Weakly supervised causal representation learning

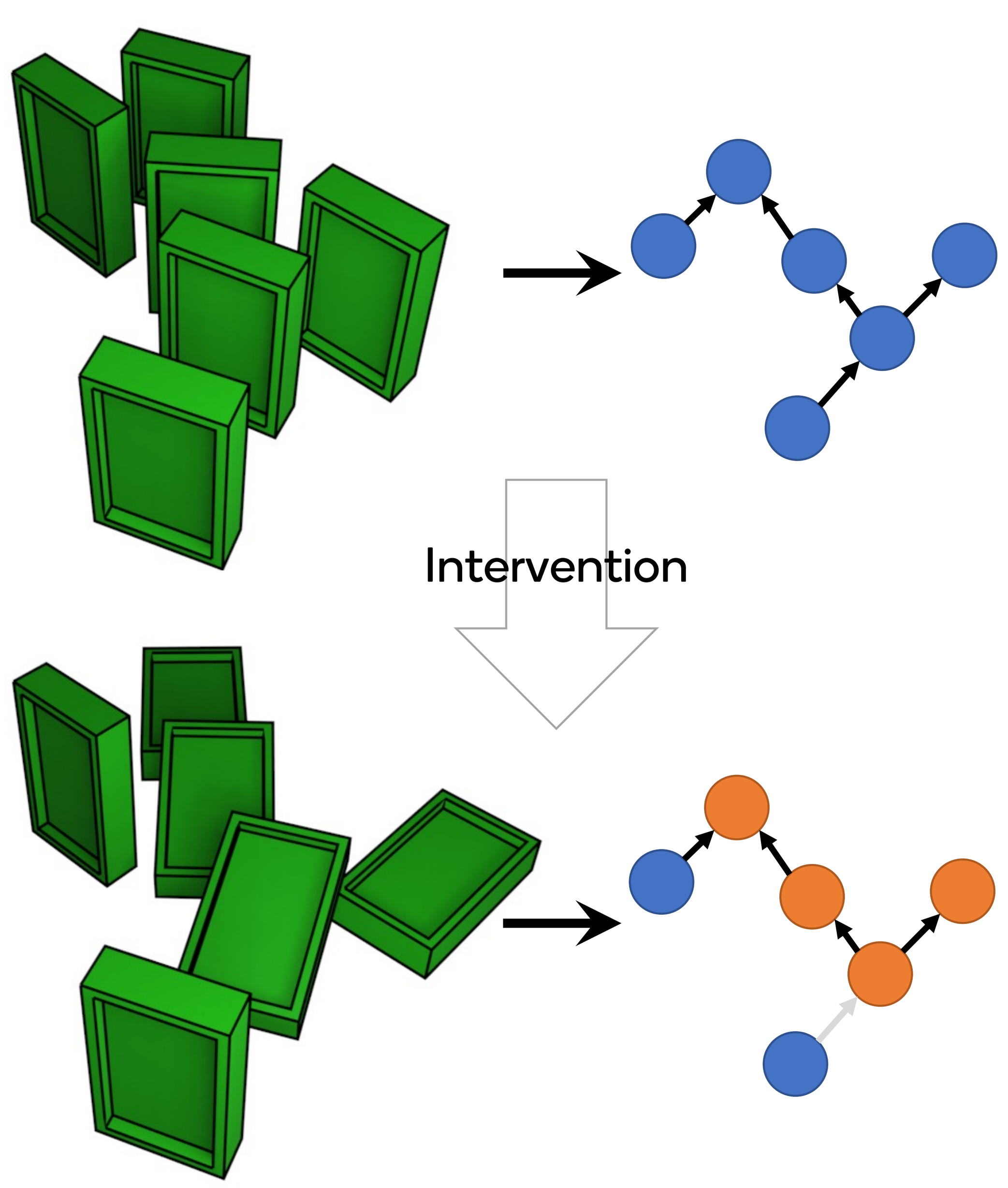
Johann Brehmer*[1], **Pim de Haan***[1,2], **Phillip Lippe**[2], and **Taco Cohen**[1]

*equal contribution [1]Qualcomm Technologies Netherlands B.V. [2]QUVA Lab, University of Amsterdam

Qualcomm AI research

Can we learn causal structure from pixels?

