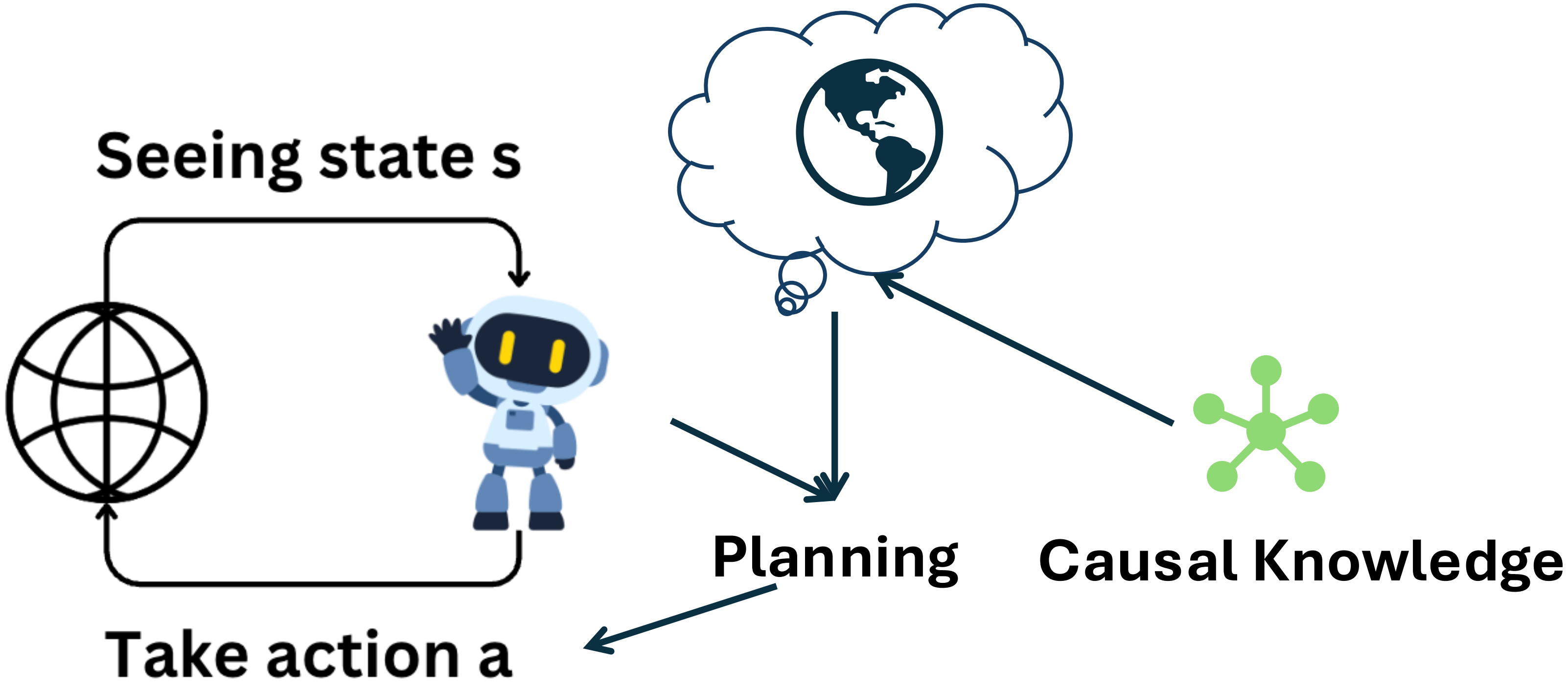
Towards Causal Foundation World Models

Mengyue Yang

Lecturer in AI

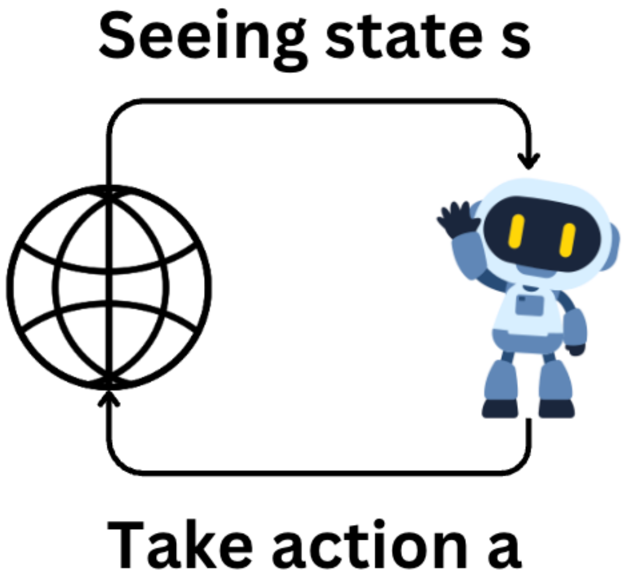
University of Bristol



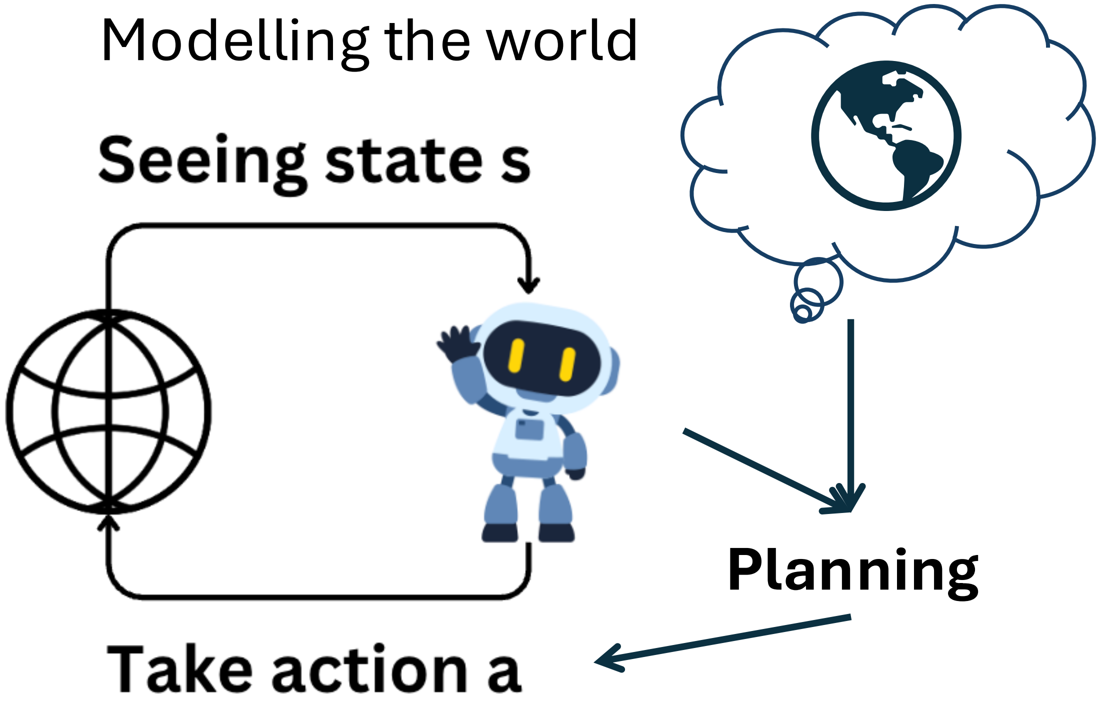
What is Agent Decision Making?

