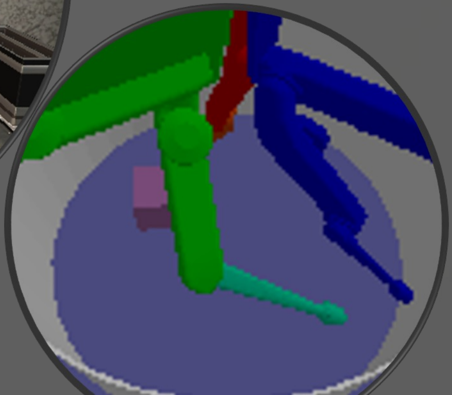
BISCUIT: Causal Representation Learning from Binary Interactions

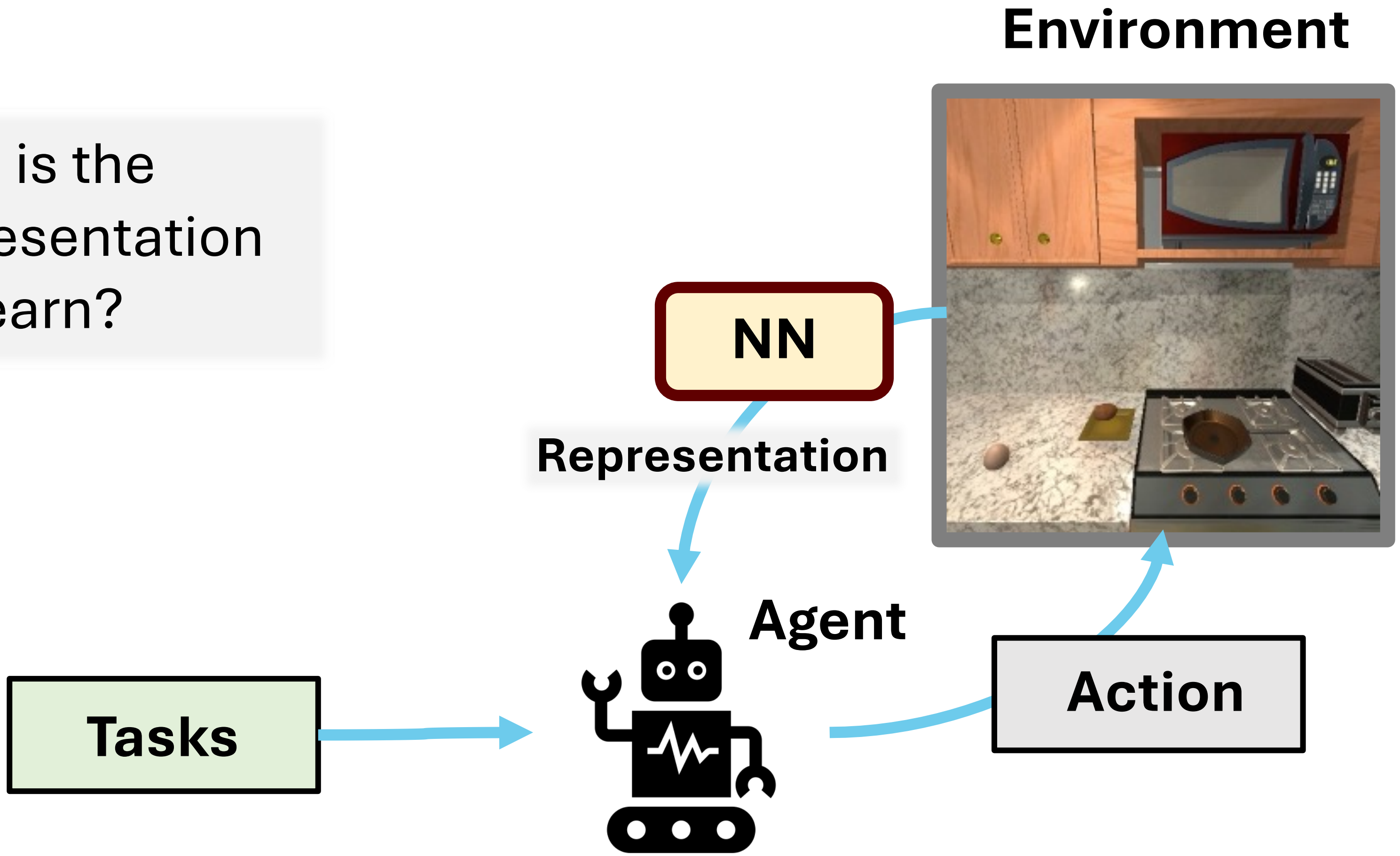


Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves

July 18, 2023

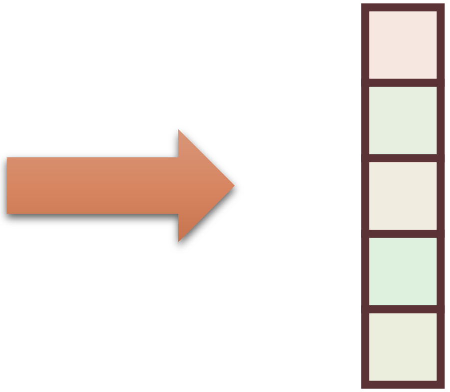
Problem Setup

What is the
best representation
to learn?



