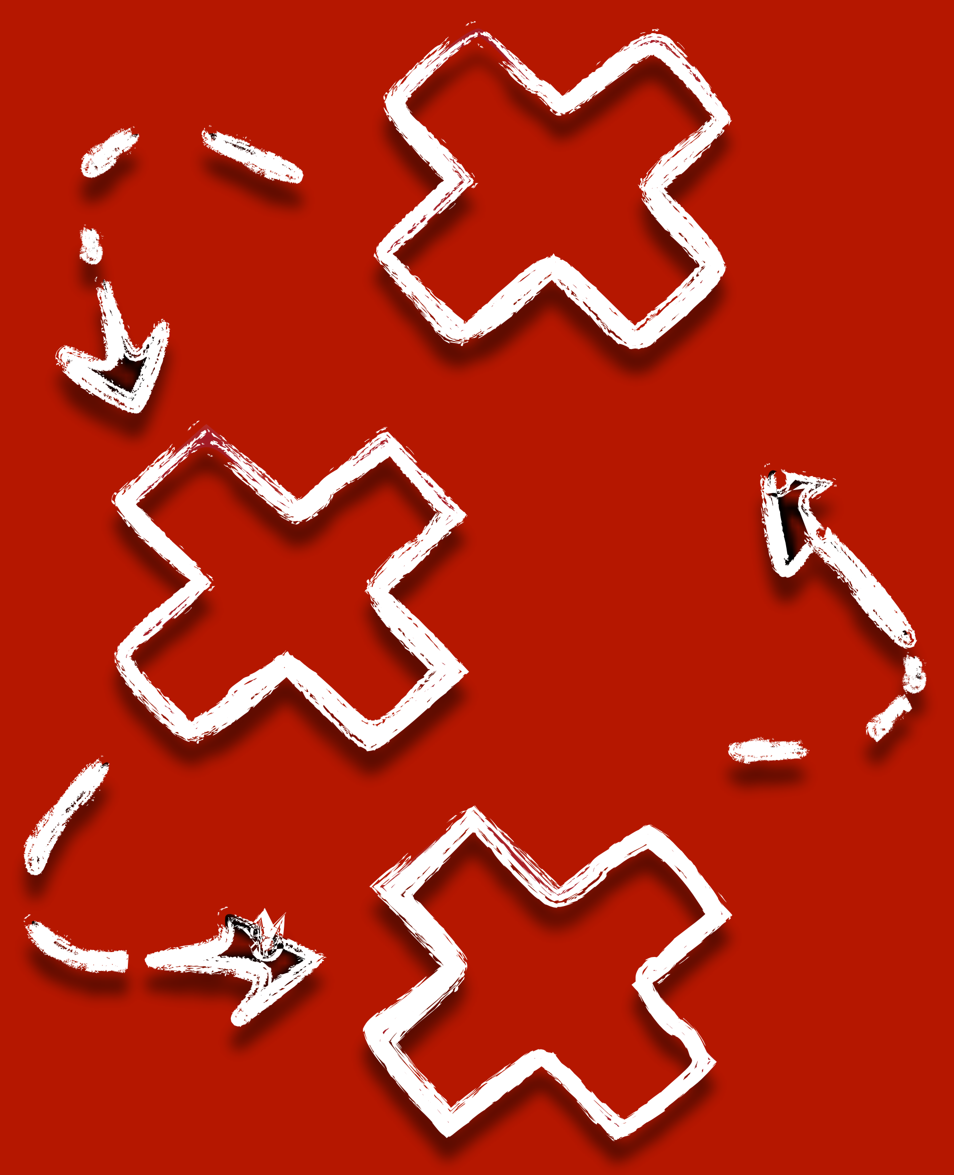


Causal representation learning in temporal settings with actions

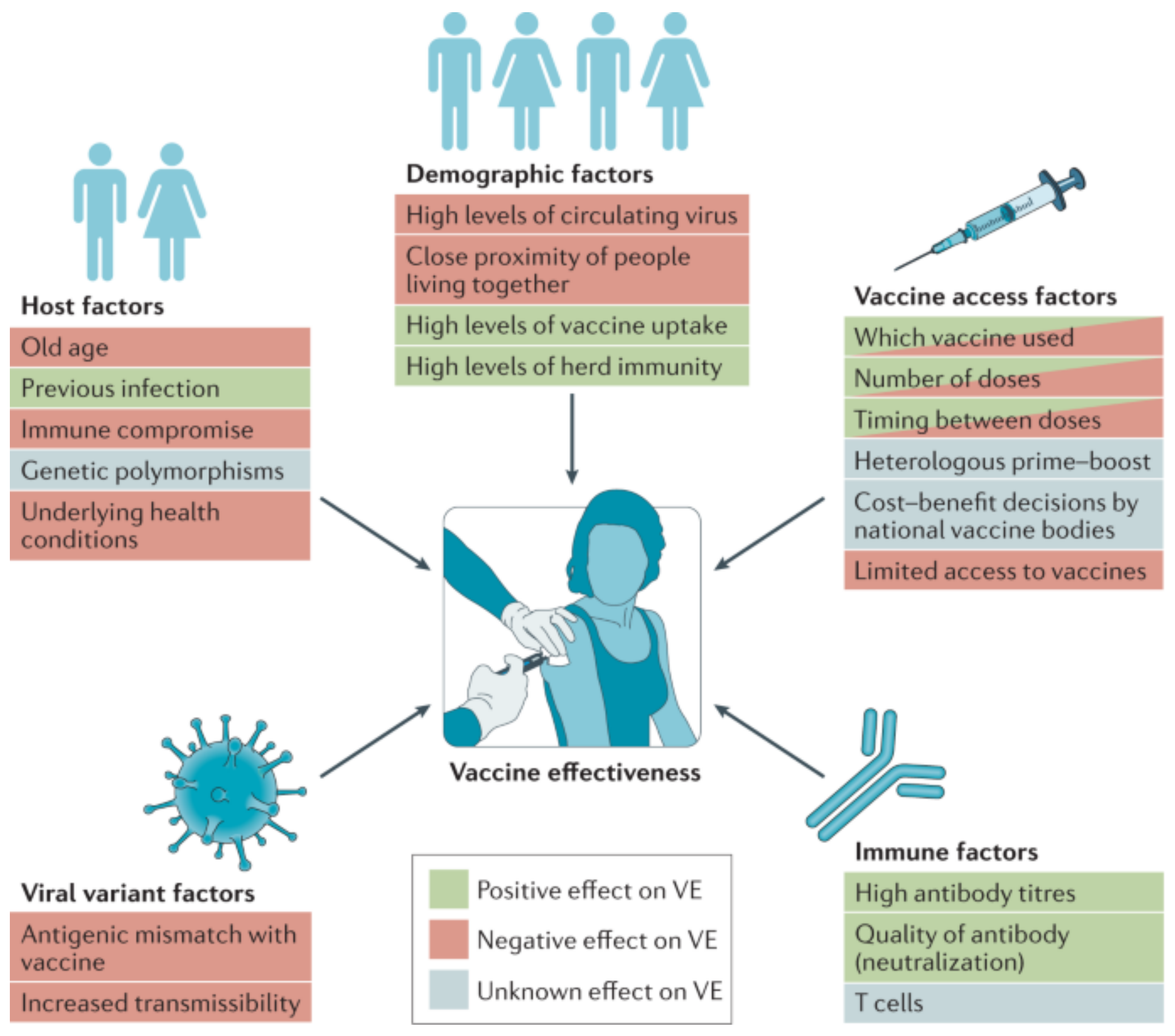
Sara Magliacane (University of Amsterdam)

Causal questions are ubiquitous

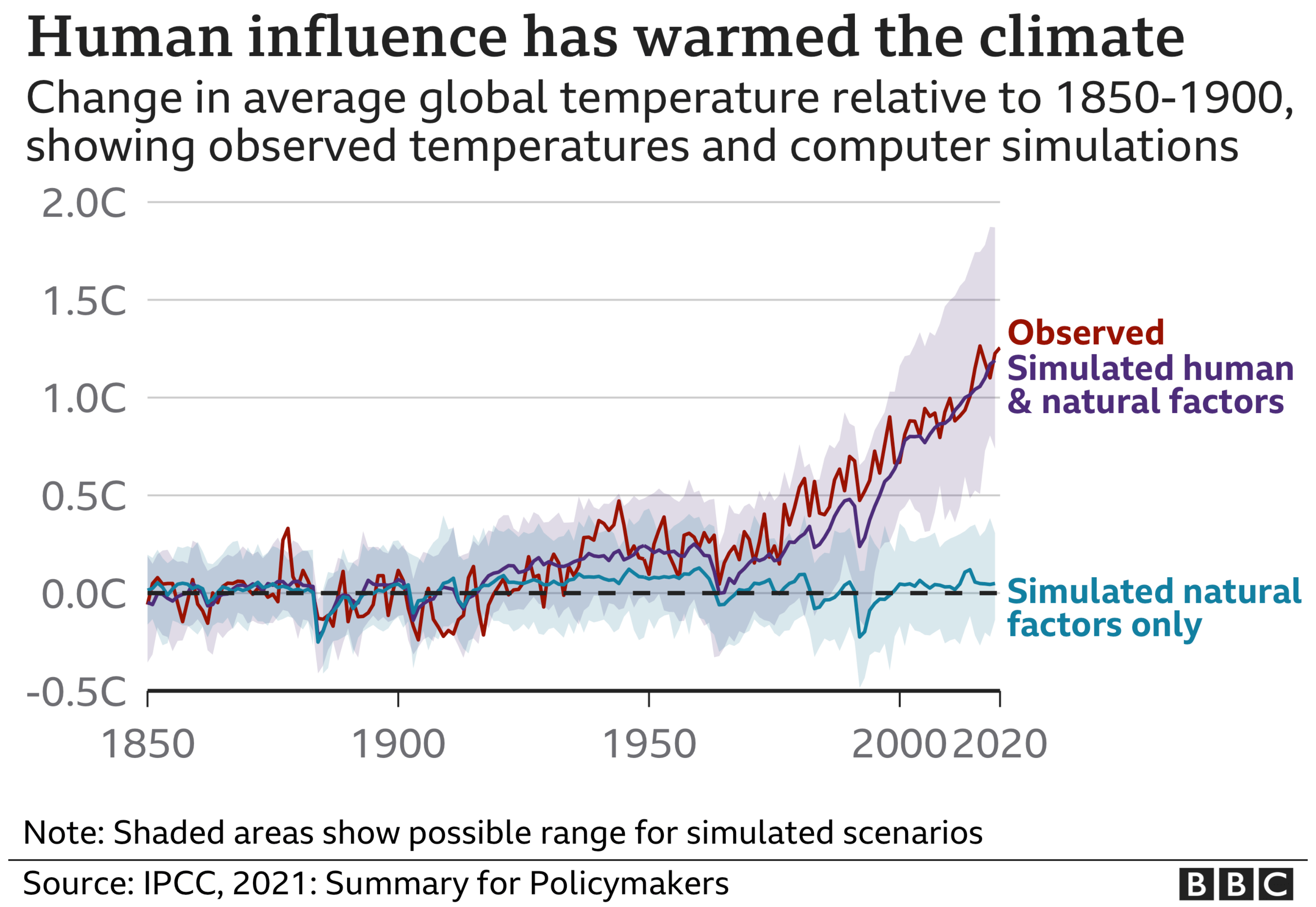
- To predict the effect of actions and decide effective policies, we need to understand: 1) **what causes what** and 2) **how**?

Causal questions are ubiquitous

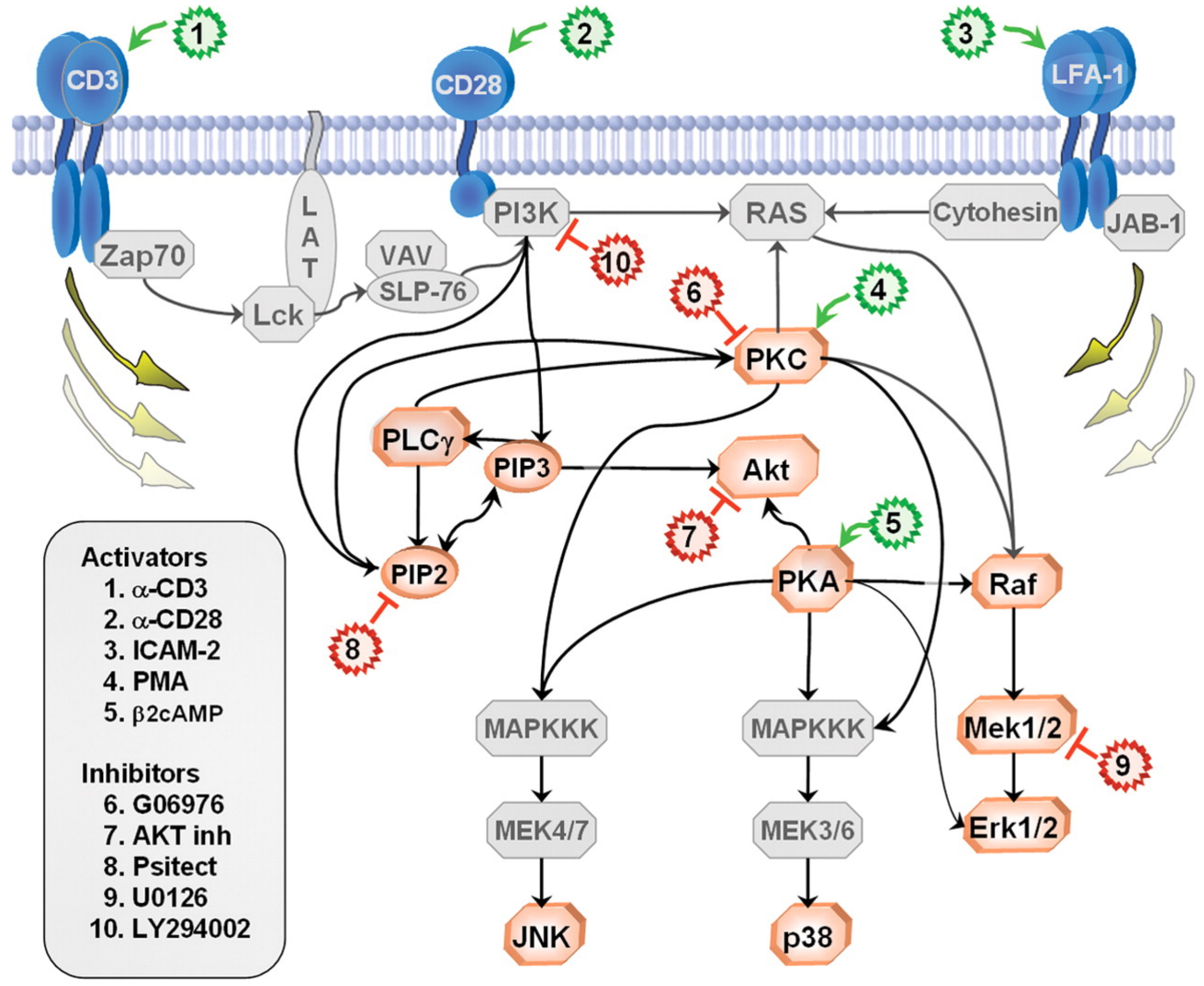
- To predict the effect of actions and decide effective policies, we need to understand: 1) **what causes what** and 2) **how**?



https://www.nature.com/articles/s41577-021-00592-1



https://www.bbc.com/news/science-environment-58600723



https://www.science.org/doi/abs/10.1126/science.1105809

Vaccine effectiveness

Climate change policy

Protein signalling networks

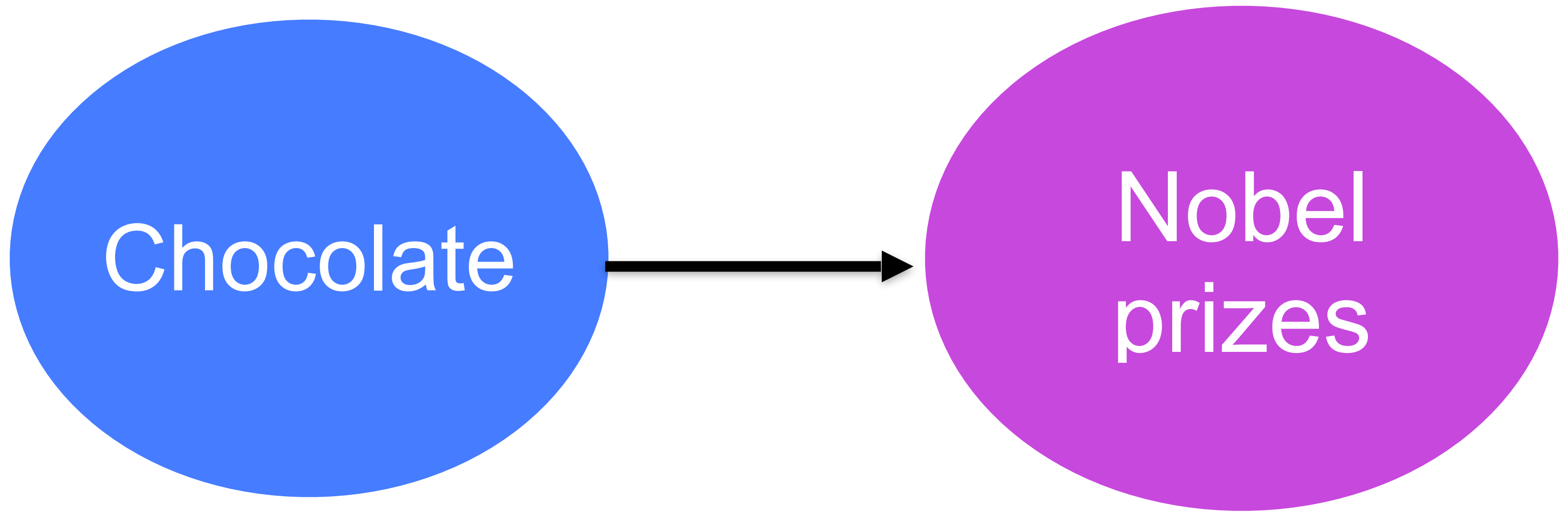
A working definition of causality in machine learning