Agent policy $\pi(a|s)$: The way that agent to achieve the goal

Model Free



Modelling the world



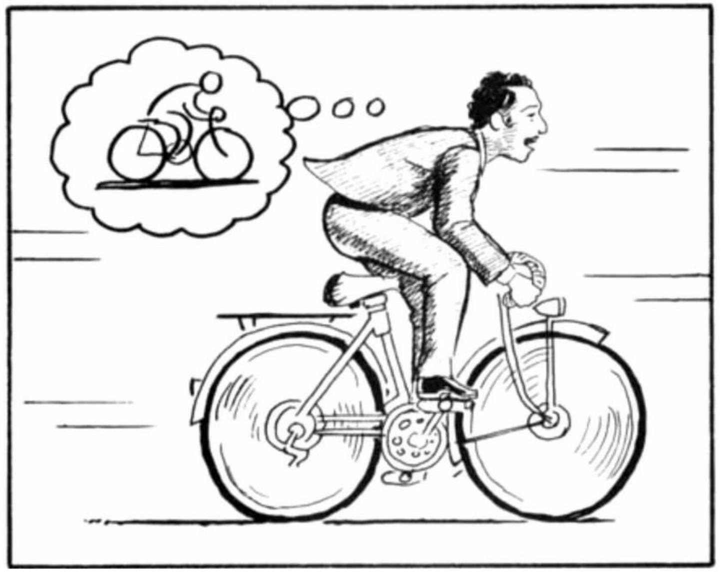
Improve policy by exploration

Planning from the knowledge and thinking

Why Modelling World is Necessary

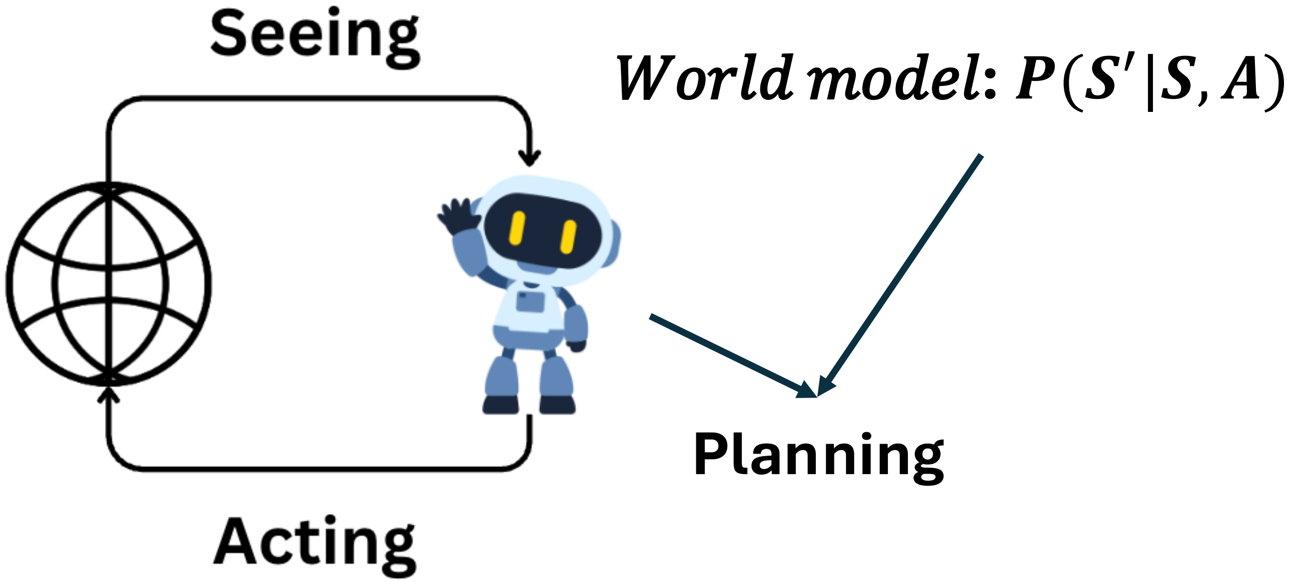
What is World Models

World Models modelling the **environment internal transition model**, makes agent 1. predict the future and 2. reflect the past possible.



What if I ride slower now? Can I achieve the goal?
Was it a good choice to start earlier?

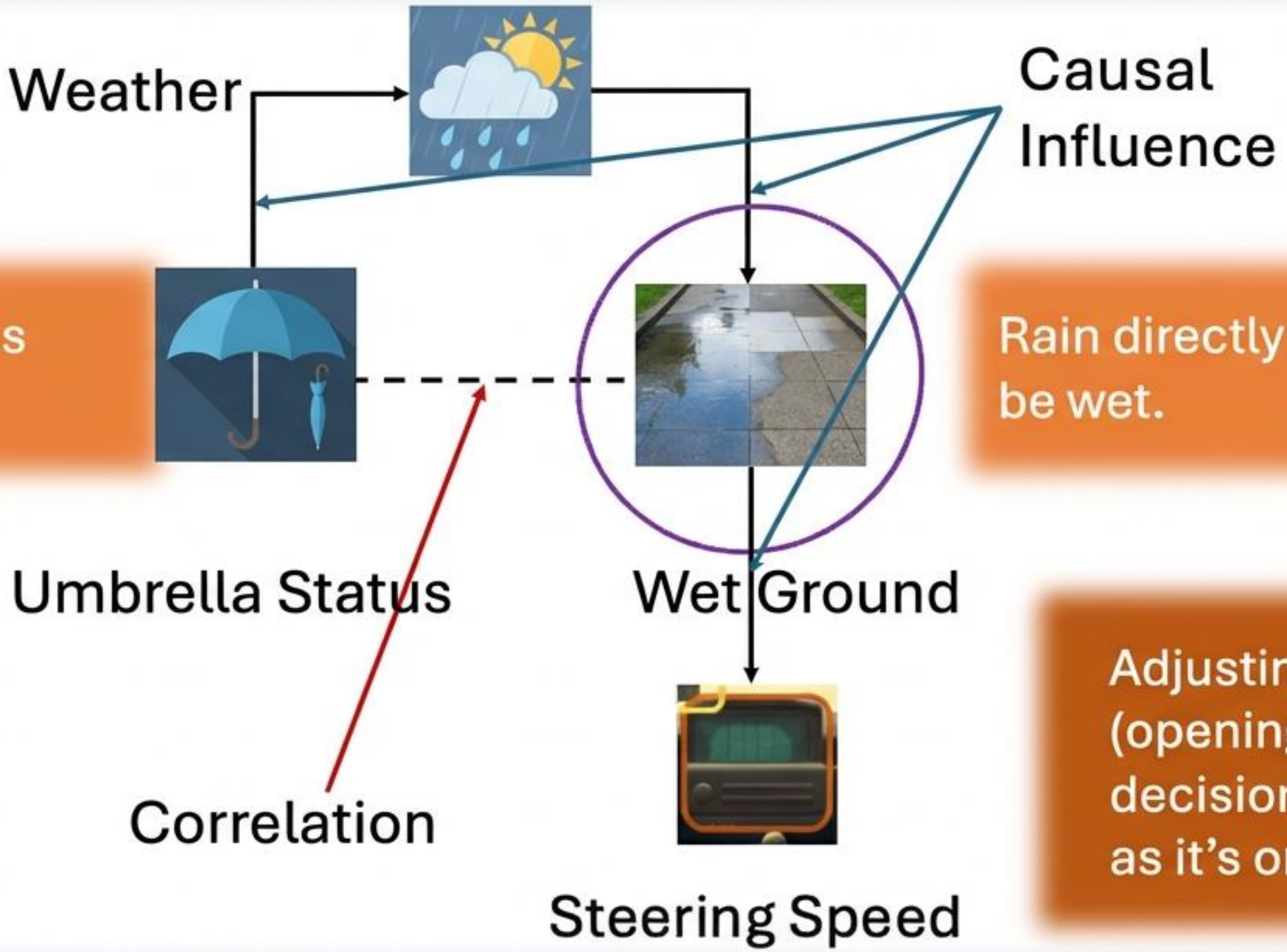
Picture borrow from Ha and Schmidhuber



Modelling the World ≠ Understanding the World

Is current world model perfect?

