Disentanglement and Causality in Representation Learning

Generative and Deep Learning (GDL)

Riccardo Massidda (riccardo.massidda@di.unipi.it)
Davide Bacciu (davide.bacciu@unipi.it)

UNIVERSITÀ DI PISA

Outline

- **Desiderata** for Representation Learning
- **Unsupervised** Approaches and Theoretical Issues
- **Non-identifiability** of Disentangled Representations
- **Weak Supervision** for Disentangled Representation Learning
- **Causal** Representation Learning

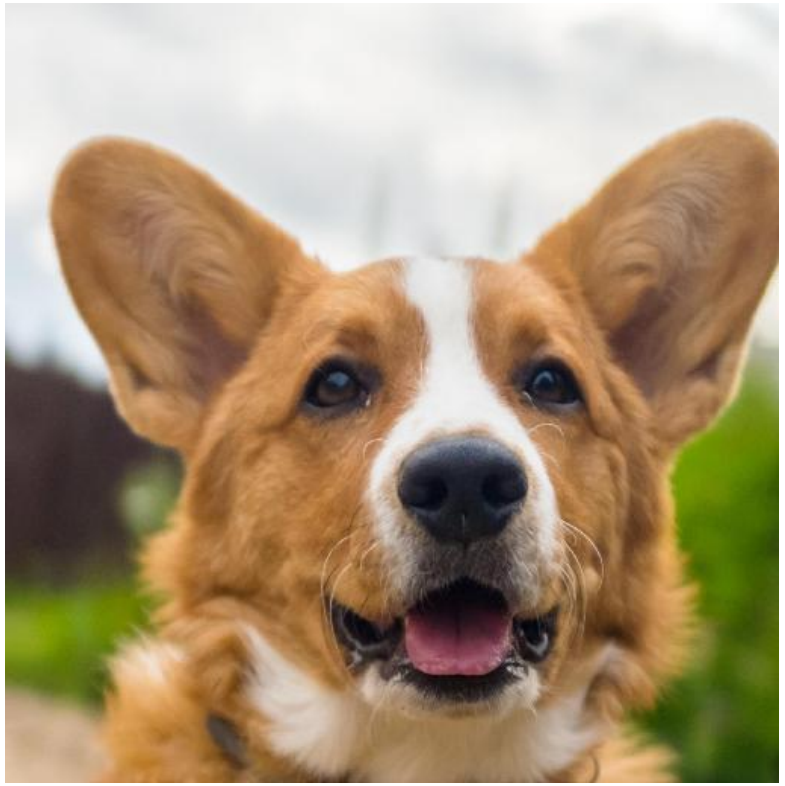
Disentangled Representations

UNIVERSITÀ DI PISA

UNIVERSITÀ
DI PISA

Representation Learning

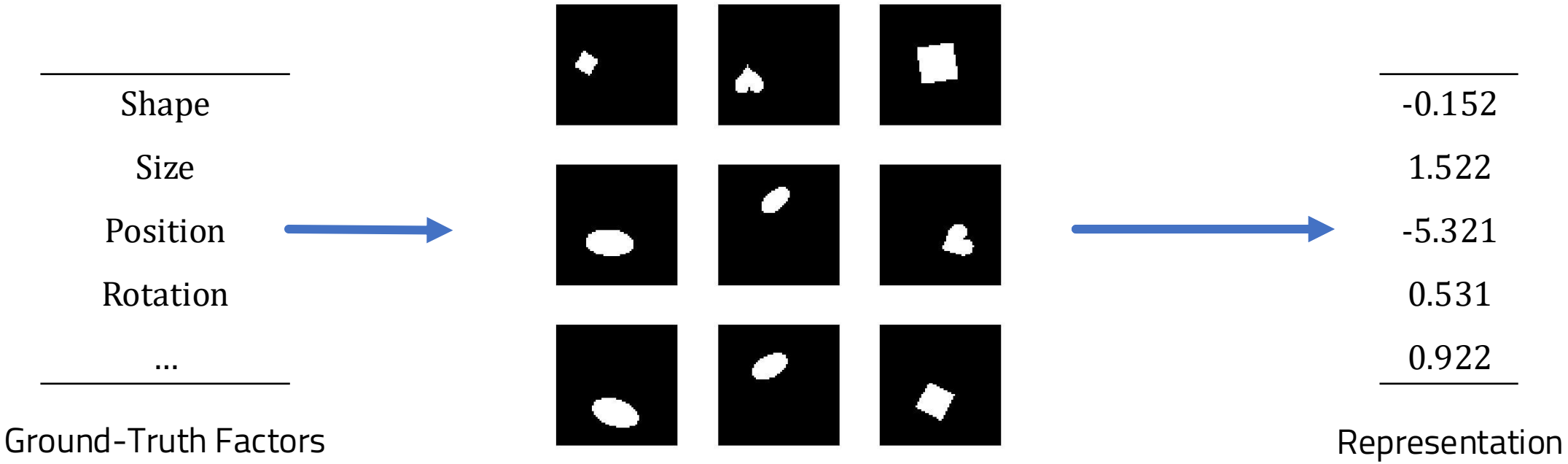
Ground-Truth Factors s
Nose Shape
Background
Ear Length
Fur Color
…



Observation x

Representation z
-0.152
1.522
-5.321
0.531
0.922

Representation Learning



Factors of Variations

- Factors of variations are the ground-truth **data-generating features**.
- The factors are assumed to be **independent**, i.e., when **intervening** on a factor the other factors remain unchanged.
- A representation is **disentangled** whenever it matches in a one-to-one relation a subset of ground-truth factors of variation.
- Intervening on a feature of a disentangled representation **only** affects a particular ground-truth factor.
- Disentangled representations are fundamental for **fairness**, **interpretability**, and **compositionality**.

Unsupervised Disentanglement

UNIVERSITÀ DI PISA

UNIVERSITÀ
DI PISA

β-VAE

- In a VAE, **Kullback-Leiber divergence** pushes the posterior toward the prior $\mathrm{N}(0, \mathrm{I})$.
- If the prior is factorized and posterior matches it, does it learn **independent latents**?

$$\log p(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}(q_\phi(z|x) \,\|\, p(z)) =: \mathcal{L}_{\mathrm{ELBO}}$$

- Beta-VAEs **constrain** the KL divergence to be sufficiently small.

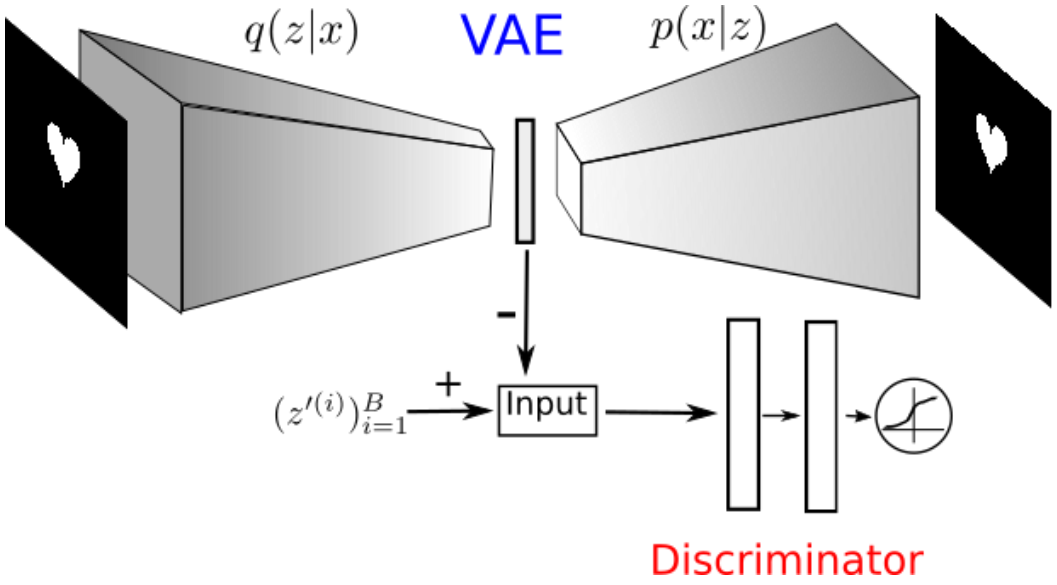
$$\max_{\phi,\theta} \; \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \quad \text{subject to} \quad \mathrm{KL}(q_\phi(z|x) \,\|\, p(z)) < \epsilon$$