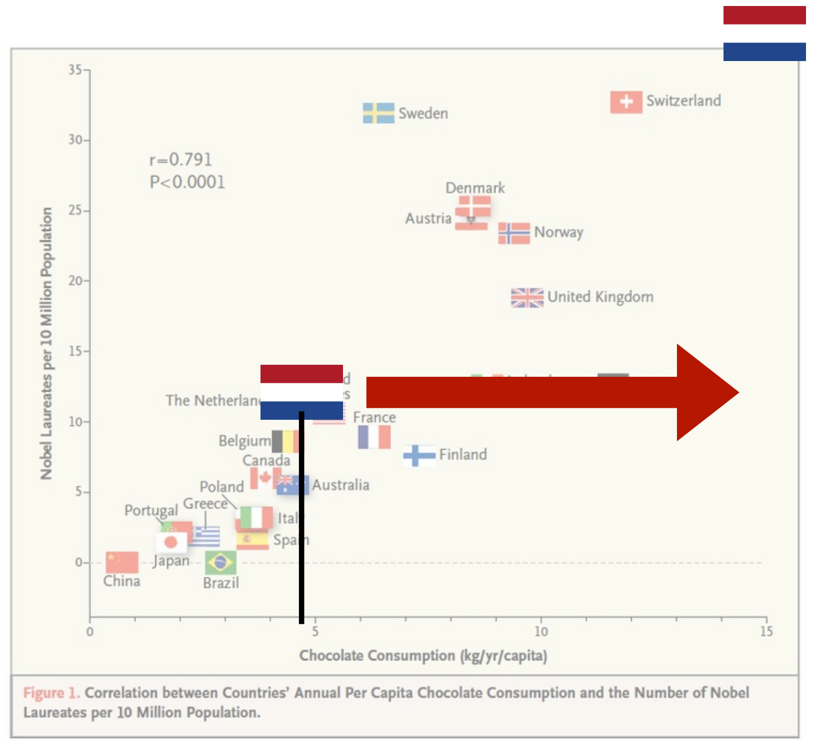
Informal definition: A variable X causes another variable Y, if changing (the distribution of) X, e.g. by fixing its value, changes (the distribution of) Y

Intervention

Challenge: estimate the causal effect of an intervention, when we do not have (all possible) interventional data **(e.g. observational data)**

Representation: We can represent causal relations in **causal graphs:** nodes are random variables, edges causal relations





Can we learn causal variables from unstructured 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

Causal Representation Learning (CRL)

Can we **predict the effect of interventions** if the causal variables are **not directly observed** and we **do not have labels for them**, but we have **high-dimensional observations of the system**?

Causal Representation Learning (CRL)

Can we **predict the effect of interventions** if the causal variables are **not directly observed** and we **do not have labels for them**, but we have **high-dimensional observations of the system**?

Task 1: identify/disentangle the causal variables from observations

Causal Representation Learning (CRL)

Can we **predict the effect of interventions** if the causal variables are **not directly observed** and we **do not have labels for them**, but we have **high-dimensional observations of the system**?

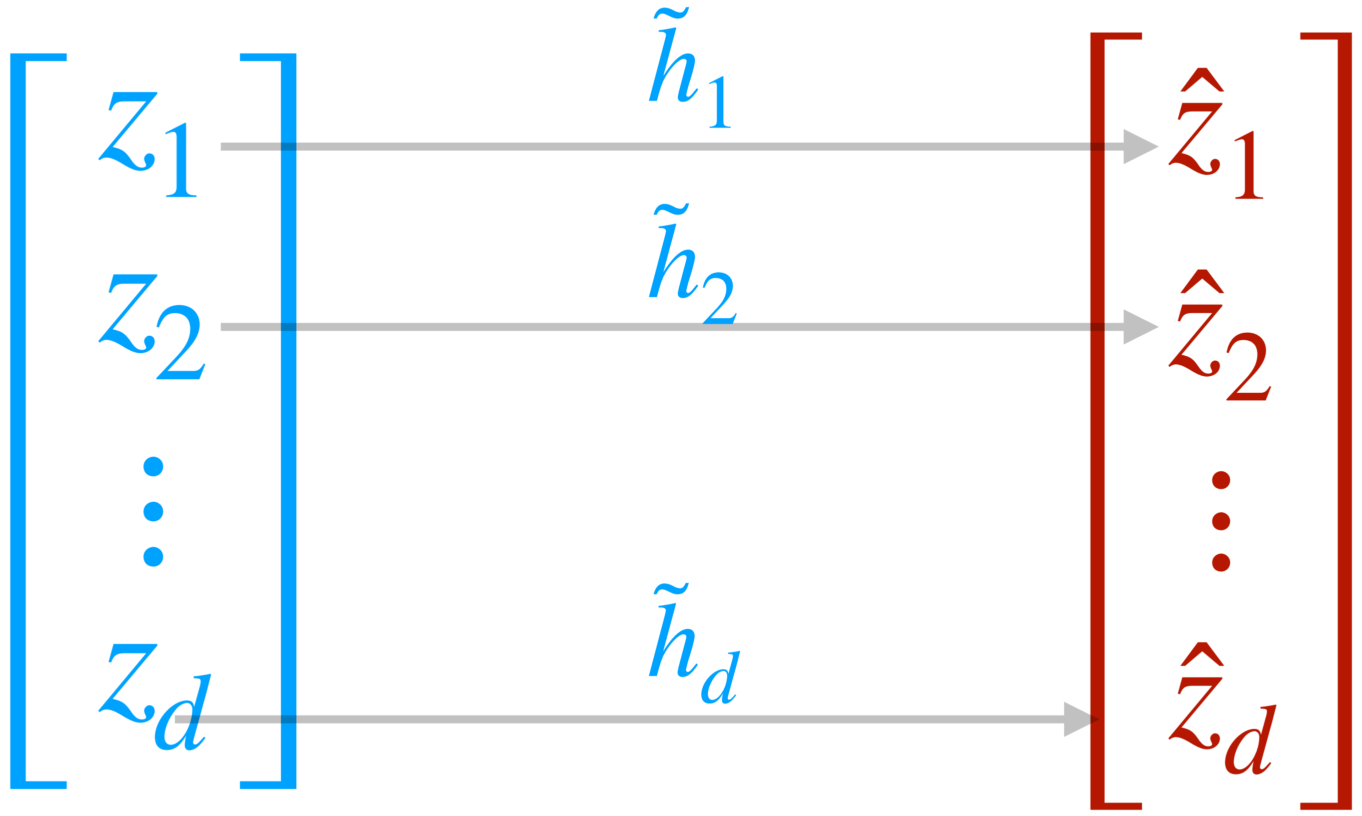
Task 1: identify/disentangle the causal variables from observations

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

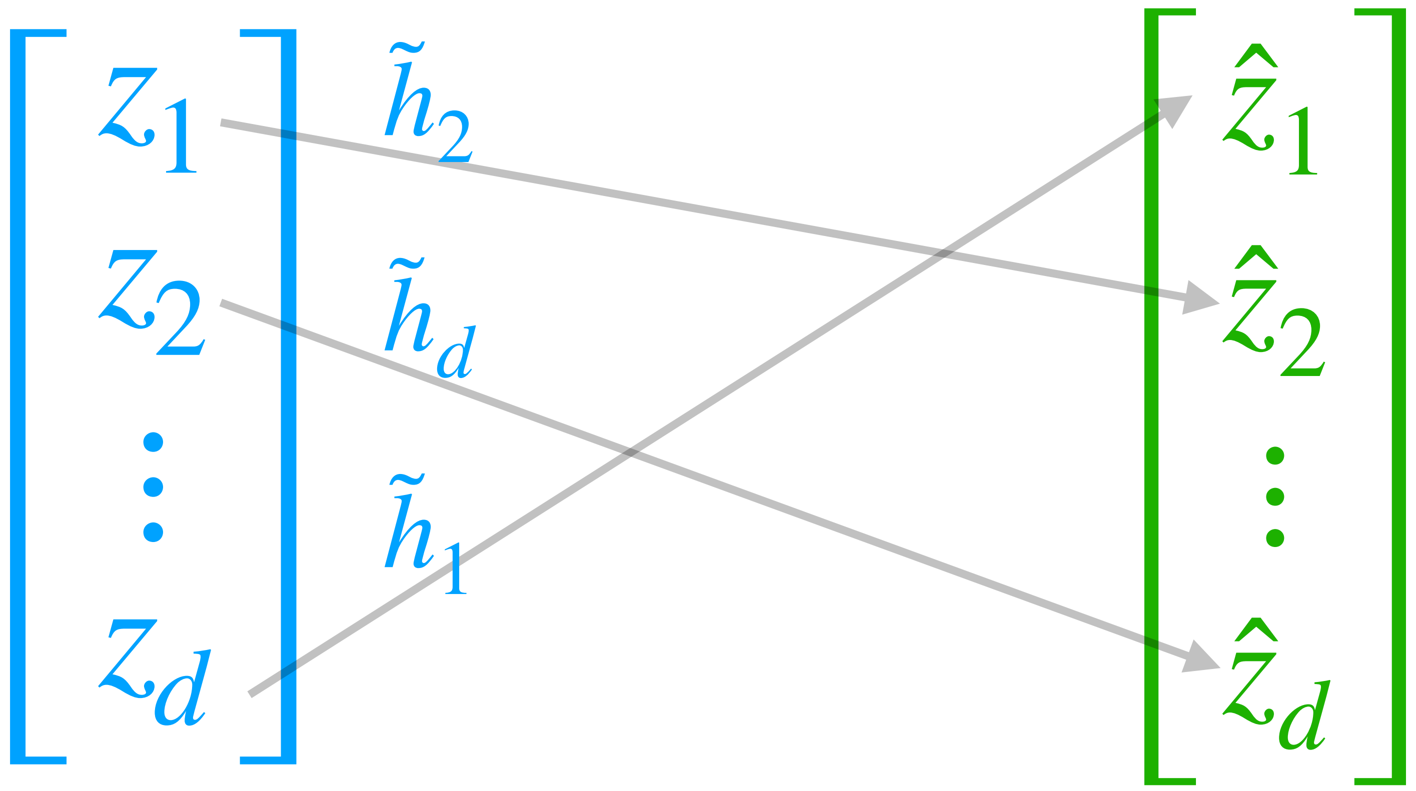
How well can we actually recover these causal variables?

- **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

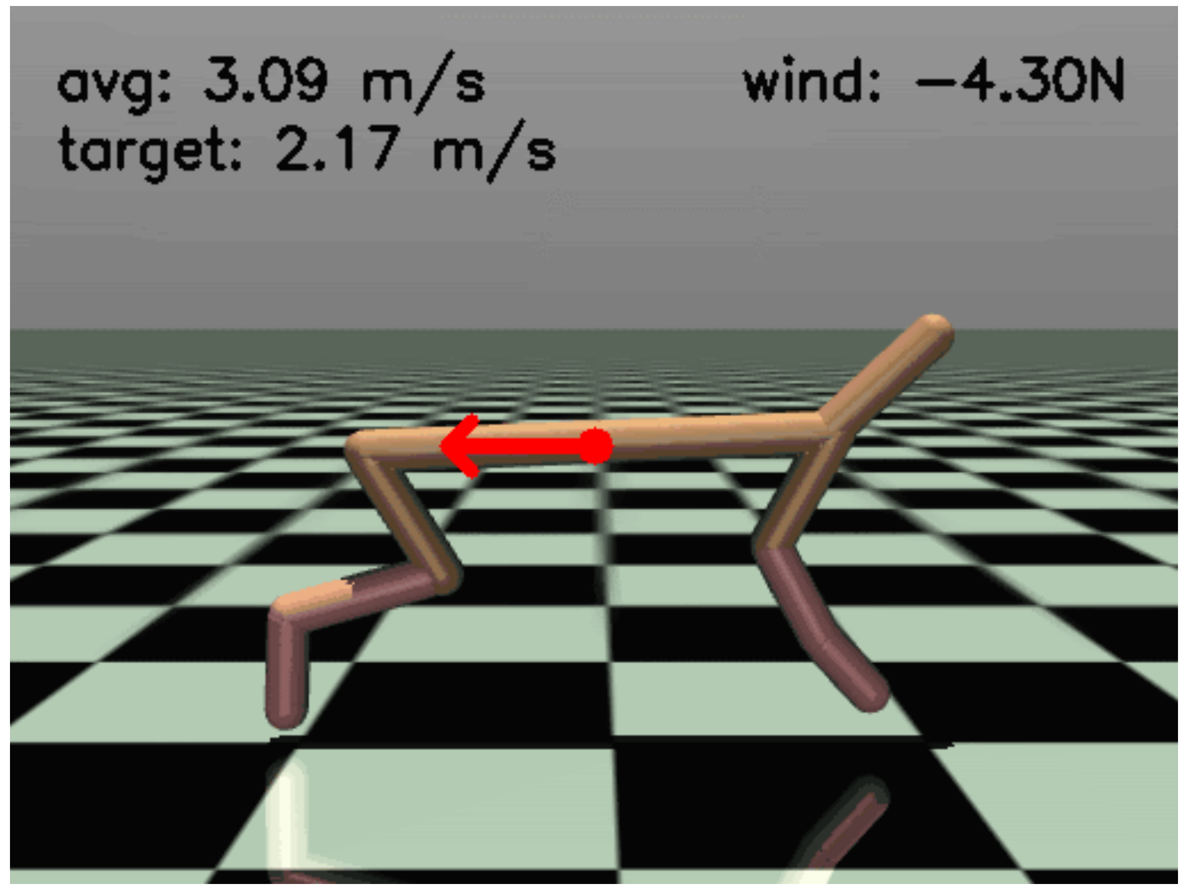


Identifiability up to permutation and component-wise transformations



CRL in temporal settings with actions

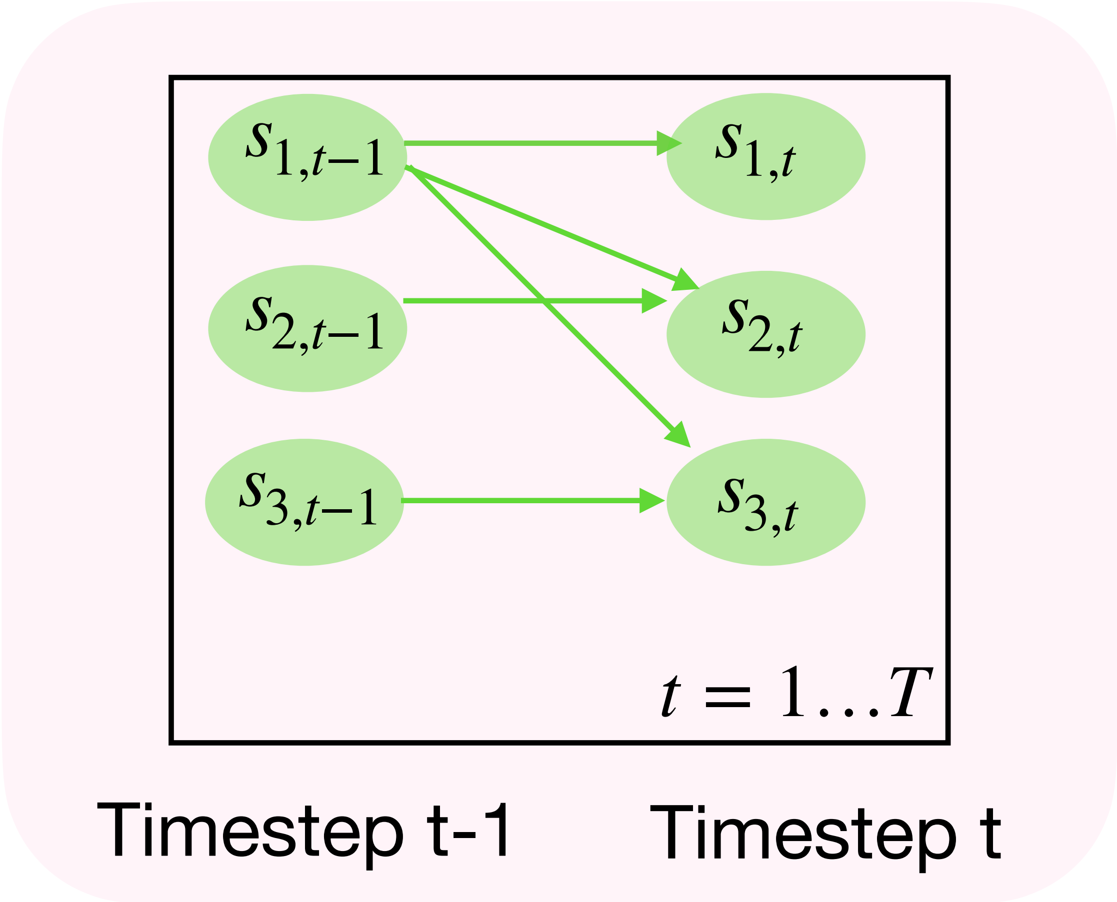
- Natural setting for learning from interventions/actions: "before" and "after"
  - E.g. sequential decision making, RL, planning, robotics, …



