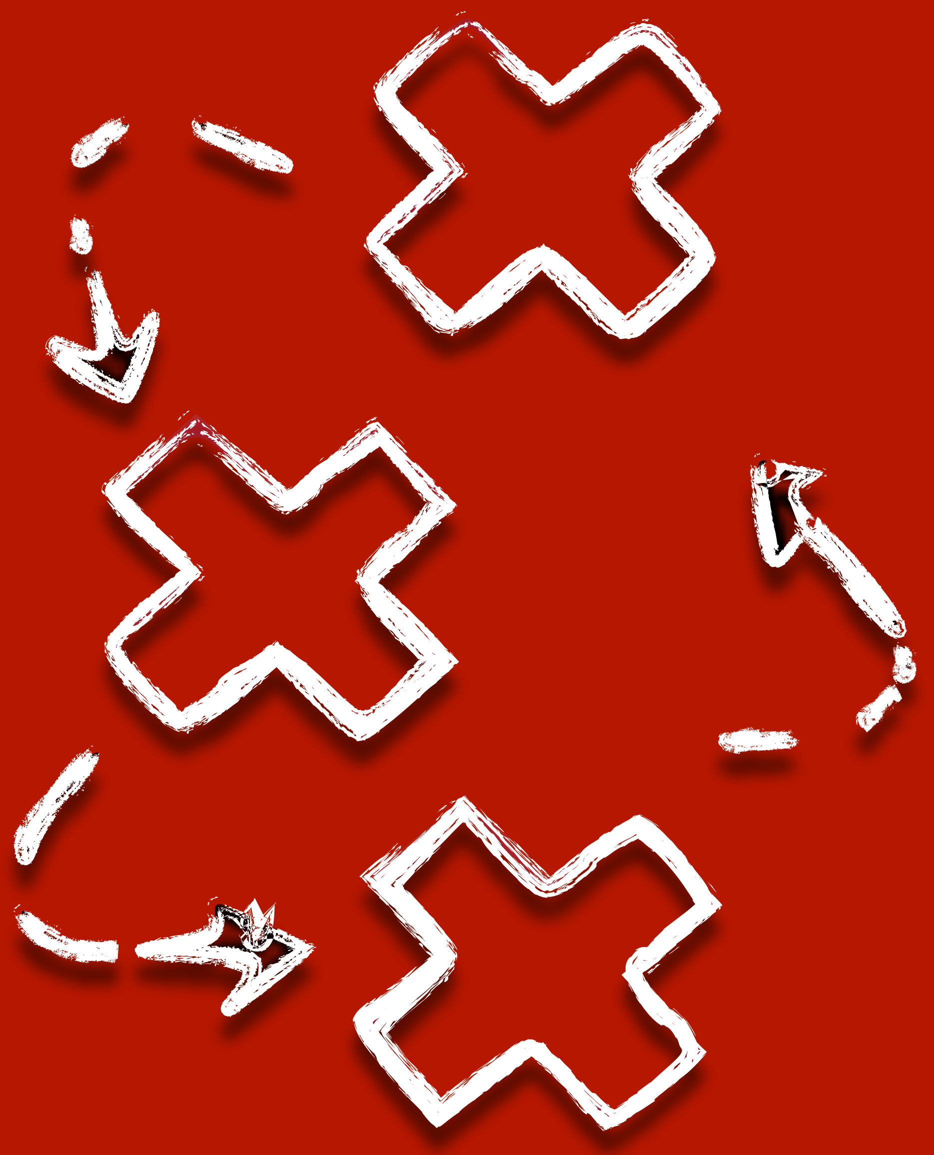


Seminar: Advanced topics in causality and causal ML research

Lecture 3: Causal Representation Learning

Lecturer: Sara Magliacane

UdS - Summer 2026

Last time: Causal discovery (structure learning) - simplest setting

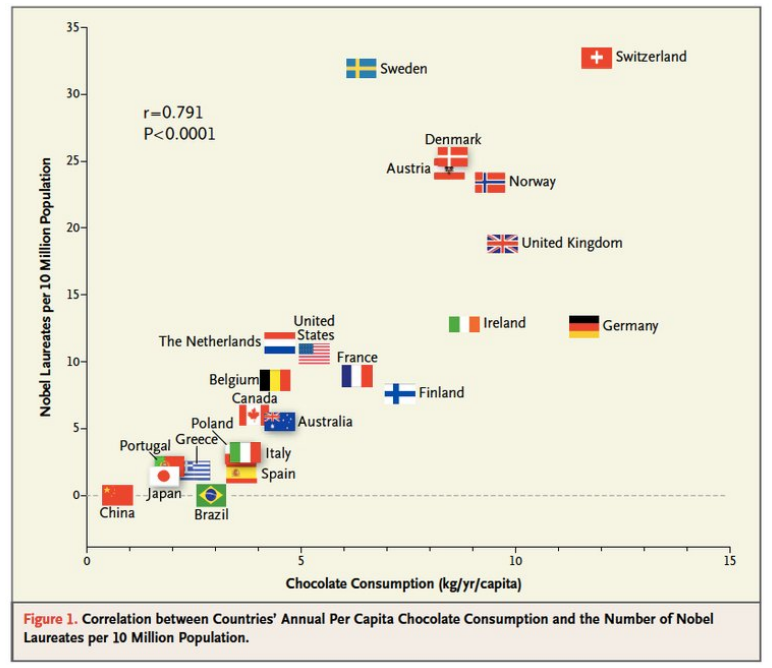


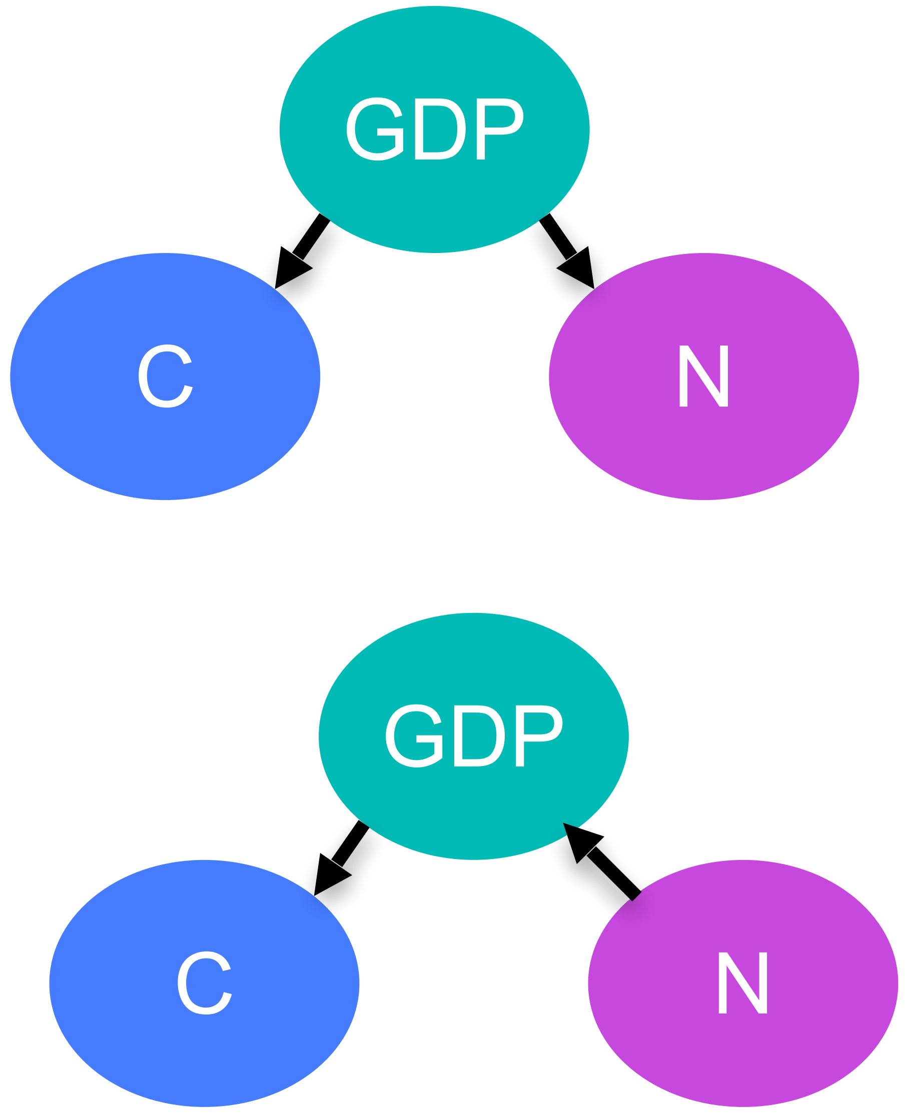
Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

C	N	GDP
4.5	5	33k
12	30	86k
10	20	46k
….	…	…

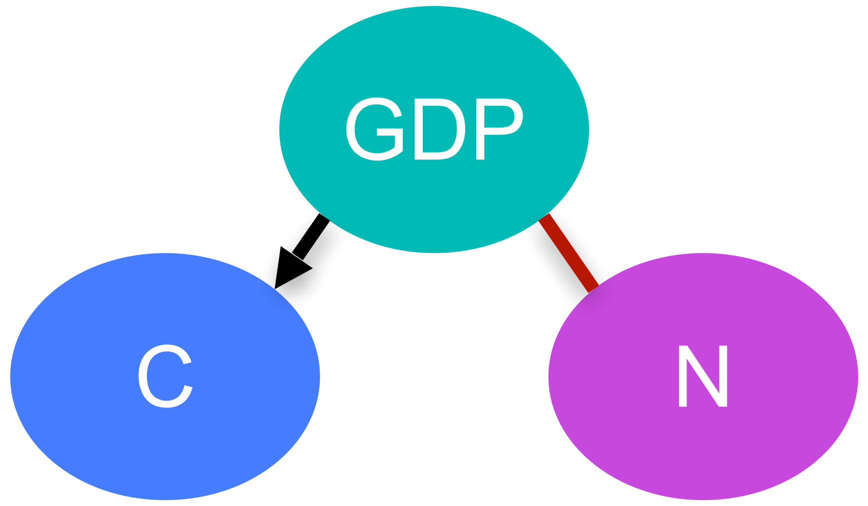
Observational data

$C \nrightarrow GDP$

[Optional] Background knowledge

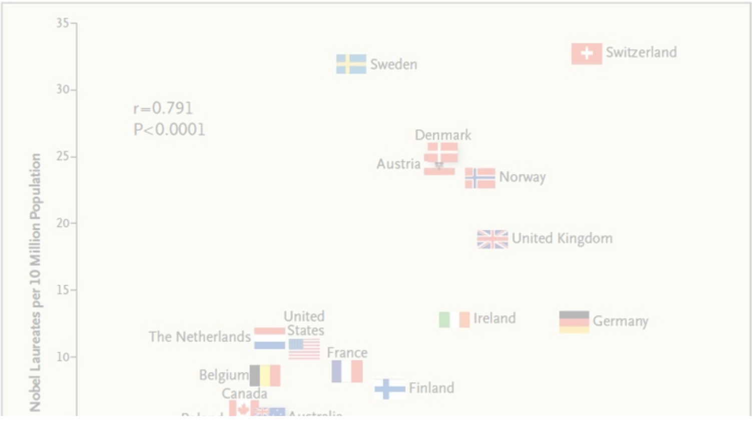


Sets of graphs that fit the data and background knowledge

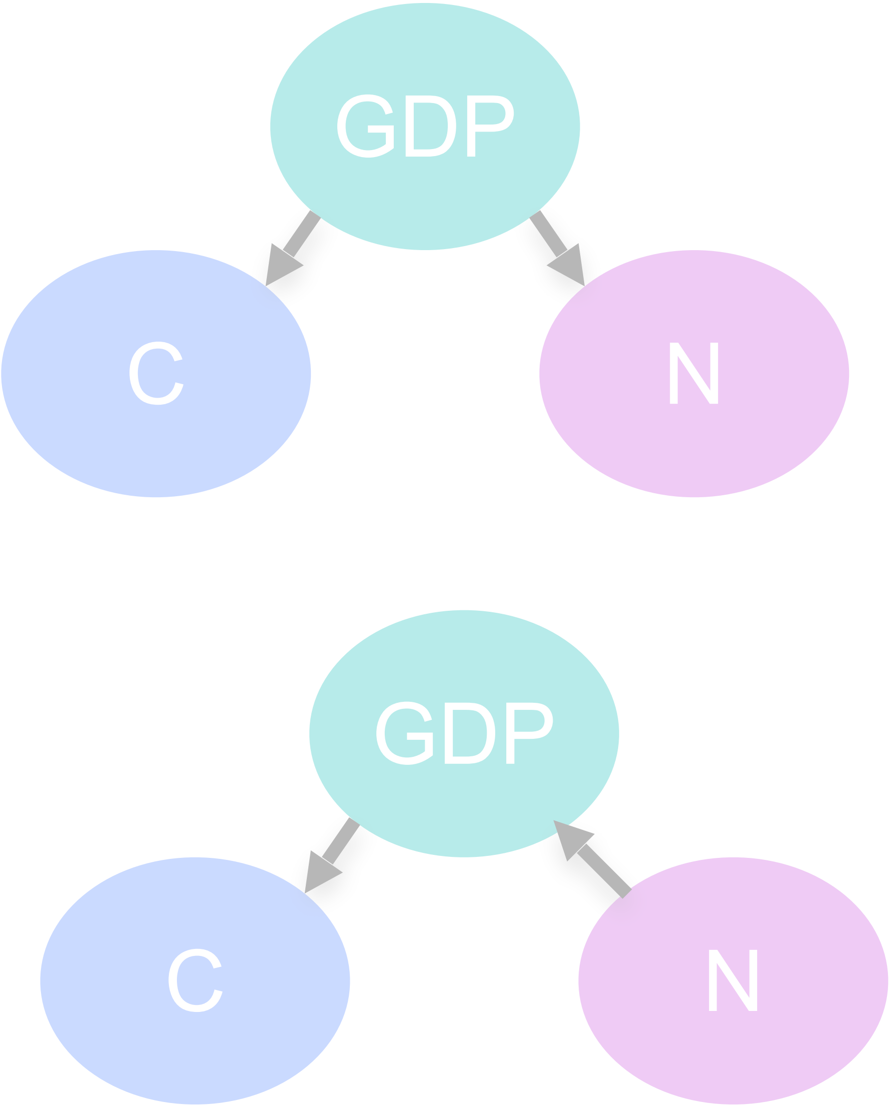


Summary graph

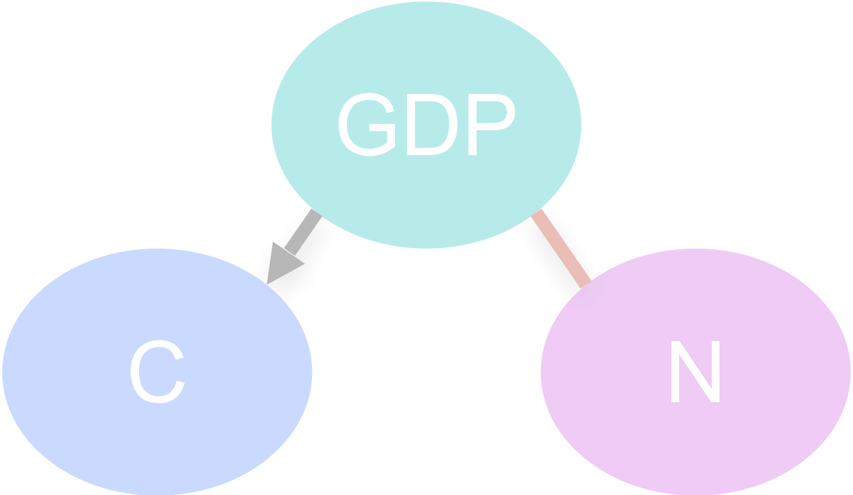
Causal Representation Learning



What if the causal variables are not directly observed (but we have high-dimensional observations, e.g. images)?

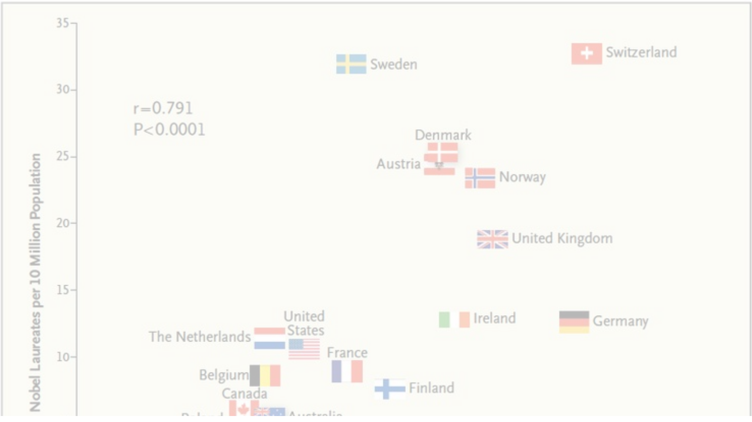


Sets of graphs that fit the data and background knowledge



Summary graph

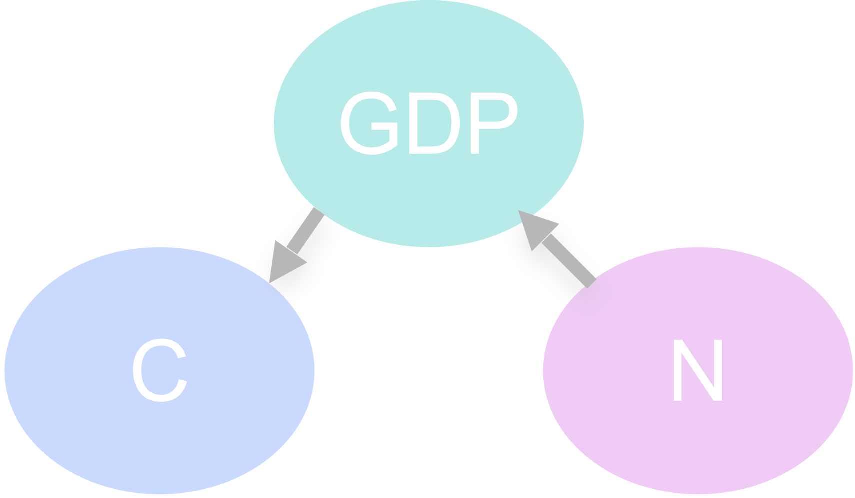
Causal Representation Learning



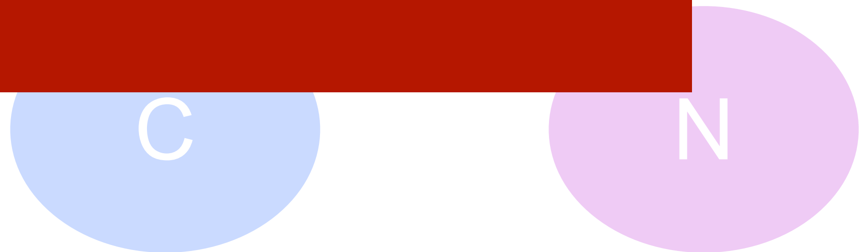
What if the causal variables are not directly observed (but we have high-dimensional observations, e.g. images)?

[Optio… …ledge

Task 1: identify/disentangle the causal variables from observations

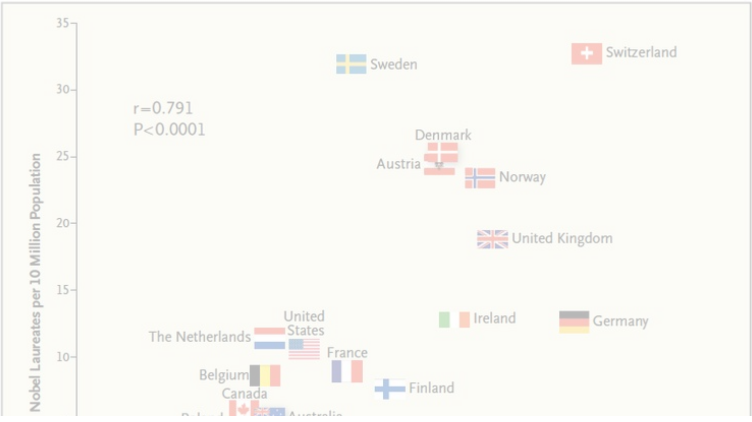


Sets of graphs that fit the data and background knowledge



Summary graph

Causal Representation Learning



What if the causal variables are not directly observed (but we have high-dimensional observations, e.g. images)?

Task 1: identify/disentangle the causal variables from observations

Task 2: learn causal relations between them from data (causal discovery)

Can we learn causal variables from high-dimensional data?

Towards Causal Representation Learning

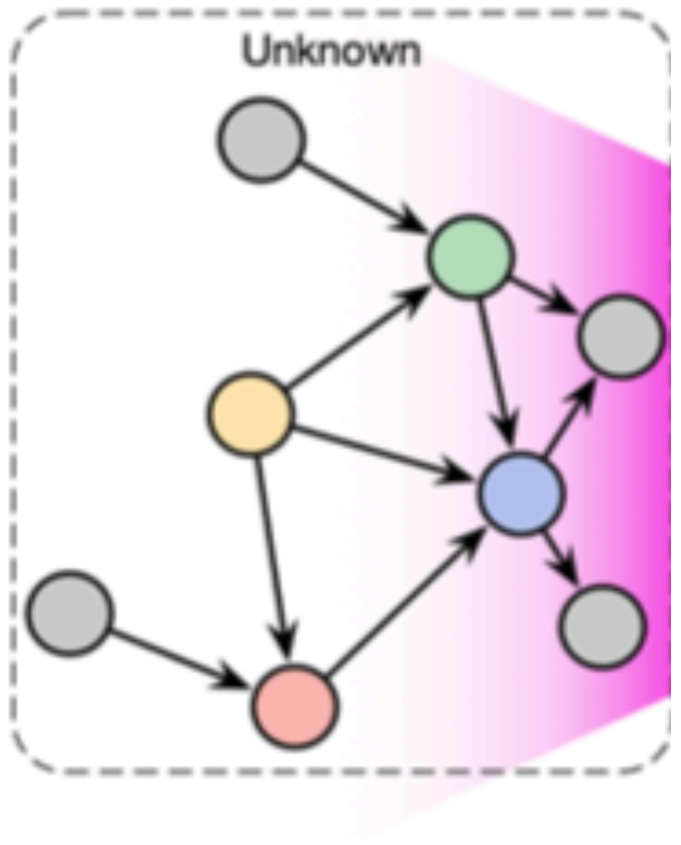
Bernhard Schölkopf †, Francesco Locatello †, Stefan Bauer *, Nan Rosemary Ke *, Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Abstract—**The two fields of machine learning and graphical causality arose and developed separately. However, there is now cross-pollination and increasing interest in both fields to benefit from the advances of the other. In the present paper, we review fundamental concepts of causal inference and relate them to crucial open problems of machine learning, including transfer and generalization, thereby assaying how causality can contribute to modern machine learning research. This also applies in the opposite direction: we note that most work in causality starts from the premise that the causal variables are given. A central problem for AI and causality is, thus, causal representation learning, the discovery of high-level causal variables from low-level observations. Finally, we delineate some implications of causality for machine learning and propose key research areas at the intersection of both communities.**

et al., 2018], and speech recognition [Graves et al., 2013], a substantial body of literature explored the robustness of the prediction of state-of-the-art deep neural network architectures. The underlying motivation originates from the fact that in the real world there is often little control over the distribution from which the data comes from. In computer vision [Geirhos et al., 2018, Shetty et al., 2019], changes in the test distribution may, for instance, come from aberrations like camera blur, noise or compression quality [Hendrycks and Dietterich, 2019, Karahan et al., 2016, Michaelis et al., 2019, Roy et al., 2018], or from shifts, rotations, or viewpoints [Azulay and Weiss, 2019, Barbu et al., 2019, Engstrom et al., 2017, Zhang, 2019]. Motivated by this, new benchmarks were proposed to

https://arxiv.org/abs/2102.11107

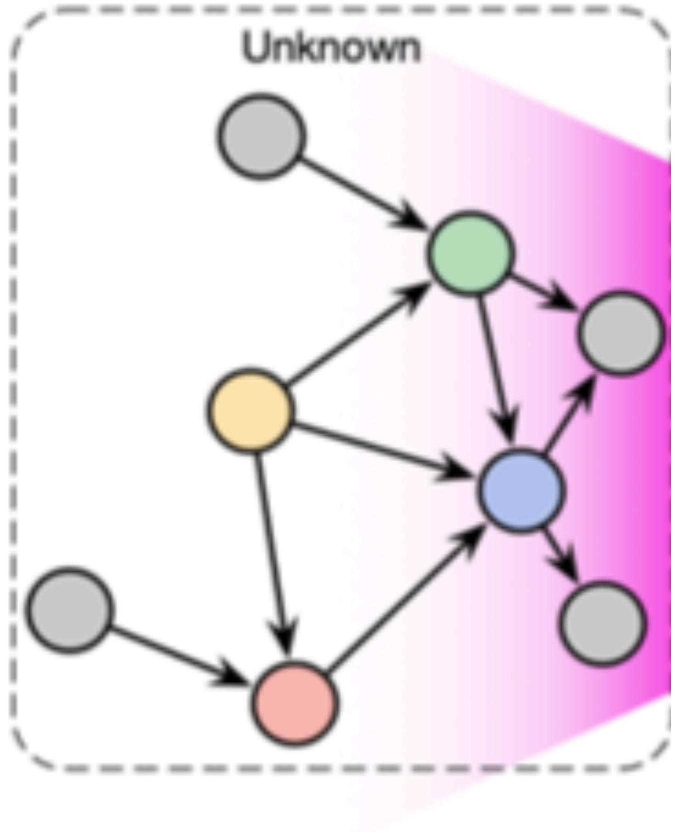
The causal representation learning problem



Unknown causal graph over unknown causal variables

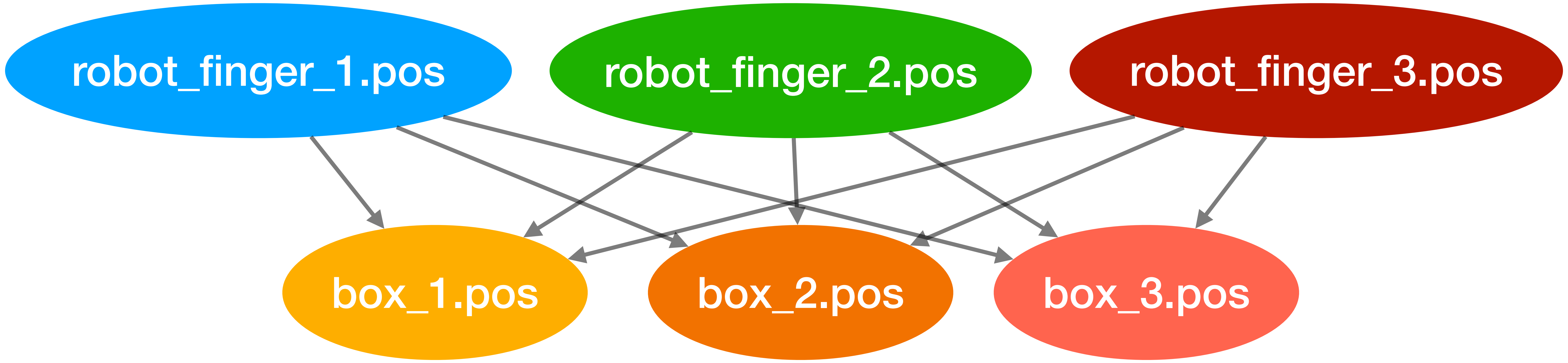
$Z_1, \ldots, Z_d$

The causal representation learning problem

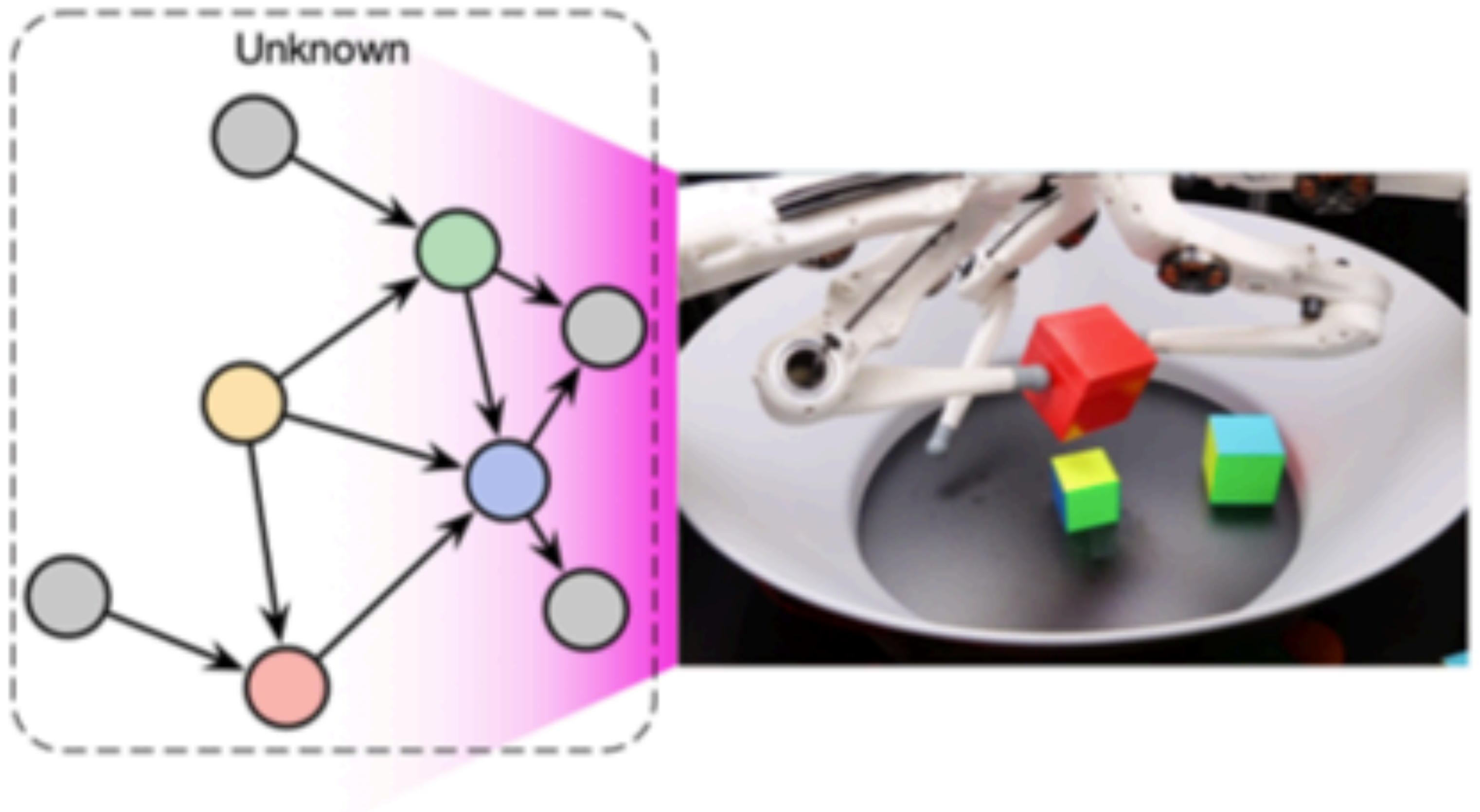


Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$



The causal representation learning problem (general version)