- Leading to the following relaxed Lagrangian formulation:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta\, \mathrm{KL}(q_\phi(z|x) \,\|\, p(z))$$

FactorVAE

- β-VAE matches the **posterior** $Q_\varphi(Z \mid X)$ to have independent components .
- This does **not imply that the marginal** $Q_\varphi(Z)$ will be independent.
- **FactorVAE** enforces this property, by approximating the **Total Correlation** with a discriminator and targeting:



$$\mathcal{L}_{\text{FactorVAE}} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \,\|\, p(z)) - \gamma\, \text{TC}(q_\phi(z))$$

$$\text{TC}(q_\phi(z)) = \text{KL}\left(q_\phi(z) \,\middle\|\, \prod_j q_\phi(z_j)\right) \approx \mathbb{E}_{q_\phi(z)}\left[\log \frac{D(z)}{1-D(z)}\right]$$

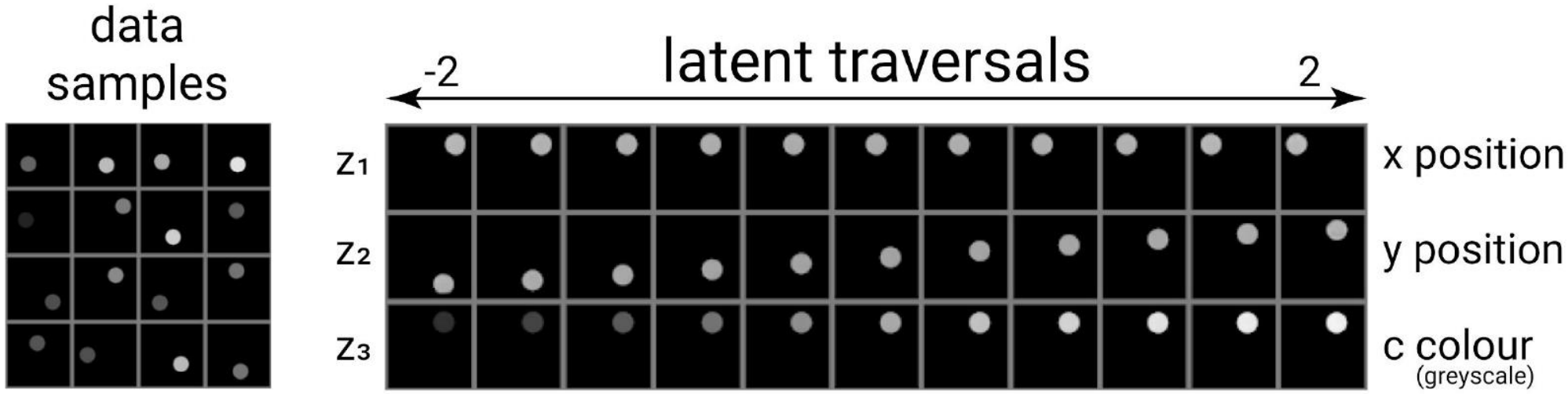
β-TCVAE

- β-TCVAE combines the two approaches by **decomposing the KL divergence**. Into:
  - Mutual Information between X and Z,
  - Total Correlation of the marginal distribution of Z, and
  - Component-Wise KL divergence between the posterior and the prior.

$$\mathcal{L}_{\beta\text{-TCVAE}} = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \alpha\, I(x;z) - \beta\, \mathrm{TC}(q_\phi(z)) - \gamma \sum_j \mathrm{KL}(q_\phi(z_j) \,\|\, p(z_j))$$

- Typically disentangles *better*, at the cost of **more hyperparameters**.

Latent Traversal for Qualitative Comparisons



Towards a Definition of Disentangled Representations – Higgins et al. (2018)

Unsupervised Disentanglement

- Using these variants of the Variation Autoencoder, we retrieve **unsupervised marginally independent** latent representations.
- Qualitative tests suggest they capture something from the factors.
- Marginal independence is a **necessary condition** of the ground-truth distribution of the sources: is it enough?

Indeterminacy in Unsupervised Disentanglement

- Using these variants of the Variation Autoencoder, we retrieve **unsupervised marginally independent** latent representations.
- Qualitative tests suggest they capture something from the factors.
- Marginal independence is a **necessary condition** of the ground-truth distribution of the sources: is it enough?
- **No!** In fact, there exist **infinite entangled solutions** with the same loss value.

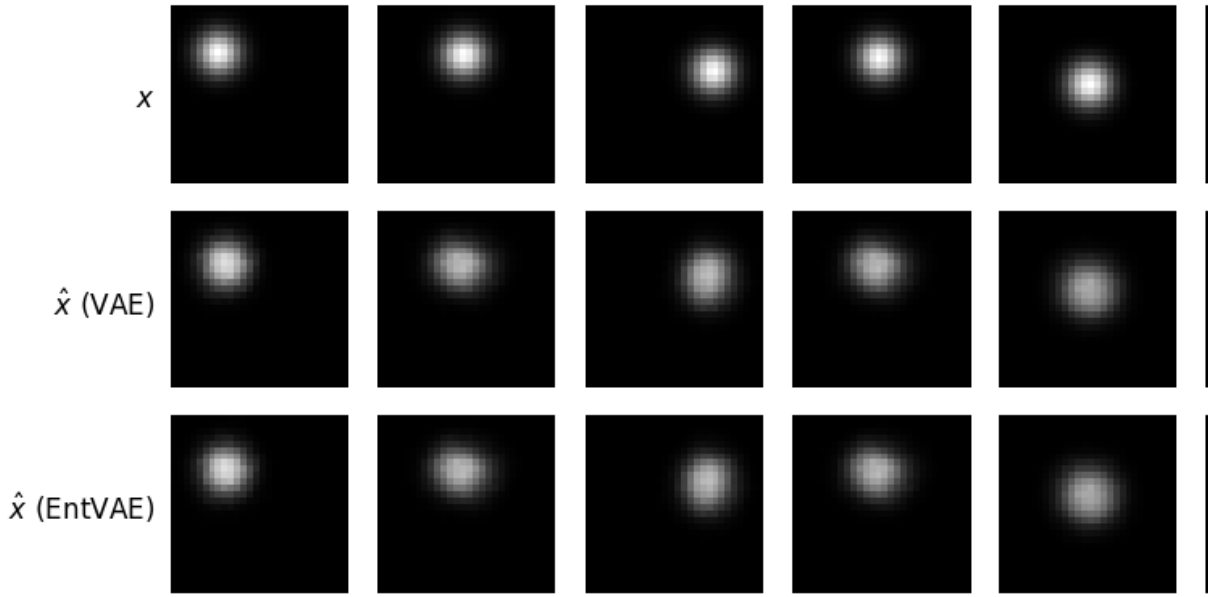
Indeterminacy in Unsupervised Disentanglement

Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\boldsymbol{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \boldsymbol{u}) = P(f(\mathbf{z}) \leq \boldsymbol{u})$ for all $\boldsymbol{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations – Locatello et al. (2019)

