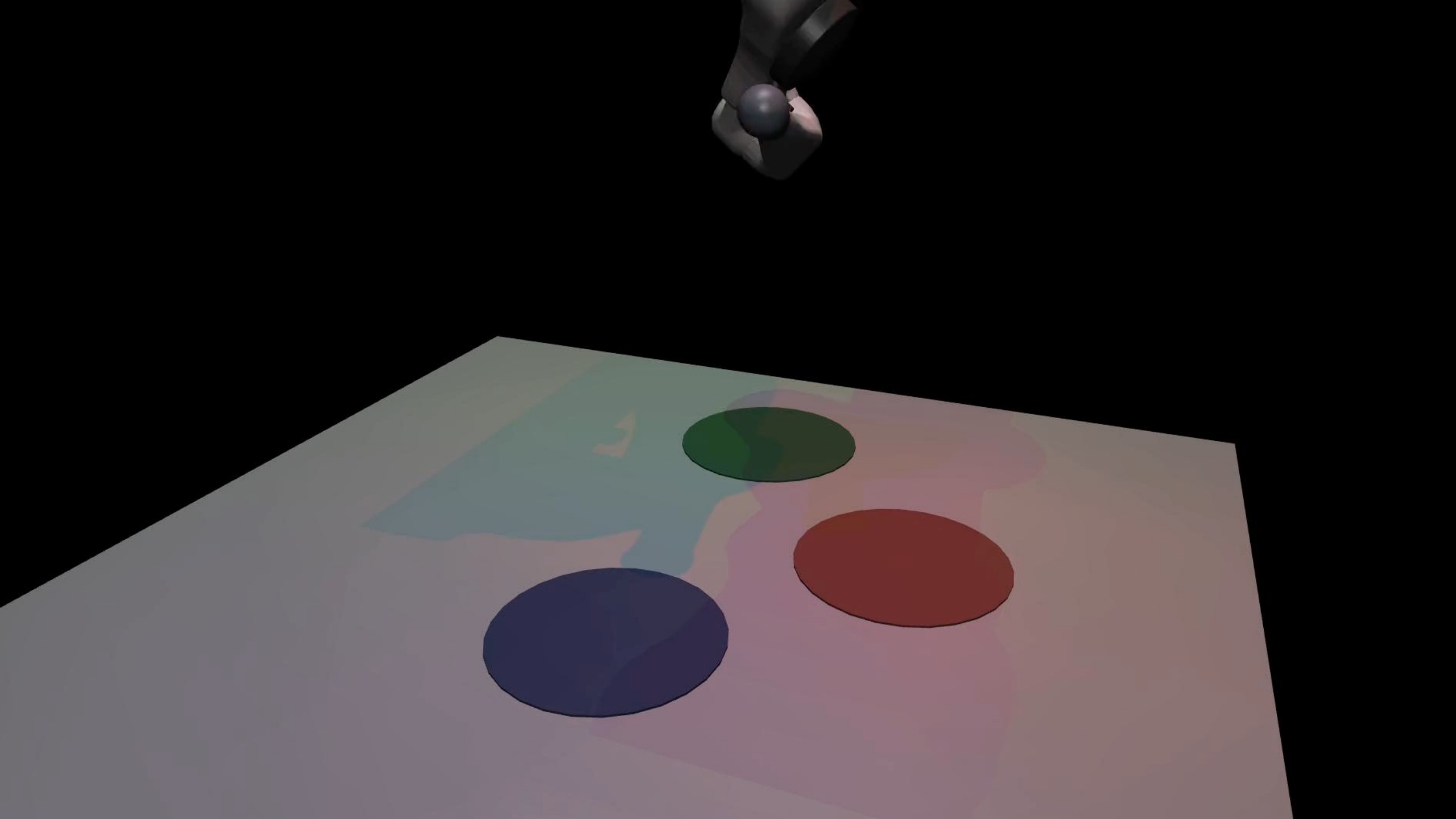
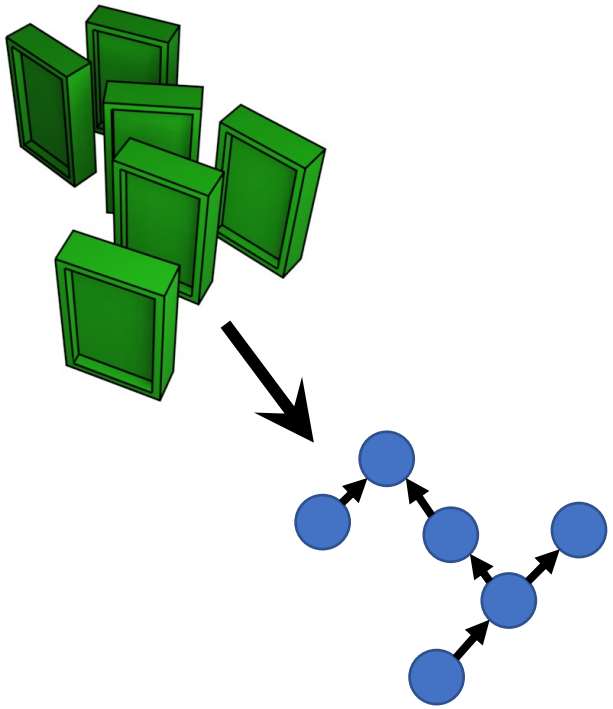
Causal representations and how to learn them

