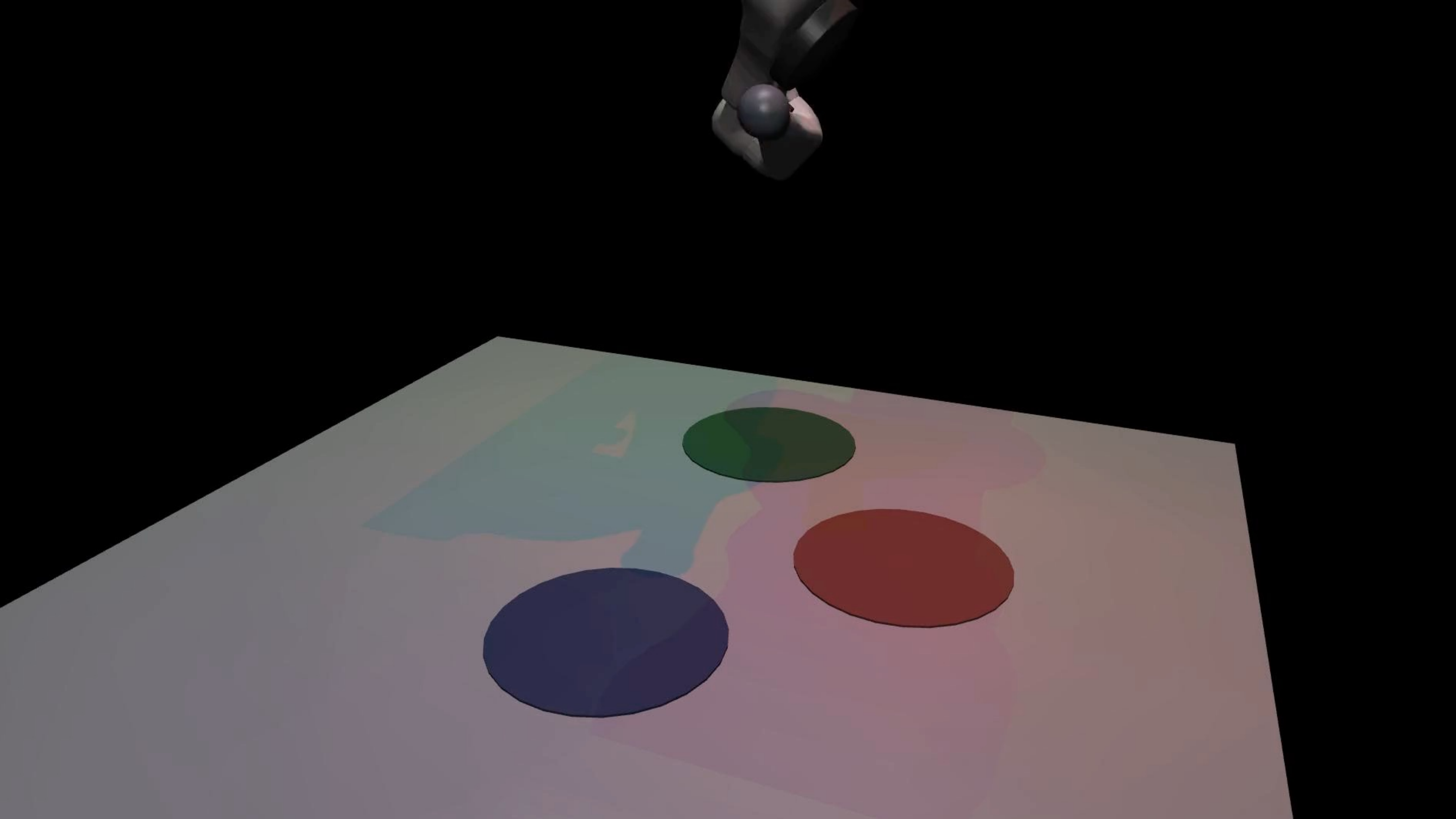
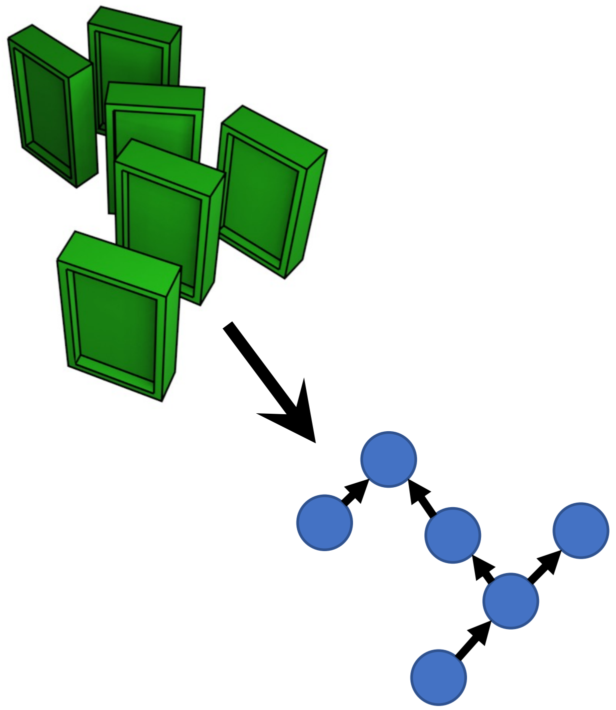
Weakly supervised causal representation learning

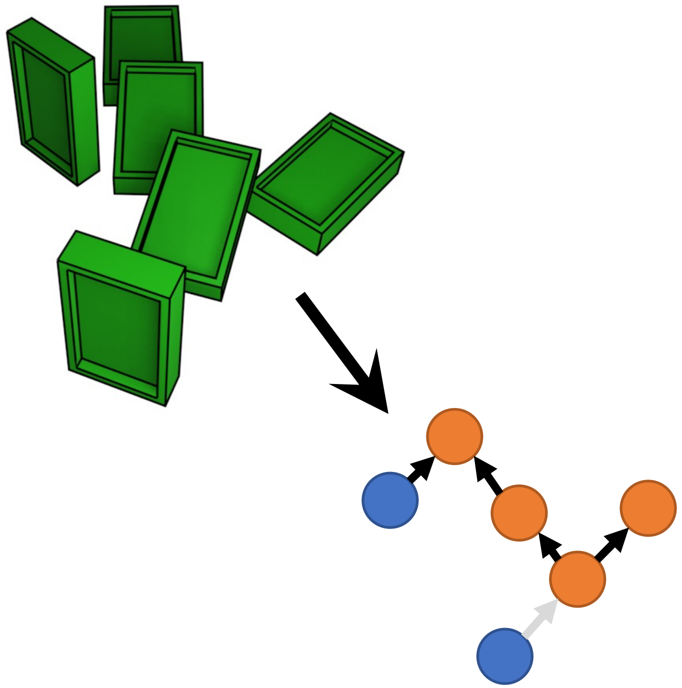
Johann Brehmer

Qualcomm Technologies Netherlands B. V.

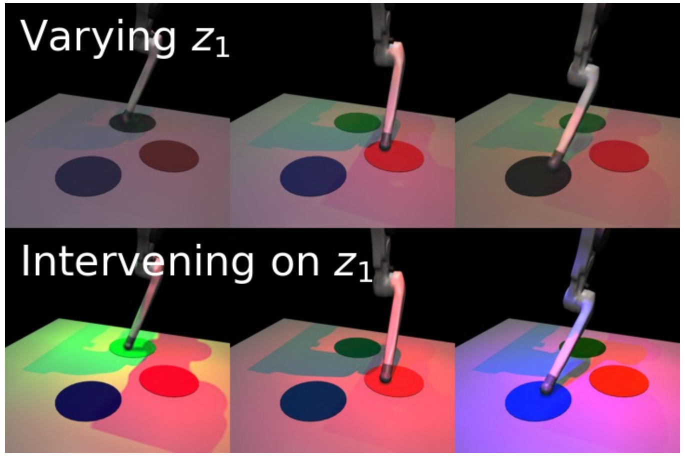




Can we **learn causal variables & causal structure from pixels**, without labels?



We prove: this is possible with **weak supervision**, when observing effects of interventions

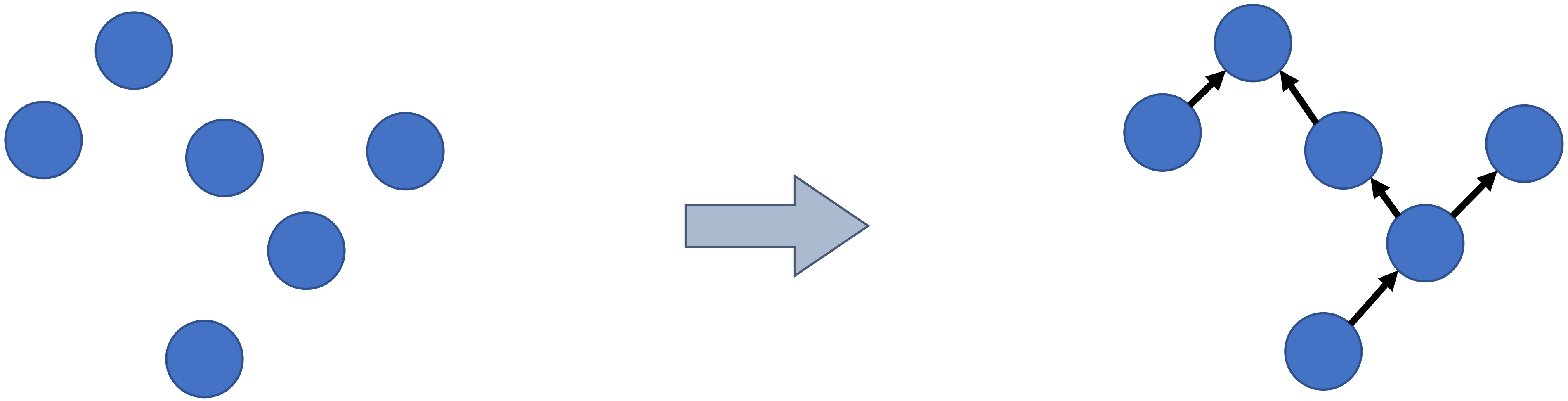


In practice, **implicit latent causal models** can identify the causal structure in image datasets

Problem

Can we learn causal representations from pixels?

Causal discovery / inference

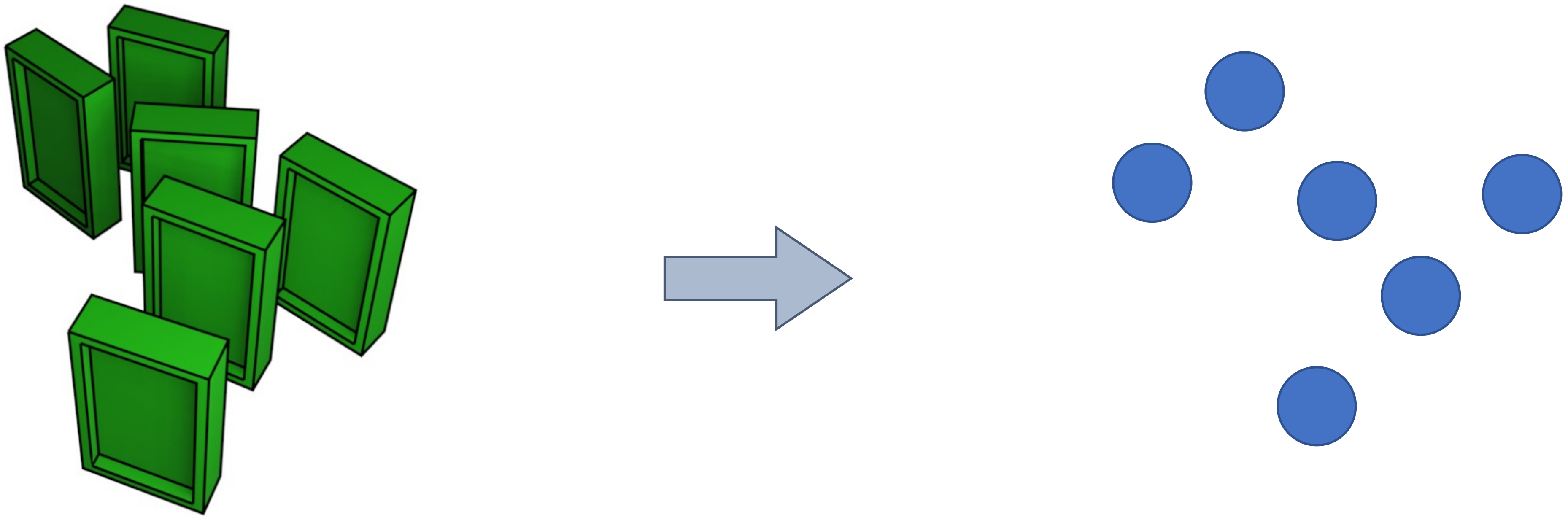


Given: dataset in terms of
high-level causal variables

Goal: learn the
causal structure

But: what if we don’t observe the causal variables?