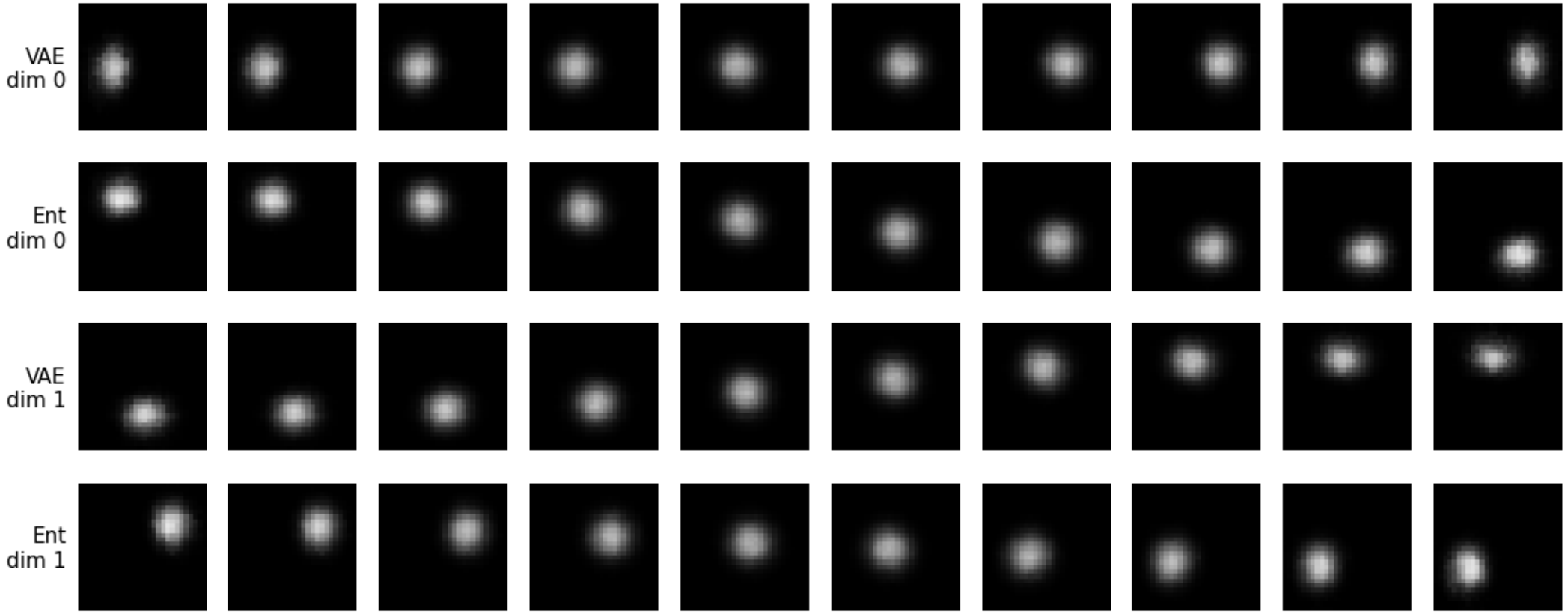
Indeterminacy in Unsupervised Disentanglement

- Given a Variational Autoencoder, we choose one of possible infinite transformation as described by Locatello et al. (2019).
- In practice, we apply f after encoding and f^{-1} before decoding.




```
Baseline VAE:
loss=89.6392  BCE=68.6759
KL=2.0963
Entangled VAE:
loss=89.6392  BCE=68.6759
KL=2.0963
Loss difference:
7.32e-07
Max |VAE.recon(x) - eVAE.recon(x)|:
1.79e-07
```

Indeterminacy in Unsupervised Disentanglement



Indeterminacy in Unsupervised Disentanglement

- Different representations can still be **marginally independent**, have the **same loss value**, but be **entangled** in respect to the ground-truth sources.
- Similar results were known in **non-linear ICA** (Hyvarinen and Pajunen, 1999), Locatello et al. (2019) connect them to the notion of disentanglement in Deep Learning.
- If they are indistinguishable, why do some unsupervised methods disentangle better than others?
  - **Inductive bias** is fundamental! (nice intuitions here, here, and here)
- What if we want **formal guarantees** to end up in a disentangled representation?

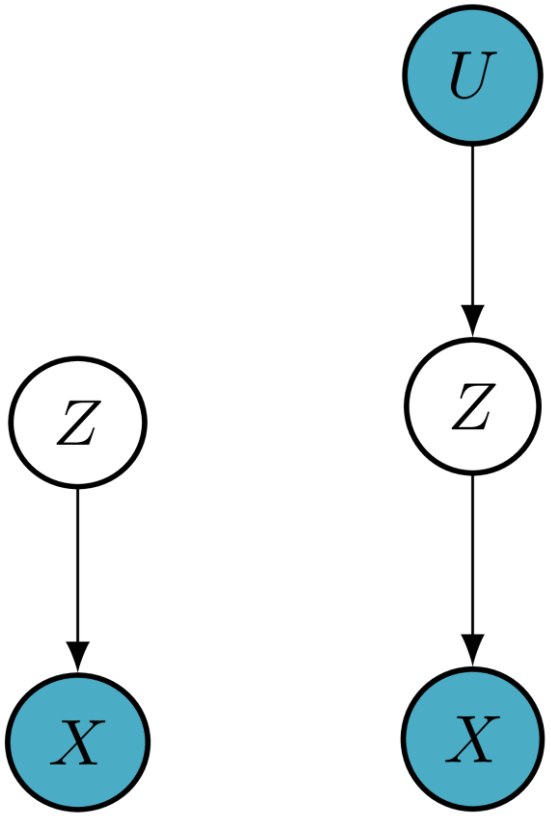
Weakly-Supervised Disentanglement

UNIVERSITÀ DI PISA

UNIVERSITÀ
DI PISA

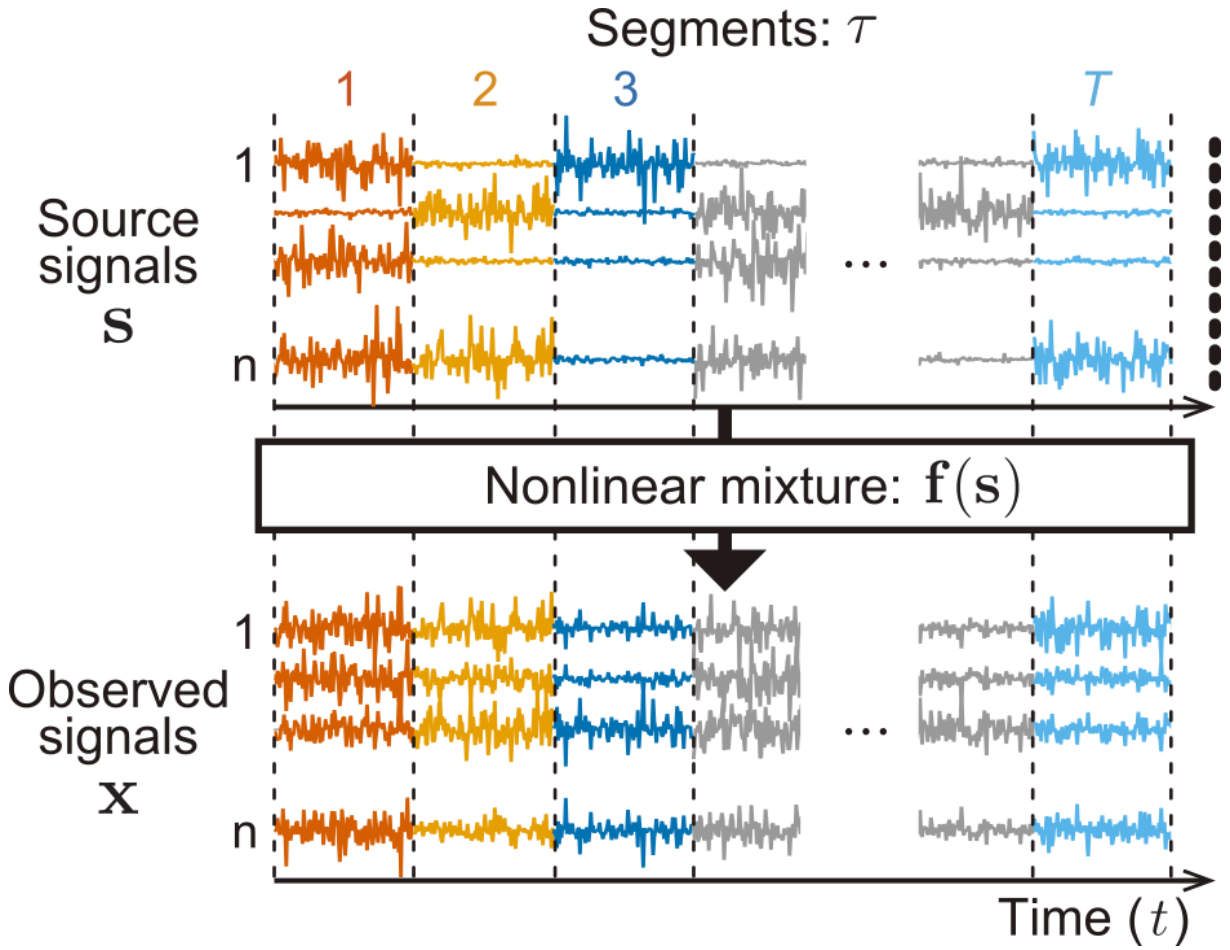
Disentanglement with Auxiliary Variables