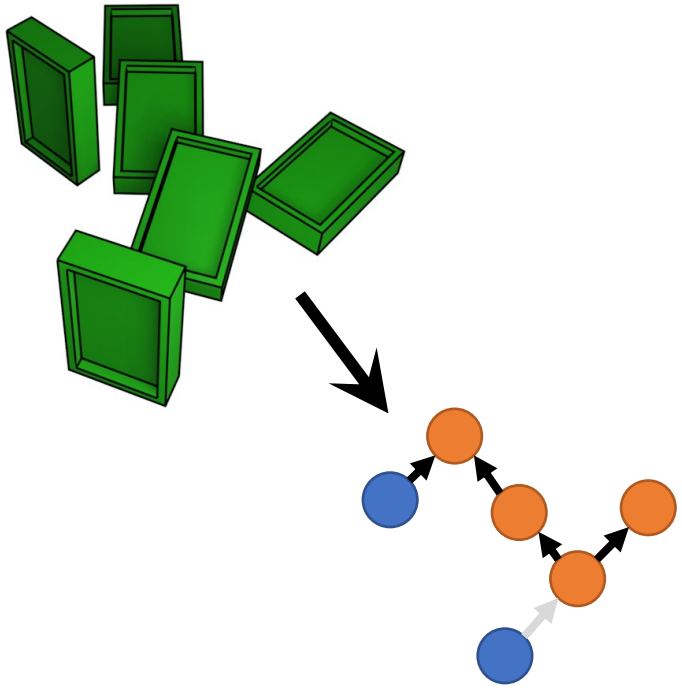
Johann Brehmer
Qualcomm AI Research

Work with Pim de Haan, Phillip Lippe, and Taco Cohen (NeurIPS 2022)