Mimic the world data doesn’t means model understand the world



Rainy weather causes people to open umbrellas and also makes the ground wet.

People open umbrellas when it rains.

Rain directly causes the ground to be wet.

Adjusting Umbrella Status (opening/closing) is a wrong decision to change Wet Ground, as it’s only a correlation.

Pearl’s Causal Hierarchy

Seeing	Doing	Imagining
	Predict something haven’t happened	Imagine based on something already happened
Association	Intervention	Counterfactual
Question: What is?	Question: What if?	Question: Was it?
What does the smoking tell us about the lung cancer.	What will happen if someone keep smoking	Would the lung cancer got worse if someone smoking.

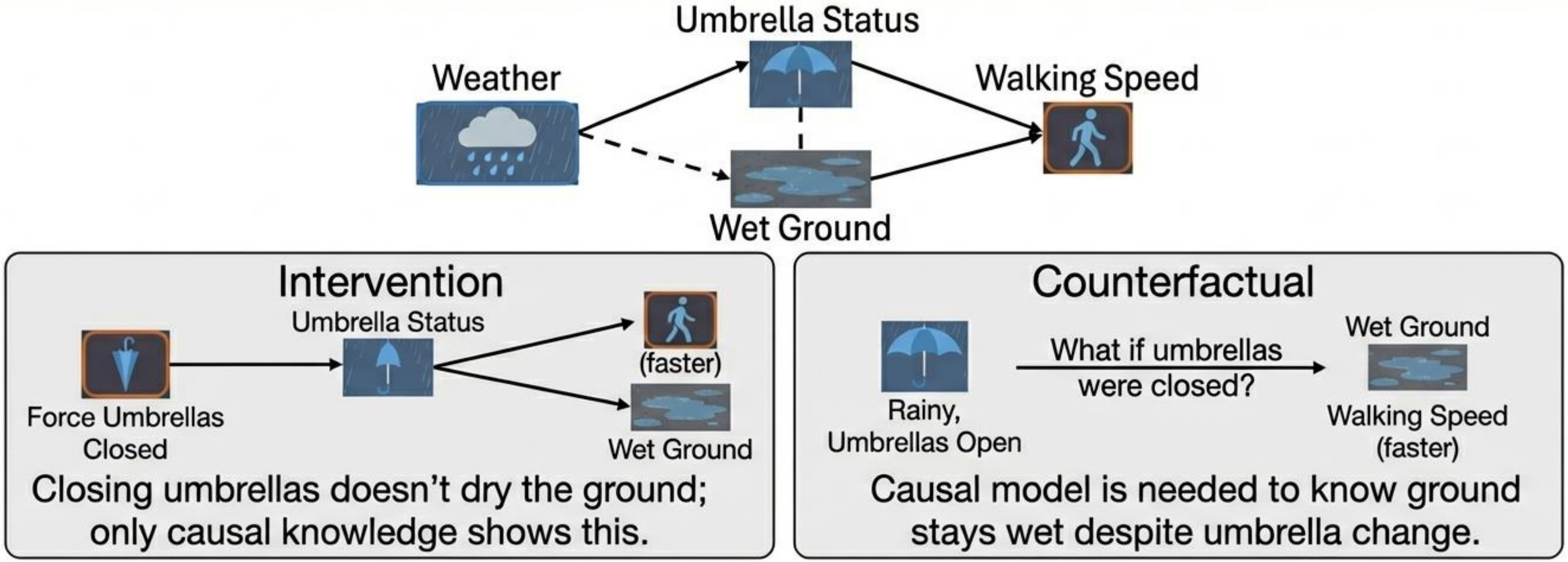
Predicting of future
Reflecting of past

=

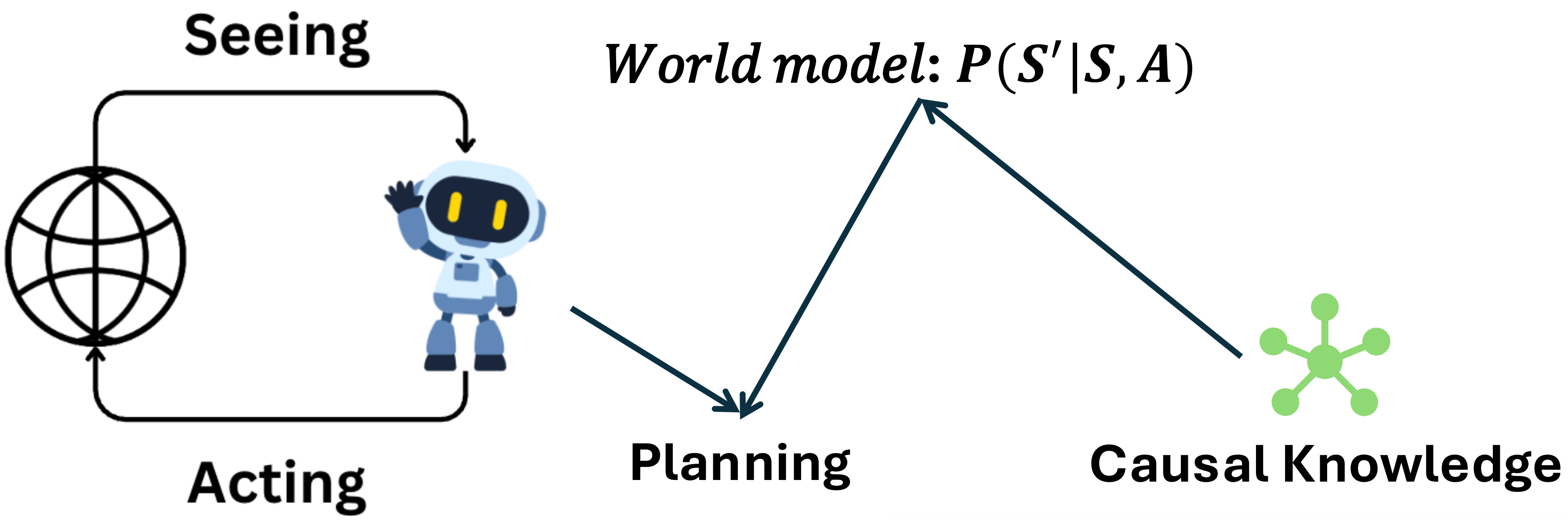
Intervention
Counterfactual

Causal Diagram

Understanding the underling causal model is a prerequisite for inferring intervention and counterfactual

