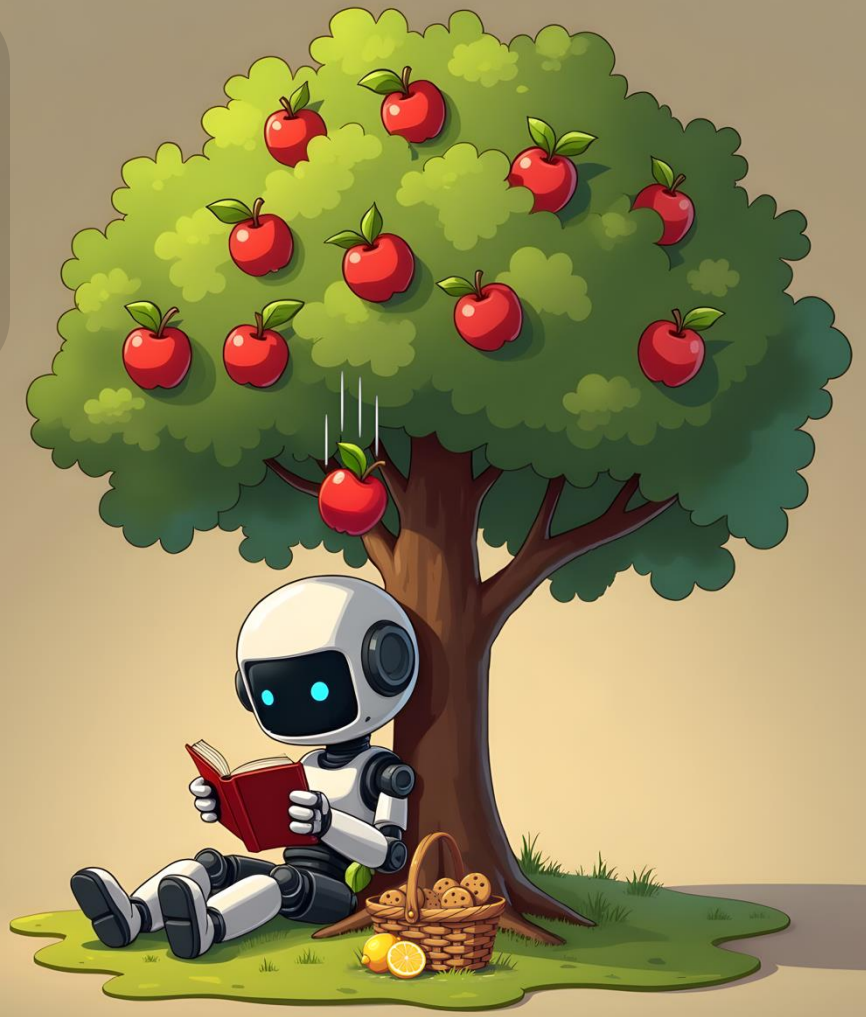
Causal Representation Learning across Multiple Environments

Phillip Lippe

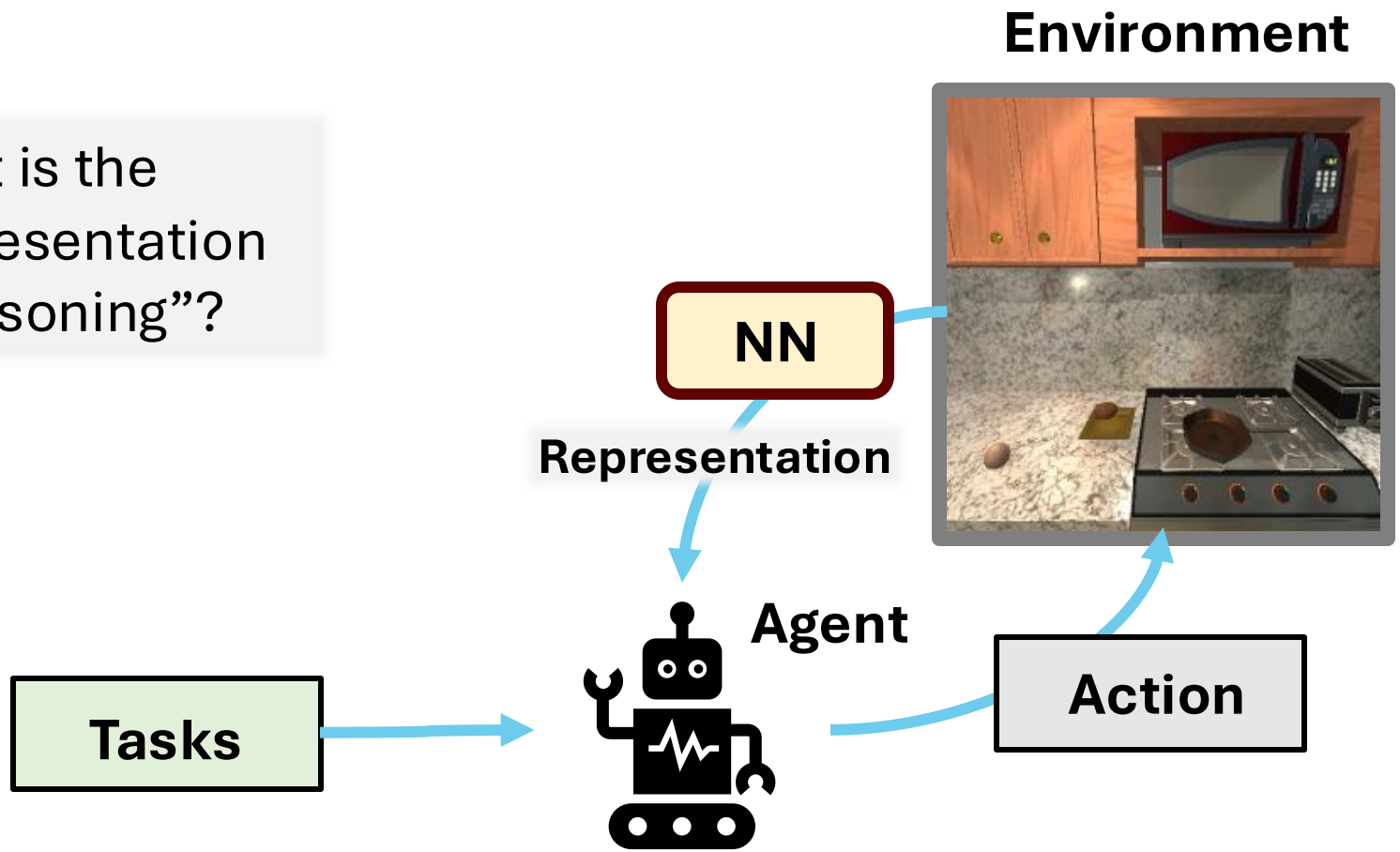
Causality Meetup

04. Dec 2024



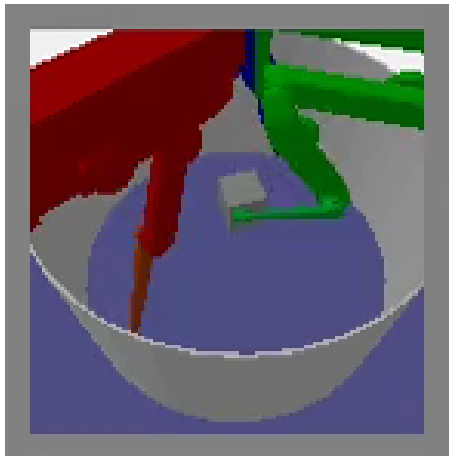
Problem Setup

What is the best representation for “reasoning”?