Unknown causal graph over unknown causal variables

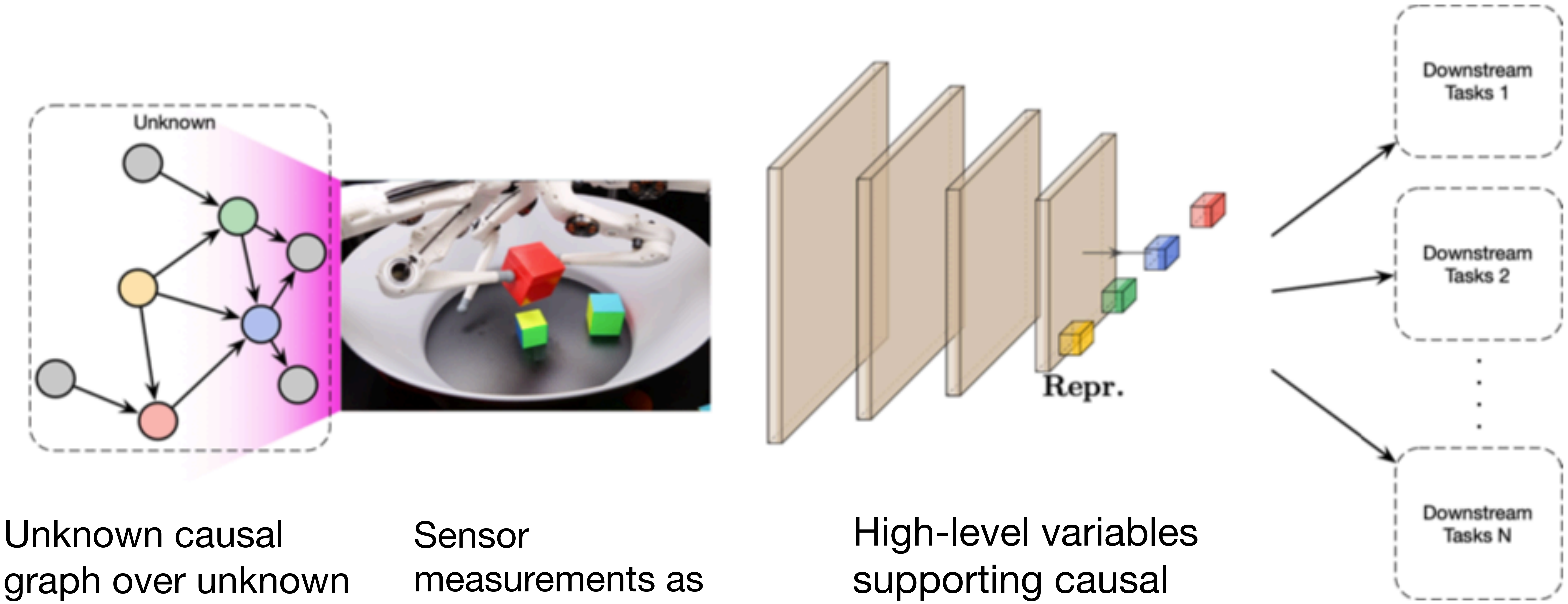
Sensor measurements as an entangled view

$Z_1, \ldots, Z_d$

$X_1, \ldots, X_p$

$p \gg d$

The causal representation learning problem (general version)



Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$
$p \gg d$

High-level variables supporting causal statements

$S_1, \ldots, S_n$

The causal representation learning problem (simplified)

Mixing function **Encoder** **Decoder**

Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$

$p \gg d$

Causal graph over reconstructed causal variables

$\hat{Z}_1, \ldots, \hat{Z}_n$

Reconstructed sensor measurements

$\hat{X}_1, \ldots, \hat{X}_p$

$p \gg d$

The causal representation learning problem (simplified)

Mixing function **Encoder** **Decoder**

Unknown causal graph over unknown causal variables

$Z_1, \ldots, Z_d$

Sensor measurements as an entangled view

$X_1, \ldots, X_p$

$p \gg d$

Causal graph over reconstructed causal variables

$\hat{Z}_1, \ldots, \hat{Z}_n$

Reconstructed up to:

- **Component-wise transformation**
- **Permutation**

Reconstructed sensor measurements

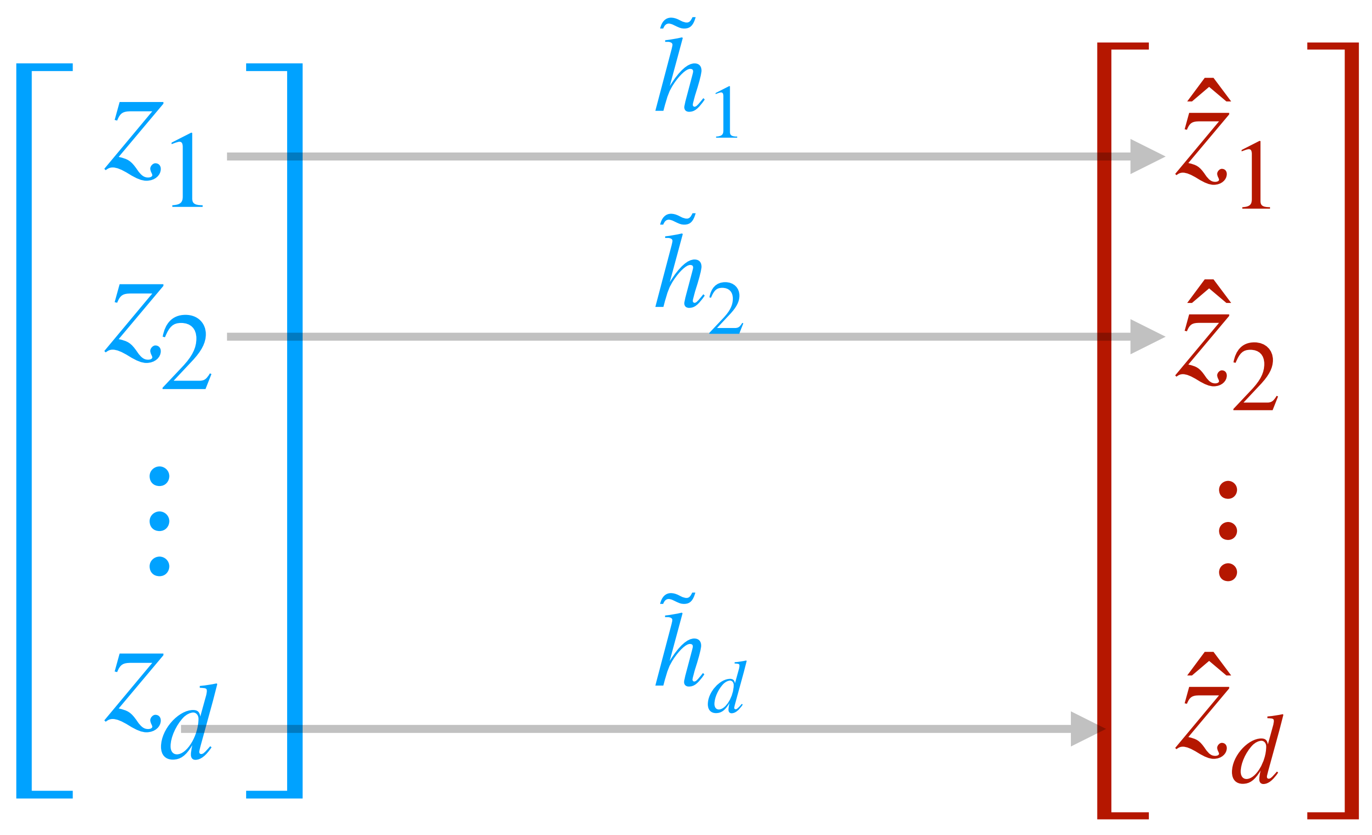
$\hat{X}_1, \ldots, \hat{X}_p$

$p \gg d$

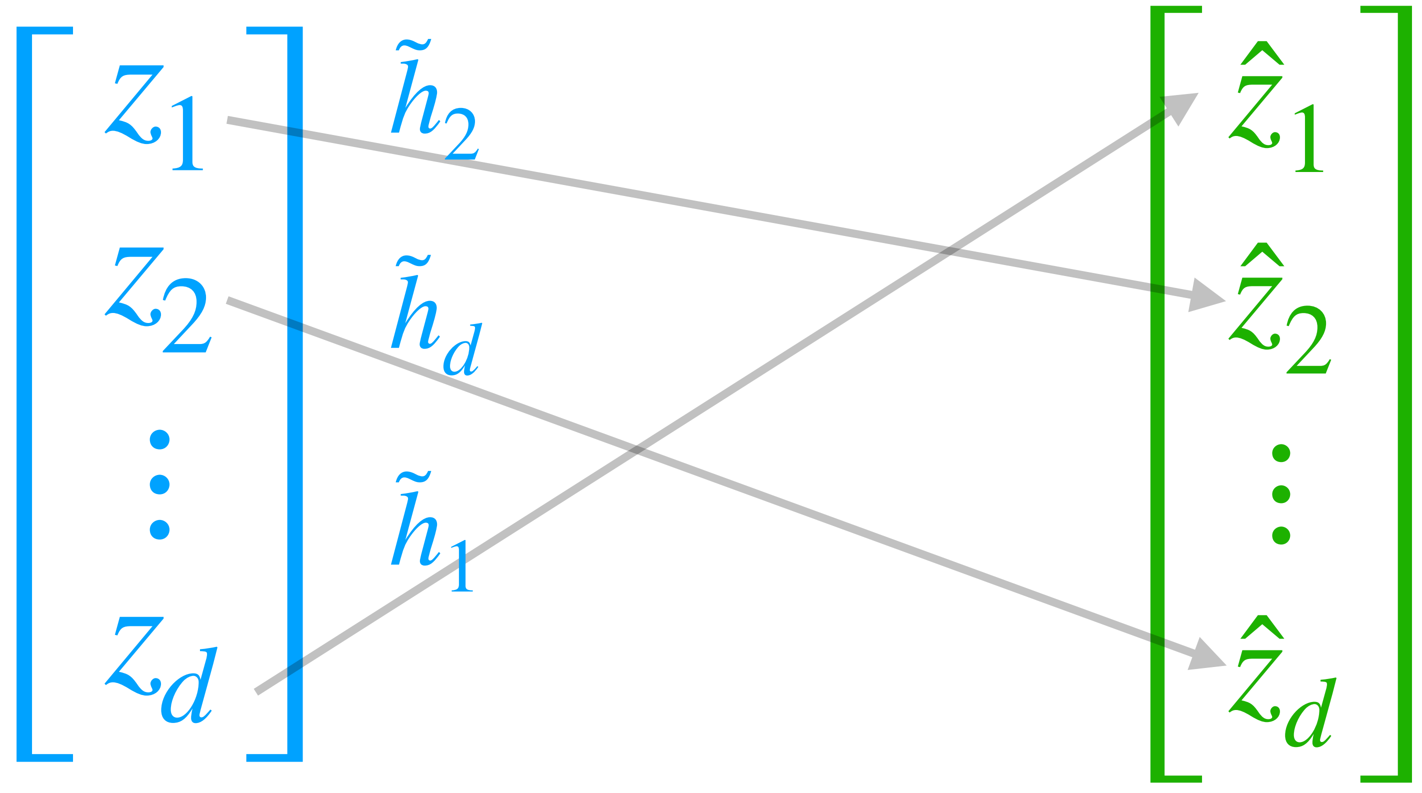
How well can we actually recover these causal variables?

- **A lot of work in CRL** focuses on **theoretical guarantees** in learning high-level causal variables from low-level observations, under different assumptions
- In general without any supervision, we cannot identify the exact causal variables, but we can **identify** them **up to an equivalence class**

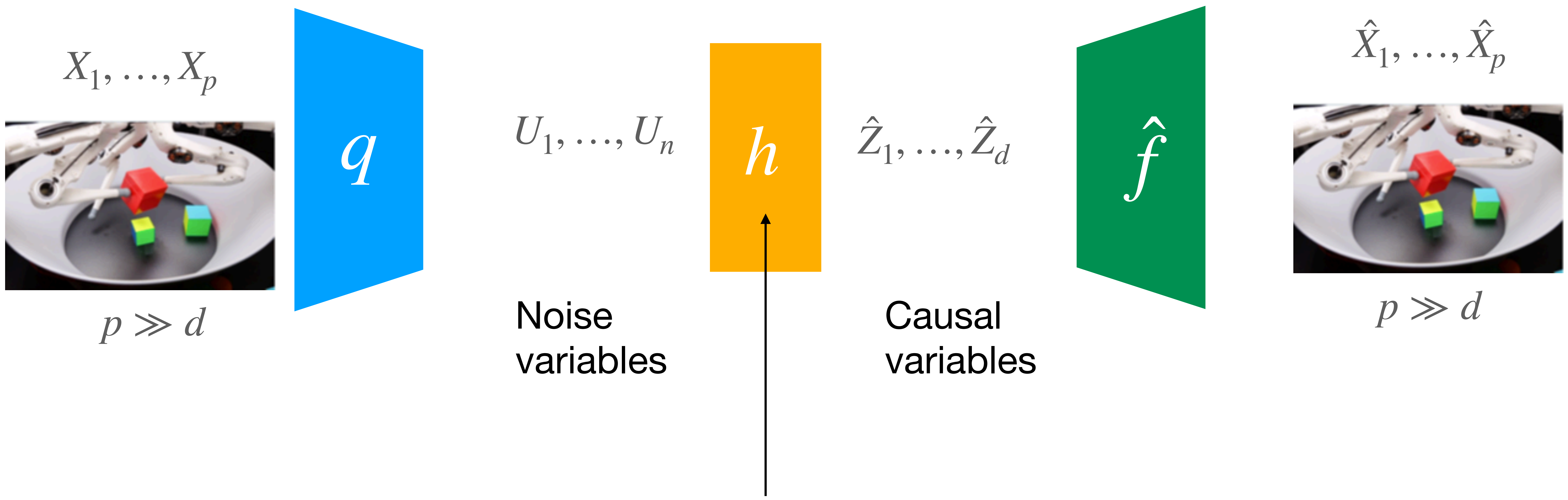
Identifiability up to component-wise transformations



Identifiability up to permutation and component-wise transformations



Causal representation learning with additional assumptions



$$\forall i = 1,\ldots,d : \hat{Z}_i = h_i(Z_{Pa_G(i)}, U_i)$$

Known structural causal model

Assumption: no latent confounders

[Leeb et al. 2020]

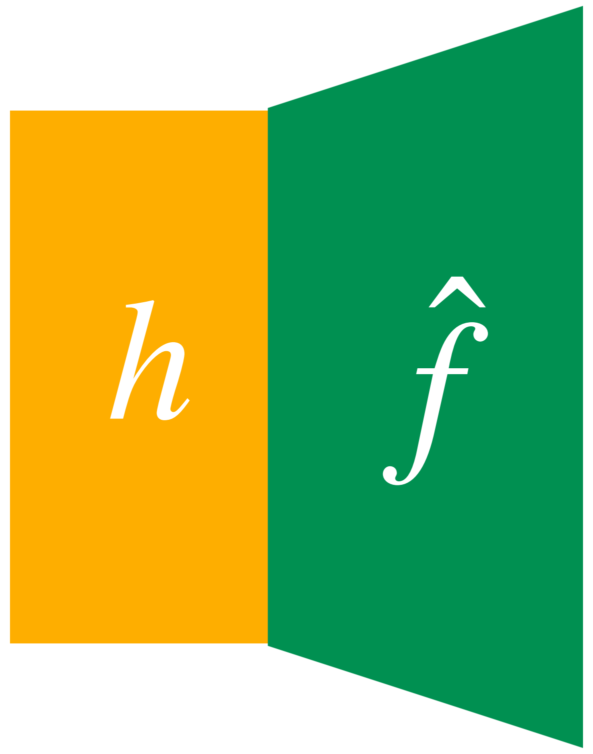
Causal representation learning **without** additional assumptions

$X_1, \ldots, X_p$

$p \gg d$

$U_1, \ldots, U_n$

Noise variables



Cannot be learnt separately

$\hat{X}_1, \ldots, \hat{X}_p$

$p \gg d$

Unknown structural causal model

In general ill-defined, infinitely many solutions

Only one causal

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions/actions

Strategy 4: assume temporal data

- Unknown vs known targets
- Perfect vs soft interventions

- Sufficient variability

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions/actions

Strategy 4: assume temporal data

Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

Output 3: Up to differentiable component-wise transformations + permutation

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions/actions

Strategy 4: assume temporal data

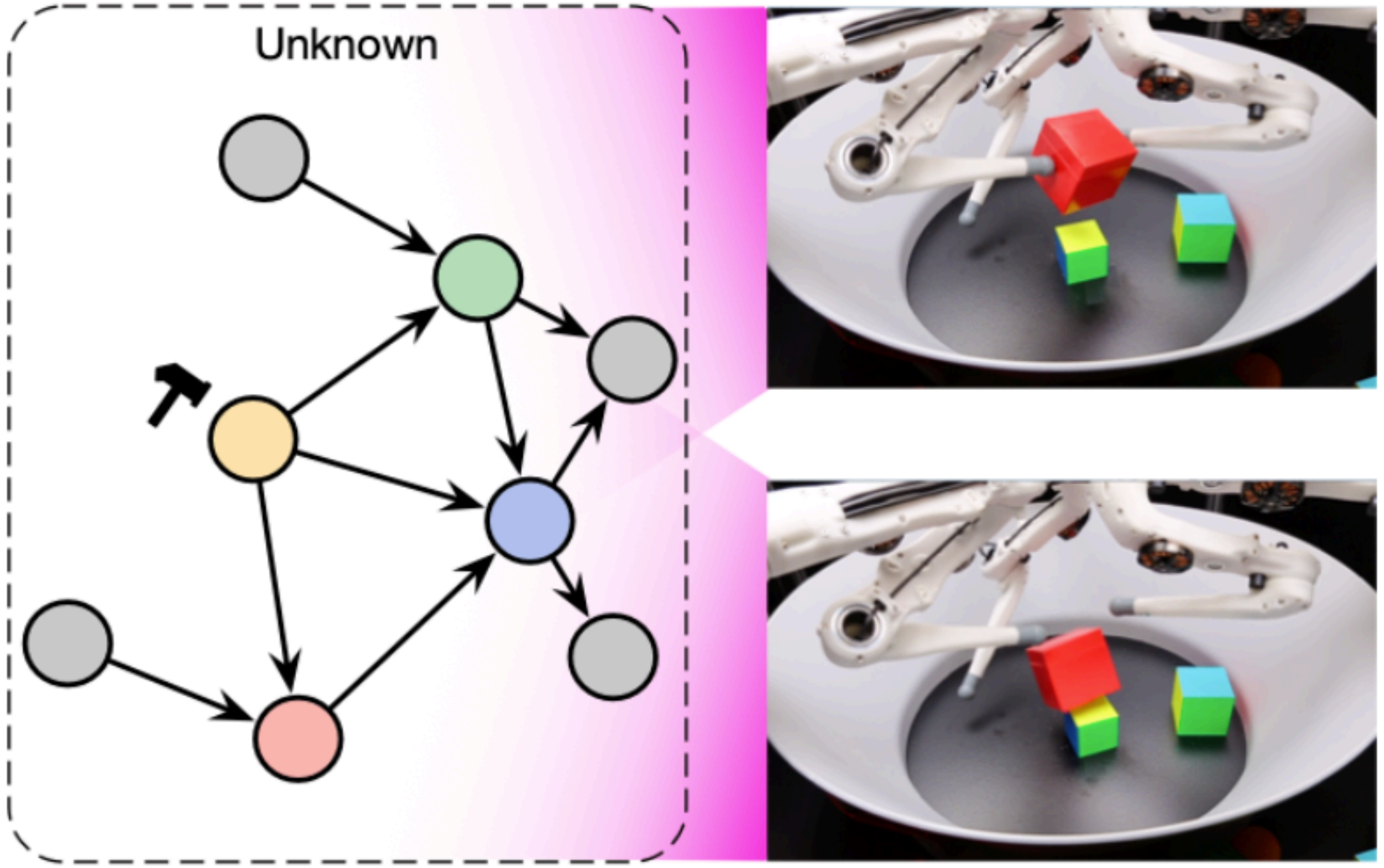
Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

Output 3: Up to differentiable component-wise transformations + permutation

Causal representation learning with additional assumptions

What if we can perform actions or observe different environments?



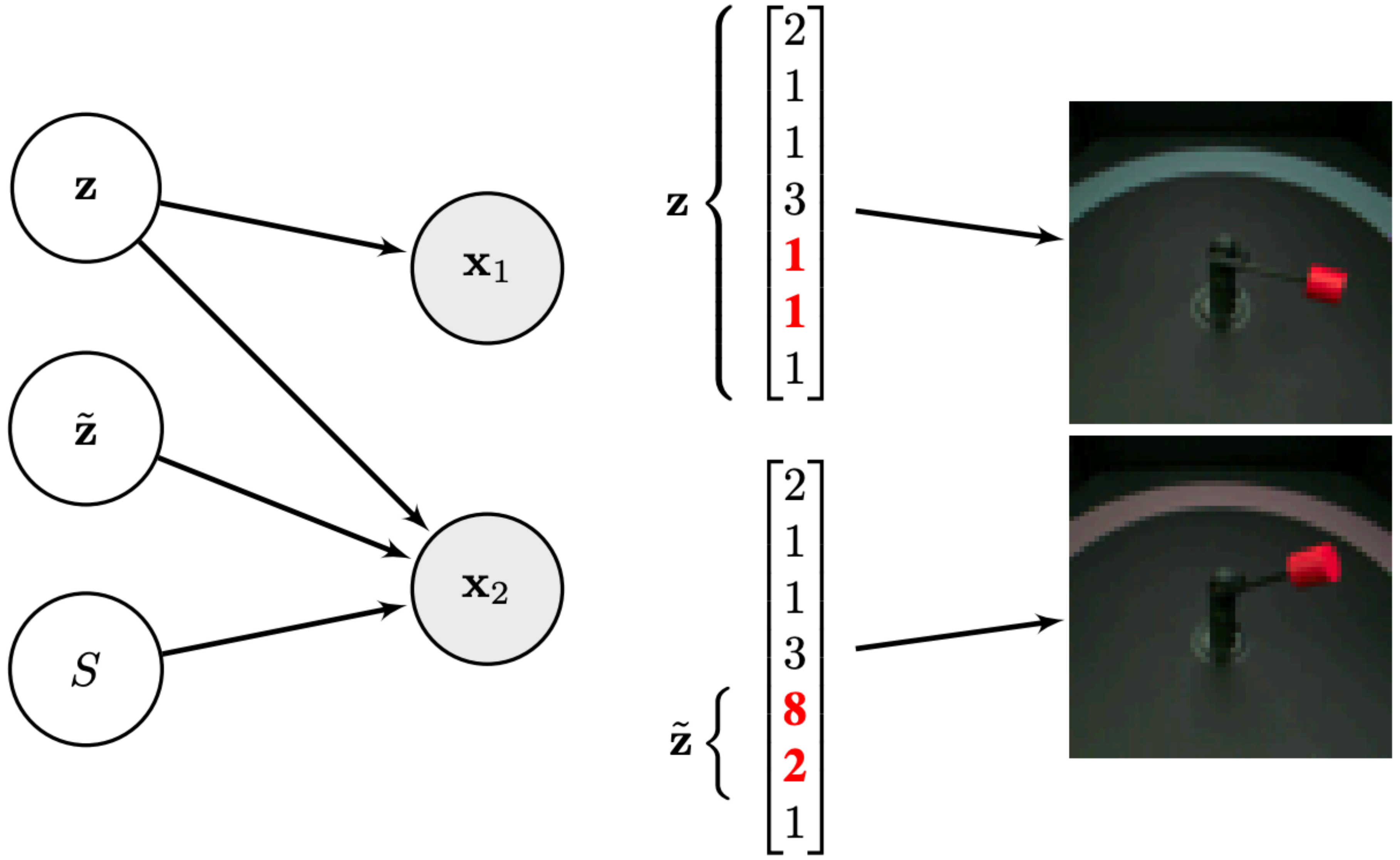
Sparse Mechanism Shift example: change in one causal variable for the position of the finger (sparse) vs. change in the image (distributed)

Weakly-Supervised Disentanglement Without Compromises

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, Michael Tschannen

- **Setting**: we observe pairs of observations x_1 and x_2 that represent the same system with several fixed causal factors

- z is the original state
- $\tilde{z}$ represents the changed values
- S selects the unchanged dimensions



https://arxiv.org/abs/2002.02886

Weakly-Supervised Disentanglement Without Compromises

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, Michael Tschannen

- **Main result**: If z has a continuous distribution, the mixing function is a diffeomorphism, we know the number of fixed causal variables $|S|$ and:
  - If for every $i \in 1,\ldots,d$, we have $S, S' \sim p(S)$ s.t. $P(S \cap S' = \{i\}) > 0$

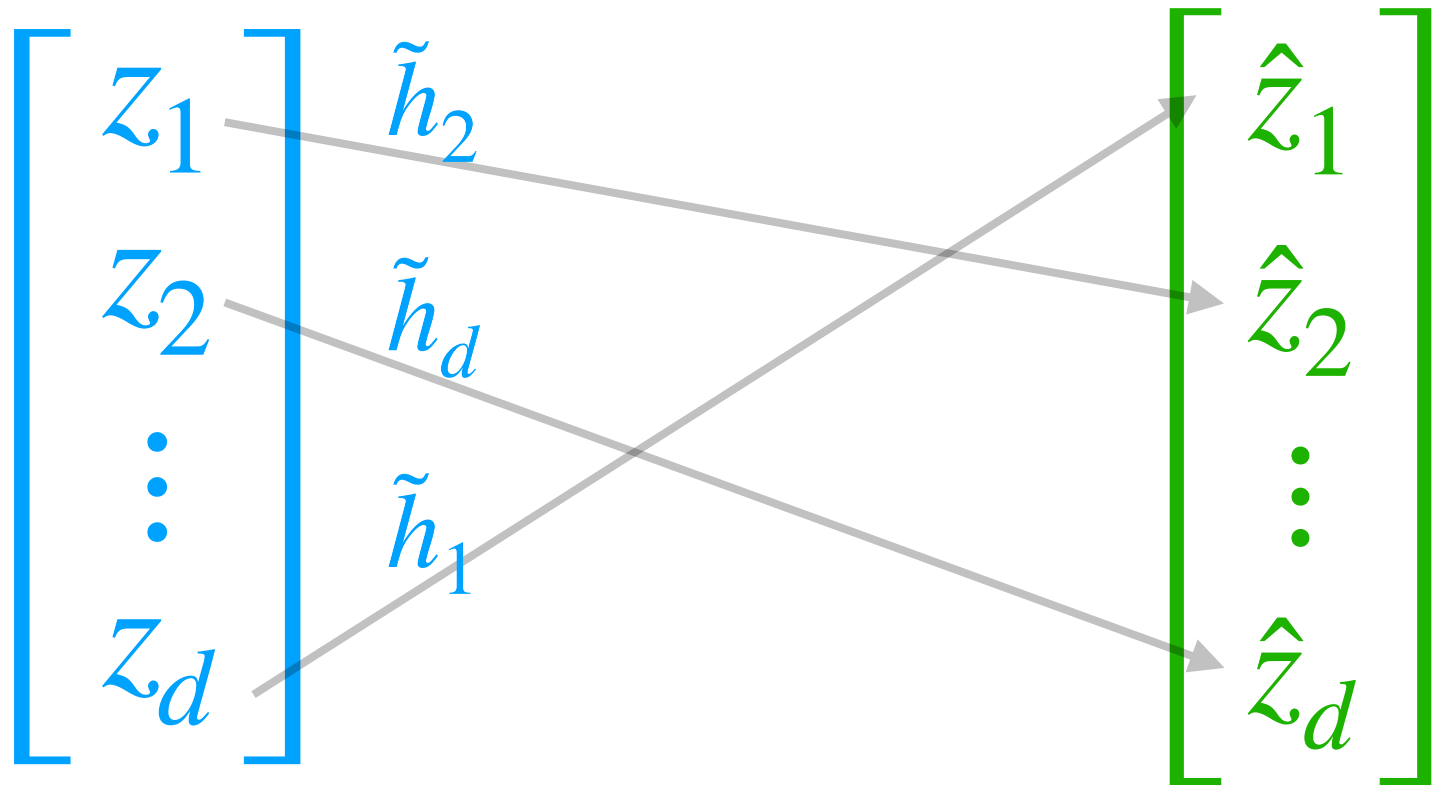
S	2	2		2	1
	1	1		1	2
	1	1		3	1
	1	3		4	4
	1	8	S'	5	5
	1	2		3	3
	1	1		2	2

https://arxiv.org/abs/2002.02886

Weakly-Supervised Disentanglement Without Compromises

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, Michael Tschannen

- **Main result**: If z has a continuous distribution, the mixing function is a diffeomorphism, we know the number of fixed causal variables $|S|$ and:
  - If for every $i \in 1,\ldots,d$, we have $S, S' \sim p(S)$ s.t. $P(S \cap S' = \{i\}) > 0$
  - **Then:** what we learn with a standard reconstruction error on the pairs are the true factors **up to component-wise transformations and permutation**