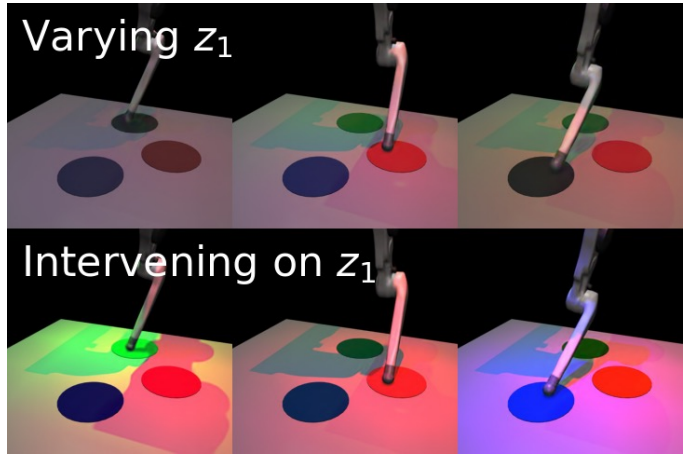




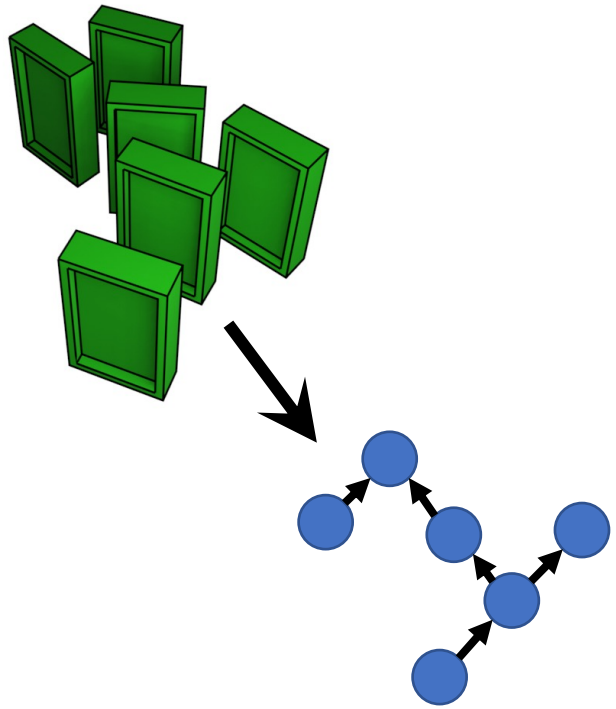
Can we **learn causal variables & causal structure from pixels**, without labels?



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

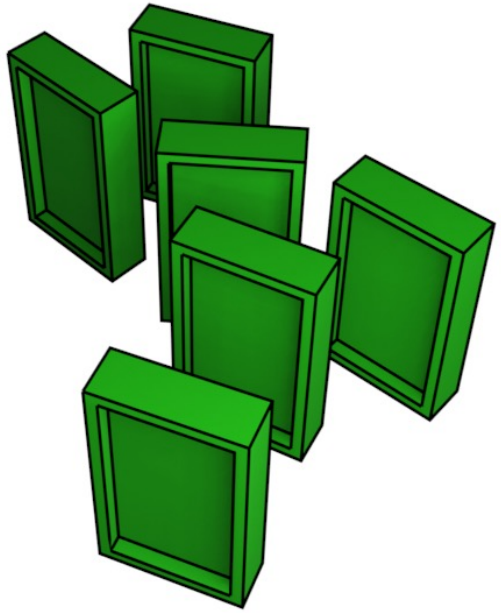


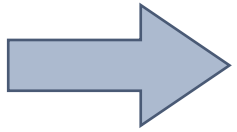
Can we **learn causal variables & causal structure from pixels**, without labels?

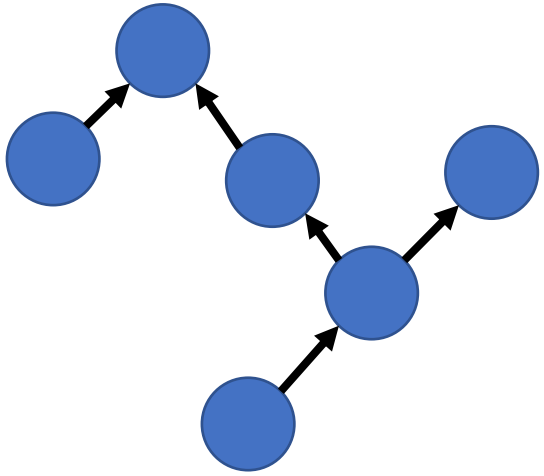
We prove: this is possible with **weak supervision**, when observing effects of interventions