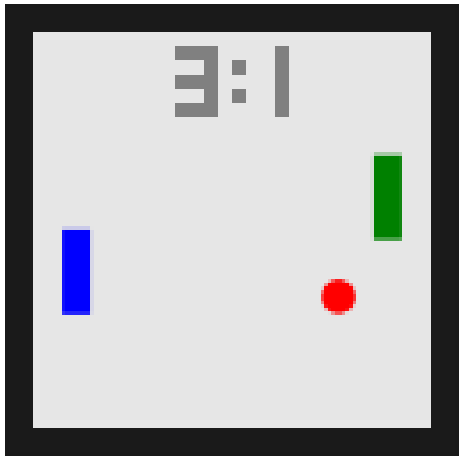
Causal Representation Learning

CausalWorld

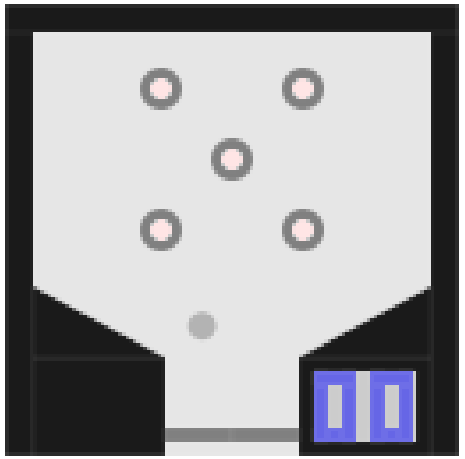


iTHOR

Pong

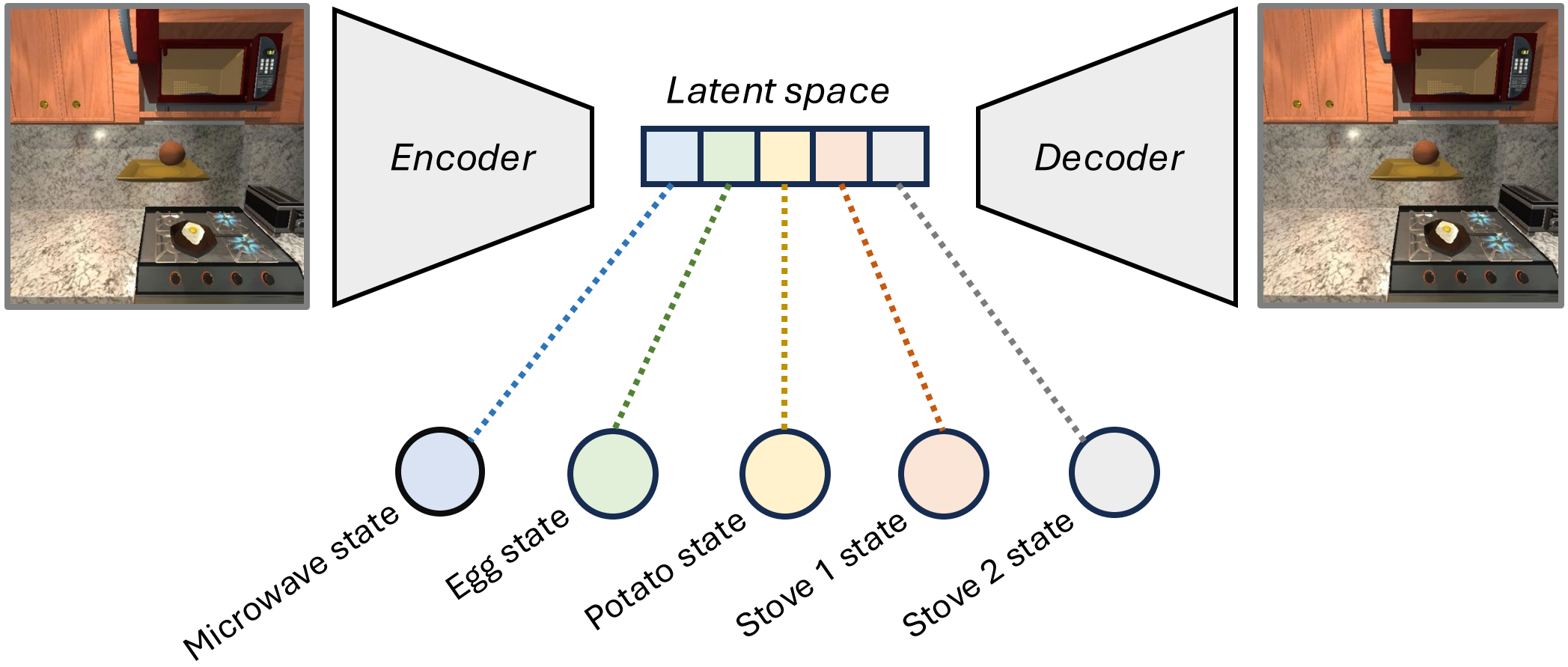


Pinball

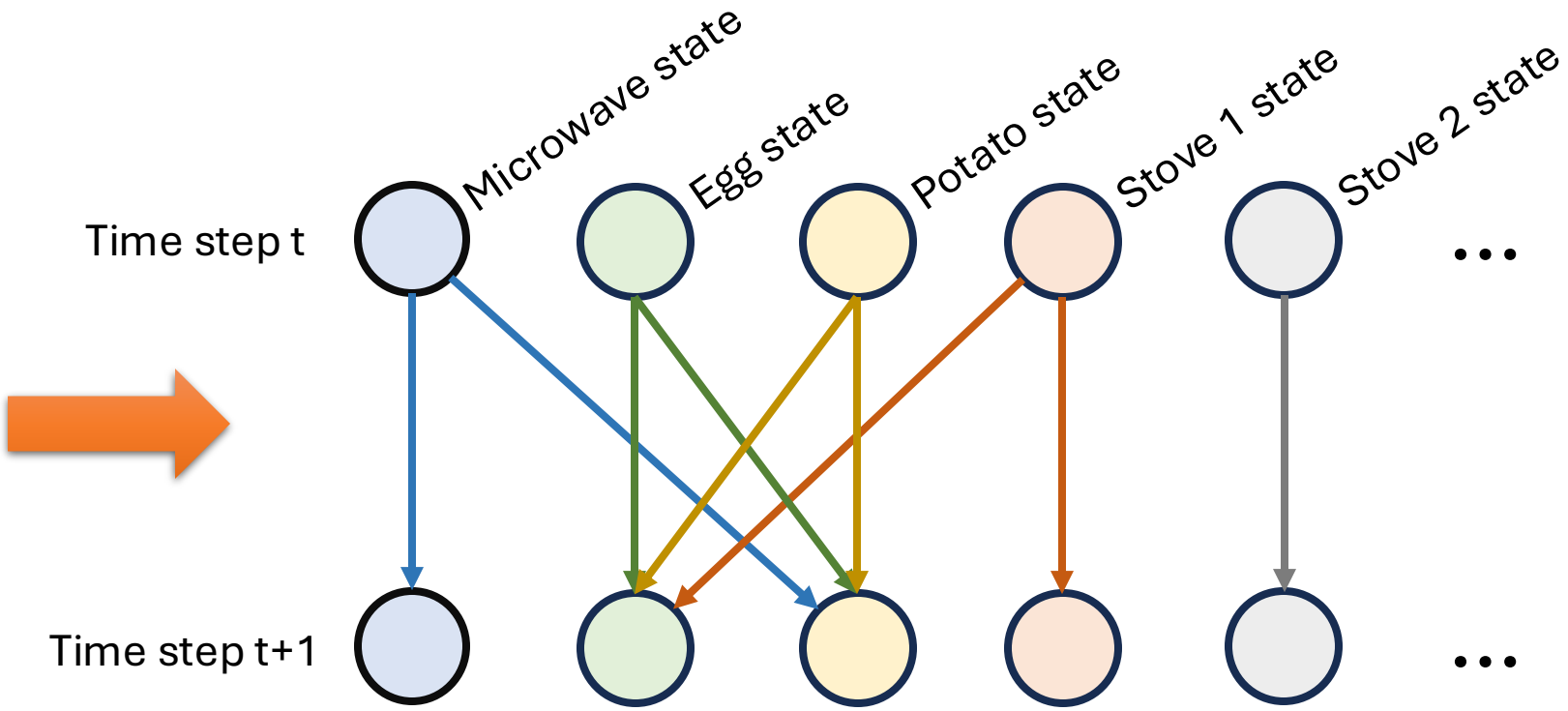


Causal Representation Learning
Identify the (1) Causal Variables and (2) Causal Structure from high-dimensional observations (e.g. videos).