Causal Representation Learning

Dense Representation

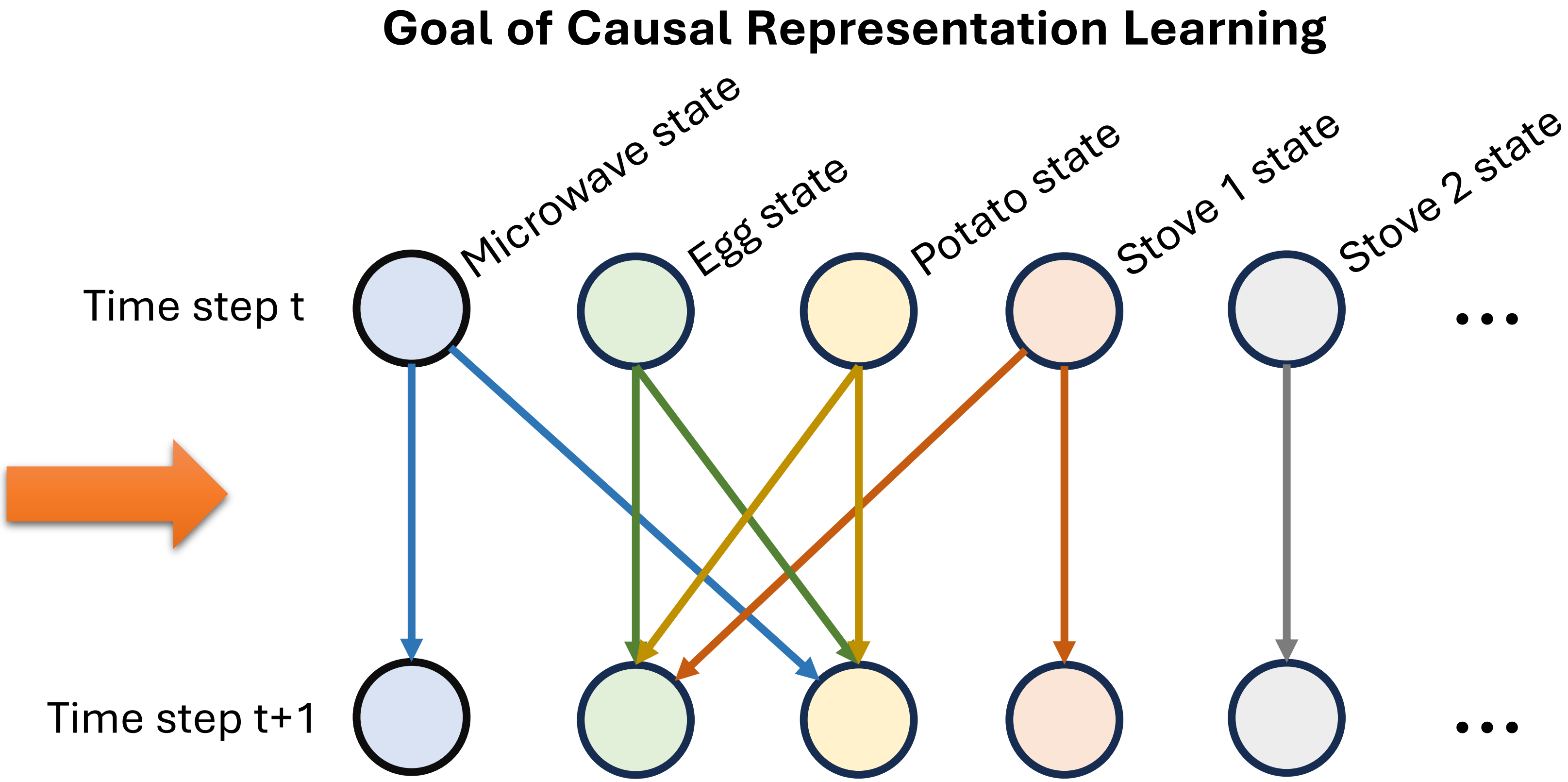


Interpretable?

