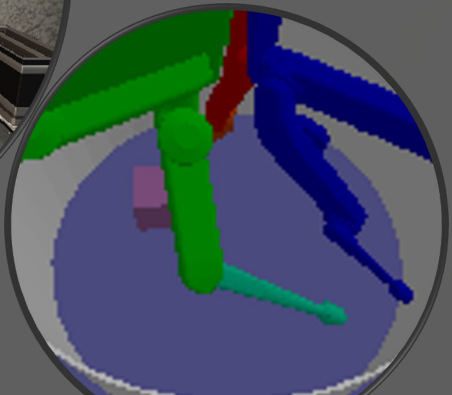
BISCUIT: Causal Representation Learning from Binary Interactions

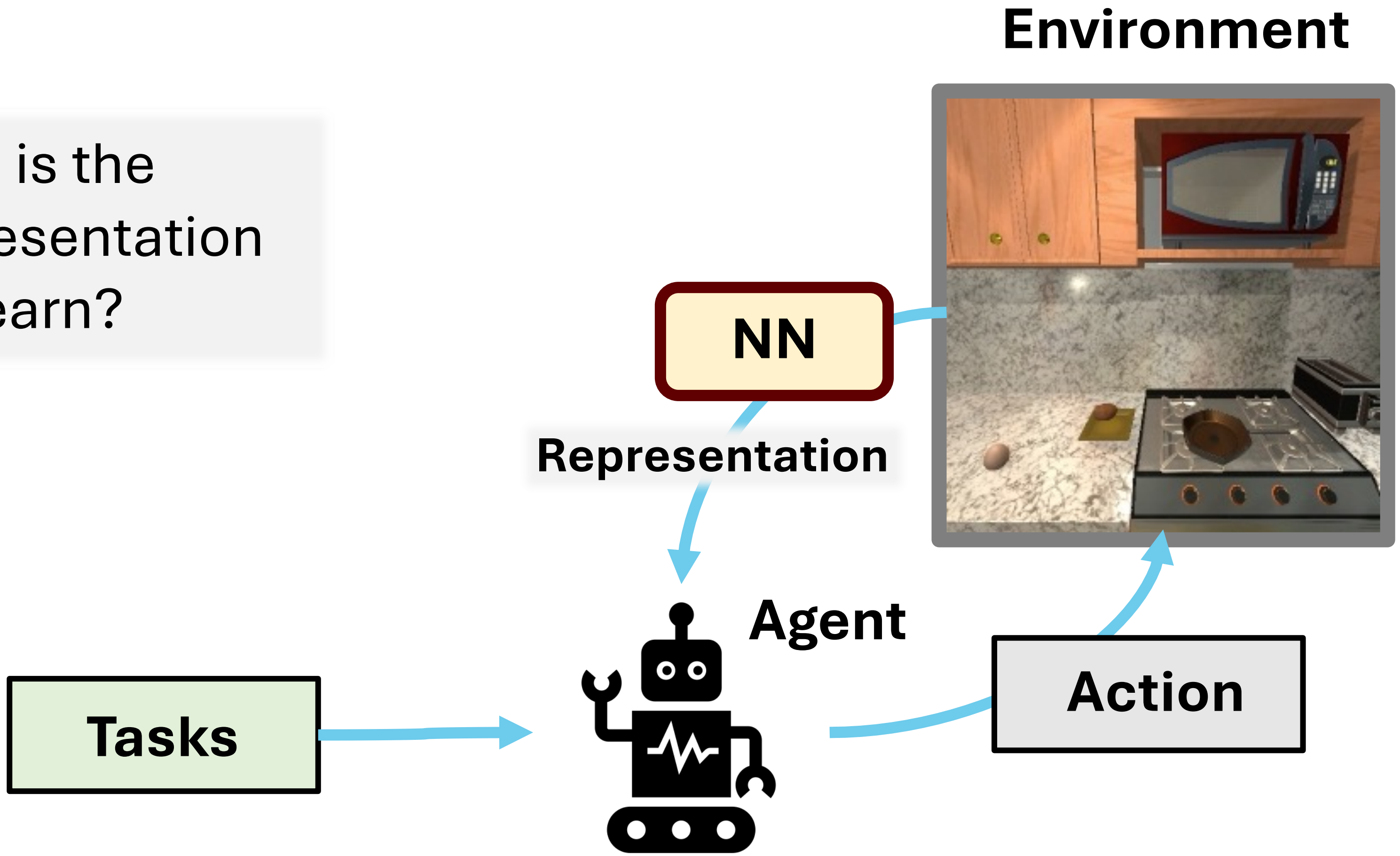


Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves

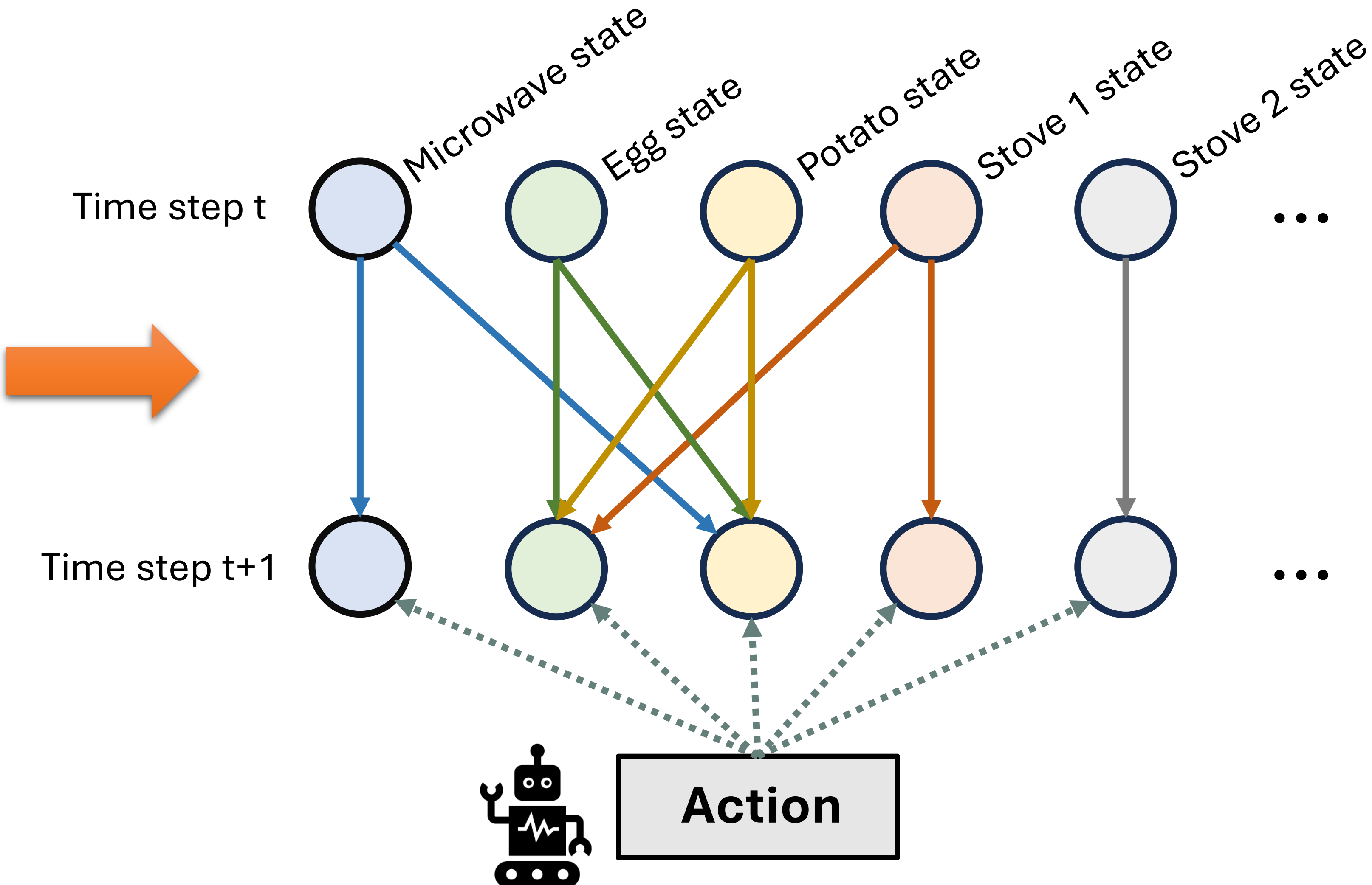
Feb 22, 2024

Problem Setup

What is the
best representation
to learn?

