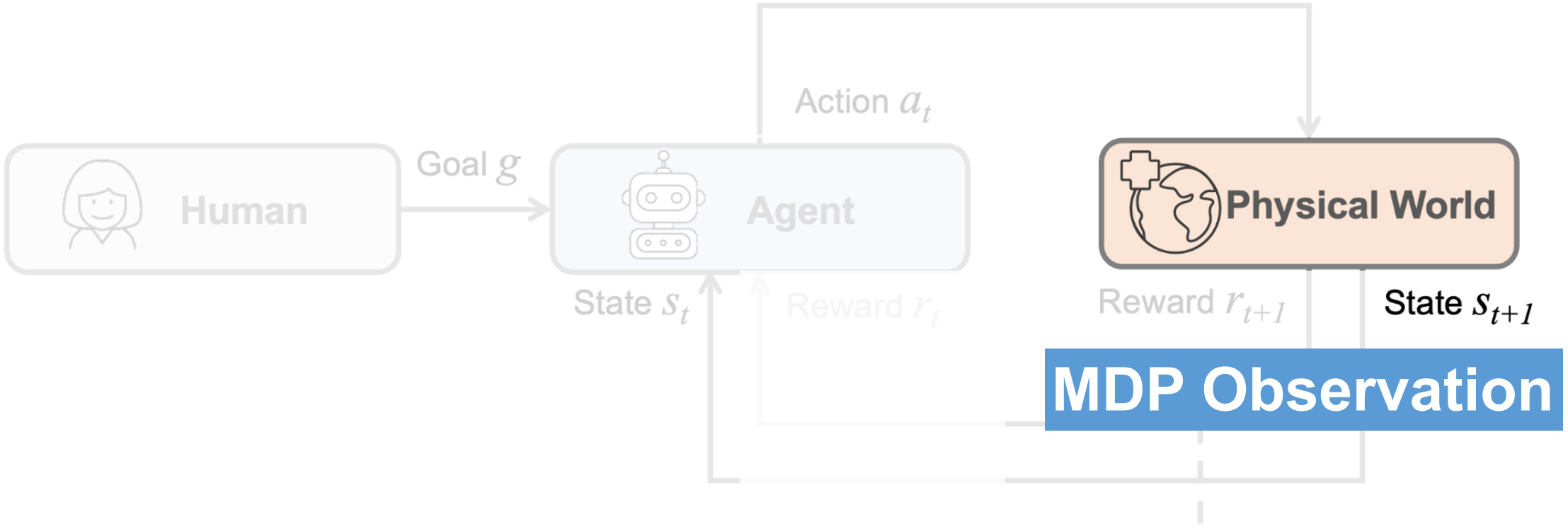
Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



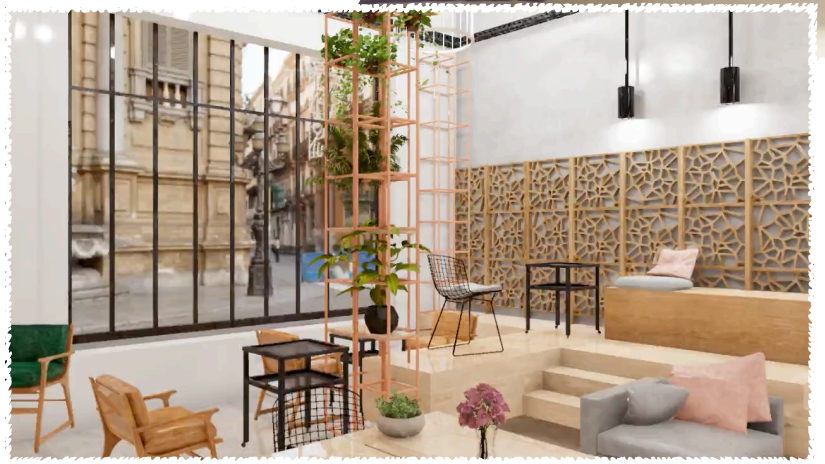
Environment：Observation

Environment：Observation (Rendered 2D Images)

Environment：Observation (Rendered 2D Images)

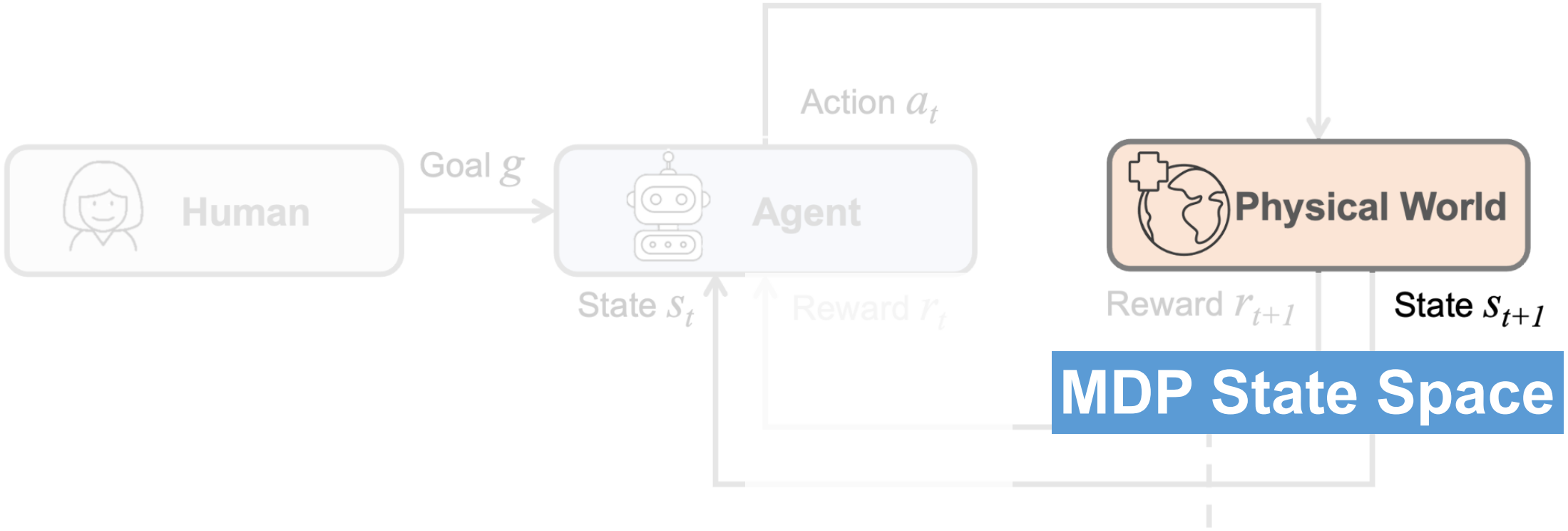
Enabled by large dataset of realistic interactive scenes and objects

50 Scenes



https://behavior.stanford.edu/

Let us go back to MDPs (Markov Decision Processes)



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

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

50 Scenes

Enabled by large dataset of realistic interactive scenes and objects



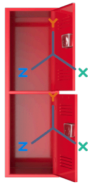



10000 Objects

Semantic

Properties: cookable, sliceable, freezable, burnable, deformable
…
Cooking temperature: 58°C
…