Disentangled representation learning

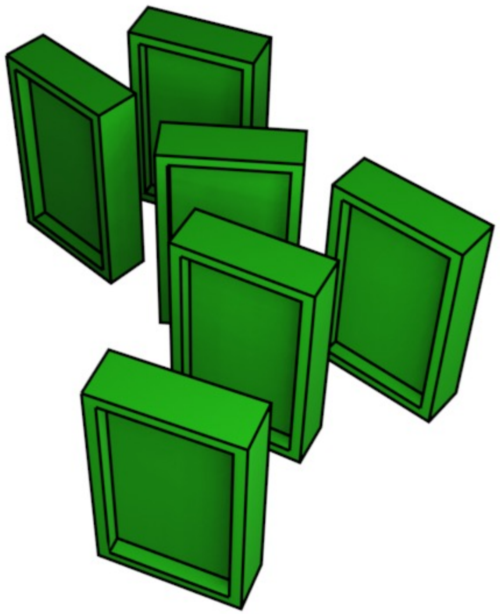


Given: **low-level, unstructured data representation** (e.g. pixels)

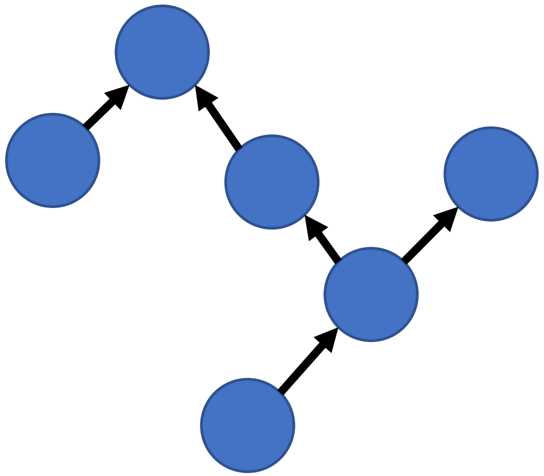
Goal: learn encoder to **high-level variables** (e.g. object positions, states, ...), usually **assuming independence**

But: useful high-level concepts are rarely independent

Causal representation learning

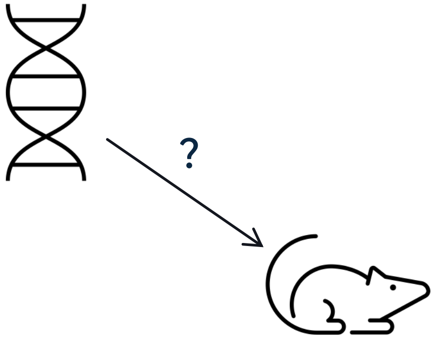


Given: **low-level, unstructured data representation**
(e.g. pixels)

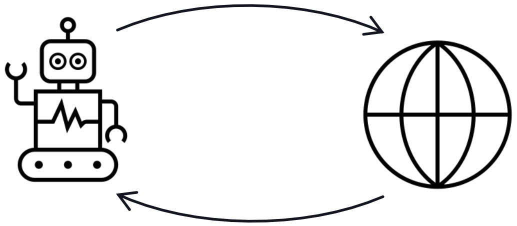


Goal: learn encoder to
high-level variables
(e.g. object positions, states, ...)
and their relations / causal structure

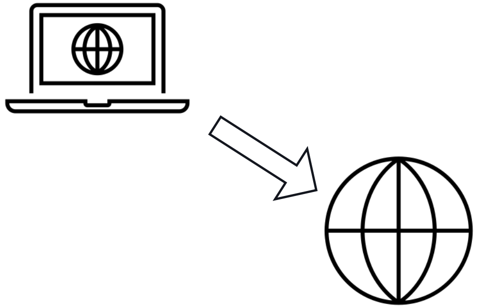
Why learn causal representations?



Causal structure may be of **scientific interest**



Causal representations are **abstractions** that may be **useful for planning**



Causal models may be more **robust to changes**

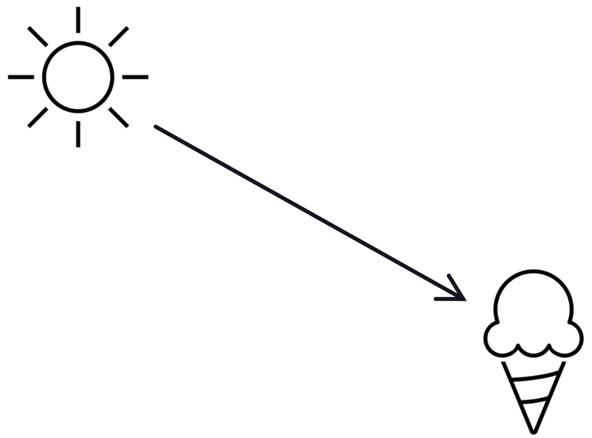
Arguably, these potential benefits have not yet been clearly demonstrated

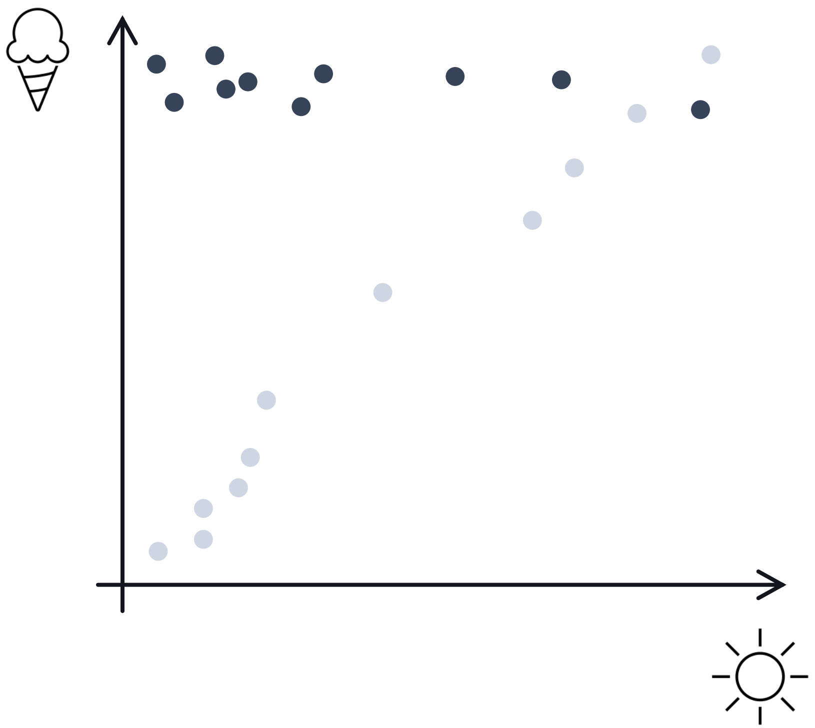
[Recent review: B. Schölkopf et al, "Towards causal representation learning", IEEE Advances in Machine Learning and Deep Neural Networks 2021]

Background

Causality and identifiability

Causality





Semantically, causal models label relations between random variables as **cause-effect relations**

Functionally, causal models describe **probability distributions and how they change** under changing conditions

Structural causal models (SCMs)

- **SCM:**

Causal variables

Noise variables with base distribution $\varepsilon_i \sim p_i(\varepsilon_i)$

$$z_1 = f_1(\varepsilon_1) \qquad z_2 = f_2(\varepsilon_2; z_1)$$

Causal mechanisms

Causal parents in acyclic graph

- **Solution:**

$$z = s(\varepsilon) \;\Rightarrow\; z \sim p_z(z)$$

Solution function
(= successively applying
causal mechanisms)

Observational
distribution

- **Interventions:**

$$f_i(\varepsilon_i; z_{\text{parents}}) \to \tilde{f}_i(\varepsilon_i) \qquad \Rightarrow \; z \sim \tilde{p}_z^i(z)$$

New mechanism
(perfect intervention: no parents)

Interventional distribution

Identifiability

- An representation / SCM $\mathcal{M}$ is **identifiable** if

Any two model
(from some family)

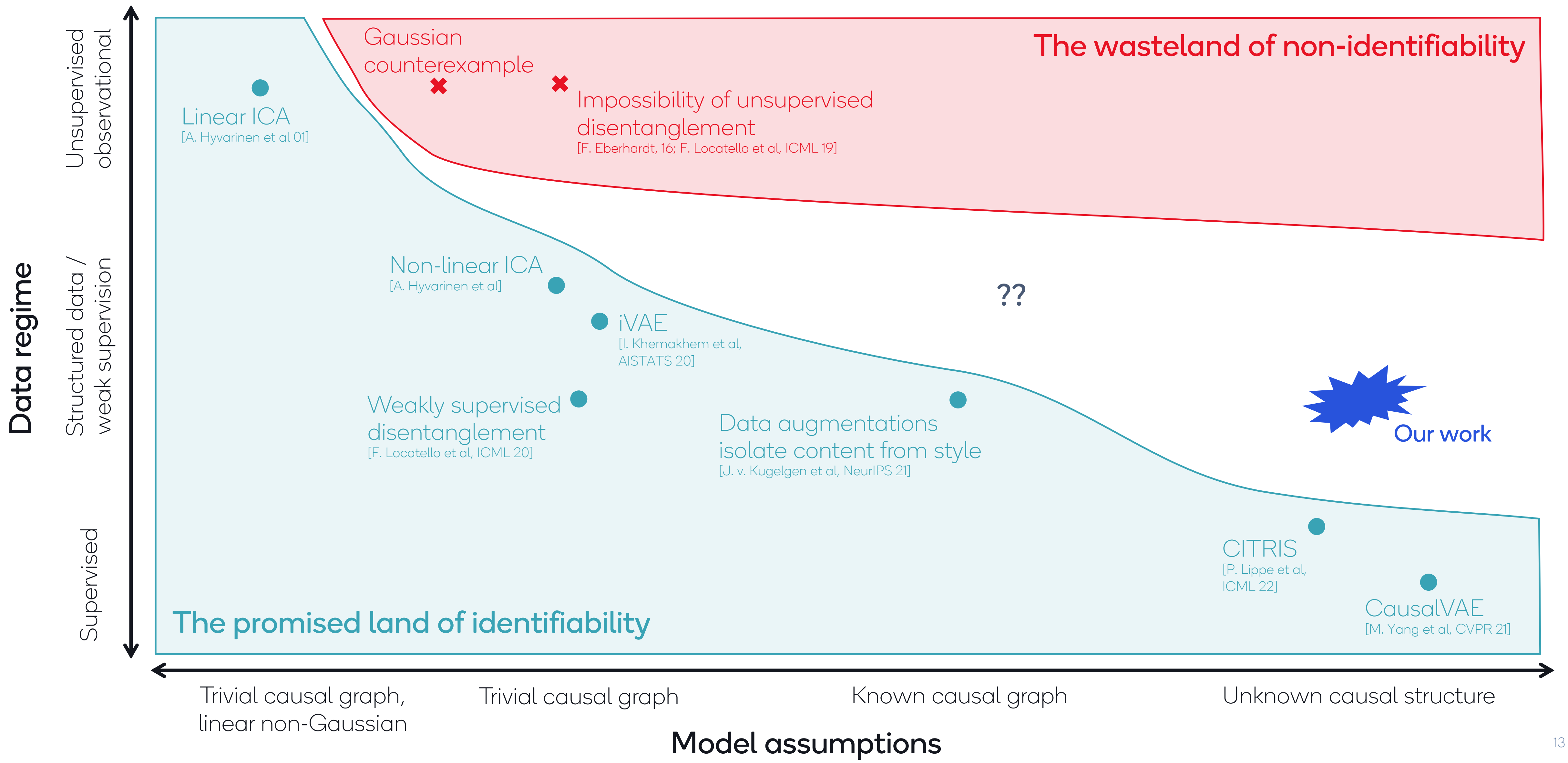
$$p_{\mathcal{M},x}(x) = p_{\mathcal{M}',x}(x) \implies \mathcal{M} \sim \mathcal{M}'$$