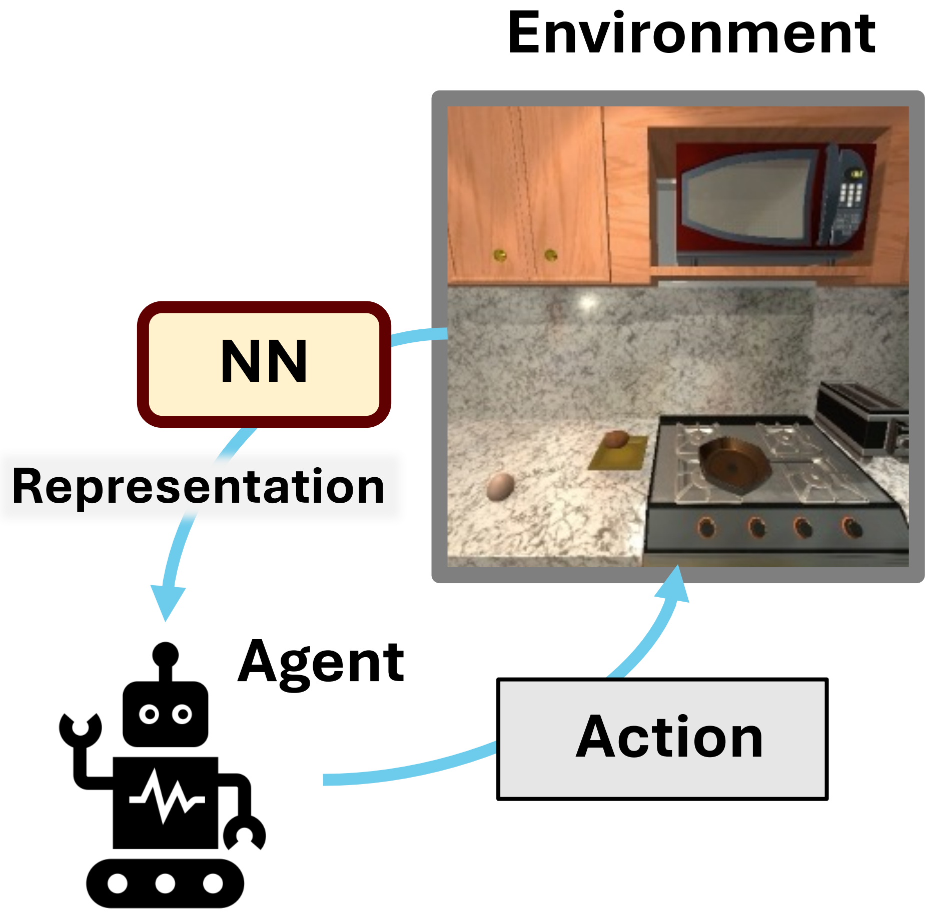
Generalizable?

Reasoning-oriented?

Causal Representation Learning



Causal Representation Learning



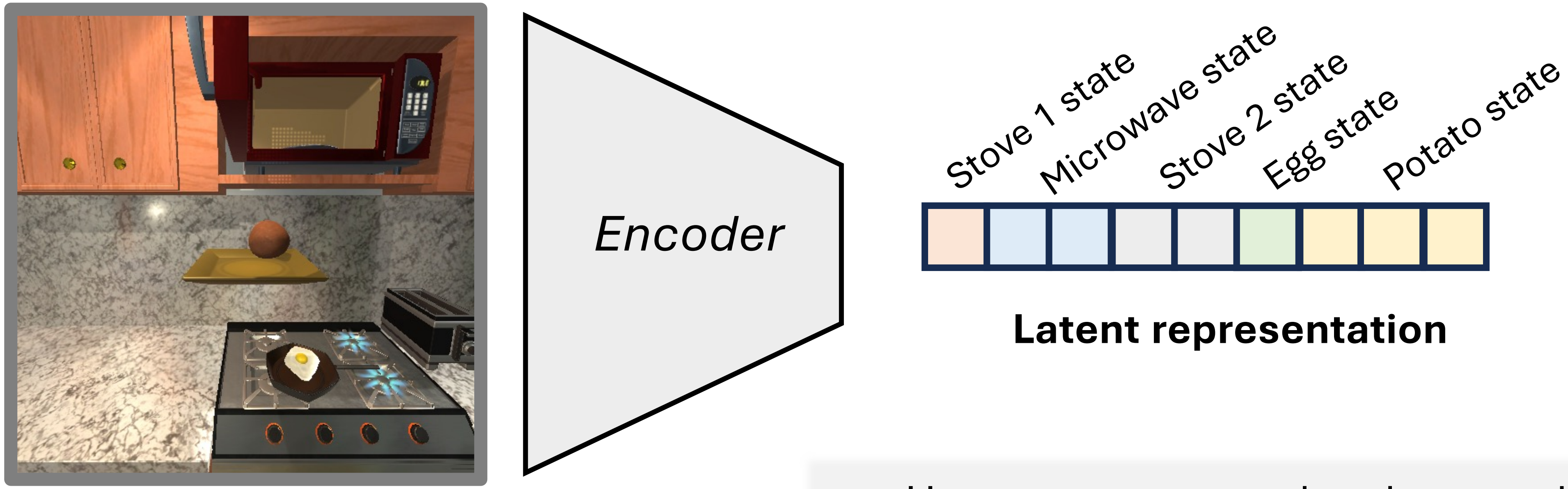
Representation Learning Tasks

What are the causal variables of the environment?

How do they interact with each other?

How can the agent intervene on causal variables?