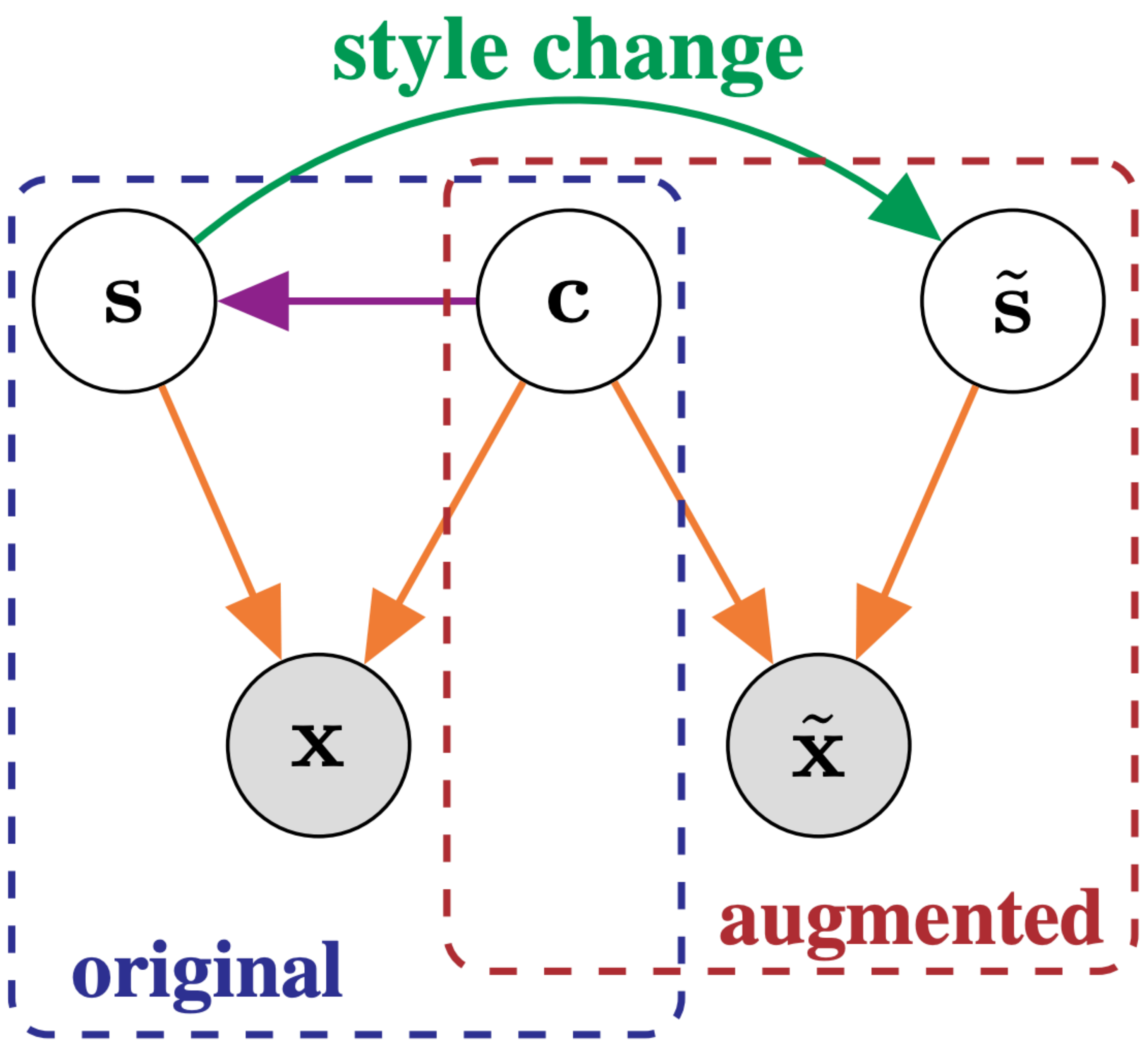
https://arxiv.org/abs/2002.02886

Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, Francesco Locatello

- Similar setting, but separate z in blocks of content c and style $s, \tilde{s}$
- **Causal factors are not independent anymore**

- **Main results:**
  - Block identification
    - We can identify the blocks of s and c, but not individual variables
  - Contrastive learning with InfoNCE asymptotically isolates content

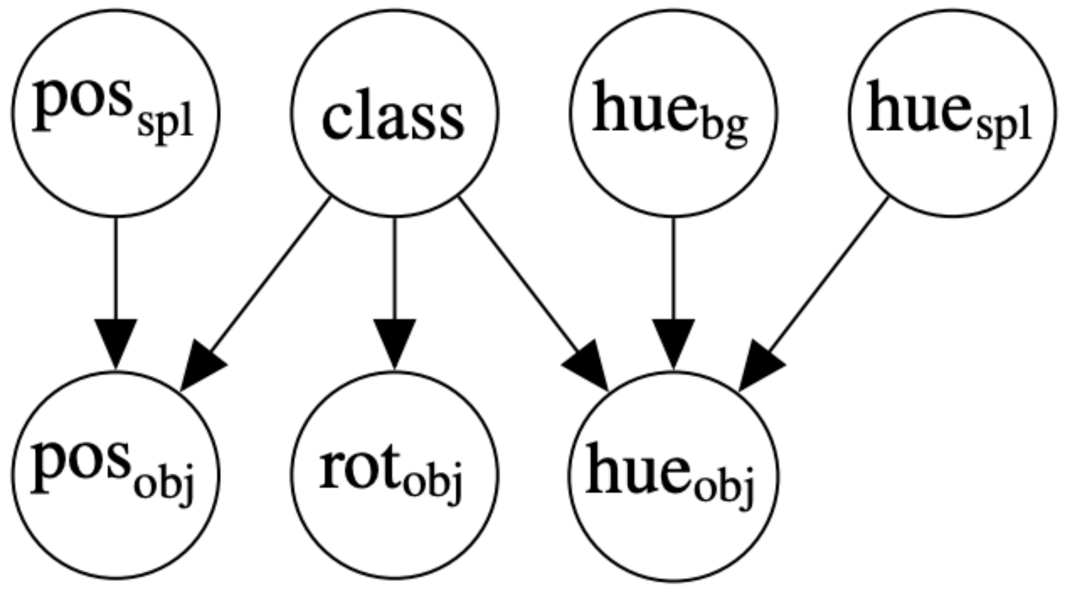


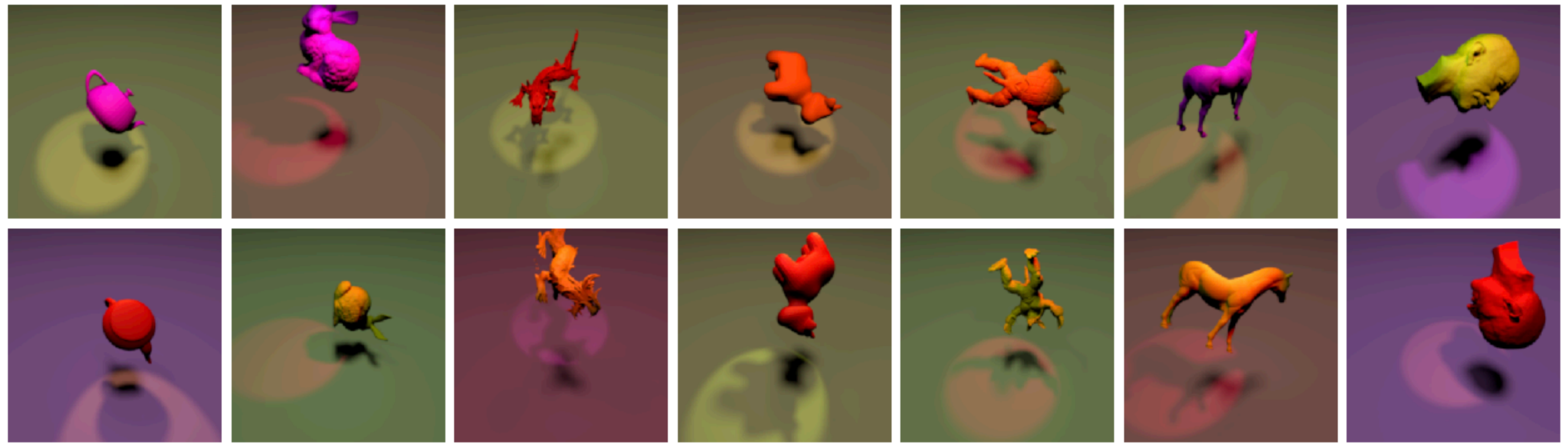
https://arxiv.org/abs/2106.04619

Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, Francesco Locatello

- **Experiments**: Causal3DIdent benchmark





object class	$\mu(\mathrm{hue})$
Teapot	0
Hare	$\frac{\mathrm{hue}_{\mathrm{bg}}+\mathrm{hue}_{\mathrm{spl}}}{2}$
Dragon	$-\frac{\mathrm{hue}_{\mathrm{bg}}+\mathrm{hue}_{\mathrm{spl}}}{2}$
Cow	-0.35
Armadillo	0.7
Horse	-0.7
Head	0.35

object class	$\mu(x)$	$\mu(y)$	$\mu(z)$
Teapot	0	0	0
Hare	$-\sin(\mathrm{pos}_{\mathrm{spl}})$	$-\cos(\mathrm{pos}_{\mathrm{spl}})$	0
Dragon	$-\sin(\mathrm{pos}_{\mathrm{spl}})$	$-\cos(\mathrm{pos}_{\mathrm{spl}})$	0
Cow	$\sin(\mathrm{pos}_{\mathrm{spl}})$	$\cos(\mathrm{pos}_{\mathrm{spl}})$	0
Armadillo	$\sin(\mathrm{pos}_{\mathrm{spl}})$	$\cos(\mathrm{pos}_{\mathrm{spl}})$	0
Horse	$-\sin(\mathrm{pos}_{\mathrm{spl}})$	$-\cos(\mathrm{pos}_{\mathrm{spl}})$	0
Head	$\sin(\mathrm{pos}_{\mathrm{spl}})$	$\cos(\mathrm{pos}_{\mathrm{spl}})$	0

https://arxiv.org/abs/2106.04619

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions/actions

Strategy 4: assume temporal data

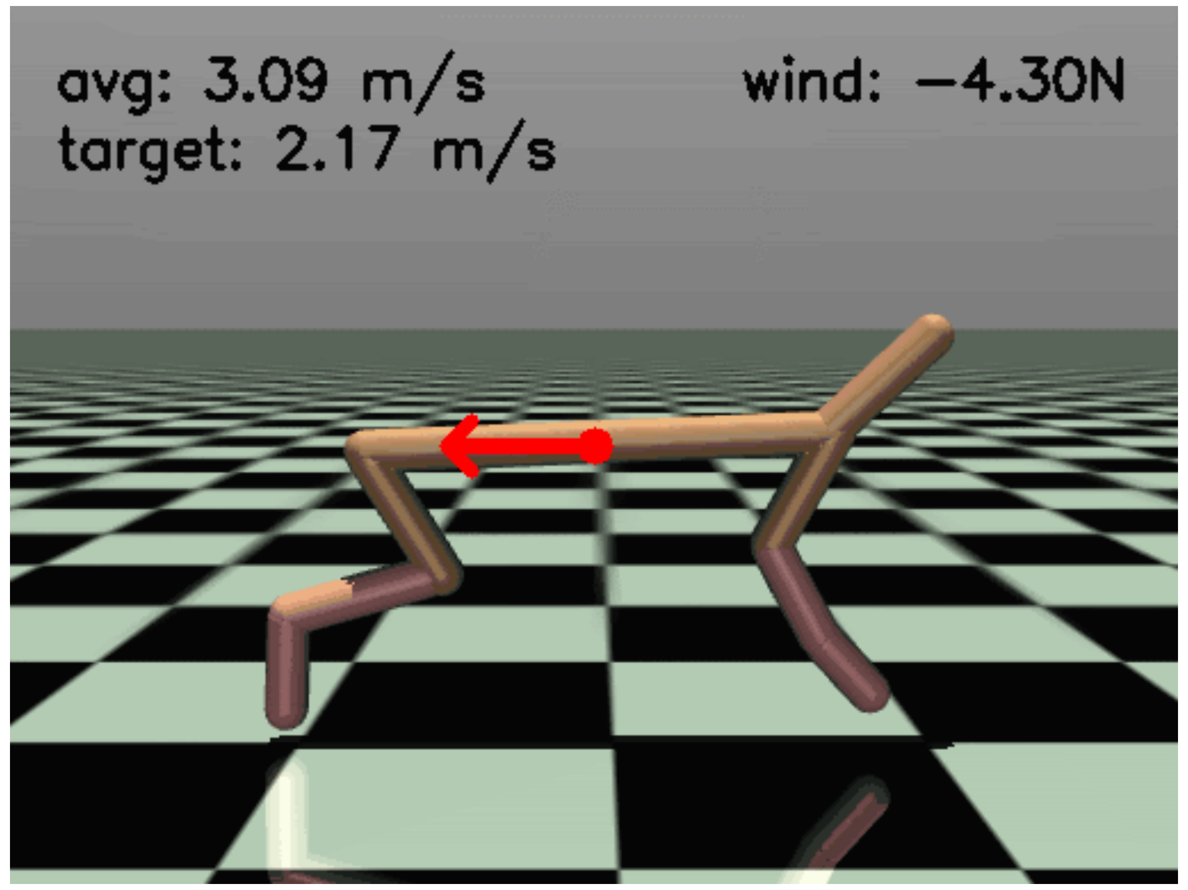
Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

Output 3: Up to differentiable component-wise transformations + permutation

CRL in temporal settings with actions

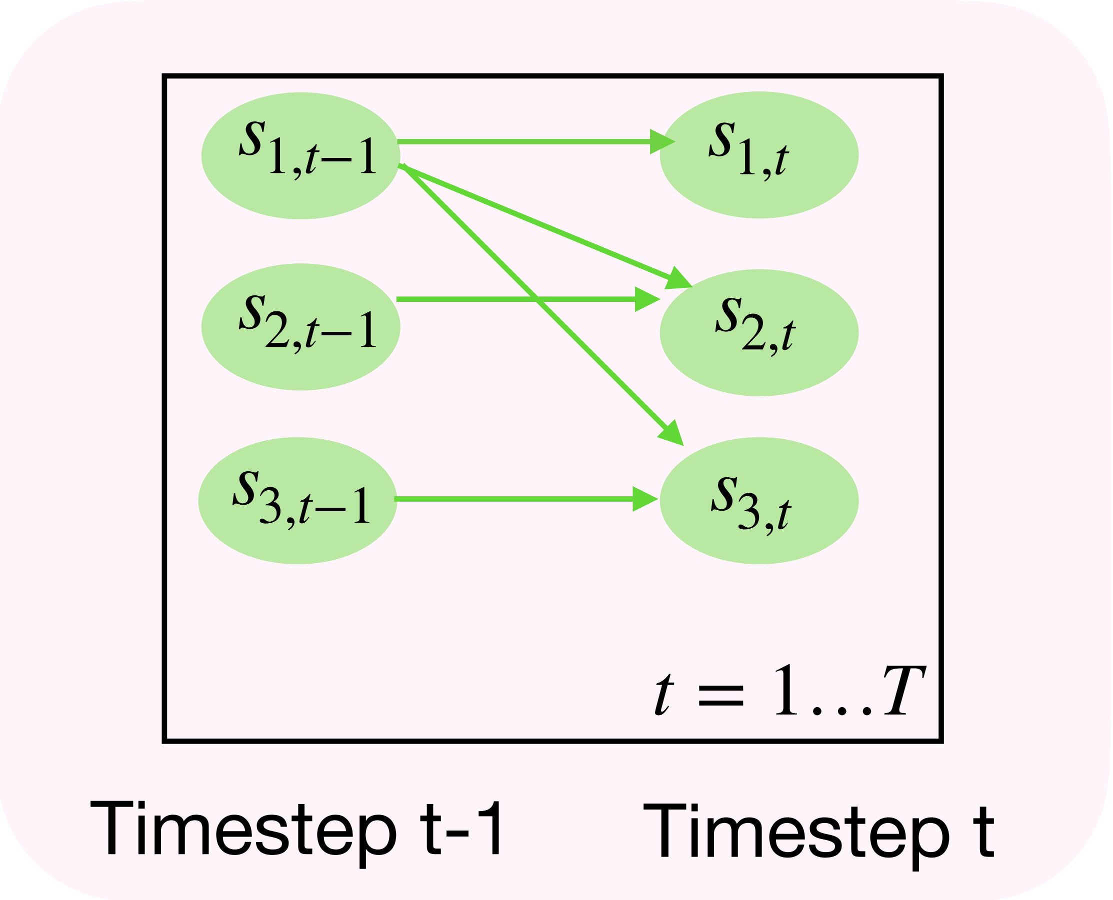
- Natural setting for learning from interventions/actions: “before” and “after”
  - E.g. sequential decision making, RL, planning, robotics, …



- Often we want to extract **semantic features from images** in an unsupervised way
  - **Causal representation learning (CRL) -** learn high level causal variables and causal relations between them from low level observations

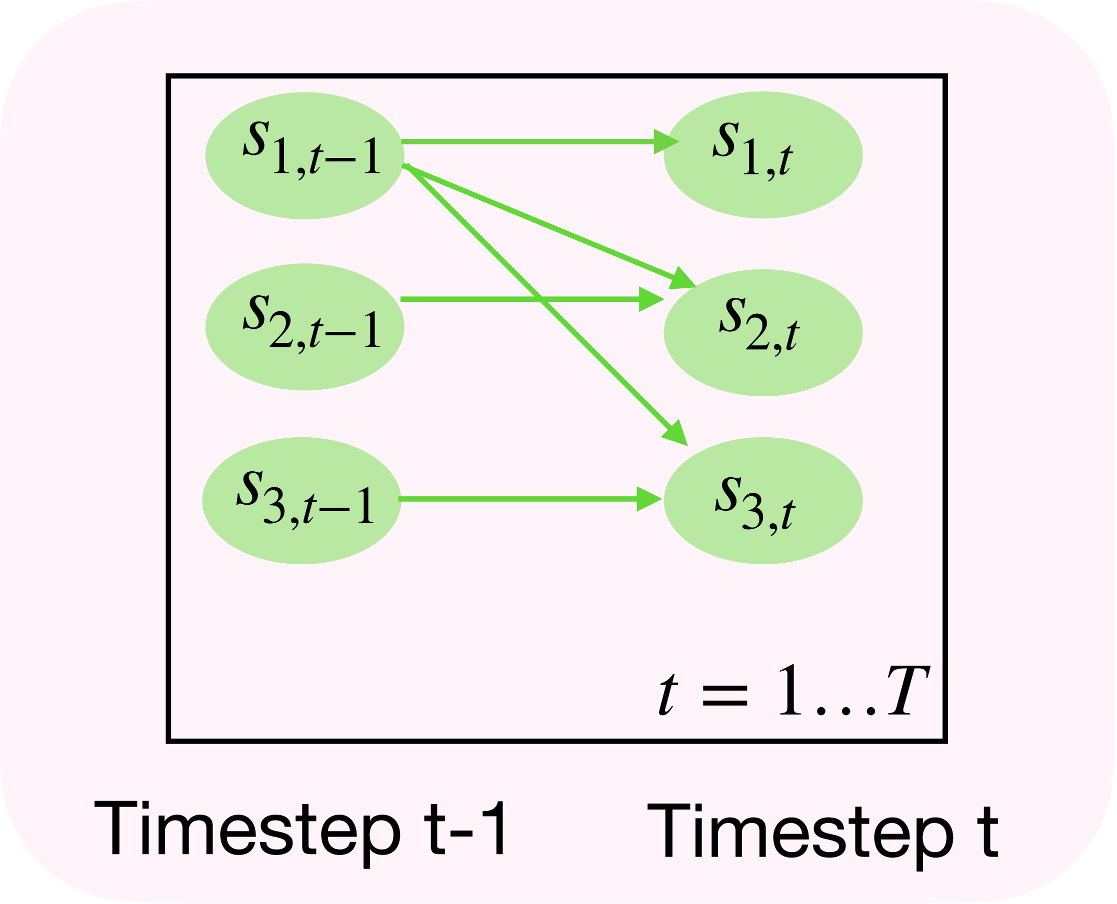
Relic from a bygone age - Dynamic Bayesian Networks

- Extension of BNs to temporal settings, also Two-Slice BNs - a type of **template PGM**
- Classic assumptions for DBNs:
  - **1-Markov assumption**: only vars from t-1 (1 timestep back) can cause vars at t
  - **Stationarity/Time invariance:** the transition model (edges) are the same across pairs of time steps

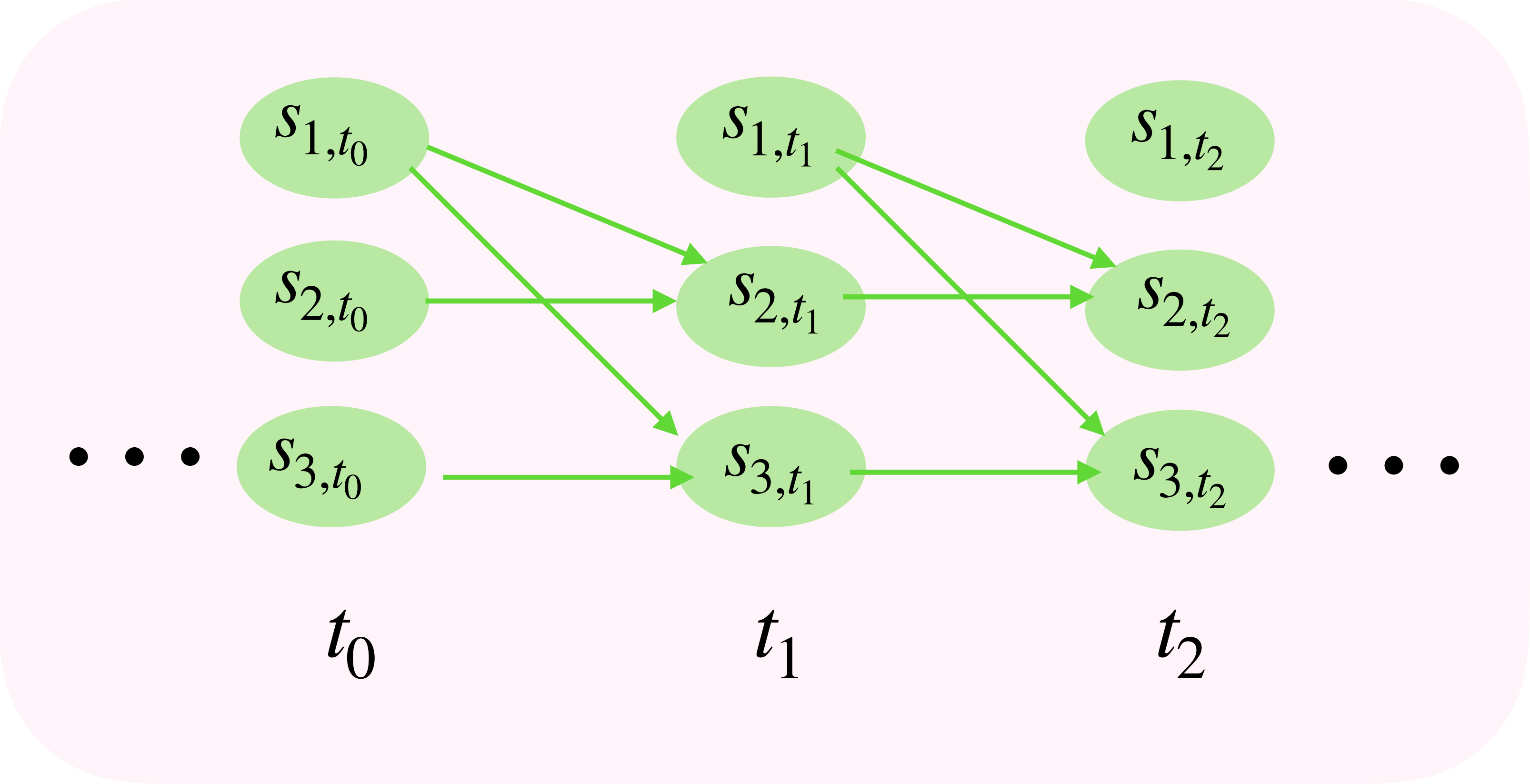


Relic from a bygone age - Dynamic Bayesian Networks

- Extension of BNs to temporal settings, also Two-Slice BNs - a type of **template PGM**
- Classic assumptions for DBNs:
  - **1-Markov assumption**: only vars from t-1 (1 timestep back) can cause vars at t
  - **Stationarity/Time invariance:** the transition model (edges) are the same across pairs of time steps

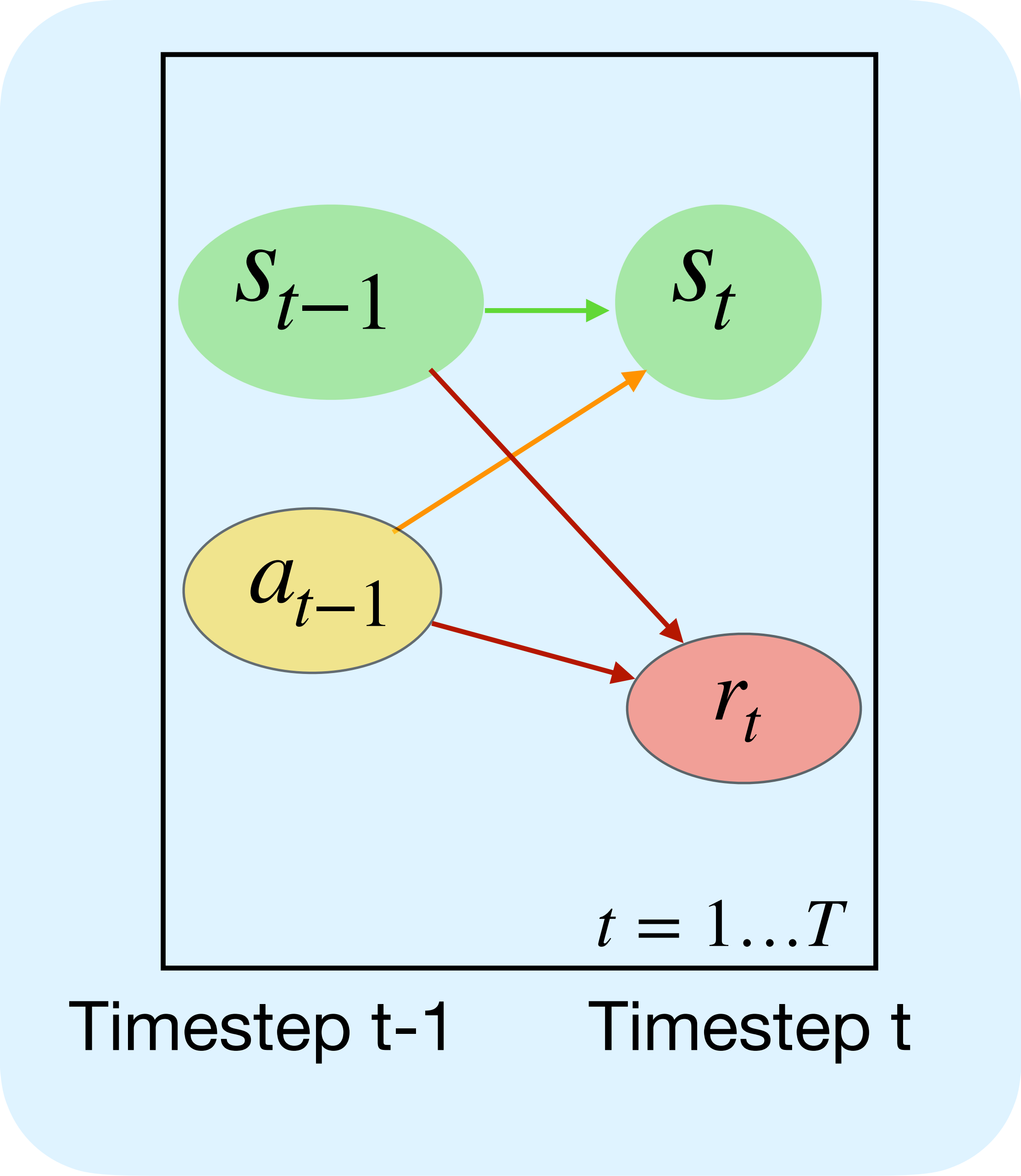


$\equiv$



MDPs vs Factored MDPs

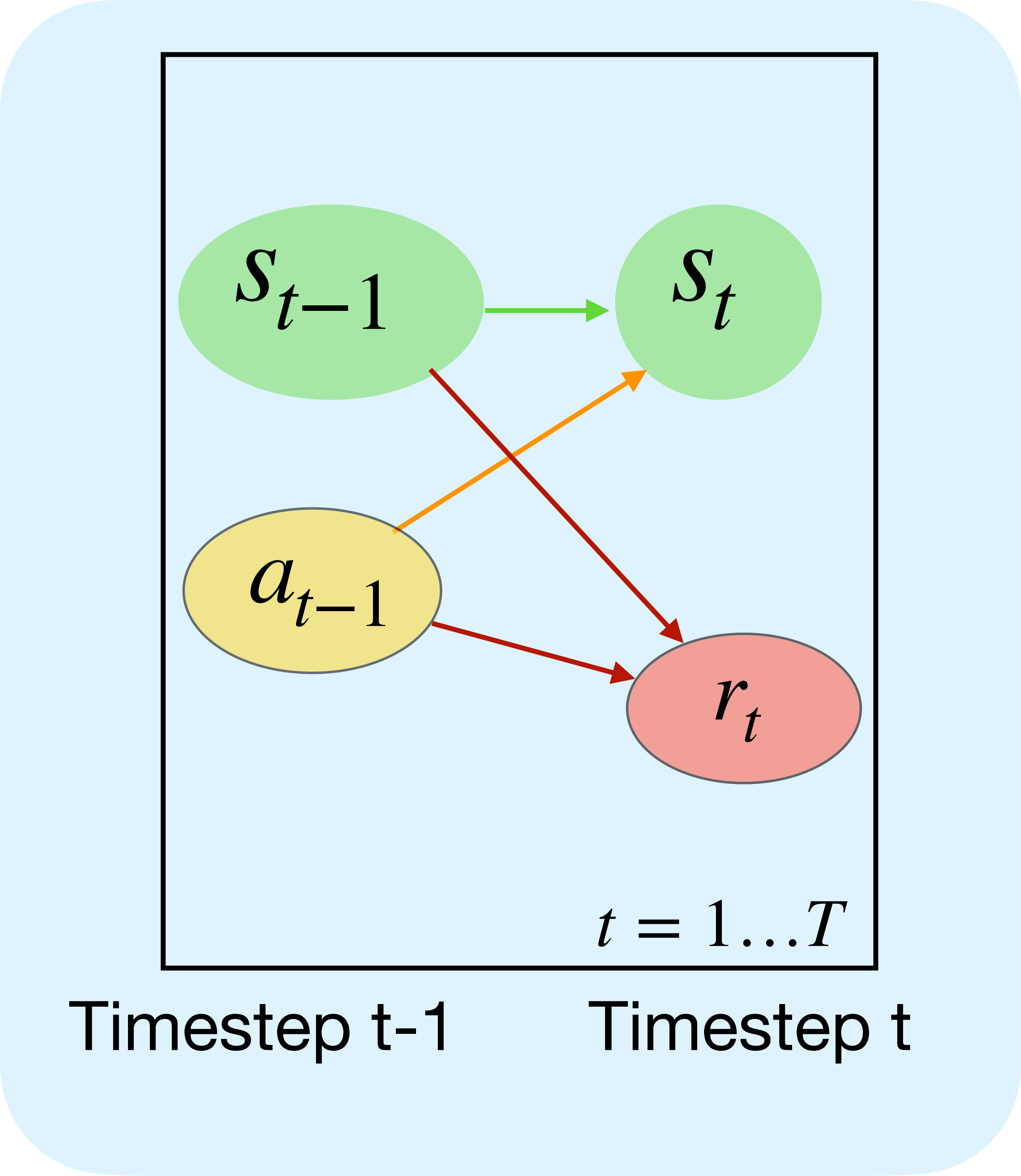
- A particularly interesting application of Dynamical Bayesian Networks are (Factored) Markov Decision Processes - **(Factored) MDPs** - used in RL