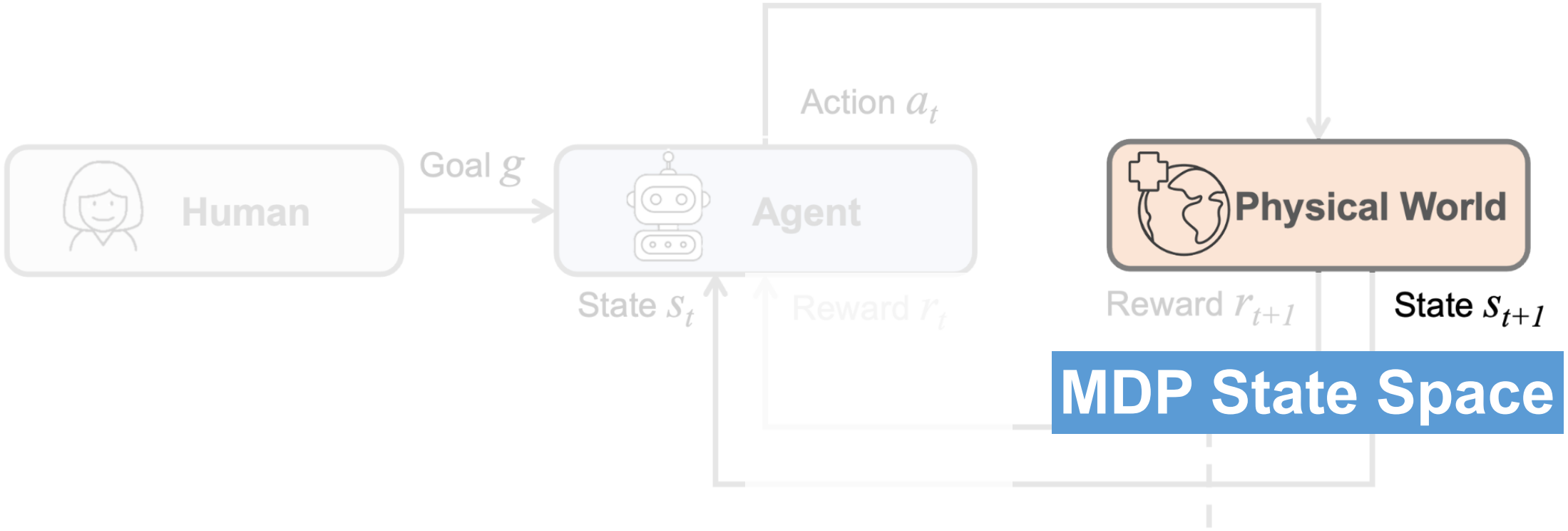
Physical



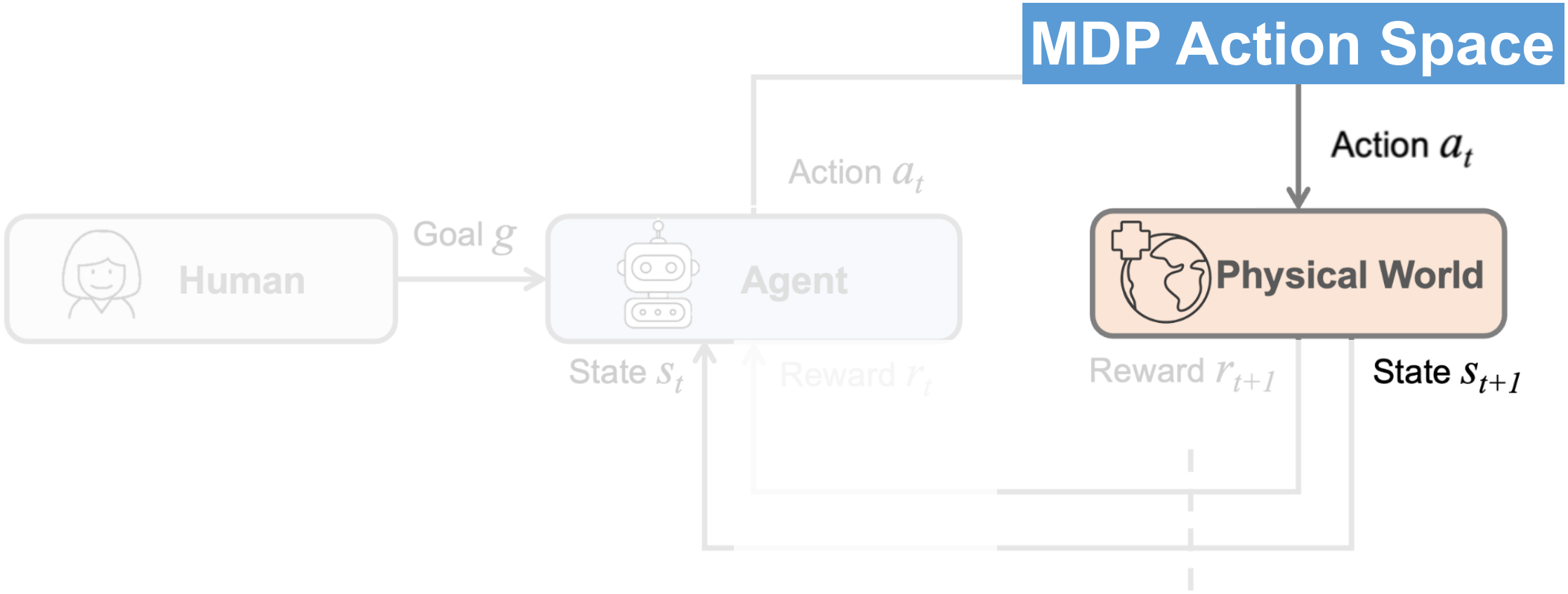
Articulation annotation (joint type, origin, axis, limit)
Mass, friction, CoM, …
Canonical size and orientation

https://behavior.stanford.edu/

Let us go back to MDPs (Markov Decision Processes)

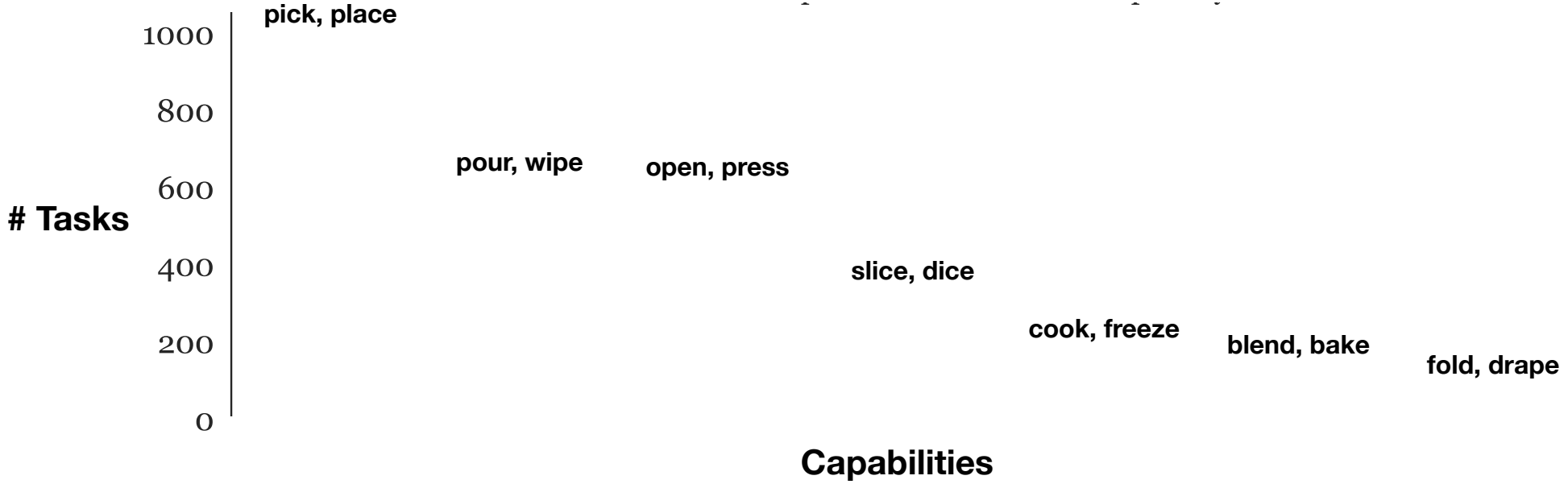


Let us go back to MDPs (Markov Decision Processes)