- Many systems can be described with high-level causal factors and causal relations between them, but only an unstructured low-level representation (like pixels) is observed
- Learning causal representations and causal structure from pixel data may be important for problems in robotics and autonomous driving [1]
- Unfortunately, this is impossible without supervision or prior assumptions [2,3]



Yes – with weak supervision

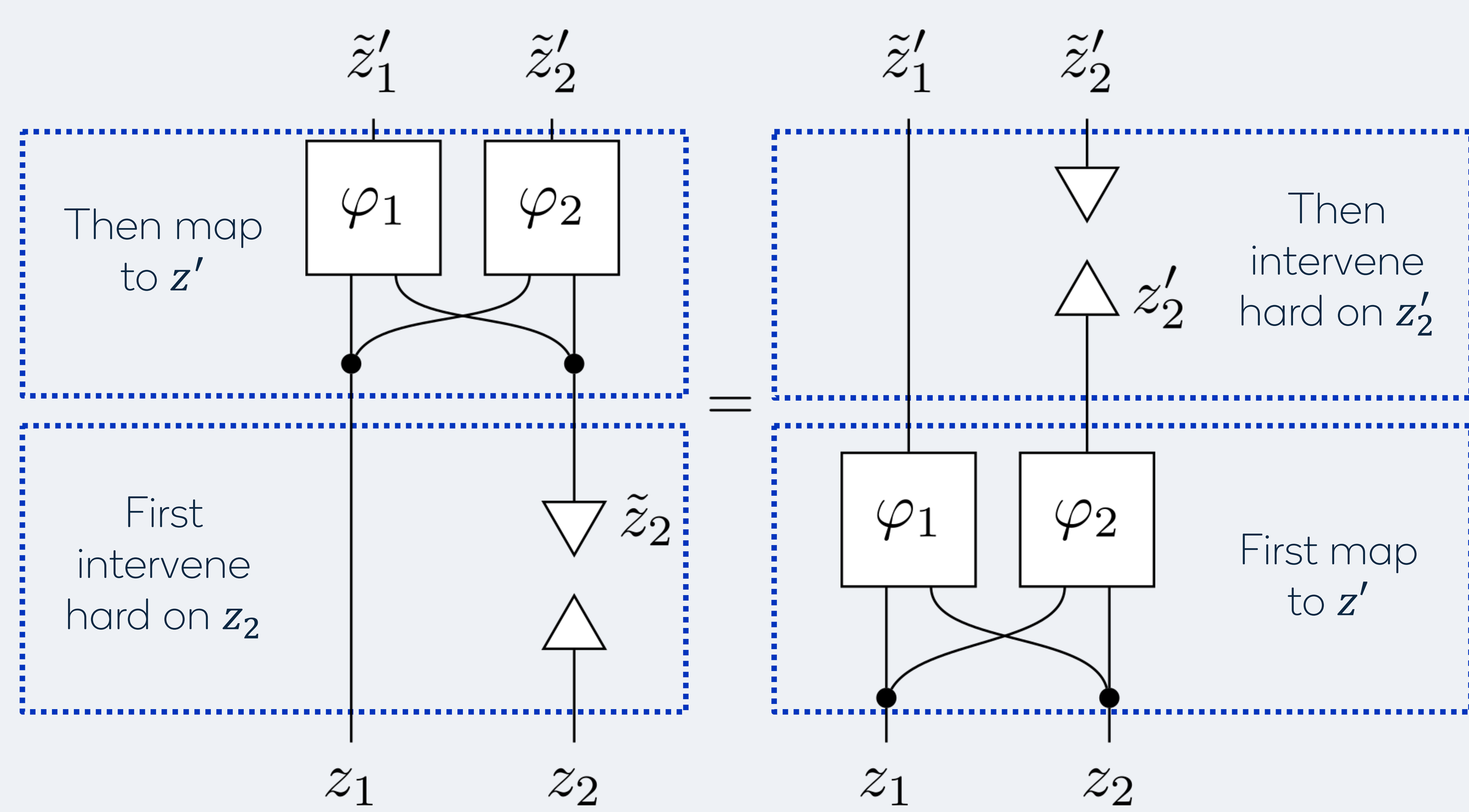
- We consider the setting in which we **observe a system before and after interventions**
- The interventions are random, not chosen by the algorithm
- **Neither labels on the causal variables nor on the intervention targets are required**
- A similar weakly supervised setting has been studied for independent factors of variations [4], we generalize that to arbitrary causal structures