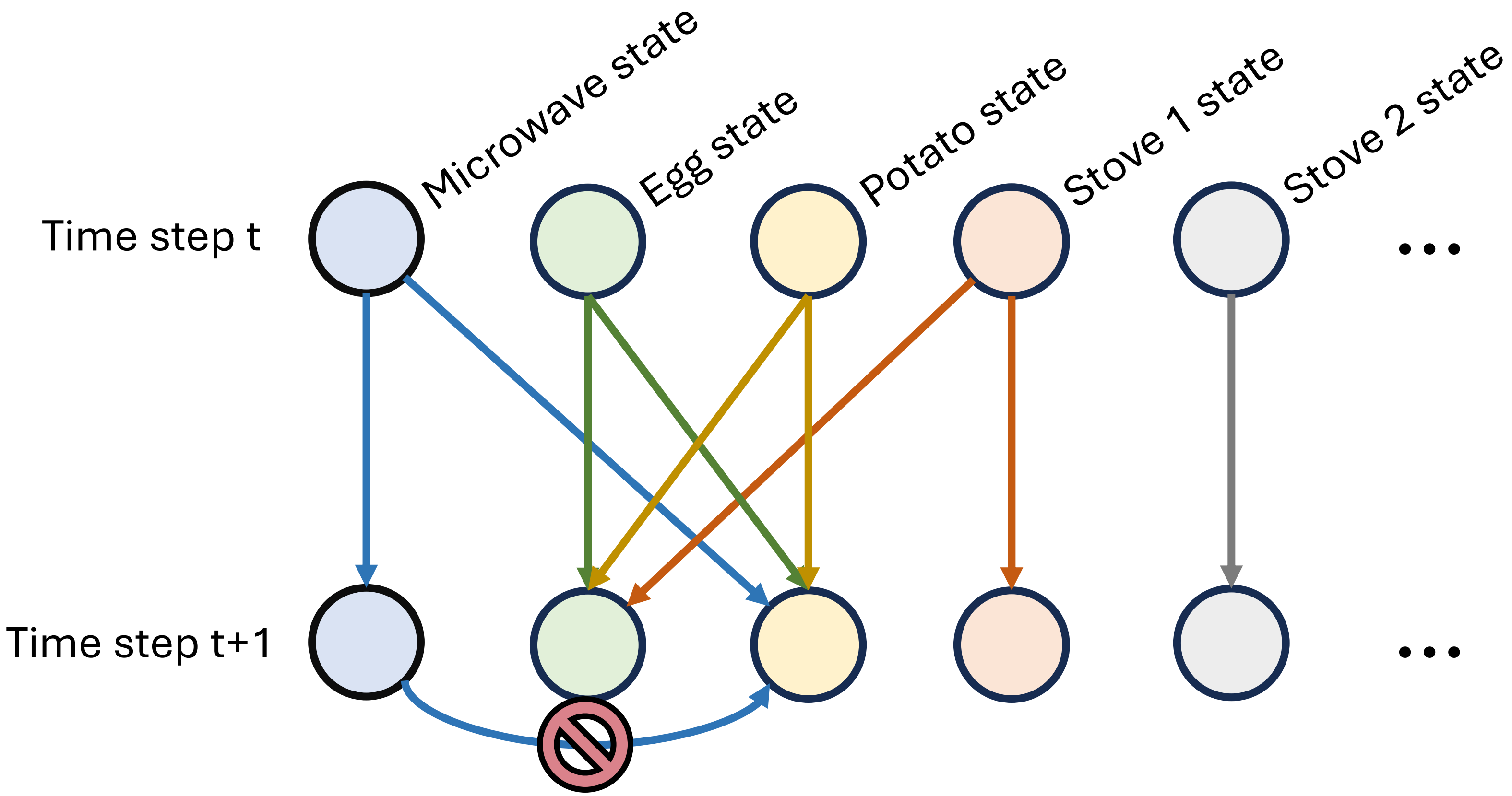
Challenges in Causal Representation Learning



How can we ensure that the causal variables are identified in the latent space?

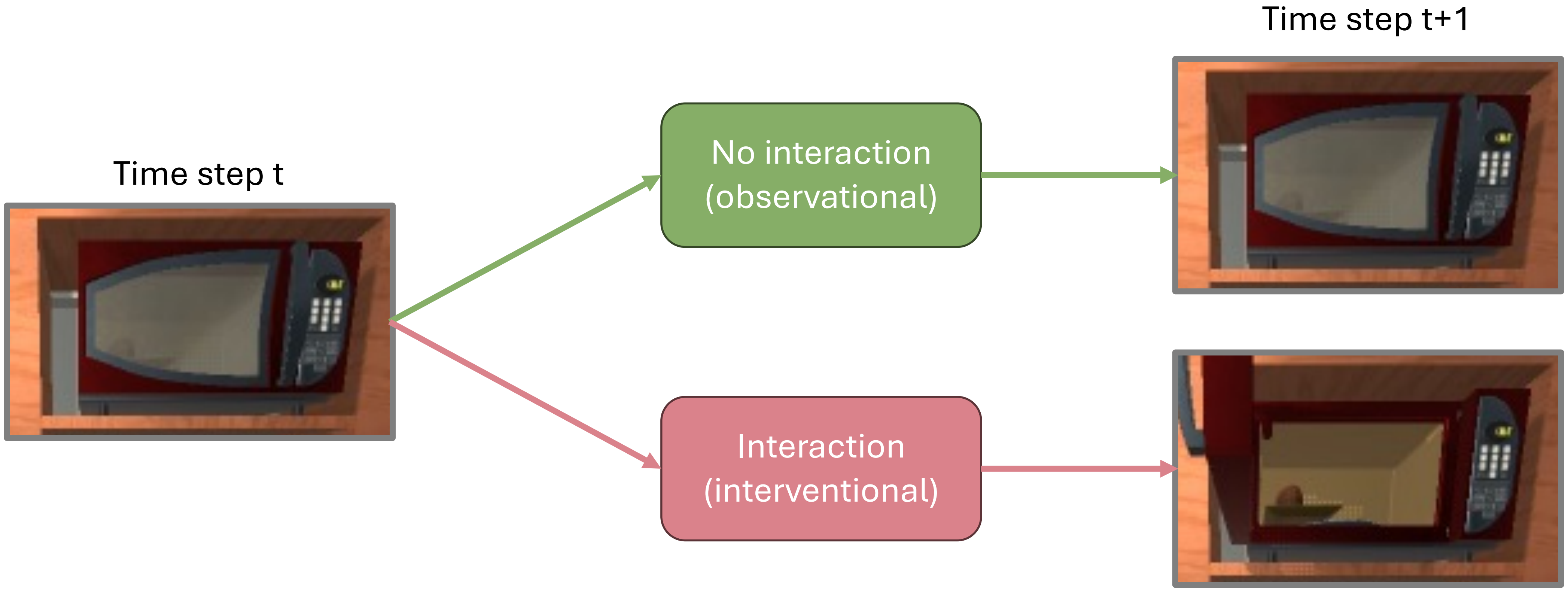
🍪 BISCUIT: Learning Causal Representations