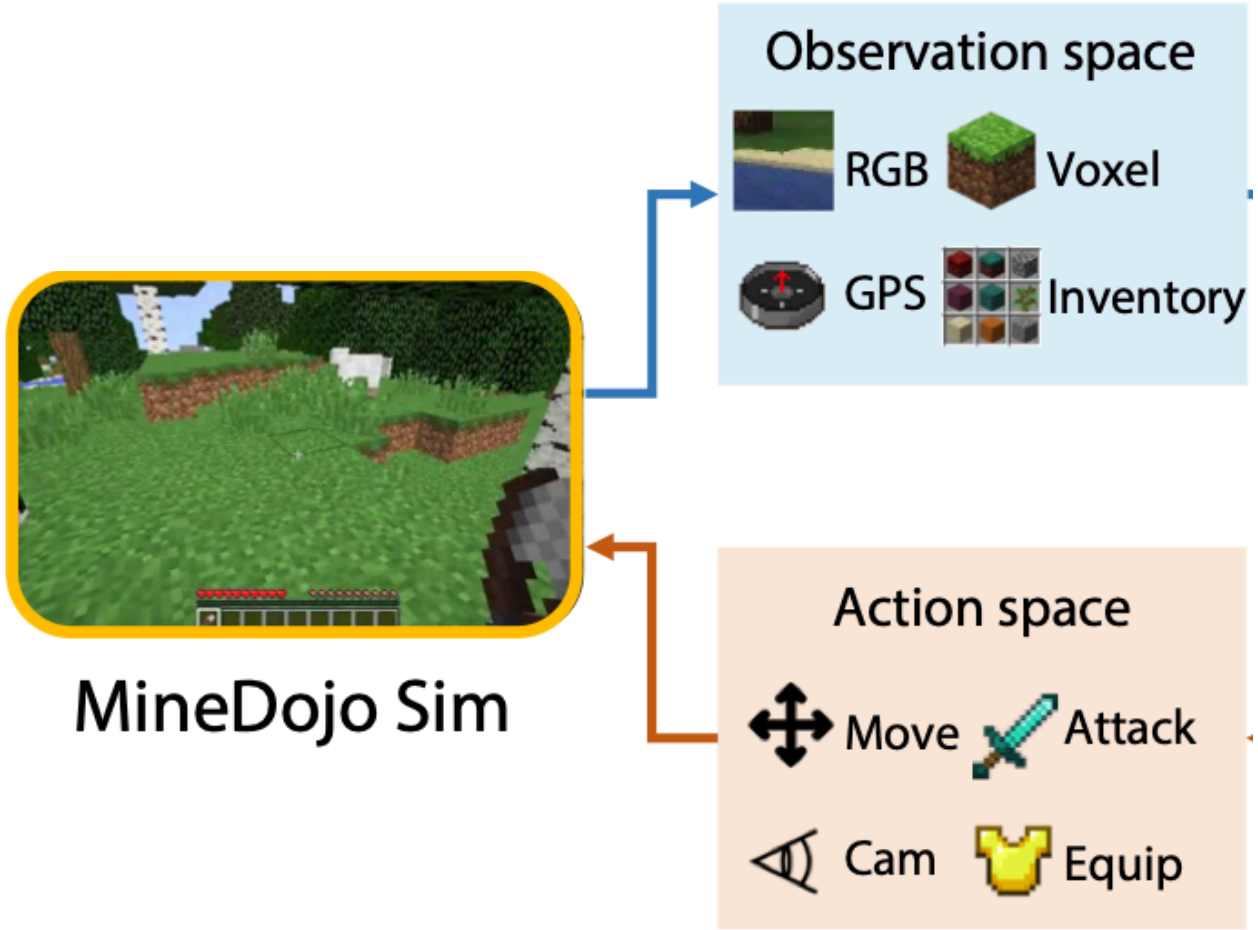
Action : can robots learn to solve these tasks?

→ What capabilities are needed?