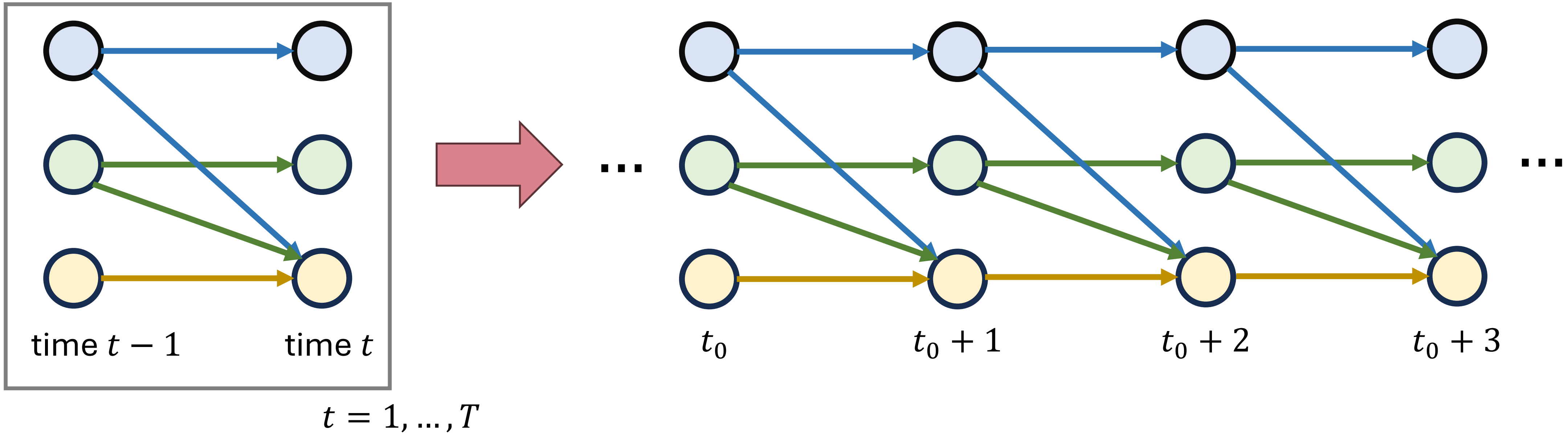


Temporal Causal Representation Learning



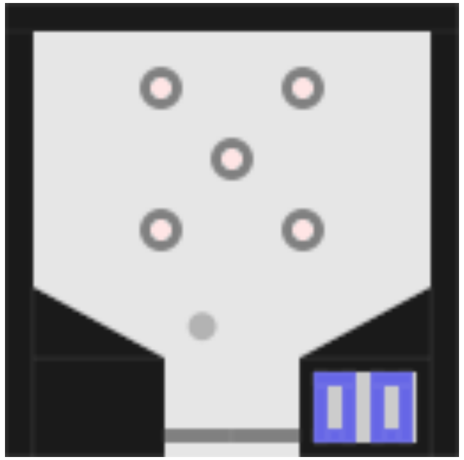
Temporal Causal Representation Learning

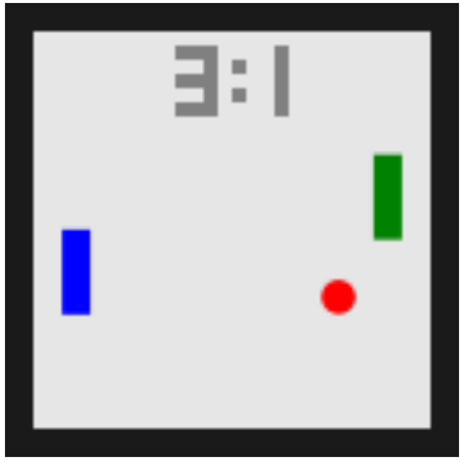
- Dynamic Bayesian Network
- Standard assumptions
  - **N-Markov**: only variables from the last N time steps can cause variables at time t
  - **Stationary/Time Invariance**: transition model stays the same across time steps

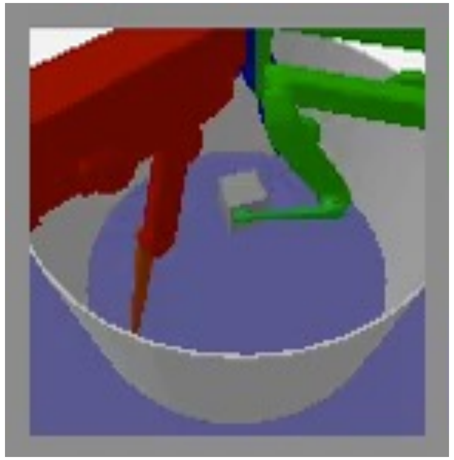


Temporal Causal Representation Learning

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

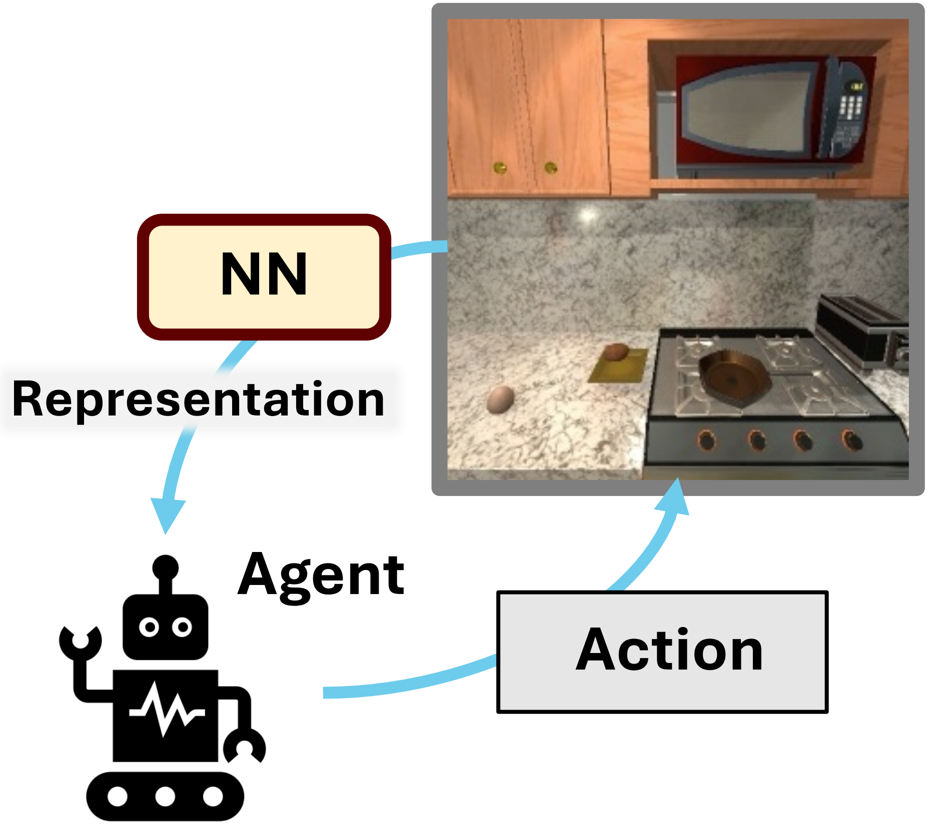






Temporal Causal Representation Learning

Environment



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?

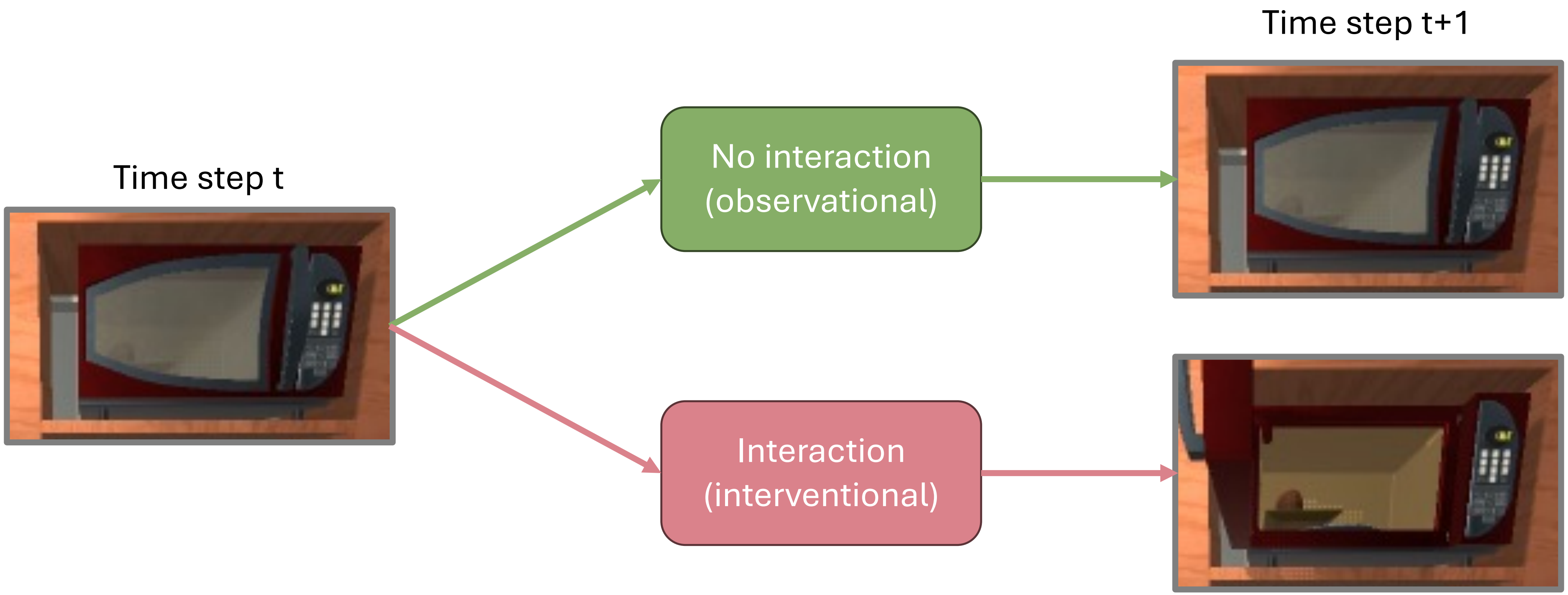
Temporal Causal Representation Learning

- **iVAE** [Khemakhem et al., 2020] – temporality as auxiliary variable, parametric assumptions
- **DMS** [Lachapelle et al., 2022] – graphical assumption (mechanism sparsity), exponential family
- **LEAP** [Yao et al., 2022ab] – sufficient mechanism variability over regimes/environments
- **Properties of Mechanisms** [Ahuja et al., 2022] – known functional form of mechanisms
- **CITRIS** [Lippe et al., 2022] – non-parameteric, known intervention targets
  - **iCITRIS** [Lippe et al., 2023a] – instantaneous effects