Data regime
(e.g. observational
distribution on pixel level)

Equivalence relation
(e.g. same up to
permutations)

- Identifiability means we can **find ground-truth causal structure** through maximum-likelihood training
  - if it is within the specified model family
  - up to the equivalence relation
  - in the limit of infinite data
  - assuming perfect training

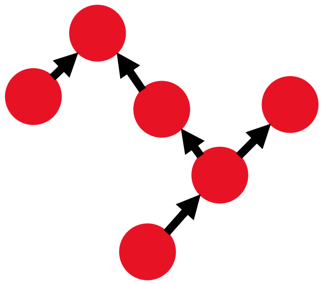
When are causal representations are identifiable?



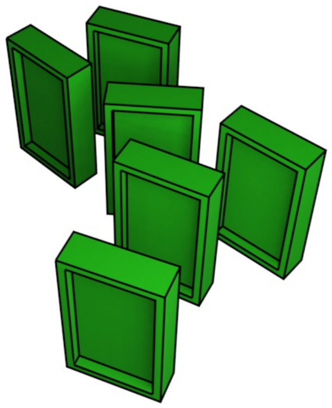
Theory

Causal representations can be identified from weak supervision

Latent causal model



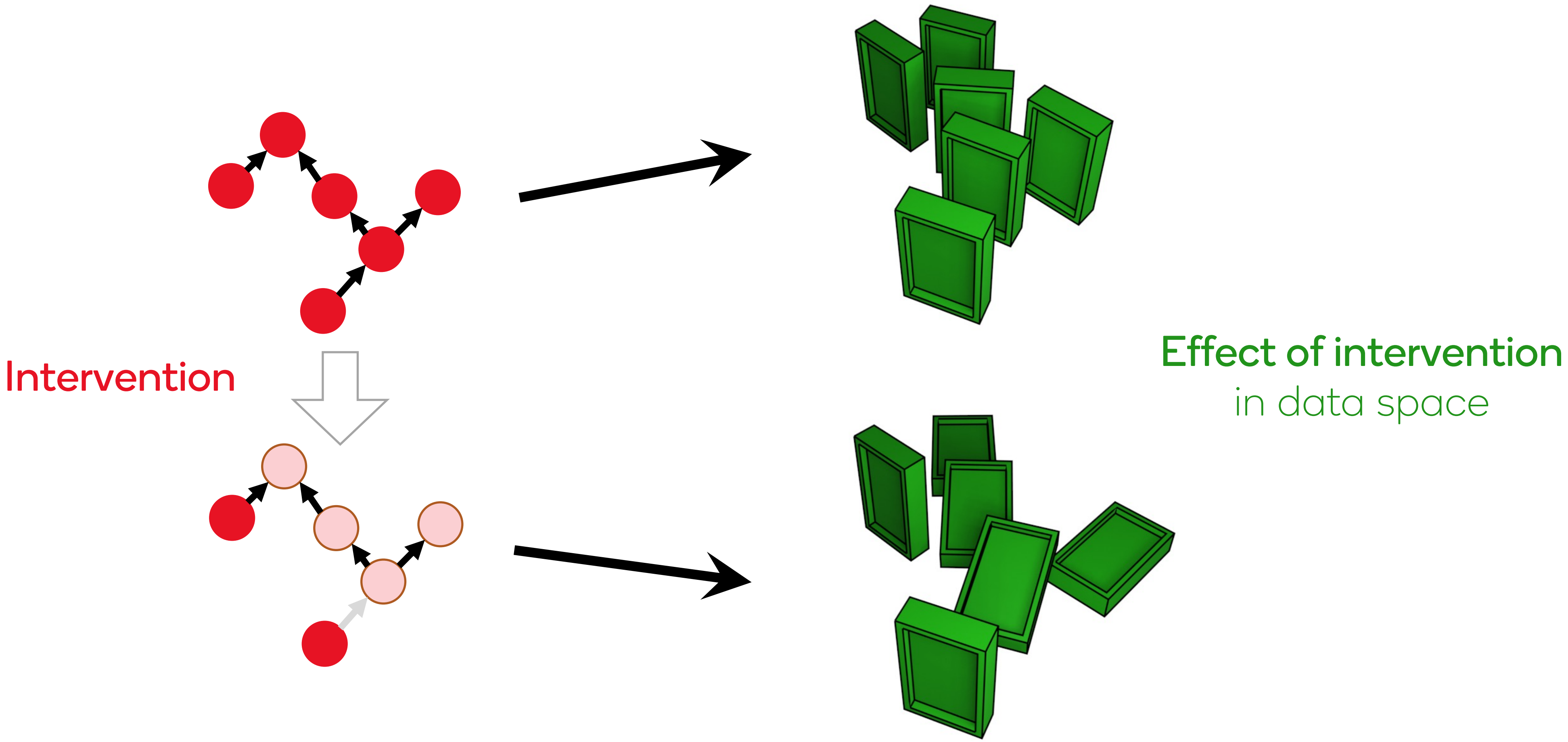
Decoder / rendering function



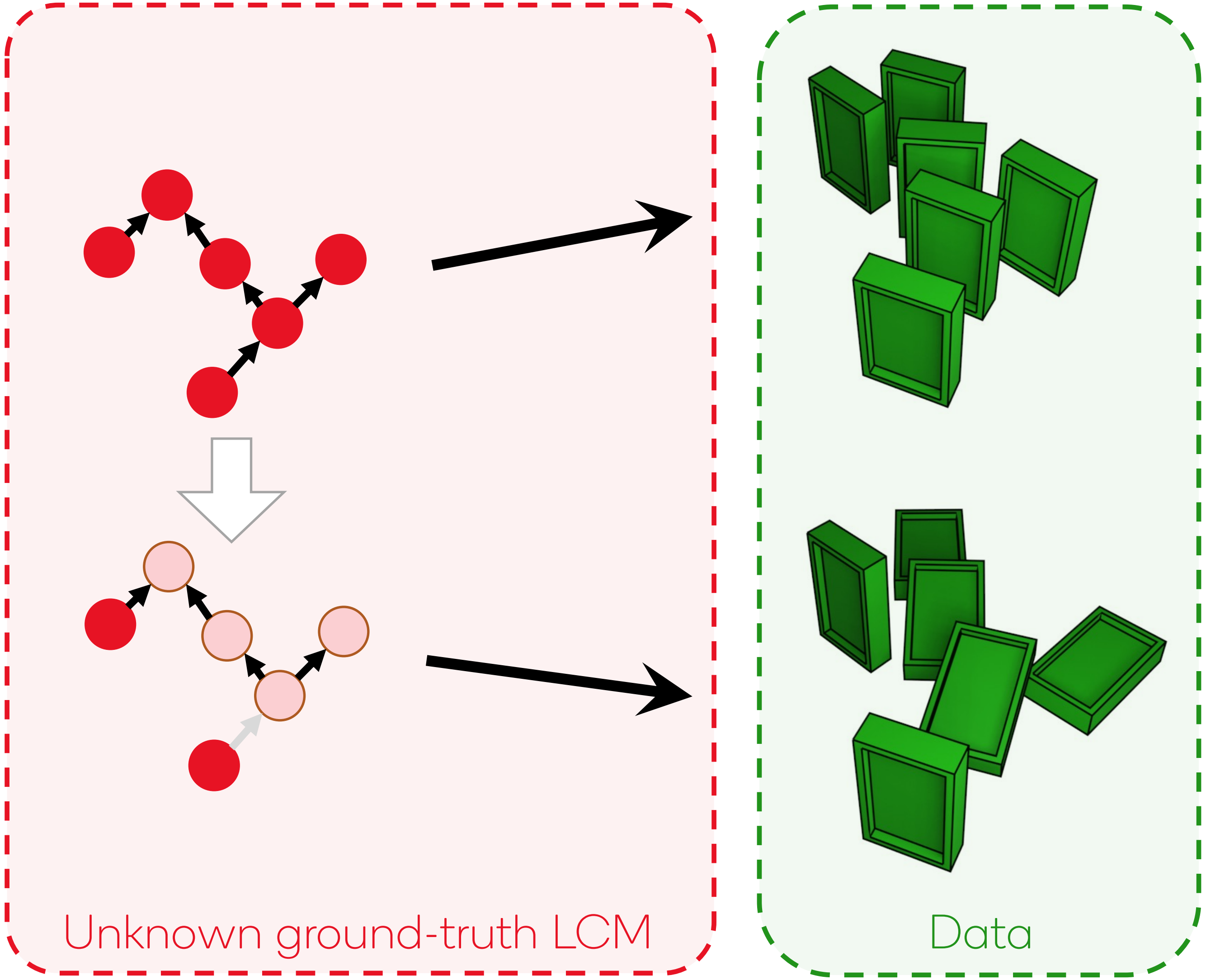
High-level variables with a structural causal model between them

Low-level data (pixels)