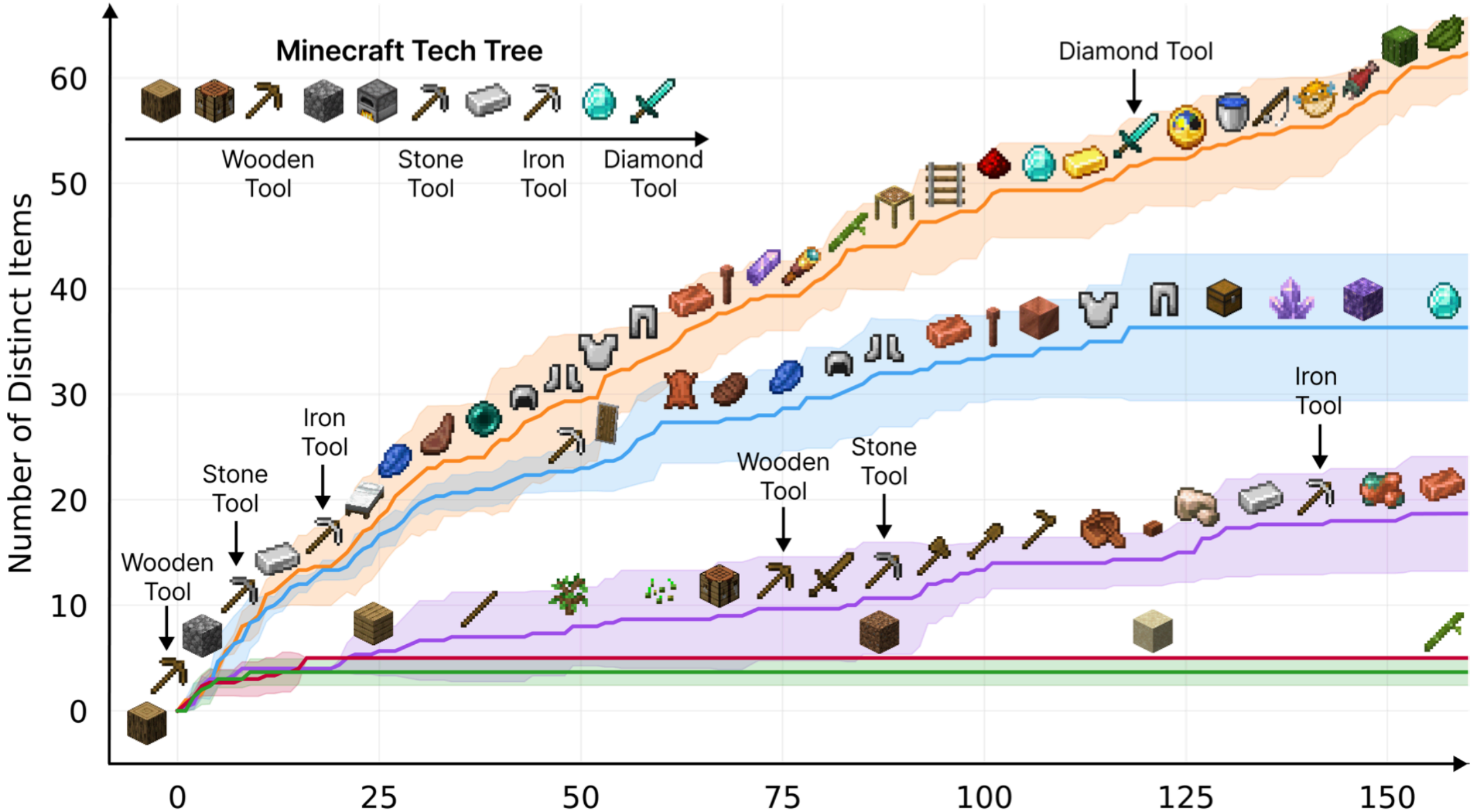


https://behavior.stanford.edu/

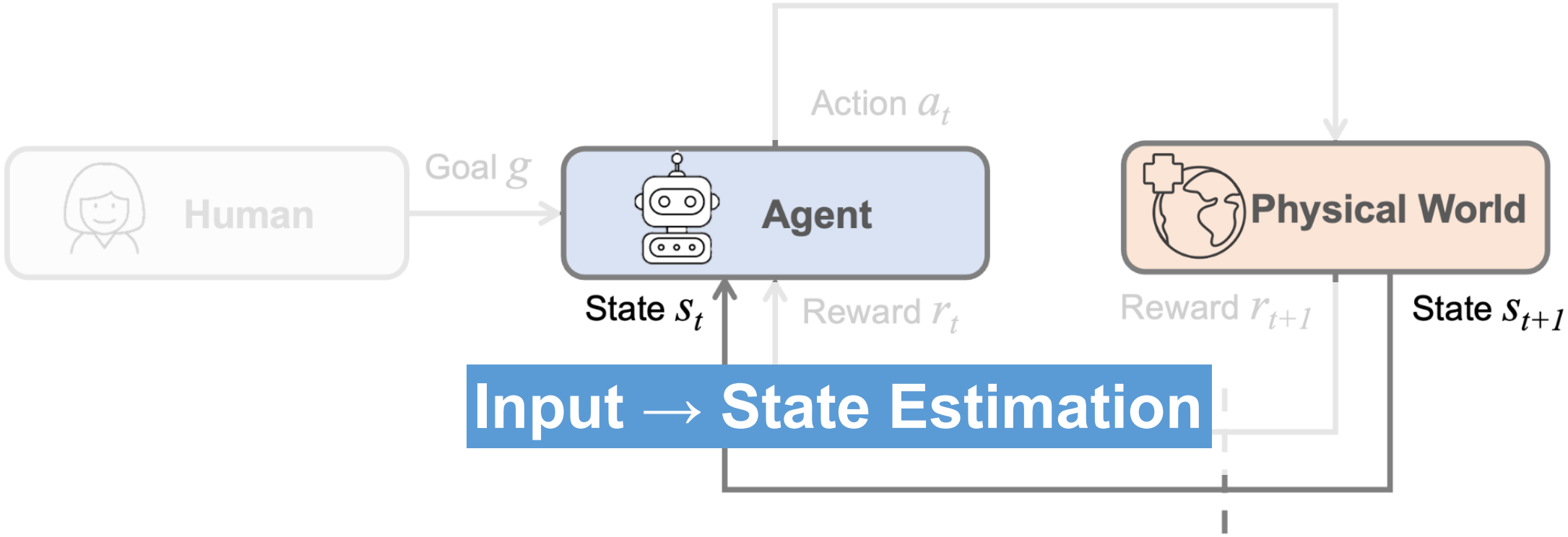
MDP Action Space: Skills



MDP Action Space: Skills



Let us go back to MDPs (Markov Decision Processes)



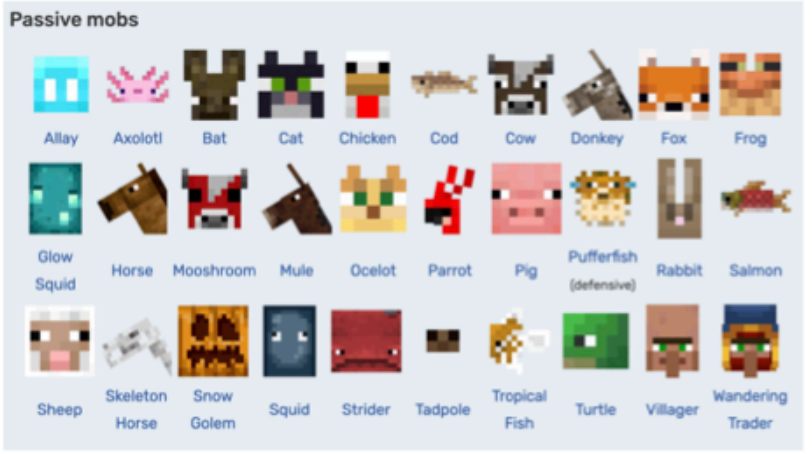
Perception / State Estimation

$o \rightarrow s$

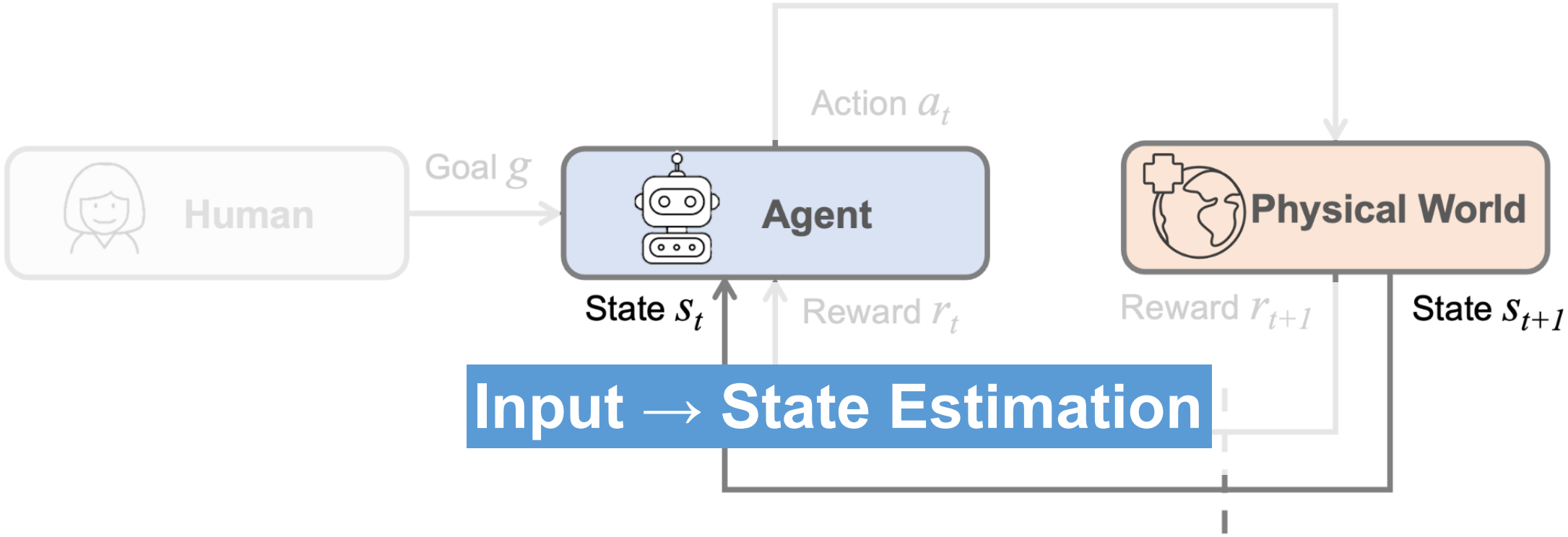
Observation → State Estimation (Assets & States)

Observation (2D rendered scenes)