Assumption 1: Causal Relations can be resolved over time



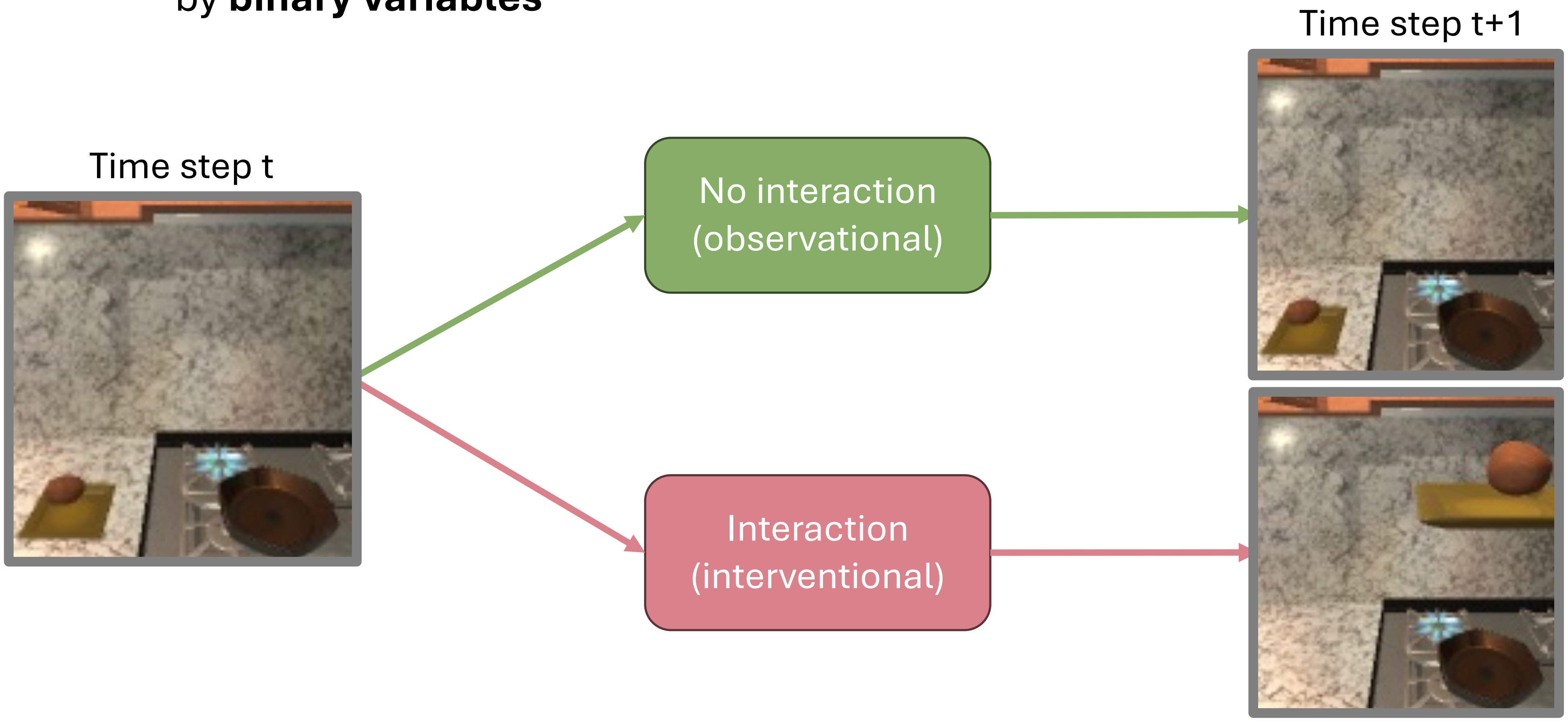
BISCUIT: Learning Causal Representations

Assumption 2: Interactions between the agent and causal variables can be described by **binary variables**



BISCUIT: Learning Causal Representations

Assumption 2: Interactions between agent and causal variables can be described by **binary variables**