Interventions

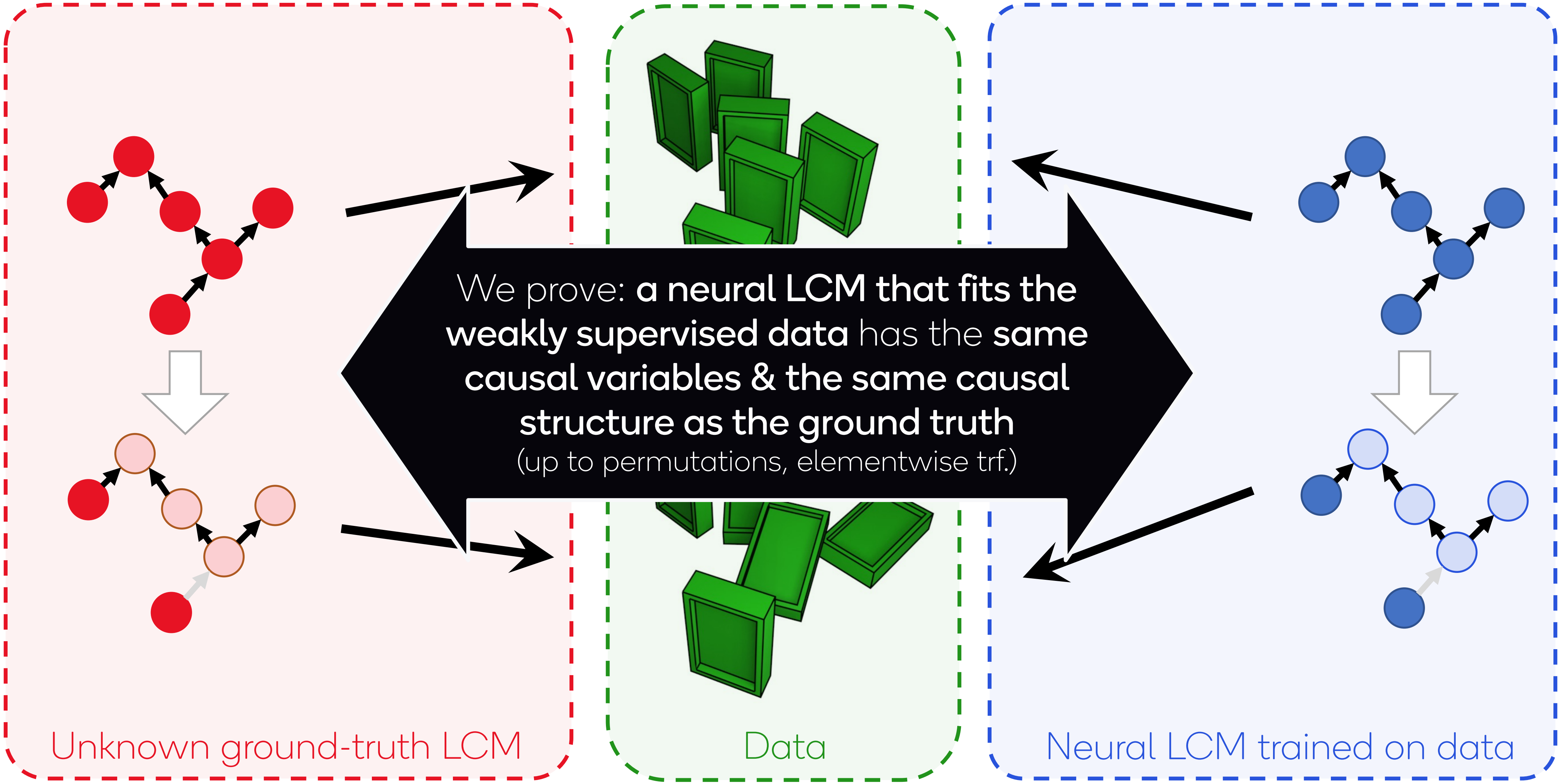


Weakly supervised data setting



- We assume access to **data pairs of the system before and after interventions**
  - Equivalent to counterfactuals
  - Causal abstraction of time-series data
- Otherwise, **no labels**
  - Only pixel-level data is observed
  - Intervention targets are unknown

Identifiability theorem



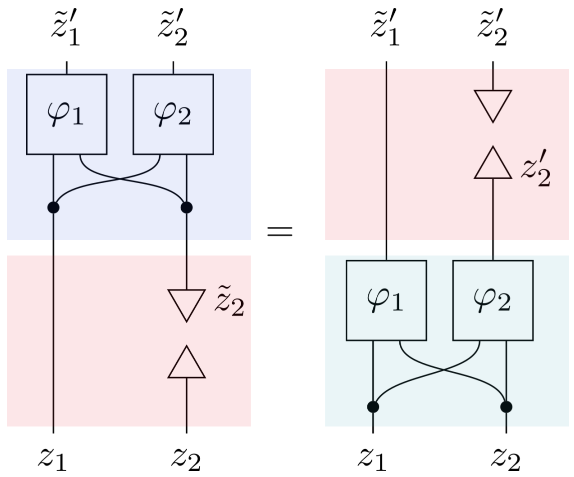
Proof sketch

1. Consider two LCMs with causal variables z and z', both matching the data.
   Define $\varphi : z \to z'$.

2. Interventions commute with φ:

$$\begin{tikzcd} z \arrow[r, "\text{intervene}"] \arrow[d, "\varphi"] & \tilde{z} \arrow[d, "\varphi"] \\ z' \arrow[r, "\text{intervene}"] & \tilde{z}' \end{tikzcd}$$

3. We assume perfect interventions. Then then $\tilde{z}_i'$ is independent of z_j . For 2 variables:



4. We assume $\mathbb{R}$-valued variables. Statistical independence then implies functional independence. Thus, $\varphi_i(z_i, z_j)$ must be constant in z_j.

5. Since this holds for any i, φ must be a permutation plus elementwise transformations.

6. Finally, we can show that the causal graphs and intervention targets in the two models are consistent with this transformation.

7. Thus the two models are isomorphic.

Assumptions

Assumption	Possible relaxation
Weakly supervised data is available	Maybe (work in progress)
Causal variables are $\mathbb{R}$-valued	Maybe (work in progress)
Causal mechanisms are diffeomorphic	Difficult
No hidden confounders	Difficult
Decoder is deterministic	Plausible (as in iVAE)
Interventions are perfect (Post-intervention values of intervention targets are independent of pre-intervention state)	Difficult (counterexamples)
Interventions are complete (The dataset contains interventions on any single causal variable)	Relaxation to n-target interventions plausible (incomplete interventions $\rightarrow$ partial identifiability)

Practice

Implicit is better than explicit

Explicit and implicit representations of causal structure

Explicit representation
through graph & causal mechanisms:

Causal variables

Noise variables with base distribution $\varepsilon_i \sim p_i(\varepsilon_i)$

$$z_i = f_i(\varepsilon_1; z_{\text{parents}})$$

Causal mechanisms

Causal parents in acyclic graph

Implicit representation
through solution function:

Causal variables

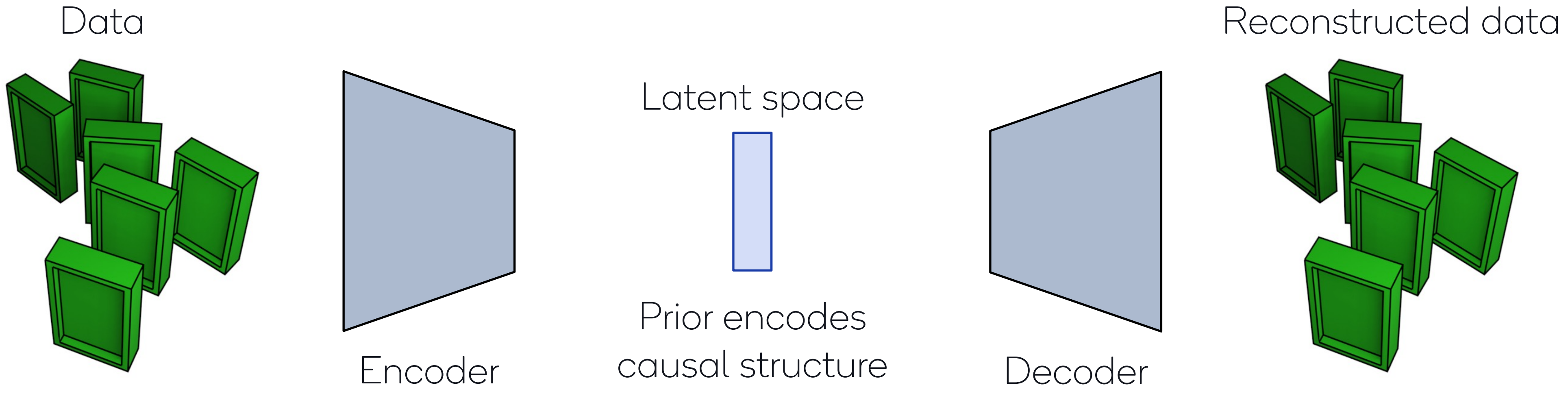
Noise variables with base distribution $\varepsilon_i \sim p_i(\varepsilon_i)$

$$z = s(\varepsilon)$$

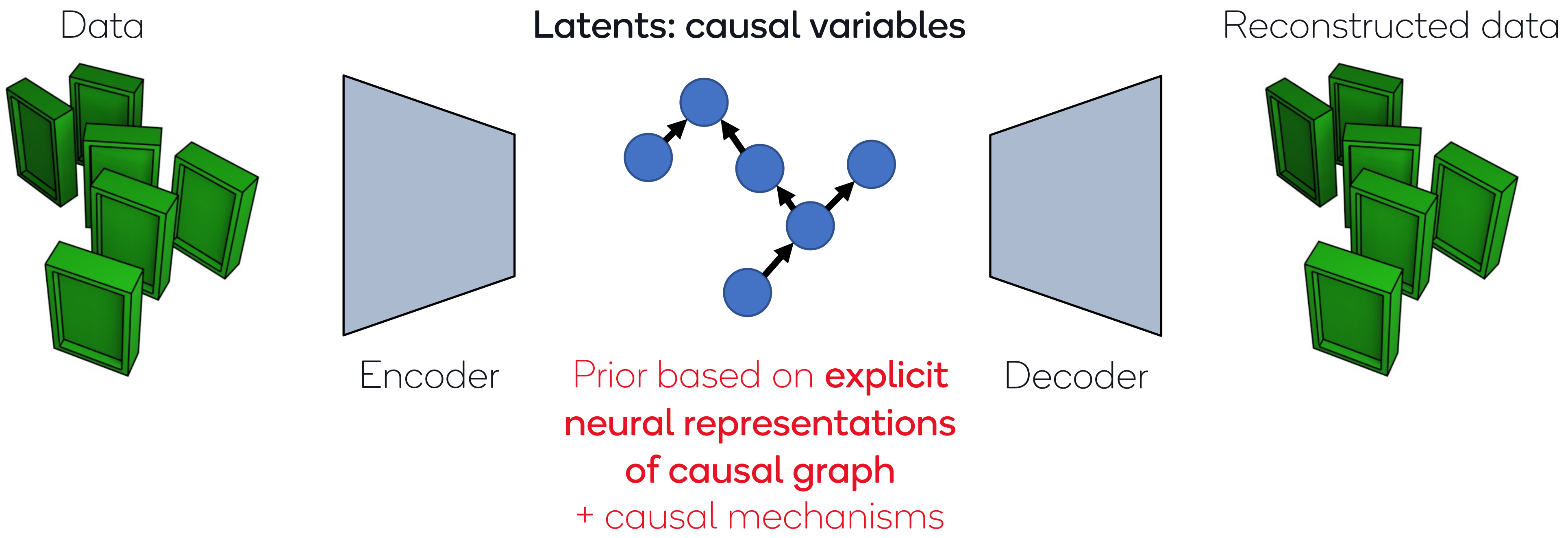
Solution function (= successively applying causal mechanisms)

Under our assumptions, explicit and implicit representation **contain the same information**

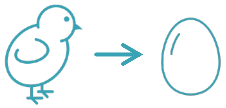
Operationalizing latent causal models



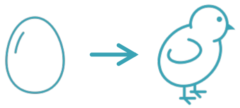
Explicit latent causal models



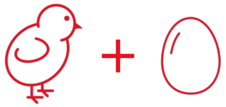
Explicit latent causal models in practice