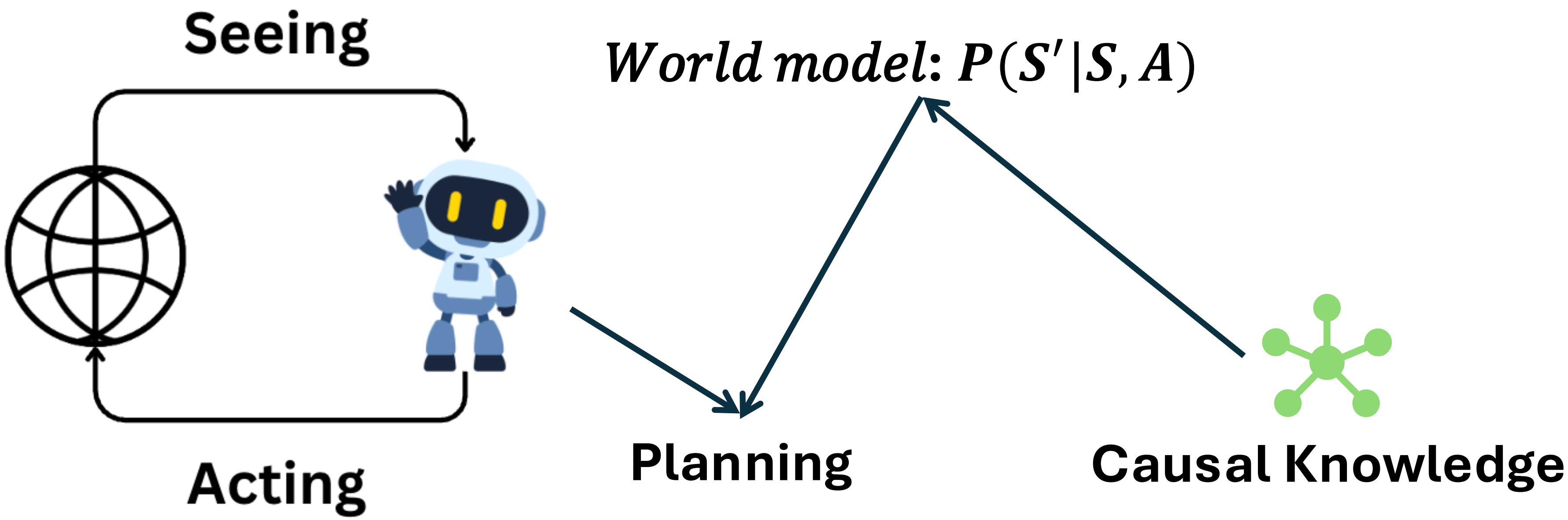


Causal World Models



Remove spurious correlation
where S,A have causal effect on S’

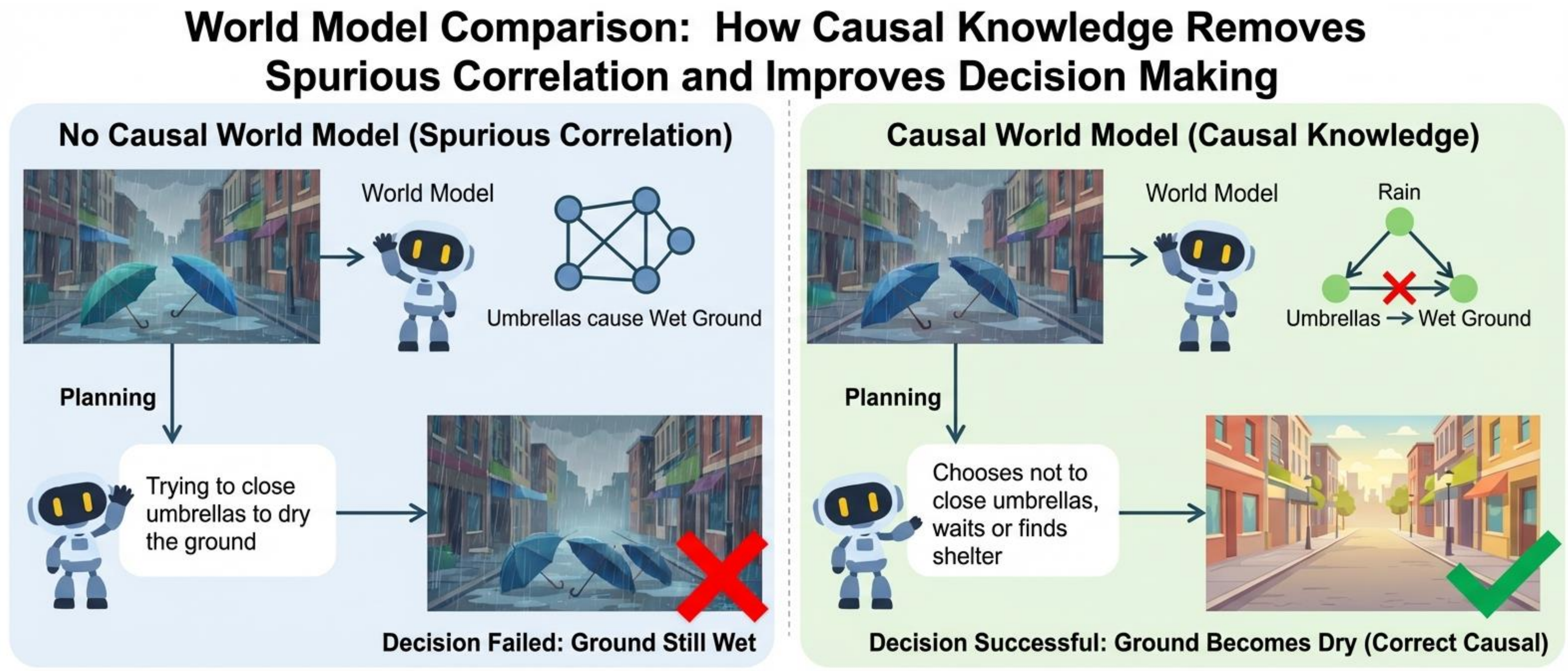
Causal World Models



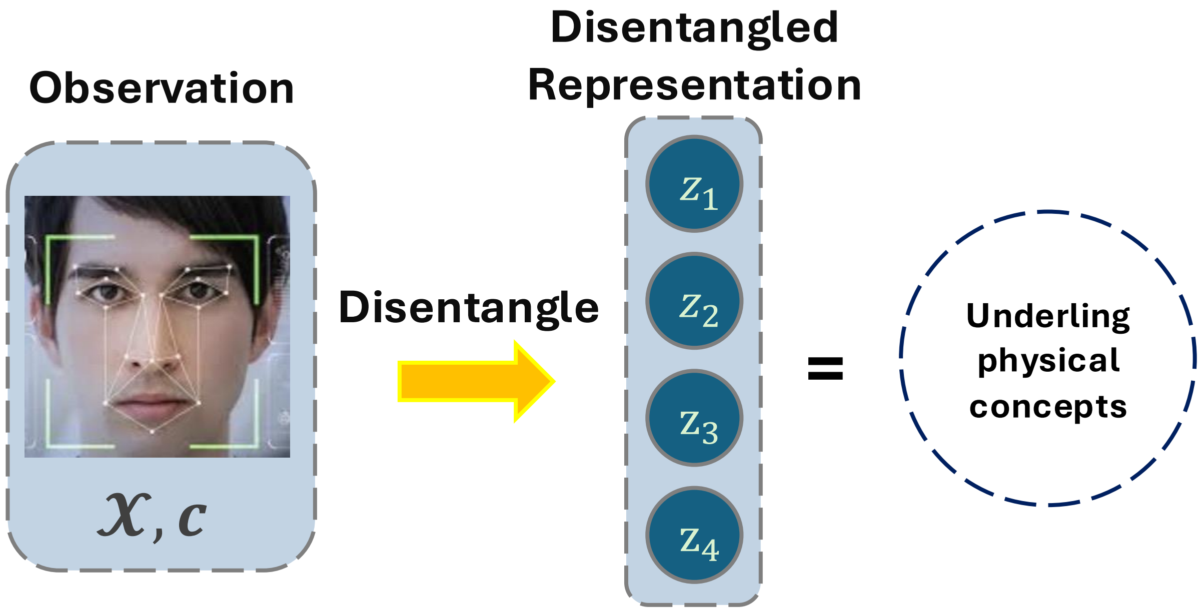
Focus planning by the true causal relations