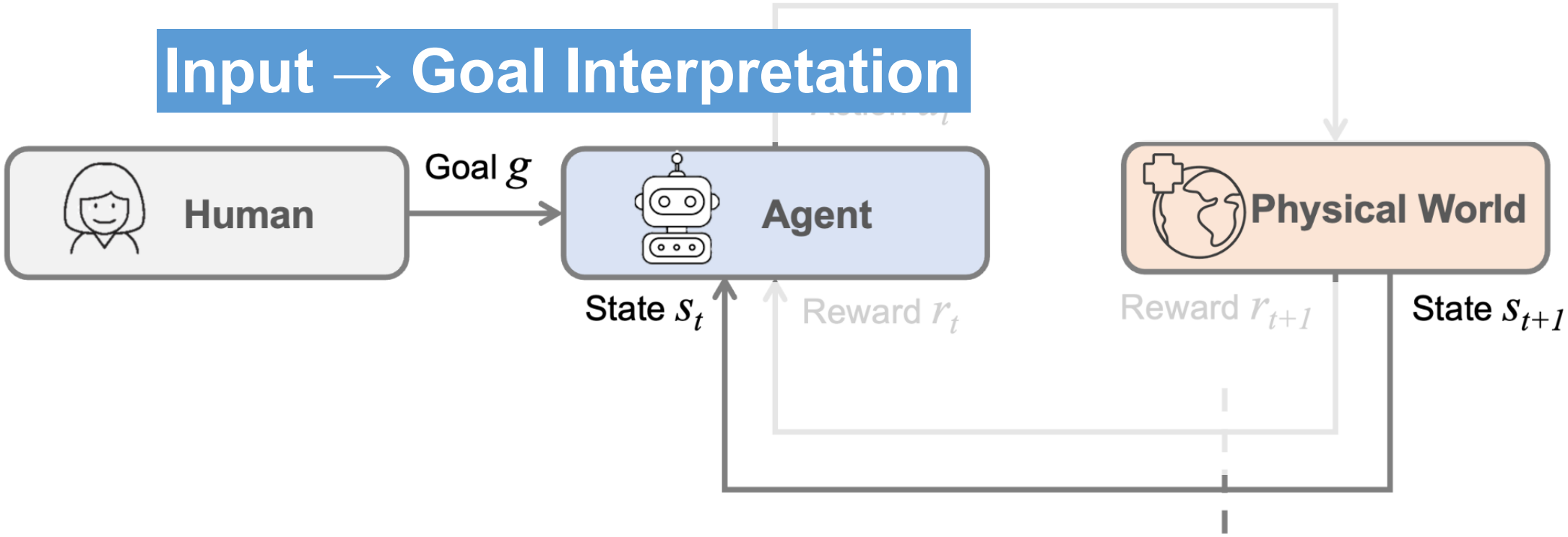
State (3D assets)



Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Goal Interpretation

g

Set up the table.

...

on_top_of (turkey, table)

...

LLM

Set up the table.

...

on_top_of (turkey, table)

...

LLM

Use the plates.

...

on_top_of (plate, table)

on_top_of (turkey, plate)

...

LLM

Goal : defines a task?

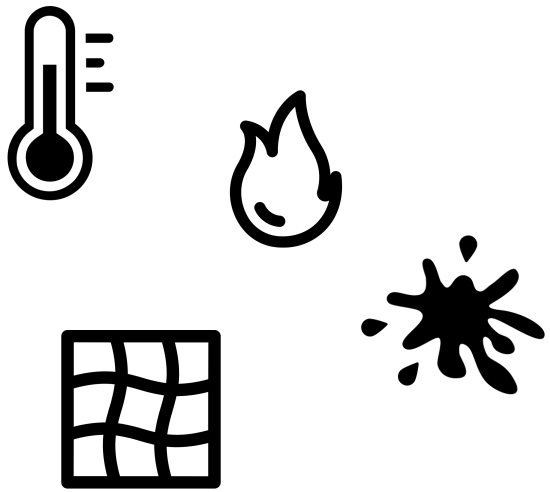
wikiHow to do anything

YouTube

“Pack lunch”

What objects?

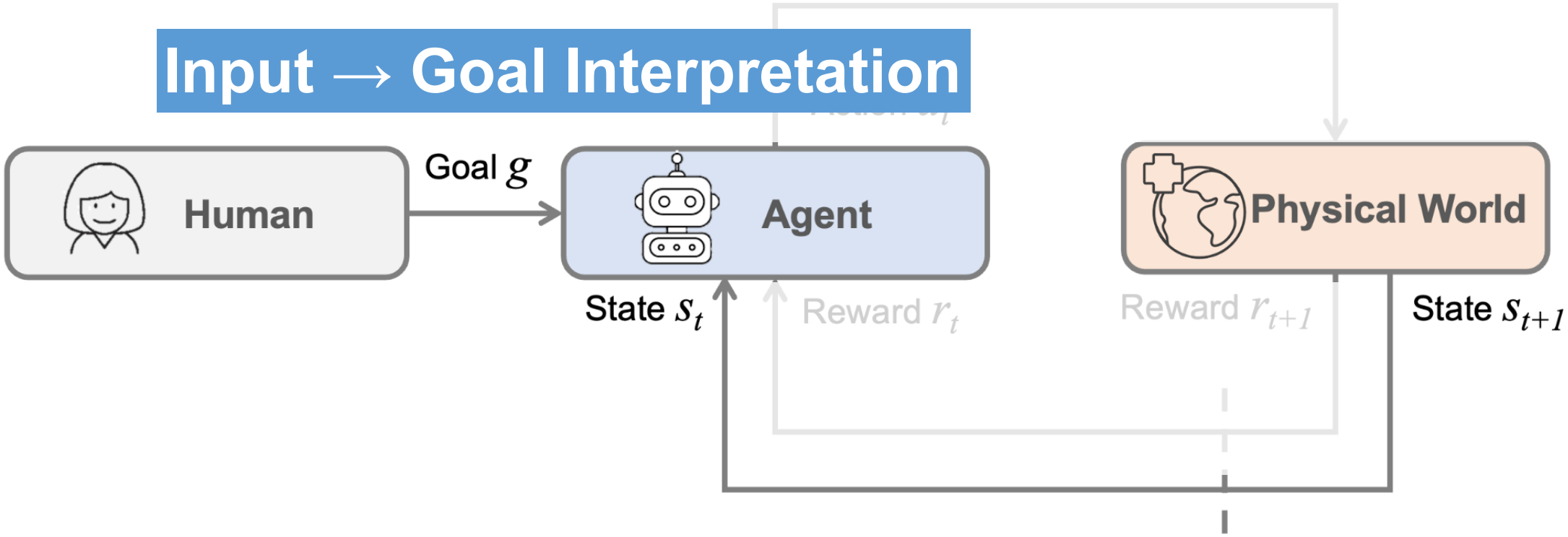
What properties?



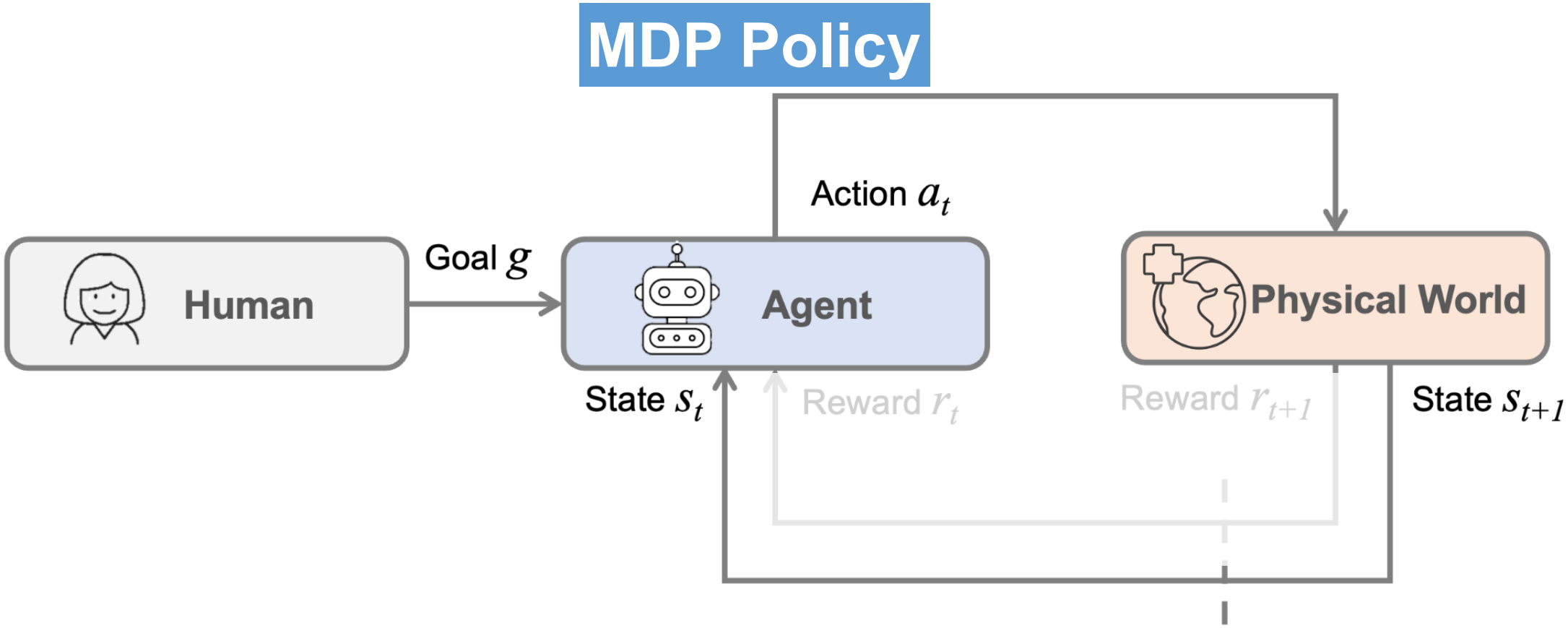
Start & Goal?