BISCUIT – non-parameteric, arbitrary graphs, unknown <u>binary</u> interactions

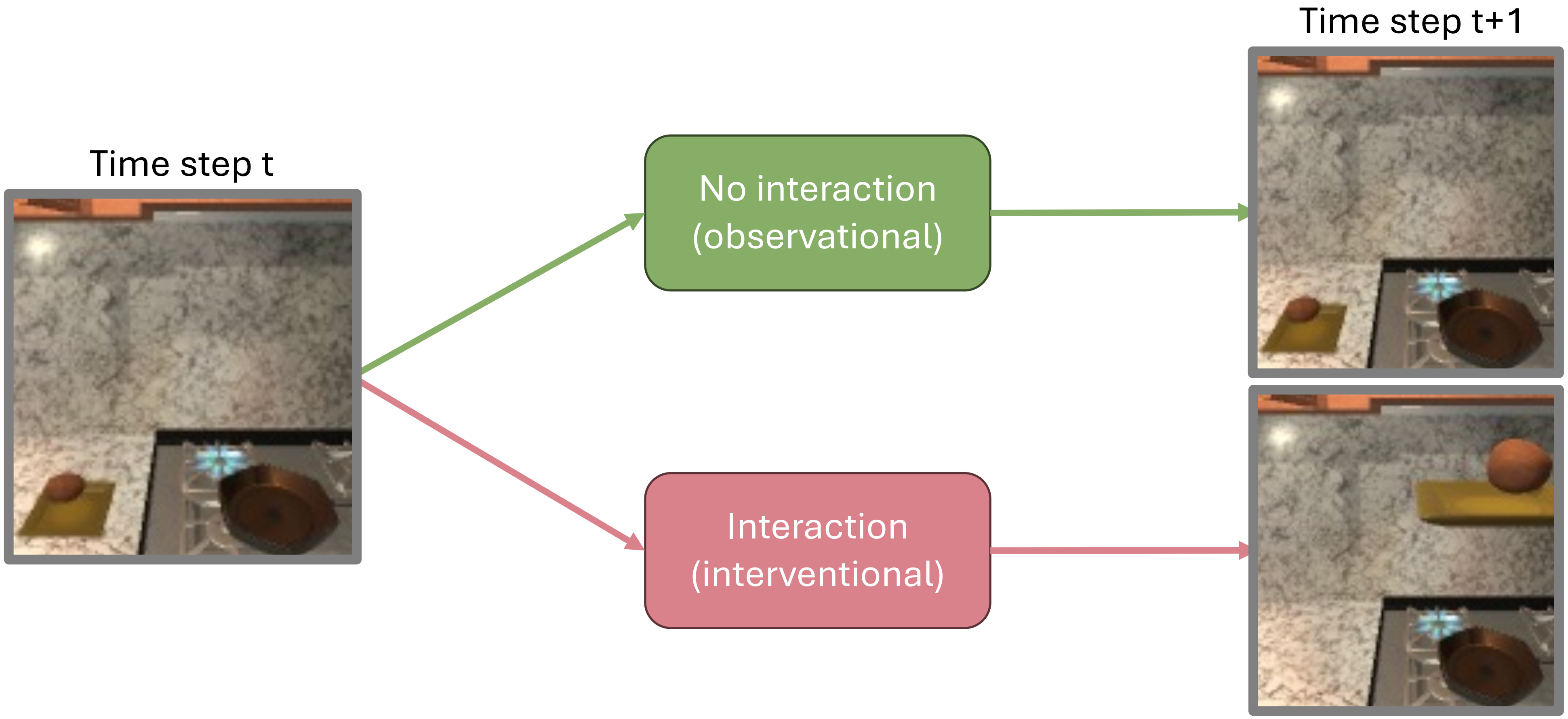
BISCUIT: Binary Interactions

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



BISCUIT: Binary Interactions

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

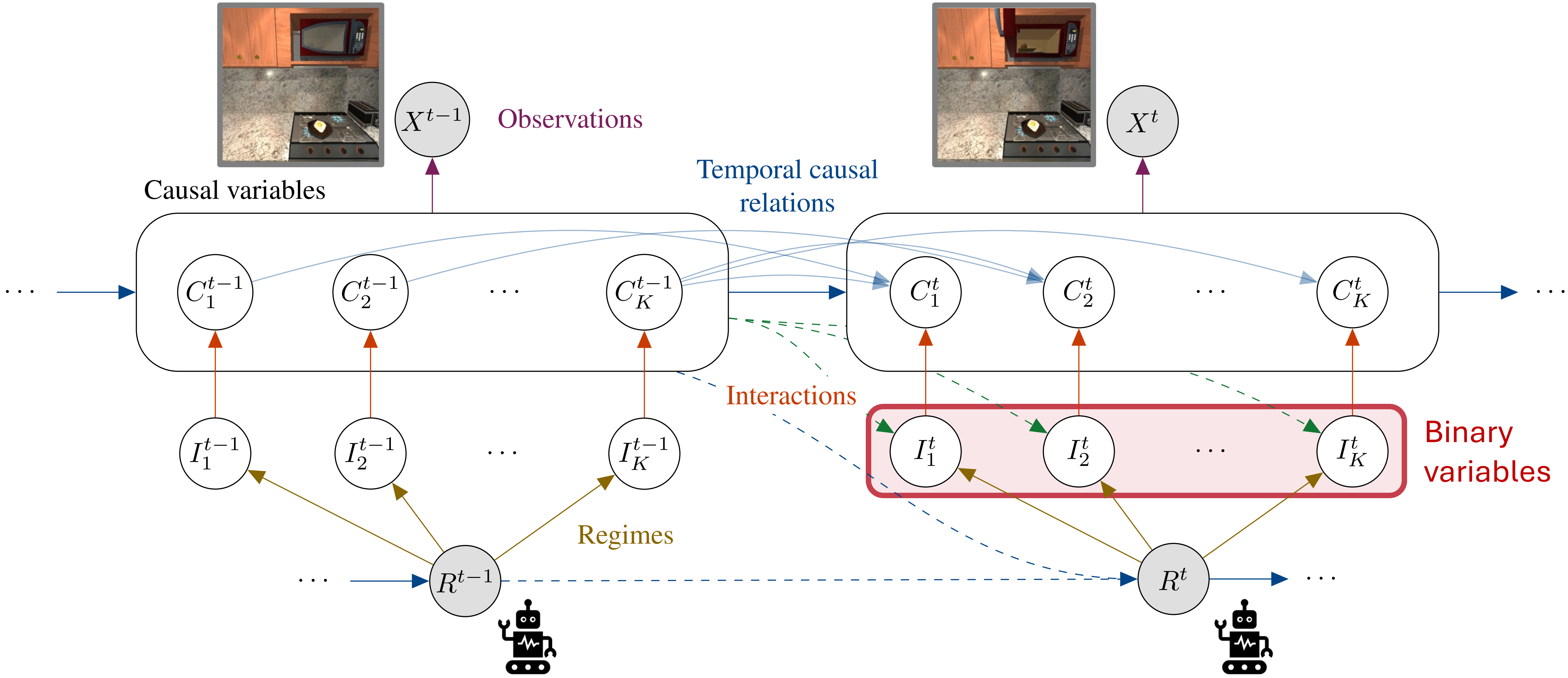


BISCUIT: Binary Interactions

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

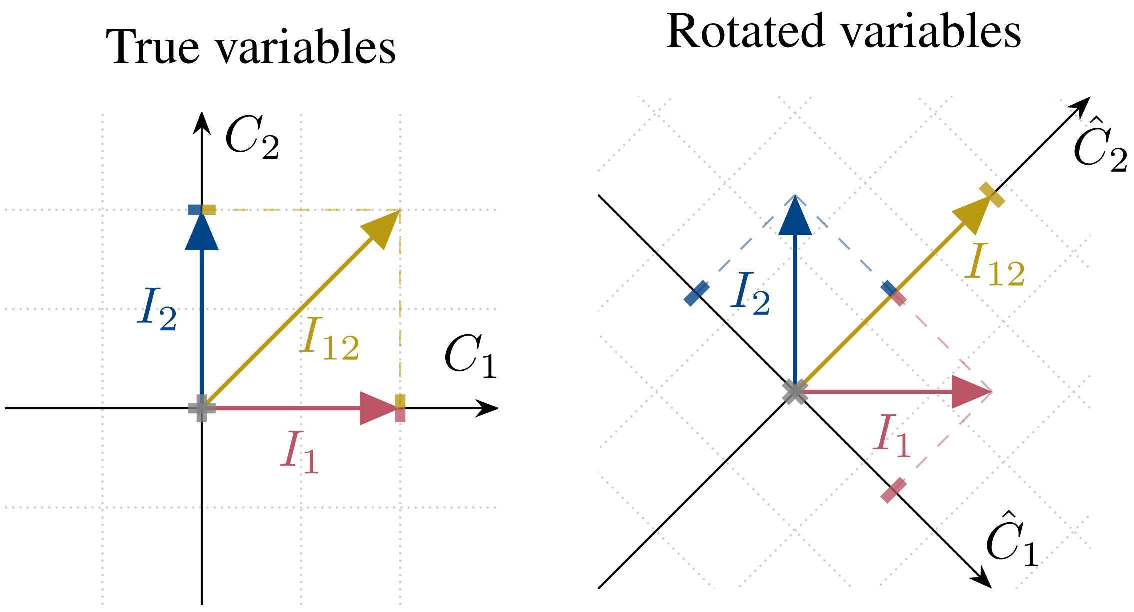
- Causal variables can be continuous values, evolving stochastically over time
- Certain interactions cause unknown interventions, changing corresponding mechanisms
- Realistic assumption in many RL environments:
  observational = no agent-variable interaction,
  interventional = agent interacting with variable