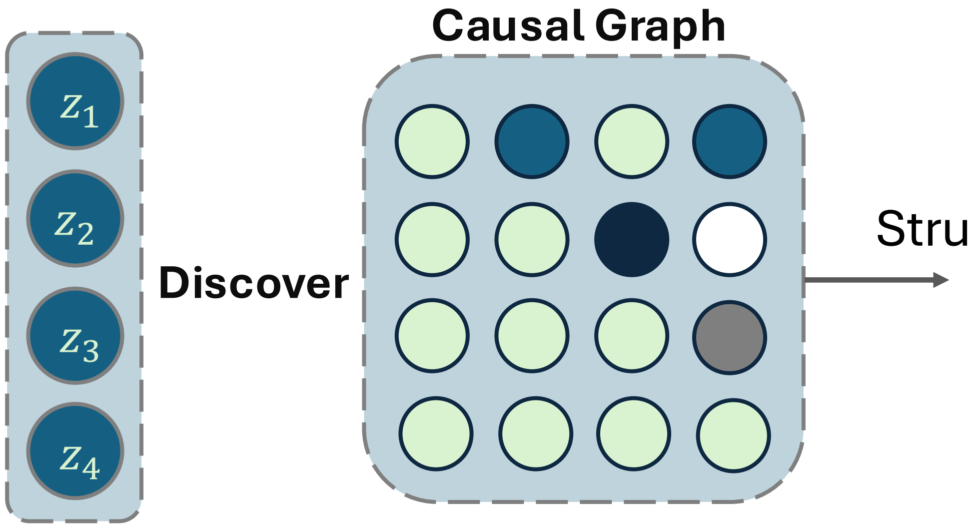
Predicting of future Reflecting of past	=	Intervention Counterfactual

Causal World Models



Causal World Models – Representation + Causal Structure

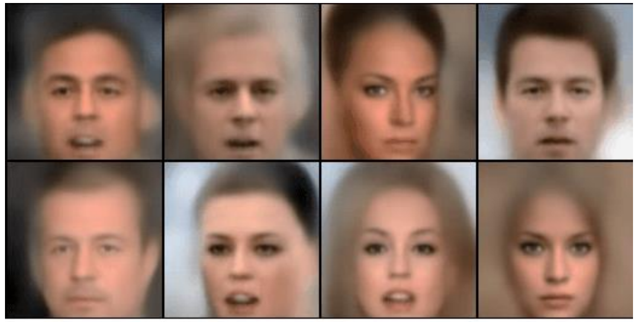




Structure causal model

$$Z_i = f_i(pa_i, U_i)$$

Intervene causes Smile, Mouth Open will change



Smile = -0.75

But Intervene effect concept Mouth Open, Smile will not be influenced

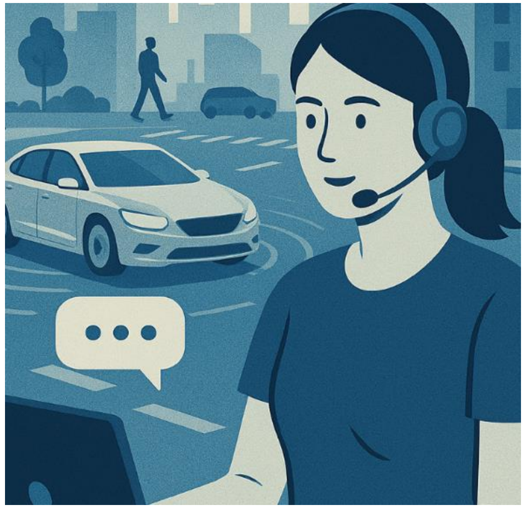
Mouth Open = -0.75

CausalVAE: Disentangled representation learning via neural structural causal models. CVPR 2021