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

Causal representation learning





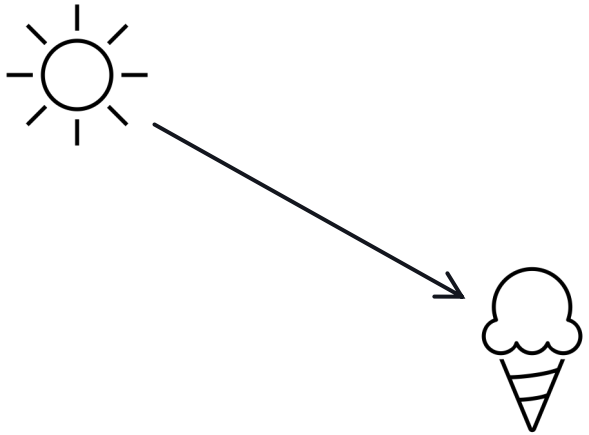


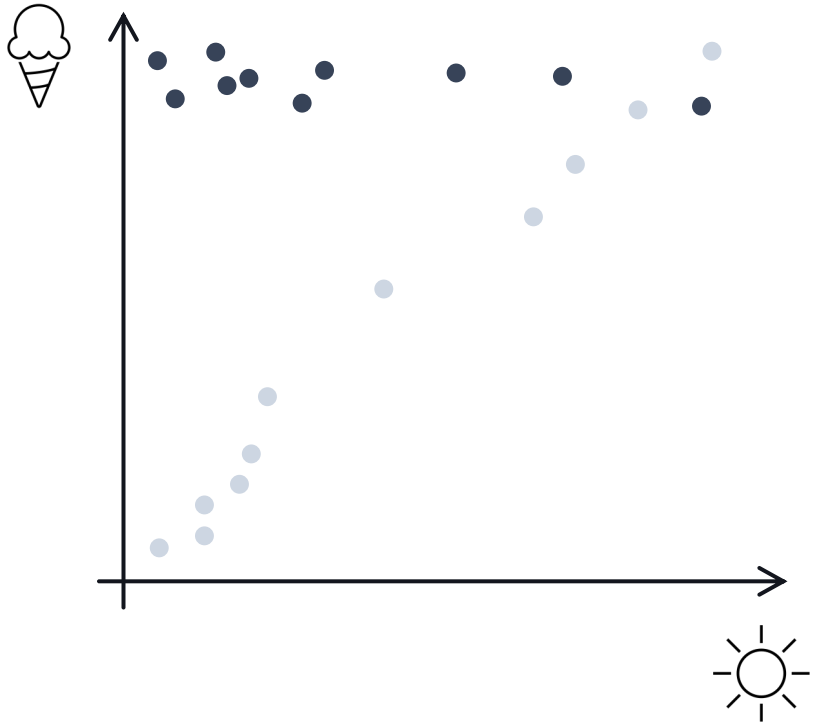
Given: **low-level, unstructured data representation**
(e.g. pixels)

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

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

What are causal representations?

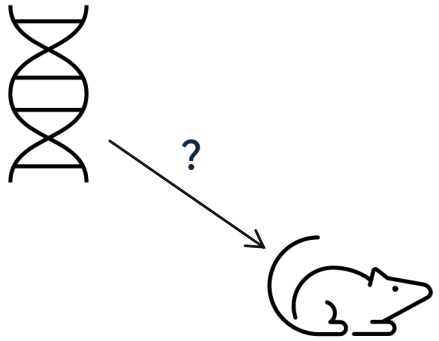




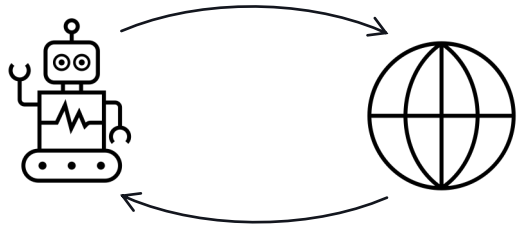
Semantically, causal models label relations between random variables as **cause-effect relations**

Functionally, causal models describe **probability distributions and how they change** under
actions / interventions

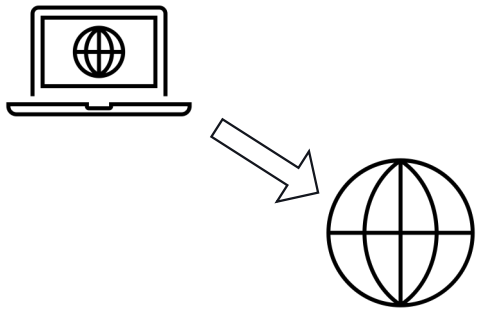
Why learn causal representations?