Causal Representation Learning

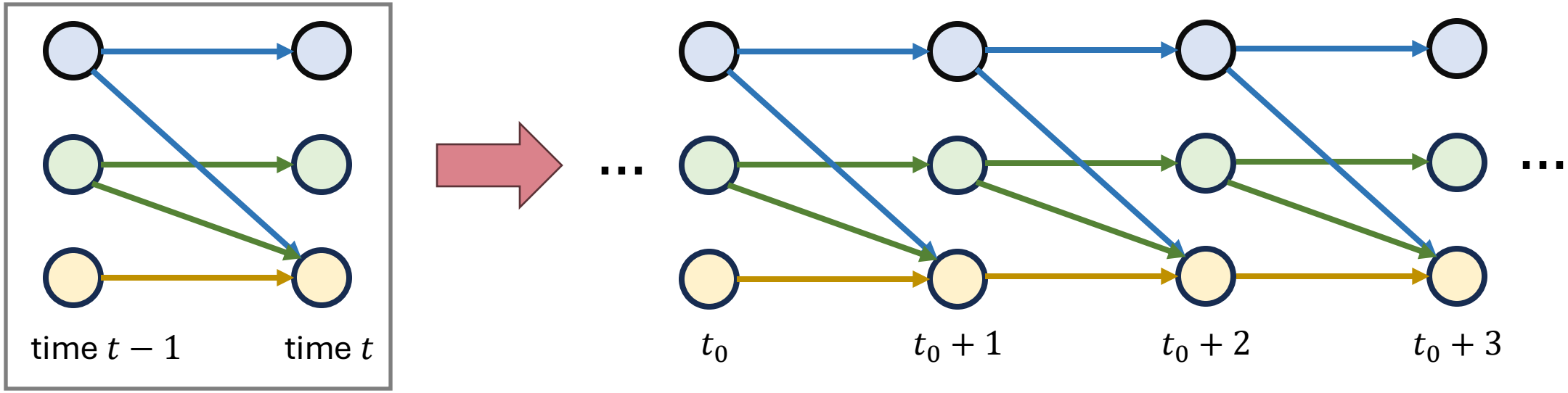


Temporal Causal Representation Learning



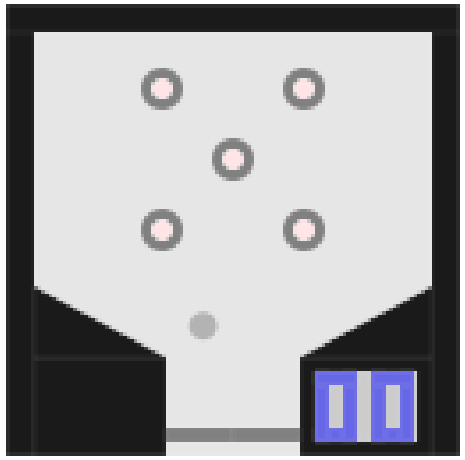
Temporal Causal Representation Learning

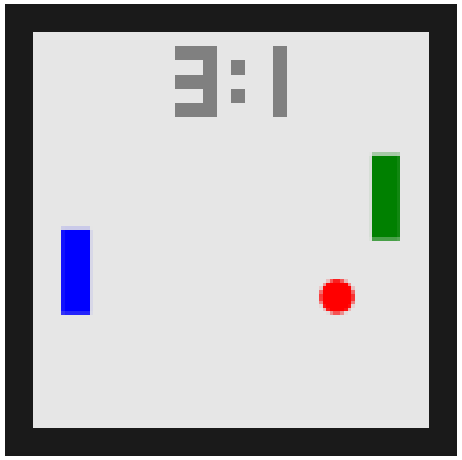
- Dynamic Bayesian Network
- Standard assumptions
  - ***N*-Markov**: only variables from the last N time steps can cause variables at time t
  - **Stationary/Time Invariance**: transition model stays the same across time steps

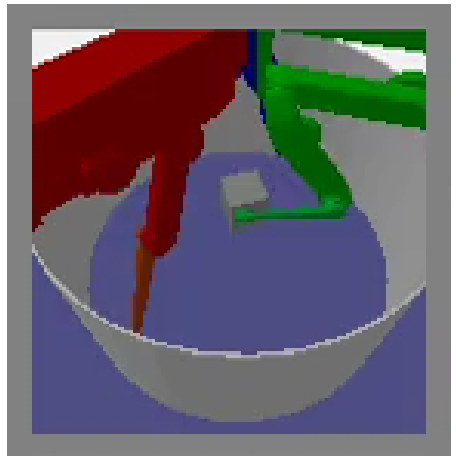


Temporal Causal Representation Learning

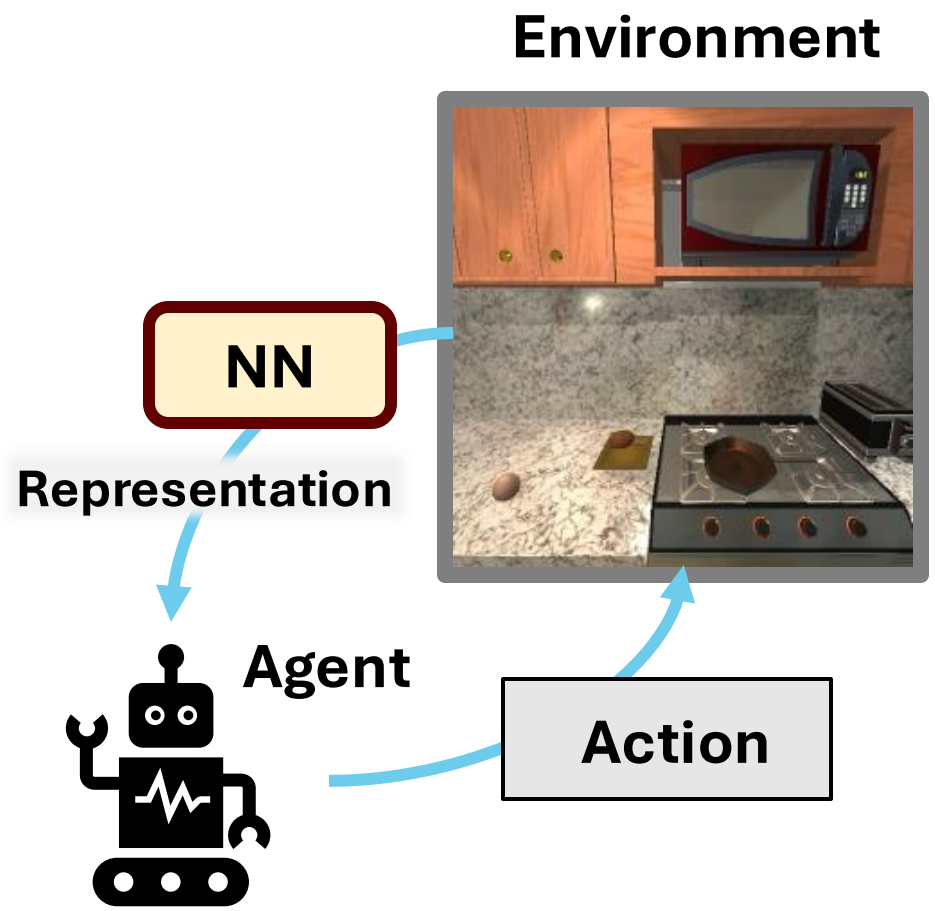
- All causal variables evolve over time and may differ between two time steps







Temporal Causal Representation Learning



Representation Learning Tasks

What are the causal variables of the environment?

How do they interact with each other?

How can external systems (e.g. an agent) intervene on causal variables?

CRL – Constraints

Observation function $f(Z) = X$ Observation (Images, Video, Text, ...)

Latent space

Observation Function

(piecewise) linear, polyn.

Kivva et al., 2022, NeurIPS
Buchholz et al., 2023, NeurIPS
Squires et al., 2023, ICML
Ahuja et al., 2023, ICML

Latent Distribution

non-Gaussian, known, ...

Khemakhem et al., 2020, STATS
Kügelgen et al., 2021, NeurIPS
Ahuja et al., 2022, ICLR
Kivva et al., 2022, NeurIPS

Interventions

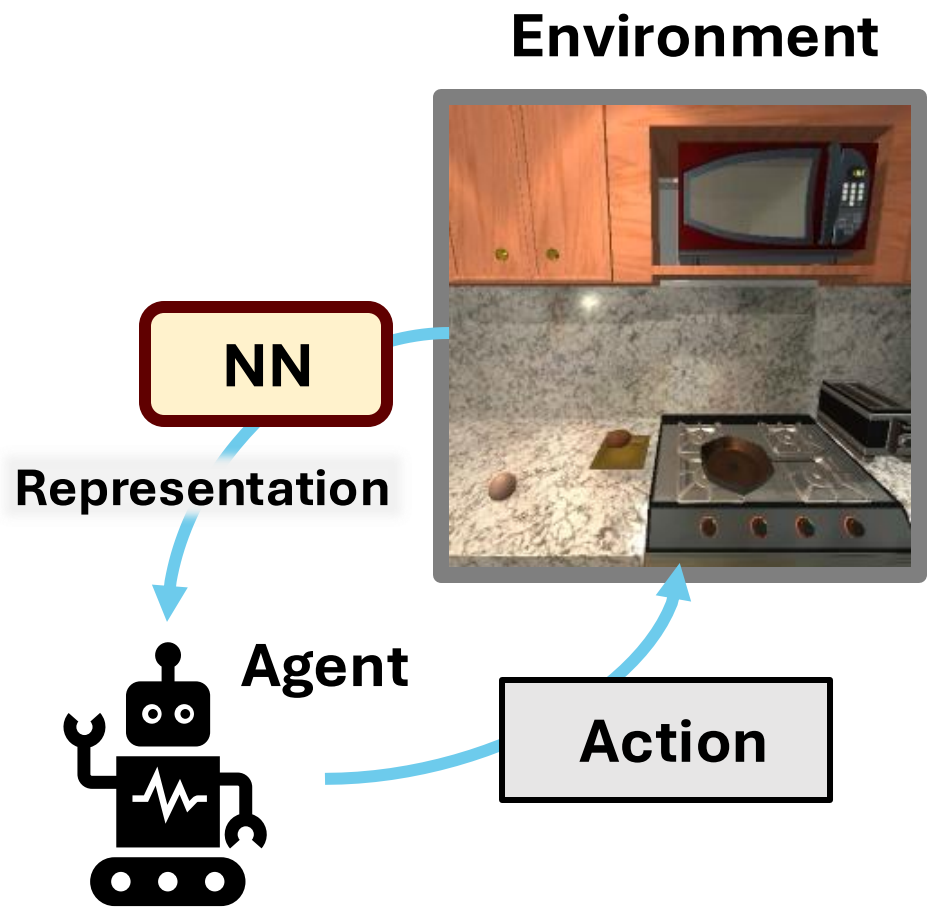
distribution shifts

***Lippe** et al., 2022, ICML*
*Brehmer **et al.**, 2022, NeurIPS*
***Lippe** et al., 2023, ICLR*
***Lippe** et al., 2023, UAI*
Kügelgen et al., 2023, NeurIPS

Interactive Environments

- Interventions naturally happen in interactive environments like in Reinforcement Learning
  - Agent performs actions on underlying system
  - Changes dynamics of causal variables
  - Effect and target of intervention unknown