BISCUIT: Causal Model



Binary Interactions enable Identifiability

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

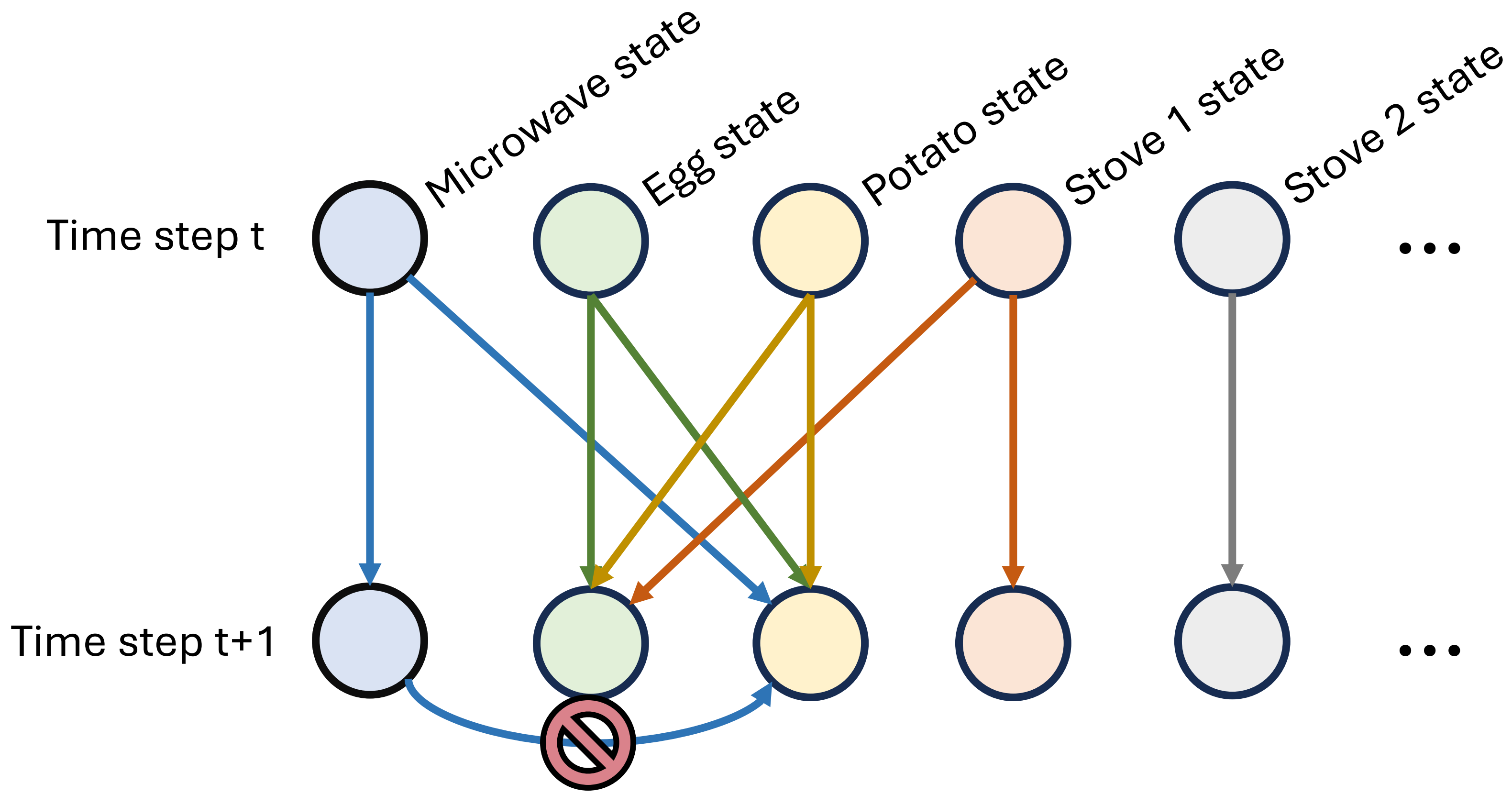


Identifiability Assumptions

- **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 $\lfloor \mathbf{log_2}\, \boldsymbol{K} \rfloor + \mathbf{2}$ actions/values of R^t
  - Example: agent with random policy

Identifiability Assumptions

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



Identifiability Assumptions

- **Assumption 4**: The causal mechanisms vary sufficiently over time or on interactions
  - Prevents cases like interventional and observational distribution being identical
  - Supports many common setups like additive Gaussian noise models or more complex distributions

A. *(**Dynamics Variability**) Each variable's log-likelihood difference is twice differentiable and not always zero:*

$$\forall C_i^t, \exists C^{t-1}: \frac{\partial^2 \Delta(C_i^t|C^{t-1})}{\partial (C_i^t)^2} \neq 0;$$

B. *(**Time Variability**) For any $C^t \in \mathcal{C}$, there exist $K+1$ different values of C^{t-1} denoted with $c^1, ..., c^{K+1} \in \mathcal{C}$, for which the vectors $v_1, ..., v_K \in \mathbb{R}^{K+1}$ with*

$$v_i = \left[\frac{\partial \Delta\left(C_i^t|C^{t-1}=c^1\right)}{\partial C_i^t} \quad \dots \quad \frac{\partial \Delta\left(C_i^t|C^{t-1}=c^{K+1}\right)}{\partial C_i^t}\right]^T$$

are linearly independent.

BISCUIT: Identifiability Results

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

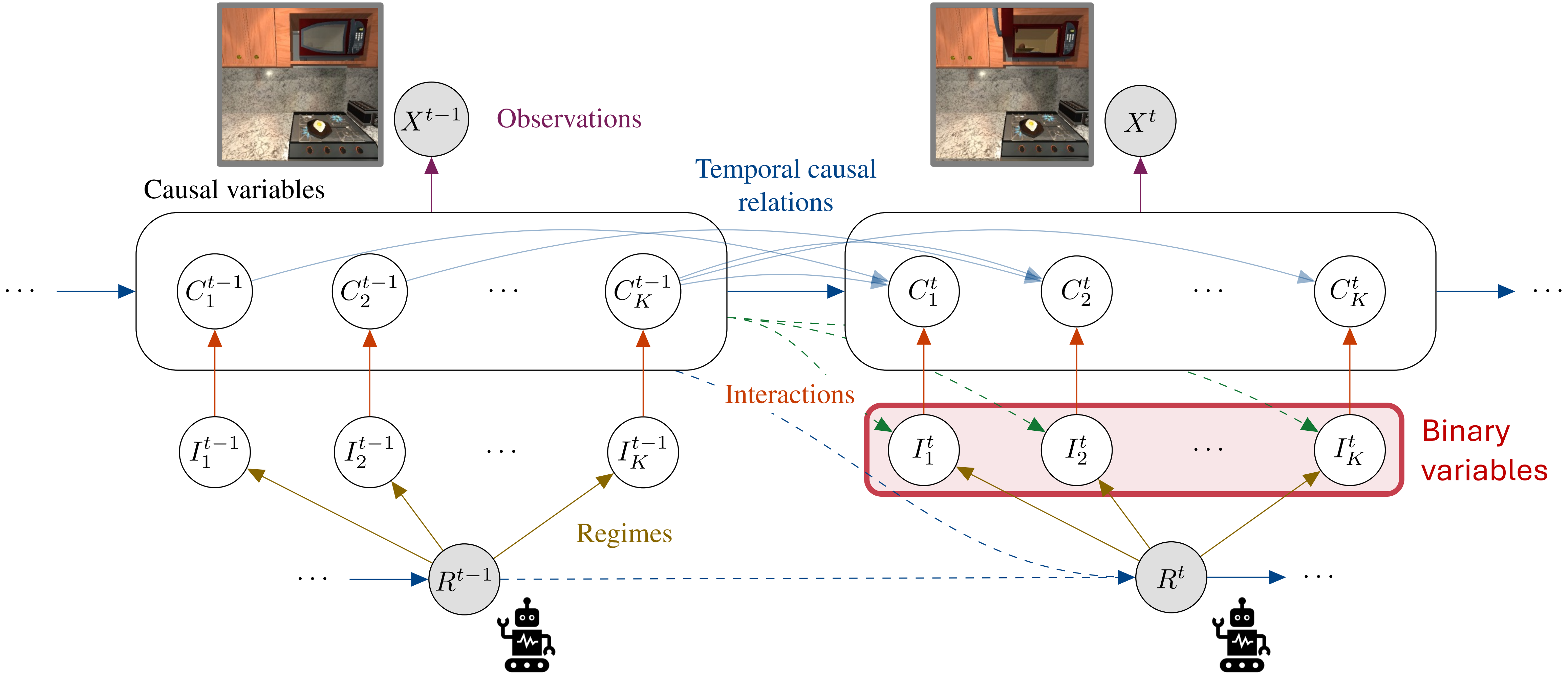
Assumption 2: All causal variables have different interaction patterns