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

Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



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)

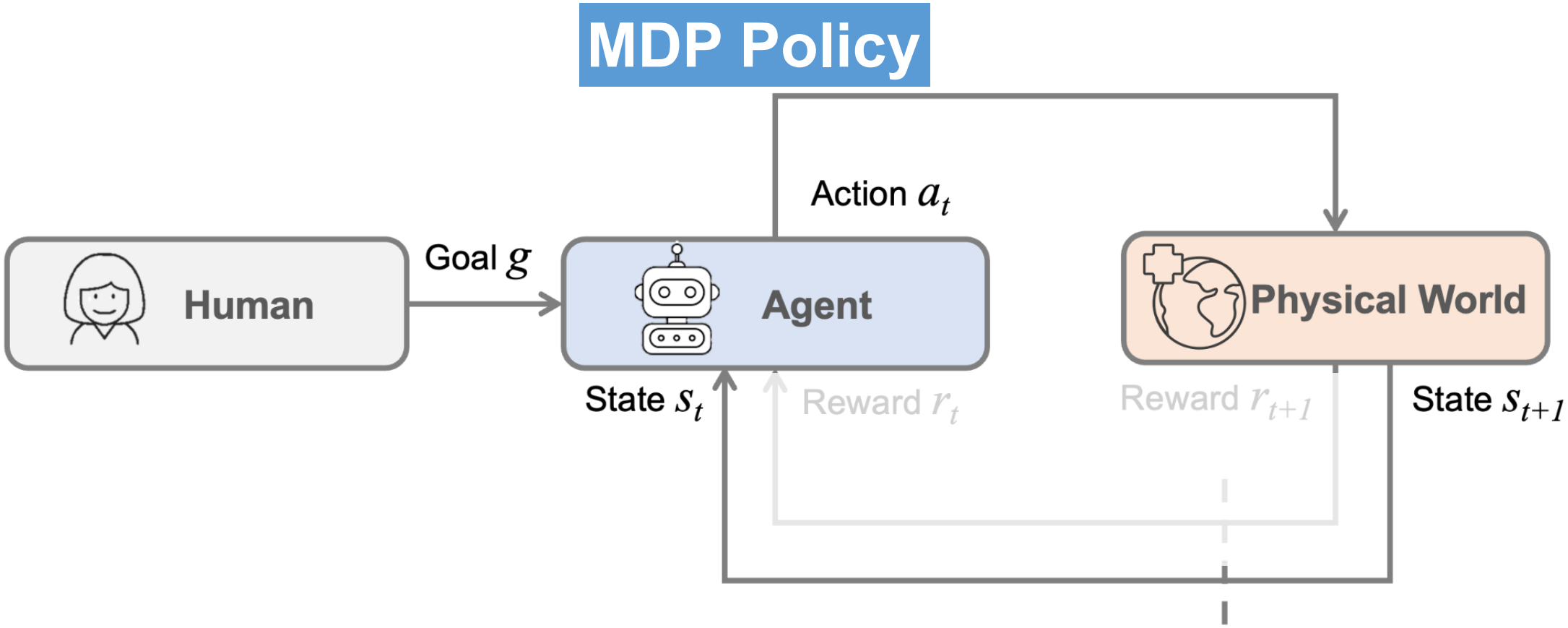
A11 COOK(strawberry)

…

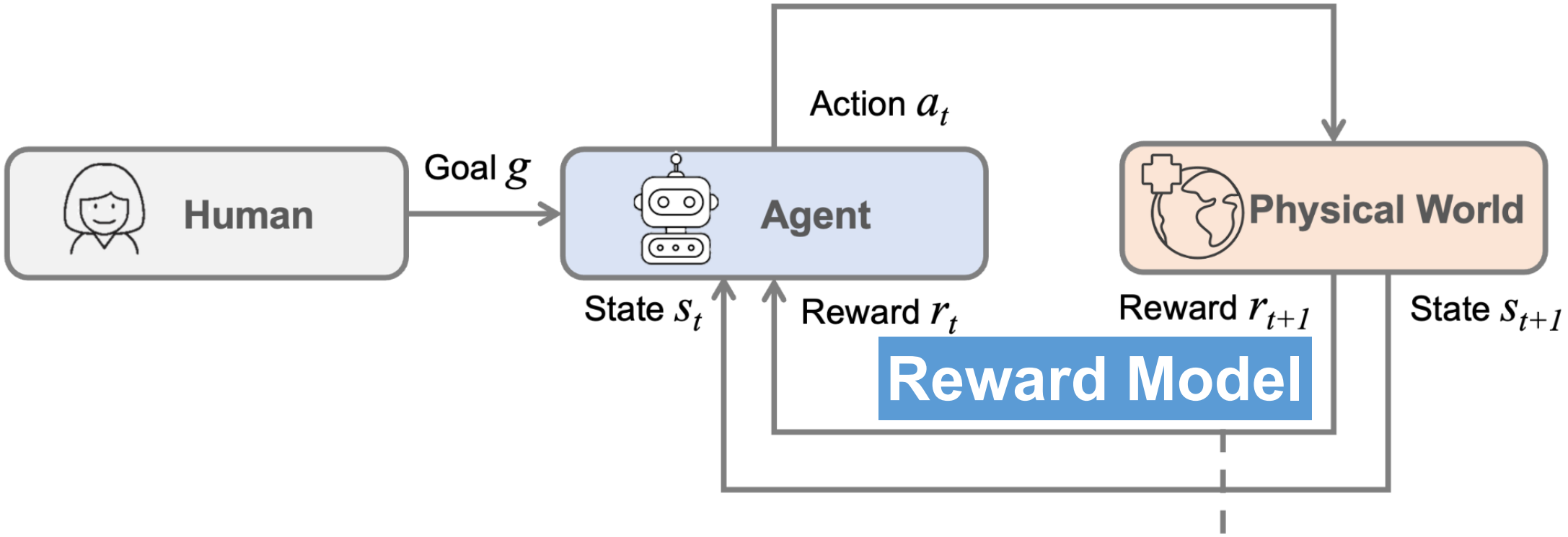
LLM Output

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.