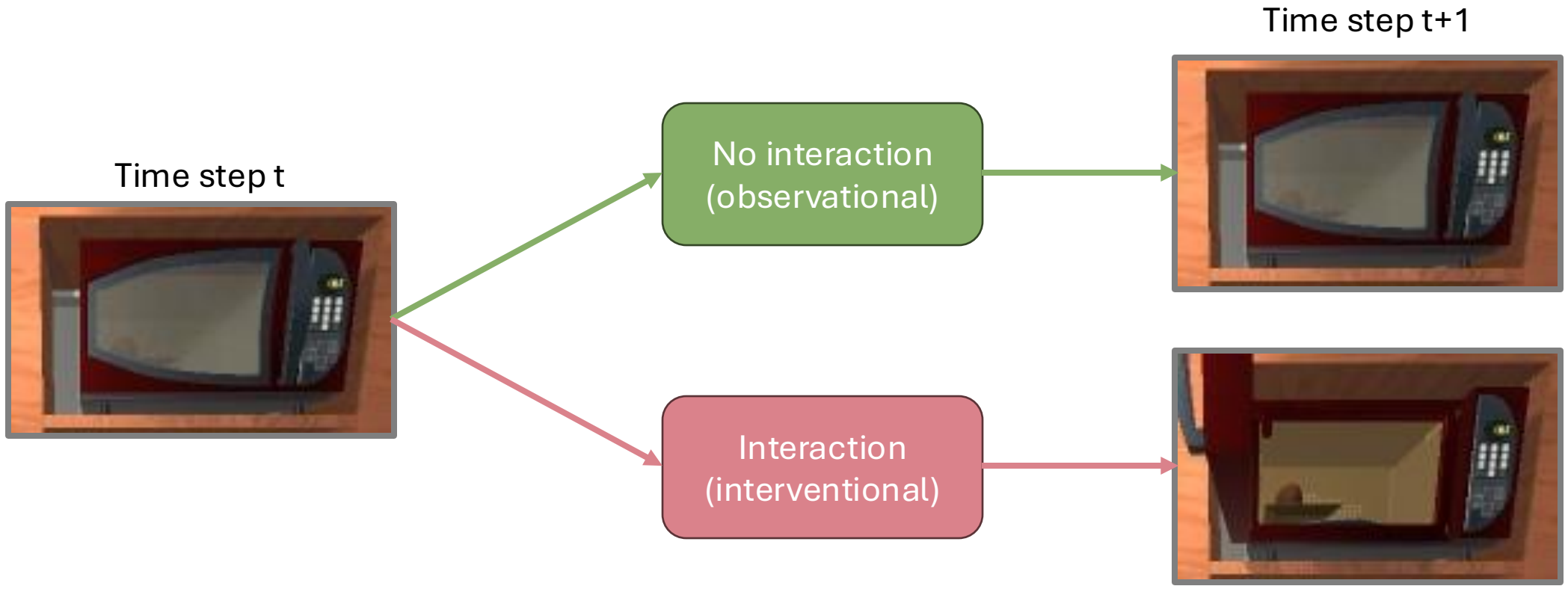
- Can we use low-level actions to identify causal variables?



BISCUIT: Binary Interactions

Lippe et al. “BISCUIT: Causal Representation Learning from Binary Interactions.” UAI, 2023.

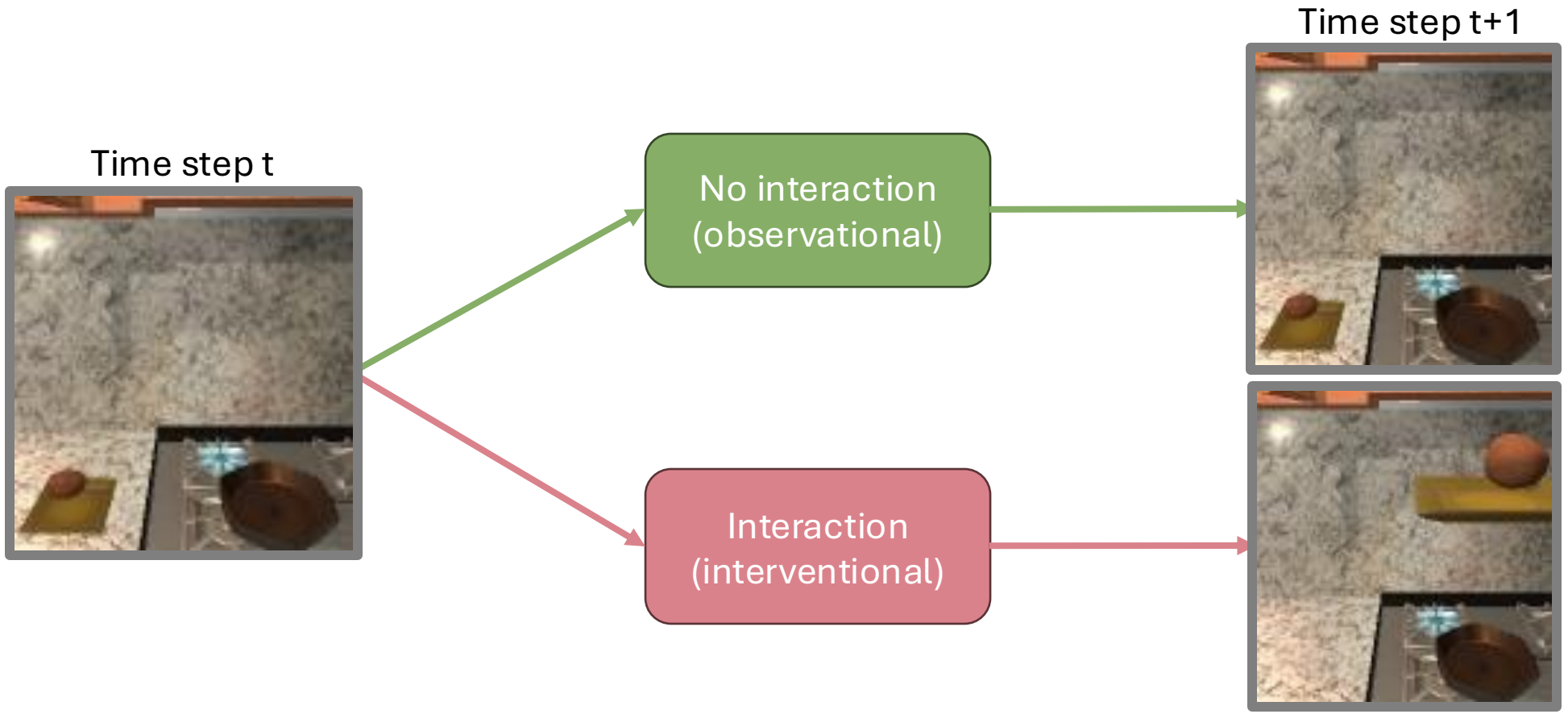
Key assumption: Interactions between the agent and causal variables can be described by **binary variables**



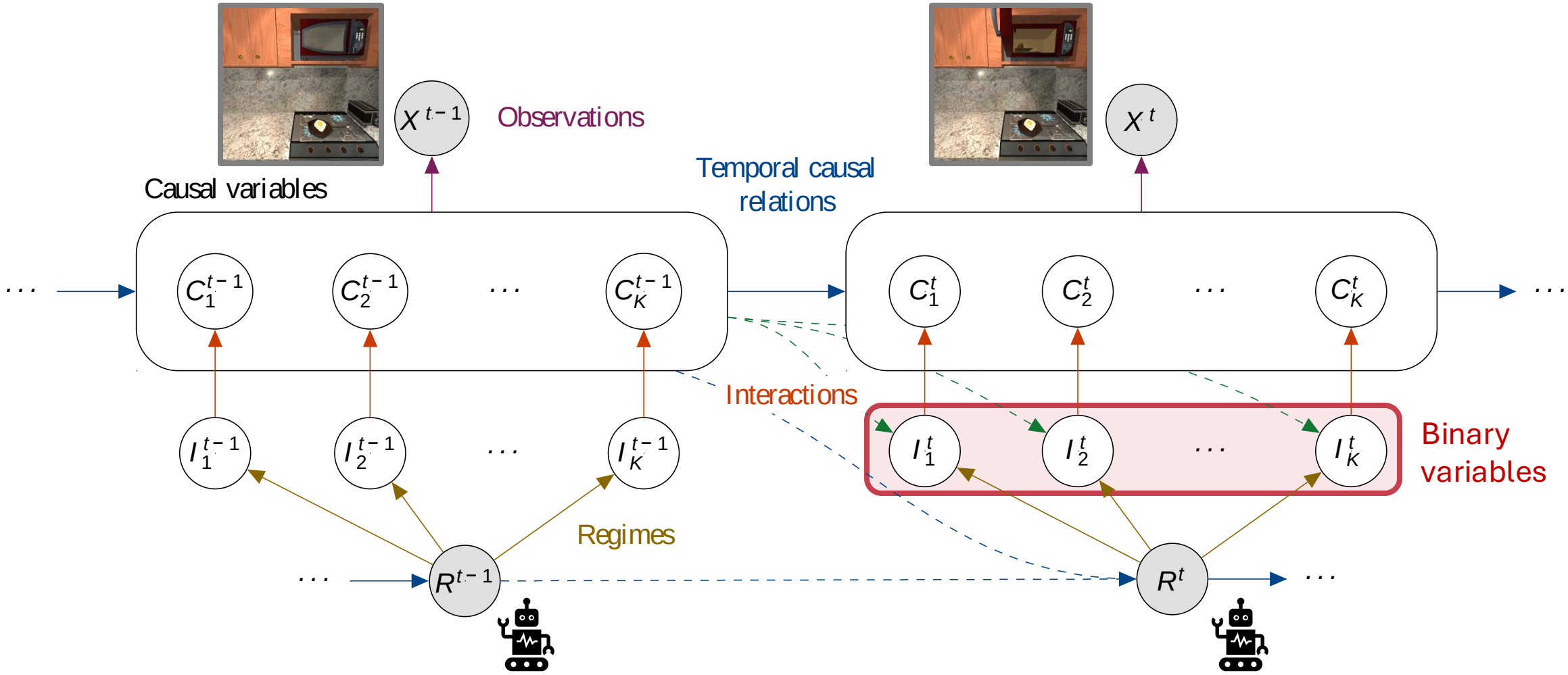
BISCUIT: Binary Interactions

Lippe et al. “BISCUIT: Causal Representation Learning from Binary Interactions.” UAI, 2023.