- Impossibility holds for a single distribution.
- Hint: if the source distribution varies across conditions, and we label which condition each observation comes from, we cannot reuse the same transformation f!
- Auxiliary variable u (time segment, environment, class): if p(s|u) changes with u, sources become identifiable.
- Only the correct factorization is simultaneously consistent with all conditions.



Time-Contrastive Learning

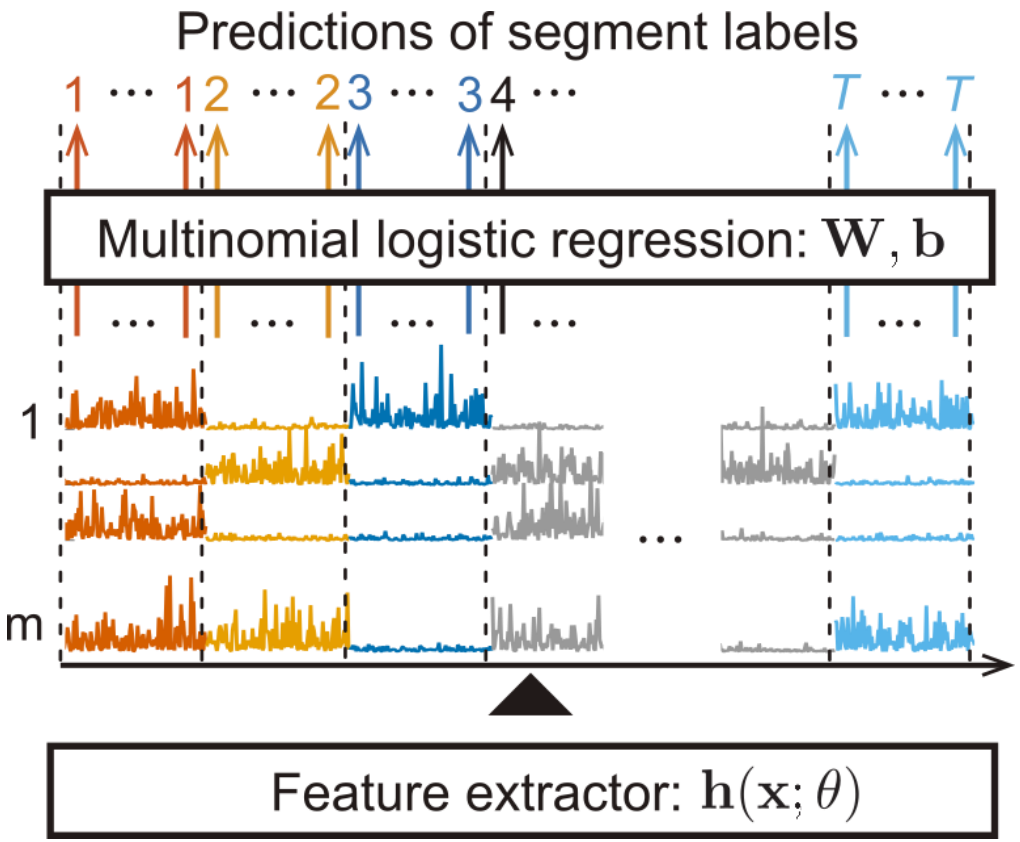
- We consider a non-stationary time series originated by source signals S and observed as a mixture X.
- The goal is to reconstruct S from X
  - …a.k.a, the cocktail party problem 🍸
- We can exploit that sources are conditional independent given time.



Time-Contrastive Learning

1. Divide a multivariate time-series into segments.
2. Associate to each data-point the corresponding segment.
3. Train a classifier to recover the segment label from the datapoint.

$$\boldsymbol{w}_\tau^T h(\boldsymbol{x}_t \mid \theta) + b_\tau$$



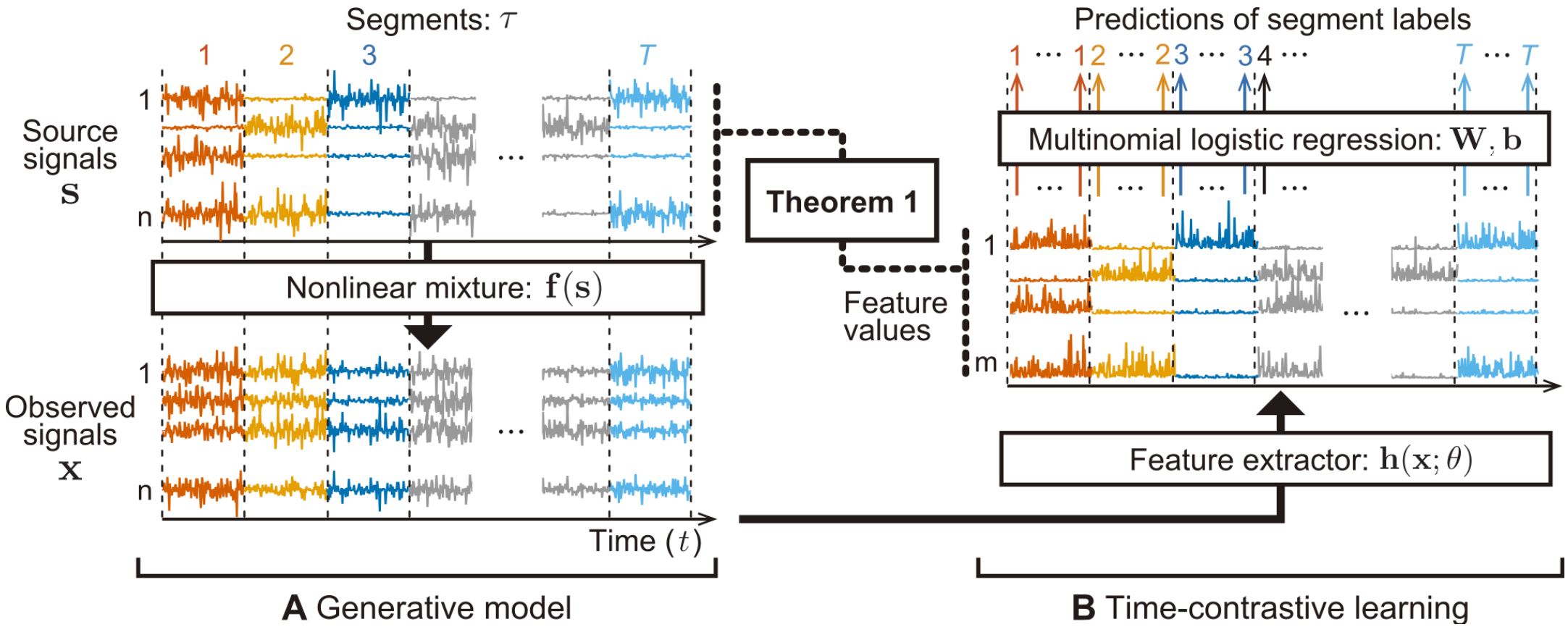
Time-Contrastive Learning

Intuitive requirements:

1. Sources come from a distribution in the **exponential family**.
2. The feature extractor has the **same dimensionality** of the data.
3. Distribution per-segment is "**different enough**".