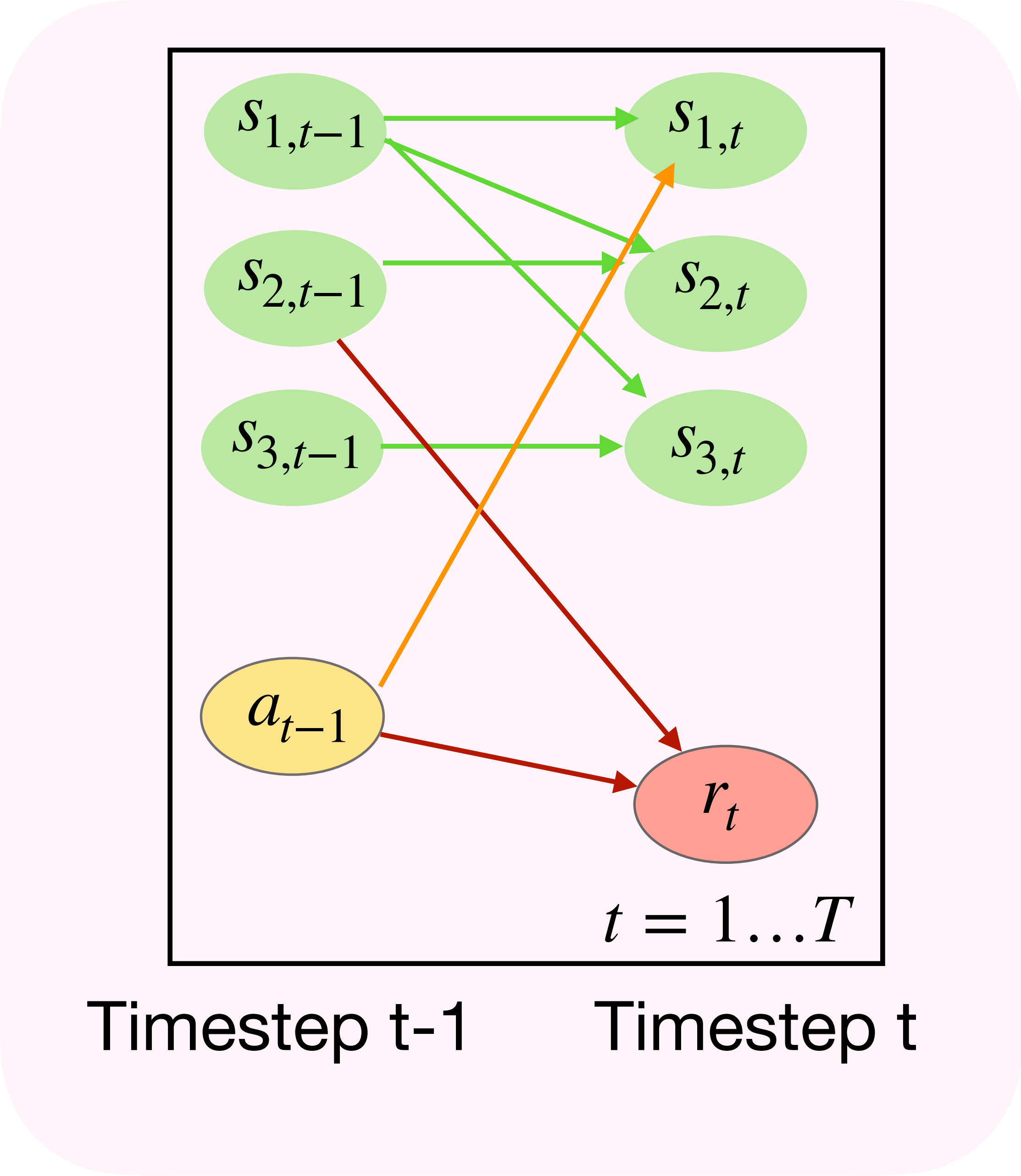
MDP

MDPs vs Factored MDPs

- A particularly interesting application of Dynamical Bayesian Networks are (Factored) Markov Decision Processes - **(Factored) MDPs** - used in RL

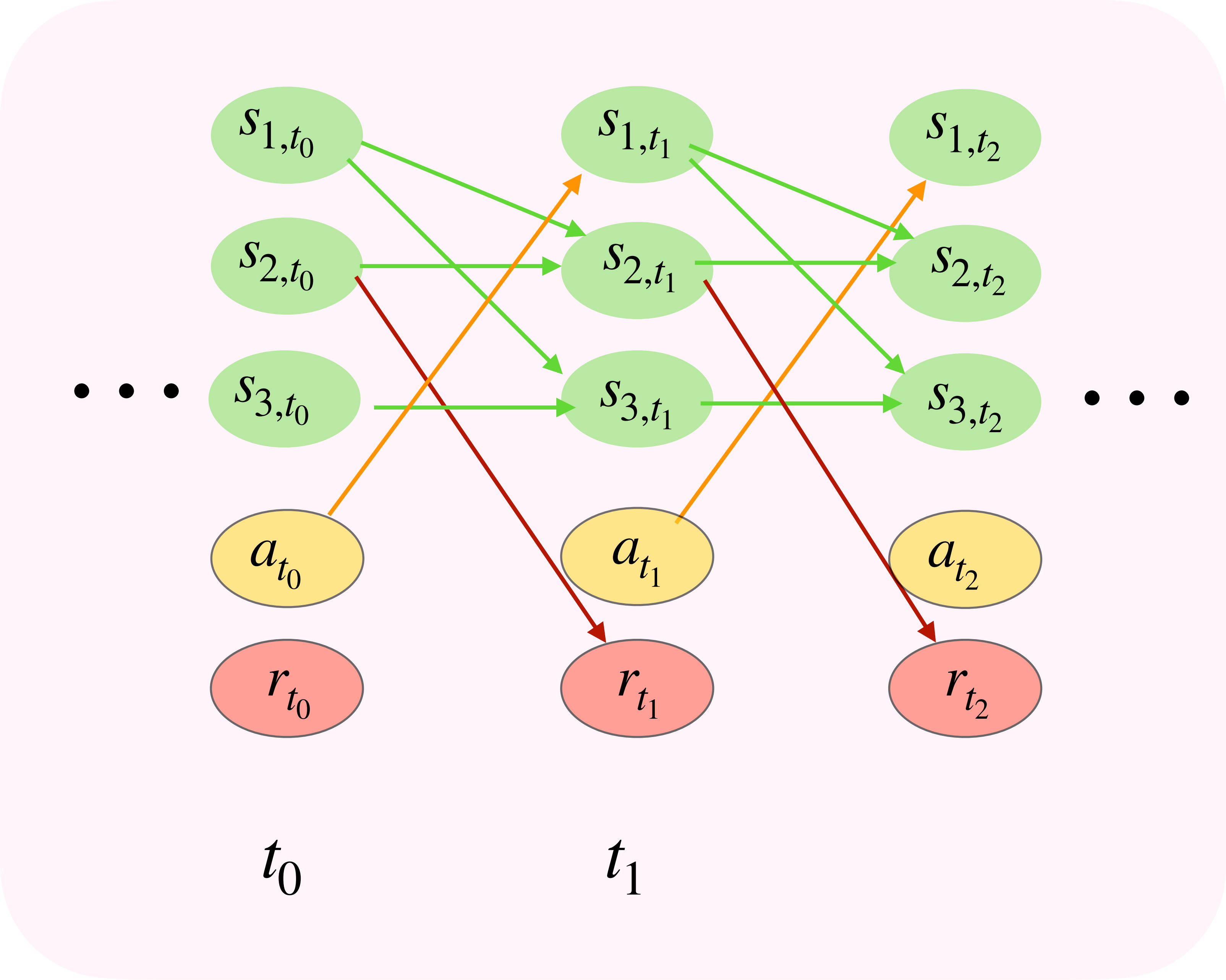


MDP



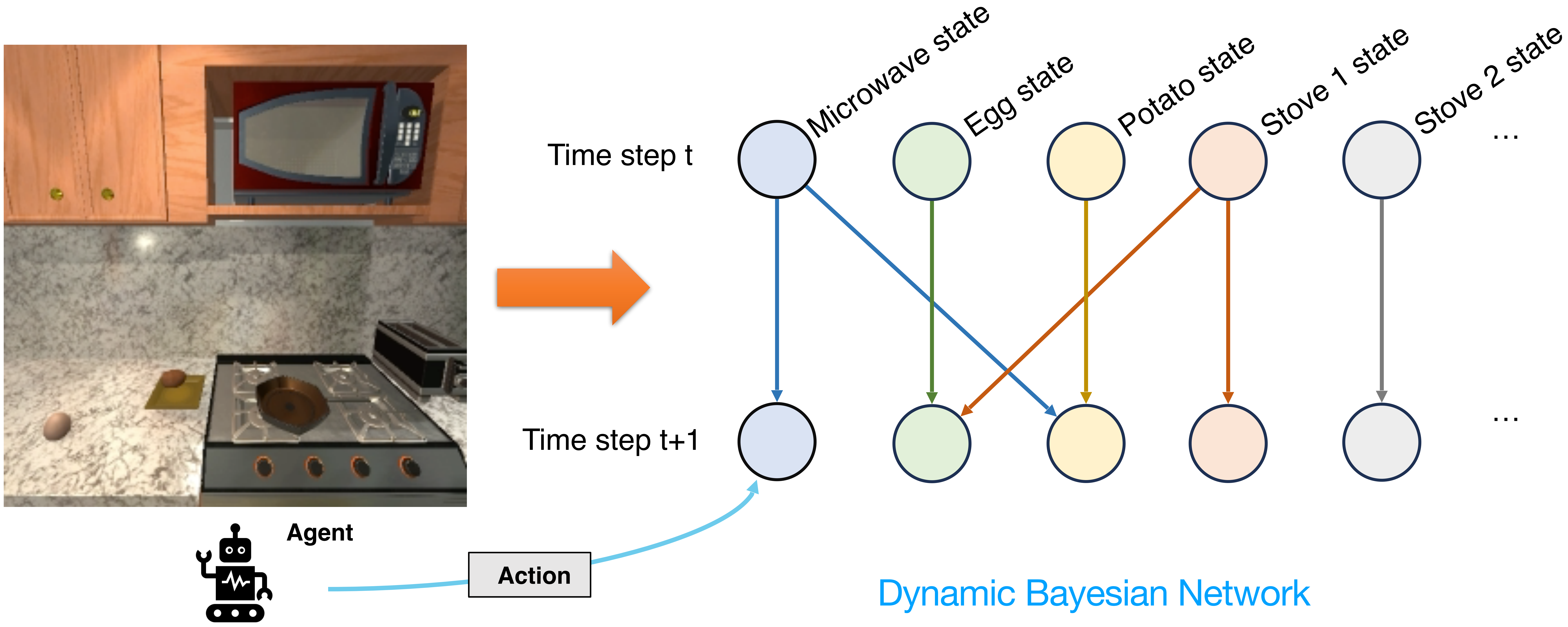
Factored MDP

$\equiv$



Factored MDP unrolled in time

Goal: CRL in temporal settings



Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 5: auxilliary variable - generalizes temporal or actions

Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

Output 3: Up to differentiable component-wise transformations + permutation

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

IVAE

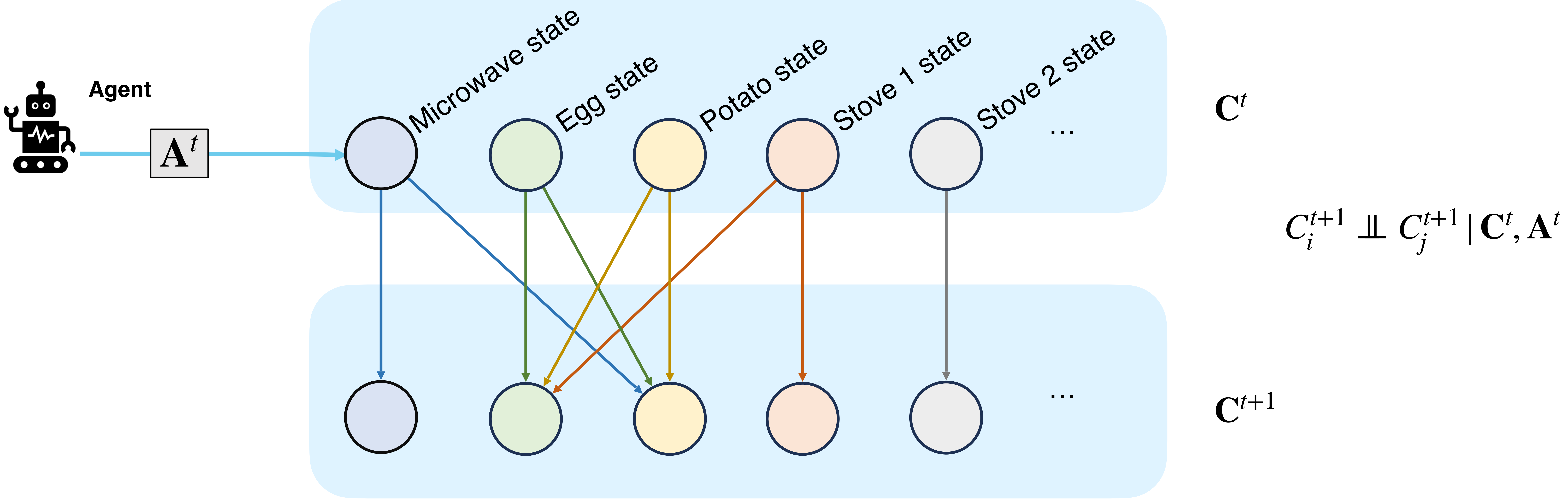
- **IVAE** assumes we observe **an auxilliary variable $\mathbf{u}$**, s.t. $z_i \perp\!\!\!\perp z_j \mid \mathbf{u}$, e.g. $\mathbf{u}$ can be time index, action and/or the state at previous timestep

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

IVAE

- **IVAE** assumes we observe **an auxilliary variable $\mathbf{u}$**, s.t. $z_i \perp\!\!\!\perp z_j \mid \mathbf{u}$, e.g. $\mathbf{u}$ can be time index, action and/or the state at previous timestep



Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

IVAE

- **IVAE** assumes we observe **an auxilliary variable $\mathbf{u}$**, s.t. $z_i \perp\!\!\!\perp z_j \mid \mathbf{u}$, e.g. $\mathbf{u}$ can be time index, action and/or the state at previous timestep
- **Conditional exponential family prior**

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) = \prod_{i}^{n} \underbrace{\frac{\overbrace{Q_i(z_i)}^{\text{Base measure}}}{Z_i(\mathbf{u})}}_{\text{Normalizing const.}} \exp\left[\sum_{j=1}^{\overbrace{k}^{k \text{ dim. suff. stat}}} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})\right]$$

$\mathbf{x} \in \mathbb{R}^d$
$\mathbf{z} \in \mathbb{R}^n$
$\mathbf{u} \in \mathbb{R}^m$

T sufficient statistic **λ natural parameter**

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

IVAE

- **IVAE** assumes we observe **an auxilliary variable $\mathbf{u}$,** s.t. $z_i \perp\!\!\!\perp z_j \mid \mathbf{u}$, e.g. $\mathbf{u}$ can be time index, action and/or the state at previous timestep
- **Conditional exponential family prior**
- **Sufficient variability**: there are enough distinct points $\mathbf{u}^0, \ldots, \mathbf{u}^{nk}$ such that $L = [\lambda(\mathbf{u}^1) - \lambda(\mathbf{u}^0), \ldots, \lambda(\mathbf{u}^{nk}) - \lambda(\mathbf{u}^0)]$ is invertible + few other ass.

$\Longrightarrow$ up to perm. & element-wise non-linear transformation

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

IVAE

- **IVAE** assumes we observe **an auxilliary variable $\mathbf{u}$**, s.t. $z_i \perp\!\!\!\perp z_j \mid \mathbf{u}$, e.g. $\mathbf{u}$ can be time index, action and/or the state at previous timestep
- **Conditional exponential family prior**
- **Sufficient variability**: there are enough distinct points $\mathbf{u}^0, \ldots, \mathbf{u}^{nk}$ such that $L = [\lambda(\mathbf{u}^1) - \lambda(\mathbf{u}^0), \ldots, \lambda(\mathbf{u}^{nk}) - \lambda(\mathbf{u}^0)]$ is invertible + few other ass.

If $k = 1$, $T_{i,1}$ non-monotonic, $\dfrac{\partial \mathbf{f}}{\partial z_i}$ continuous

If $k \geq 2$, $T_{i,k}$ twice differentiable, $\mathbf{f}$ all 2nd order cross derivatives

$\Longrightarrow$ up to perm. & element-wise non-linear transformation

Variational Autoencoders and Nonlinear ICA: A Unifying Framework

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, Aapo Hyvärinen

IVAE

- **IVAE** assumes we observe **an auxilliary variable $\mathbf{u}$**, s.t. $z_i \perp\!\!\!\perp z_j \mid \mathbf{u}$, e.g. $\mathbf{u}$ can be time index, action and/or the state at previous timestep
- **Pros:**
  - More general, doesn't only work in temporal settings
  - In temporal setting can work also with only observational data
- **Cons:**
  - Parametric assumptions (conditional (strong) exponential family prior, extra ass. for element-wise)
  - Counter-example: Gaussians with fixed variance/ Additive Gaussian

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions/actions

Strategy 4: assume temporal data

Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

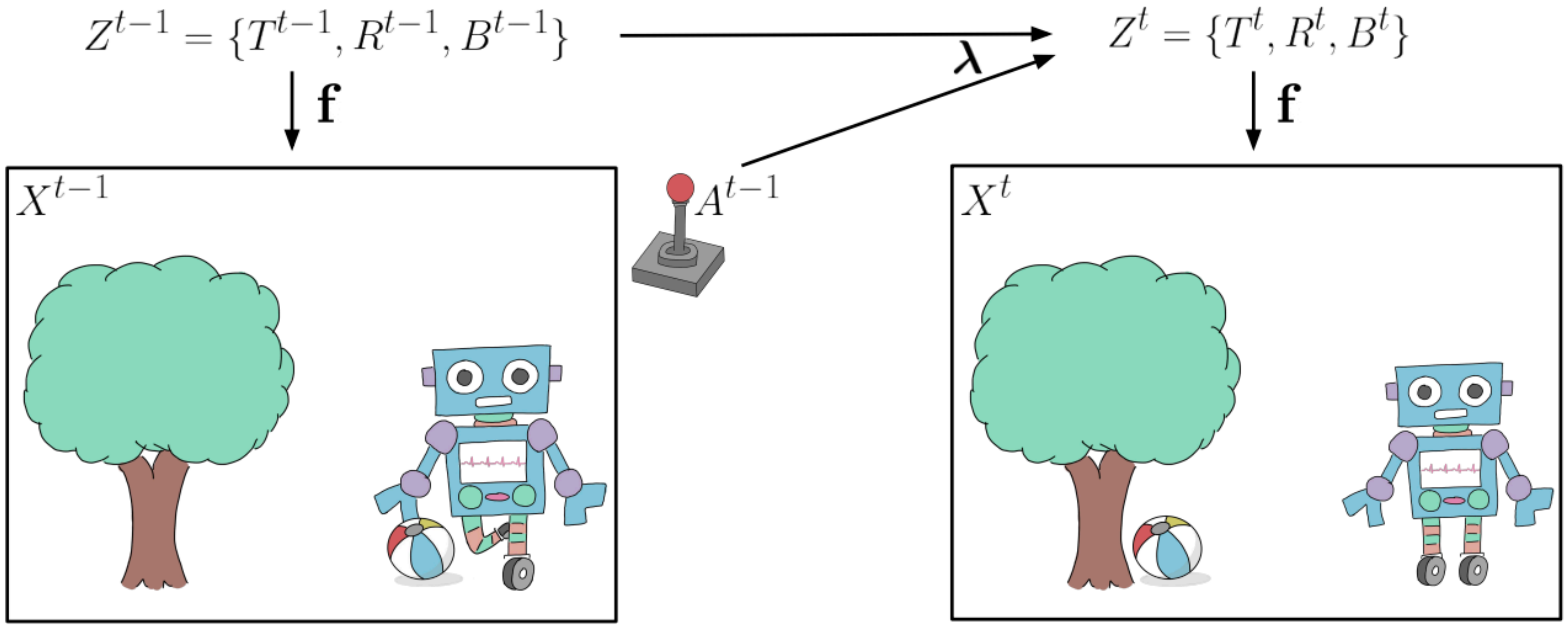
Output 3: Up to differentiable component-wise transformations + permutation

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

DMSVAE

- **Setting:** we have actions and time series, we assume a **sparse** graph



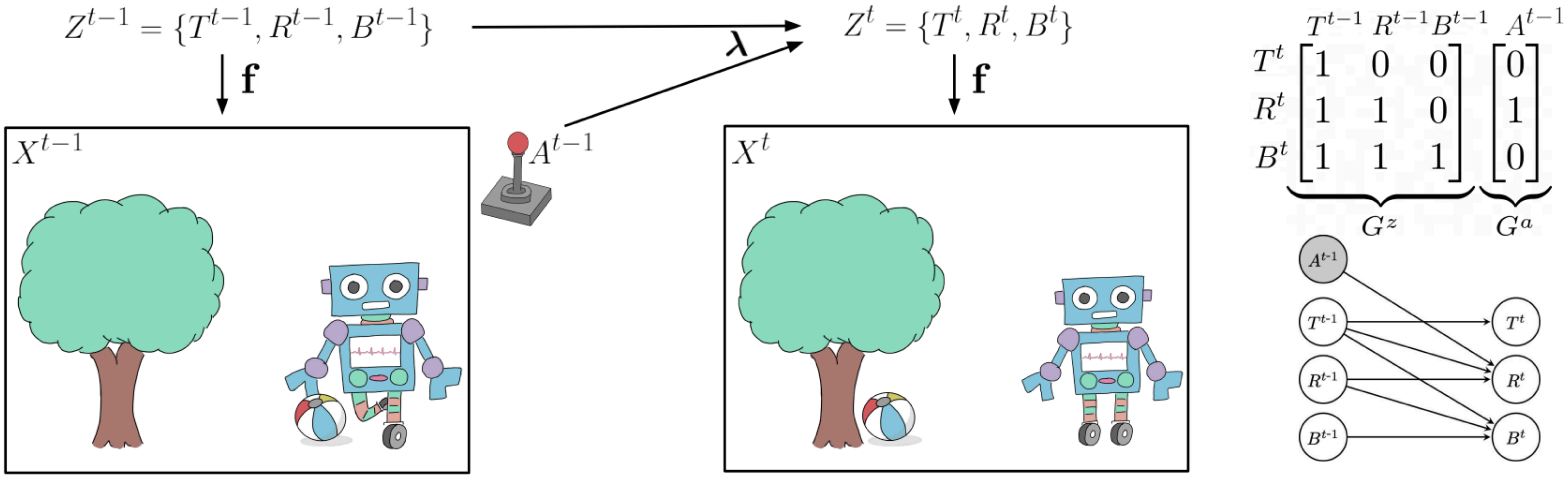
https://arxiv.org/abs/2107.10098

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

DMSVAE

- **Setting:** we have actions and time series, we assume a **sparse** graph



Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

DMSVAE

- Assume conditional independence of the latents given the past:

$$p(z^t \mid z^{<t}, a^{<t}) = \prod_{i=1}^{d_z} p(z_i^t \mid z^{<t}, a^{<t})$$

- Assume each factor above lives in an exponential family:

Binary adjacency matrices removing dependencies (causal graph)

$$p(z_i^t \mid z^{<t}, a^{<t}) = h_i(z_i^t) \exp\{\mathbf{T}_i(z_i^t)^\top \underbrace{\lambda_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t})} - \psi_i(z^{<t}, a^{<t})\}$$

Base measure

Sufficient statistic

Transition function/mechanism outputting natural parameter

Normalizing constant

Slide by Sébastien Lachapelle, https://arxiv.org/abs/2107.10098

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

DMSVAE

Main result: ident. up to permutation & element-wise linear transformations if:

1. **Conditional exponential family prior with k=1**
2. **Sufficient time variability:** the Jacobian of the ground truth transition function λ with respect to z varies "sufficiently"
3. **Sufficient action variability:** the ground truth transition function λ is affected "sufficiently strongly" by each action

https://arxiv.org/abs/2107.10098

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

DMSVAE

Main result: ident. up to permutation & element-wise linear transformations if:

1. **Conditional exponential family prior with k=1**
2. **Sufficient time variability:** the Jacobian of the ground truth transition function λ with respect to z varies "sufficiently"
3. **Sufficient action variability:** the ground truth transition function λ is affected "sufficiently strongly" by each action
4. **Graphical assumption:** in G, for each i, there exists a way to describe it as the intersection of sets of parents, children or variables affected by actions.
5. The estimated graph $\hat{G}$ is **sparser than true graph G (regularization)**

https://arxiv.org/abs/2107.10098

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

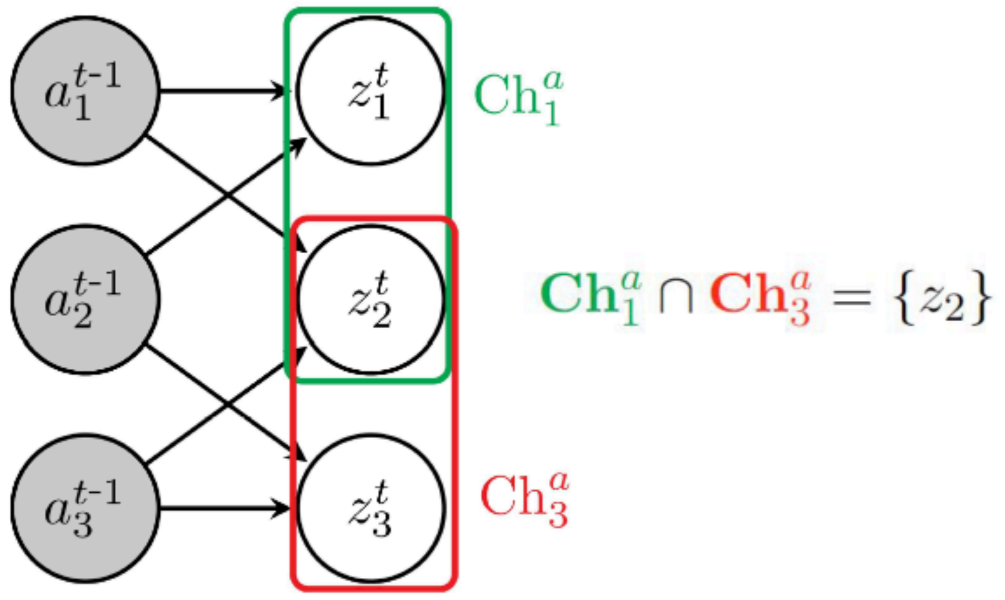
DMSVAE

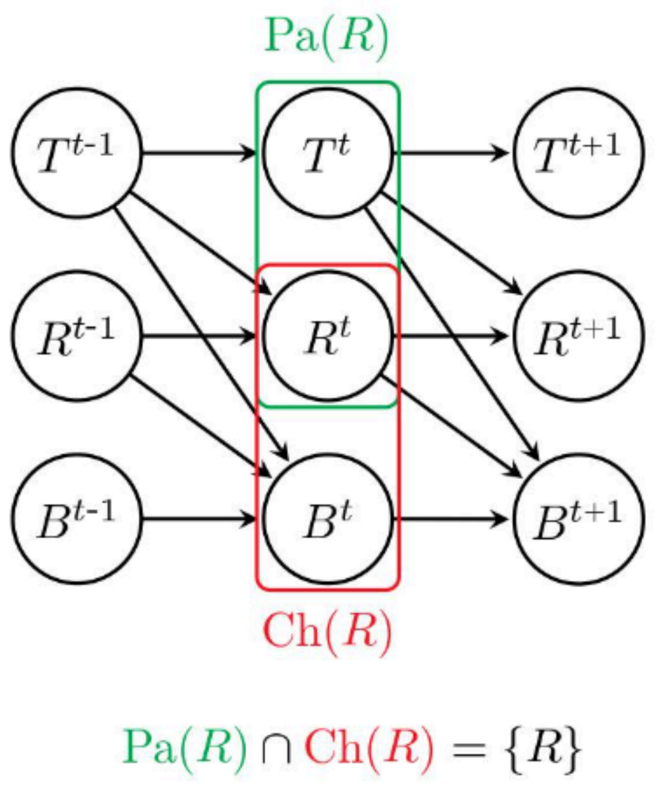
Understanding the graphical criterion

5. ***[Graphical criterion]*** *For all $p \in \{1, ..., d_z\}$, there exist sets $\mathcal{I}, \mathcal{J} \subset \{1, ..., d_z\}$ and $\mathcal{L} \subset \{1, ..., d_a\}$ such that*

$$\left(\bigcap_{i\in\mathcal{I}} \mathbf{Pa}_i^z\right) \cap \left(\bigcap_{j\in\mathcal{J}} \mathbf{Ch}_j^z\right) \cap \left(\bigcap_{\ell\in\mathcal{L}} \mathbf{Ch}_\ell^a\right) = \{p\},$$

where $\mathbf{Pa}_i^z$ and $\mathbf{Ch}_i^z$ are the sets of parents and children of node z_i in G^z, respectively, while $\mathbf{Ch}_\ell^a$ is the set of children of a_ℓ in G^a.