Key assumption: Interactions between the agent and causal variables can be described by **binary variables**

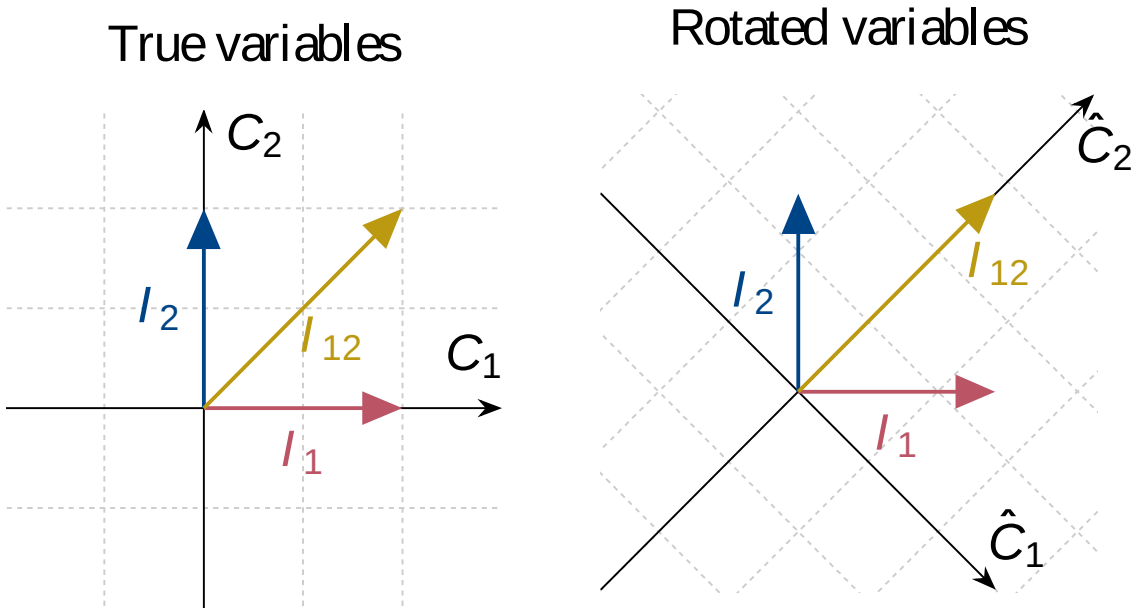


BISCUIT: Causal Model



Binary Interactions enable Identifiability

- Knowing each variable has only two mechanisms helps identify difficult cases
- Example: Additive Gaussian Noise – $C_i^t = \mu_i\left(C^{t-1}, I_i^t\right) + \epsilon_i,\ \epsilon_i \sim \mathcal{N}(0, \sigma^2)$
  - Both true and rotated variables model the same distribution, but under interventions, only the true variables have two means

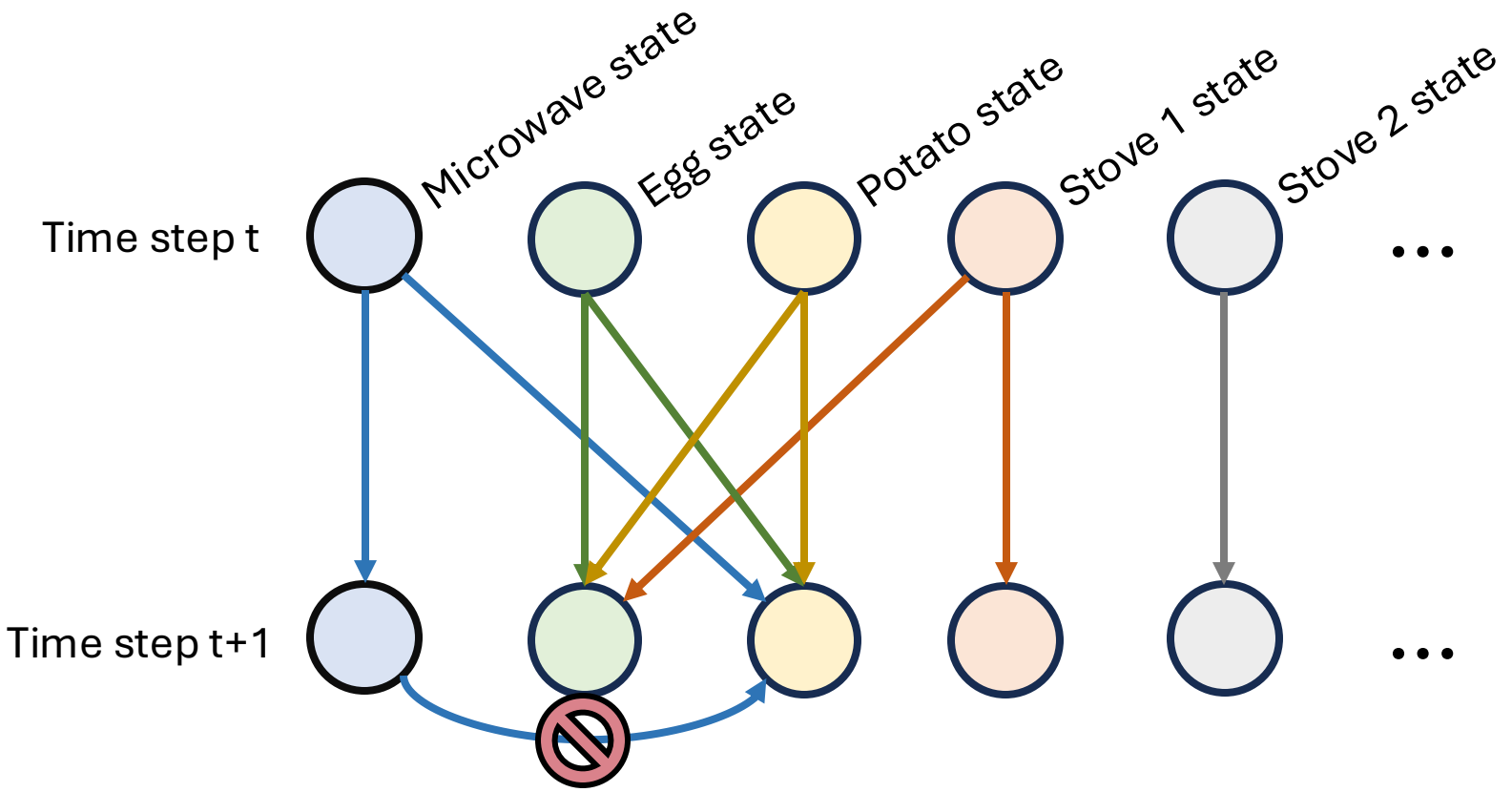


Identifiability Assumptions

- **Assumption 2**: interaction variables of different causal variables are not deterministic functions of each other
  - Implies that two variables are not always interacted with at the same time
  - Distinct interaction patterns
- If the interaction variables I_i^t are independent of C^{t-1}, only requires $\lfloor \log_2 K \rfloor + 2$ actions/values of R^t
  - Example: agent with random policy

Identifiability Assumptions

- **Assumption 3**: Causal Relations can be resolved over time



BISCUIT: Identifiability Results

Assumption 1: Interactions between agent and causal variables can be described by **binary variables**