Assumption 3: Causal Relations can be resolved over time

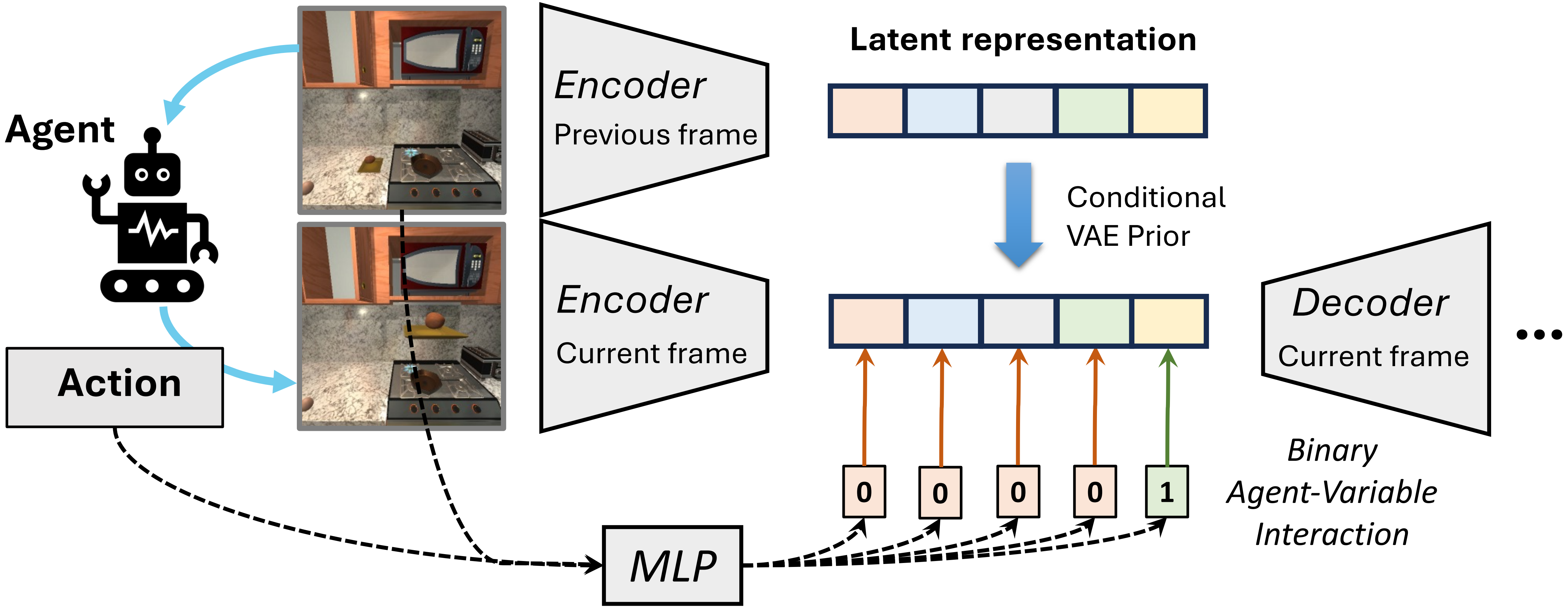
Assumption 4: The causal mechanisms vary sufficiently over time or on interactions

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:

$$p_\omega(z^t|z^{t-1},R^t) = \prod_i p_\omega\left(z_i^t|z^{t-1},\underbrace{f_i(R^t,z^{t-1})}_{\textit{Binary function output}}\right)$$

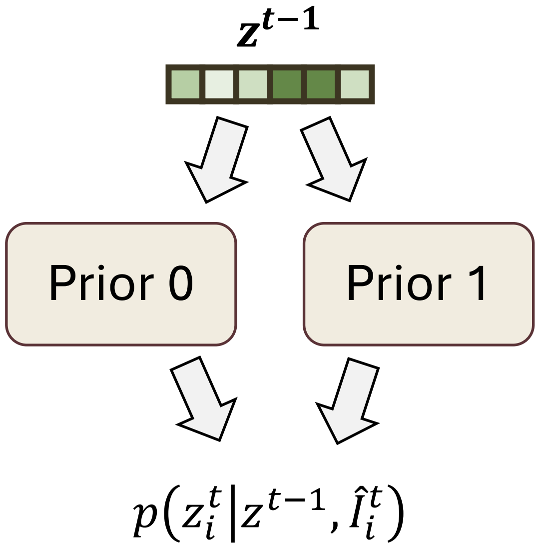
BISCUIT: Learning Binary Variables

- Prior $p(z_i^t | z^{t-1}, \hat{I}_i^t)$
  - $\hat{I}_i^t = f_i(z^{t-1}, R^t)$

BISCUIT: Learning Binary Variables

- Prior $p(z_i^t|z^{t-1}, \hat{I}_i^t)$
  - $\hat{I}_i^t = f_i(z^{t-1}, R^t)$

- Option 1: Marginalizing
  - $p\left(z_i^t \middle| z^{t-1}, \hat{I}_i^t\right) = p\left(\hat{I}_i^t = 0| \dots\right)p\left(z_i^t \middle| z^{t-1}, 0\right) + p(\hat{I}_i^t = 1| \dots)p\left(z_i^t \middle| z^{t-1}, 1\right)$
  - Converges to $p\left(z_i^t \middle| z^{t-1}, 0\right) = p\left(z_i^t \middle| z^{t-1}, 1\right)$

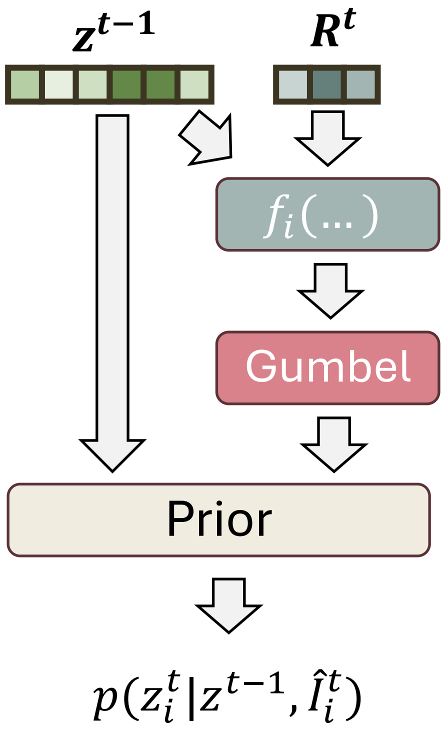


BISCUIT: Learning Binary Variables

- Prior $p(z_i^t | z^{t-1}, \hat{I}_i^t)$
  - $\hat{I}_i^t = f_i(z^{t-1}, R^t)$

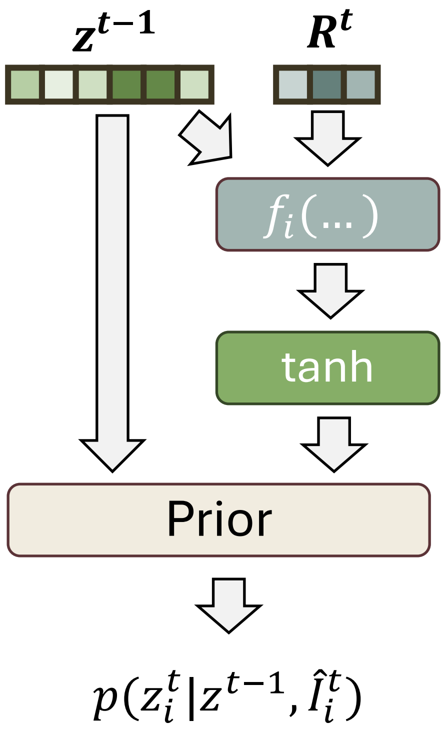
- Option 1: Marginalizing

- Option 2: Gumbel Sigmoid
  - $\hat{I}_i^t = \text{GumbelSigmoid}\left(f_i(z^{t-1}, R^t)\right)$
  - High variance causes local minima

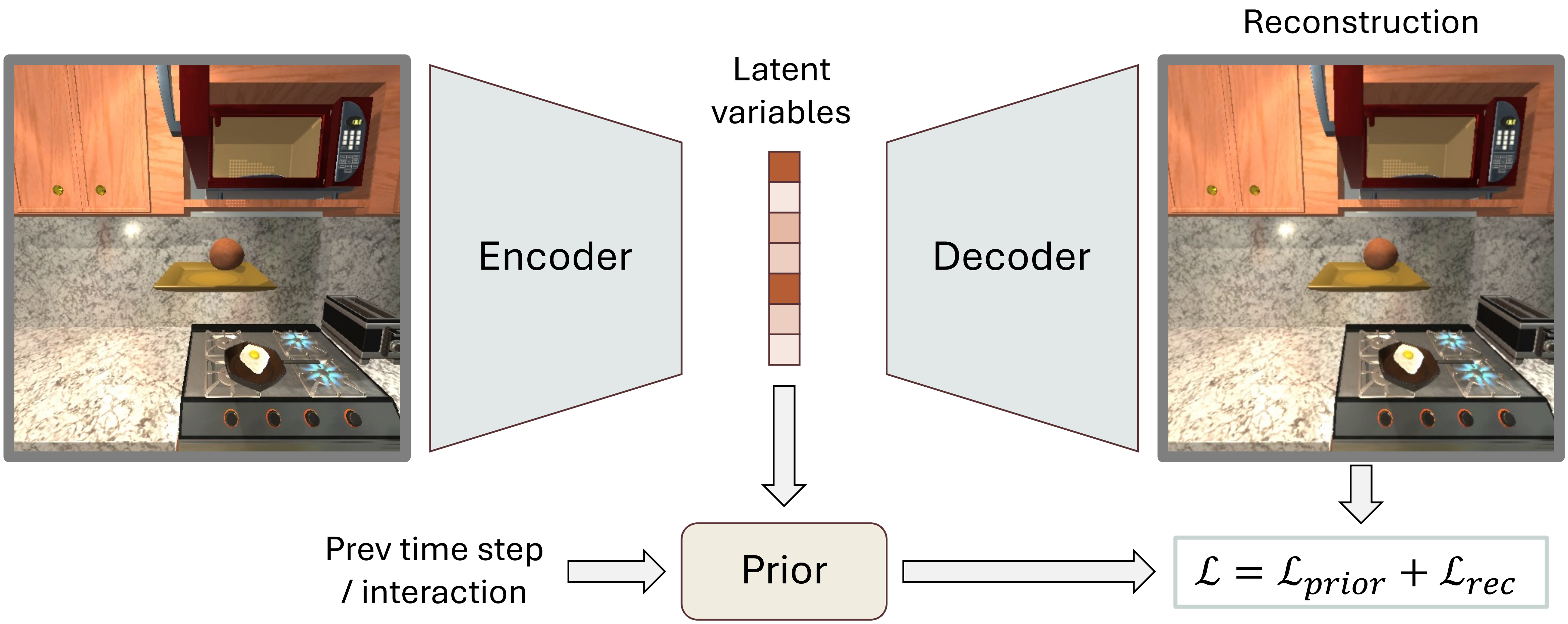


BISCUIT: Learning Binary Variables

- Prior $p(z_i^t|z^{t-1},\hat{I}_i^t)$
  - $\hat{I}_i^t = f_i(z^{t-1},R^t)$
- Option 1: Marginalizing
- Option 2: Gumbel Sigmoid
- Option 3: Continuous Relaxation
  - $\hat{I}_i^t = \tanh\left(\frac{f_i(z^{t-1},R^t)}{\tau}\right)$
  - Smooth optimization
  - Decrease temperature over training