Identifiability result: in the weakly supervised setting, causal variables, SCMs, and interventions are identifiable

- We define **latent causal models (LCMs)** as structural causal models (SCMs) and a diffeomorphic decoder from causal variables to a data space
- We prove: **if two LCMs have the same data distribution in the weakly supervised setting, they are identical** (up to a permutation of the causal variables and an elementwise reparameterization)
- Key assumptions:
  - Causal variables are $\mathbb{R}$-valued
  - Interventions are perfect: post-intervention values of intervention targets are independent of pre-intervention state
  - Interventions are complete: the dataset contains interventions on any single causal variable

Proof sketch

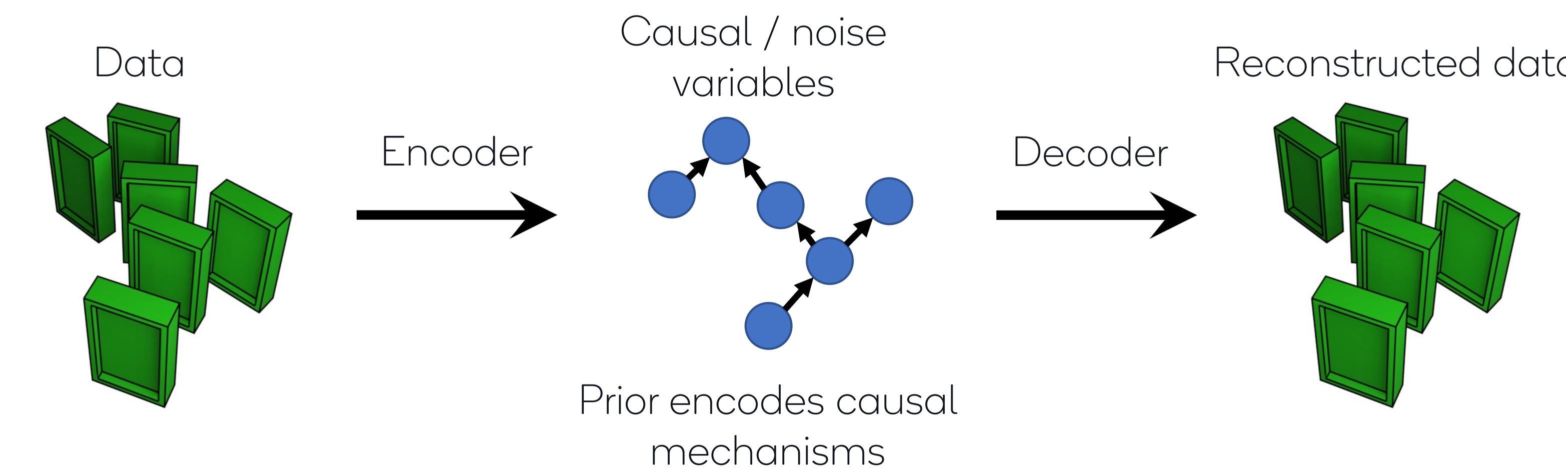
- Given two sets of causal variables z and z' both matching the data, and a map $\varphi : z \to z'$, we get the same conditional if we



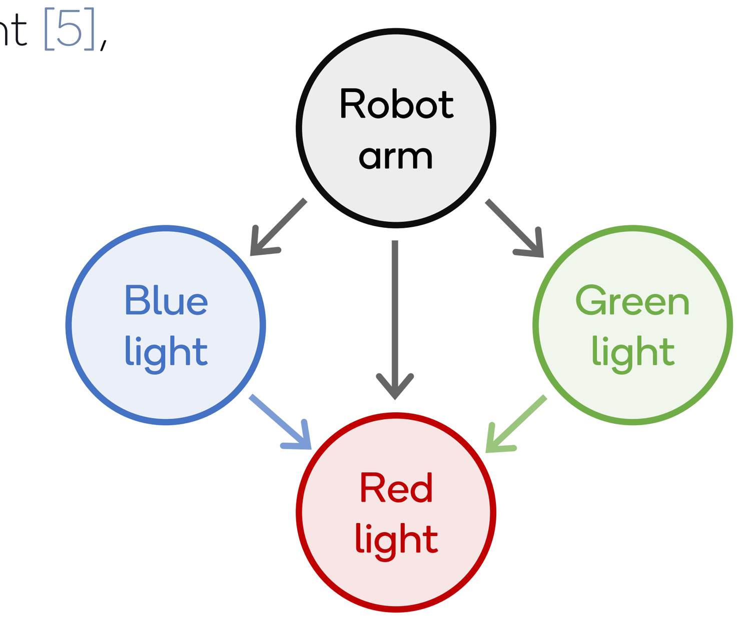
- For $\mathbb{R}$-valued variables, $\varphi_2(z_1, z_2)$ must be constant in z_1.
- Then z and z' are related by permutation + pointwise reparameterization