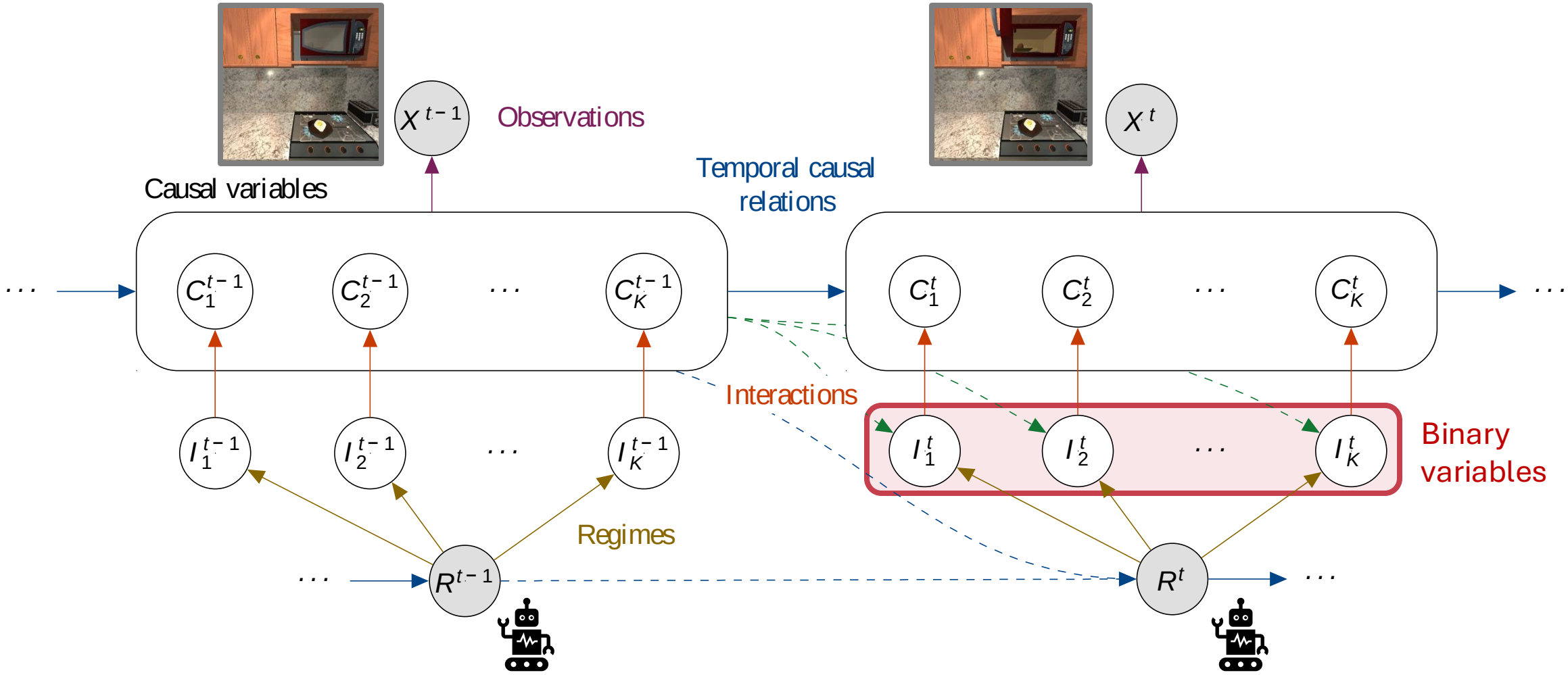
Assumption 2: All causal variables have different interaction patterns

Assumption 3: Causal Relations can be resolved over time

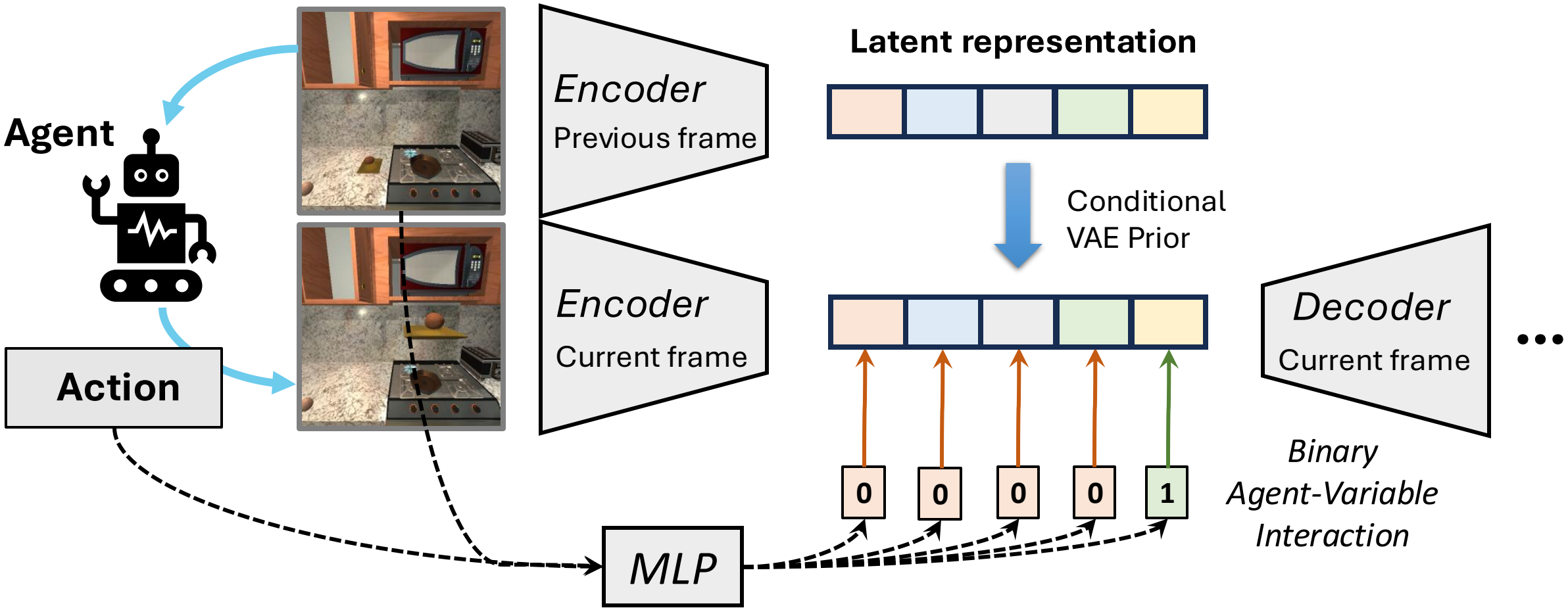
Assumption 4: The causal mechanisms vary sufficiently over time or on interactions

Identifiability Result
The causal variables can be identified up to permutation and element-wise transformations.

BISCUIT: Causal Model (Reminder)



BISCUIT: Architecture



BISCUIT: Architecture

- Loss function:

$$\mathcal{L}_t = \underbrace{-\mathbb{E}_{q_\phi(z^t|x^t)}[\log p_\theta(x^t|z^t)]}_{\text{Reconstruction}} + \underbrace{\mathbb{E}_{q_\phi(z^{t-1}|x^{t-1})}\left[KL\left(q_\phi(z^t|x^t)||p_\omega(z^t|z^{t-1},R^t)\right)\right]}_{\text{Prior modeling}}$$

Encoder Decoder Prior

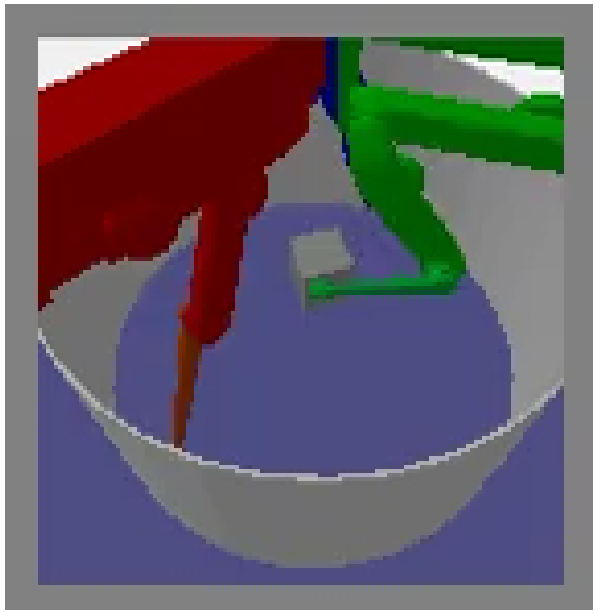
- Prior structure:

$$p_\omega(z^t|z^{t-1},R^t) = \prod_i p_\omega\left(z_i^t|z^{t-1}, f_i(R^t,z^{t-1})\right)$$

Binary function output

CausalWorld – Robotic Trifinger

- Tri-finger robot interacting with its environment and objects
  - Causal variables include object position, frictions, colors, etc.
- Action: 9-dimensional motor angles (3 per finger)
- BISCUIT identifies causal variables accurately



Accuracy of learned causal variables
(higher is better / lower is better)

Models	CausalWorld
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00
LEAP (Yao et al., 2022b)	0.30 / 0.00
DMS (Lachapelle et al., 2022b)	0.32 / 0.00
BISCUIT-NF (Ours)	**0.97** / 0.01

iTHOR

- Kitchen environment with 10 causal variables
  - Cabinet (open/closed)
  - Microwave (open/closed)
  - Microwave (on/off)
  - Egg (position, broken, cooked)
  - Plate/potato (position)
  - 4x Stove burner (on/off, burning)
  - Toaster (on/off)
- Actions represented as x-y coordinate of a randomly sampled object pixel