🍪 BISCUIT: Theoretical Results

Assumption 1: Causal Relations can be resolved over time

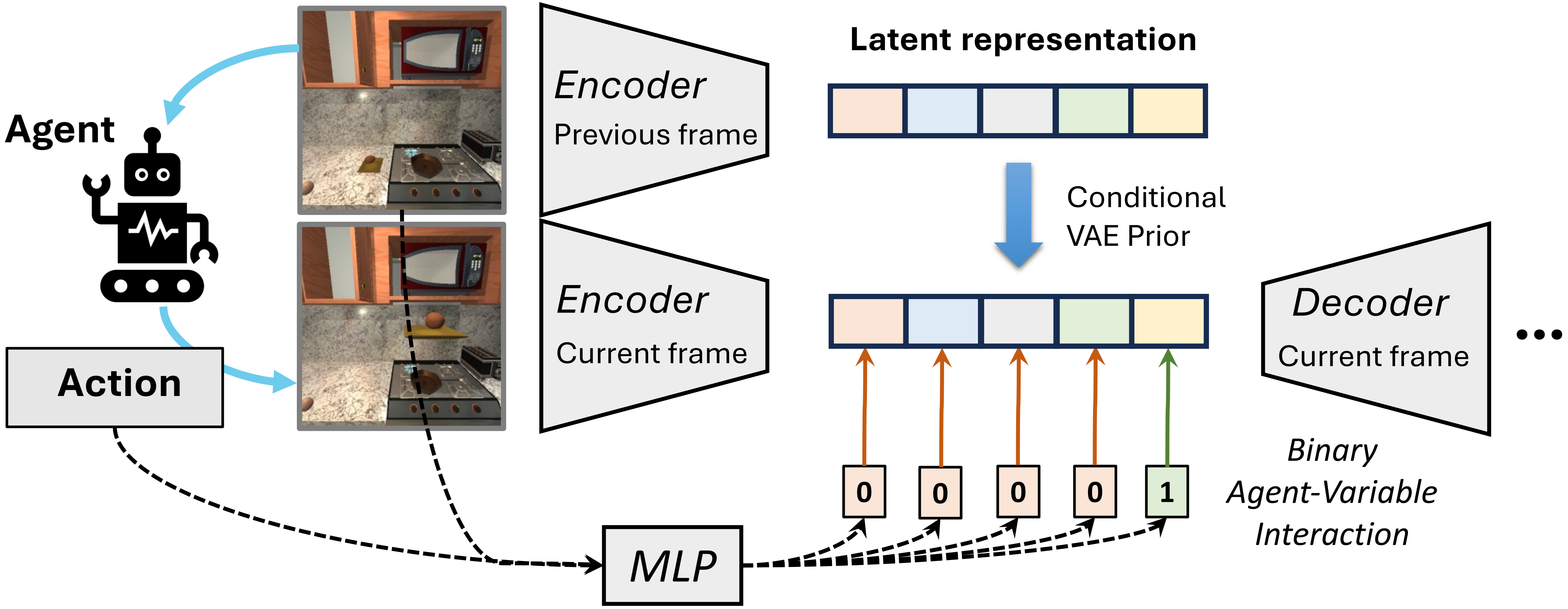
Assumption 2: Interactions between agent and causal variables can be described by **binary variables**

Assumption 3: All causal variables have different interaction patterns

Assumption 4: The causal mechanisms need to sufficiently vary on *interventions* or *over time* (allows for additive Gaussian noise models)

BISCUIT Theoretical Result
Under these assumptions, causal variables can be identified from videos with low-level actions.