Causal structure may be of **scientific interest**

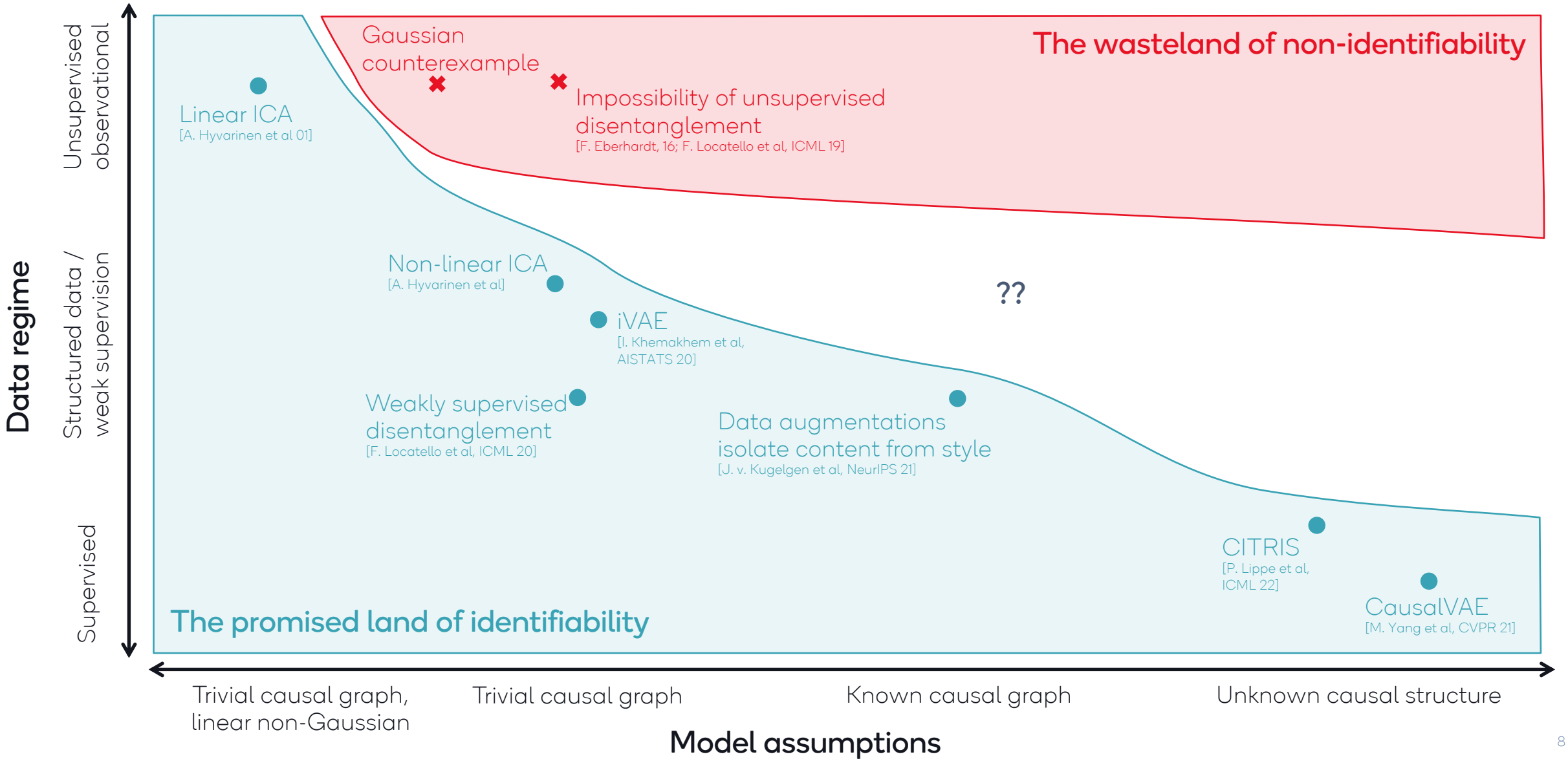


Causal representations are **abstractions** that may be **useful for planning**

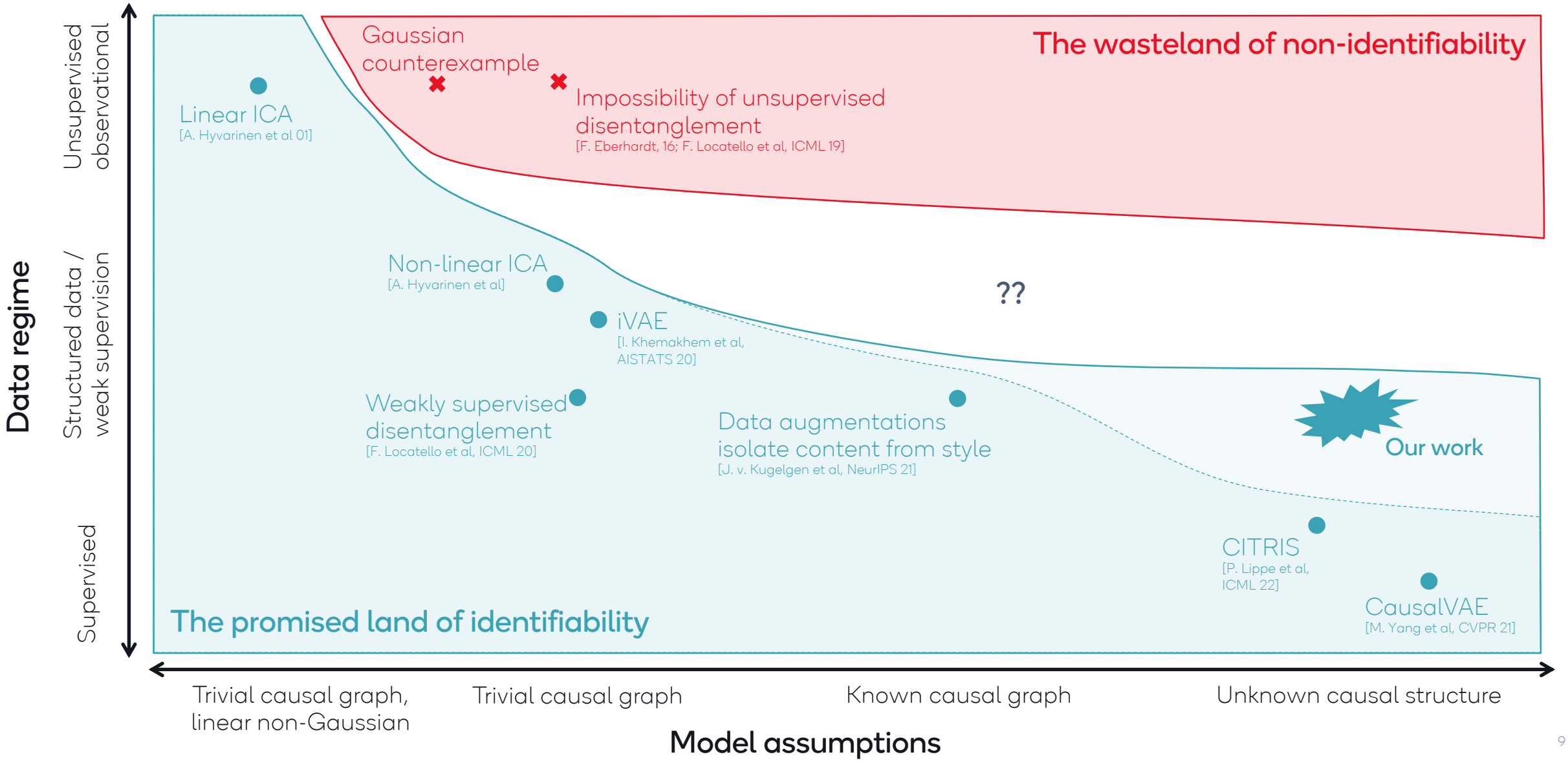