Then, we can recover the original sources up to permutation and a strictly monotonic component-wise transformation.

Time-Contrastive Learning



Identifiable Variational Autoencoders (iVAE)

- iVAE adapts the identifiability results of TCL to Variational Autoencoders.
- For simplicity, we consider the latent representation to be zero-centered with log-variance depending on the auxiliary variable. Formally,

$$p(z|u) = \prod_{i=1}^{n} \sqrt{\frac{-\lambda_{i,1}(u)}{\pi}} \exp\left(z_i^2 \cdot \lambda_{i,1}(u)\right)$$

- Where the lambda function is modeled by a neural network, and has form

$$\lambda_{i,1}(u) = -\frac{1}{2\sigma_i^2(u)}$$

Identifiable Variational Autoencoders (iVAE)

- We still train by maximizing the ELBO, which does not compare anymore against the marginal probability of Z, but with the conditional probability given the context U.

$$\underbrace{\mathbb{E}_{q_\phi(z|x,u)}\left[\log p_\theta(x|z)\right]}_{\text{reconstruction}} - \underbrace{\mathbb{E}_{q_\phi(z|x,u)}\left[\log q_\phi(z|x,u) - \log p_\lambda(z|u)\right]}_{\mathrm{KL}[q_\phi(z|x,u)\|p_\lambda(z|u)]}$$

Identifiable Variational Autoencoders (iVAE)

Intuitive requirements in the simplified Gaussian scenario:

1. The relation between sources and observations is injective.
2. The family of distributions Z|U contains the ground-truth distribution.
3. Each $\lambda(u)$ leads to "sufficiently different" distributions of Z.

Then, we can recover the original sources up to permutation and a component-wise non-linear transformation.

Causal Representation Learning

UNIVERSITÀ DI PISA

UNIVERSITÀ
DI PISA

Causal Representation Learning

- Assuming their independence, we limit the choice of factors.
- We would like to uncover sources even if they are **causally related**.
- Causal Representation Learning assumes that the factors are related by a Structural Causal Model (SCM) and tries to recover it from observations.
- Several causal graphs could correspond to the observational distribution. Therefore, as with causal discovery, we need **interventions** to uncover the causal graph.

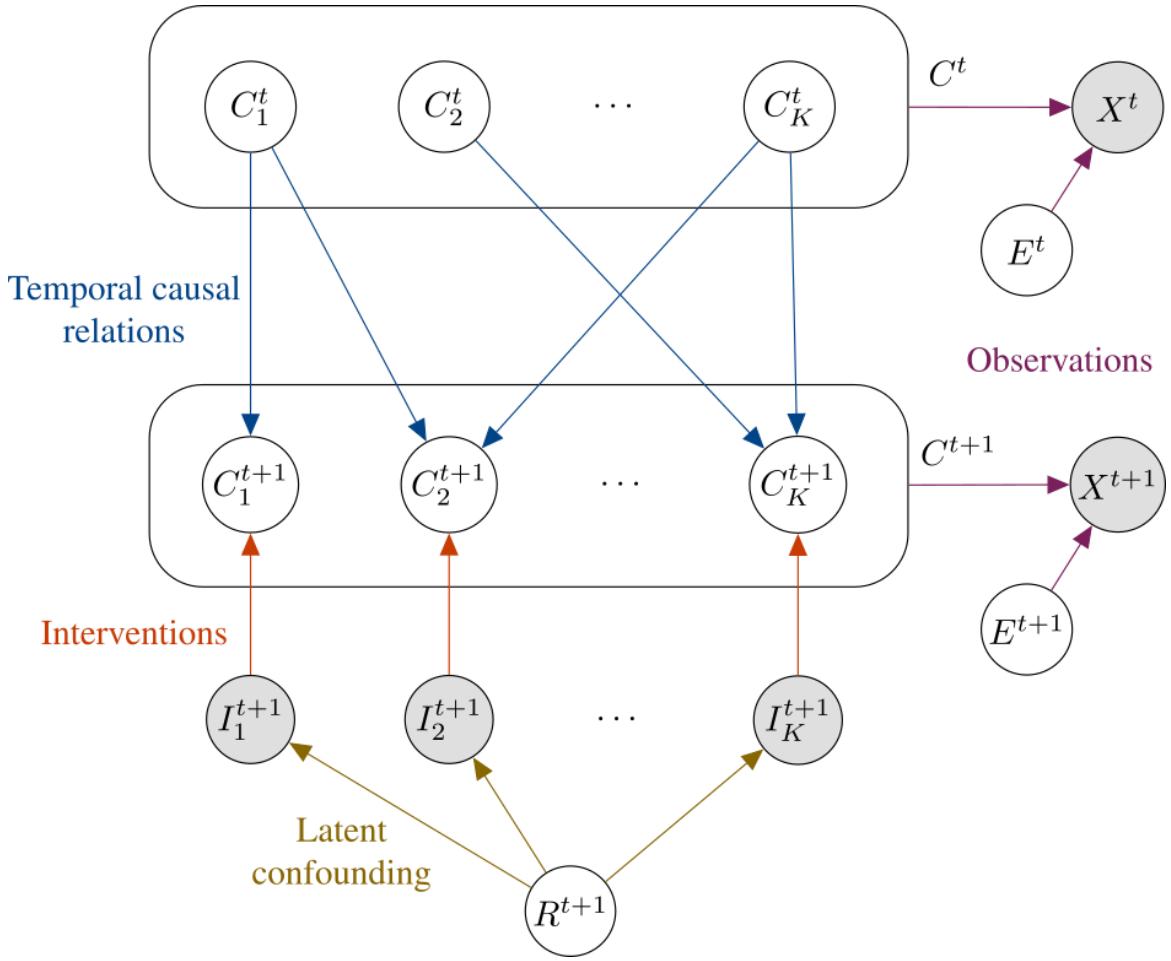
Temporal Intervened Sequences (TRIS)

- Observations depend on the causal variables and an exogenous term through a mixing function

$$\boldsymbol{x}^t = h(\boldsymbol{c}_1^t, \dots, \boldsymbol{c}_k^t, \boldsymbol{e}^t)$$