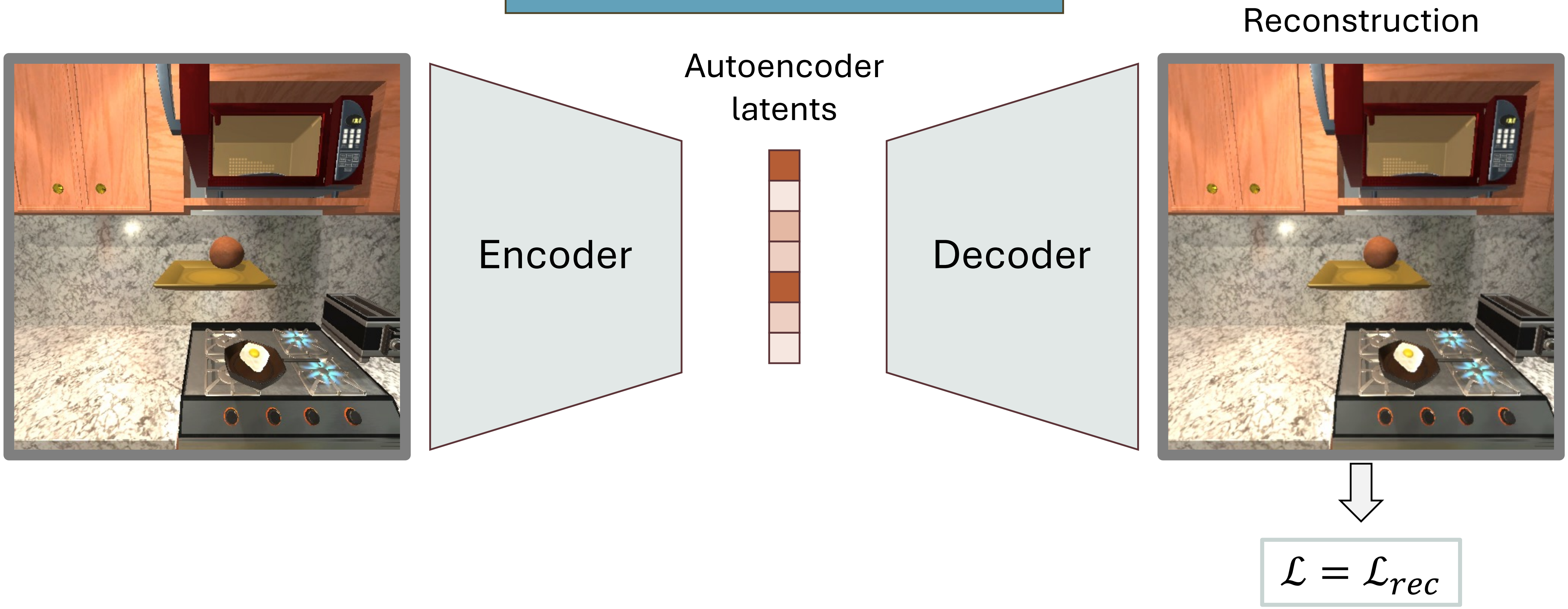


VAE: Competing Losses

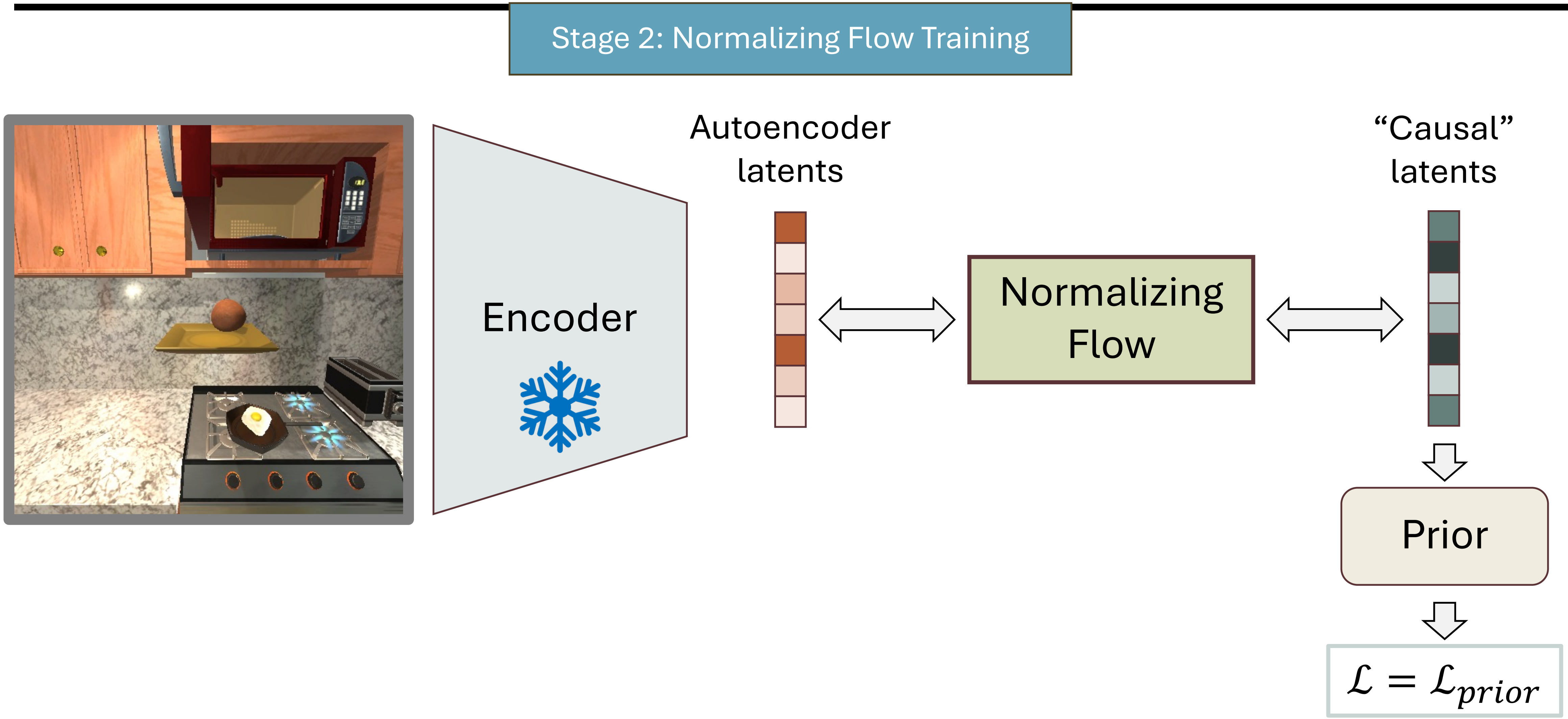


AE+NF: Splitting Objectives

Stage 1: Autoencoder Training



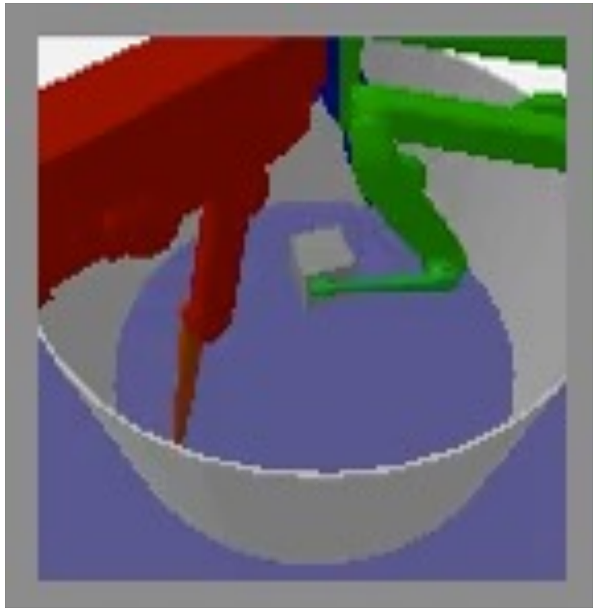
AE+NF: Splitting Objectives



Experiments

Synthetic Environment

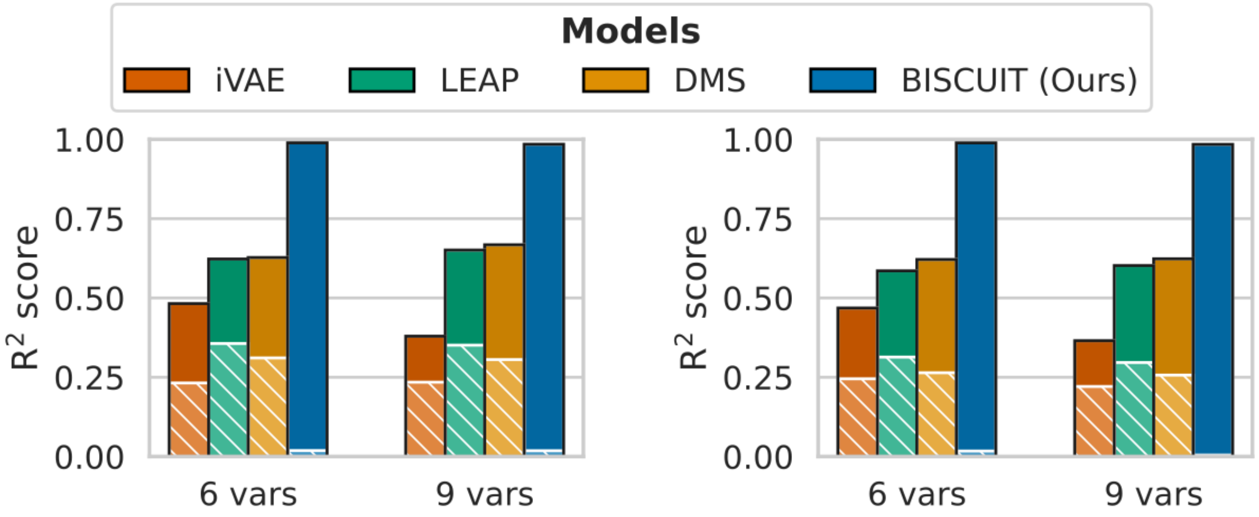
CausalWorld



iTHOR

Synthetic Environments

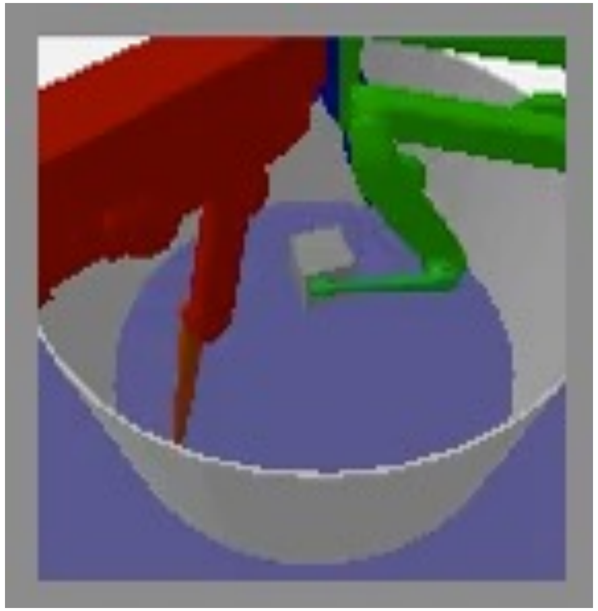
- Evaluated on synthetic dataset with additive Gaussian noise model
- Identifies causal variables well, also under mininal bound of interactions



(a) Random Interactions (b) Minimal Interactions

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

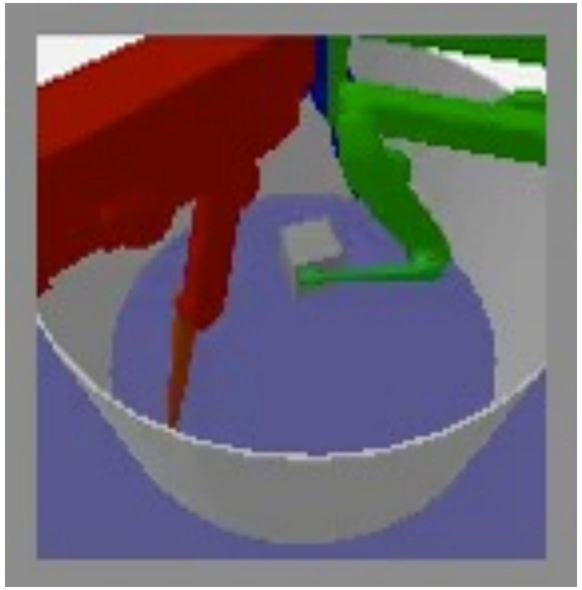


Accuracy of learned causal variables
(higher is better / lower is better)

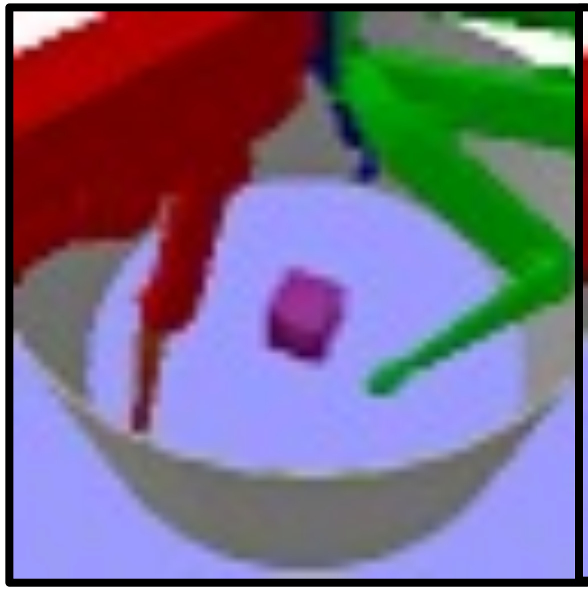
Models	**CausalWorld**
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00
LEAP (Yao et al., 2022b)	0.30 / 0.00
DMS (Lachapelle et al., 2022b)	0.32 / 0.00
BISCUIT-NF (Ours)	**0.97** / 0.01