Models	iTHOR
iVAE (Khemakhem et al., 2020a)	0.48 / 0.35
LEAP (Yao et al., 2022b)	0.63 / 0.45
DMS (Lachapelle et al., 2022b)	0.61 / 0.40
BISCUIT-NF (Ours)	**0.96 / 0.15**

higher better / lower better

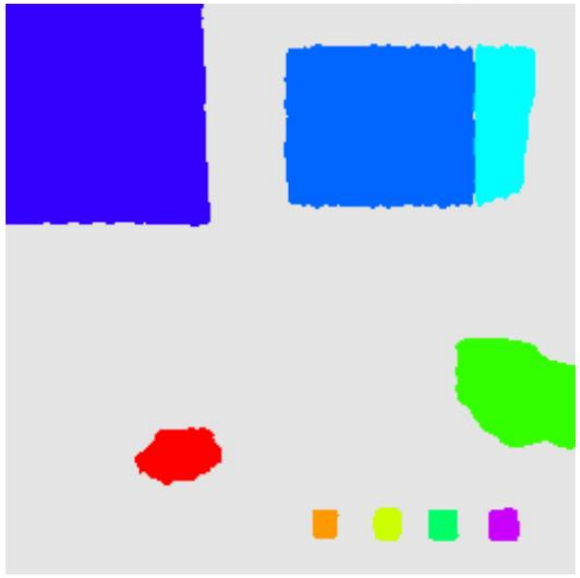
iTHOR – Interaction Maps

- Visualize learned interaction variables by the x-y locations they are active
- Each causal variable shown in different color

Original image

Overlapped image

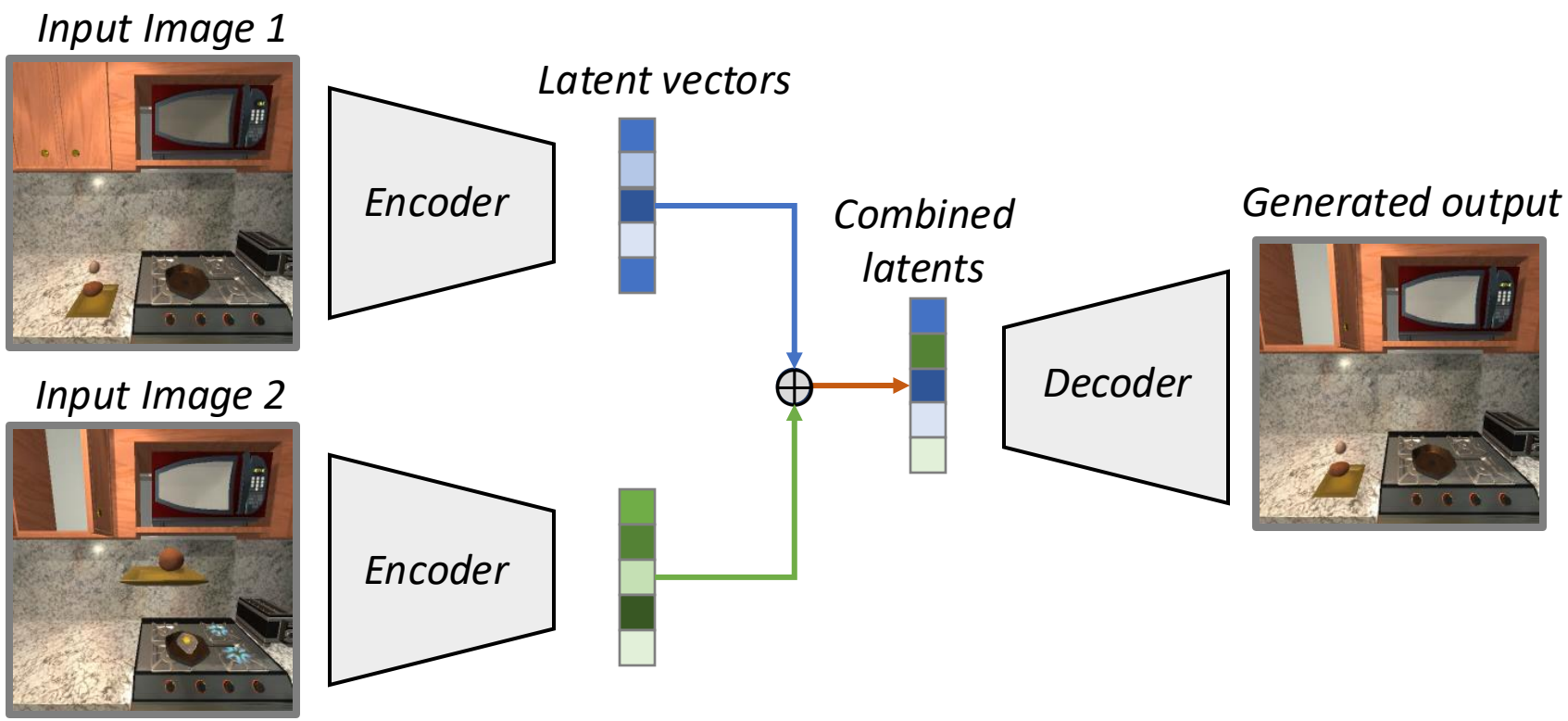
Interaction map



iTHOR – Triplet Evaluation

- Test compositional generation ability of latent space
- Suitable across various identifiability classes

Goal
Open Cabinet
Turn on Microwave
Keep other variables fixed



iTHOR – Triplet Evaluation

Input image 1

Input image 2

Generated Output

Latents from image 2

Microwave Open

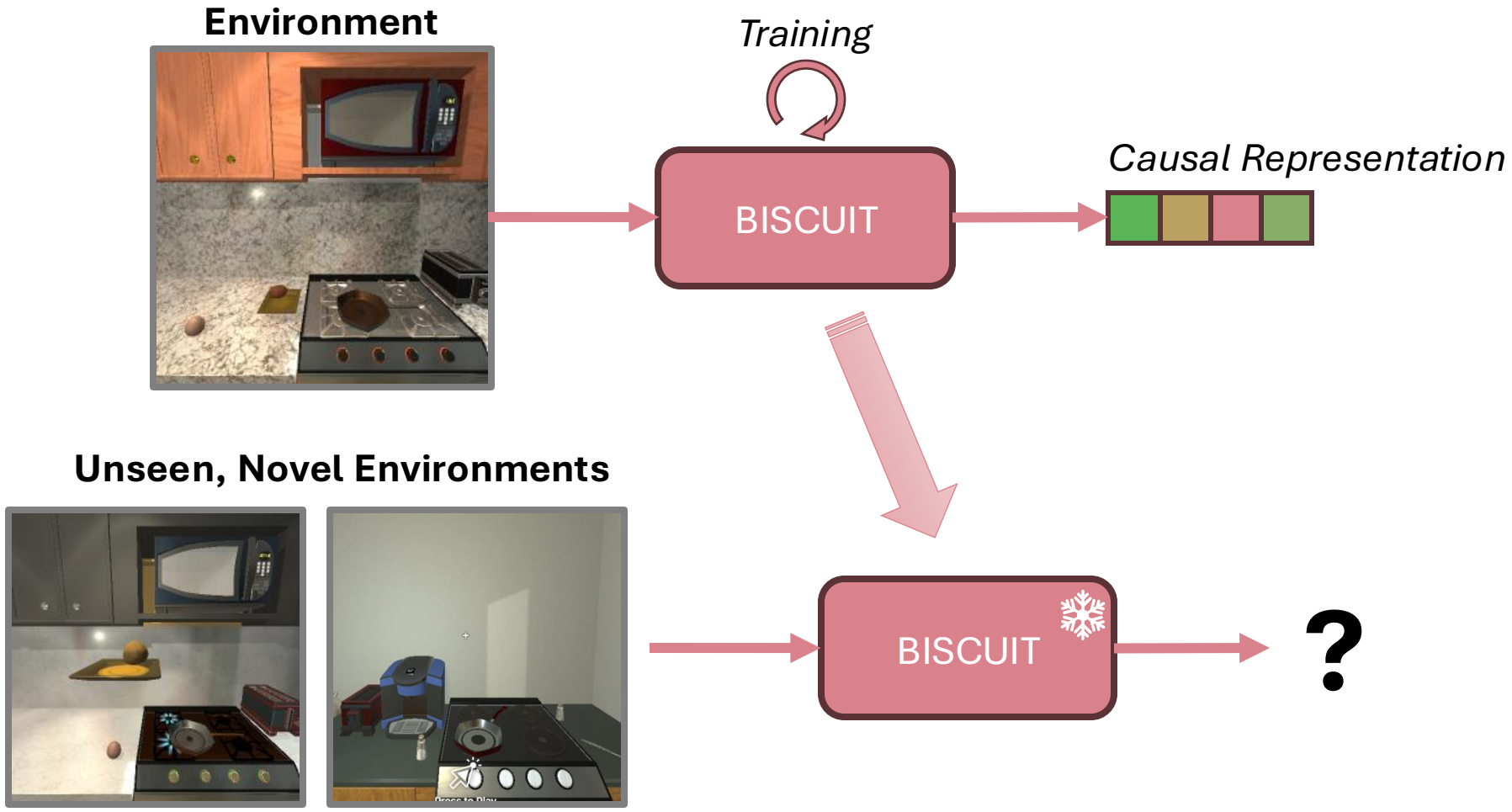
iTHOR – Triplet Evaluation

Input image 1	Input image 2	Generated Output	Latents from image 2
			Stove (front-left)

iTHOR – BISCUIT Demo

Demo: https://colab.research.google.com/github/phlippe/BISCUIT/blob/main/demo.ipynb