The Challenge of Scaling in Open-ended Environments

Open-ended world

Infinite state, action, multi-agent dynamic - exploding of strategy



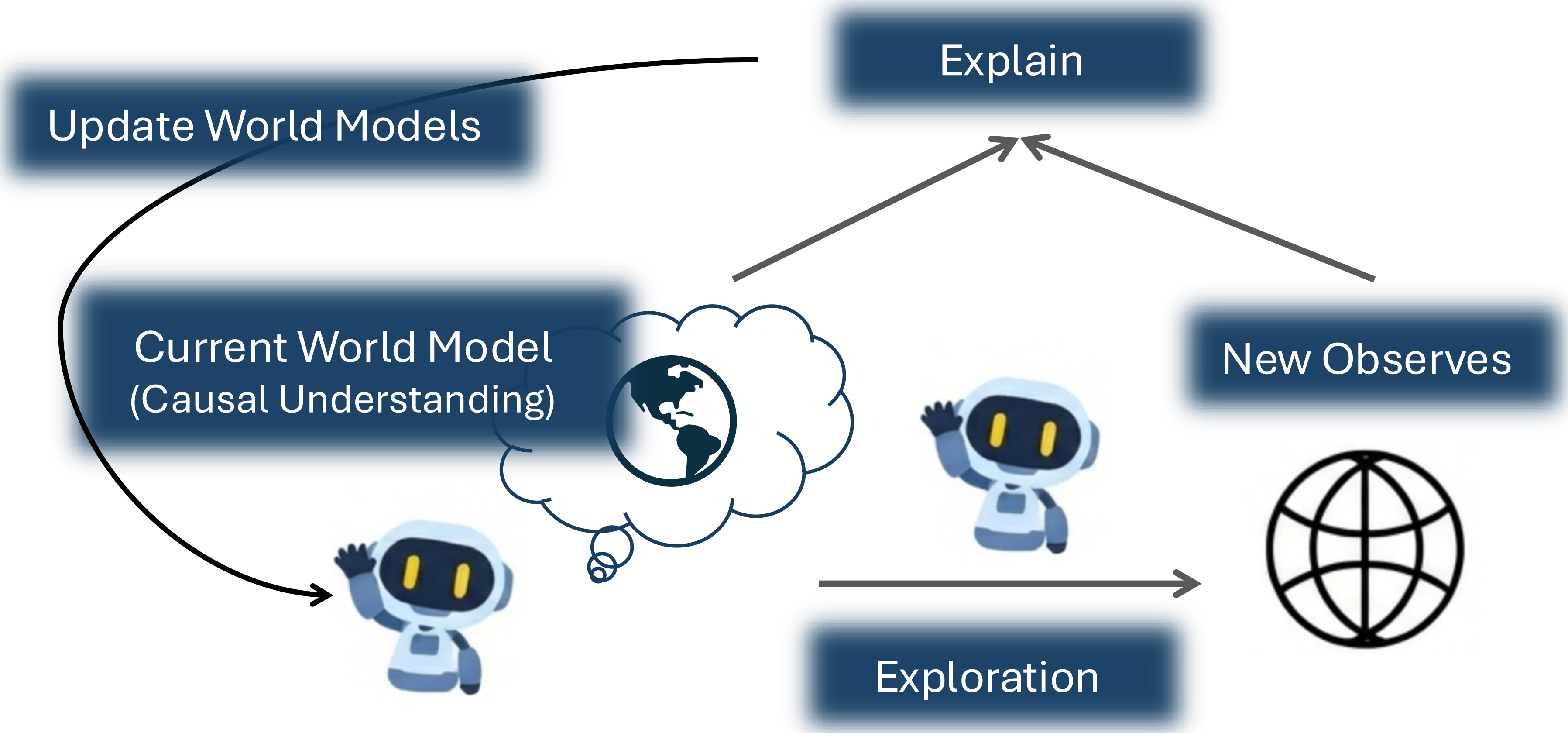
Chat Agent

Automatic driving in real-world

Football AI

Partial observation, causality changing through the observation window....

Continual Causal Learning in Open-Ended Worlds



Causal Discovery in Open-Ended World

[NeurIPS 2025]

Curious Causality-Seeking Agents in Open-Ended World





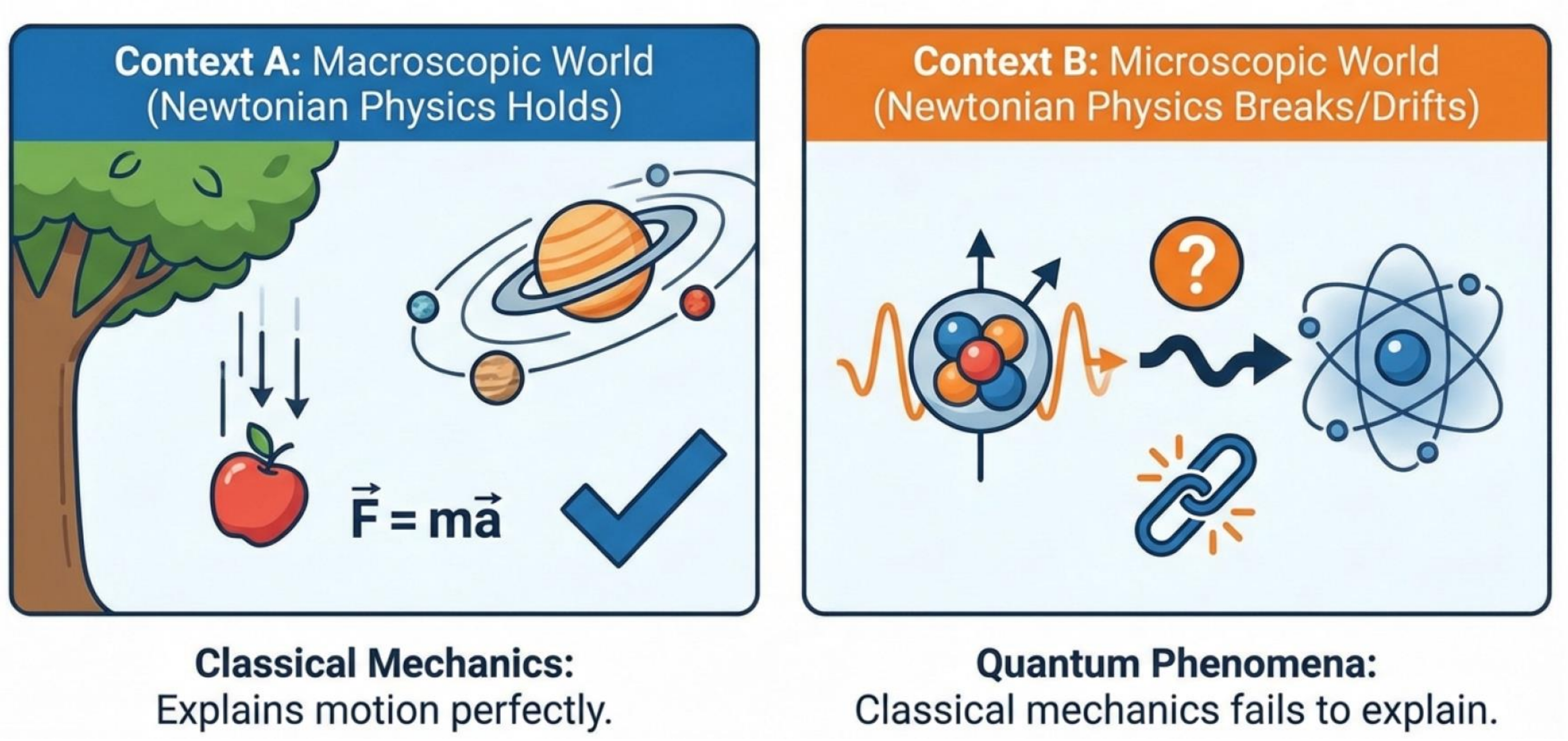


Zhiyu Zhao, Haoxuan Li, Haifeng Zhang, Jun Wang, Francesco Faccio, Jürgen Schmidhuber, Mengyue Yang