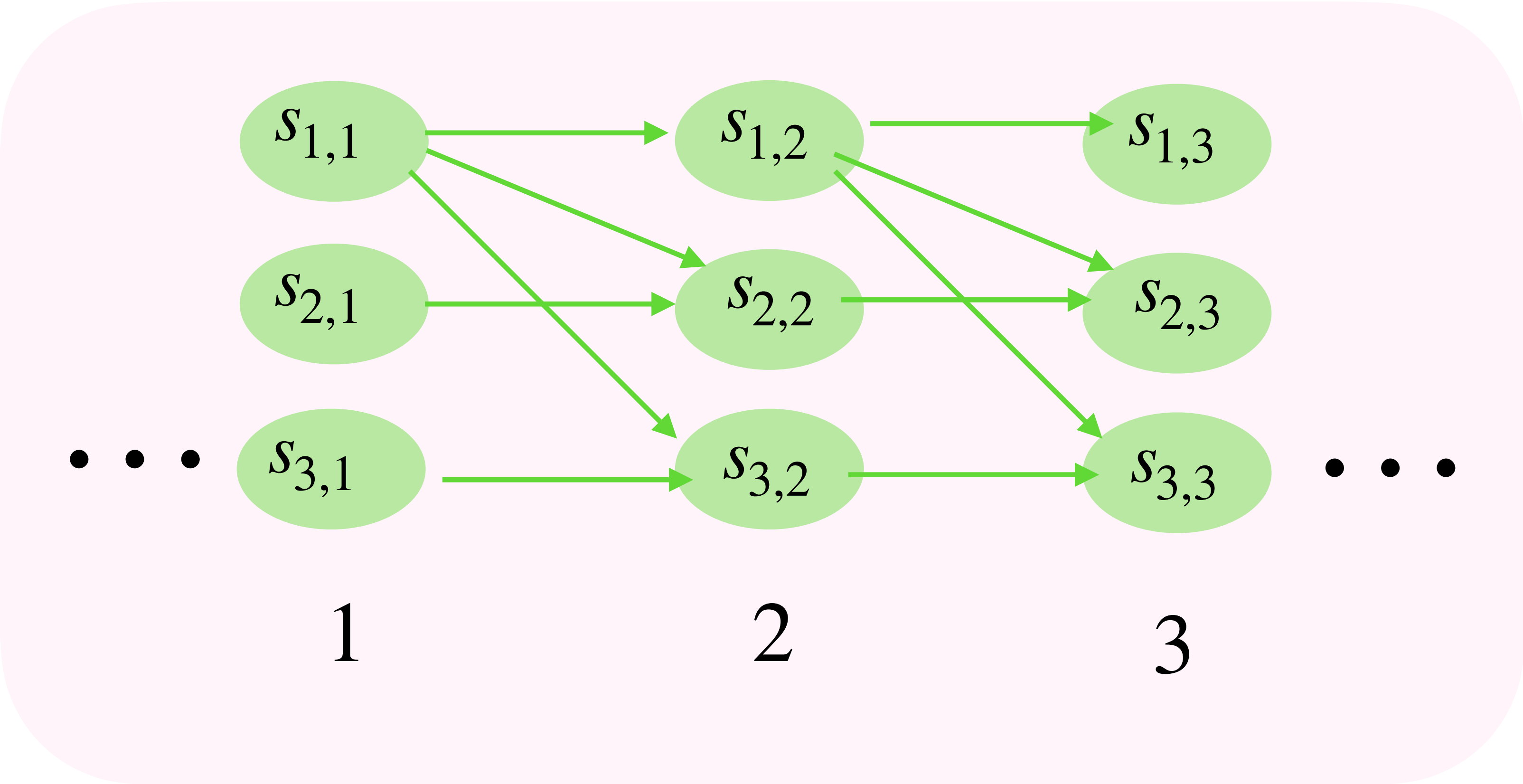
- Often we want to extract **semantic features from images** in an unsupervised way
  - **Causal representation learning (CRL) -** learn high level causal variables and causal relations between them from low level observations

Modelling causality in time series: Dynamic Bayesian Networks

- Extension of Bayesian networks to temporal settings, a type of template graph.
- Common assumptions for Dynamic Bayesian Networks:
  - **1-Markov assumption**: only vars from t-1 (1 timestep back) can cause vars at t
  - **Stationarity:** the transition model (edges) are the same across pairs of time steps
  - **No instantaneous effects:** there are no edges between vars at same timestep

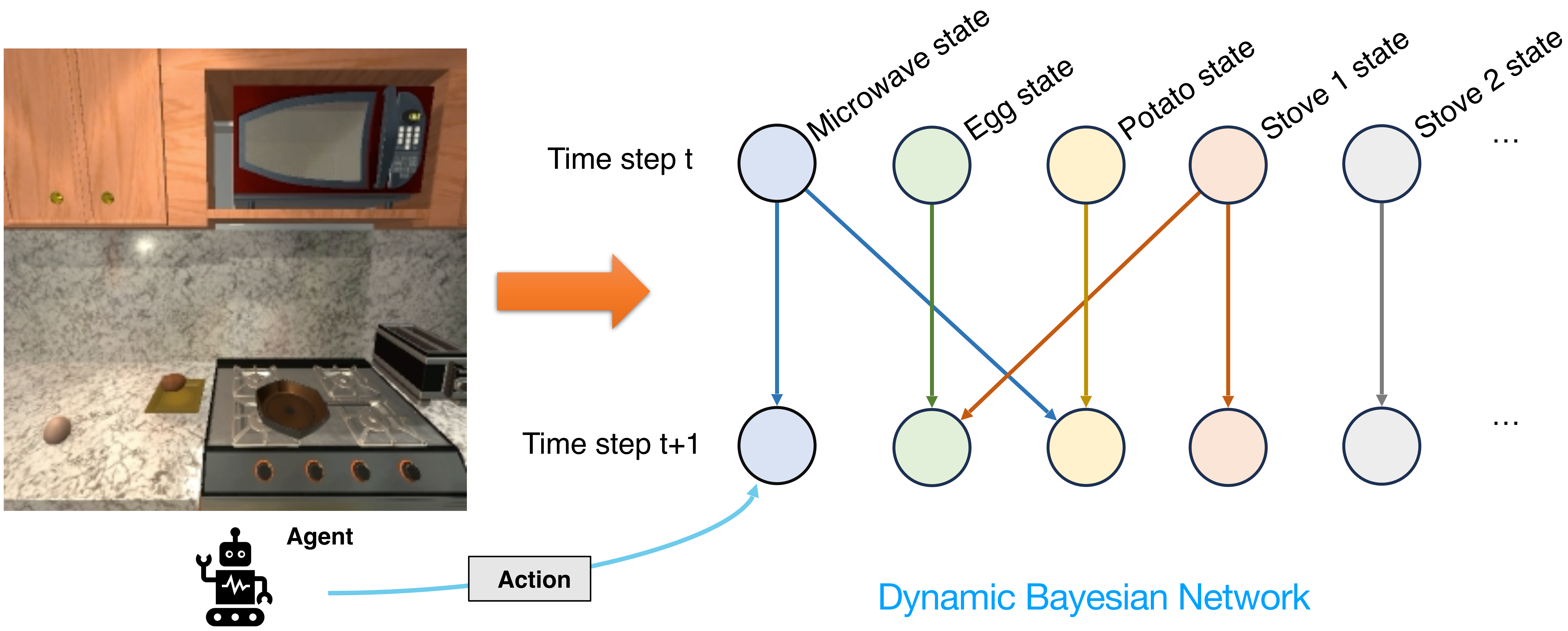


$\equiv$

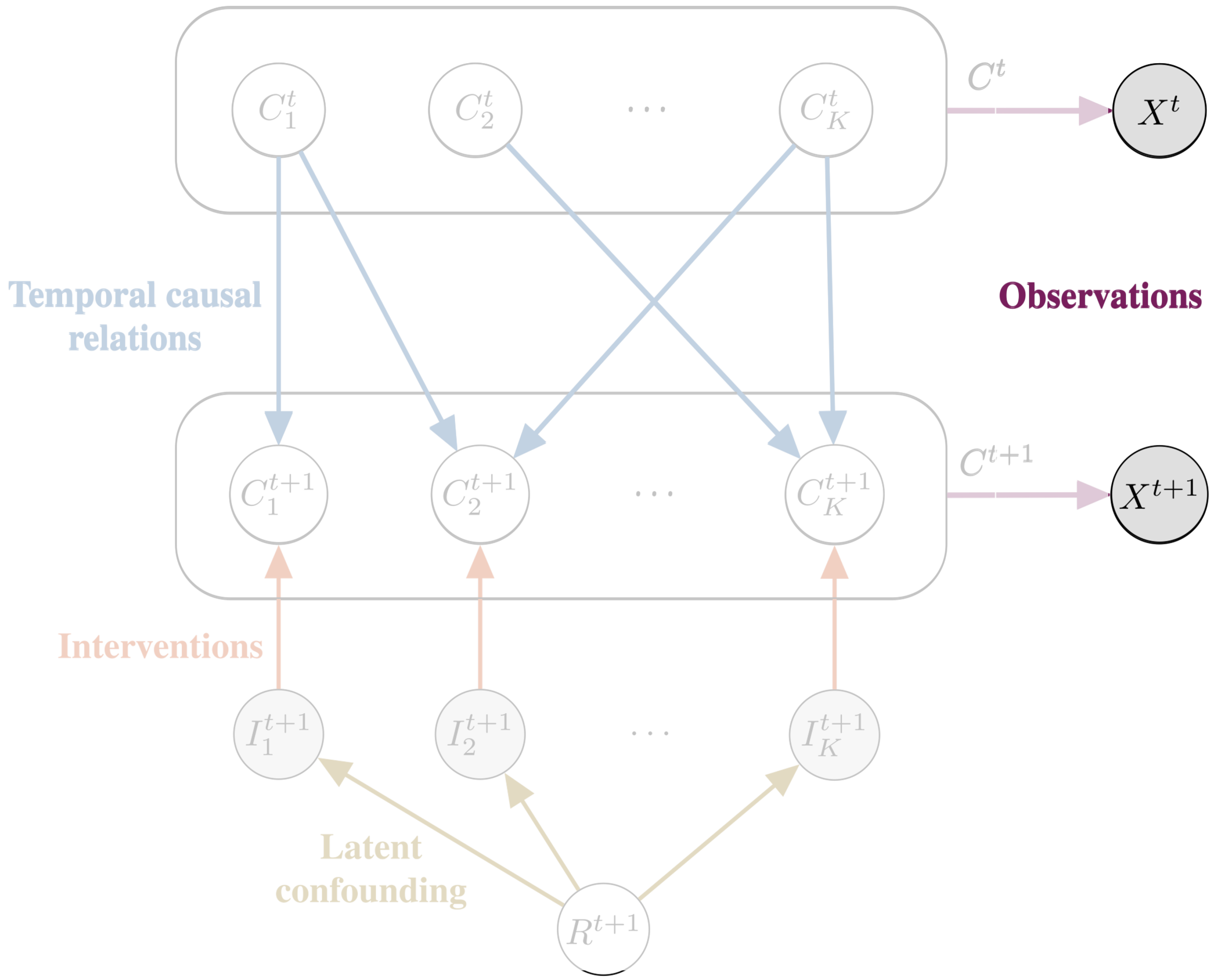


MDPs in RL are an example

Goal: CRL in temporal settings

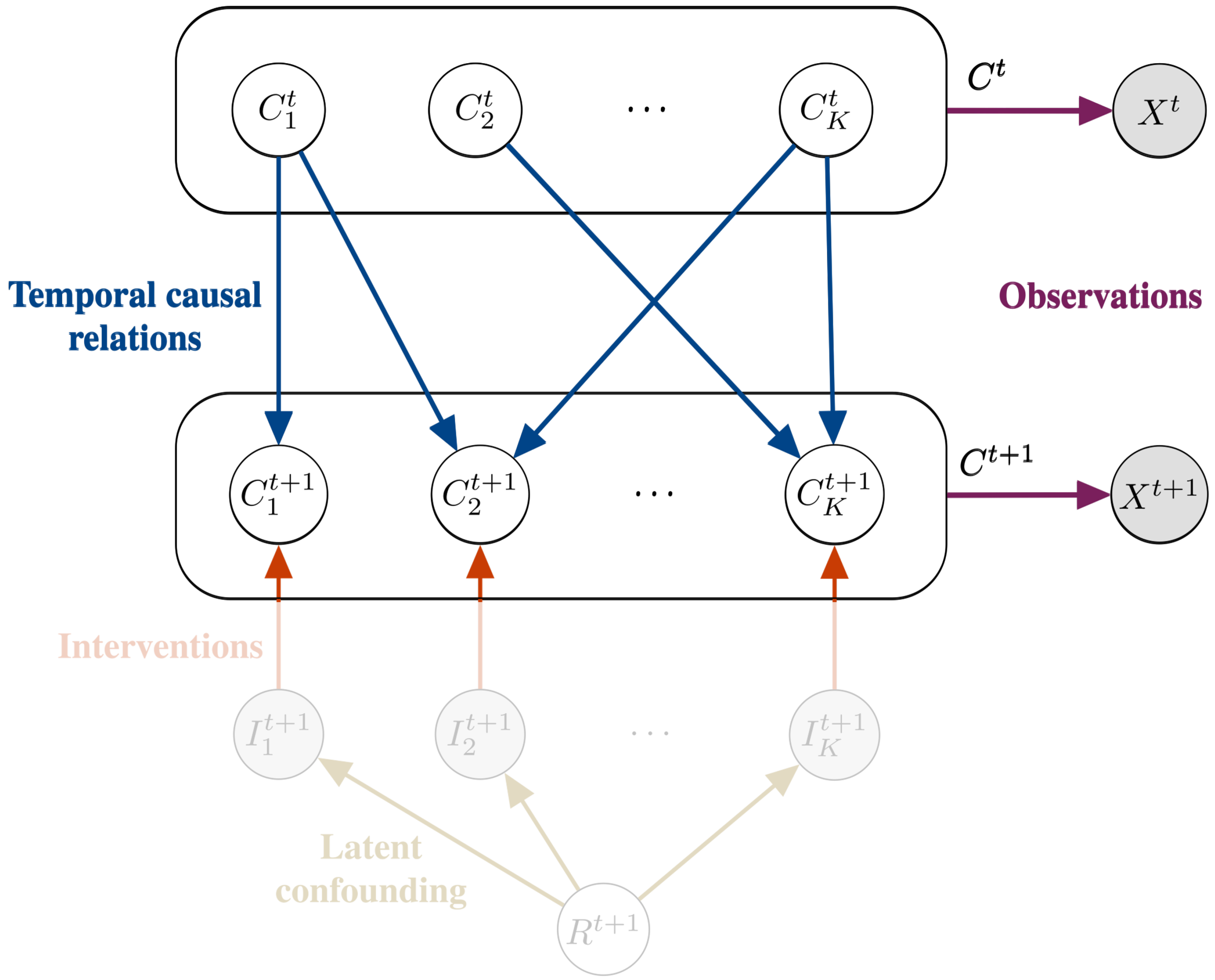


Temporal Intervened Sequences (TRIS)



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

Temporal Intervened Sequences (TRIS)

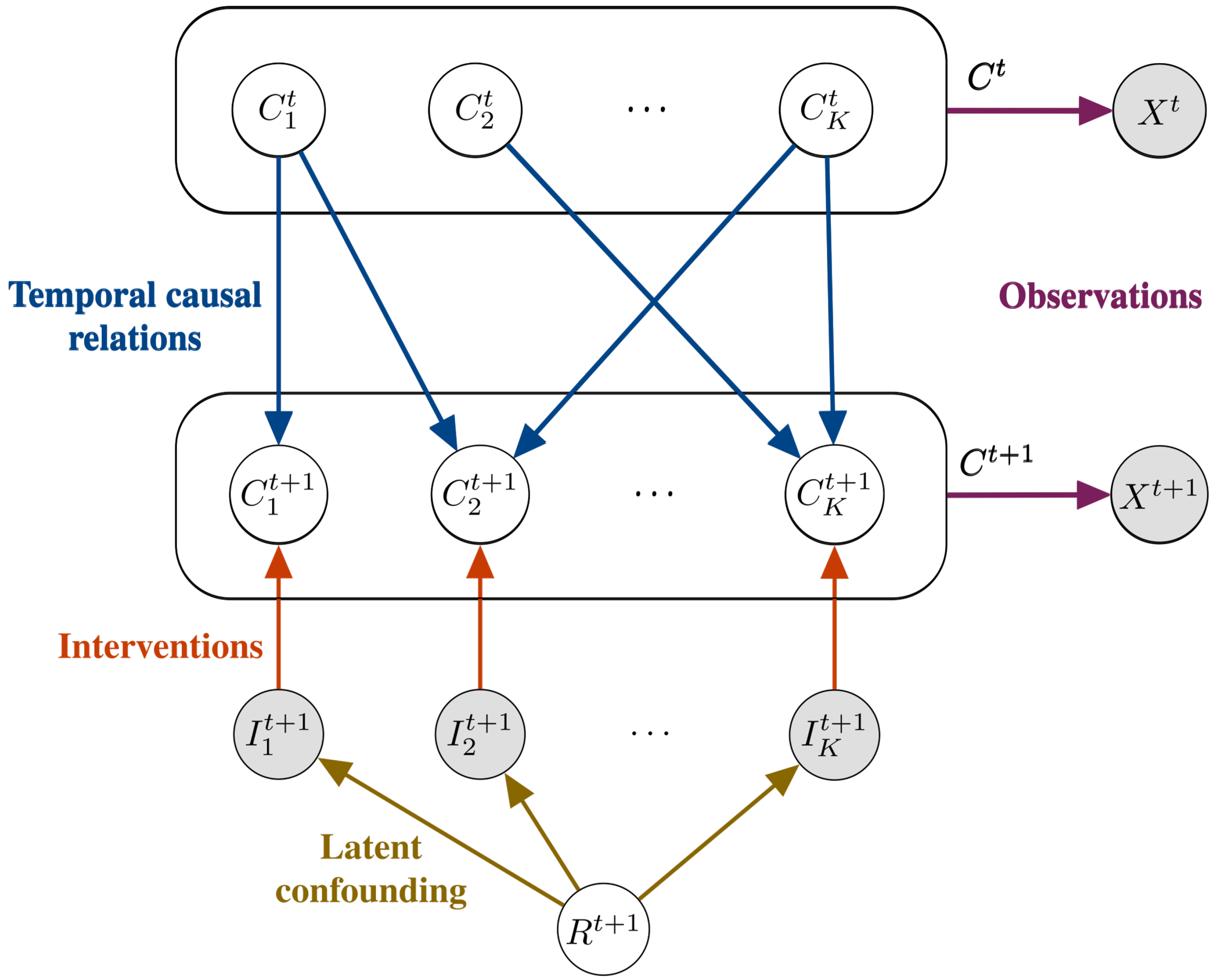


- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images

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

$$X^t = h(C_1^t, \ldots C_K^t, E^t)$$

Temporal Intervened Sequences (TRIS)