Easy to learn graph given representations



Easy to learn representations given graph

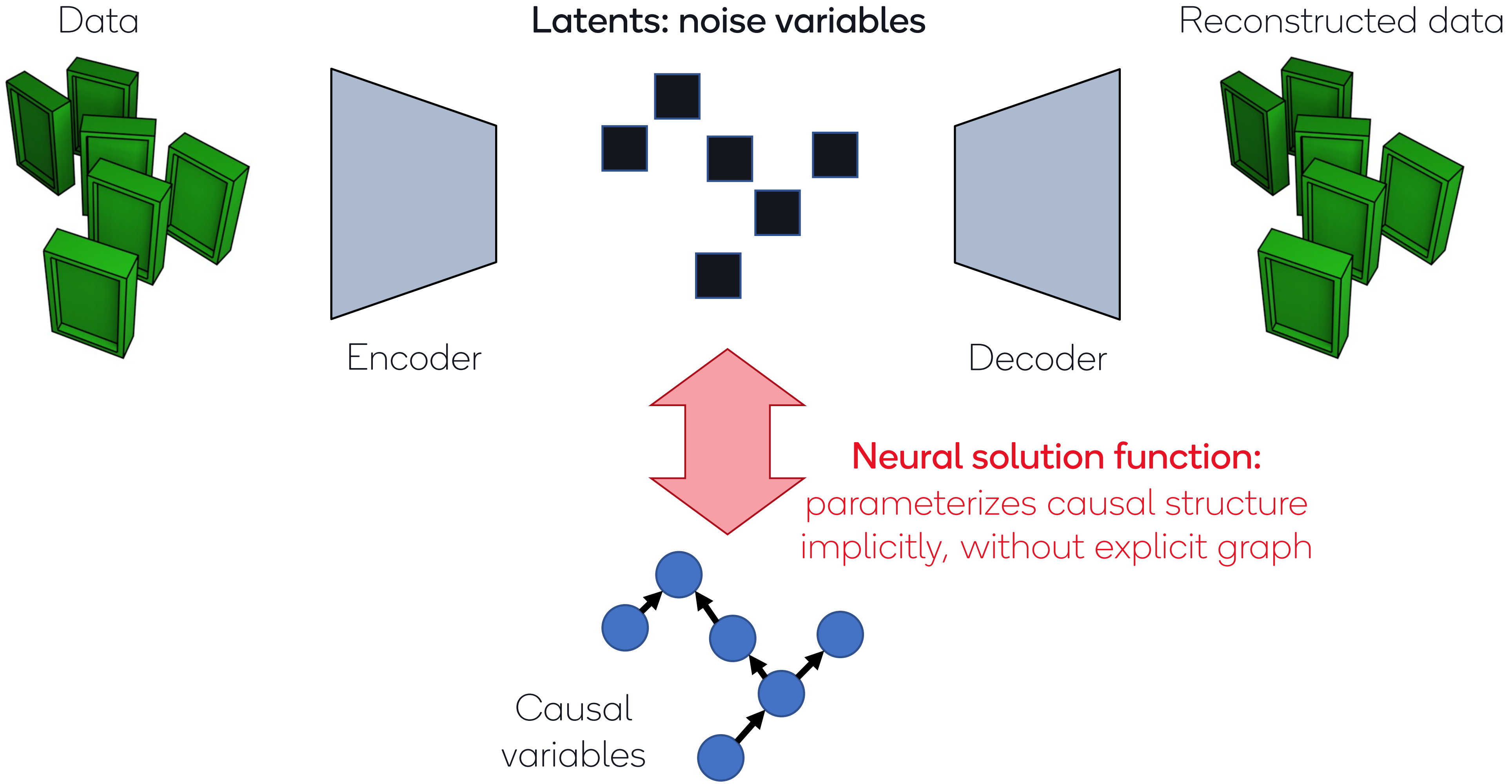


Difficult to learn graph and representation simultaneously

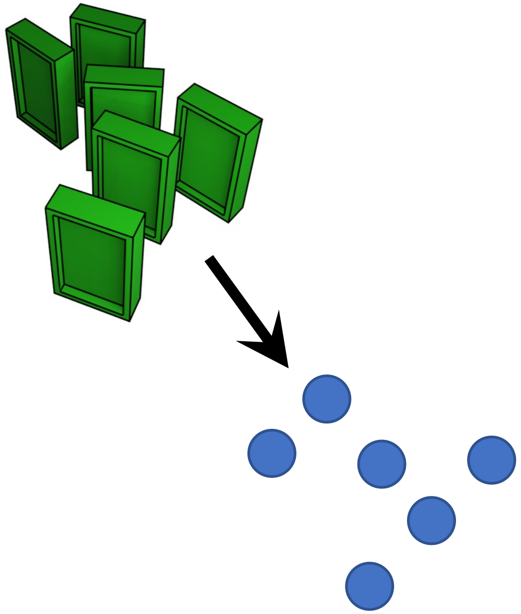
(Evidence for **local minima** in the loss landscape corresponding to wrongly oriented graph edges)

⇒ don't learn explicit graphs if you don't have to

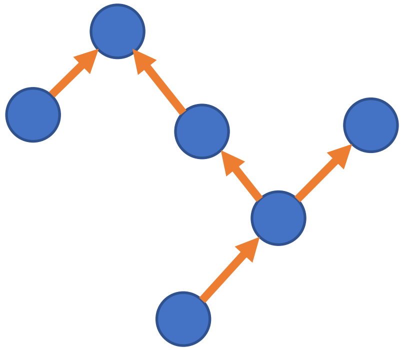
Implicit latent causal models



What can you do with ILCMs?

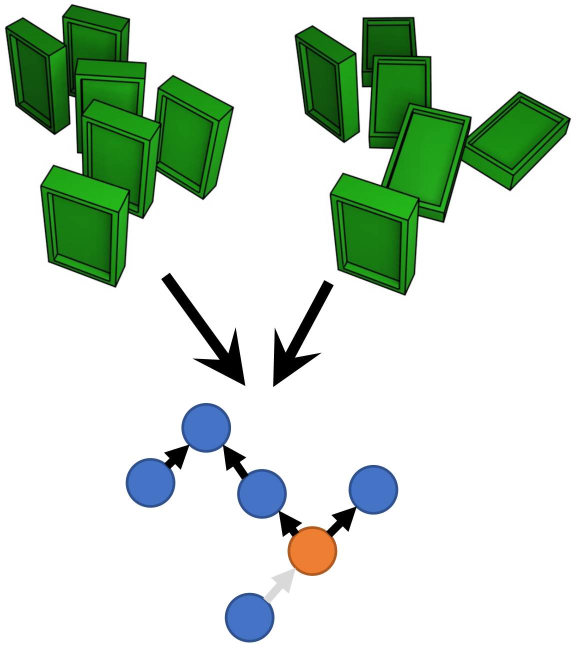


Map pixels to causal variables

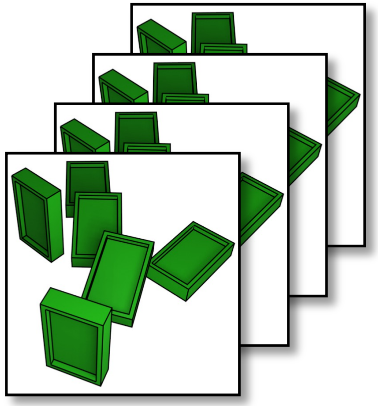


Find the causal graph

- ILCM-E: with off-the-shelf causal discovery algorithm ENCO
- ILCM-H: with our new heuristic



Infer interventions from data pairs



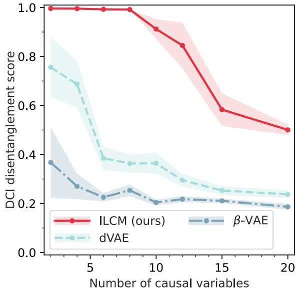
Generate observational, interventional, and counterfactual data

Experiments

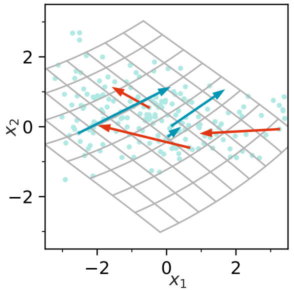
Things work, mostly

Experiments

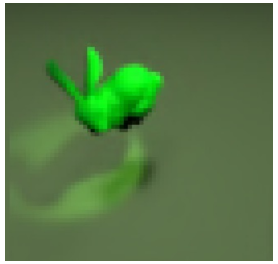
Complexity of causal system



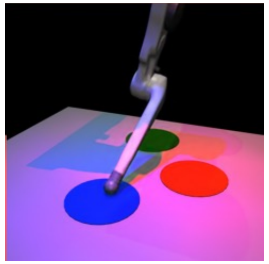
Scaling experiment



2D toy example



Causal3DIdent



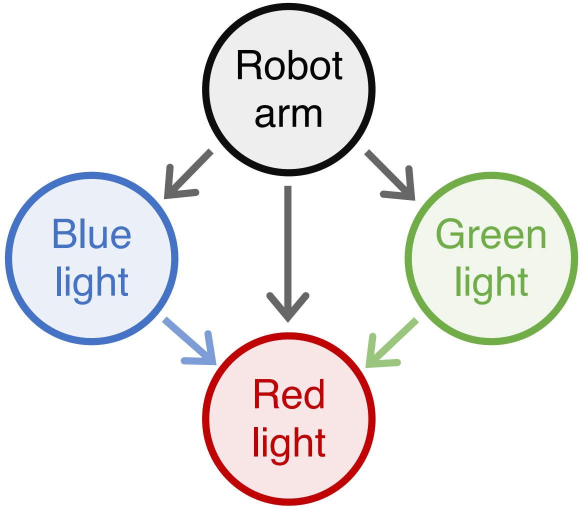
CausalCircuit

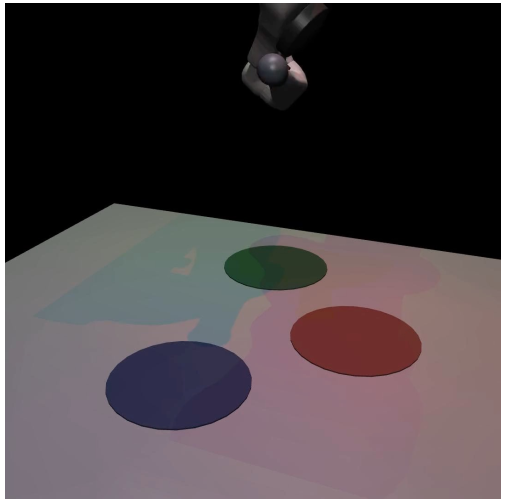
Data dimension

[Causal3DIdent: J. von Kügelgen et al, "Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style", NeurIPS 2021]