Let us go back to MDPs (Markov Decision Processes)



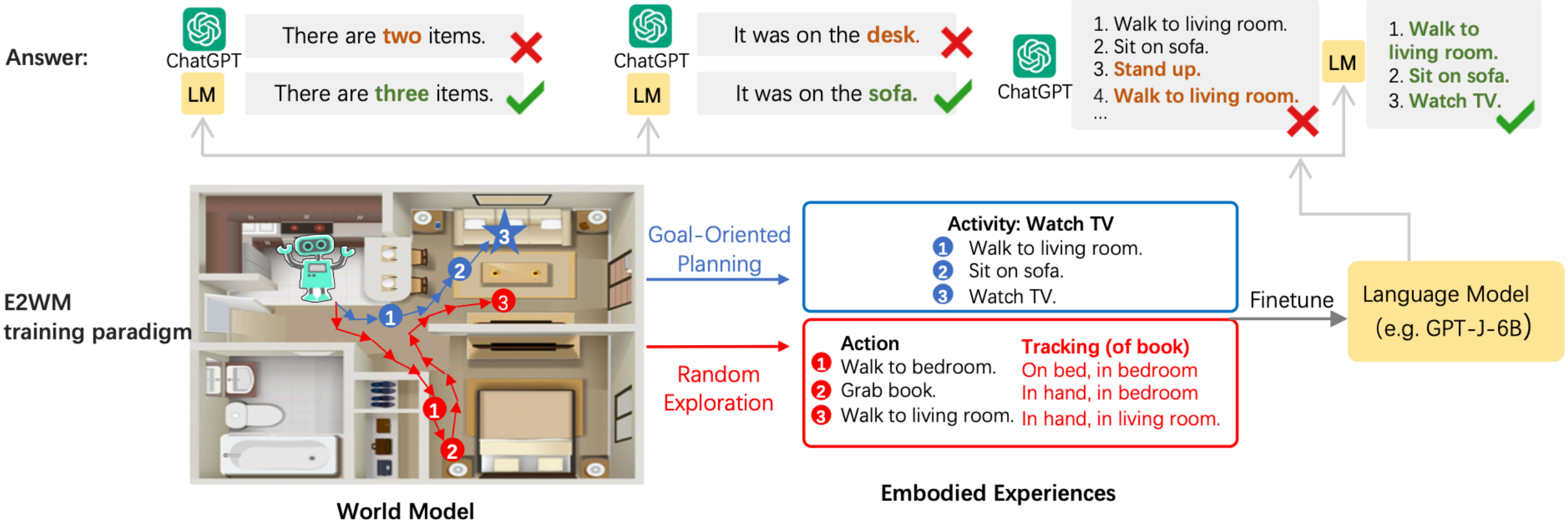
Let us go back to MDPs (Markov Decision Processes)



Reward Model

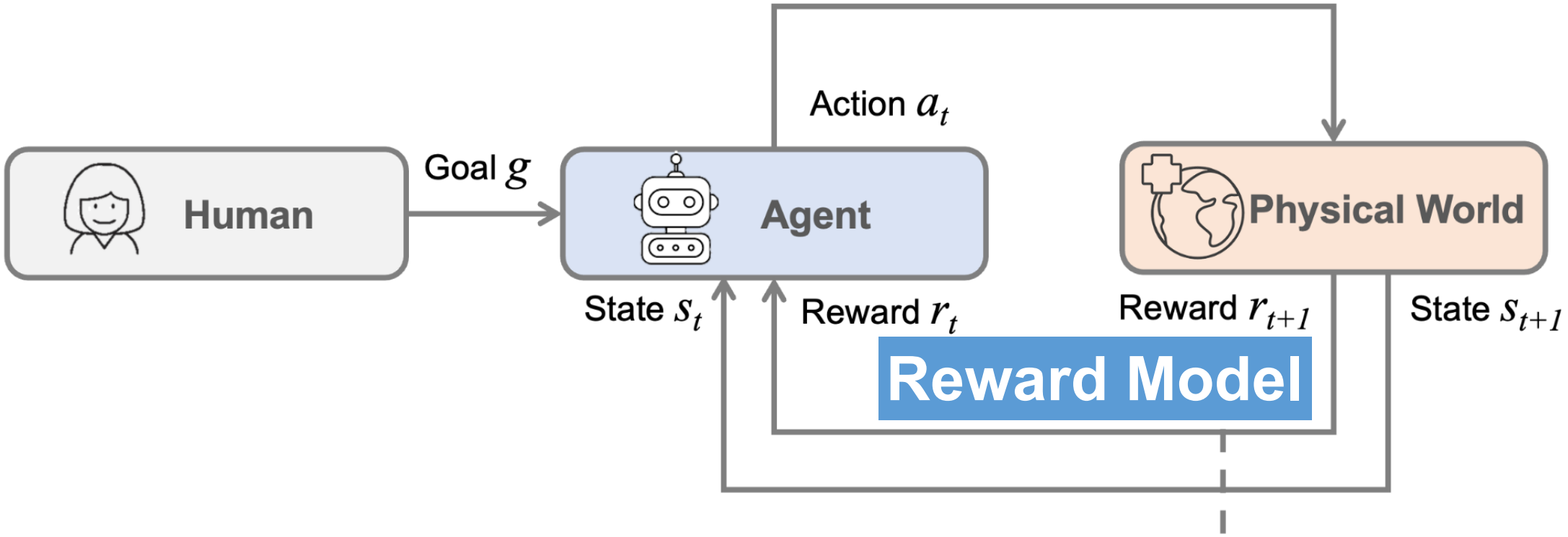
$o, a \rightarrow r$

Reward Model

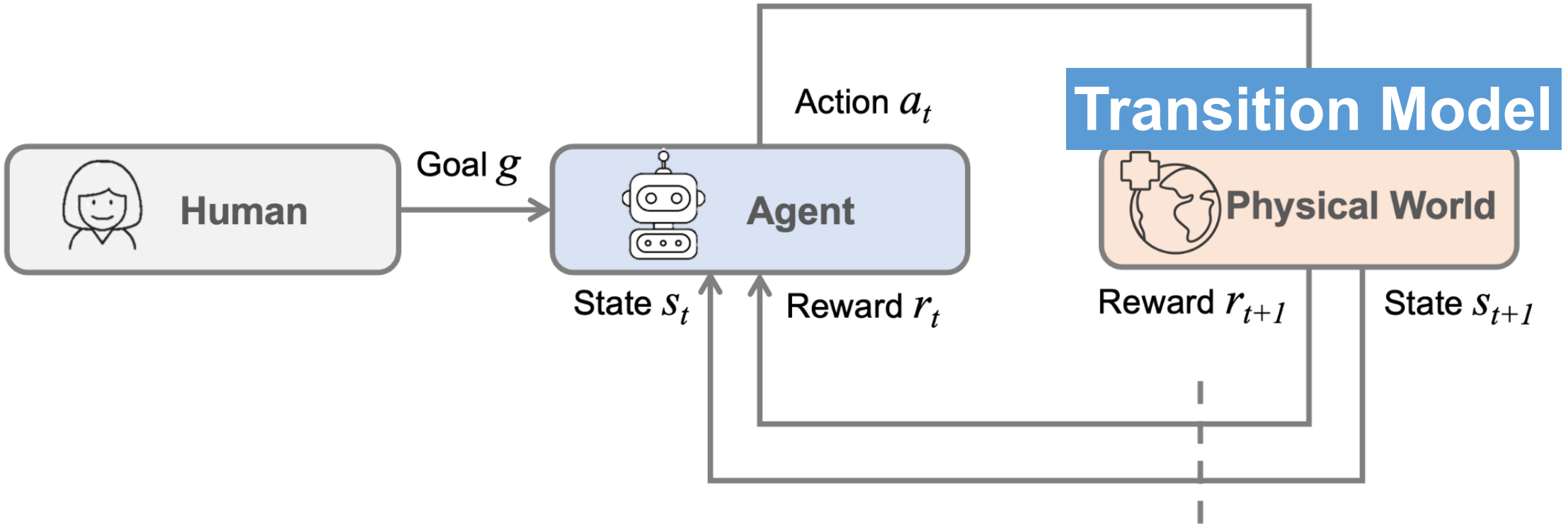


https://arxiv.org/pdf/2305.10626

Let us go back to MDPs (Markov Decision Processes)



Let us go back to MDPs (Markov Decision Processes)