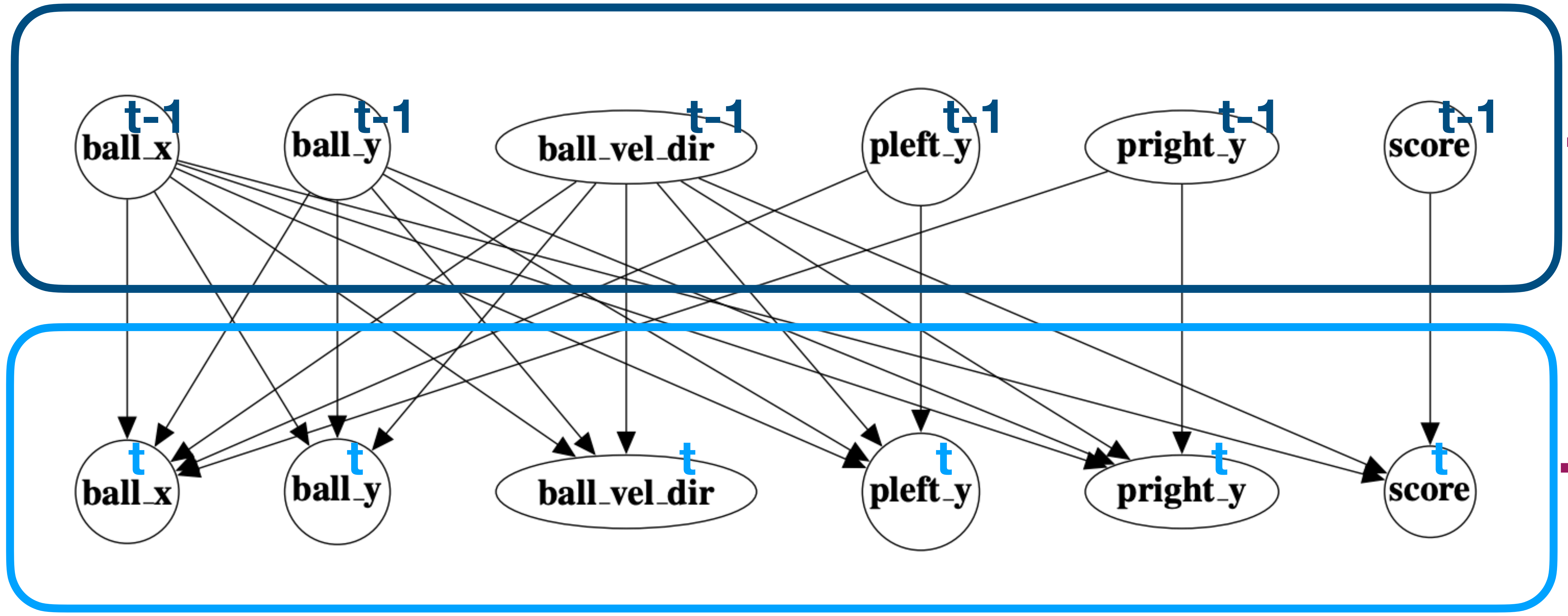


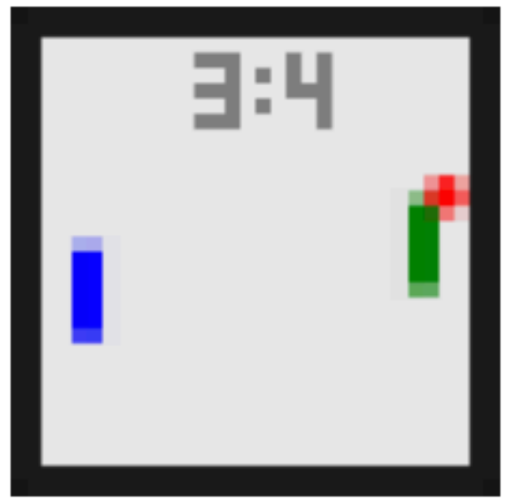
- We want to learn the underlying causal process from **temporal sequences** of **high-dimensional data** $\{X^t\}_{t=1}^T$, e.g. images
- We assume that the **latent** causal process is a **Dynamic Bayesian network** with **K multidimensional causal variables**
- We assume that **(soft or perfect) interventions**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

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

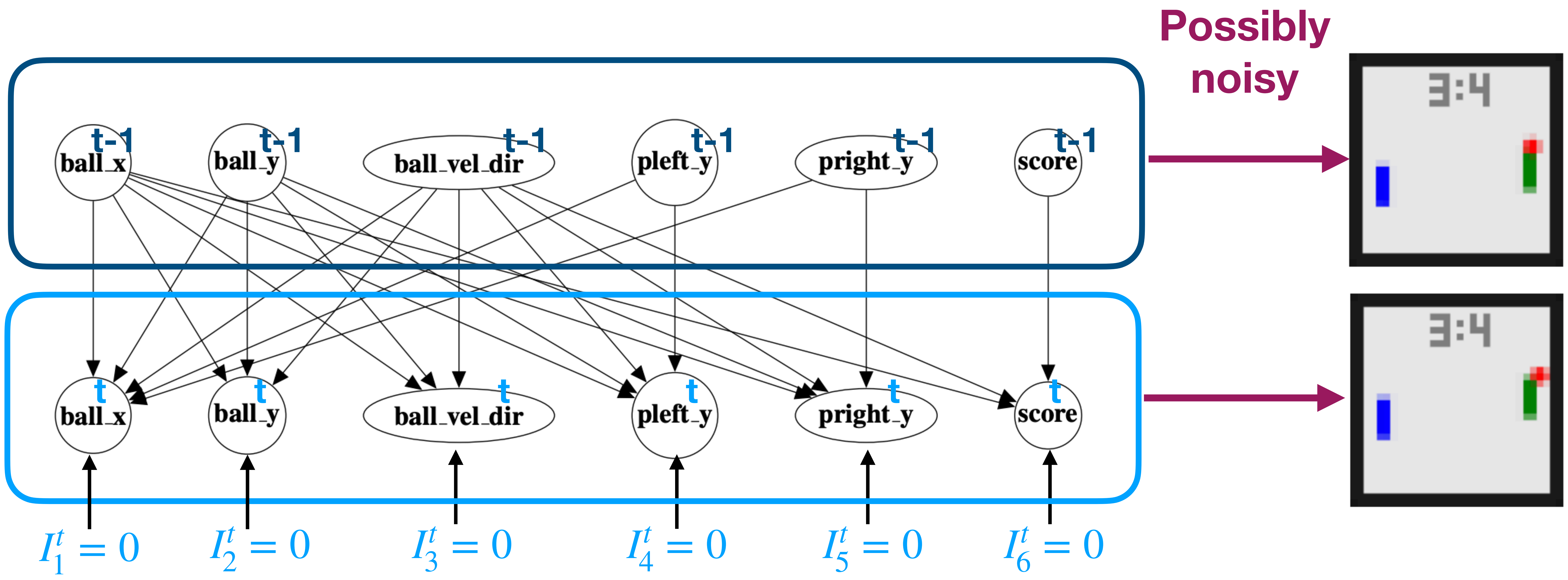
Possibly noisy





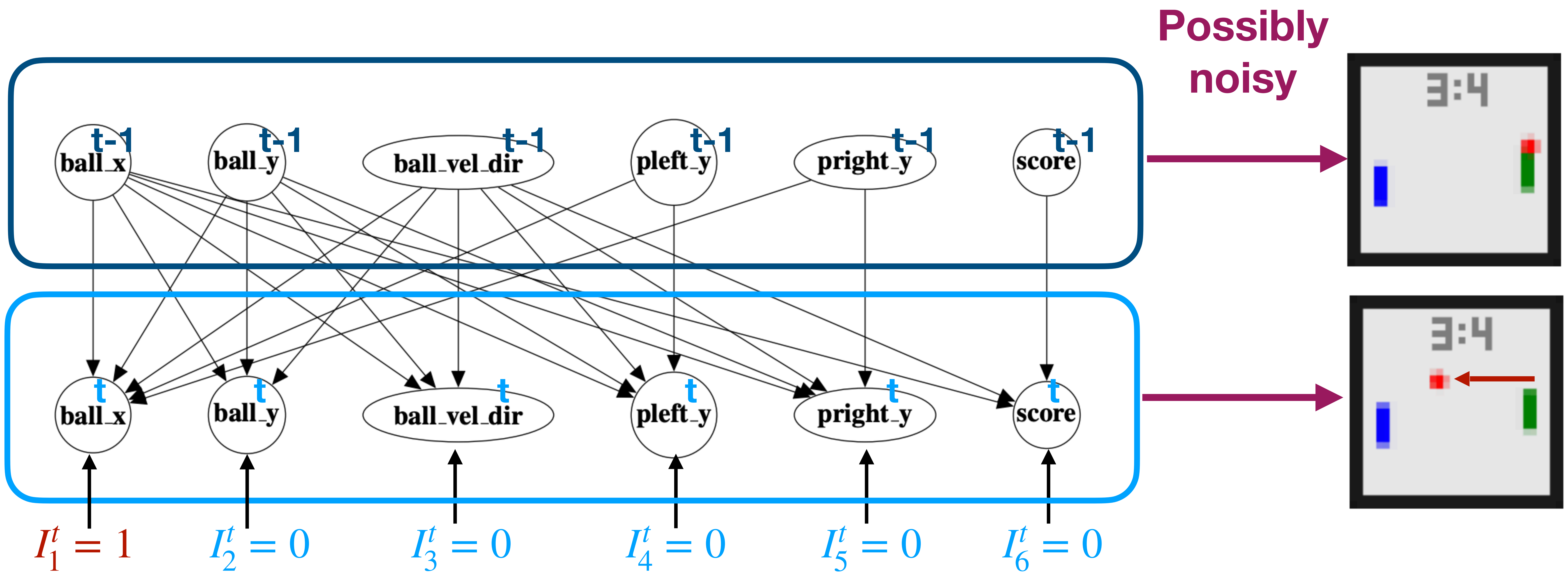
CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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



CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

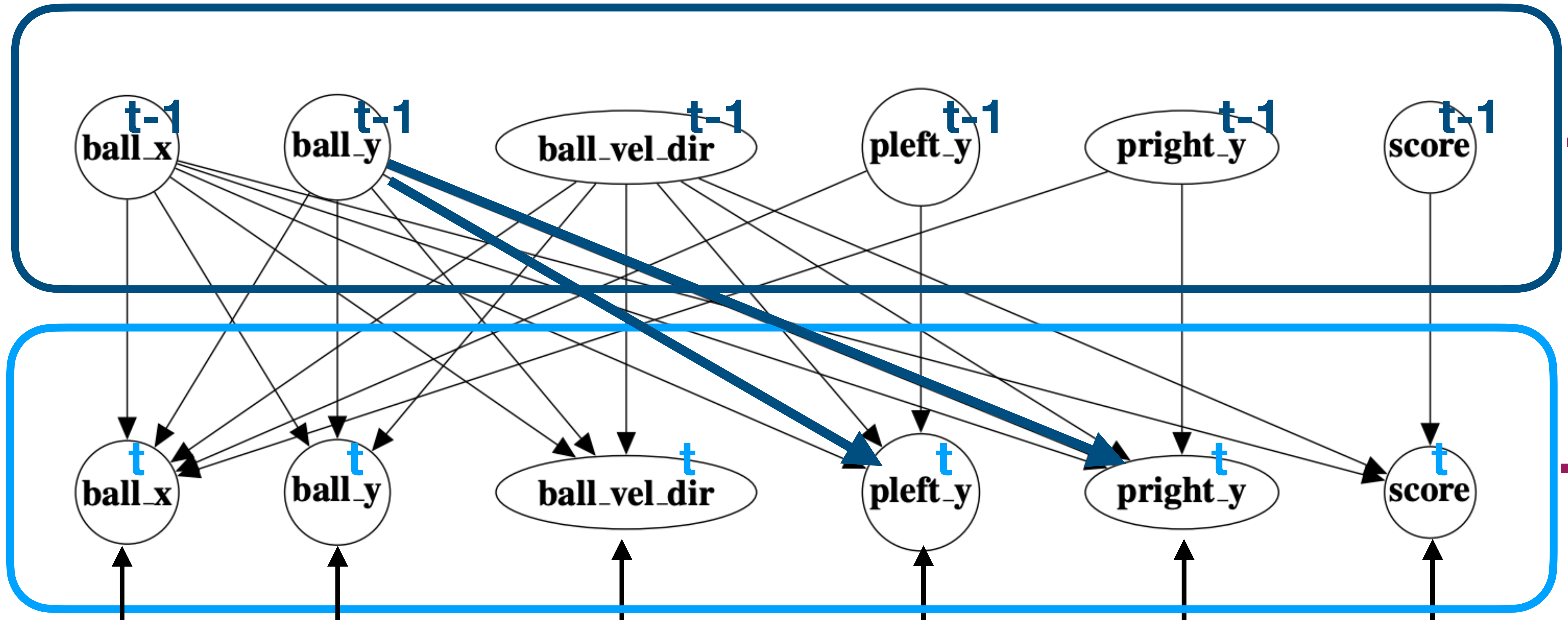


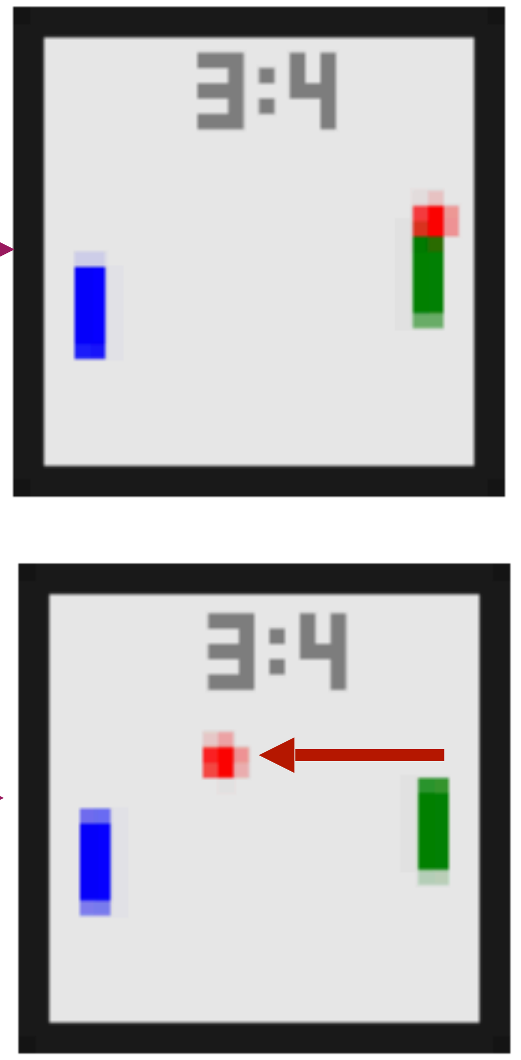
Stochastic intervention
(we don't know where the ball will be)

CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

Possibly noisy





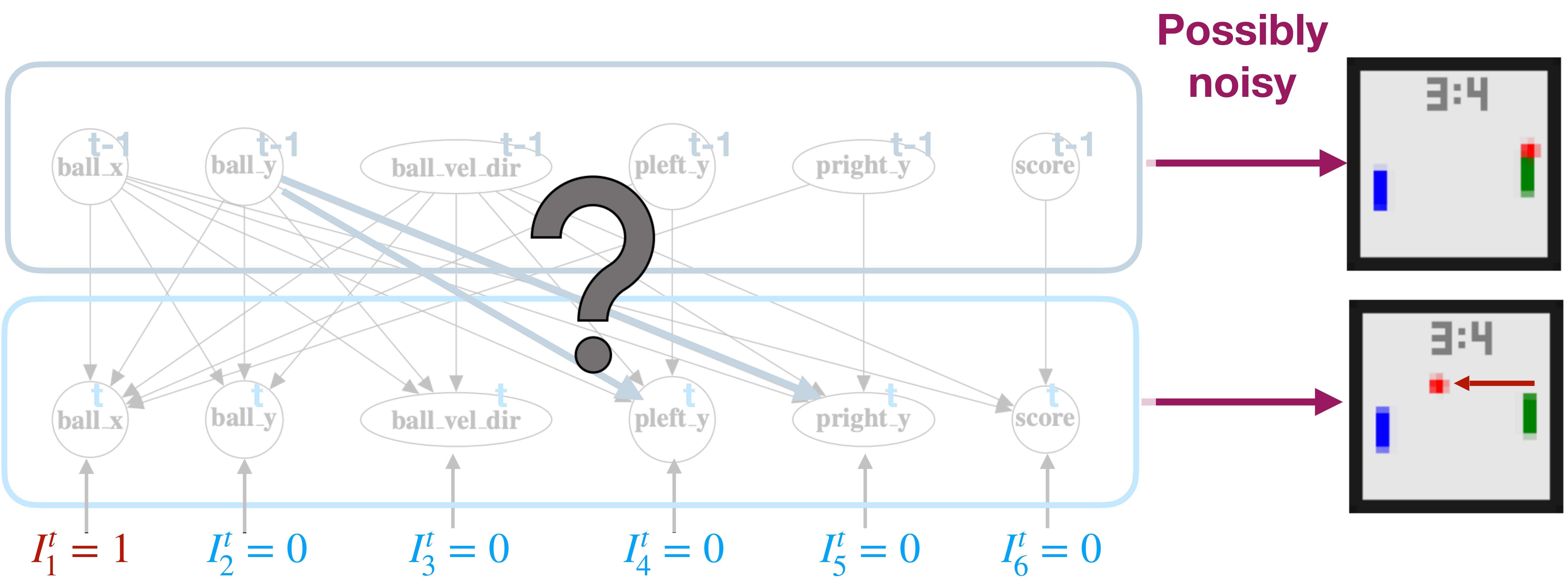
$I_1^t = 1$ $I_2^t = 0$ $I_3^t = 0$ $I_4^t = 0$ $I_5^t = 0$ $I_6^t = 0$