CausalWorld – AE + NF

Example Sequence

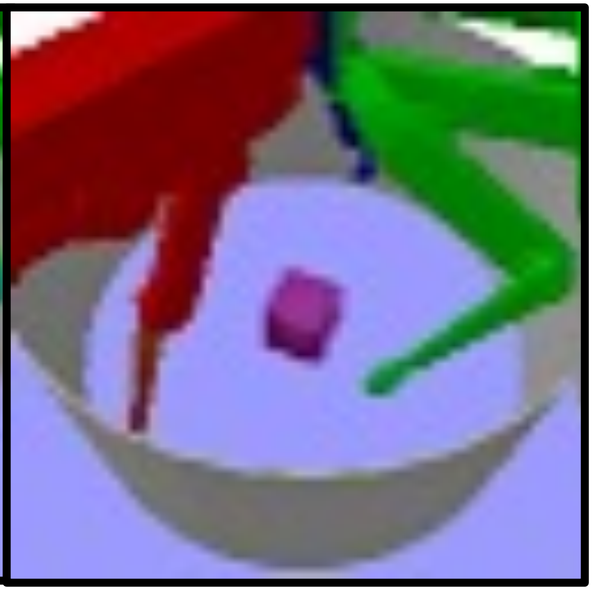


Ground Truth

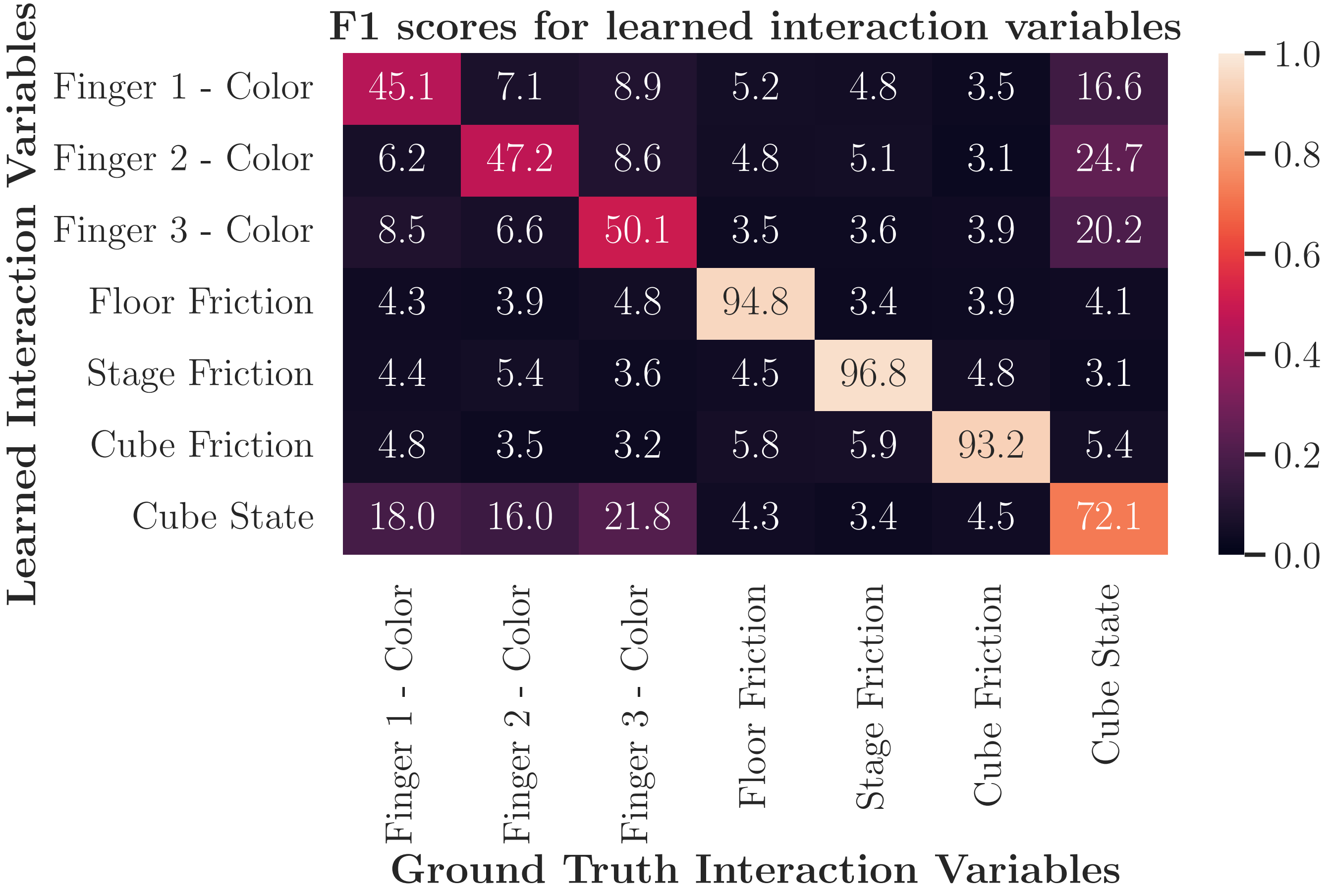


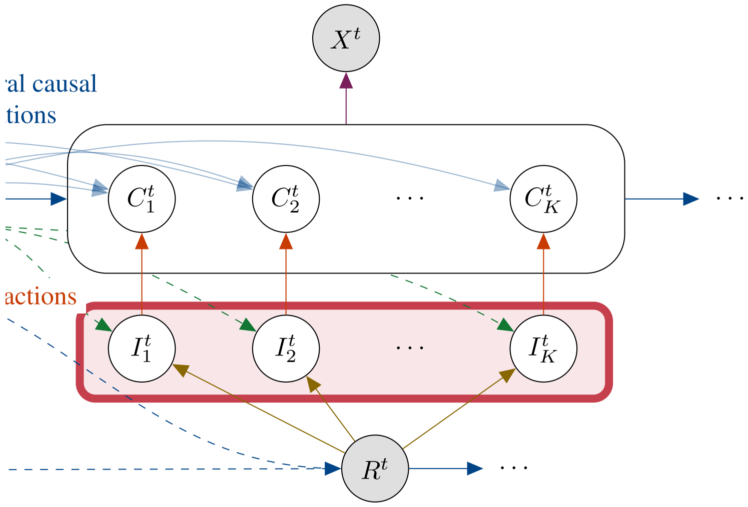
β-VAE

BISCUIT – AE + NF



CausalWorld – Learned Interactions





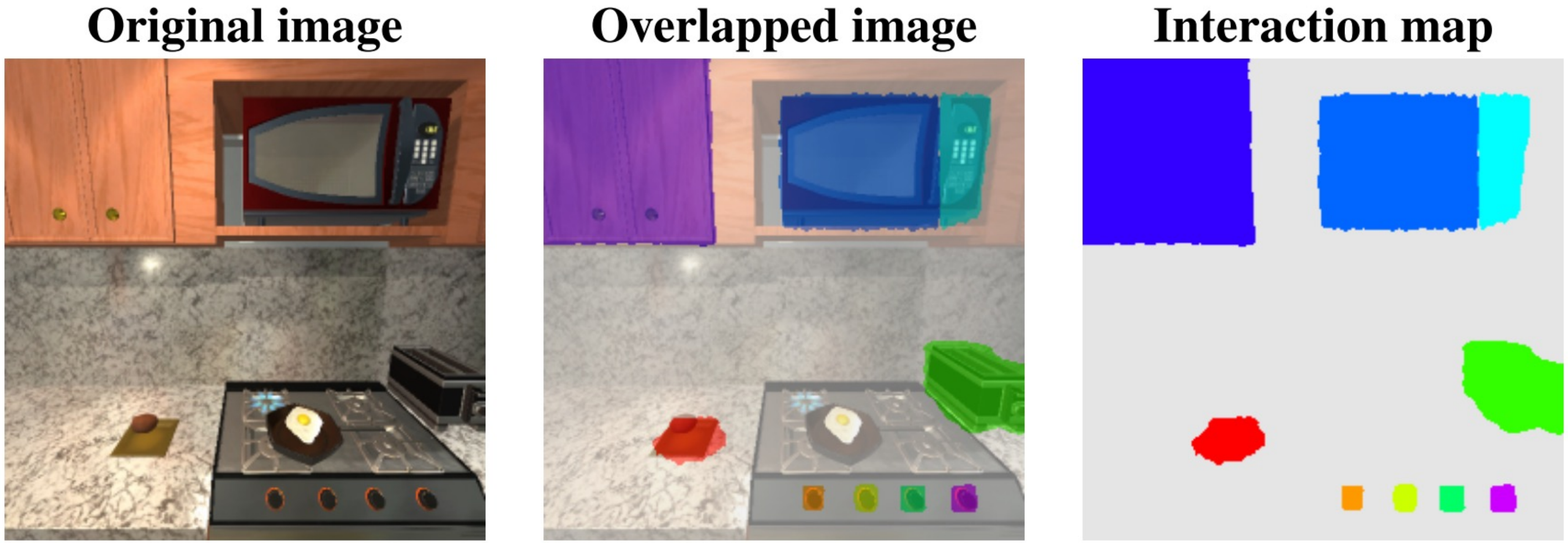
iTHOR

- 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**

iTHOR – Interaction Maps

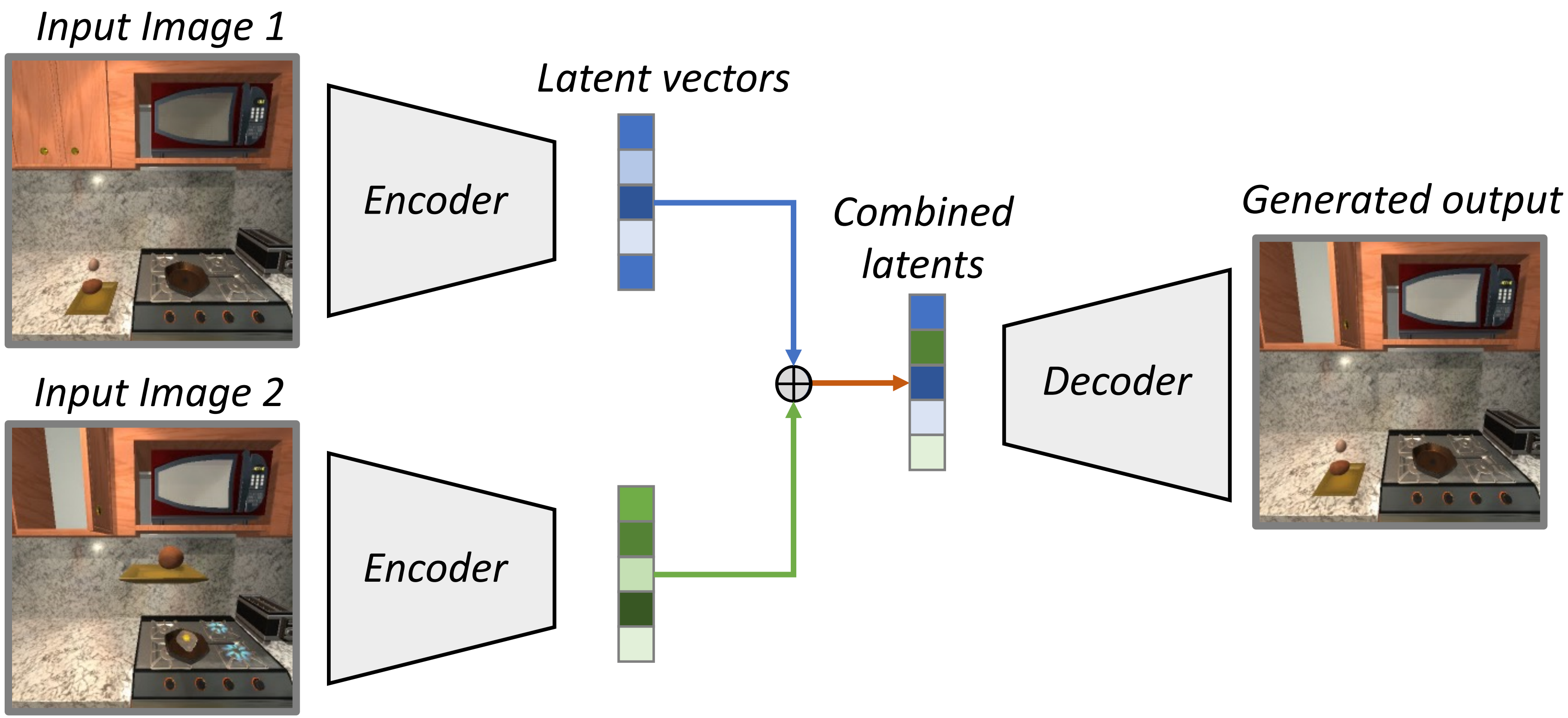
- Visualize learned interaction variables by the x-y locations they are active
- Each causal variable shown in different color