The complete and empty graphs are two counterexamples

Slide by Sébastien Lachapelle, https://arxiv.org/abs/2107.10098

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, Simon Lacoste-Julien

DMSVAE

- **Pros:**
  - Can work with also only observational data or unknown intervention targets
  - Can do Gaussians with fixed variance / Additive Gaussian

- **Cons:**
  - Graphical assumption (although actions can make it easier to fulfill)
    - UAI 2022 workshop paper shows **partial identifiability** results without this assumption
  - Parametric assumptions (conditional exponential family prior, k=1)
    - [Lachapelle et al. 2024] is **nonparametric**, but ident. up to $\mathbf{a}/\mathbf{z}$ consistency
  - Potentially tricky to optimize for images

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions

Strategy 4: assume temporal data

Output 1: Up to permutation and scale

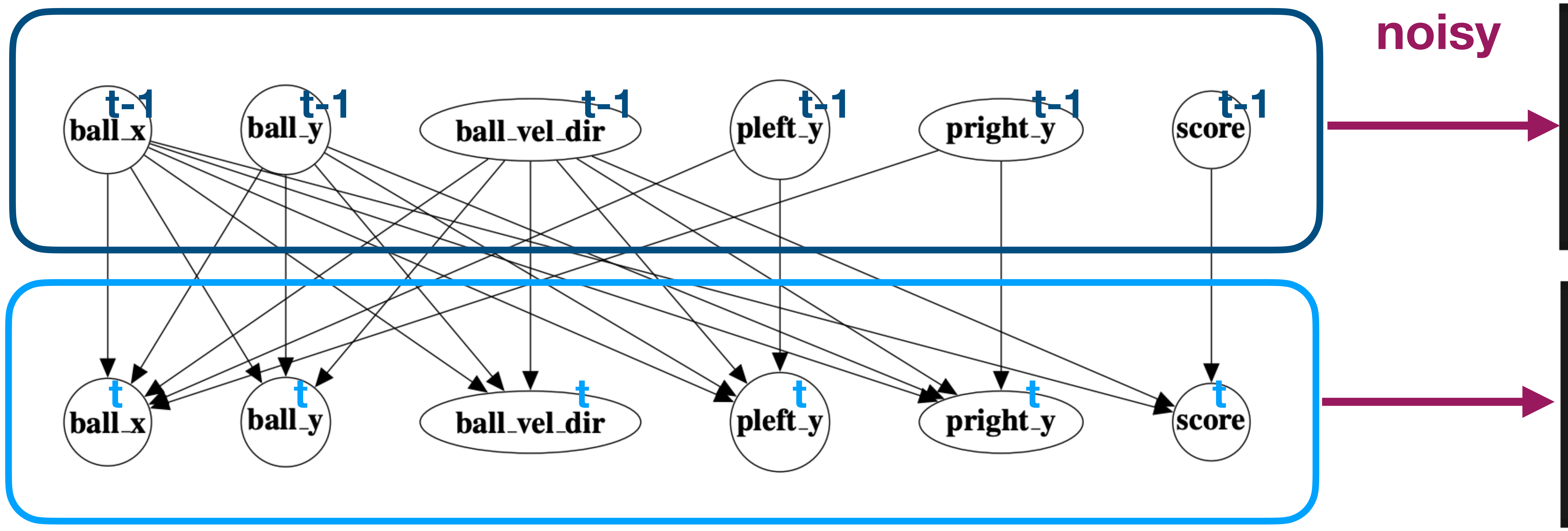
Output 2: Up to differentiable component-wise transformations

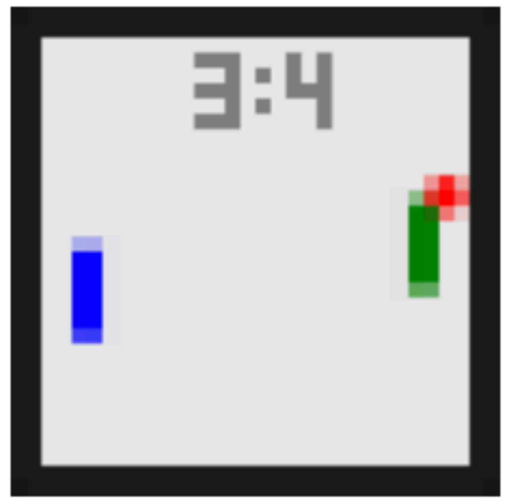
Output 3: Up to differentiable component-wise transformations + permutation

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

Possibly noisy

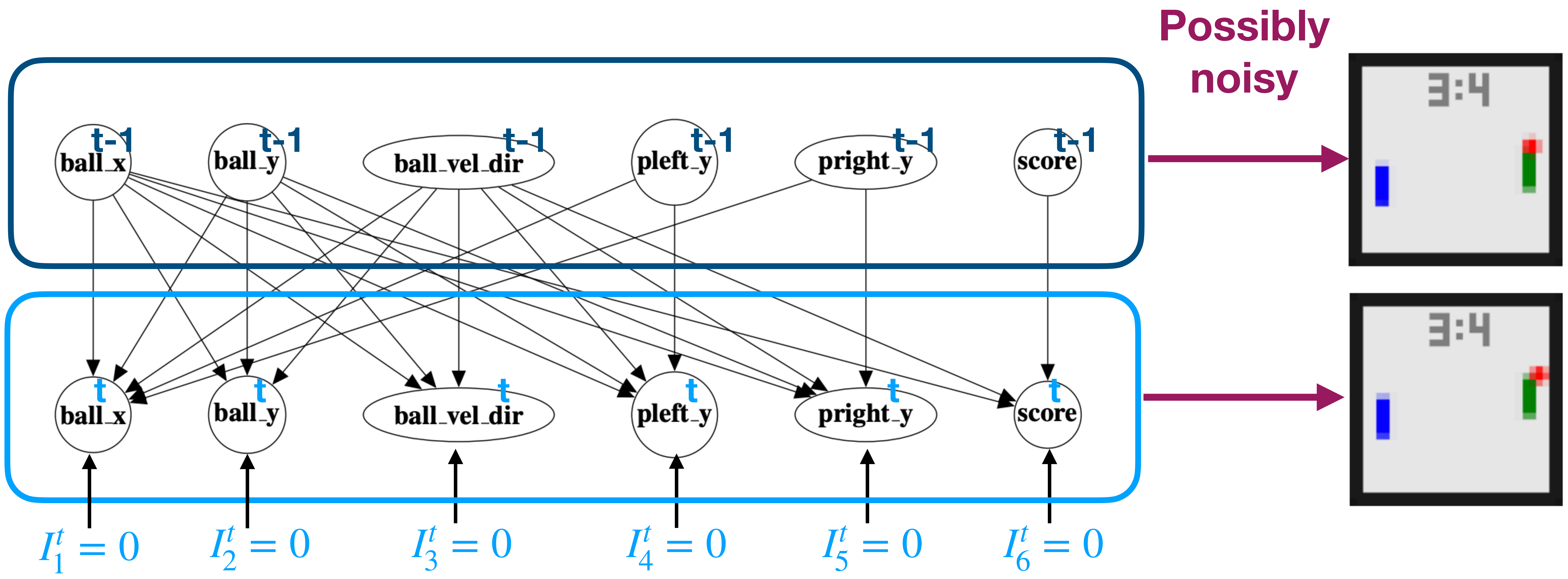




https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

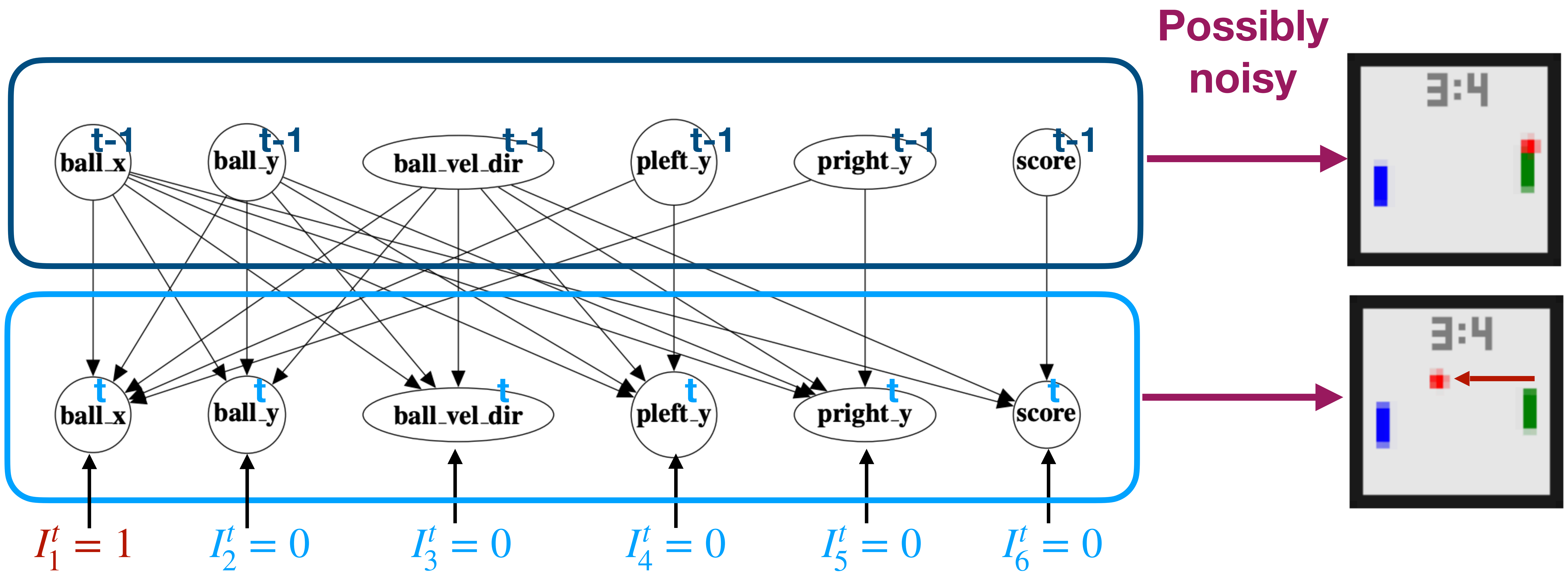
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves



https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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



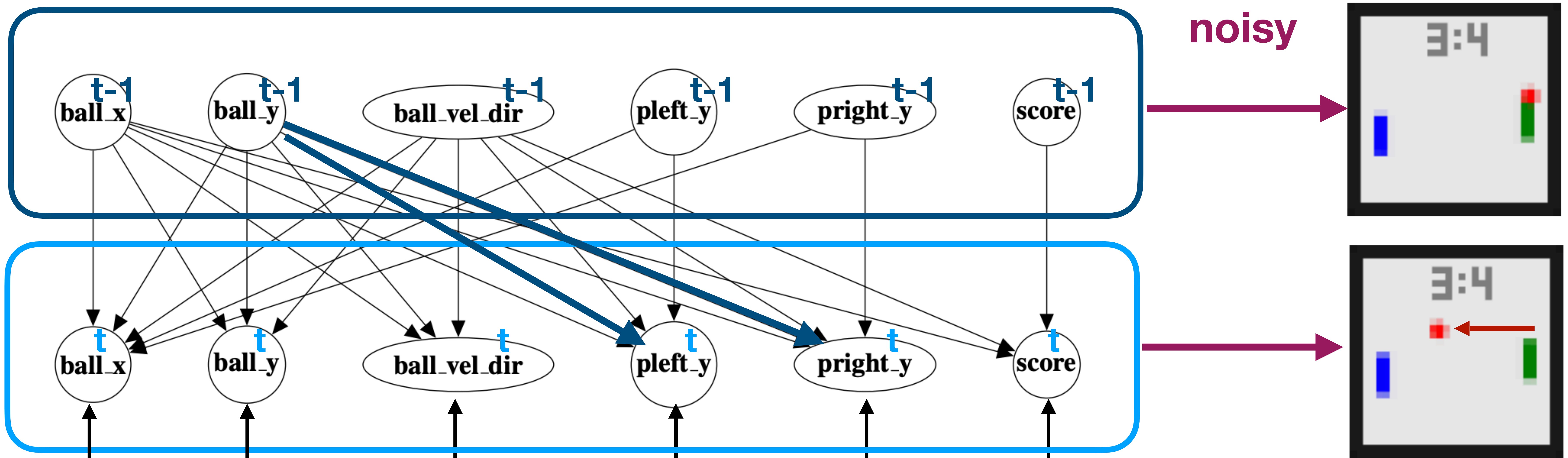
Stochastic intervention
(we don't know where the ball will be)

https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

Possibly noisy



$I_1^t = 1$ $I_2^t = 0$ $I_3^t = 0$ $I_4^t = 0$ $I_5^t = 0$ $I_6^t = 0$

Stochastic intervention
(we don't know where the ball will be)

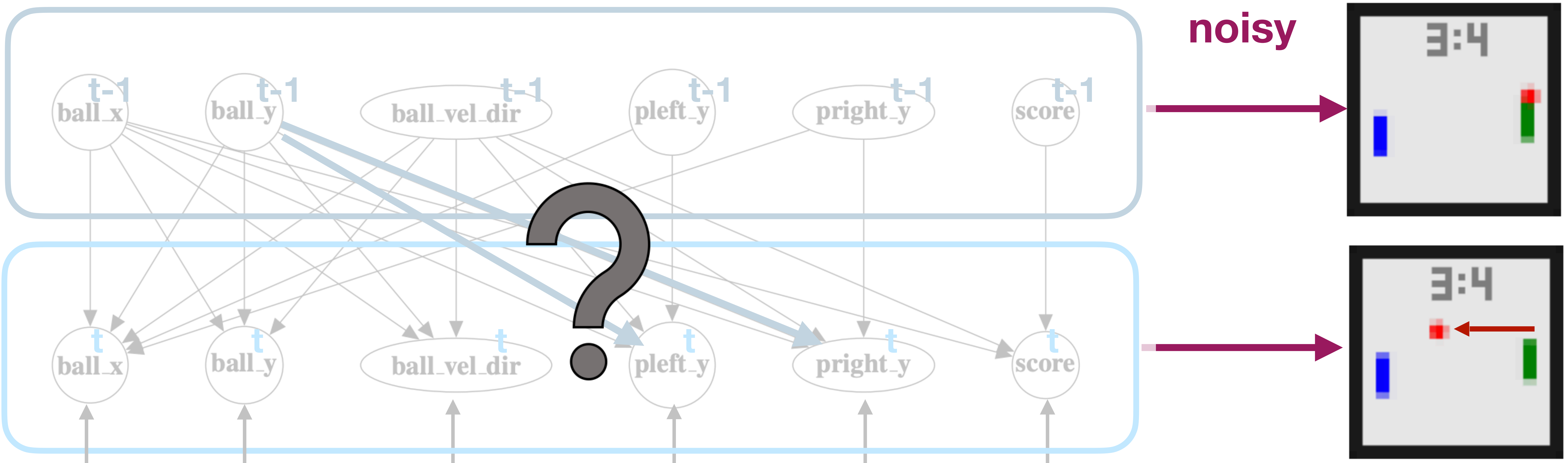
The paddles continue moving as usual (not counterfactual)

https://arxiv.org/abs/2202.03169

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

Possibly noisy



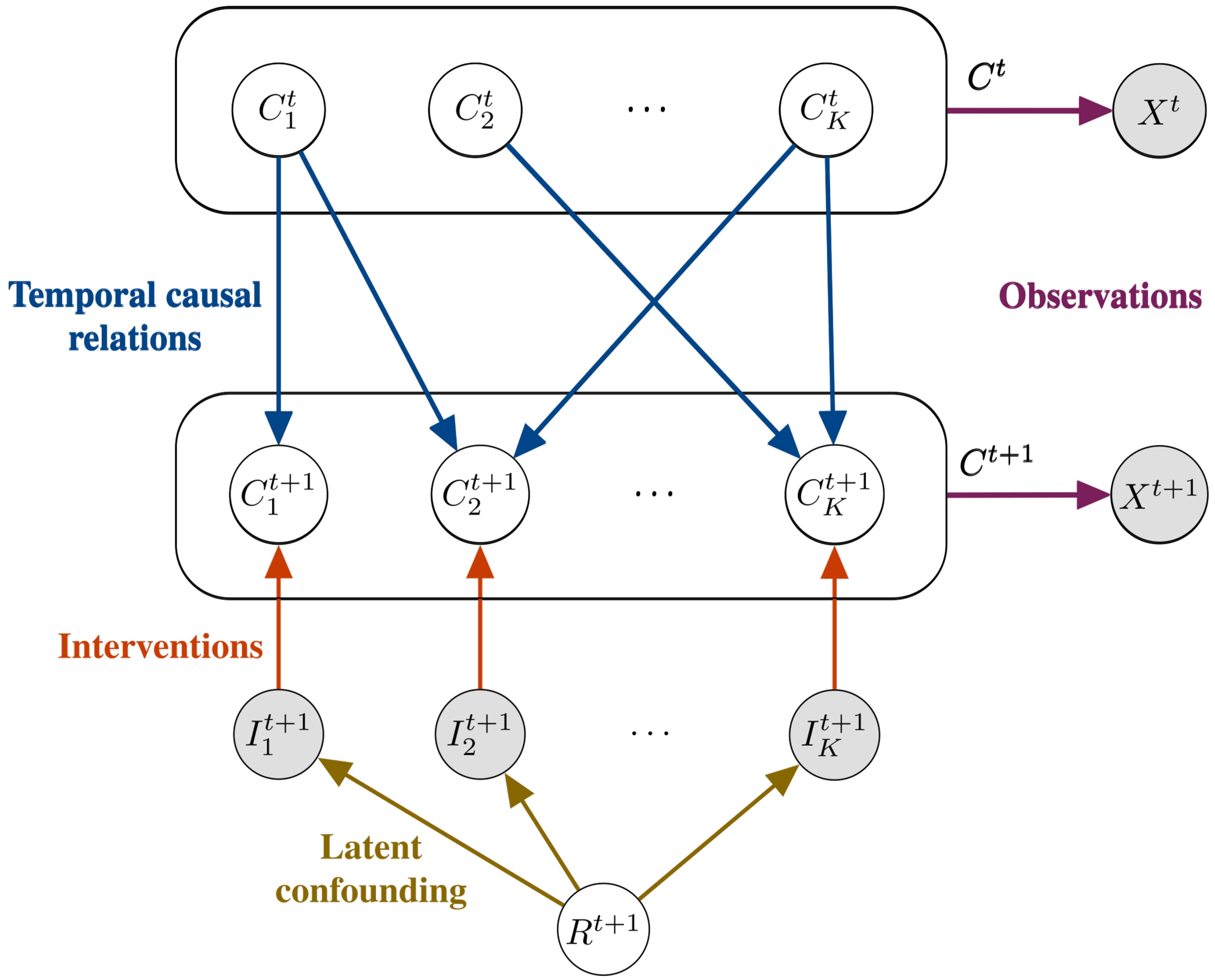
$I_1^t = 1$ $I_2^t = 0$ $I_3^t = 0$ $I_4^t = 0$ $I_5^t = 0$ $I_6^t = 0$

Stochastic intervention

(we don’t know where the ball will be)

https://arxiv.org/abs/2202.03169

Temporal Intervened Sequences (TRIS)

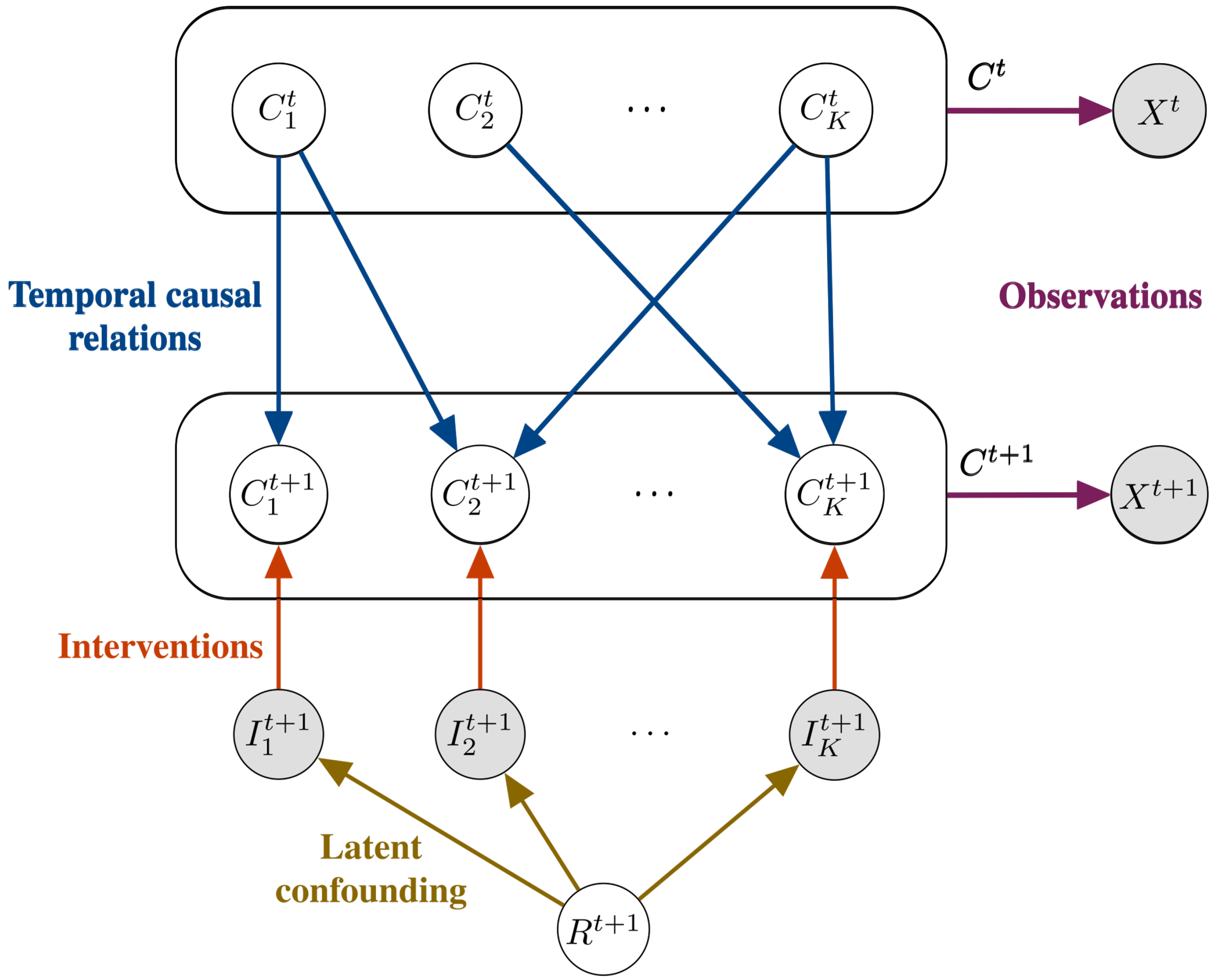


- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images

- We assume that the **latent** causal process is a **Dynamic Bayesian network** with **K multidimensional causal variables**

$$X^t = h(C_1^t, \dots C_K^t, E_0^t)$$

Temporal Intervened Sequences (TRIS)

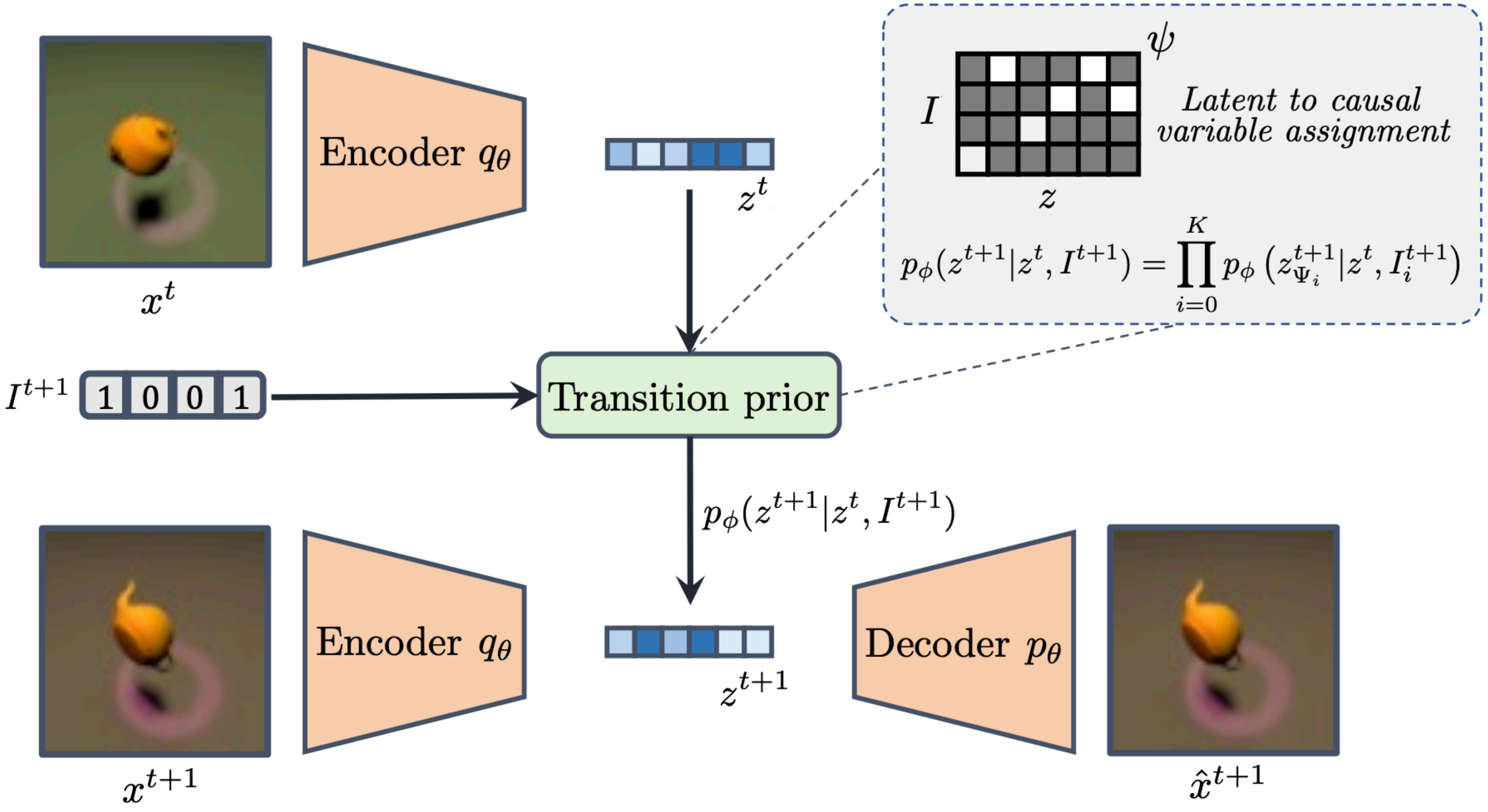


- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images
- We assume that the **latent** causal process is a **Dynamic Bayesian network** with **K multidimensional causal variables**
- We assume that **(soft or perfect) interventions** can happen on the underlying system and **we observe the targets** I_i^t
  - $I_i \to C_i$