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

The paddles continue moving as usual (not counterfactual)

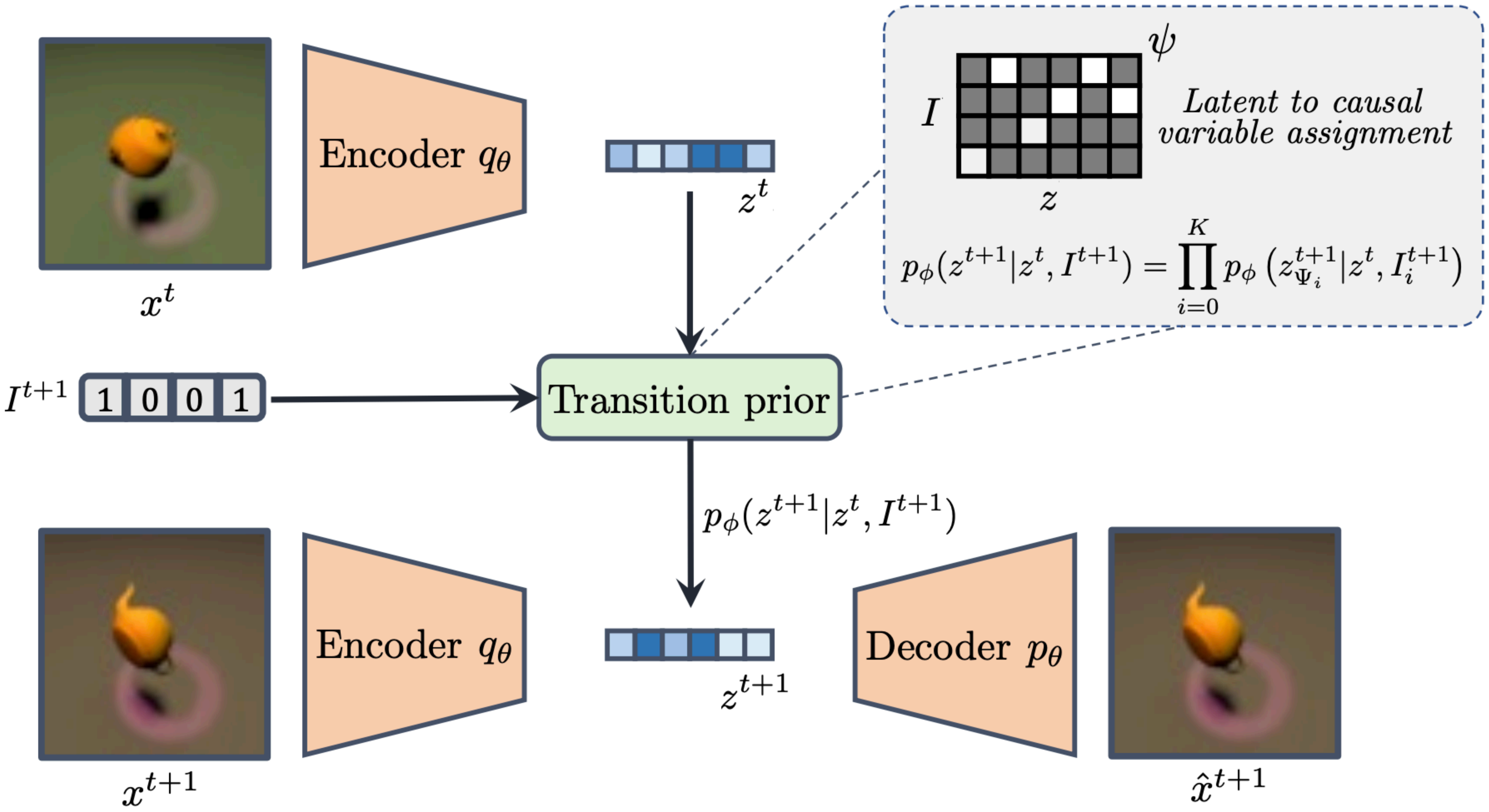
CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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



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

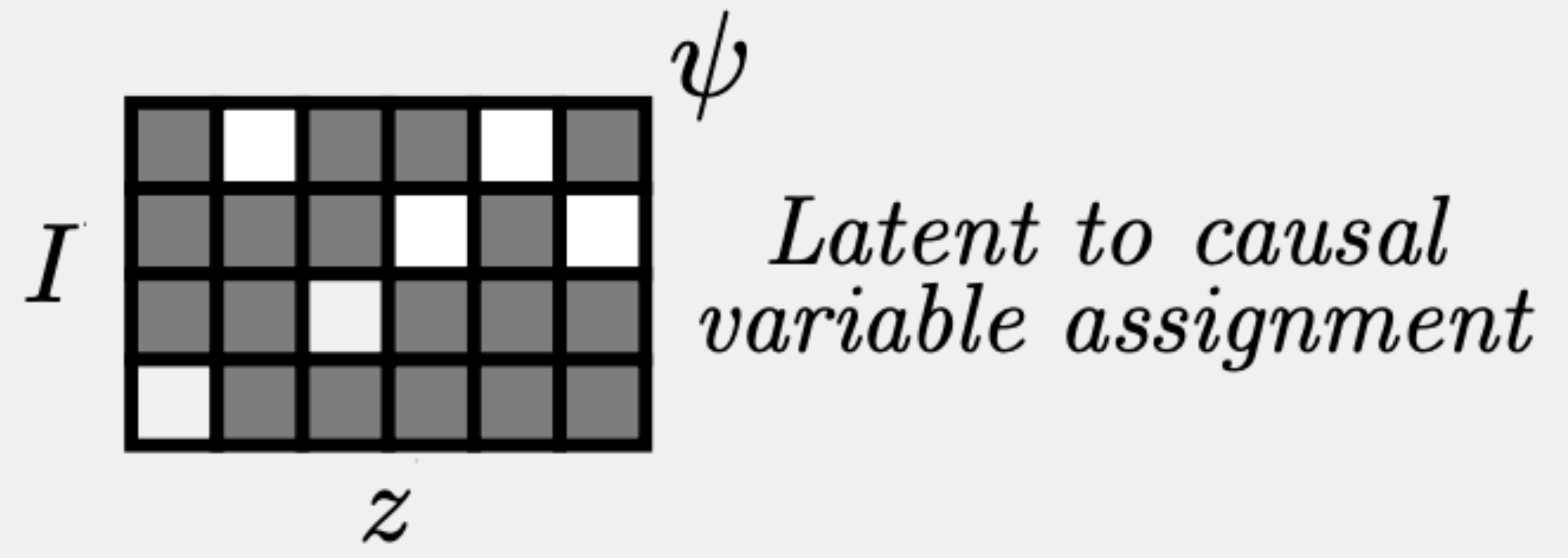
A variational autoencoder architecture: CITRIS-VAE



CITRIS: Causal Identifiability from TempoRal Intervened Sequences

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

- We have **multidimensional causal factors**, so we need to learn an assignment function ψ that matches each C_i with the assigned latents



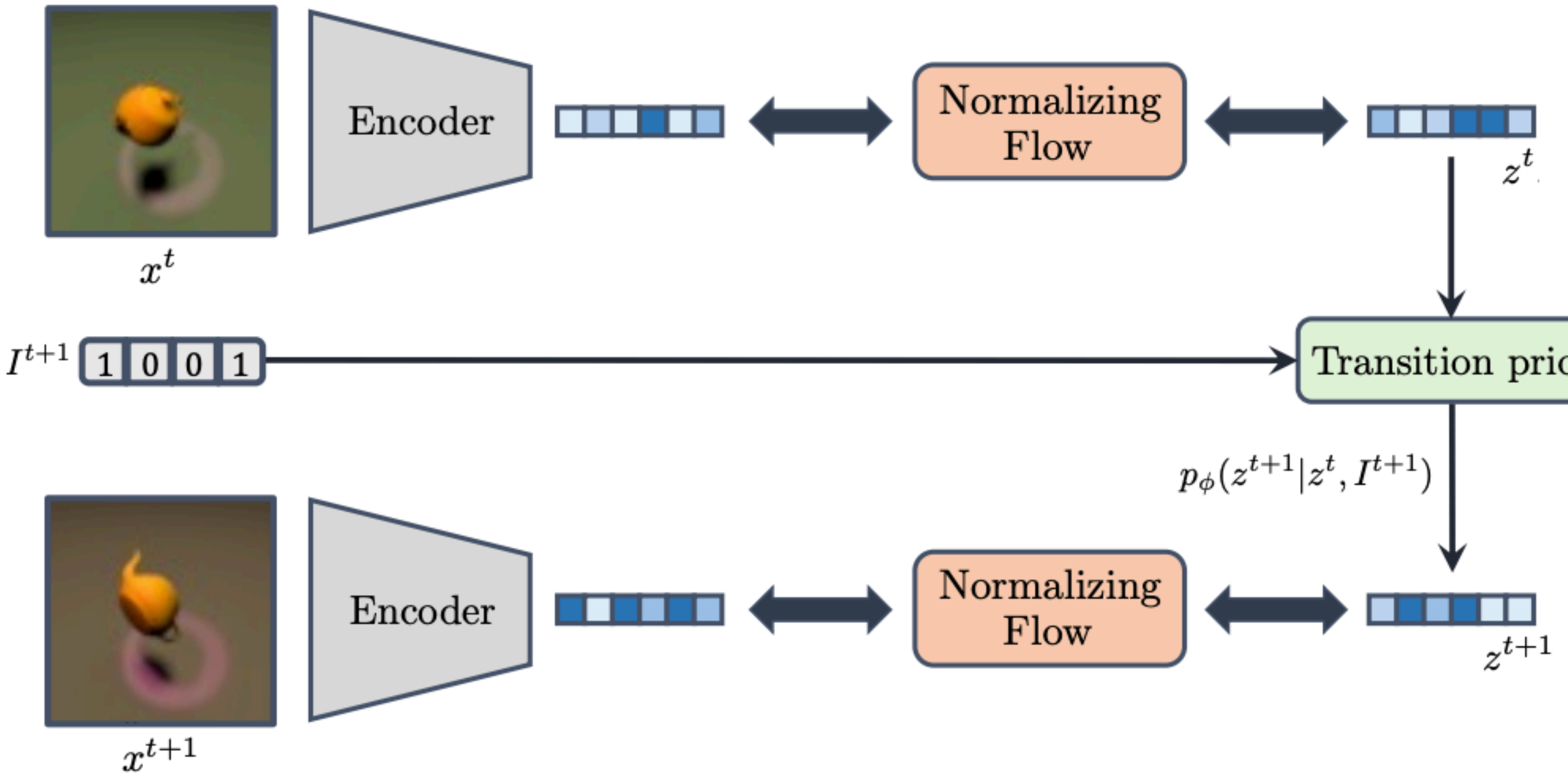
$$p_\phi(z^{t+1}|z^t, I^{t+1}) = \prod_{i=0}^{K} p_\phi\left(z^{t+1}_{\Psi_i}|z^t, I^{t+1}_i\right)$$

$$C_i \longrightarrow z_{\Psi_i}$$

z_{Ψ_0} “junk” variables

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)

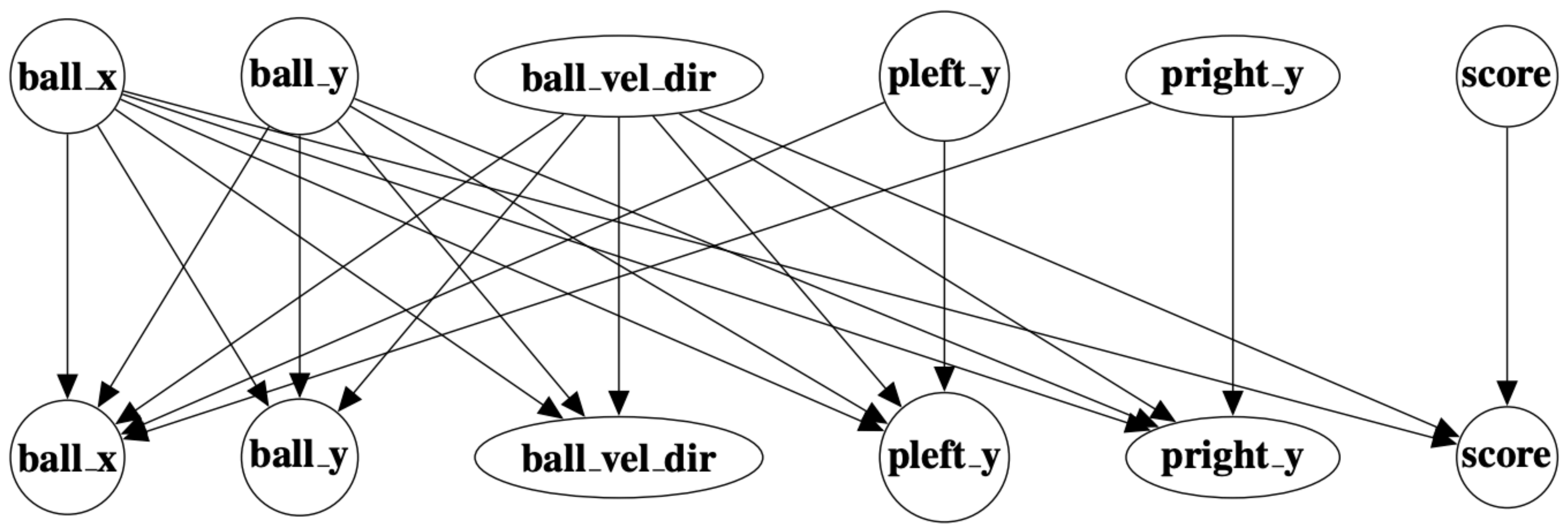


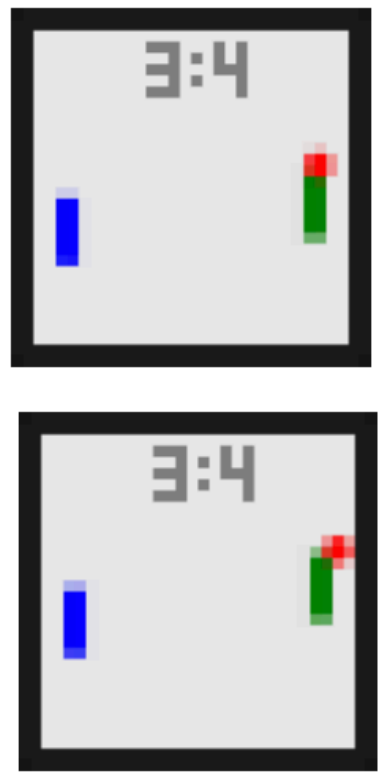
CITRIS: Simplified identification results

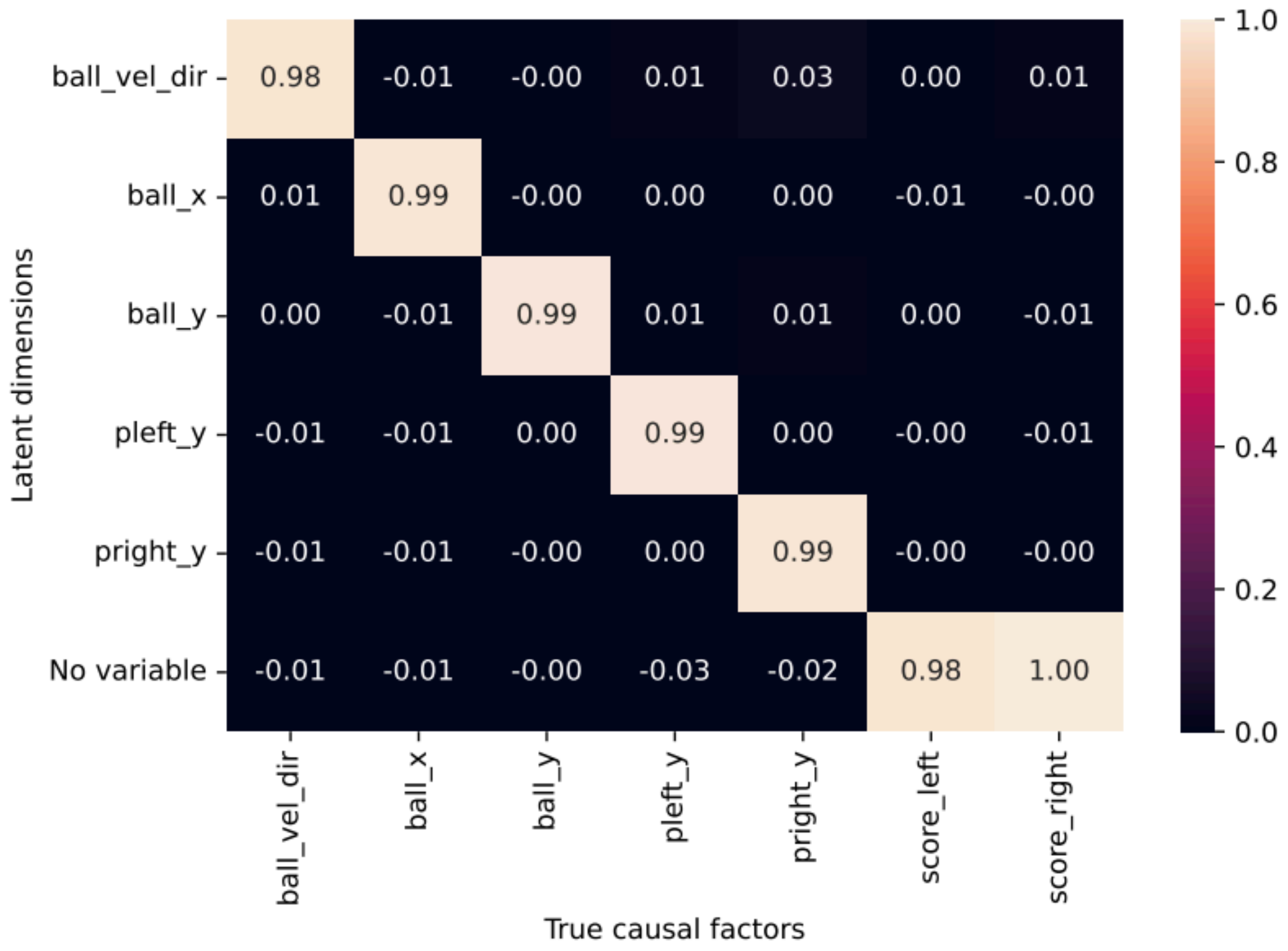
- TRIS setting, sufficient latent dimensions
- Assumption 1: Each I_i^t is not a deterministic function of I_j^t
- **[Simplification] Assumption 2:** Interventions have an effect on **all components** of any multi-dimensional causal variable
- **[Simplification] Assumption 3:** There are **"enough"** different types of interventions ($O(log_2 K)$)
- Then we can identify causal variables $C_1, \ldots, C_K$ **up to unknown invertible element-wise transformations**

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.