Motivation: Causal "Drift" in Open Worlds

Causal Discovery in Open-Ended World

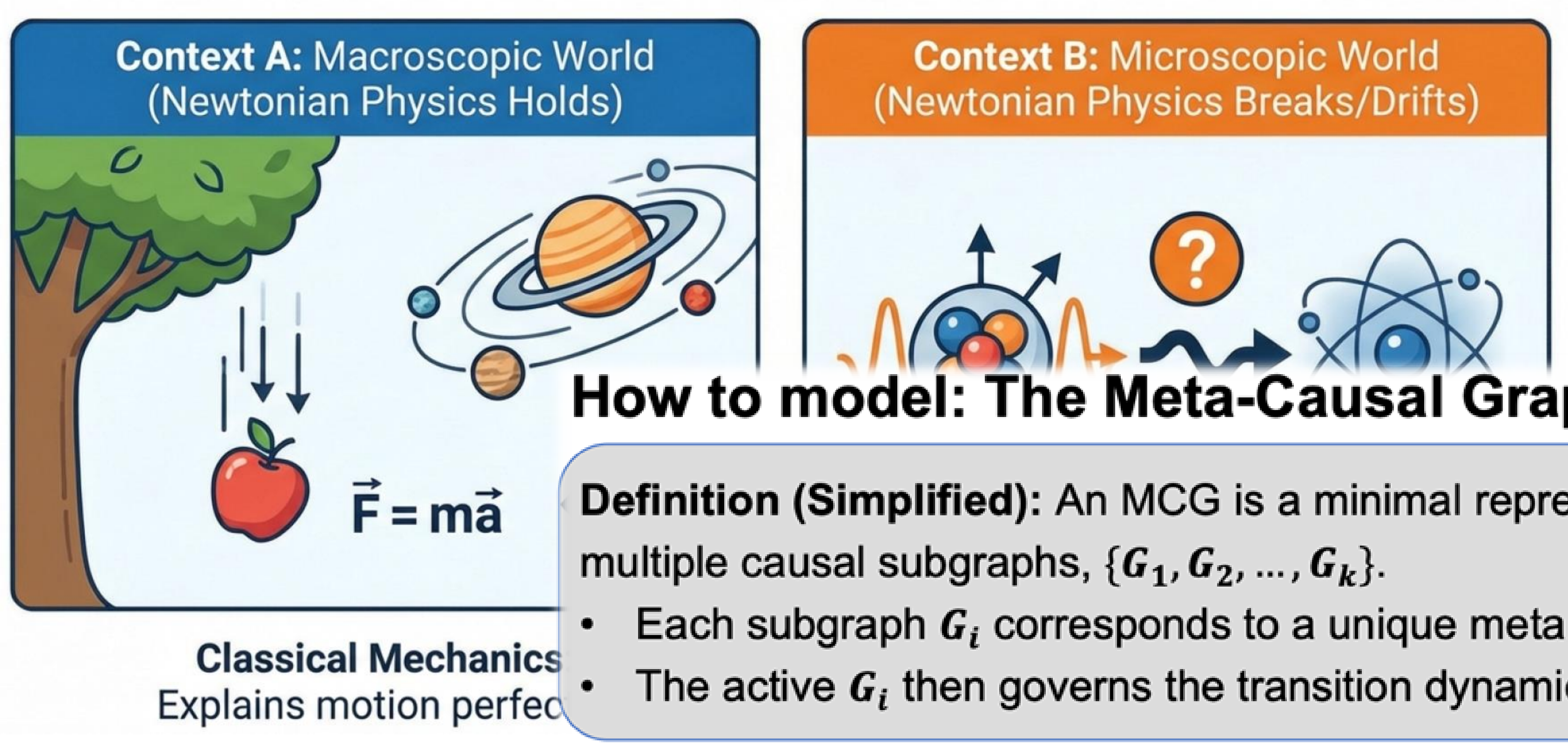
Causal “Drift” in Open Worlds



Curious Causality-Seeking Agents in Open-Ended World. NeurIPS 2025

Causal Discovery in Open-Ended World

Causal “Drift” under Different Condition



How to model: The Meta-Causal Graph (MCG)

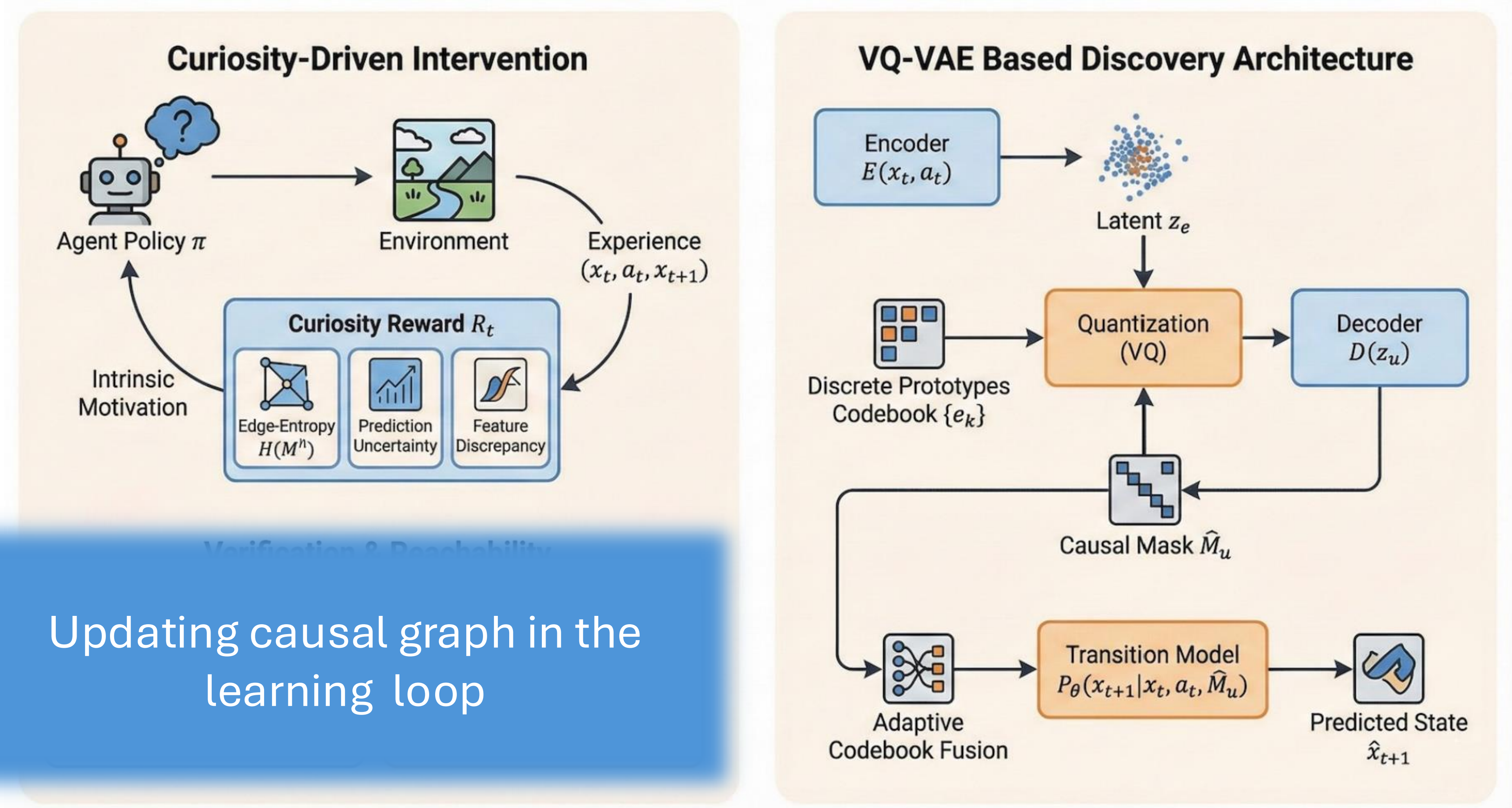
Definition (Simplified): An MCG is a minimal representation that contains multiple causal subgraphs, $\{\boldsymbol{G_1}, \boldsymbol{G_2}, \dots, \boldsymbol{G_k}\}$.

- Each subgraph $\boldsymbol{G_i}$ corresponds to a unique meta state $\boldsymbol{m_i}$.
- The active $\boldsymbol{G_i}$ then governs the transition dynamics: $P(s_{t+1} | s_t, a_t)$.