A variational autoencoder architecture: CITRIS-VAE

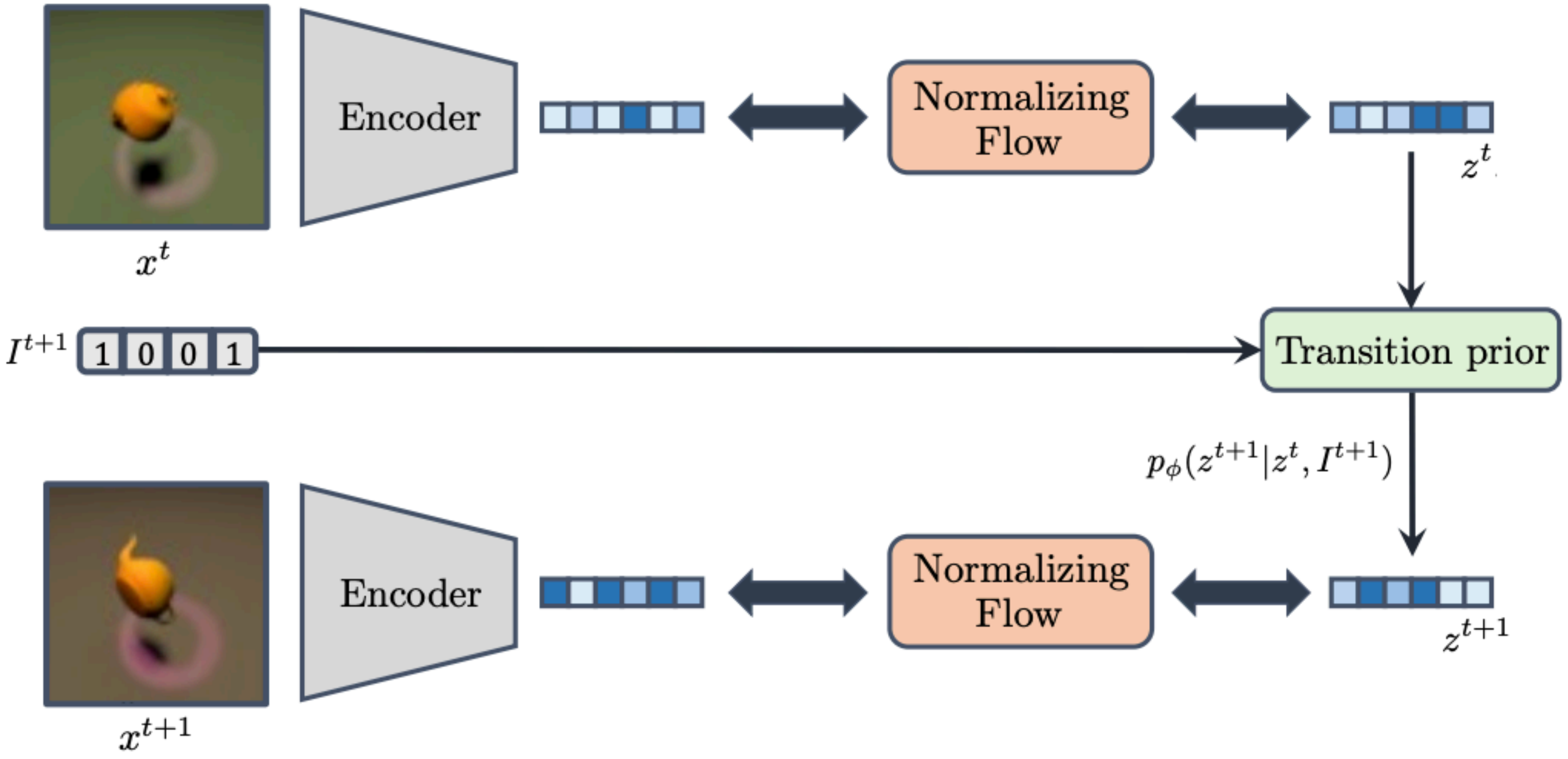
- Encoder q_θ from image x to latent z
- Decoder p_θ from latent z to image x
- **Assignment ψ from latent z to causal C**
  - $\forall i = 1,\ldots,K$: z_{ψ_i} corresponds to C_i
  - z_{ψ_0} corresponds to the noise
- Transition prior p_ϕ between causal at t and t+1

$$\mathcal{L}_{\text{ELBO}} = -\mathbb{E}_{z^{t+1}}\left[\log p_\theta\left(x^{t+1}|z^{t+1}\right)\right] + \mathbb{E}_{z^t,\psi}\left[\sum_{i=0}^{K} D_{\text{KL}}\left(q_\theta(z_{\psi_i}^{t+1}|x^{t+1})||p_\phi(z_{\psi_i}^{t+1}|z^t, I_i^{t+1})\right)\right]$$



A normalizing flow architecture: CITRIS-NF

- We can leverage a pretrained autoencoder to get a low-dimensional latent space
    - Can be trained on observational data
- Then we train a normalizing flow to disentangle the variables (with a transition prior)

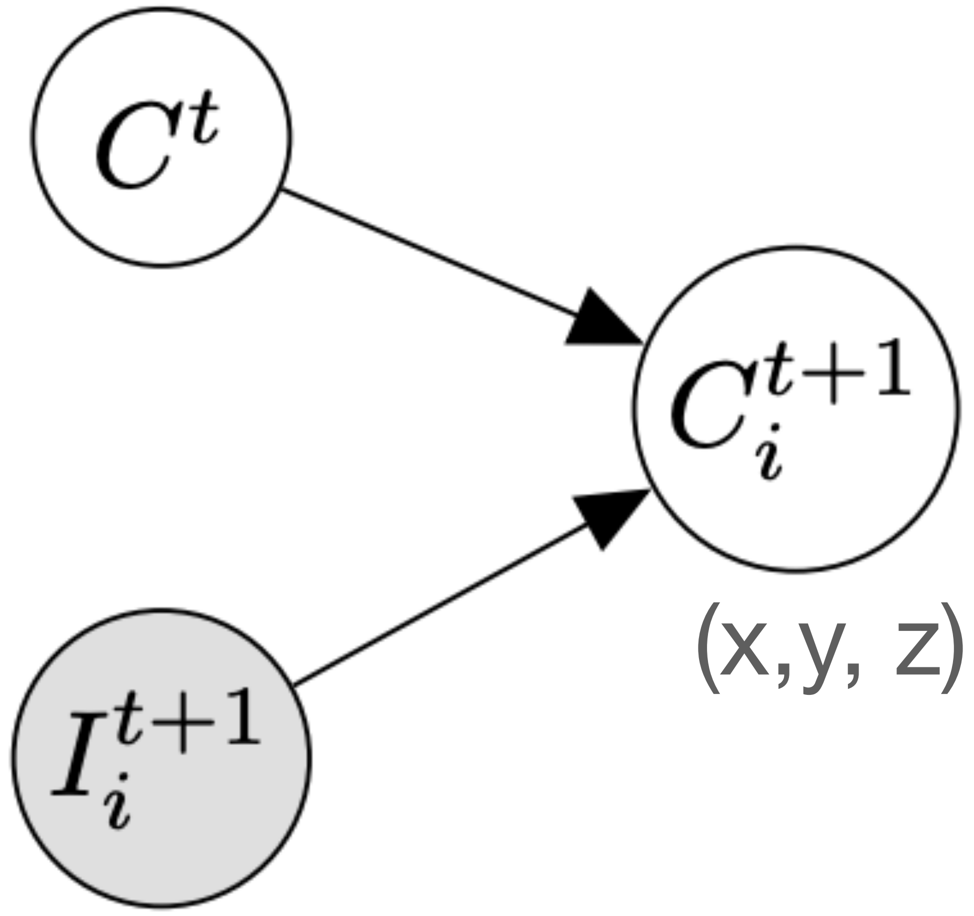


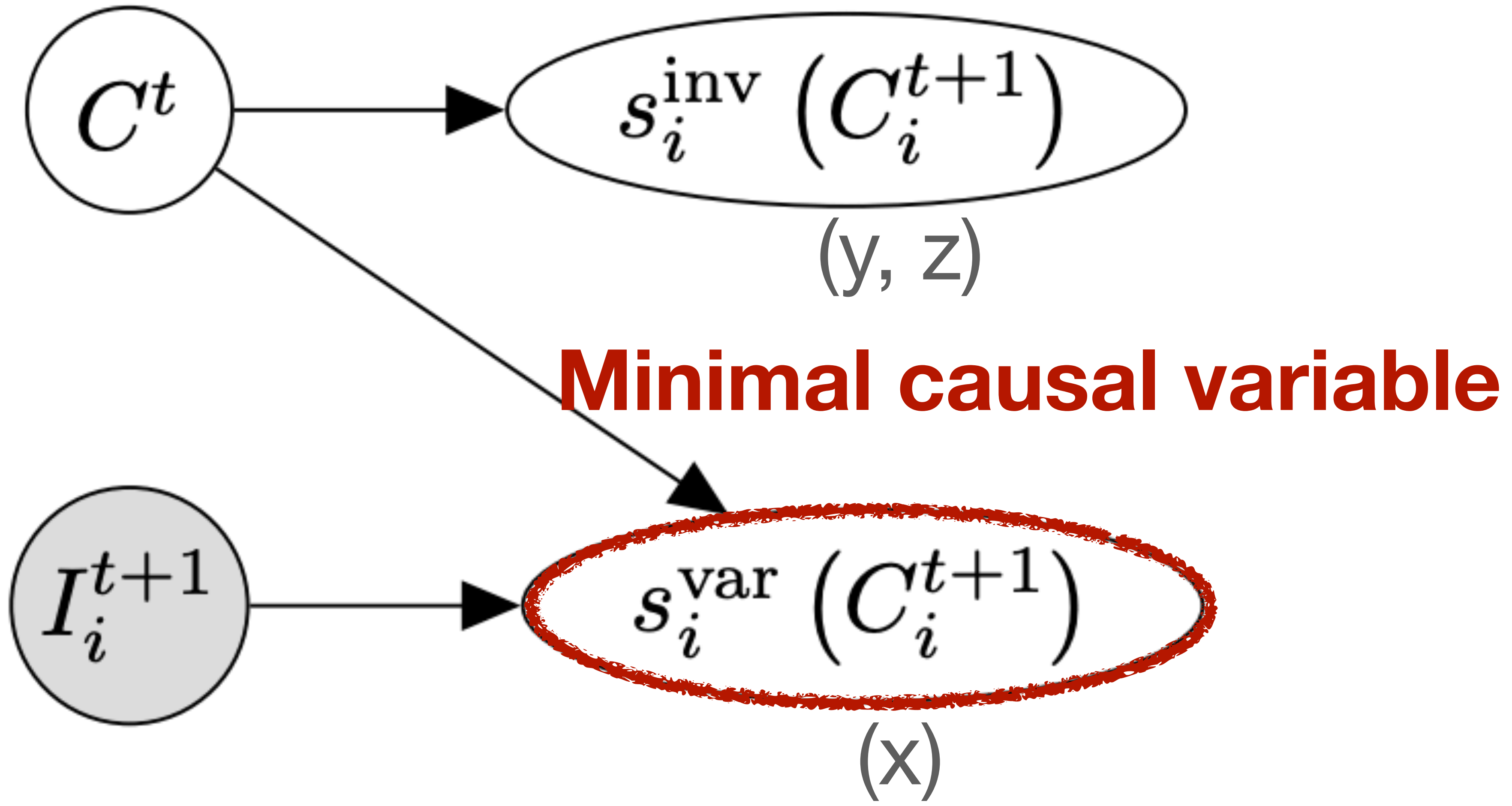
CITRIS: Simplified identification results

- TRIS setting, sufficient latent dimensions
- Assumption 1: Each I_i^t is not a deterministic function of I_j^t
- **Assumption 2:** Interventions have an effect on **all components** of any multi-dimensional causal variable
- **Assumption 3:** There are **"enough"** interventions ($O(log_2 K)$)
- Then if we maximize both the likelihood $p(x^{t+1} \mid x^t, I^{t+1})$ and the information content of the latents assigned to the noise (i.e. we assign latent variables to causal variable parsimonioulsy), we can identify causal variables $C_1, \ldots, C_K$ **up to invertible component-wise transformations**

A peek to the full results: minimal causal variables

- We model multidimensional variables, e.g. (x,y,z) coordinates for position
  - Not all might be affected by an intervention, e.g. only x is affected.
  - The **minimal causal variable** is the part of the causal variable that is **varying**

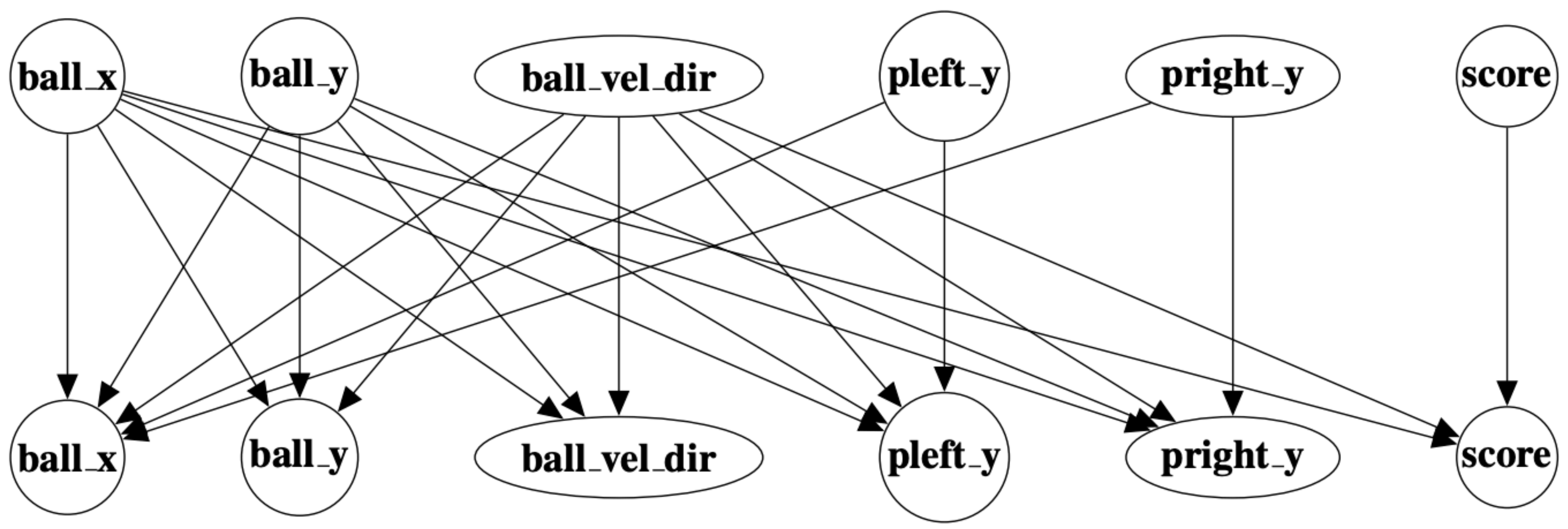


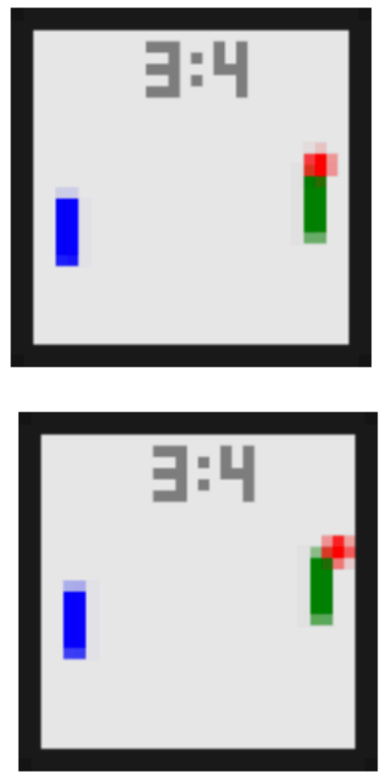


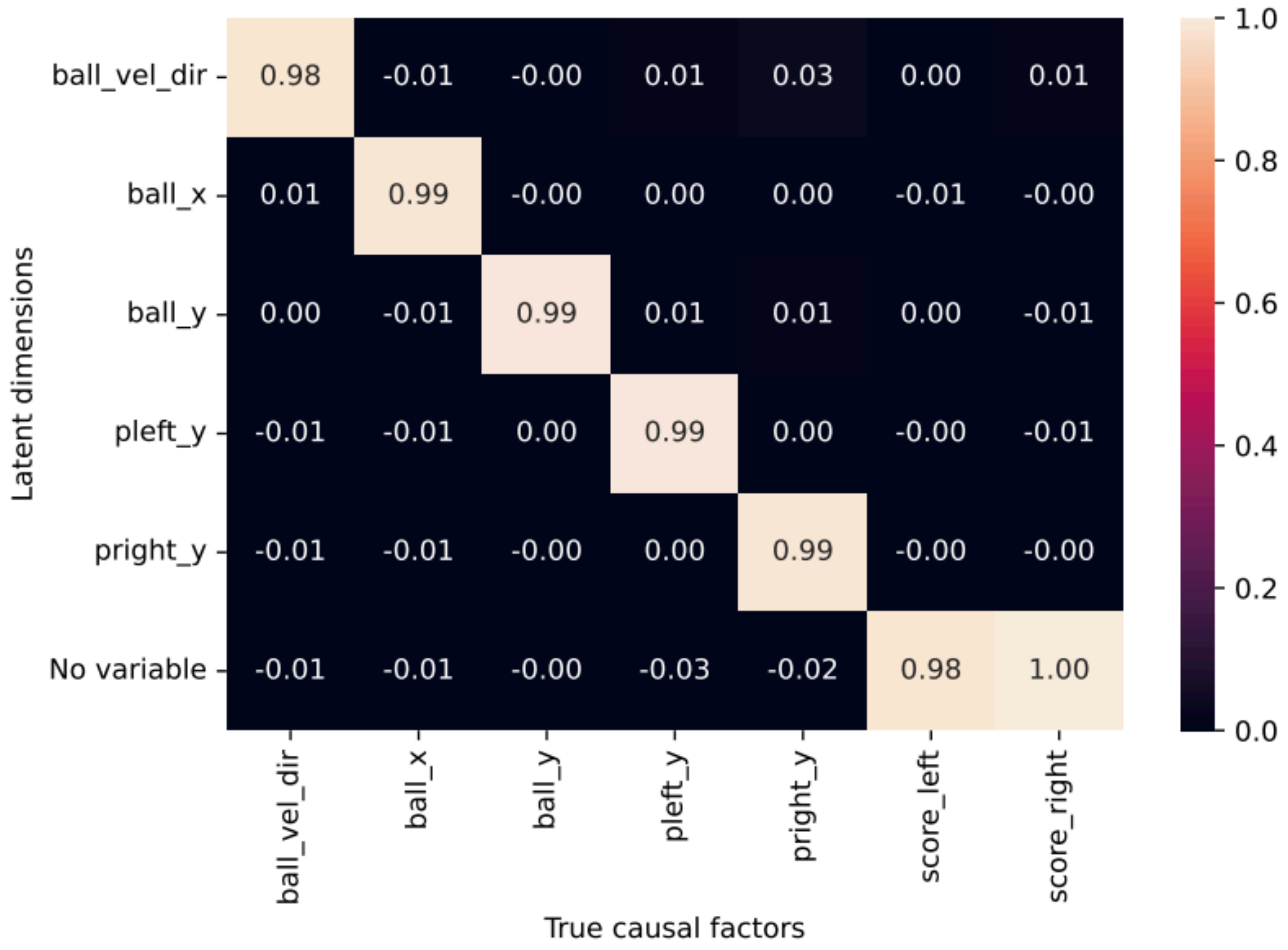
- We can identify minimal causal variables **up to invertible component-wise transformations**, if I_i^t is not a deterministic function of I_j^t

Experiments: Interventional Pong

- **Train:** We learn encoder f on a dataset with **potentially dependent** causal variables from images $\{X^t\}_{t=1}^T$ and intervention targets $\{I^t\}_{t=1}^T$ -> **unsupervised**
- **Test:** We evaluate f on a dataset with **independent** causal variables and evaluate correlation with ground truth causal variables.

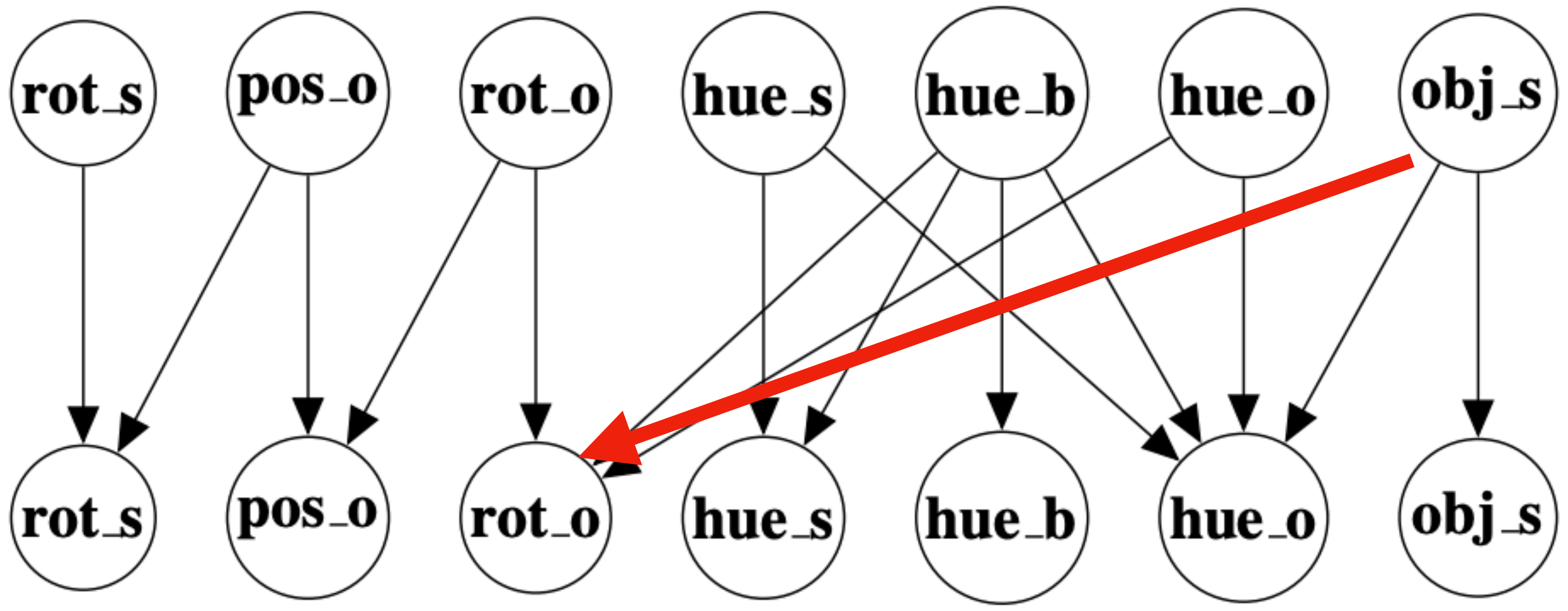






Experiments: Temporal Causal3DIdent

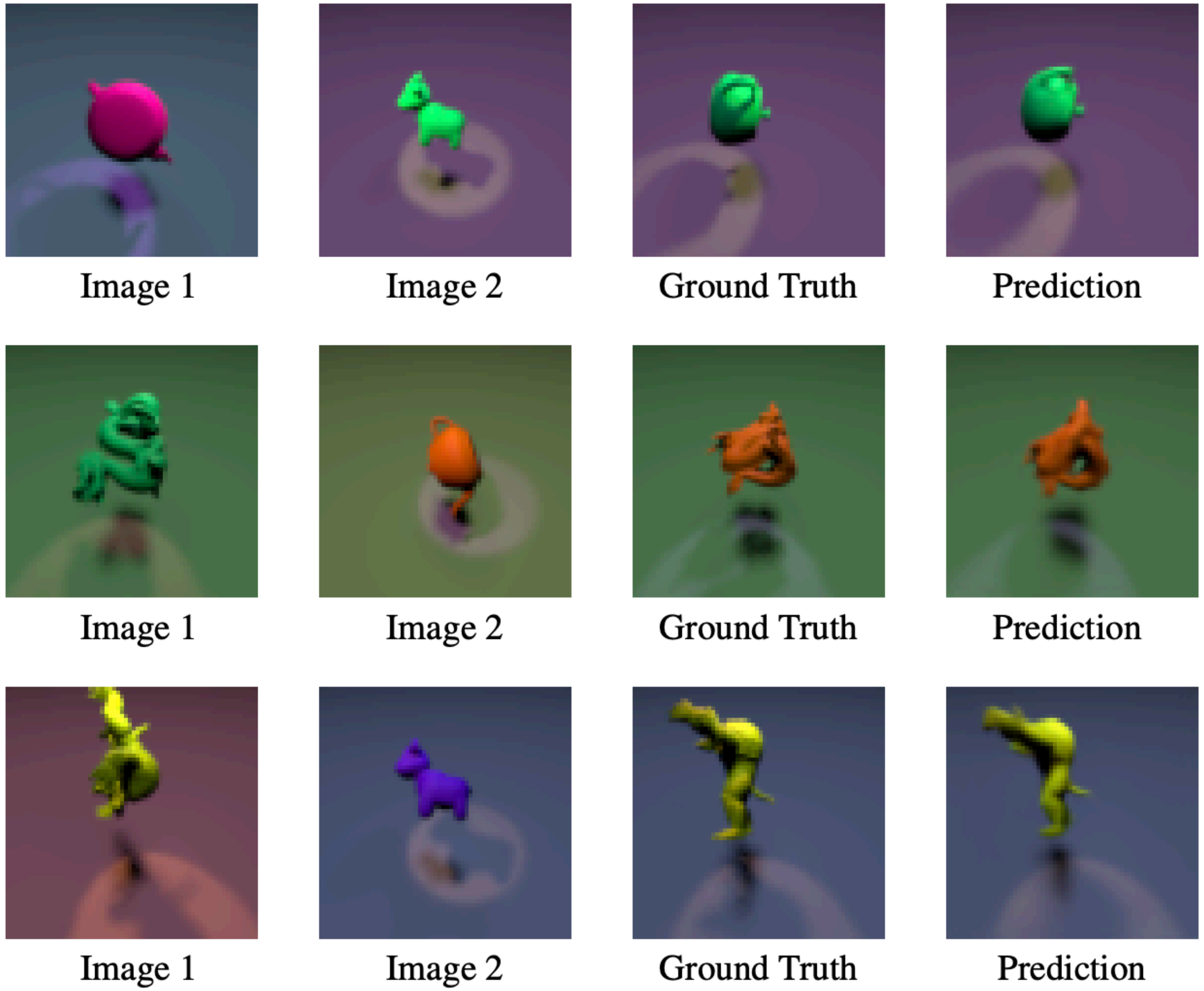
shape + spotlight colors + rot

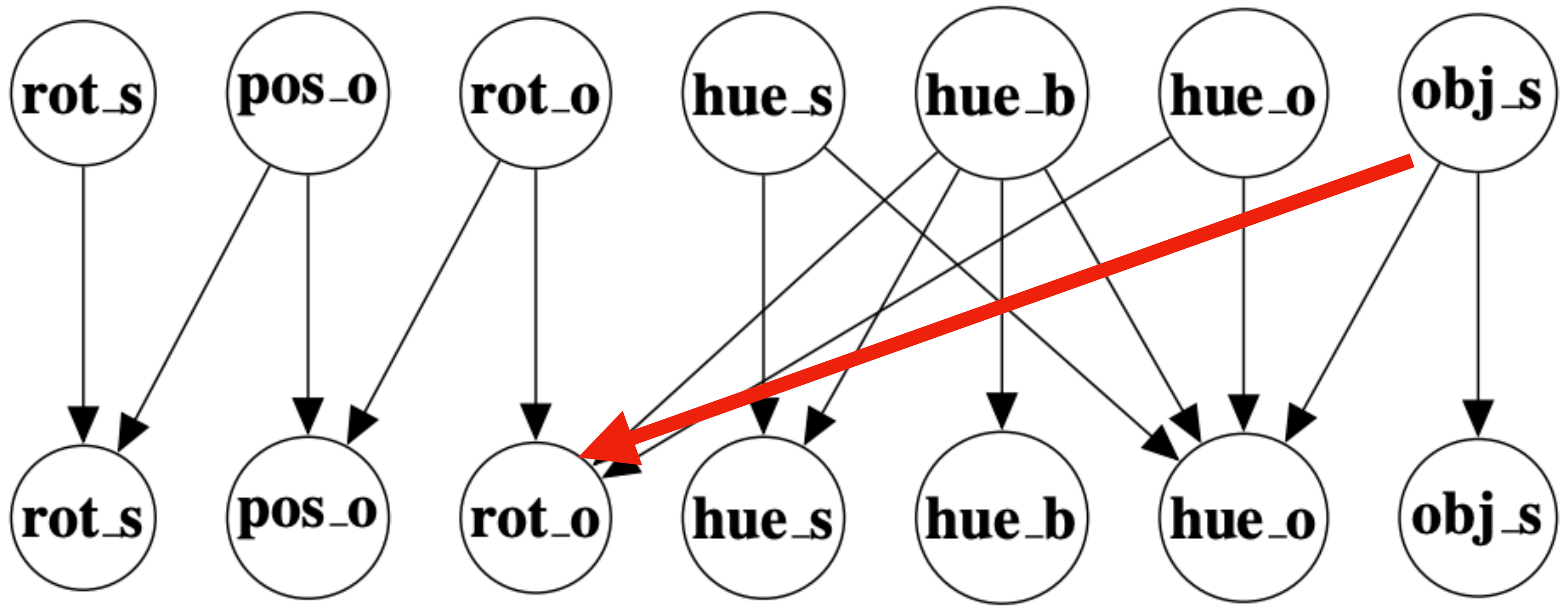


Causal graph learnt with CITRIS-NF

Experiments: Temporal Causal3DIdent

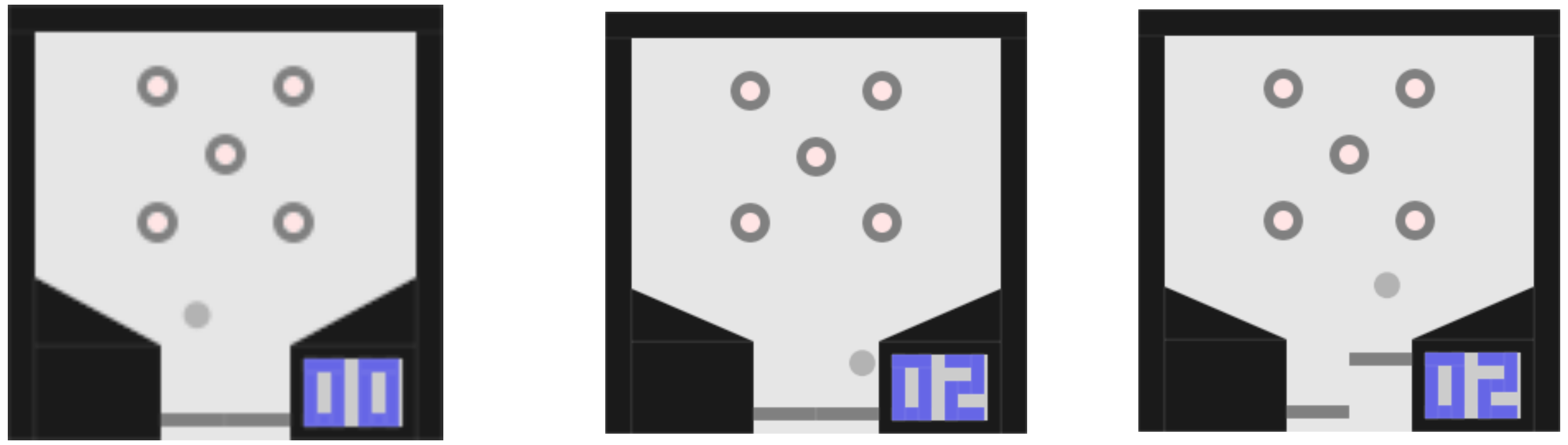
shape + spotlight colors + rot

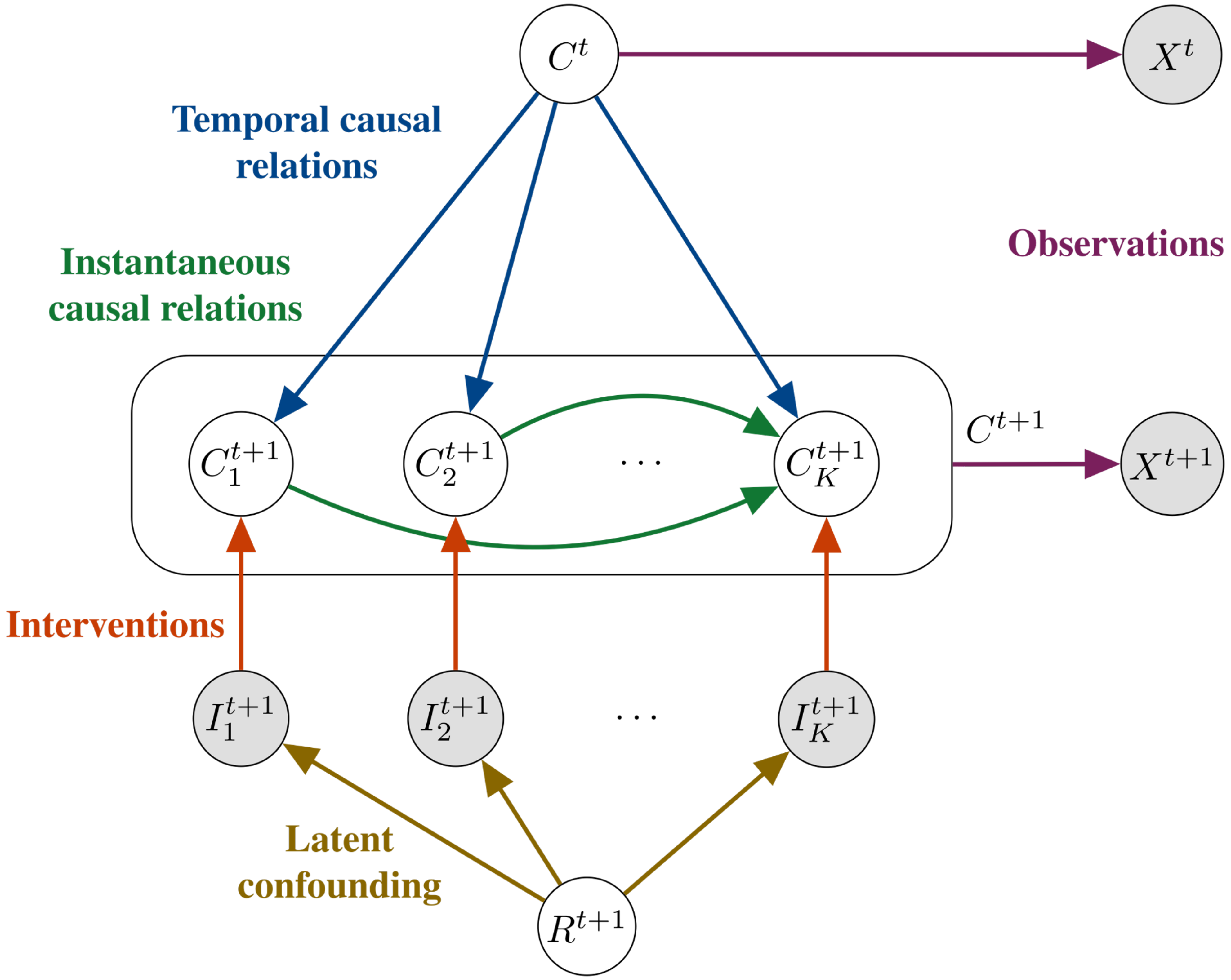




Causal graph learnt with CITRIS-NF

Instantaneous effects: iCITRIS [Lippe et al 2023a]