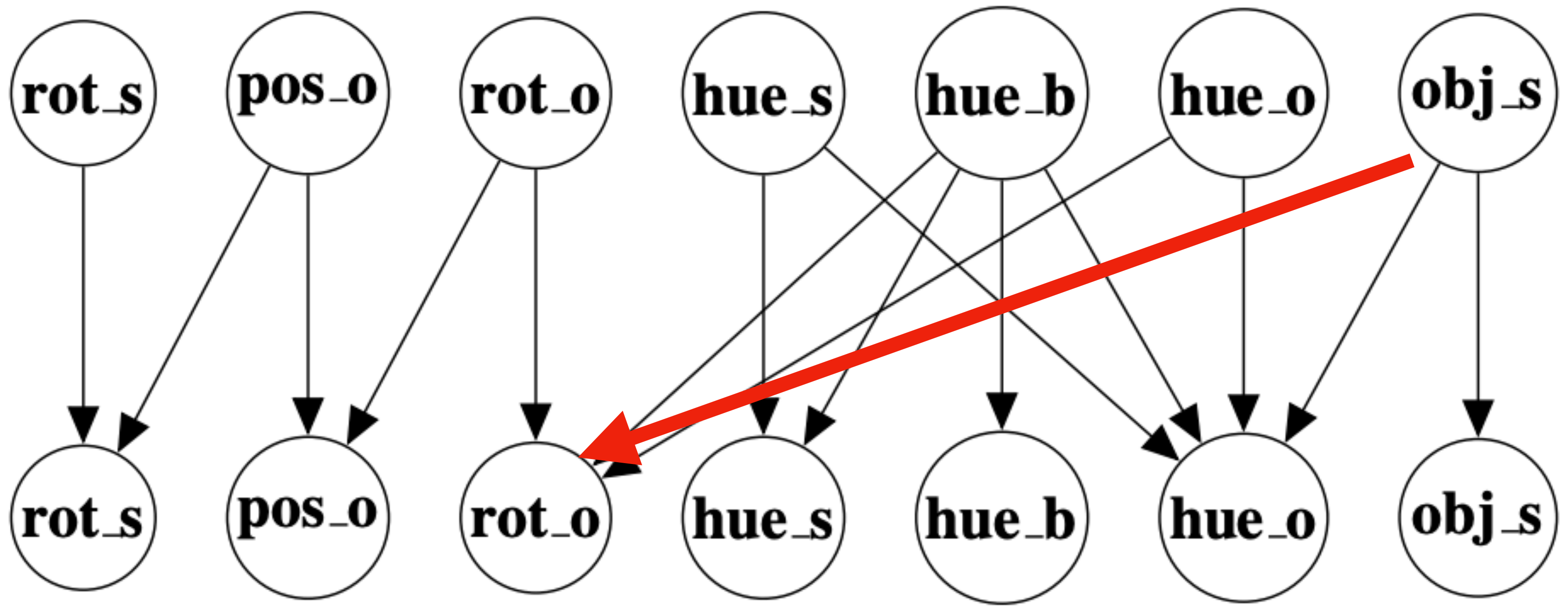






Experiments: Temporal Causal3DIdent

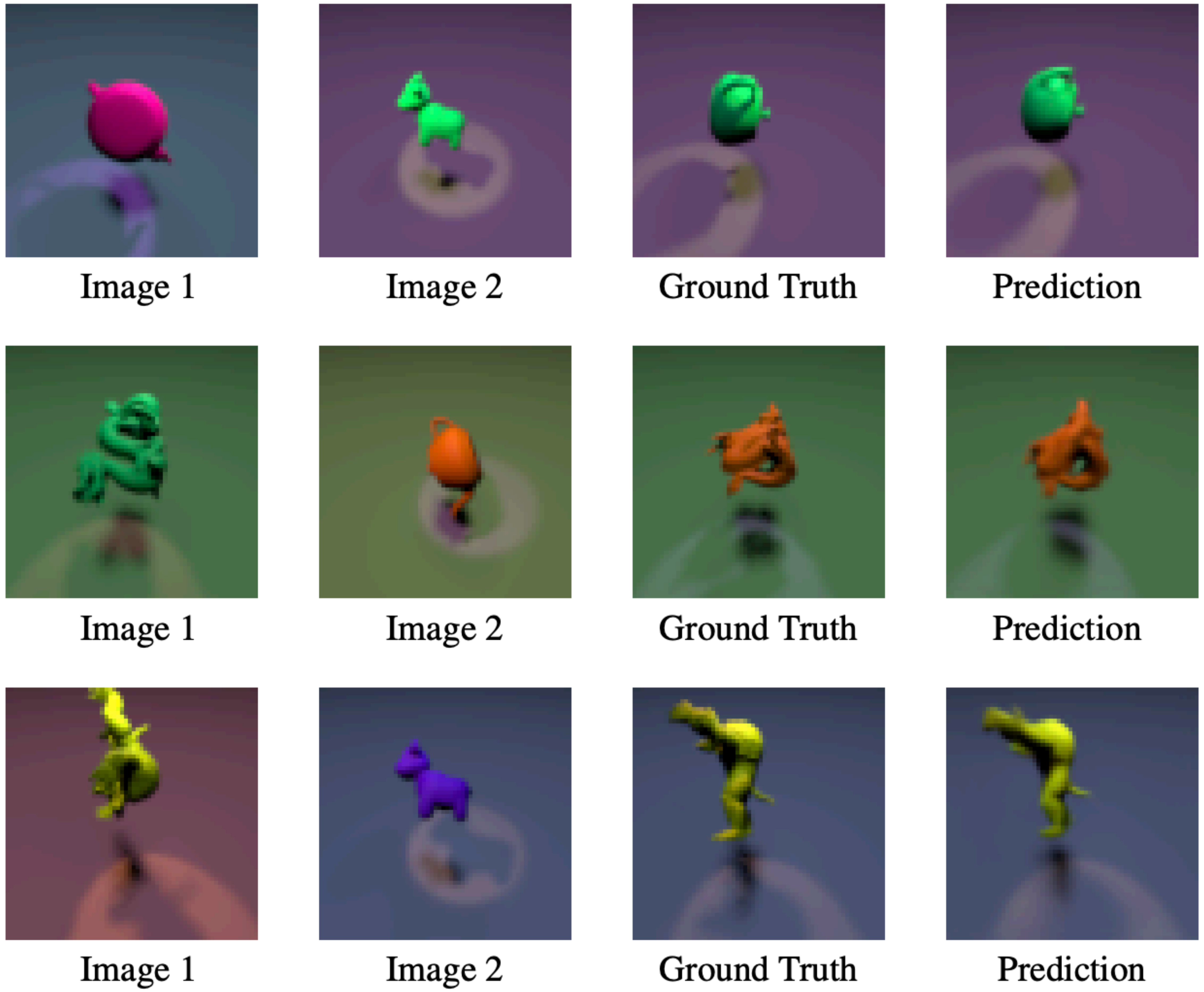
shape + spotlight colors + rot

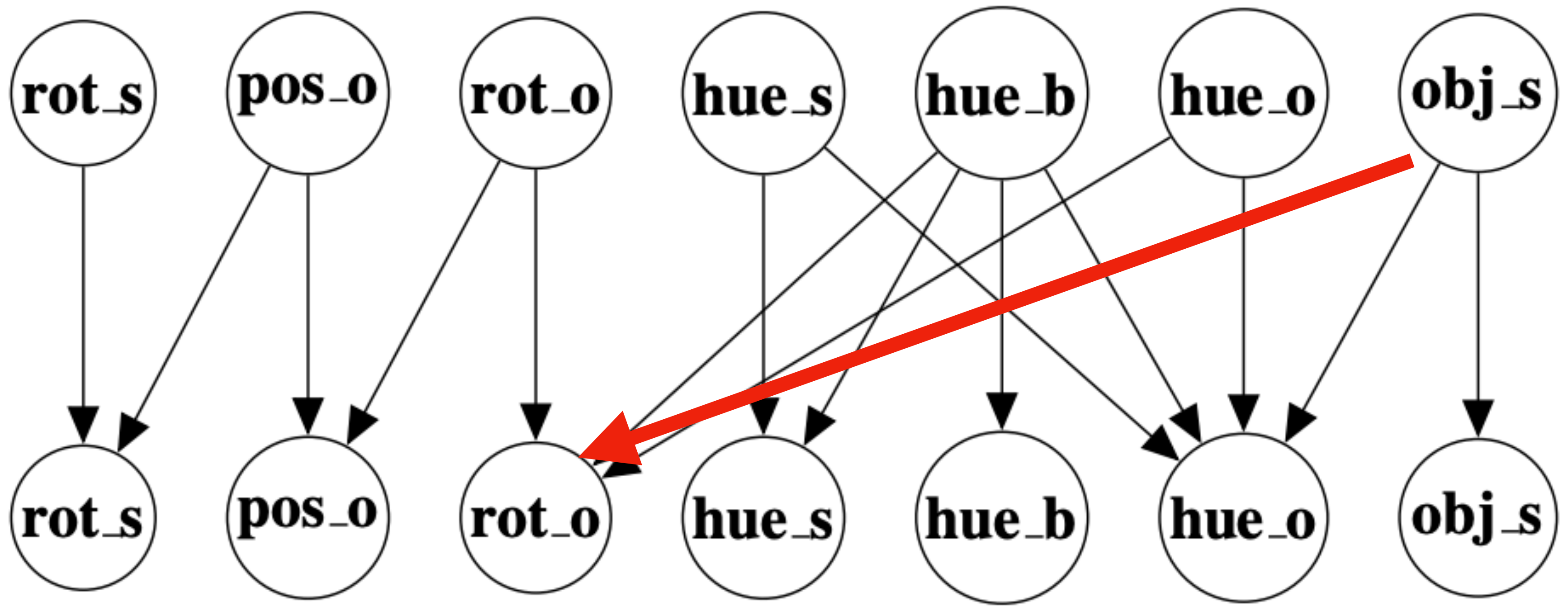


Causal graph learnt with CITRIS-NF

Experiments: Temporal Causal3DIdent

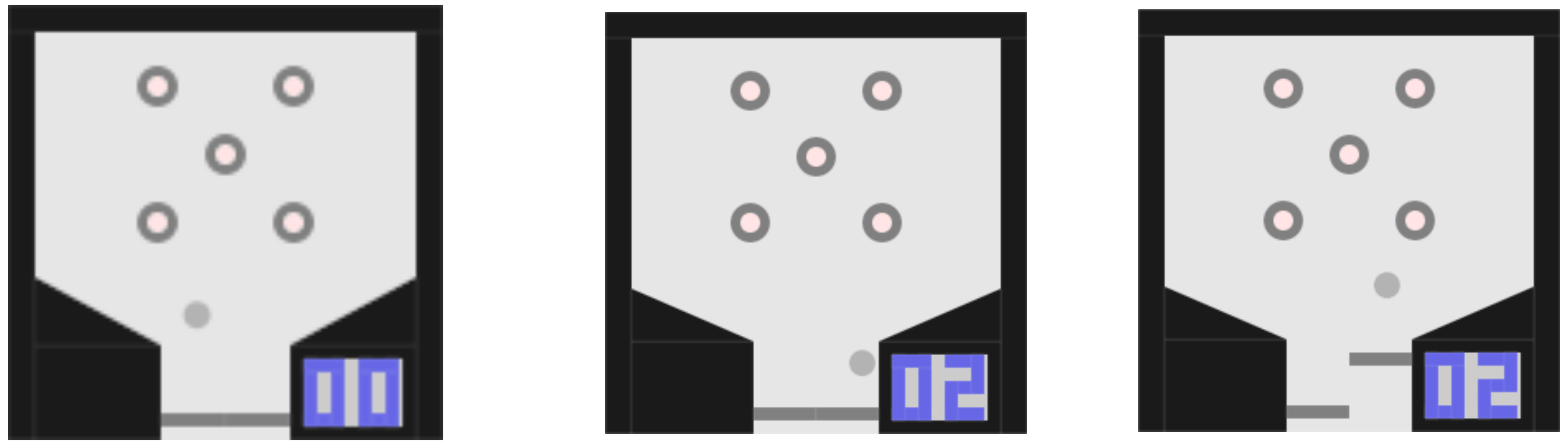
shape + spotlight colors + rot



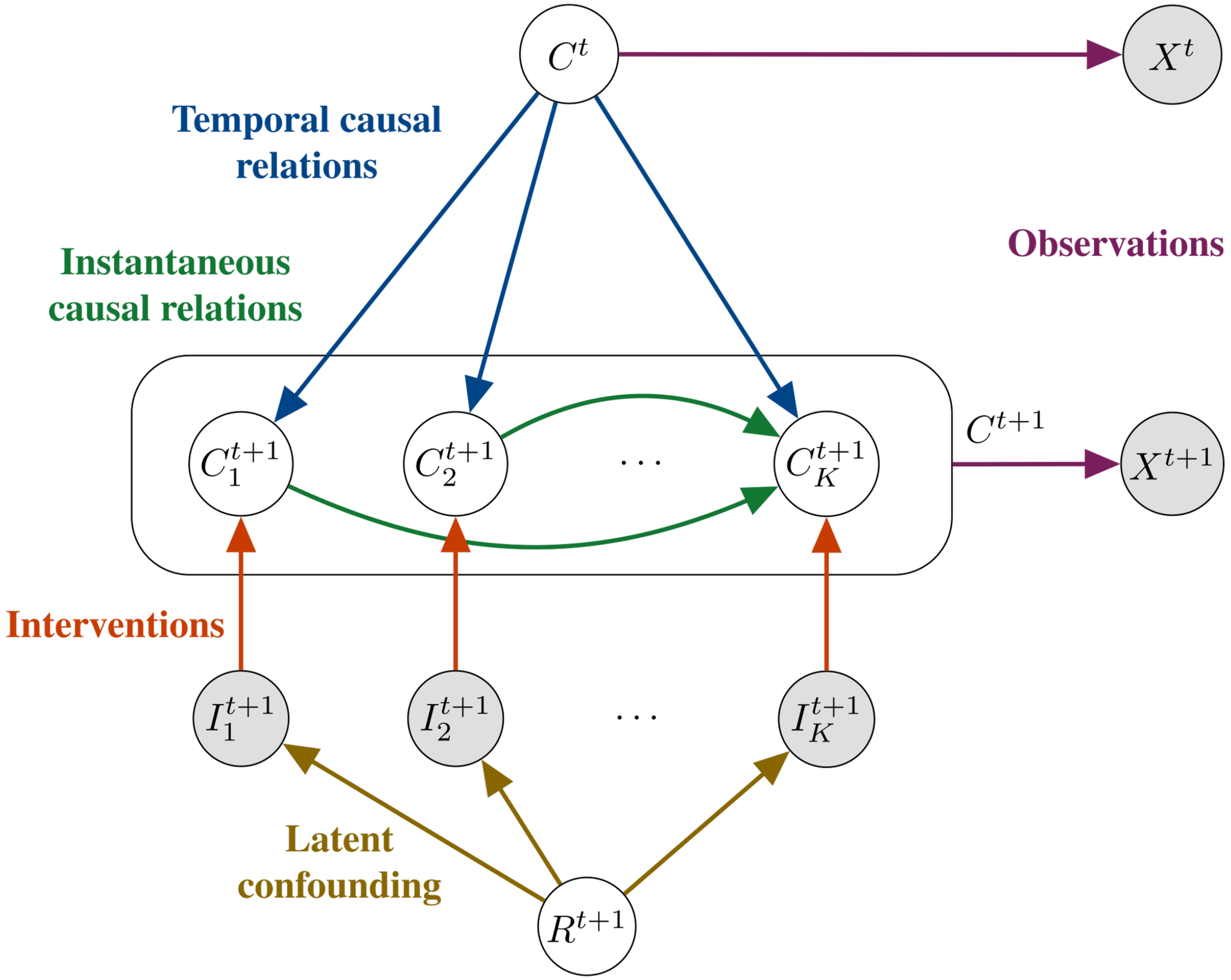


Causal graph learnt with CITRIS-NF

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

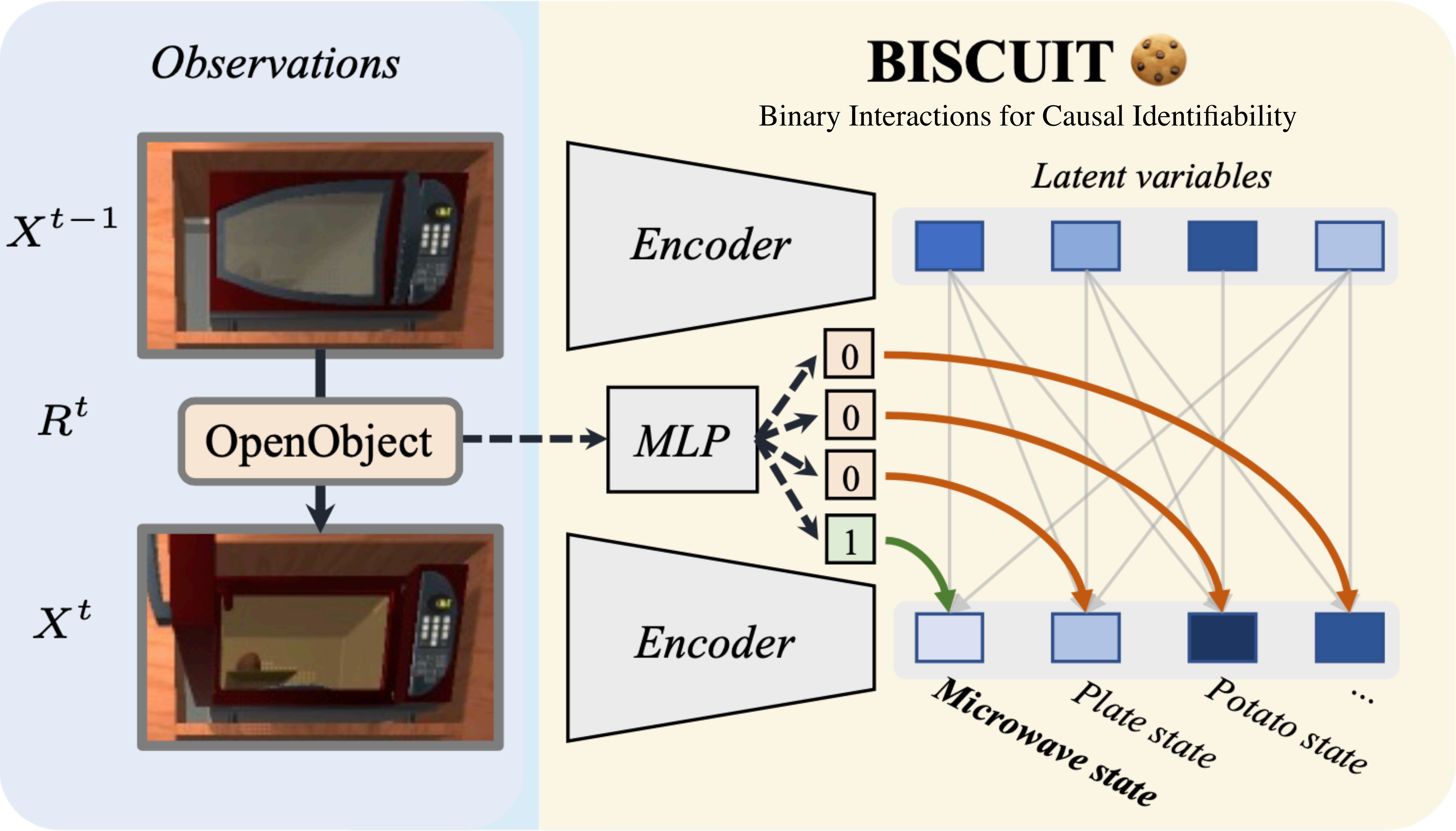


Summary CITRIS & iCITRIS

- **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?**

BISCUIT: Causal Representation Learning from Binary Interactions

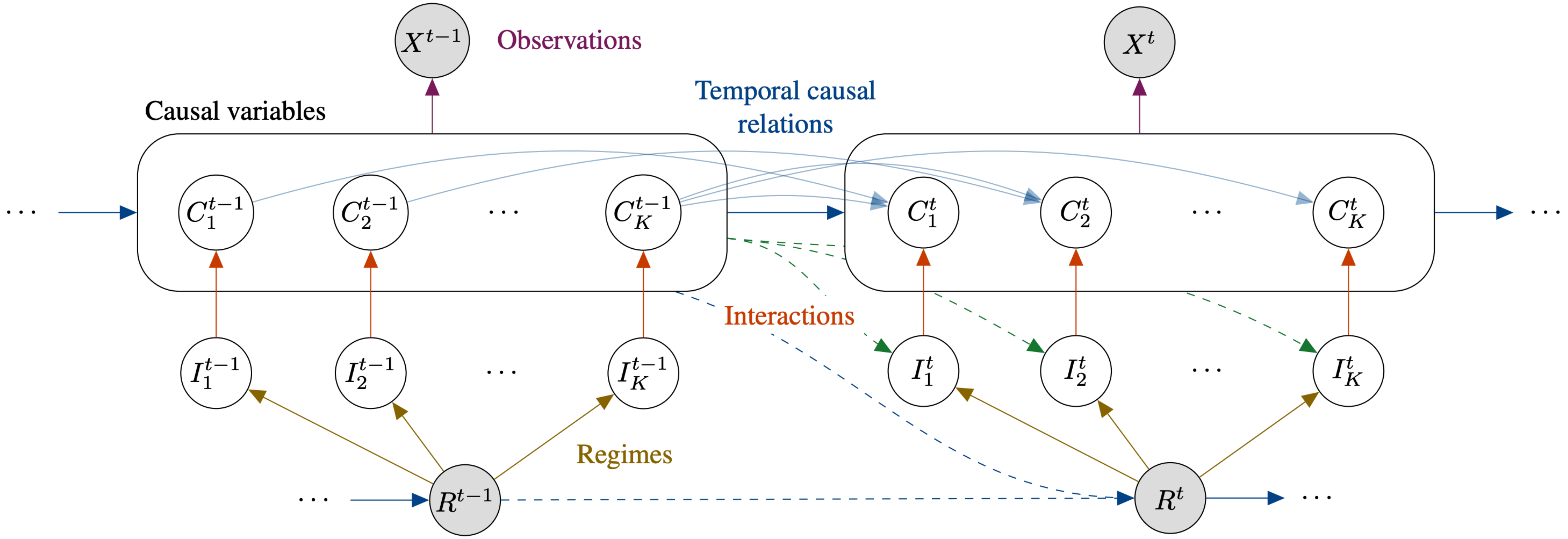
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves



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)

Time step t

No interaction