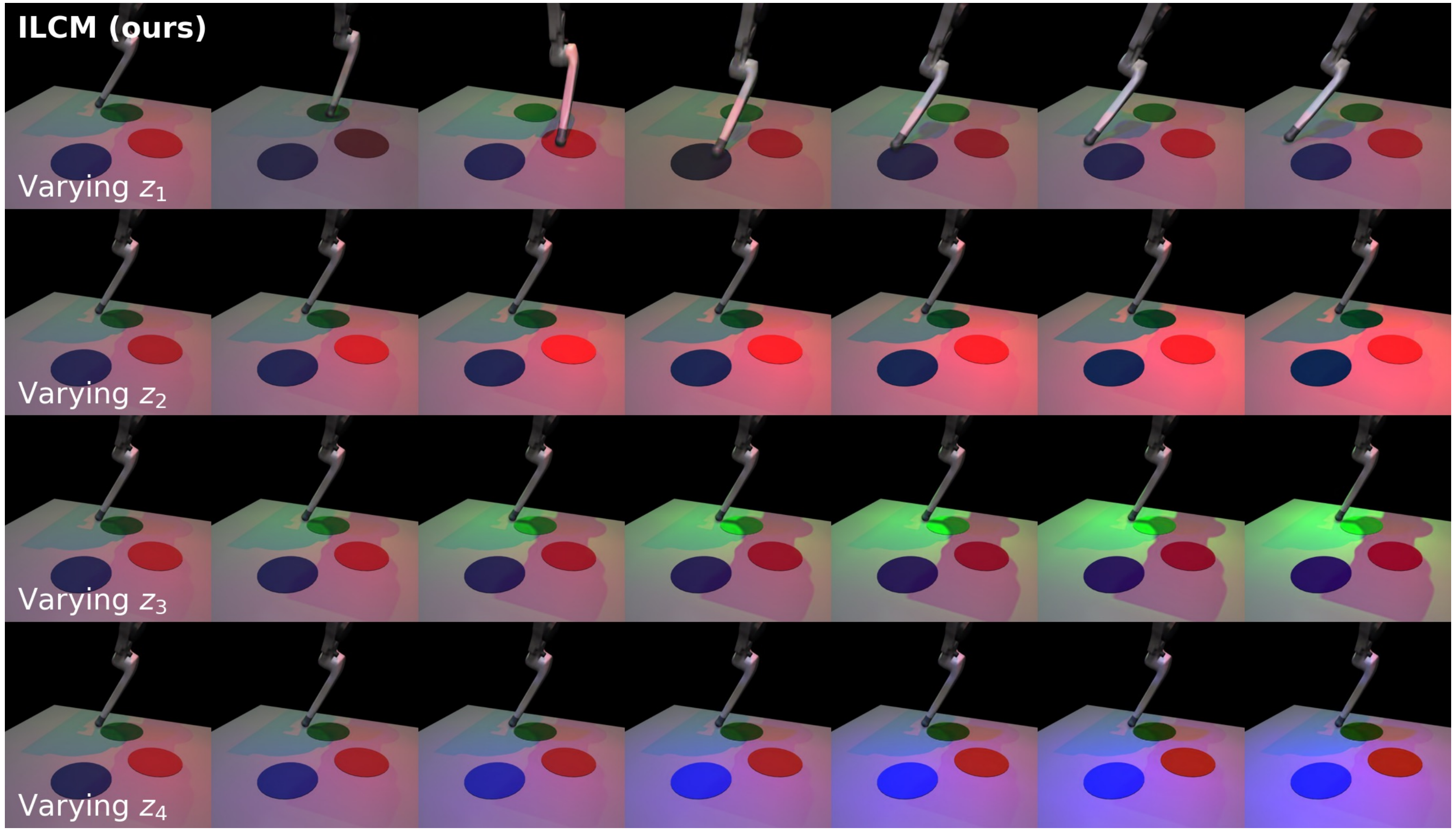
CausalCircuit

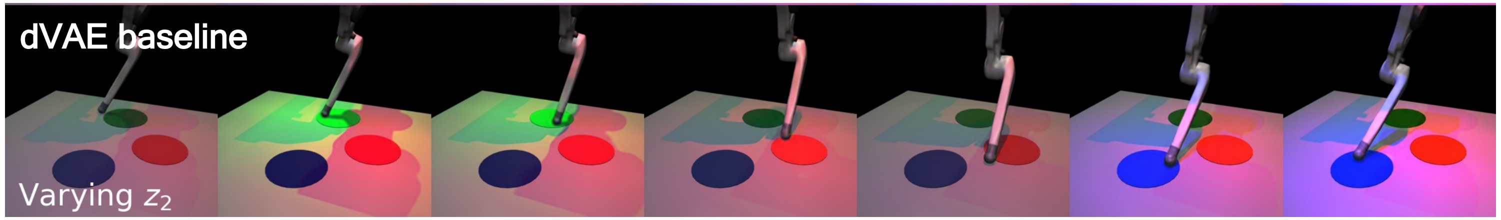
- **New dataset** with more intuitive causal structure
- **Robot arm interacts with touch-sensitive lights, which are connected with a circuit**
  - Robot arm movement based on inverse kinematic model
  - Physics + rendering with MuJoCo
  - 4 continuous causal variables: robot arm restricted to 1D arc + 3 light states
  - 512x512 images from fixed camera position
- ILCMs are trained on pre- and post-intervention data