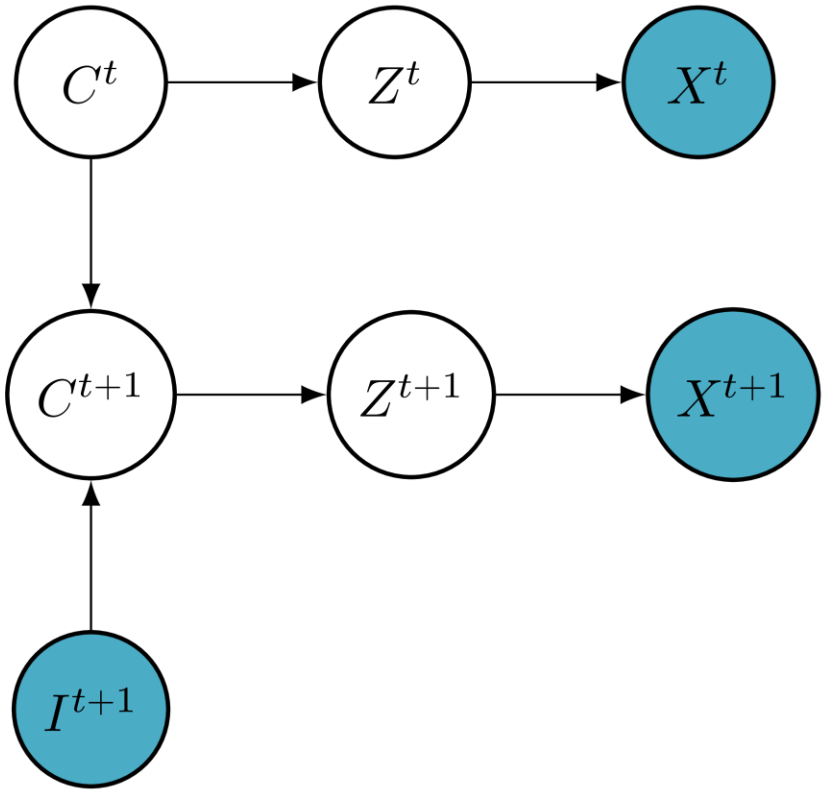
- For each causal variable, we assume to have two possible regimes, the observational and the interventional.
- For each observation X^t, we know whether a causal variable has been intervened or not, but not how!
- Overall, data is composed by triplets

$$\mathcal{D} = \{(x^1, x^2, I^2), \dots (x^{T-1}, x^T, I^T)\}$$



Representing Causal Variables

- Instead of directly representing values of causal variables, they are mapped to a m-dimensional latent space.
- The goal is to learn:
  - The unmixing function g from observations to latents.
  - The assignment function **ψ** from latent space to causal variables.



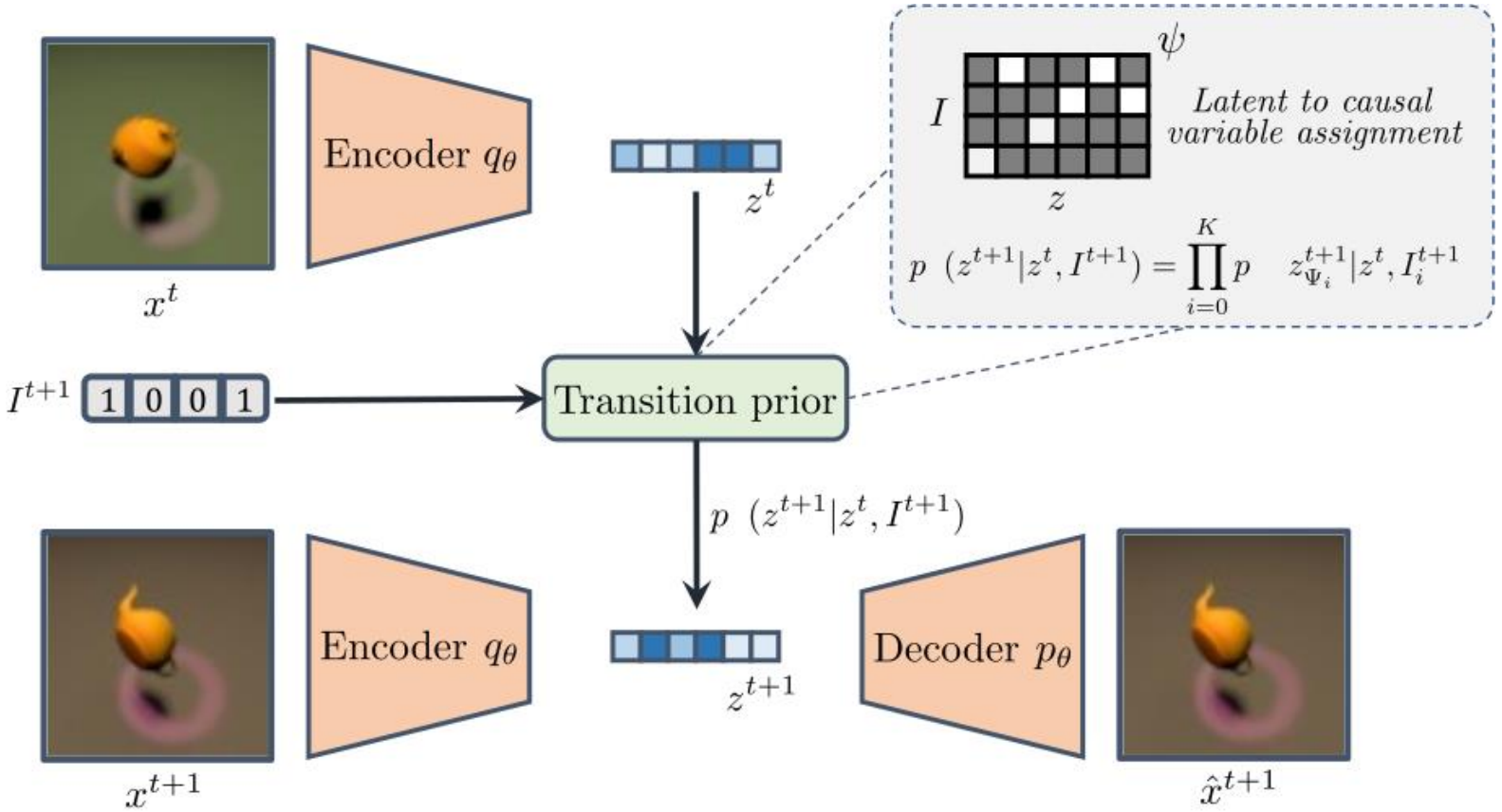
Learning TRIS using CITRIS-VAE

- As in iVAE, the intuition is to train the latent space of a VAE to match a conditional distribution instead of a fixed prior.
- In CITRIS, the transition prior depends on the previous time-step and the interventions.

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

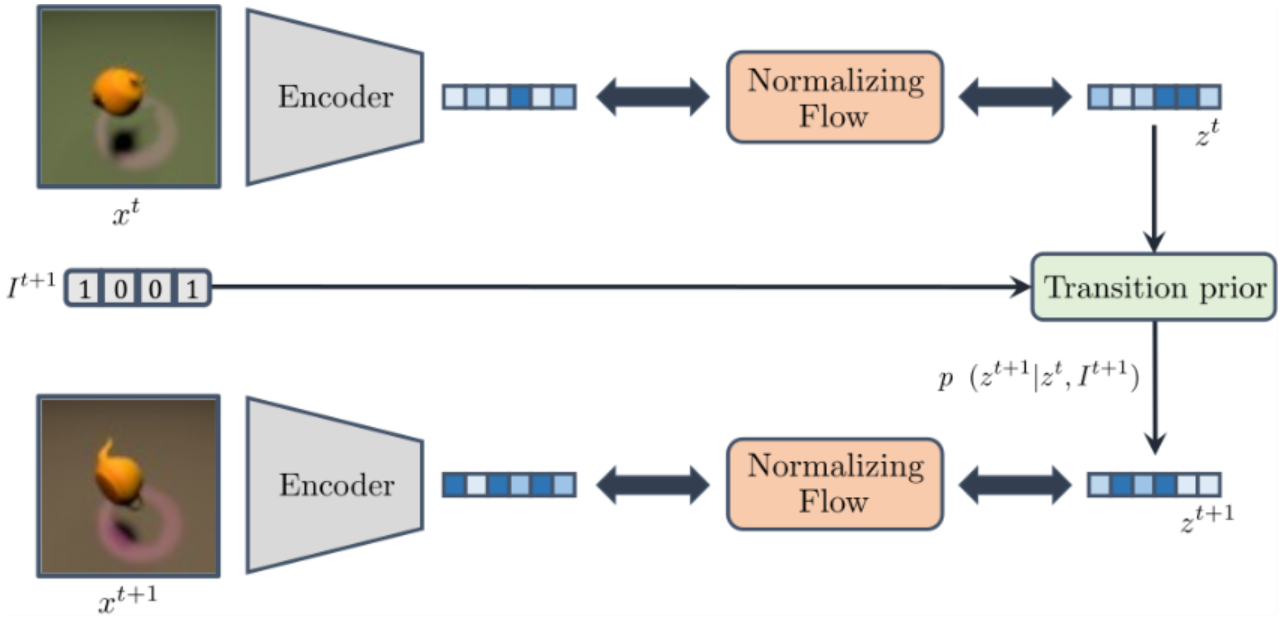
- In practice, for each feature of the latent-space, the transition prior computes the mean of a standard Gaussian distribution.
- Empirically, training in parallel a supervised Target Classifier recovering the interventions target from the latent codes helps retrieving disentangled representations.
- The causal graph can be recovered by performing causal discovery on the latent codes.