- Soft interventions are not enough to disentangle instantaneous "components"
  - **Partially perfect interventions**: a soft intervention that is perfect in terms of instantaneous parents

- Estimate jointly the causal variables and the graph

Summary CITRIS & iCITRIS

- **Pros:**
  - Multidimensional causal variables
  - No parametric assumptions
  - Work with arbitrary graphs, **even instantaneous effects**
  - **Identifiability up to component-wise transformations**

- **Cons:**
  - Need (sufficiently diverse) interventional data
  - **Need known intervention targets -> can we get rid of this?**

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use actions

Strategy 4: assume temporal data

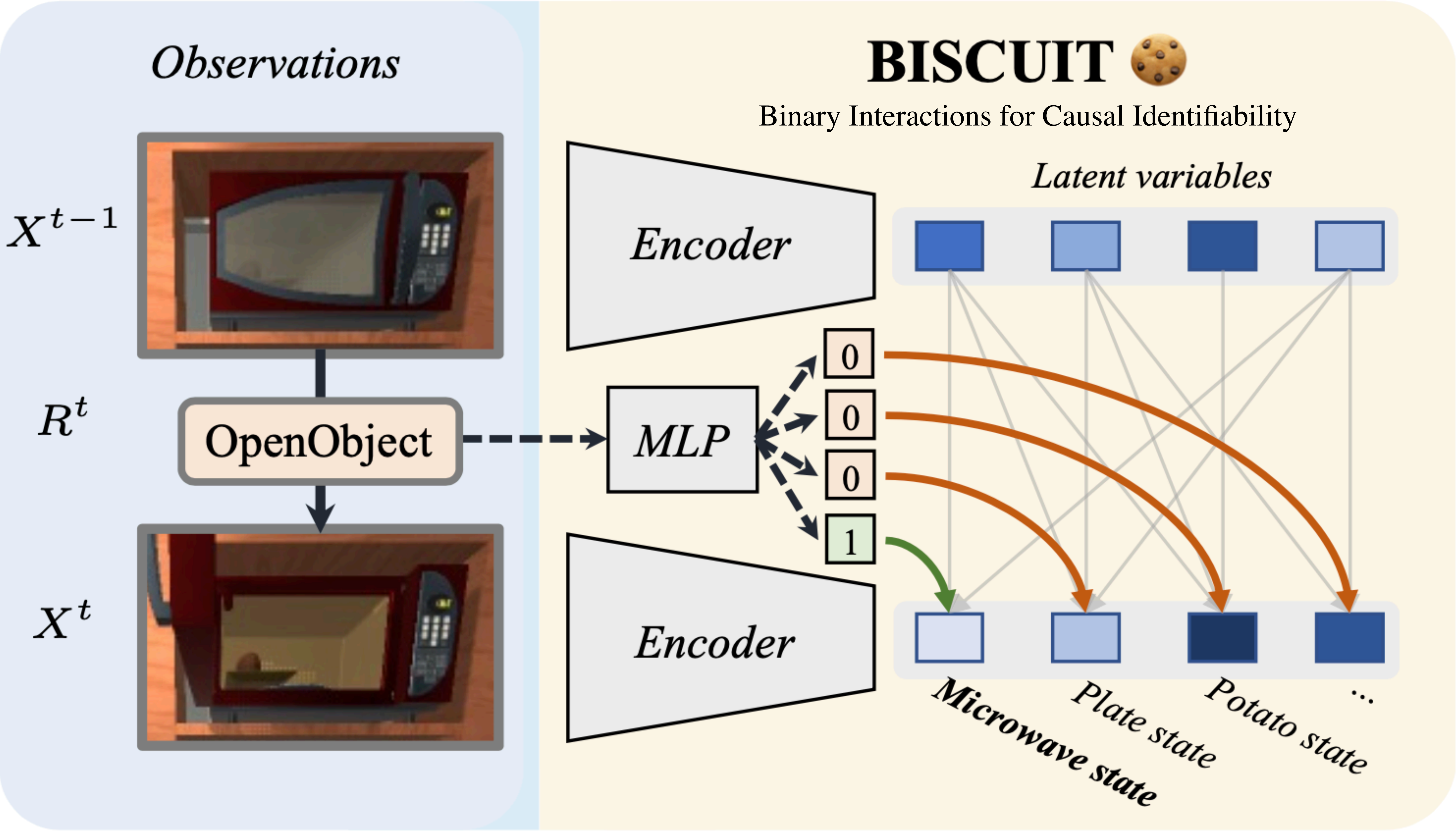
Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

Output 3: Up to differentiable component-wise transformations + permutation

BISCUIT: Causal Representation Learning from Binary Interactions

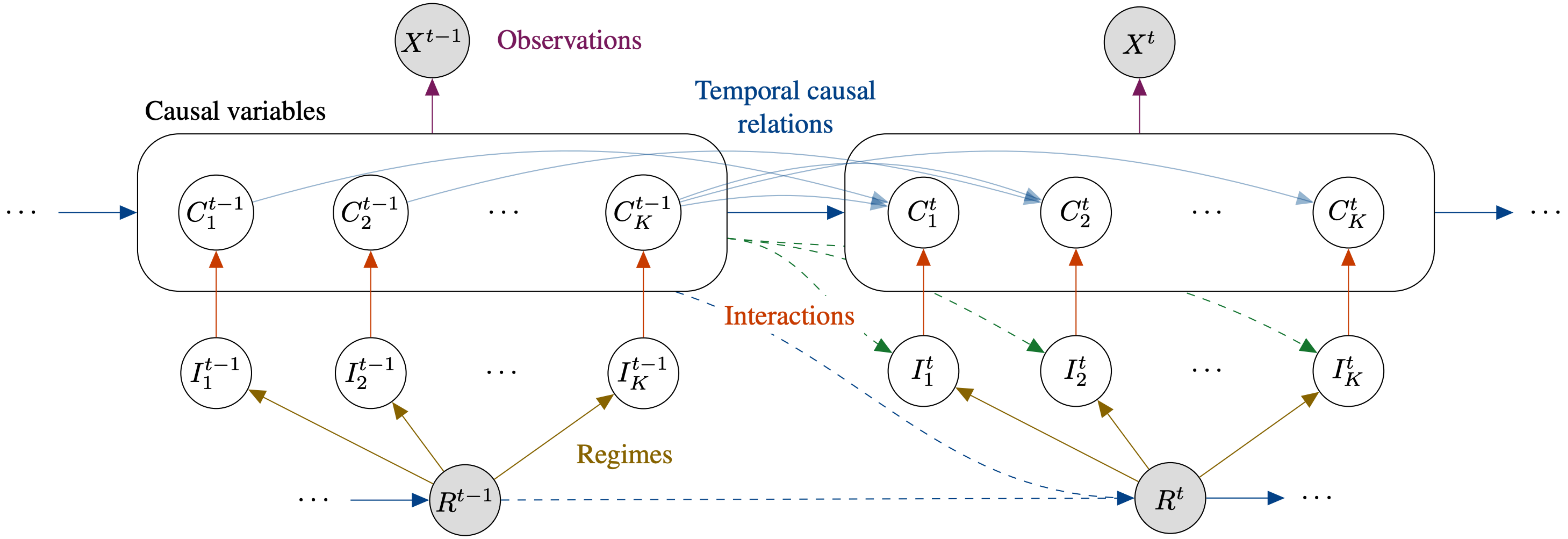
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves



https://phlippe.github.io/BISCUIT/

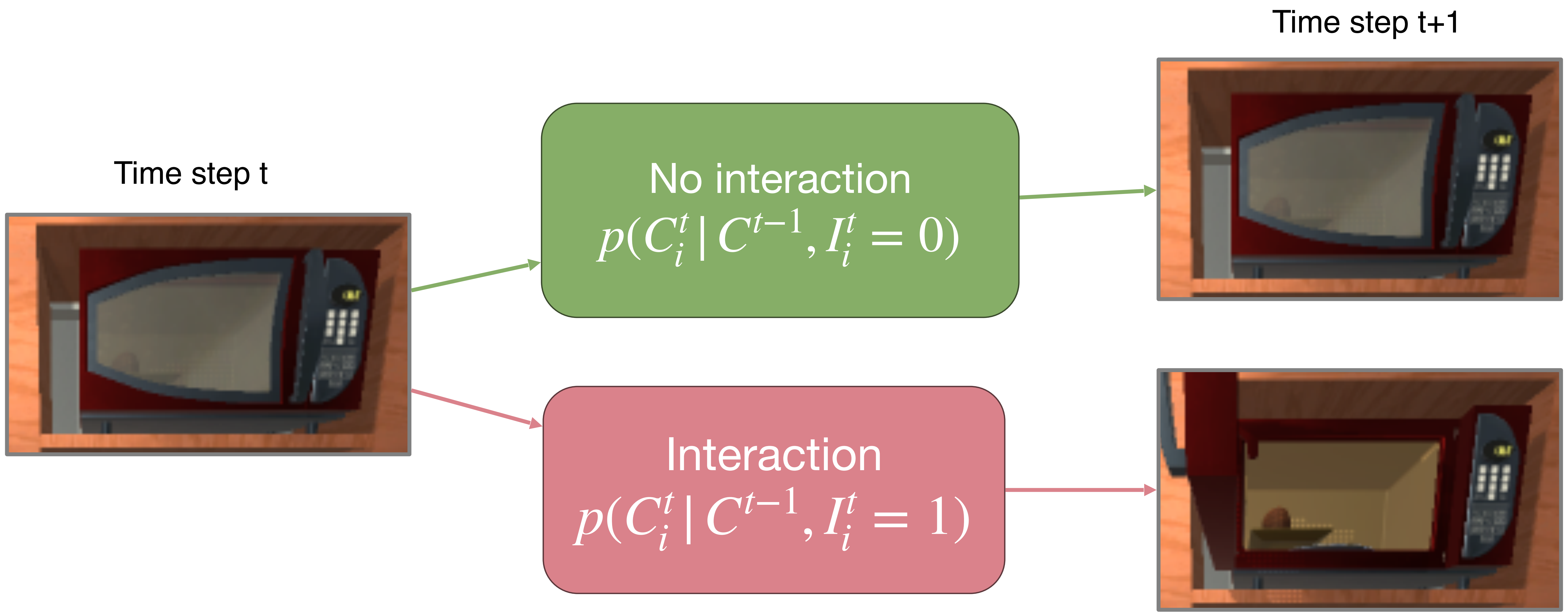
An extension of TRIS: the BISCUIT model

- The binary **intervention variables** are unobserved, but we **observe an action/regime R^t**
- The regime R^t can be **caused by the previous state C^{t-1} and previous regime R^{t-1}**
- We assume **the effect of R^t can be encoded in binary interaction variables $I^t = f(R^t, C^{t-1})$**



Assumption 1: Action/Regime can be encoded in binary interactions

- **Assumption 1:** interactions between the regime and each causal variable can be described by a binary variable (although the binding can change across timesteps based on state)
  - Each causal variable has exactly two mechanisms (same as CITRIS)

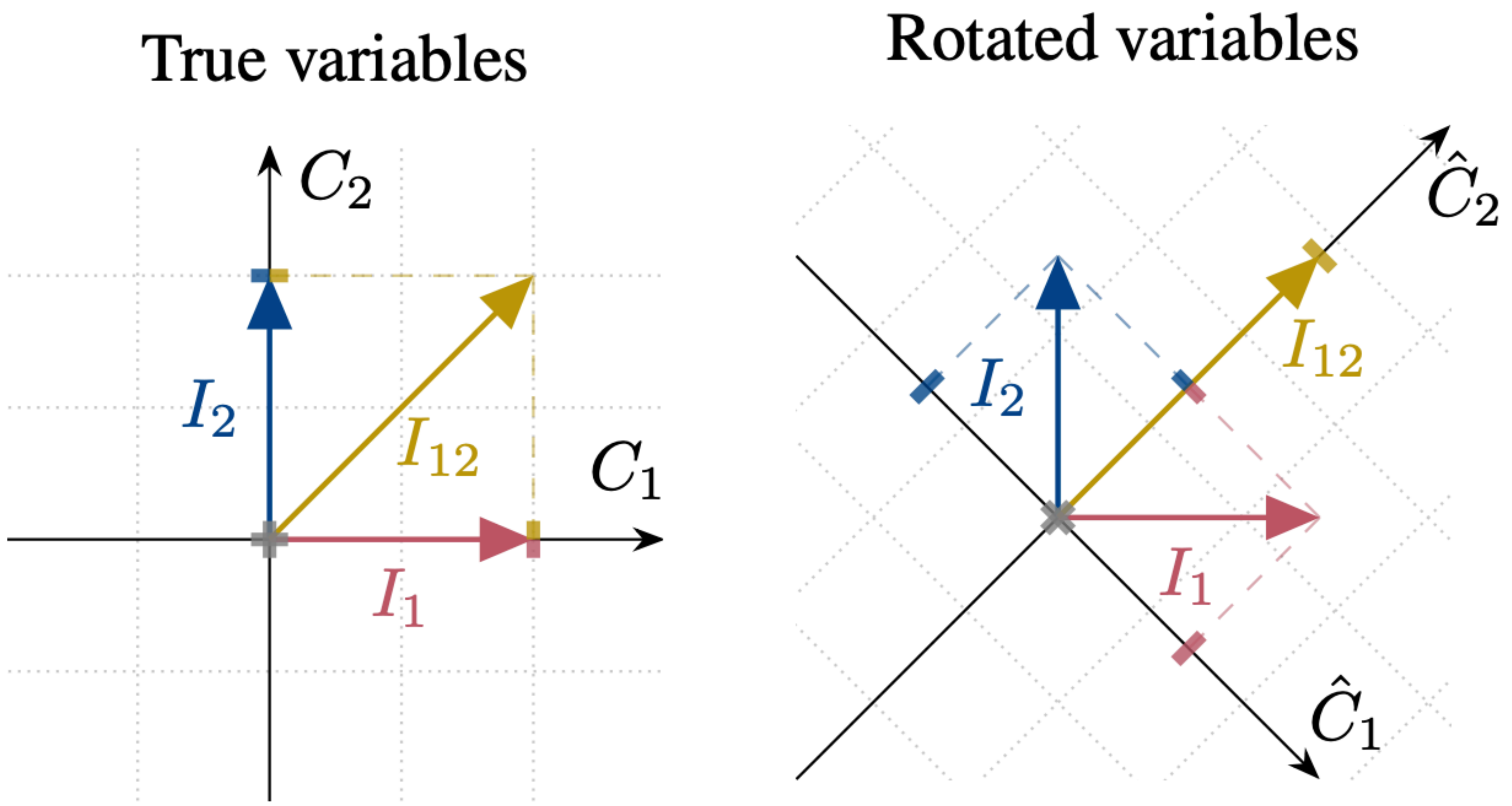


Another example: collisions between agent and objects that change dynamics of objects

These can depend on previous state (position

Binary interactions - The power of Two

- Knowing each variable has only two mechanisms helps identify difficult cases
- Example: Additive Gaussian, both true and rot. variables model the same distribution, but under interventions, only the true variables have two means



https://phlippe.github.io/BISCUIT/

Assumption 2: Distinct interaction patterns

- A causal variable C_i has a **distinct interaction pattern**, if $I_i^t = f_i(C^{t-1}, R^t)$ is **not a function of any other interaction variable I_j^t**

- **Intuitively:** if we always intervene and perturb two objects at the same time, we will not get enough information from the perturbation to distinguish them.

- If I_i^t are independent of C^t (as in CITRIS), we only need $O(logK)$ distinct values of R^t for full identifiability

BISCUIT - identifiability

- **Assumption 1:** binary interactions
- **Assumption 2:** distinct interaction patterns with “enough” types of interactons
- **Assumption 3:** the mechanisms vary sufficiently either over interactions or time

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

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

similar to ICA

$$\Delta(C_i^t \,|\, C^{t-1}) = \log \frac{p(C_i^t \,|\, C^{t-1}, I_i^t = 1)}{p(C_i^t \,|\, C^{t-1}, I_i^t = 1)}$$

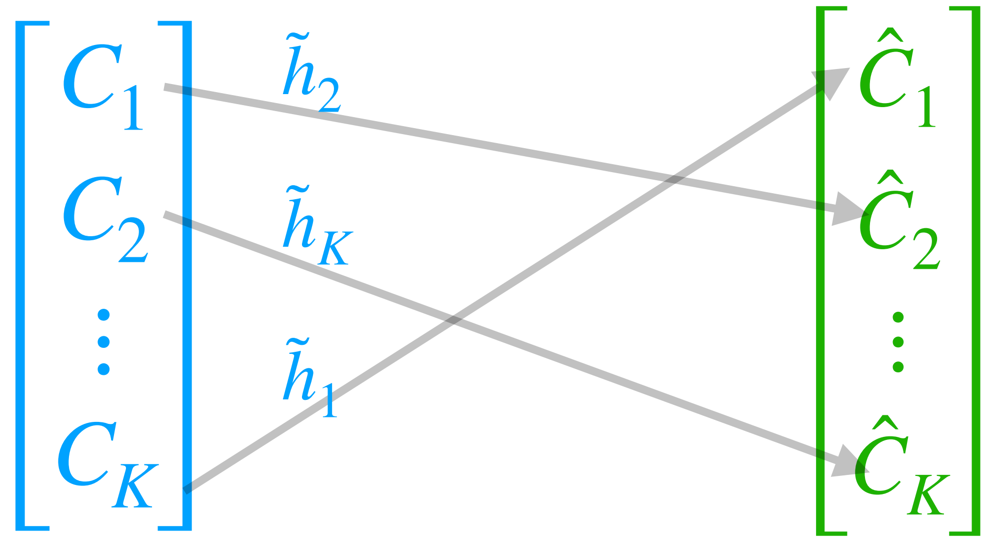
B. (***Time Variability***) *For any $C^t \in \mathcal{C}$, there exist $K+1$ different values of C^{t-1} denoted with $c^1, ..., c^{K+1} \in \mathcal{C}$, for which the vectors $v_1, ..., v_K \in \mathbb{R}^{K+1}$ with*

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

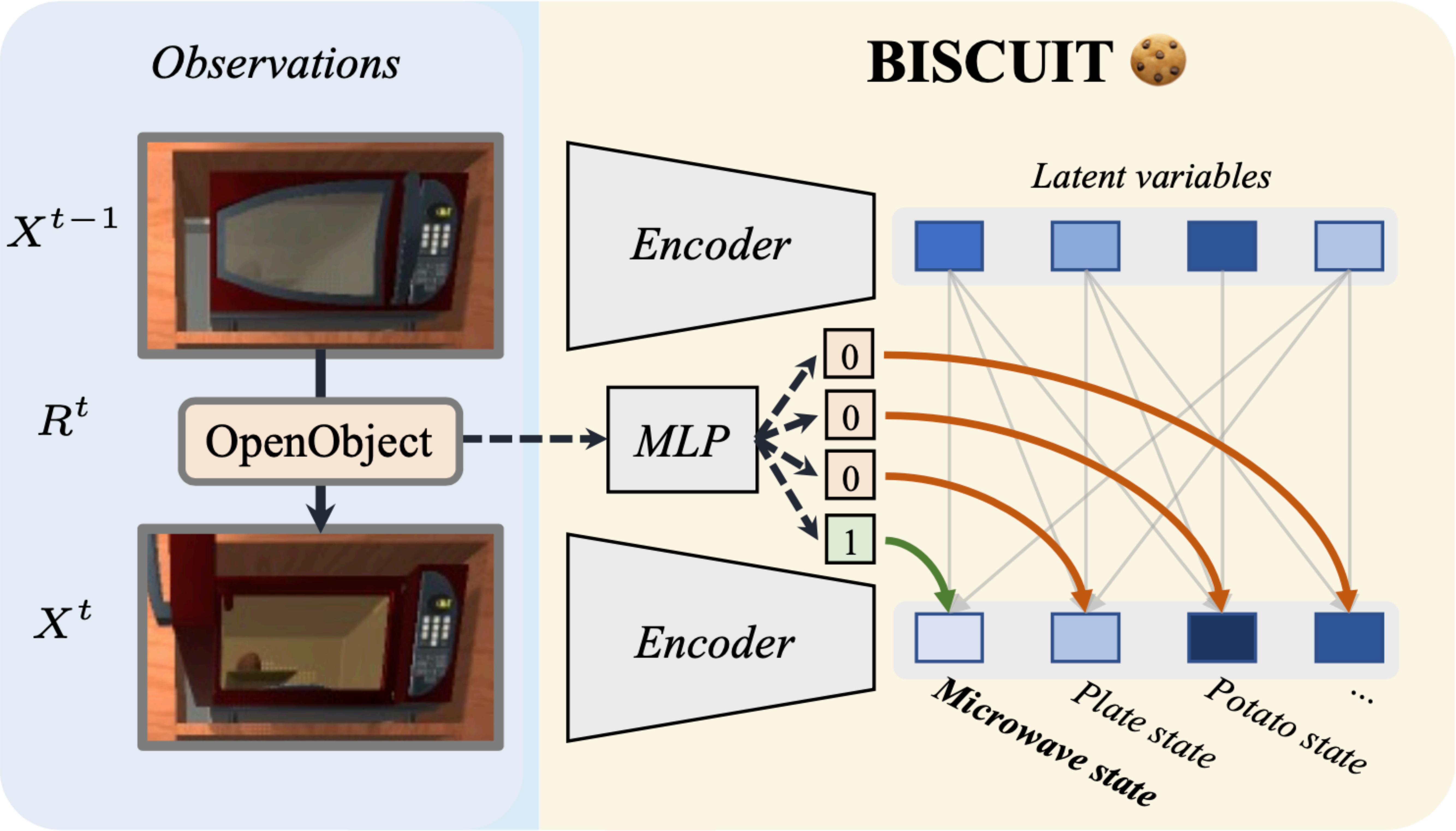
are linearly independent.

Effect of interaction given previous state is different across causal variables

$\Longrightarrow$ Maximising likelihood allows for identifiability up to permutation and component-wise transformations



BISCUIT architectures



BISCUIT-VAE

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

Transition prior:

$$p_\omega(z^t|z^{t-1}, R^t) = \prod_{i=1}^{M} p_{\omega,i}\left(z_i^t|z^{t-1}, \mathrm{MLP}_\omega^{\hat{I}_i}(R^t, z^{t-1})\right)$$

BISCUIT-NF

Leverage pretrained autoencoder
+ normalizing flows

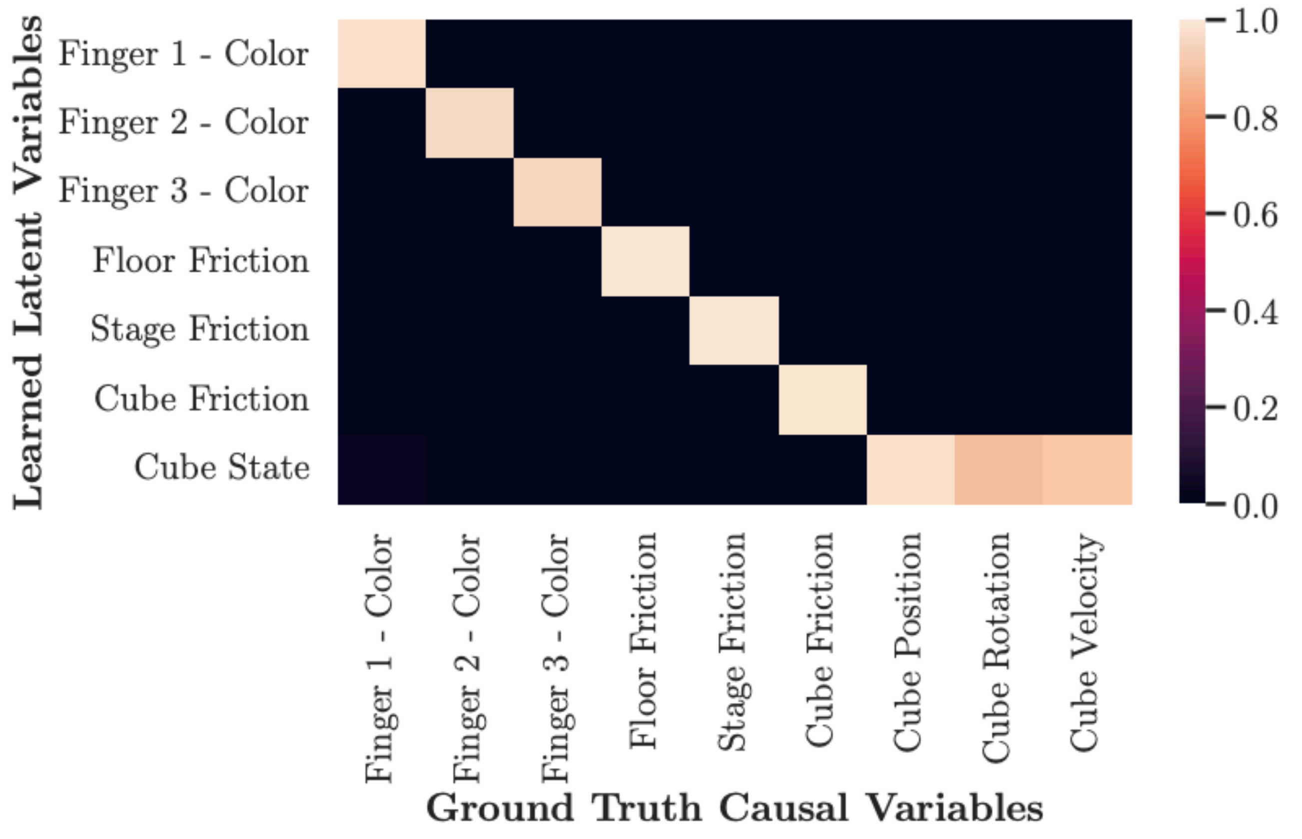
BISCUIT on CausalWorld and iTHOR

- CausalWorld - three finger robot manipulating objects
  - Variables: object position, frictions, colors, etc.
  - Action: 9-dimensional motor angles (3 per finger)
- iTHOR - kitchen environment, action is (x,y) position of click

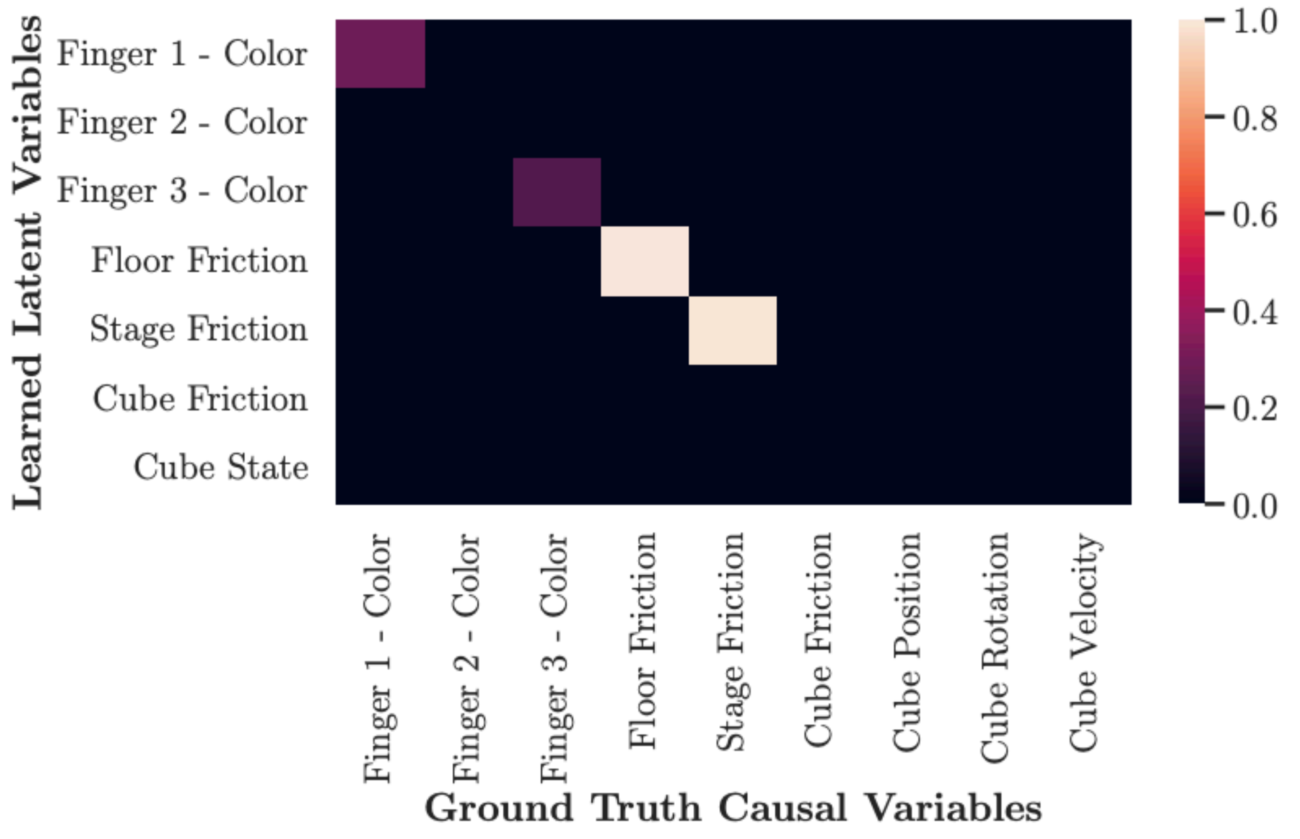
Table 1: R^2 scores (diag ↑ / sep ↓) for the identification of the causal variables on CausalWorld and iTHOR.