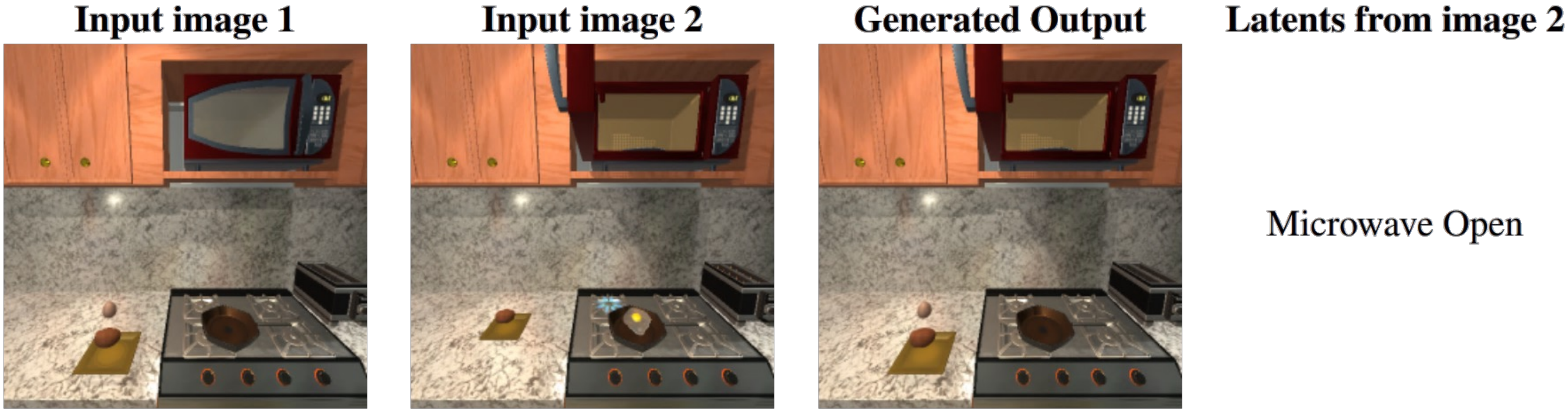
iTHOR – Triplet Evaluation

- 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



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

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/

Collaborators

Sara Magliacane

Sindy Löwe

Yuki Asano

Taco Cohen

Efstratios Gavves

University of Amsterdam
Faculty of Science

QUVA
Deep Vision Lab

Qualcomm
AI research

AMLAB
Amsterdam
Machine Learning Lab

MIT-IBM
Watson
AI Lab

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 Processing 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.