$p(C_i^t \mid C^{t-1}, I_i^t = 0)$

Interaction
$p(C_i^t \mid C^{t-1}, I_i^t = 1)$

Time step t+1

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

These can depend on previous state (position of objects)

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

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

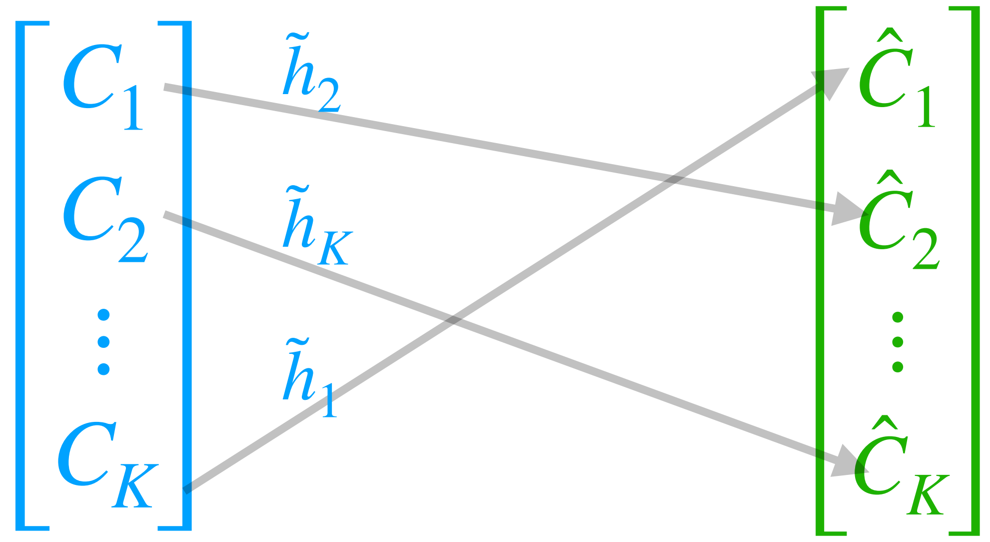
B. (***Time Variability***) *For any $C^t \in \mathcal{C}$, there exist $K+1$ different values of C^{t-1} denoted with $c^1, ..., c^{K+1} \in \mathcal{C}$, for which the vectors $v_1, ..., v_K \in \mathbb{R}^{K+1}$ with*

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

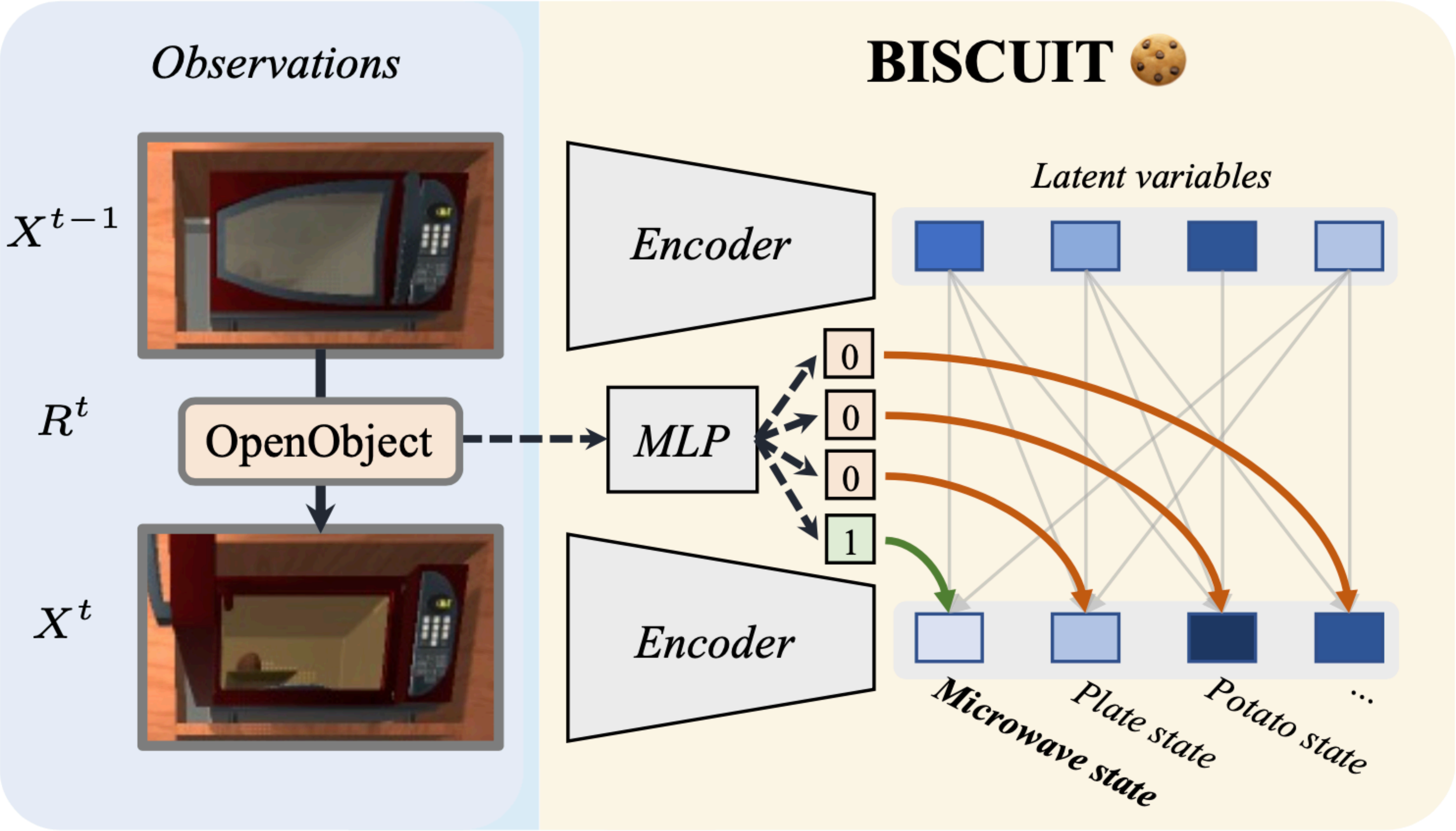
are linearly independent.

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

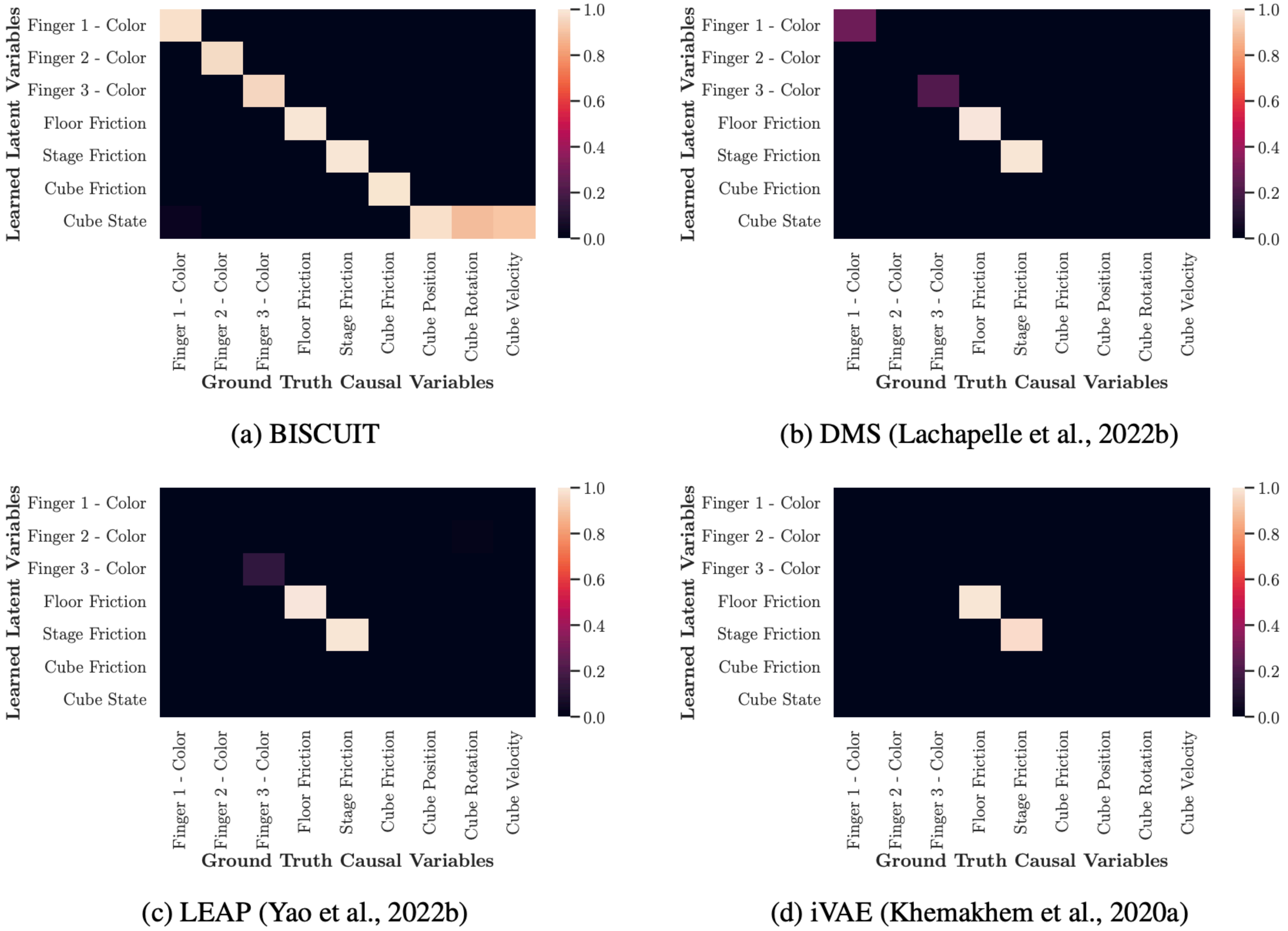
BISCUIT on CausalWorld and iTHOR

- 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.