Models	CausalWorld	iTHOR
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00	0.48 / 0.35
LEAP (Yao et al., 2022b)	0.30 / 0.00	0.63 / 0.45
DMS (Lachapelle et al., 2022b)	0.32 / 0.00	0.61 / 0.40
BISCUIT-NF (Ours)	**0.97** / 0.01	**0.96 / 0.15**

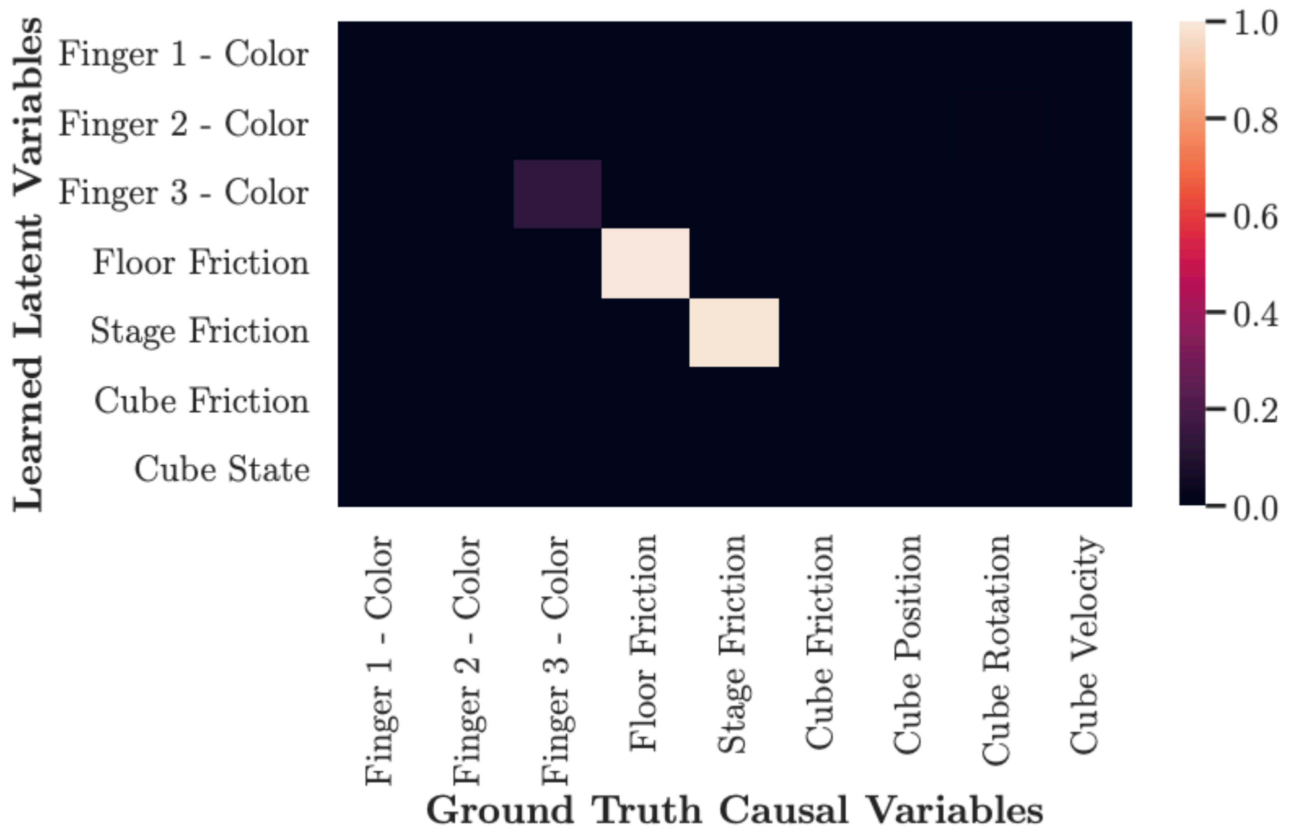
BISCUIT on CausalWorld - R^2 metric



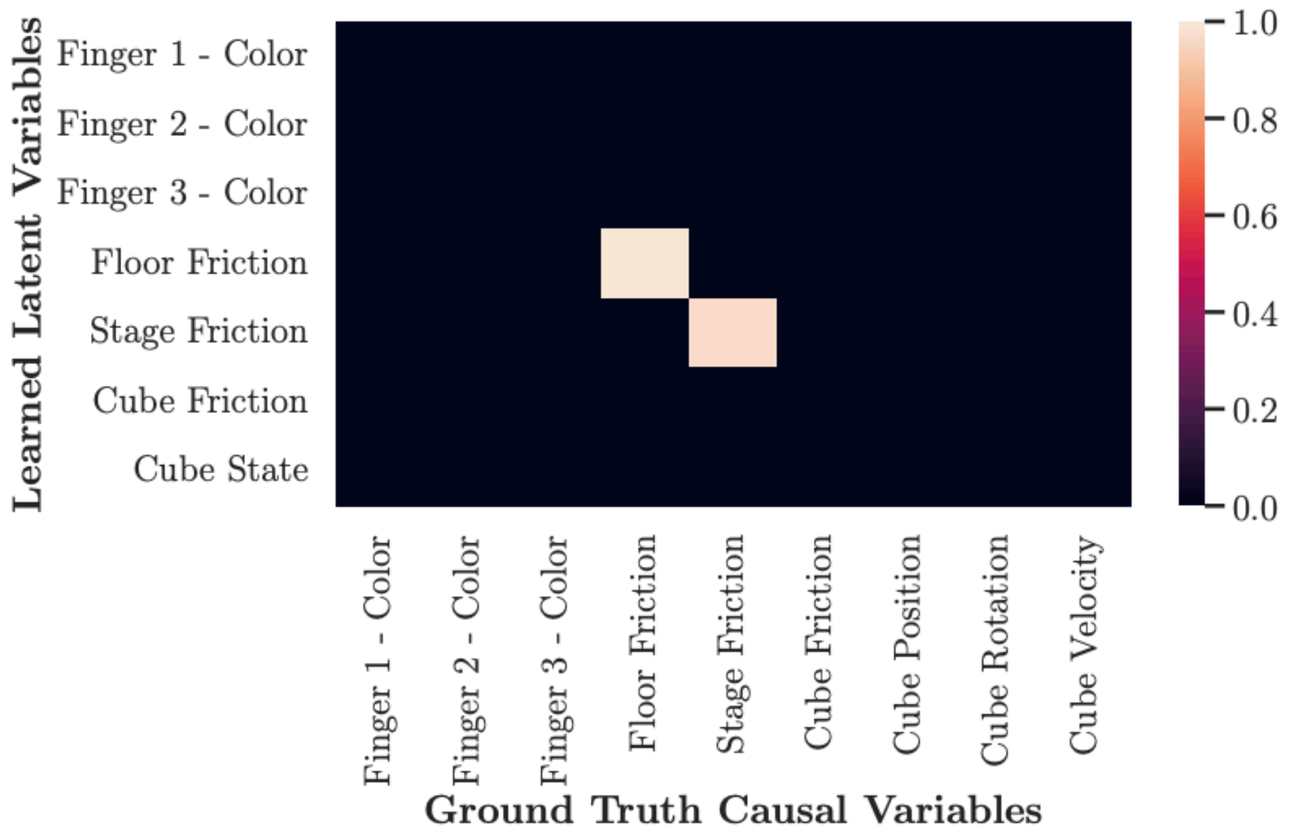
(a) BISCUIT



(b) DMS (Lachapelle et al., 2022b)



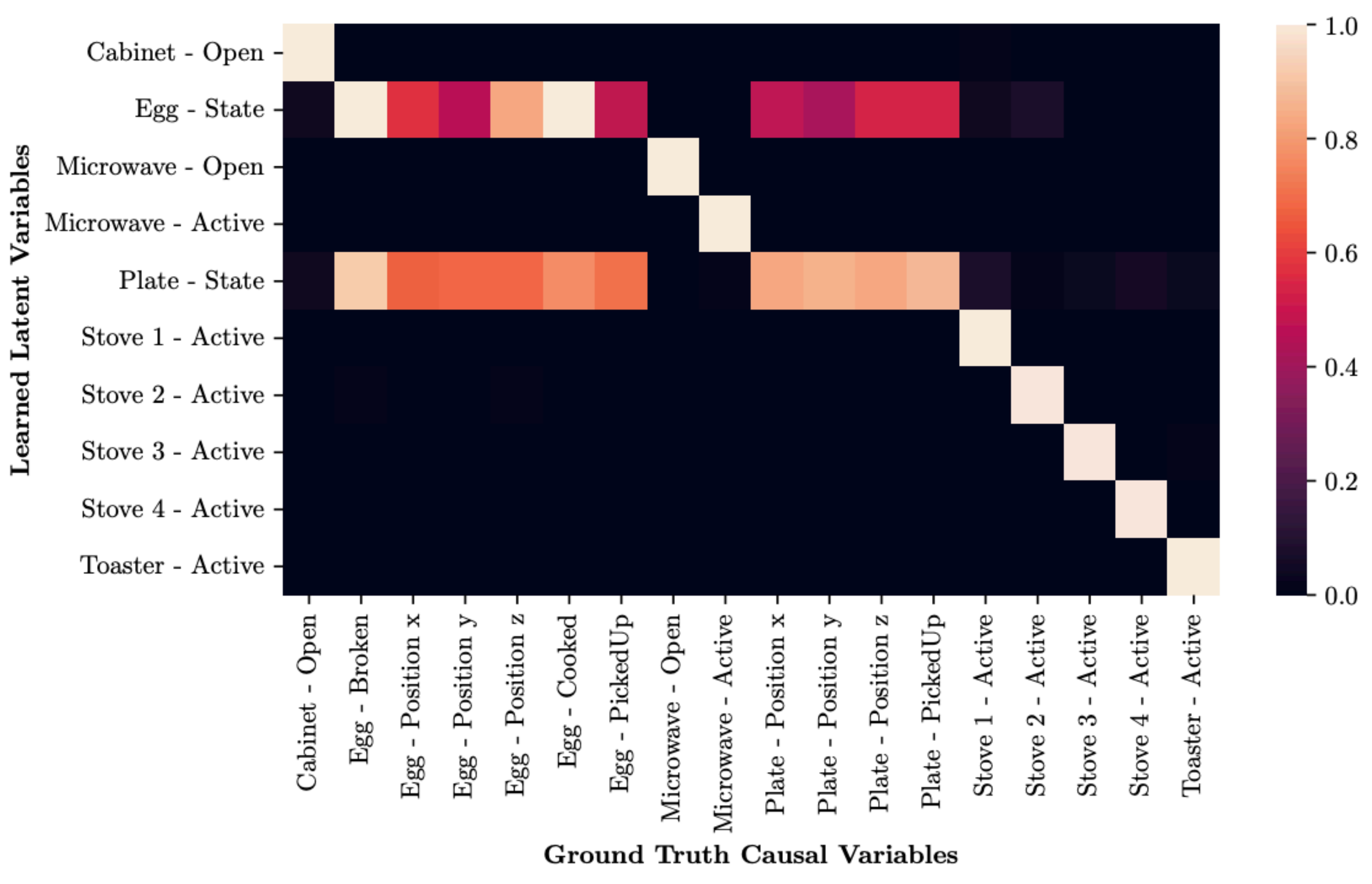
(c) LEAP (Yao et al., 2022b)



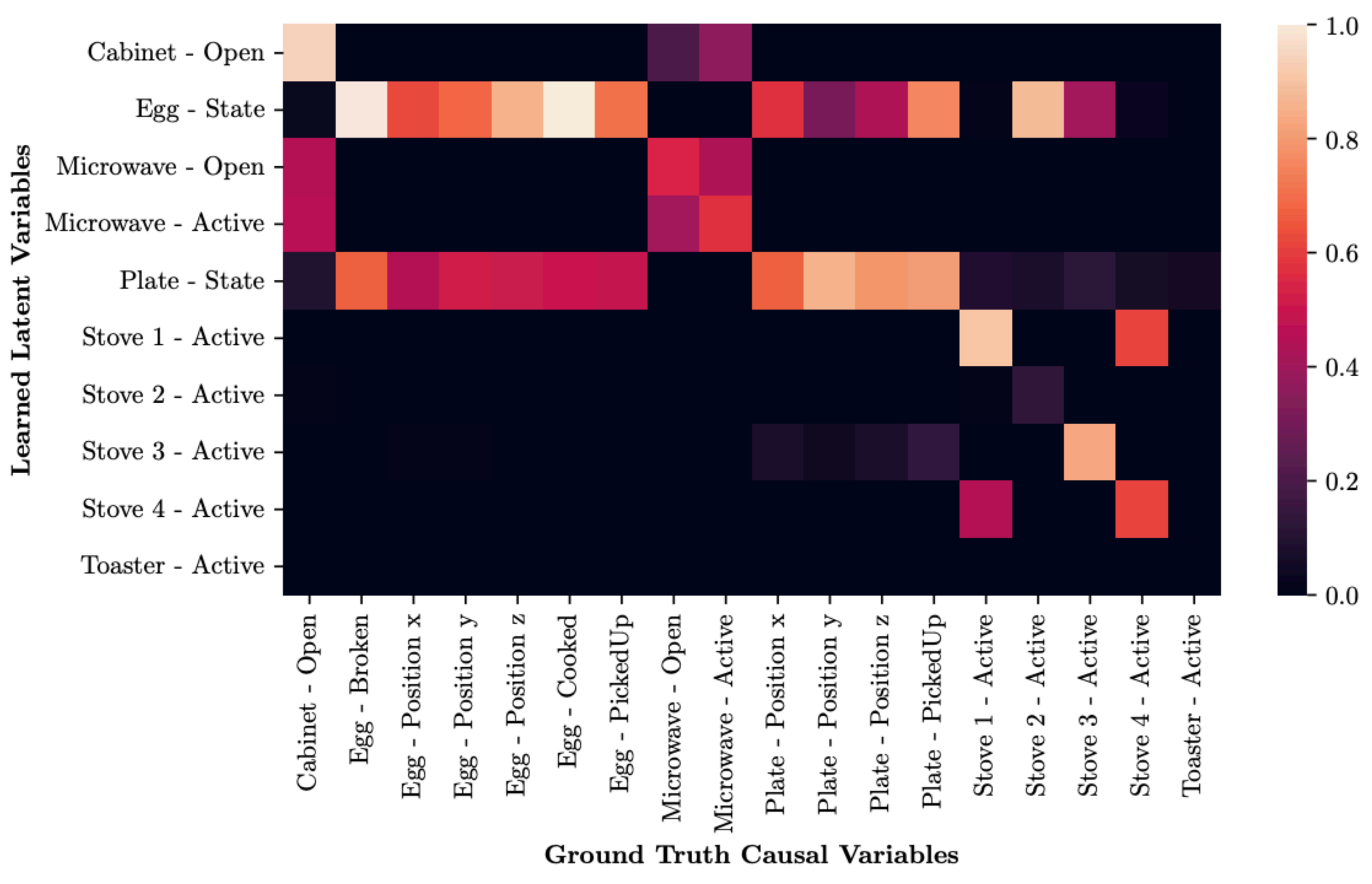
(d) iVAE (Khemakhem et al., 2020a)

Here we assign the permutation based on the most correlated latent variable

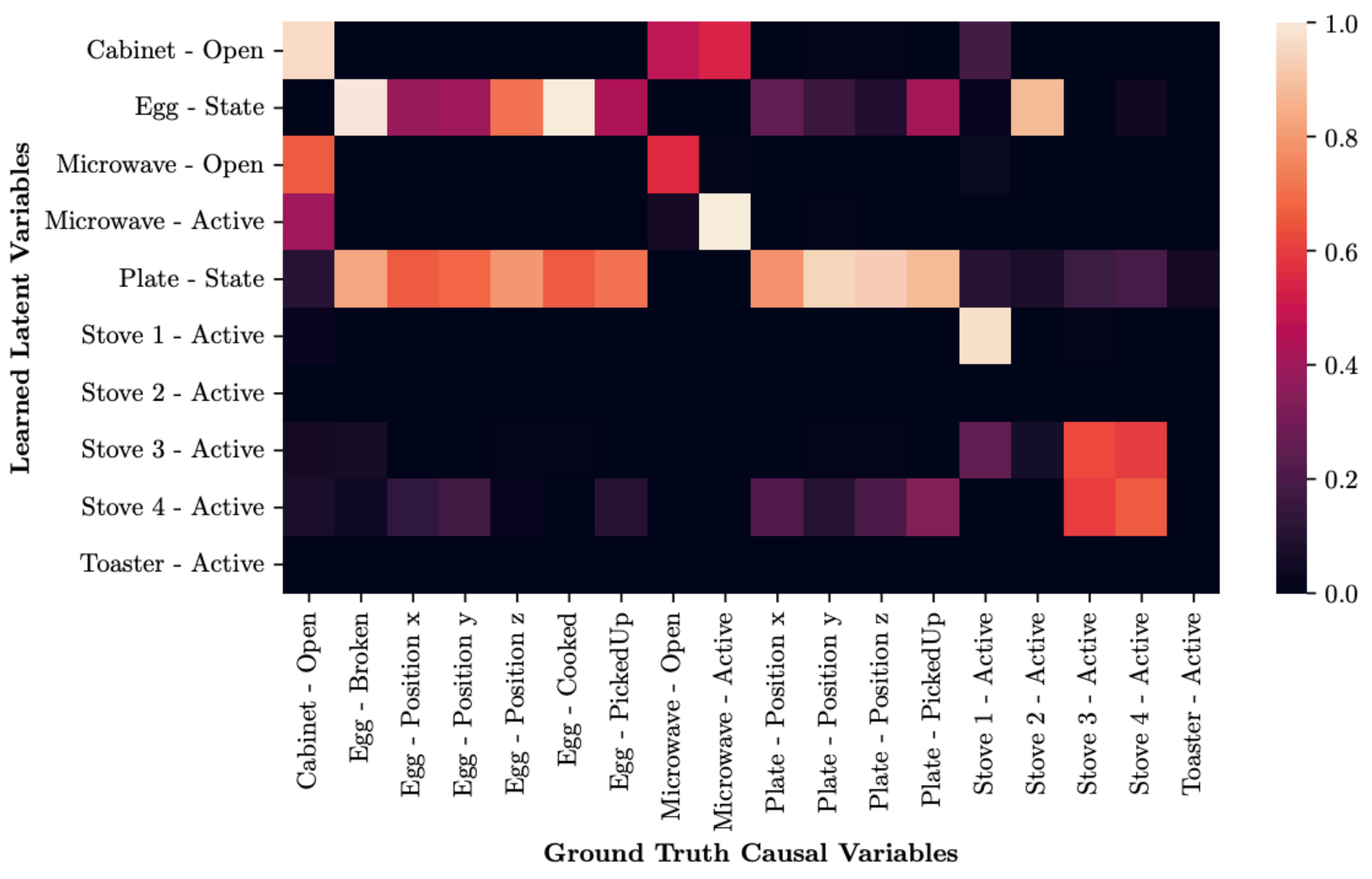
BISCUIT on iTHOR - R^2 metric



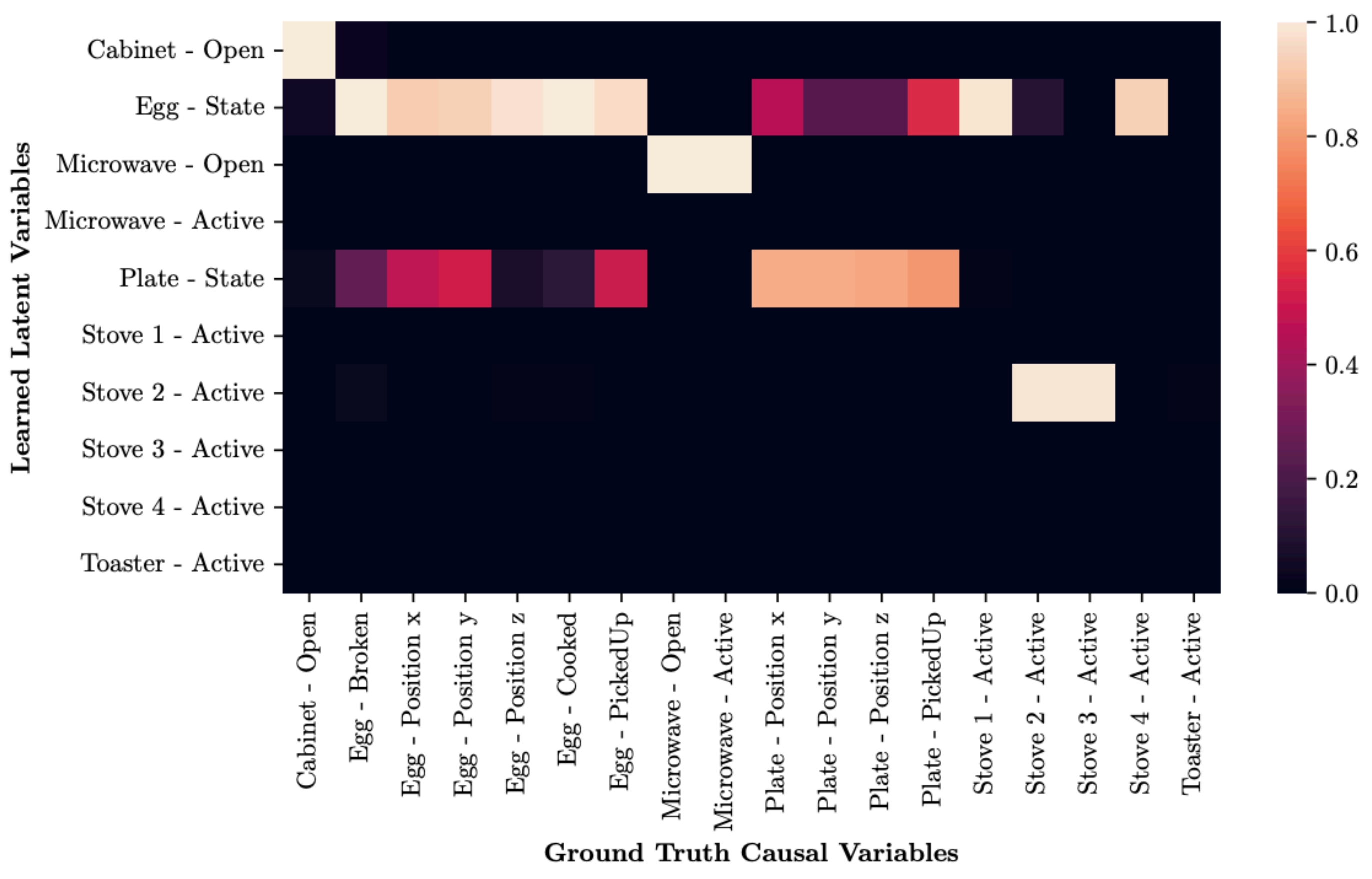
(a) BISCUIT



(b) DMS (Lachapelle et al., 2022b)



(c) LEAP (Yao et al., 2022b)



(d) iVAE (Khemakhem et al., 2020a)

Here we assign the permutation based on the most correlated latent variable

BISCUIT on iTHOR - Generating new images

Input image 1

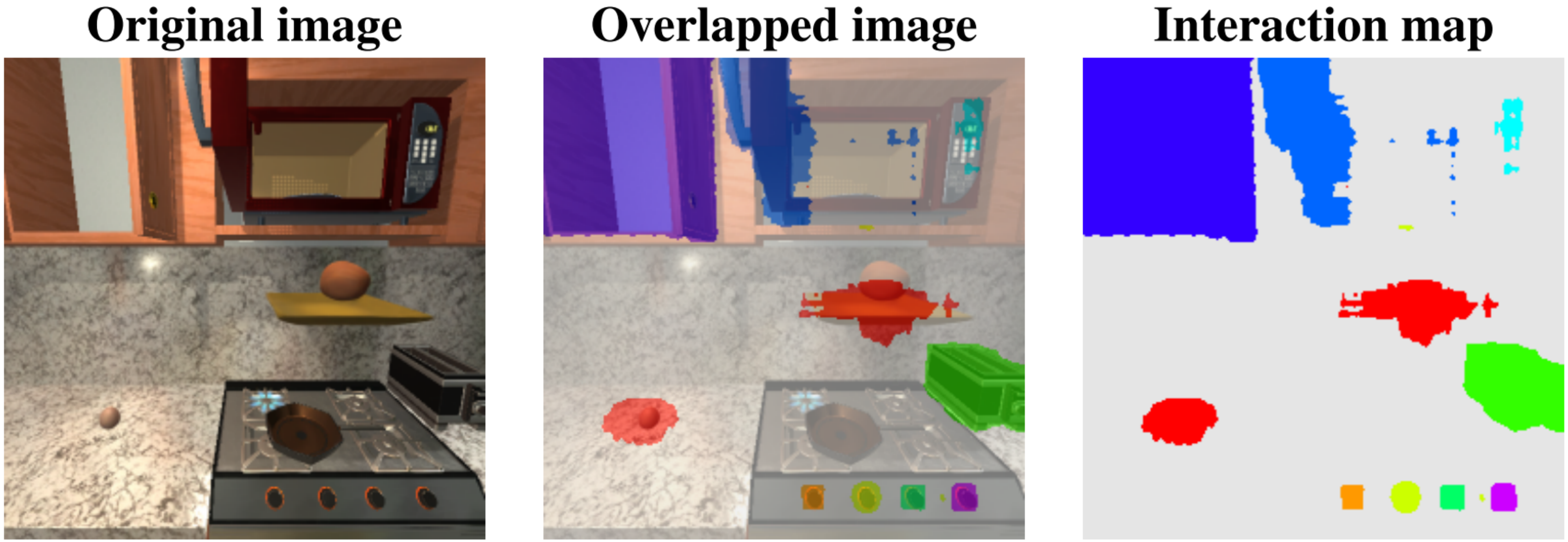
Input image 2

Generated Output

Latents from image 2

Microwave Open

BISCUIT on iTHOR - dynamic interaction map



BISCUIT [Lippe et al. 2023]

- **Pros:**
  - No parametric assumptions
  - Works also with arbitrary dense graphs
  - **No need for intervention targets**

- **Cons:**
  - Needs (sufficiently diverse) interventional data
  - Works only in a temporal setting
  - **Assumes binary interactions**

Causal Representation Learning overview

Task 1: learn the causal variables

Task 2: learn the causal graph (or equivalence class)

Strategy 1: restrict SCM (e.g. linear, Gauss/exp) or mixing function (e.g. piecewise/linear)

Strategy 2: multi-view or paired data

Strategy 3: use interventions/actions

Strategy 4: assume temporal data

Output 1: Up to permutation and scale

Output 2: Up to differentiable component-wise transformations

Output 3: Up to differentiable component-wise transformations + permutation

Conclusion

- Causal representation learning (CRL) is an exciting new field that allows us to identify causal variables from unstructured data in different settings

- **Open questions:**
  - How can we use CRL for real-world downstream tasks, e.g.
    - Causal effect estimation when we have unstructured data
      - -> needs methods that for observational data with milder assumptions
    - RL/ robotics
      - -> needs methods that can work with general actions
    - XAI or mechanistic interpretability
      - -> need scaling + methods for discrete concepts