Learning TRIS using CITRIS-VAE



Learning TRIS using CITRIS-NF

- Instead of learning a disentangled representation, we can disentangle existing representations: pre-train an unsupervised Autoencoder and map with **normalizing flows**.



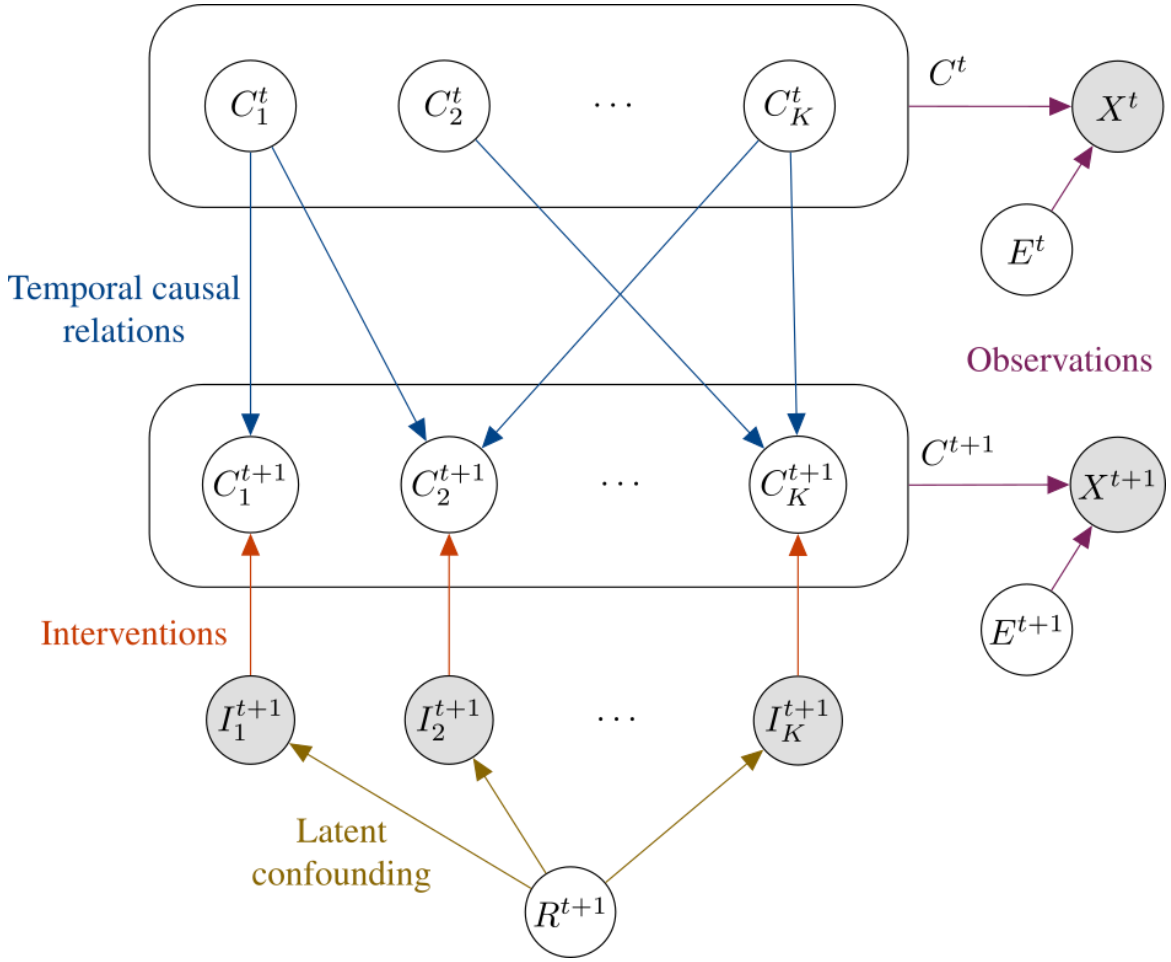
$$\mathcal{L}_{\text{NF}} = \underbrace{-\sum_{i=0}^{K} \log p_\phi\left(z_{\psi_i}^{t+1} \mid z^t, I_i^{t+1}\right)}_{\text{transition prior NLL}} \underbrace{-\log\left|\det \frac{\partial f_\theta(e^{t+1})}{\partial e^{t+1}}\right|}_{\text{flow log-det Jacobian}}$$

Identifiability of CITRIS

In the limit of infinite data, CITRIS can correctly recover the ground-truth causal variables if the following holds for each variable C_i

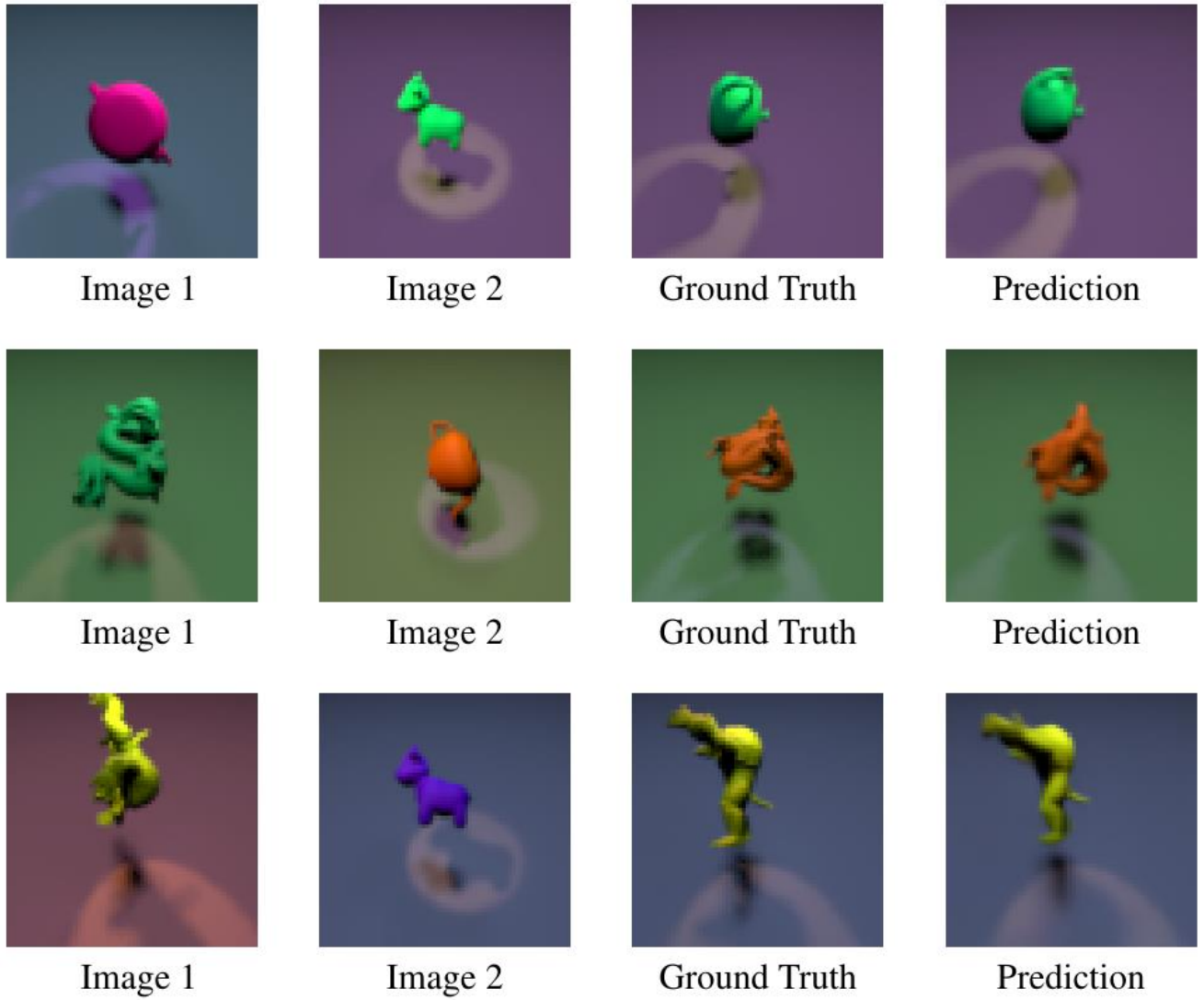
1. There exists at least a regime where C_i is intervened, and Ci is not always intervened with another variable C_j.
2. The variable always depends on its intervention variable I_i.