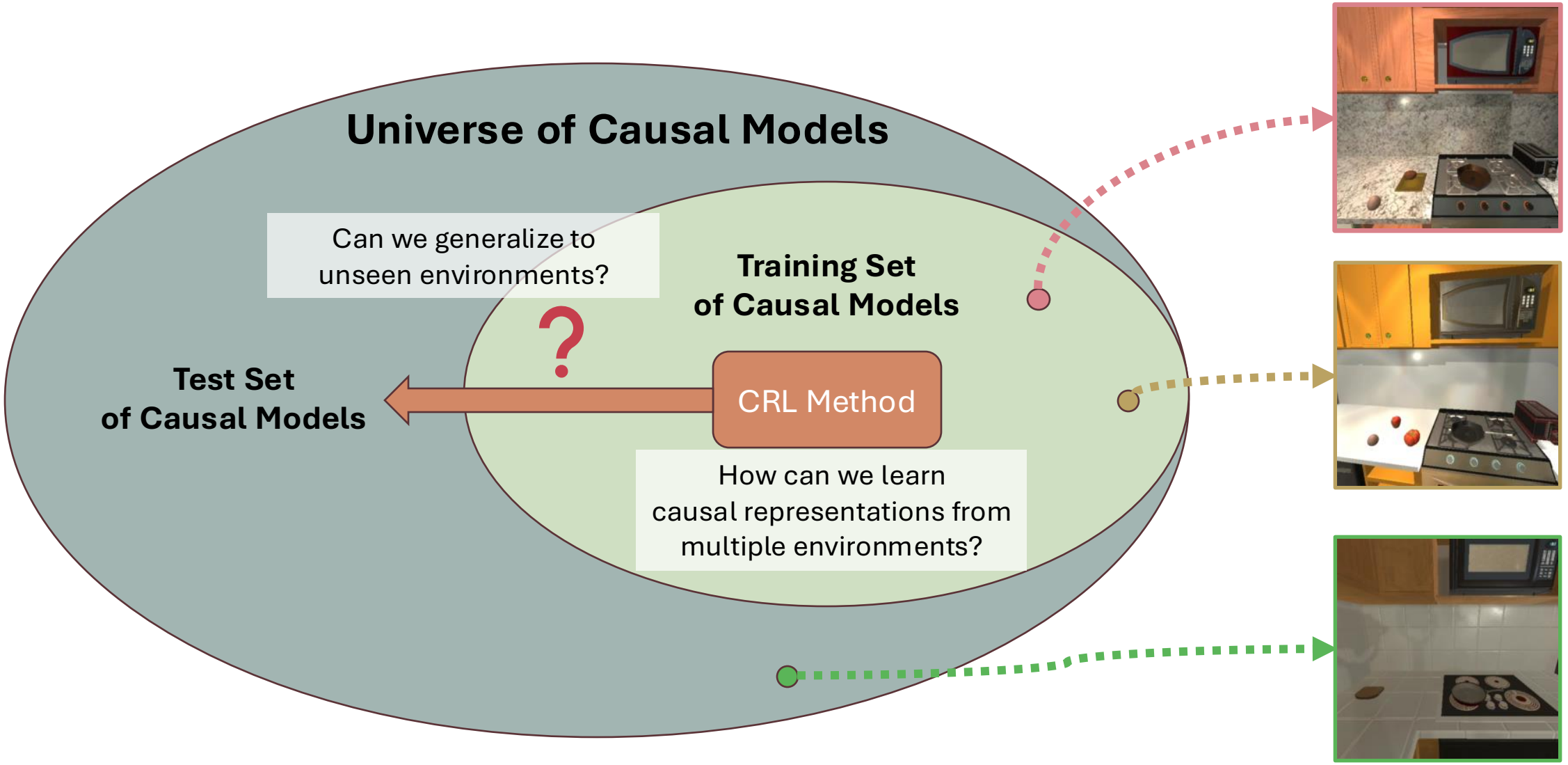
Causal Representation Learning



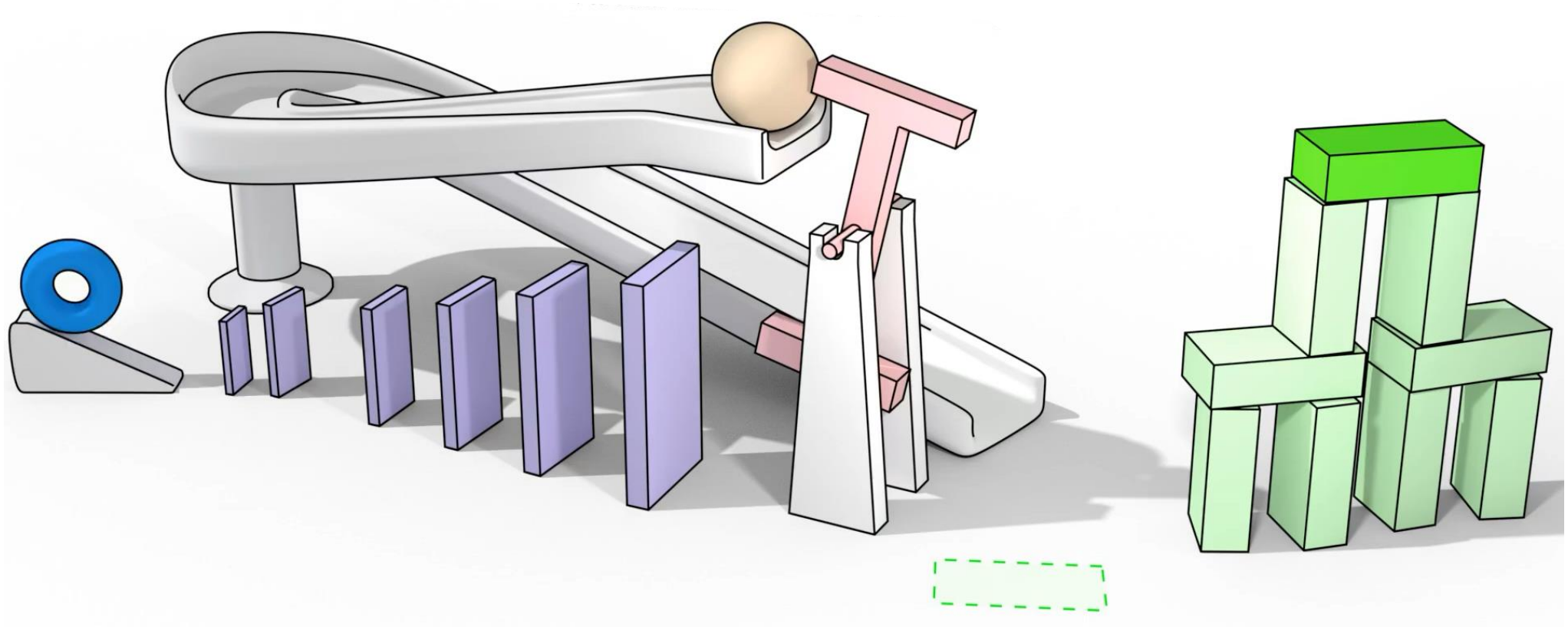
CRL across Multiple Environments

- CRL methods are commonly trained for a single causal model
- New environments require re-training models
- Can we use the generalization ability of ML to have one model to learn them all?

CRL across Multiple Environments



CRL and Object-Centric Representations



Credit: ICLR 2022 OSC Workshop

CRL and Object-Centric Representations

- Objects play a vital role in many causal systems
- Objects often constitute a group of causal variables
  - Attributes
  - Position
  - Orientation
- Identifying the objects first can give a coarse identification of the causal variables
- A lot of strong tools exist for object detection

Work in Progress

Shared current state of WIP.

References

[Lippe et al., 2023b] Lippe P, Magliacane S, Löwe S, Asano YM, Cohen T, Gavves E. BISCUIT: Causal Representation Learning from Binary Interactions. In 39th Conference on Uncertainty in Artificial Intelligence, 2023. Project page https://phlippe.github.io/BISCUIT/.

[Ahuja et al., 2022] Ahuja K, Hartford J, Bengio Y. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In International Conference on Learning Representations 2022.

[Khemakhem et al., 2020] Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the Twenty Third Inter- national Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*. PMLR, 2020.

[Lachapelle et al., 2022] Lachapelle, S., Rodriguez, P., Le, R., Sharma, Y., Everett, K. E., Lacoste, A., and Lacoste-Julien, S. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In First Conference on Causal Learning and Reasoning, 2022.

[Lippe et al., 2022] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In Proceedings of the 39th International Conference on Machine Learning, ICML, 2022.

[Lippe et al., 2023a] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. Causal representation learning for instantaneous and temporal effects in interactive systems. In The Eleventh International Conference on Learning Representations, 2023.

[Yao et al., 2022a] Yao, W., Chen, G., and Zhang, K. Temporally Disentangled Representation Learning. In Advances in Neural Information Proces sing Systems 35, NeurIPS, 2022.

[Yao et al., 2022b] Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. Learning Temporally Causal Latent Processes from General Temporal Data. In International Conference on Learning Representations, 2022.