Transition Model

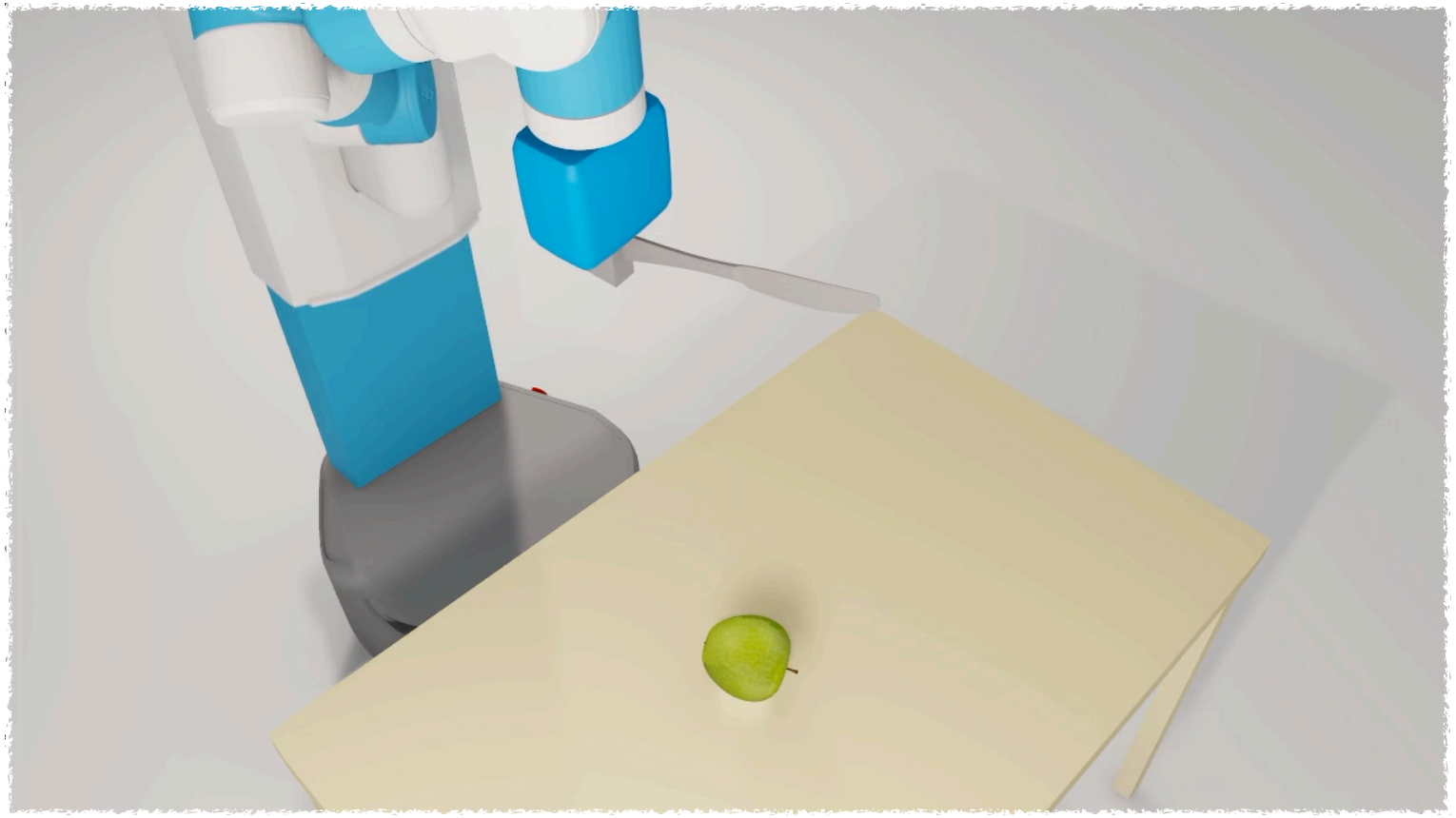
$$o_t, a \rightarrow o_{t+1}$$

"World Modeling"

OmniGibson

Physics Transition Examples

```
class SlicingRule:
```



https://behavior.stanford.edu/

OmniGibson

Physics Transition Interface

```
class TransitionRule:

def condition(self, *args) -> bool

def transition(self, *args)
```

Determines whether a transition should occur

What should happen when a transition is triggered

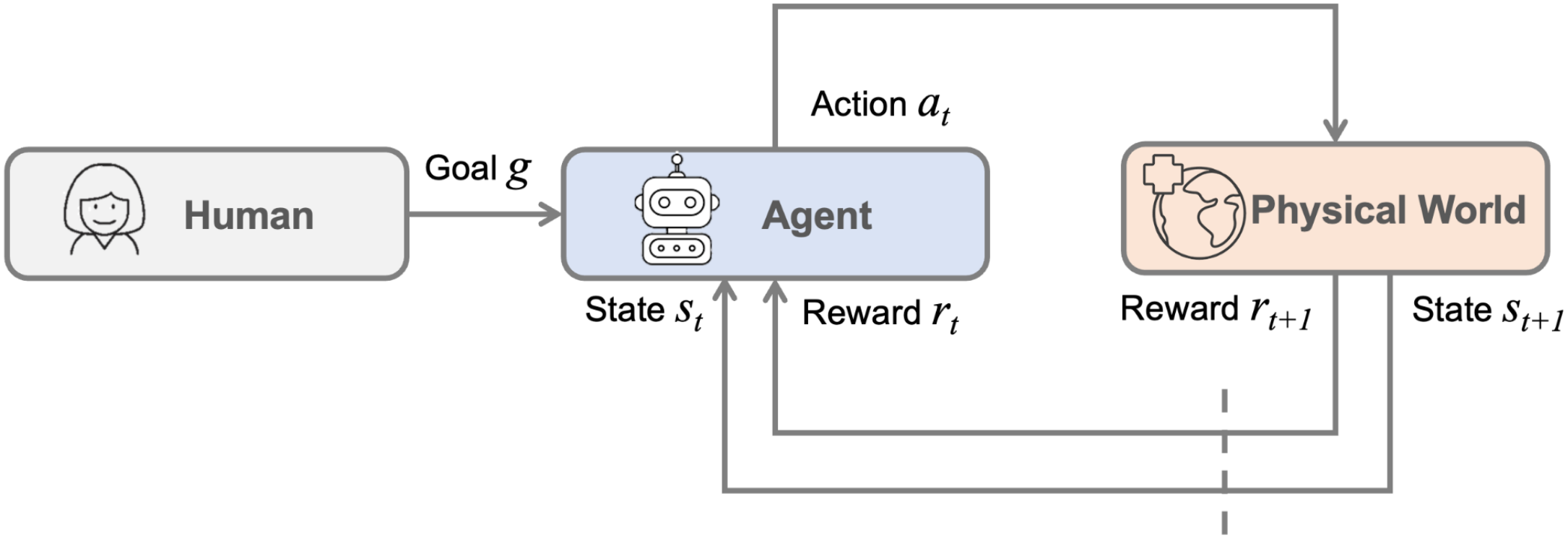
Allows us to capture arbitrarily complex **physical phenomena**!

OmniGibson

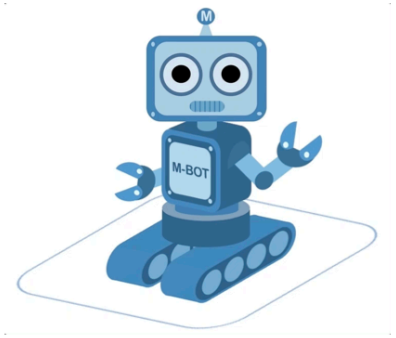
Physics Transition Examples

```
class SlicingRule:

def condition(self, *args) -> bool

    # Return True if a slicer object is touching
    a sliceable object with sufficient force


def transition(self, *args)

    # Remove the sliceable object and
    import its sliced component objects
```

Let us go back to MDPs (Markov Decision Processes)



Virtual Agents vs Physical Agents



Different Instantiations of MDP

Perfect actuations

Perfect odometry

RGB only perception

●
●
●

Noisy actuations

Noisy odometry

Multiple perceptual modalities

●
●
●

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

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
4. Robotic Foundation Models	30min	Yunzhu Li
5. Remaining Challenges	15min	Yunzhu Li
QA	30min	

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
4. Robotic Foundation Models	30min	Yunzhu Li
5. Remaining Challenges	15min	Yunzhu Li
QA	30min	

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
4. Robotic Foundation Models	30min	Yunzhu Li
5. Remaining Challenges	15min	Yunzhu Li
QA	30min	