BISCUIT: Architecture



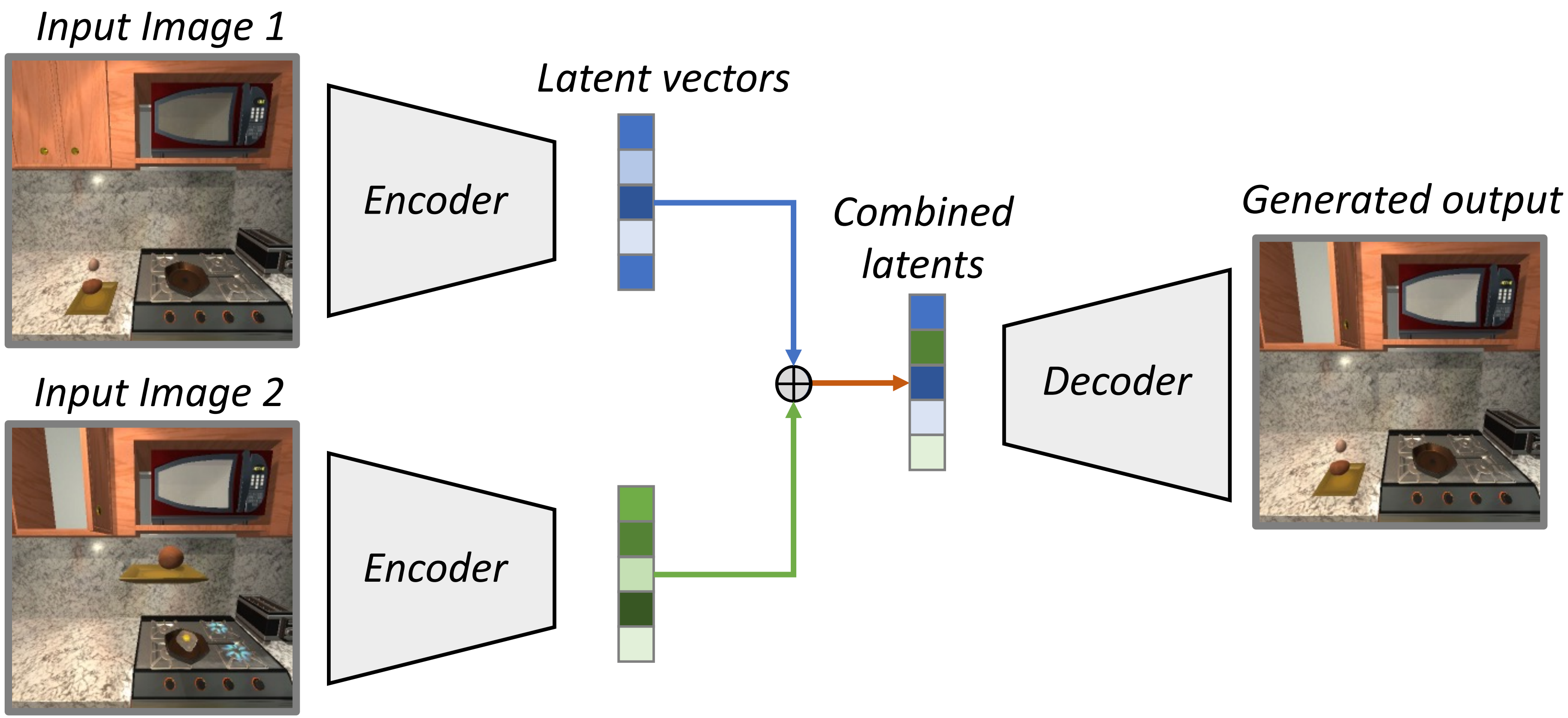
Experiments – iTHOR

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

iTHOR – Simulate Latent Interventions

- BISCUIT accurately identifies causal variables
- Validated by performing interventions in latent space

Goal
Open Cabinet
Turn on Microwave
Keep other variables fixed



iTHOR – Simulate Latent Interventions

Input image 1

Input image 2

Generated Output

Latents from image 2

Microwave Open

iTHOR – Simulate Latent Interventions

Input image 1

Input image 2

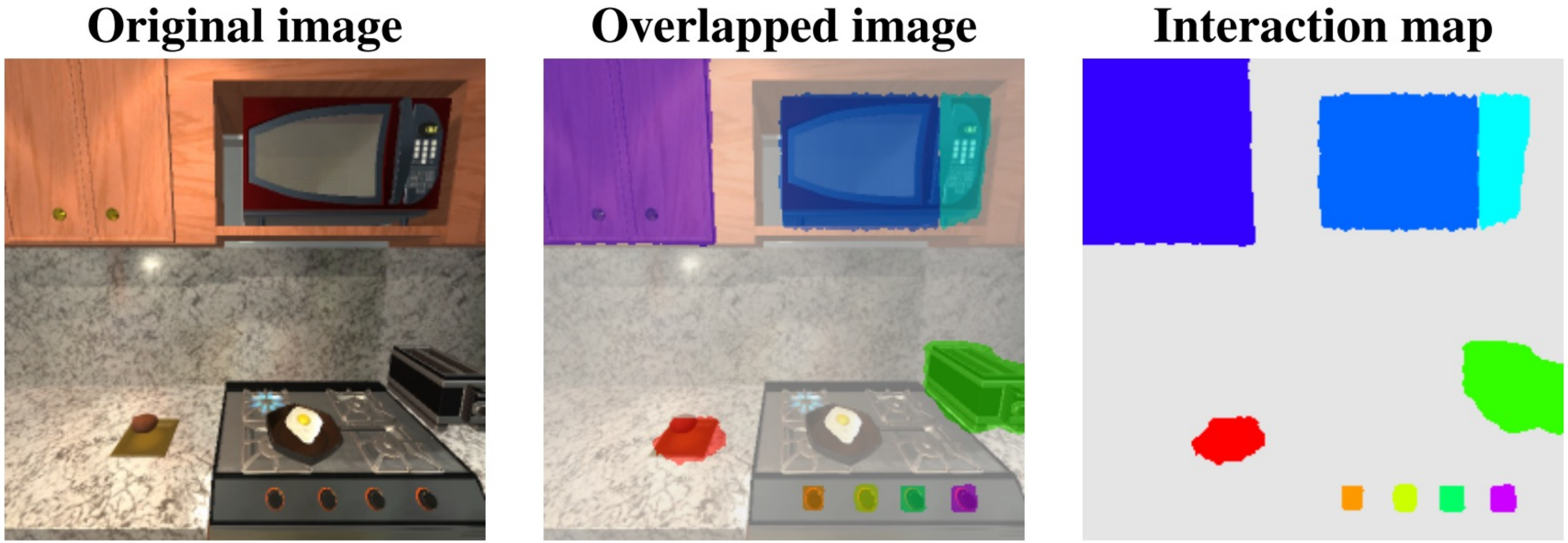
Generated Output

Latents from image 2

Stove (front-left)