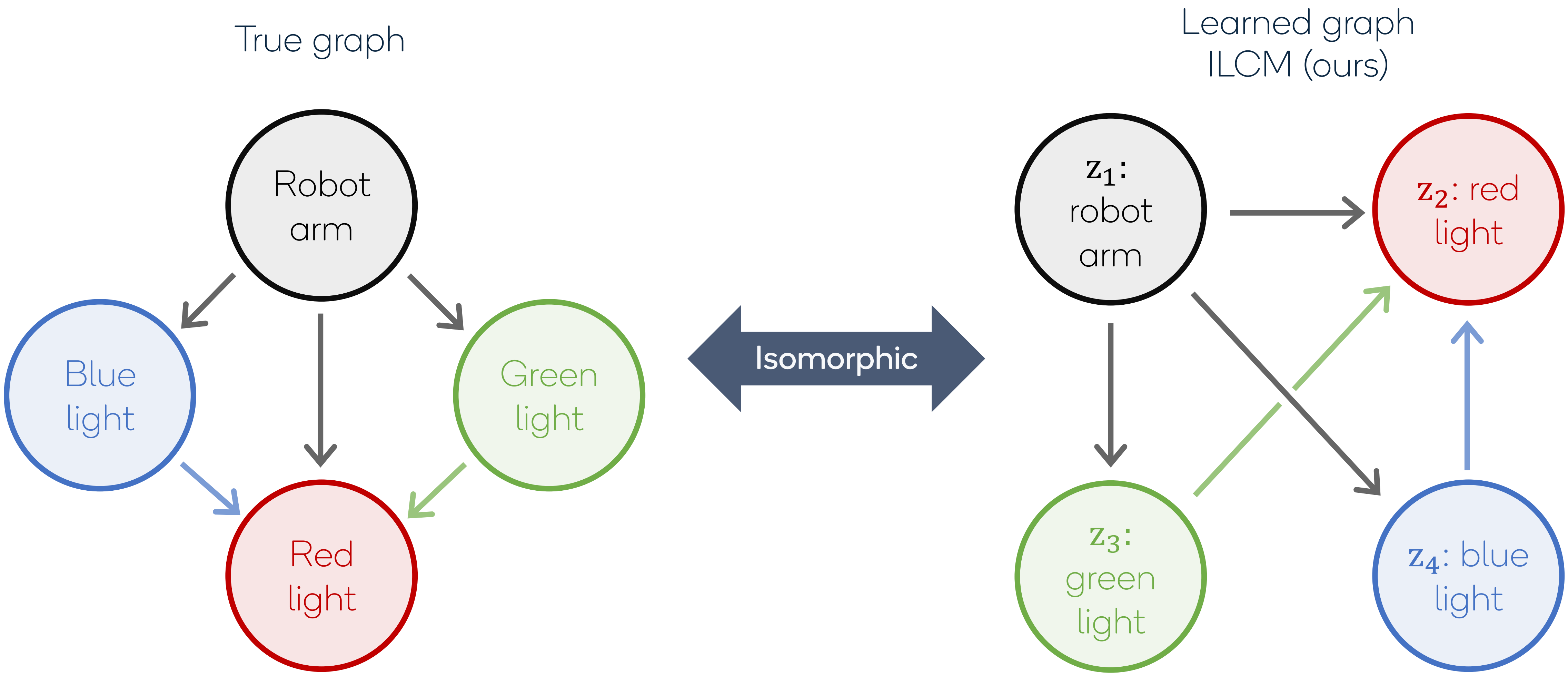


LCMs **disentangle** the causal variables



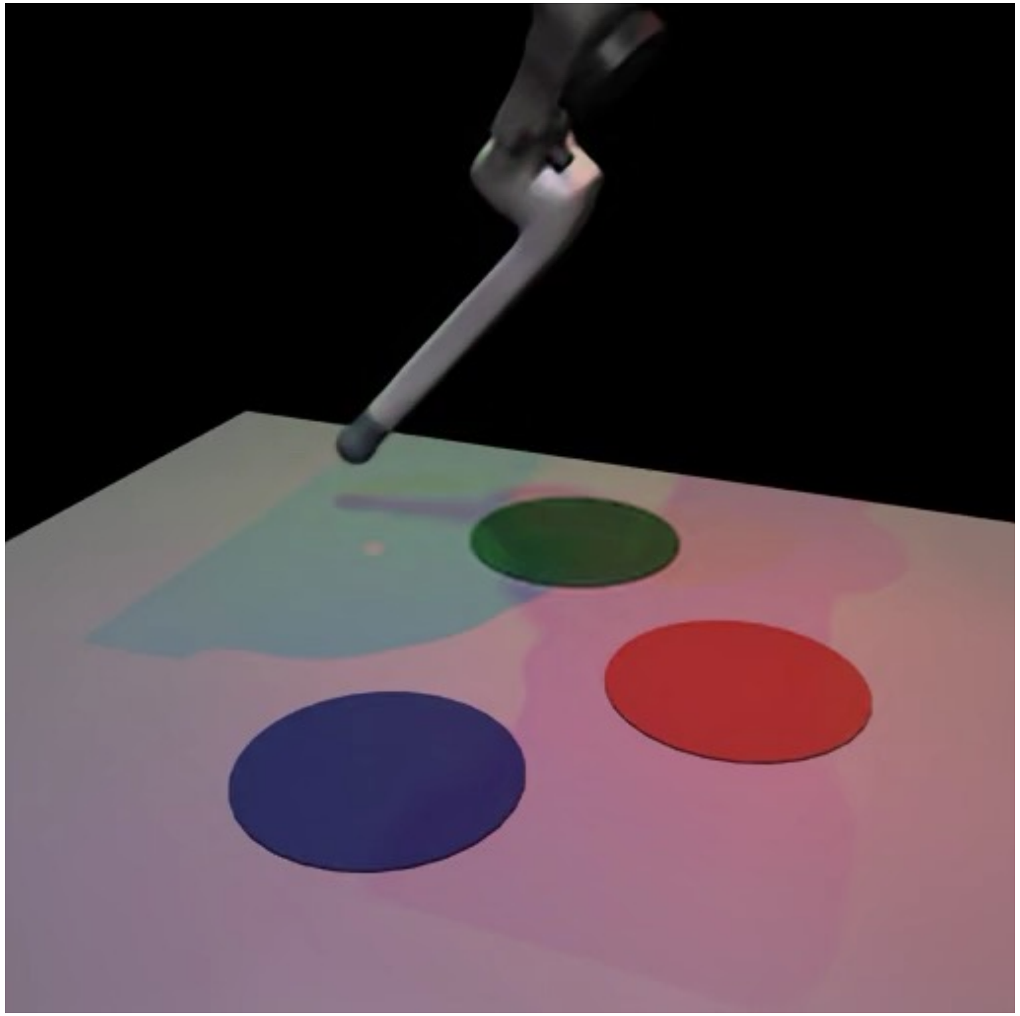


LCMs learn the **correct graph**



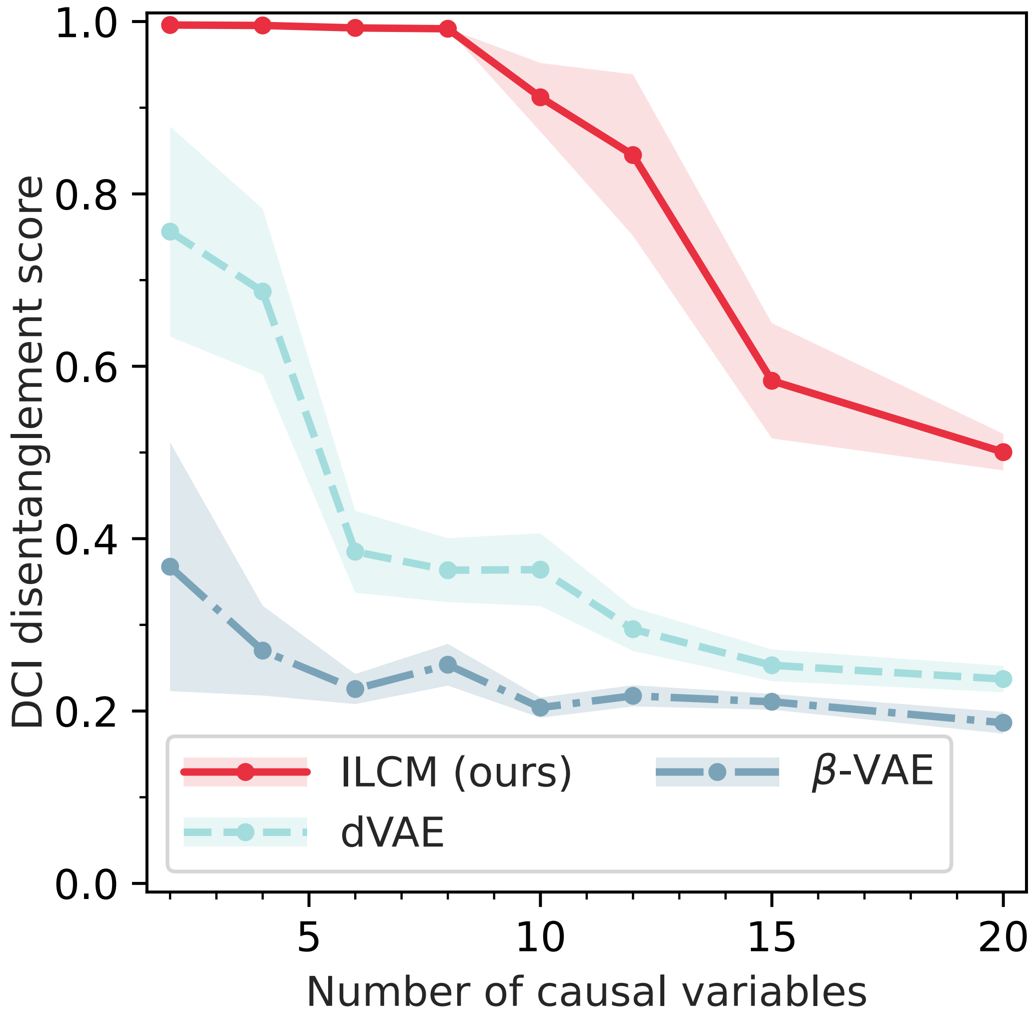
ILCMs let us **reason causally**

ILCM samples, **intervening** on a single latent (including causal effects)



Do ILCMs **scale**?

- **Toy experiment**:
  - n causal variables
  - linear causal effects
  - SO(n) decoder
- ILCM results **robust up to ~10 variables** without additional tuning

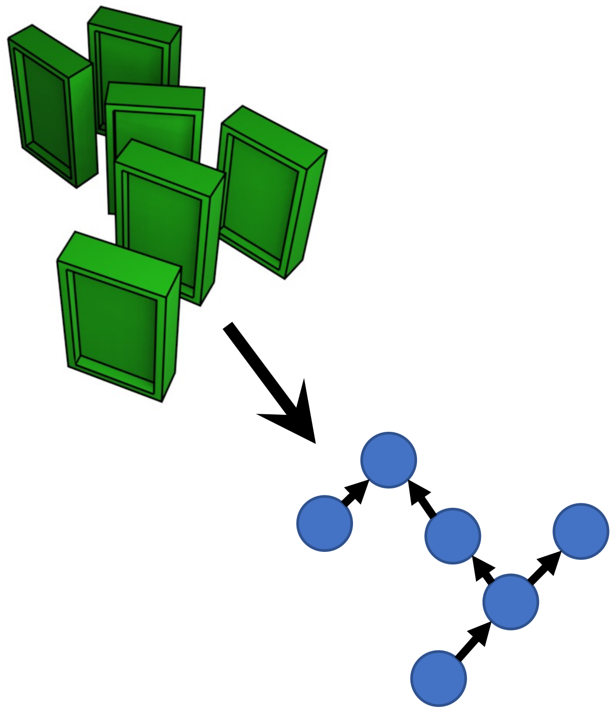


Outlook

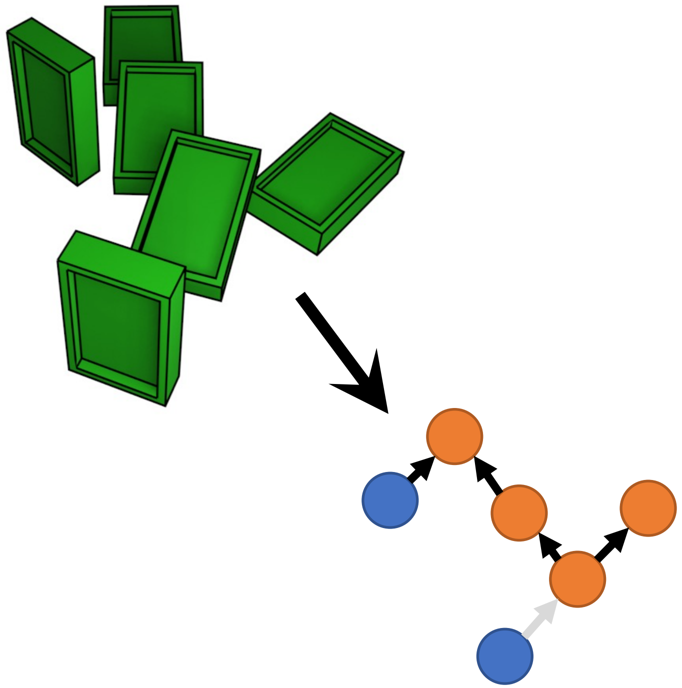
Towards useful causal representation learning

A long way to go

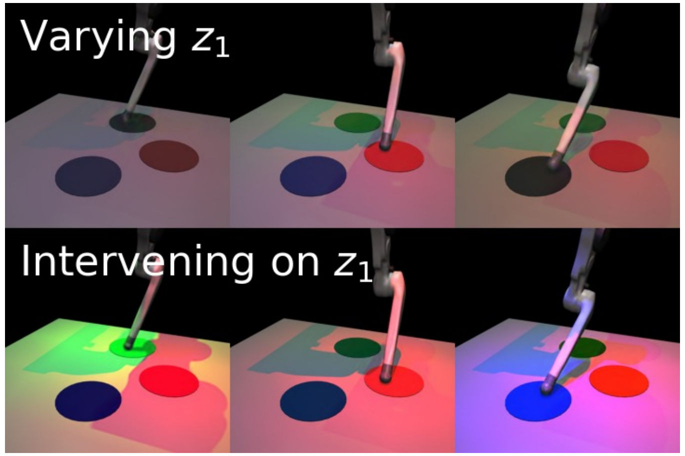
Where we are	Where we need to get
Identifiability theorems	**Demonstrate usefulness** on downstream tasks
Pre- & post-intervention data	**Realistic data regimes:** observational & interventional data, video data, ...
God-given interventions	**Learning intervention policies**
Fixed causal variables	**Variable scene composition**
Strict **DAG-based causality**	**Weaker relational structures**
Toy experiments (up to O(10) variables)	**Realistic experiments**



Can we **learn causal variables & causal structure from pixels**, without labels?



We prove: this is possible with **weak supervision**, when observing effects of interventions



In practice, **implicit latent causal models** can identify the causal structure in image datasets

Weakly supervised causal representation learning

JB, Pim de Haan, Phillip Lippe, Taco Cohen
NeurIPS 2022
arXiv:2203.16437


Pim de Haan


Phillip Lippe


Taco Cohen

Towards causal representation learning
Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, Yoshua Bengio
IEEE Advances in Machine Learning and Deep Neural Networks 2021
arXiv:2102.11107

Weakly-supervised disentanglement without compromises
Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, Michael Tschannen
ICML 2020
arXiv: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
NeurIPS 2021
arXiv:2106.04619

CITRIS: Causal identifiability from temporal intervened sequences
Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, Efstratios Gavves
ICML 2022
arXiv:2202.03169

Thank you

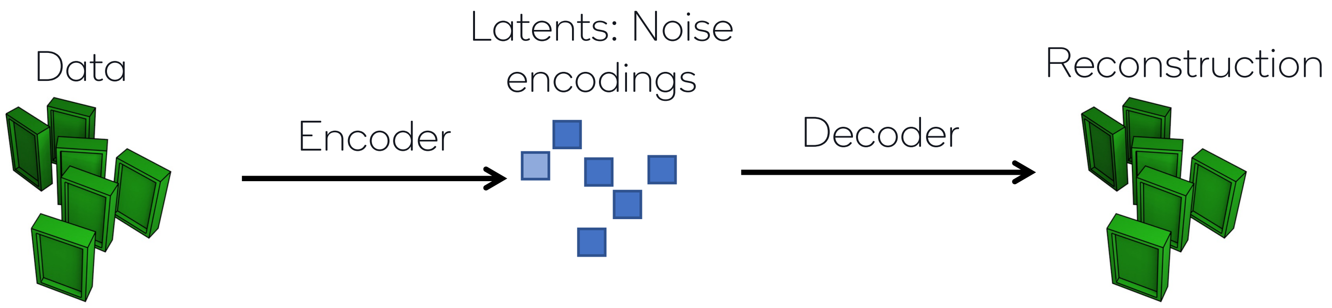
Follow us on:

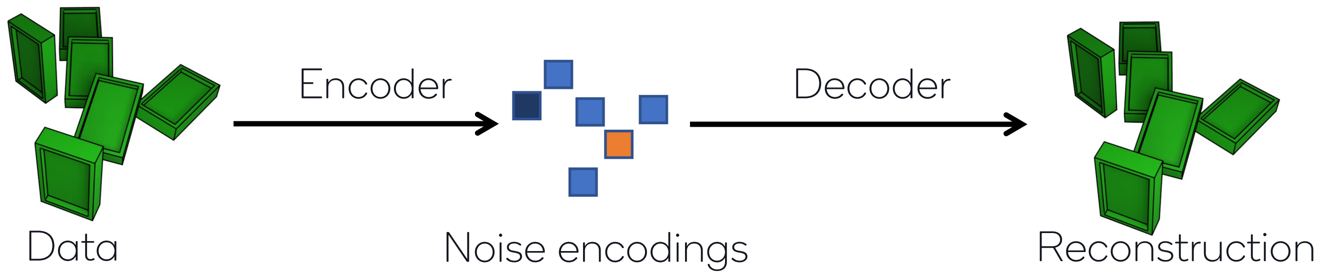
For more information, visit us at:

qualcomm.com & qualcomm.com/blog

Implicit LCMs (ILCMs)

VAE with noise encoding latents:





- Latent variables: **noise encodings**

causal variables

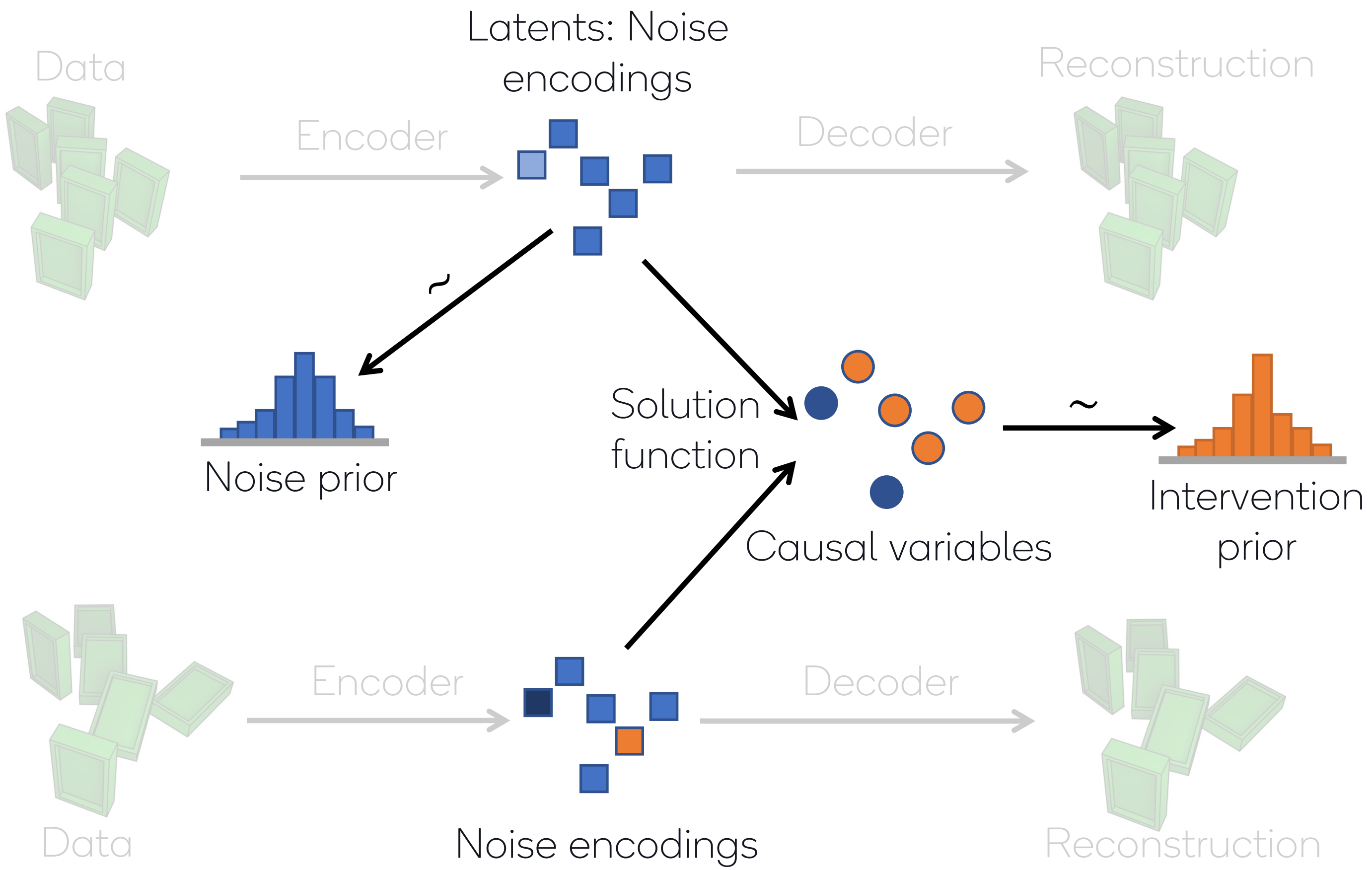
$$e = s^{-1}(z)$$

solution function: map between noise variables and causal variables in un-intervened SCM

- Convenient property: distribution factorizes in a way that **does not require the causal graph**

Implicit LCMs (ILCMs)

VAE with noise encoding latents:



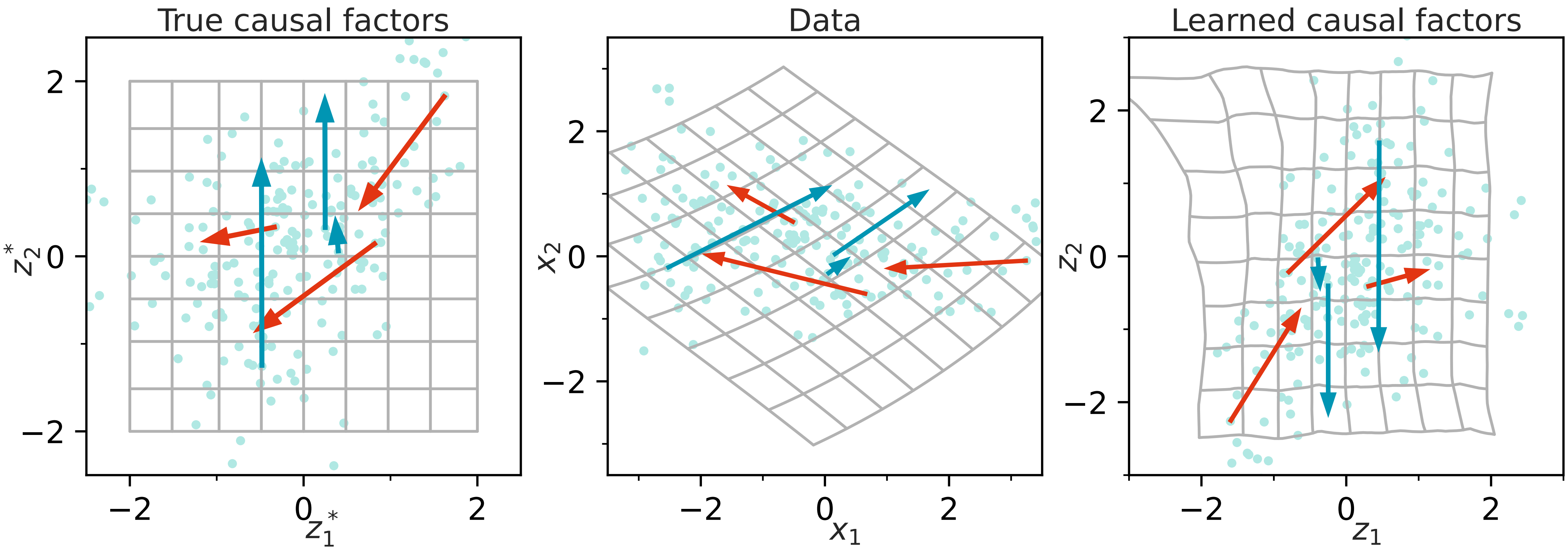
- **Prior encodes causal structure implicitly**
  - Pre-intervention: iid noise prior
  - Post-intervention: **learnable solution function** transforms noise to causal variables
- Encoder, decoder & solution function are learned end to end
- **No need for explicit graph parameterization!**
  - Circumvents optimization challenges

Experiment results

Dataset	True graph	Method	D	C	I	Int. accuracy	Learned graph	SHD
2D toy data		ILCM-E (ours)	**0.99**	**0.99**	**0.00**	**0.96**		**0**
		ILCM-H (ours)	**0.99**	**0.99**	**0.00**	**0.96**		**0**
		dVAE	0.35	0.50	0.01	**0.96**		1
		β-VAE	0.52	0.53	**0.00**	–	–	–
CausalCircuit		ILCM-E (ours)	**0.97**	**0.97**	**0.00**	**1.00**		**0**
		ILCM-H (ours)	**0.97**	**0.97**	**0.00**	**1.00**		**0**
		dVAE-E	0.34	0.55	**0.00**	**1.00**		5
		β-VAE	0.39	0.43	**0.00**	–	–	–
		Slot attention	0.39	0.82	**0.00**	–	–	–

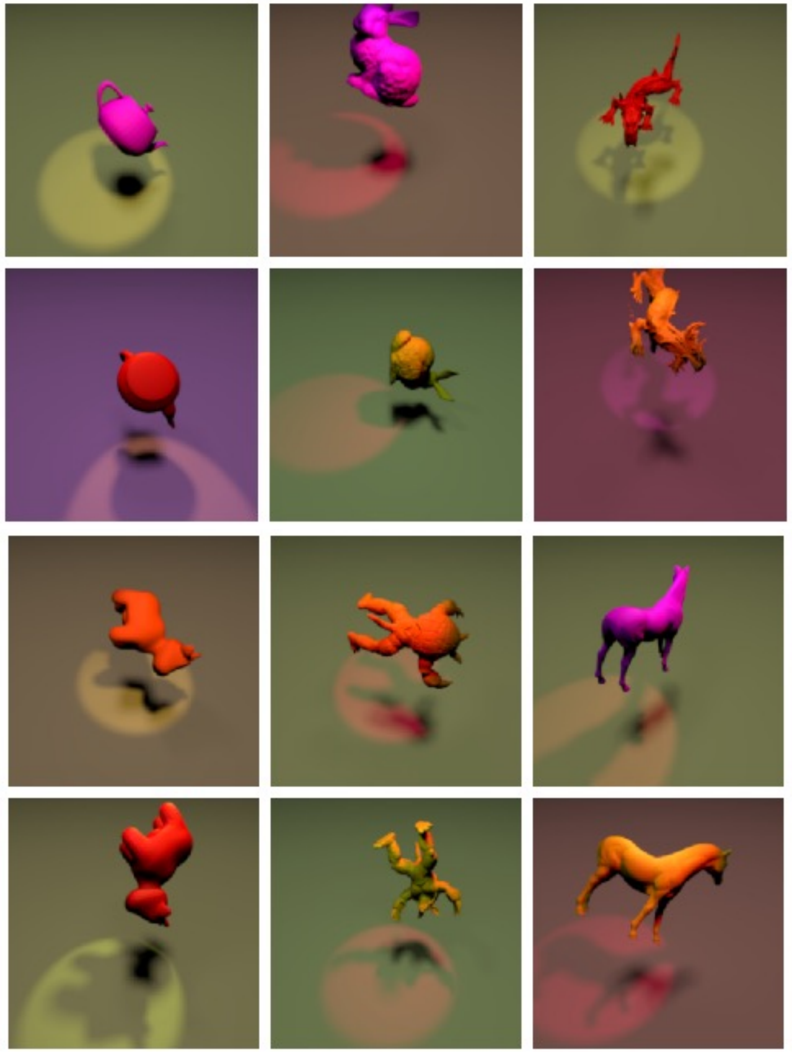
Dataset	True graph	Method	D	C	I	Int. accuracy	Learned graph	SHD
Causal3DIdent		ILCM-E (ours)	0.99	0.99	**0.00**	**0.98**		**0**
		ILCM-H (ours)	0.99	0.99	**0.00**	**0.98**		**0**
		dVAE	**1.00**	**1.00**	**0.00**	**0.98**		**0**
		β-VAE	0.94	0.94	**0.00**	–	–	–
		Slot attention	0.90	0.90	0.01	–	–	–
		ILCM-E (ours)	**1.00**	**1.00**	**0.00**	**0.98**		**0**
		ILCM-H (ours)	**1.00**	**1.00**	**0.00**	**0.98**		**0**
		dVAE	0.91	0.91	**0.00**	**0.98**		1
		β-VAE	0.92	0.92	**0.00**	–	–	–
		Slot attention	0.56	0.84	0.02		–	–
		ILCM-E (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		ILCM-H (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		dVAE	0.83	0.83	**0.00**	**0.98**		2
		β-VAE	0.63	0.71	**0.00**	–	–	–
		Slot attention	0.42	0.59	0.02	–	–	–
		ILCM-E (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		ILCM-H (ours)	**0.99**	**0.99**	**0.00**	**0.98**		1
		dVAE	0.79	0.81	**0.00**	**0.98**		2
		β-VAE	0.63	0.68	0.01	–	–	–
		Slot attention	0.87	0.87	0.03	–	–	–
		ILCM-E (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		ILCM-H (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		dVAE	0.80	0.81	0.01	**0.98**		2
		β-VAE	0.28	0.52	0.16	–	–	–
		Slot attention	0.32	0.35	0.04	–	–	–
		ILCM-E (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		ILCM-H (ours)	**0.99**	**0.99**	**0.00**	**0.98**		**0**
		dVAE	0.60	0.64	**0.00**	**0.98**		3
		β-VAE	0.57	0.61	0.01	–	–	–
		Slot attention	0.53	0.67	0.01	–	–	–

2D toy experiment



Causal3DIdent

- Recently proposed **disentanglement benchmark**
  - 3D renderings of objects under various lighting conditions
- We construct **six datasets with different causal structures**
  - **3 causal factors**: object color, light color, light position
  - Each dataset has a different causal graph, random nonlinear causal mechanisms
  - **64x64 images**
- We train ILCMs on pre- and post-intervention data



Causal3DIdent: J. von Kügelgen et al, "Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style", NeurIPS 2021

Causal3DIdent disentanglement

LCMs disentangle the causal factors…

- mean disentanglement score: 0.99 (1 is optimal)

… better than acausal baselines

• disentanglement VAEs:	disentanglement score 0.82
• beta-VAEs:	disentanglement score 0.66
• slot attention:	disentanglement score 0.60