$$C_i^{t+1} \not\perp I_i^{t+1} \mid C^t, I_j^{t+1}. \quad \forall j \neq i$$



Triplet Evaluation

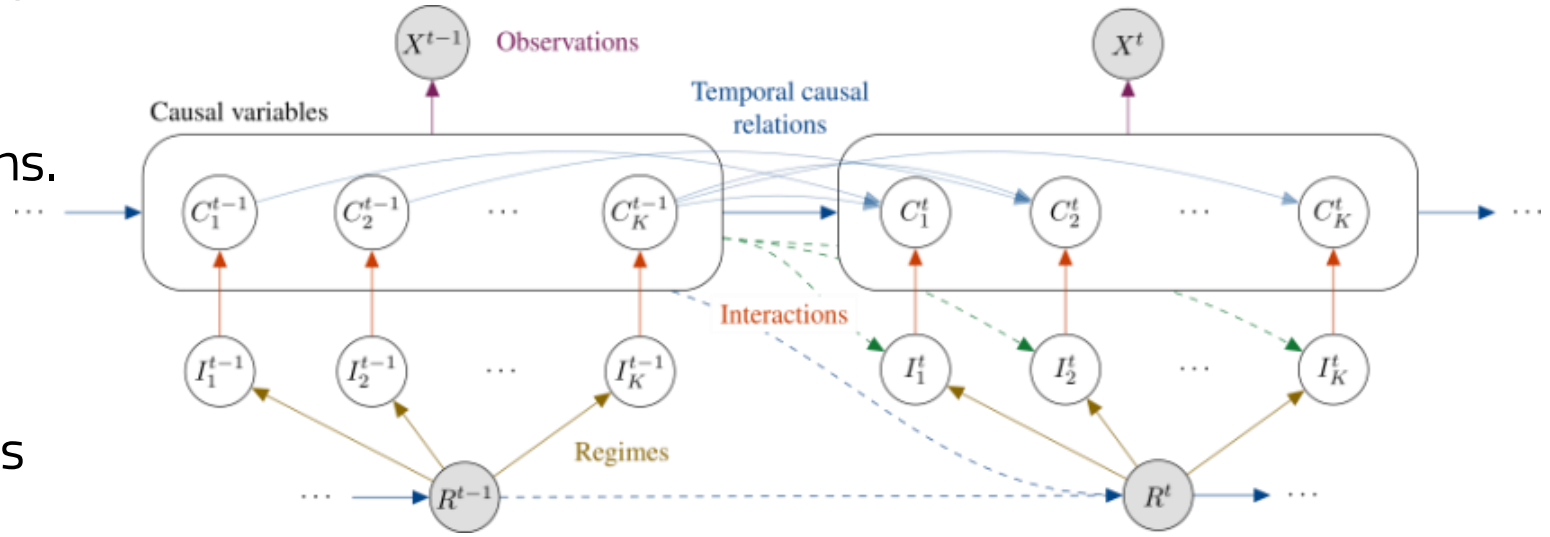
- Given a trained CITRIS model, we can perform interventions "interchange interventions", where we replace the value of a causal variable from an image with the value of a causal variable of another.



Causal Representation Learning from Binary Interactions

The assumption on known interventions can be relaxed by knowing a unique auxiliary variable, which describes the distribution of the interventions.

BISCUIT applies the same principles of CITRIS to also handle temporal dependencies between regimes.



Conclusions

- **Marginal independence** is necessary but not sufficient. Locatello et al. (2019) show an infinite family of entangled solutions matching the same observational distribution: unsupervised disentanglement is provably non-identifiable.
- **Inductive bias** (architecture, regularization, prior choice) explains the empirical success of β-VAE / FactorVAE / β-TCVAE, but yields no formal guarantees.
- **Identifiability** requires breaking the indeterminacy exploiting distributional shifts produced by an auxiliary variable (TCL, iVAE).
- **Causal Representation Learning** generalizes disentanglement to causally related factors. CITRIS recovers both the latent variables and their causal structure.

Disentangled Representations (unsupervised)

- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A. (2017). **β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework**. *ICLR*.
- Kim, H., Mnih, A. (2018). **Disentangling by Factorising**. *ICML*. [FactorVAE]
- Chen, R. T. Q., Li, X., Grosse, R., Duvenaud, D. (2018). **Isolating Sources of Disentanglement in Variational Autoencoders**. *NeurIPS*. [β-TCVAE]
- Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D., Lerchner, A. (2018). **Towards a Definition of Disentangled Representations**. *arXiv:1812.02230*.

(Non-)identifiability

- Hyvärinen, A., Pajunen, P. (1999). **Nonlinear independent component analysis: Existence and uniqueness results**. *Neural Networks* 12(3): 429–439.
- Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., Bachem, O. (2019). **Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations**. *ICML* (Best Paper).

Weak supervision via auxiliary variables

- Hyvärinen, A., Morioka, H. (2016). **Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA**. *NeurIPS*. [TCL]
- Khemakhem, I., Kingma, D. P., Monti, R. P., Hyvärinen, A. (2020). **Variational Autoencoders and Nonlinear ICA: A Unifying Framework**. *AISTATS*. [iVAE]

Causal Representation Learning

- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., Bengio, Y. (2021). **Toward Causal Representation Learning**. *Proceedings of the IEEE* 109(5): 612–634.
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., Gavves, E. (2022). **CITRIS: Causal Identifiability from Temporal Intervened Sequences**. *ICML* (Spotlight).
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., Gavves, E. (2023). **BISCUIT: Causal Representation Learning from Binary Interactions**. *UAI*.

Additional Talks

- **Francesco Locatello** on Towards Causal Representation Learning (2021)
- **Dhanya Sridhar** and **Jason Hartford** on Causal Representation Learning (2023)
- **Sara Magliacane** on Causal Representation Learning in Temporal Settings (2025)