iTHOR – Interaction Maps

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

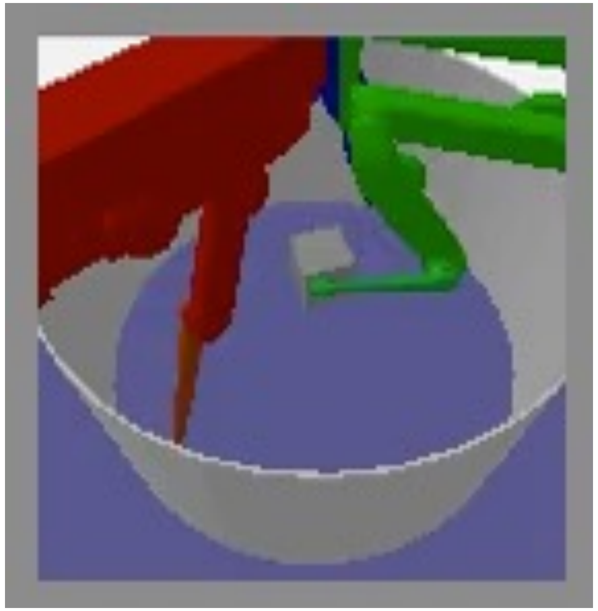


CausalWorld – Robotic Trifinger

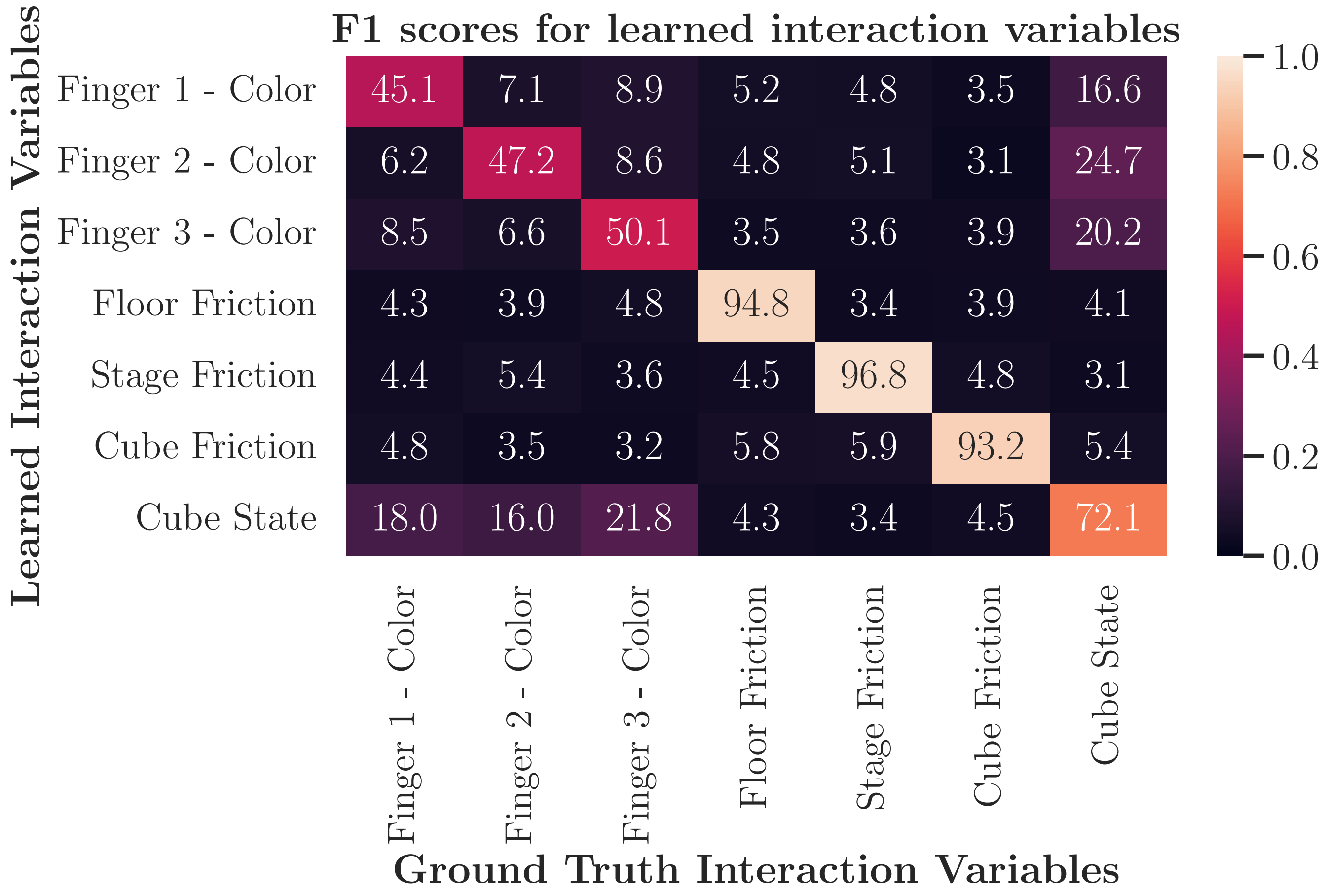
- Tri-finger robot interacting with its environment and objects
  - Causal variables include object position, frictions, colors, etc.
- Action: 9-dimensional motor angles (3 per finger)
- BISCUIT identifies causal variables accurately

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



CausalWorld – Learned Interactions



Conclusion

- BISCUIT identifies causal variables from interactive environments
- Key assumption: binary interaction variables describe agent-causal variable interactions
- Applicable to a variety of robotic and embodied AI environments
- Ability to ‘imagine’ by performing latent interventions
- Identifies actions to perform interventions

Project website and demo: phlippe.github.io/BISCUIT/