Curious Causality-Seeking Agents in Open-Ended World. NeurIPS 2025

Learning in Open-Ended World

Causal Understanding =
Statistical Learning + Active Exploration

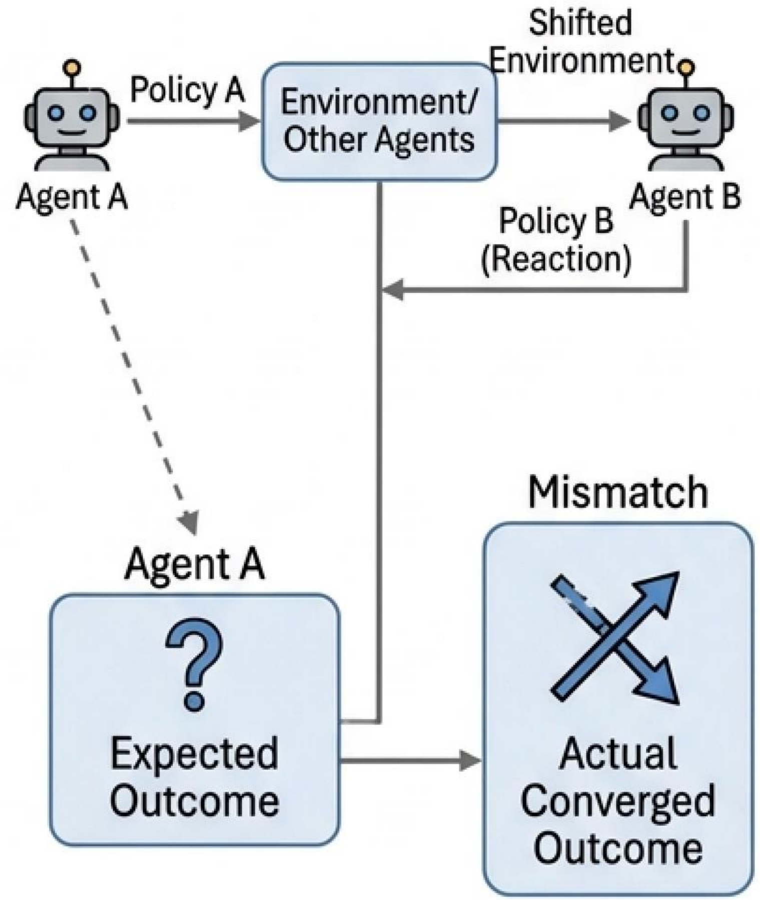


Updating causal graph in the learning loop

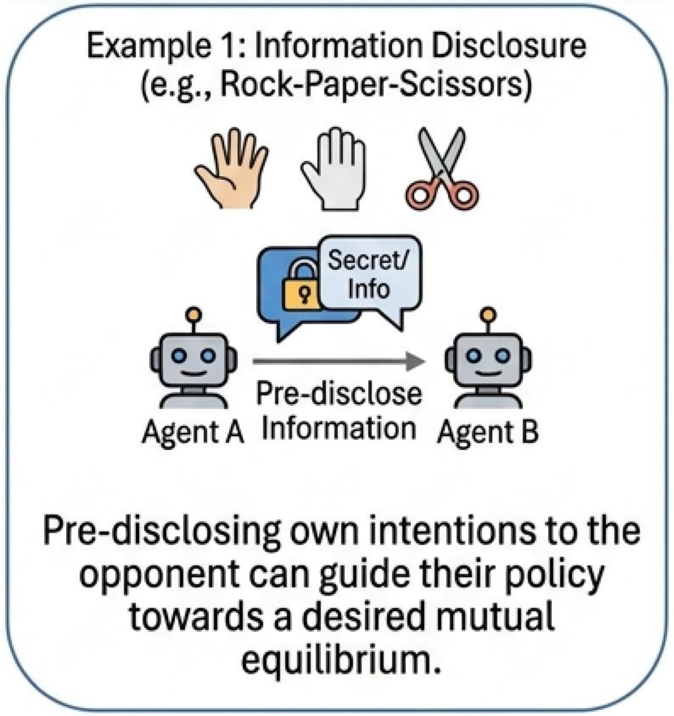
Curious Causality-Seeking Agents in Open-Ended World. NeurIPS 2025

What If the World can be Changed by Policy

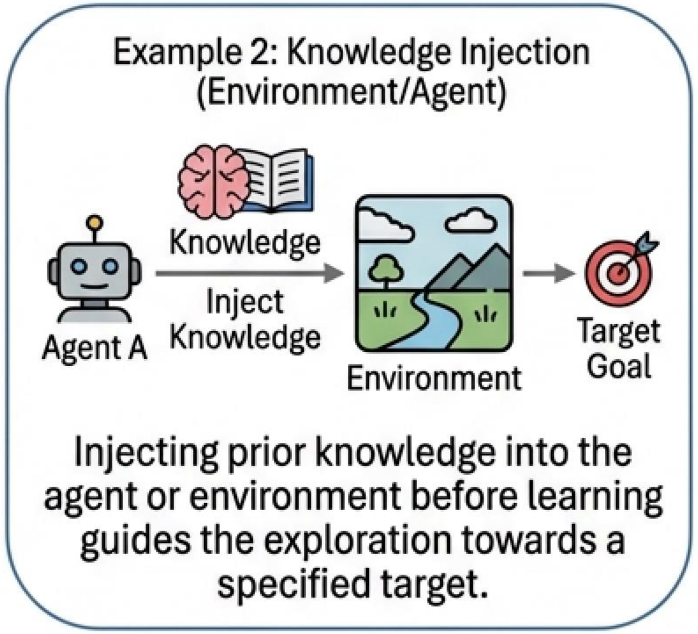
The Challenge:
Policy Interaction & Divergence



Our Explorations:
Pre-policy Intervention for Convergence



Leverage knowledge from foundation models



Attaining Human's Desirable Outcomes in Human-AI Interaction via Structural Causal Games. ICML 2024 workshop
A Principle of Targeted Intervention for Multi-Agent Reinforcement Learning. NeurIPS 2025
Efficient Reinforcement Learning with Large Language Model Priors. ICLR 2025

Future Outlook: Towards Generalizable Causal Representations

Current: Task-Specific Causal Learning
(RL Exploration)

Football AI

Chat Agent/ Driving

Limited transferability. Re-learns basic physics from scratch for each task.

Goal: Learn Universal Causal Primitives

Future: Generalizable Causal Foundation

Generic Warehouse Robot

Household Kitchen

Sci-Fi Planet Rover

- Learn invariant mechanisms across domains
- Enable zero-shot generalization to new environments.

Moving from learning **'how to play football'** to learning **'how objects interact physically'**.

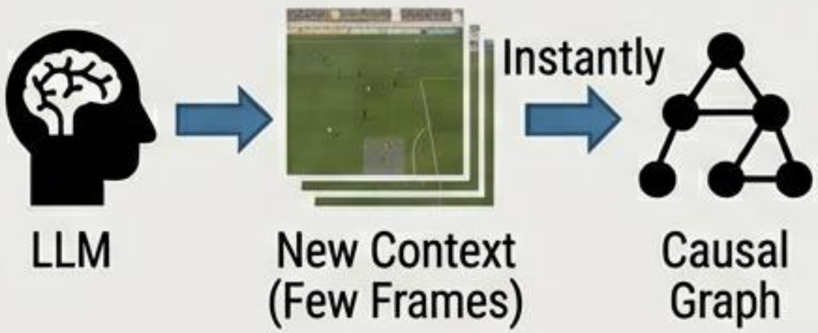
Future Outlook: Causal Foundation World Models & Scaling

Massive Multimodal Data
(Video, Text, Action)

Causal Foundation Model
(Pre-trained Transformer
with Causal Mechanisms)

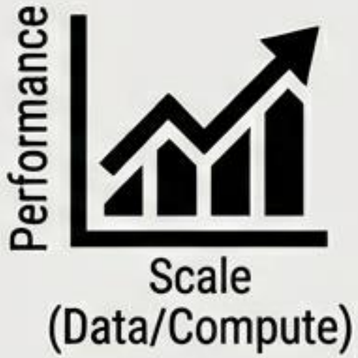
- Generalized World Simulation
- Zero-Shot Planning
- Counterfactual Reasoning

In-Context Causal Learning



Leverage LLM reasoning for fast, in-context causal discovery

The Scaling Hypothesis



Causal understanding emerges and robustifies with large-scale pre-training.

Final Goal: A unified, scalable world model with robust causal reasoning capabilities.