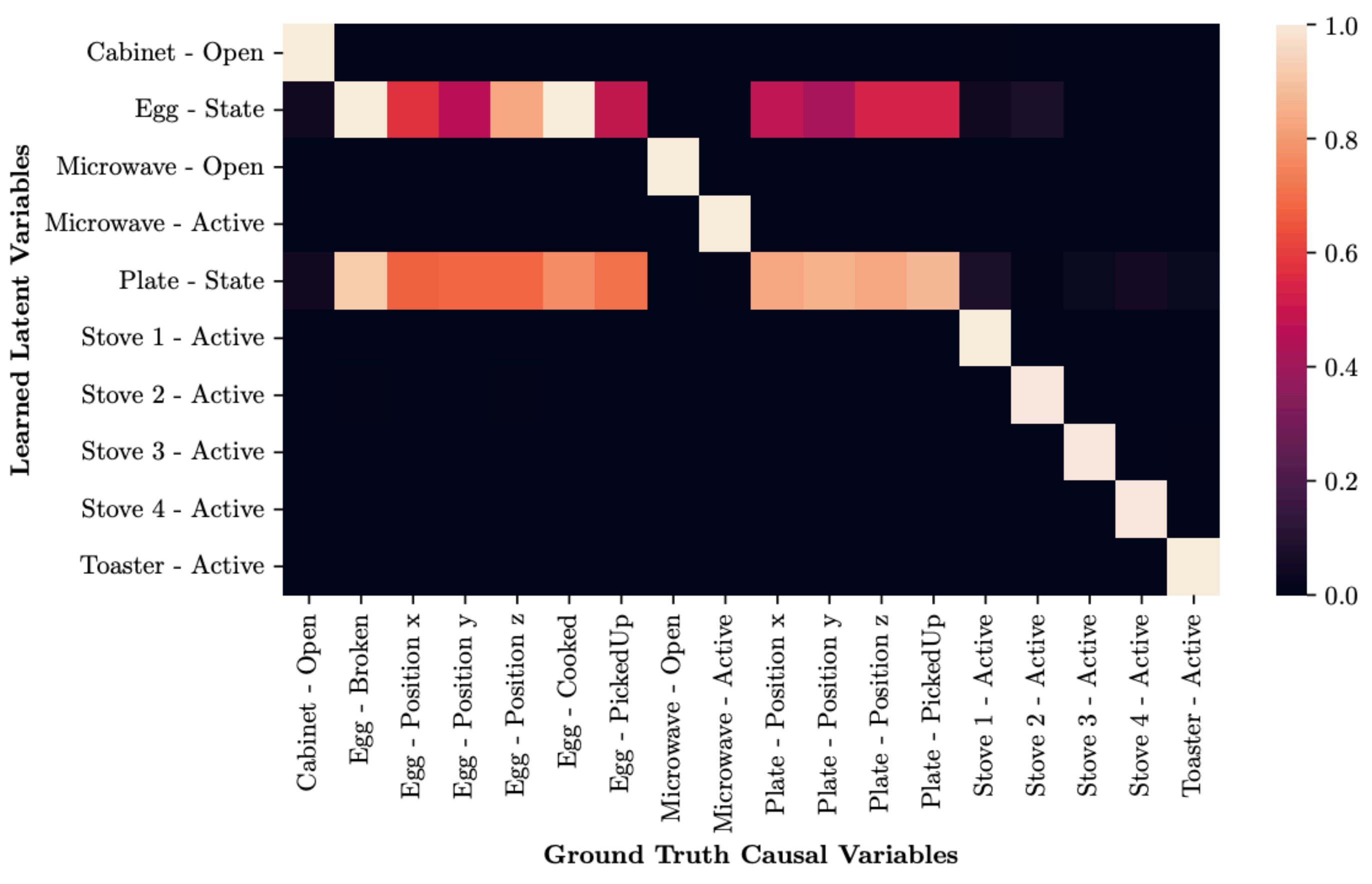
Models	CausalWorld	iTHOR
iVAE (Khemakhem et al., 2020a)	0.28 / 0.00	0.48 / 0.35
LEAP (Yao et al., 2022b)	0.30 / 0.00	0.63 / 0.45
DMS (Lachapelle et al., 2022b)	0.32 / 0.00	0.61 / 0.40
BISCUIT-NF (Ours)	**0.97** / 0.01	**0.96 / 0.15**

BISCUIT on CausalWorld - R^2 metric

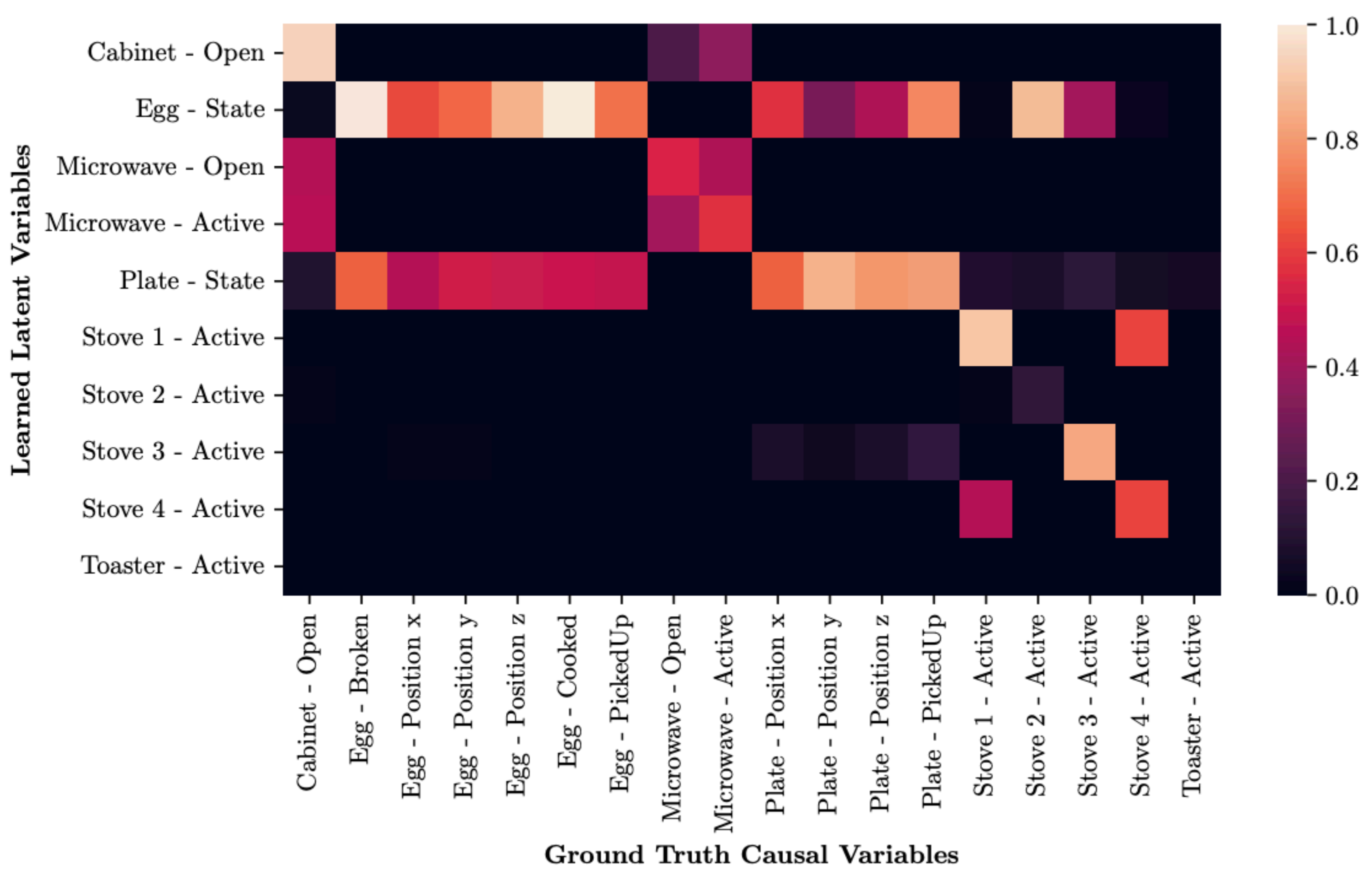


Here we assign the permutation based on the most correlated latent variable

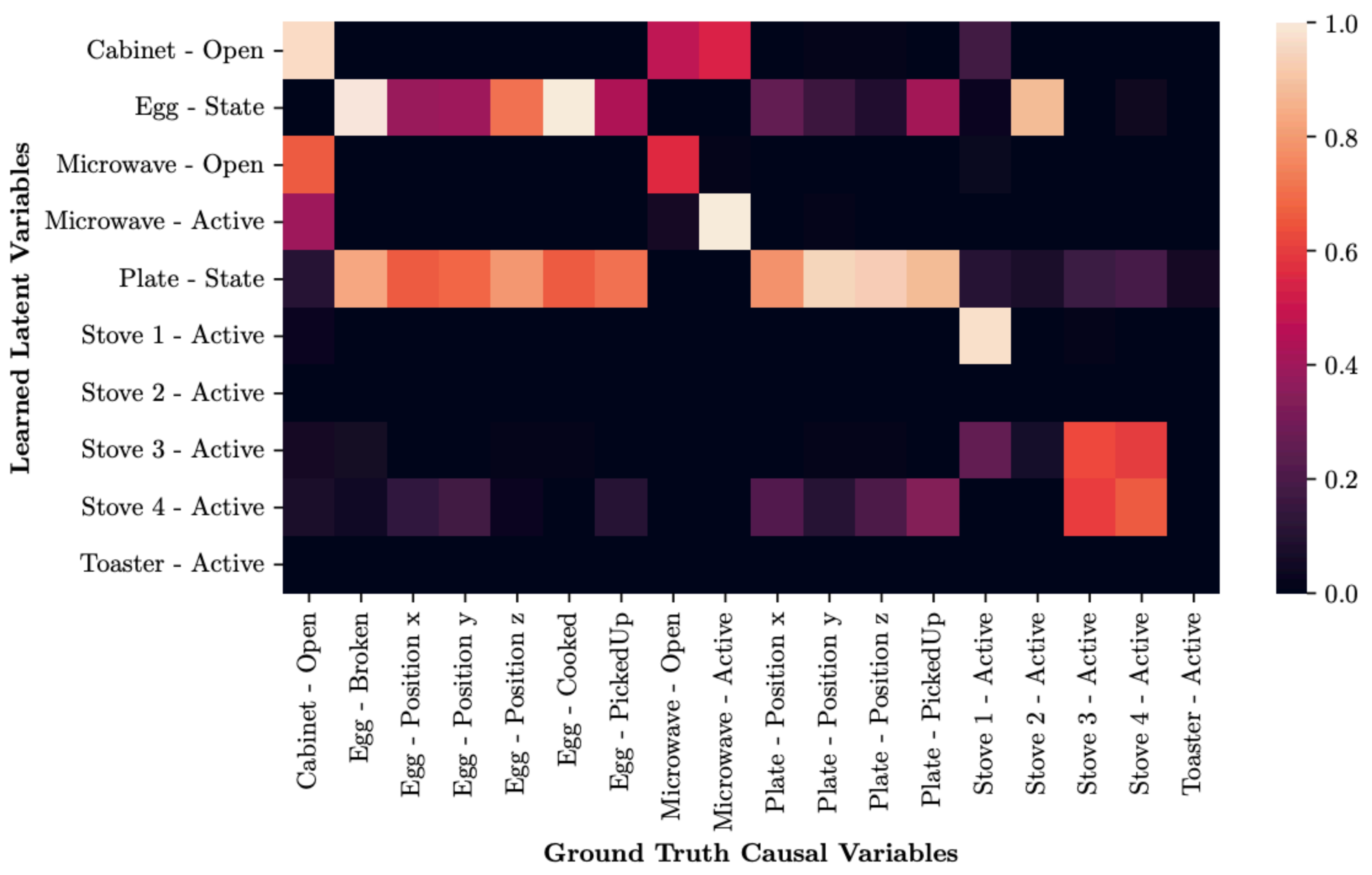
BISCUIT on iTHOR - R^2 metric



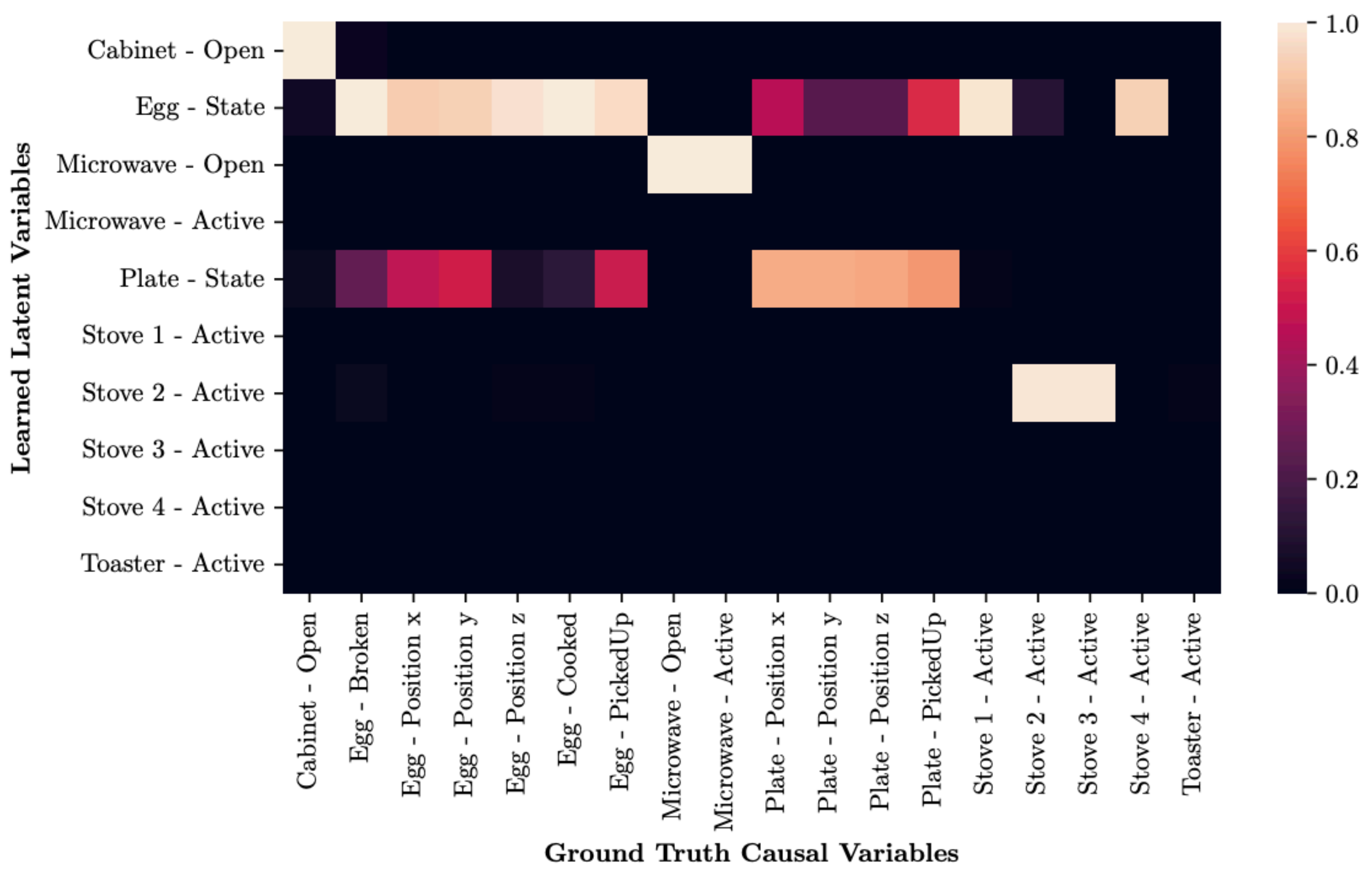
(a) BISCUIT



(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

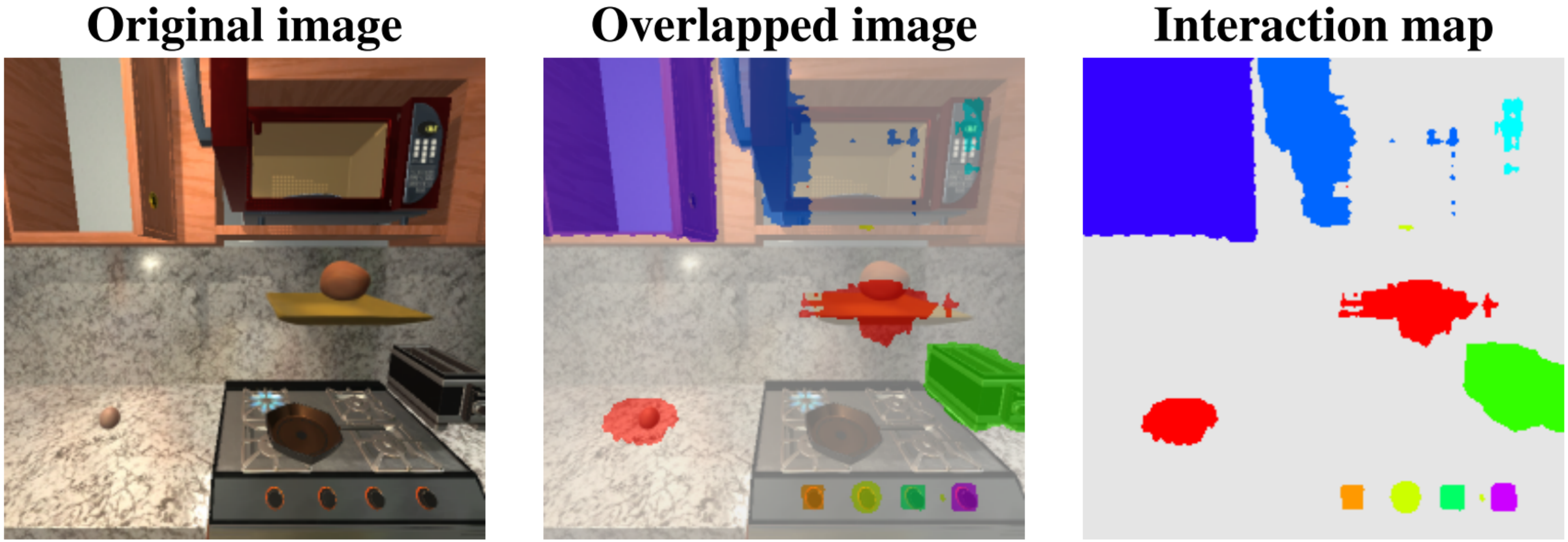
Input image 2

Generated Output

Latents from image 2

Microwave Open

BISCUIT on iTHOR - dynamic interaction map



Conclusions & Future work

- Causal representation learning (CRL) is an exciting new field that allows us to extract causal semantics from images with provable guarantees
- CRL can work on realistic images/simulators in temporal settings with actions
  - CITRIS does not have parametric or graphical assumptions, but requires knowing the intervention targets
  - **BISCUIT overcomes this limitation, requiring only a labelled action**
- **Future work:**
  - Gap between theory and real-world data -> working on CRL without actions
  - Downstream tasks for CRL -> combination with RL, XAI

References

[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.

[Lachapelle et al., 2024] Lachapelle, S., López, P. R., Sharma, Y., Everett, K., Priol, R. L., Lacoste, A., & Lacoste-Julien, S. (2024). Nonparametric Partial Disentanglement via Mechanism Sparsity: Sparse Actions, Interventions and Sparse Temporal Dependencies. *arXiv preprint arXiv:2401.04890*.

[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.

[Lippe et al., 2023b] Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. BISCUIT: Causal Representation Learning from Binary Interactions. UAI 2023.

[Yao et al., 2022] 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.