Latent causal models work in practice

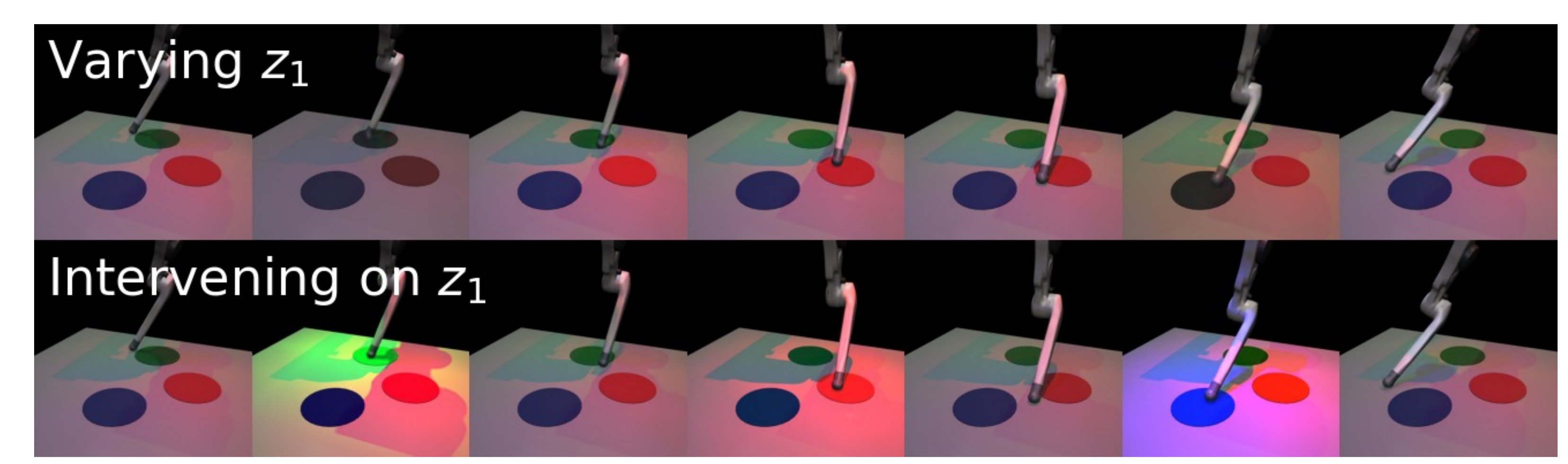
- Practical implementation of LCMs: **VAE with learnable SCM as prior**



- New **implicit** parameterization of latent causal structure
  - Latent noise variables & neural parameterization of solution function
  - No explicit graph parameterization in latent space necessary
  - Avoids optimization challenges due to explicit graph learning
- Experiments on toy data, Causal3DIdent [5], **new CausalCircuit dataset**
  - 2–4 causal factors
  - Various graphs, non-linear causal effects
  - Non-linear representations, up to 512 x 512 image data



- **LCMs identify the true causal graphs**
- **LCMs disentangle causal factors** better than acausal baselines
- **LCMs let us infer interventions and reason about them**



Varying one learned latent factor corresponds to changing a single true high-level concept; intervening on a learned latent correctly models interventions

References

[1] B. Schölkopf et al, "Towards Causal Representation Learning", IEEE proceedings 2021
[2] F. Eberhardt, "Green and grue causal variables", Synthese 2016
[3] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations", ICML 2019
[4] F. Locatello et al., "Weakly-supervised disentanglement without compromises", ICML 2020
[5] J. von Kügelgen et al., "Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style", NeurIPS 2021