Causal models may be more **robust to changes**

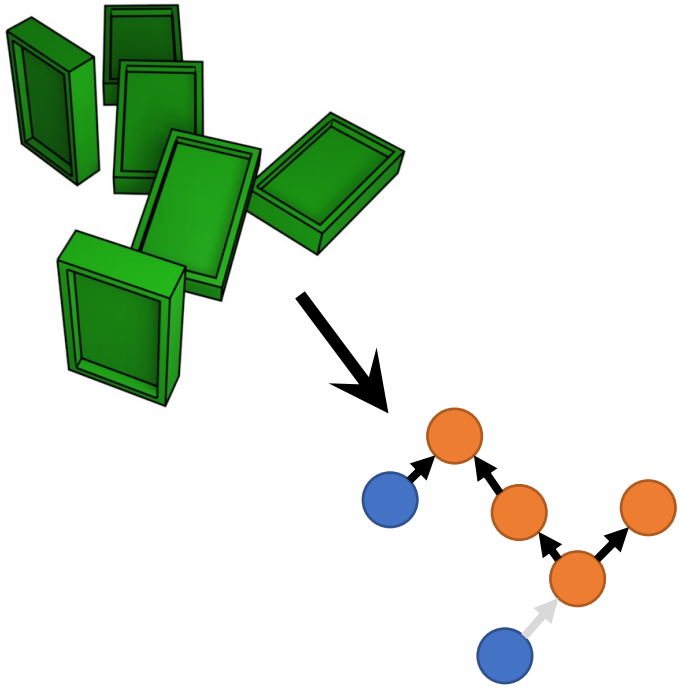
When can we learn causal representations?



When can we learn causal representations?



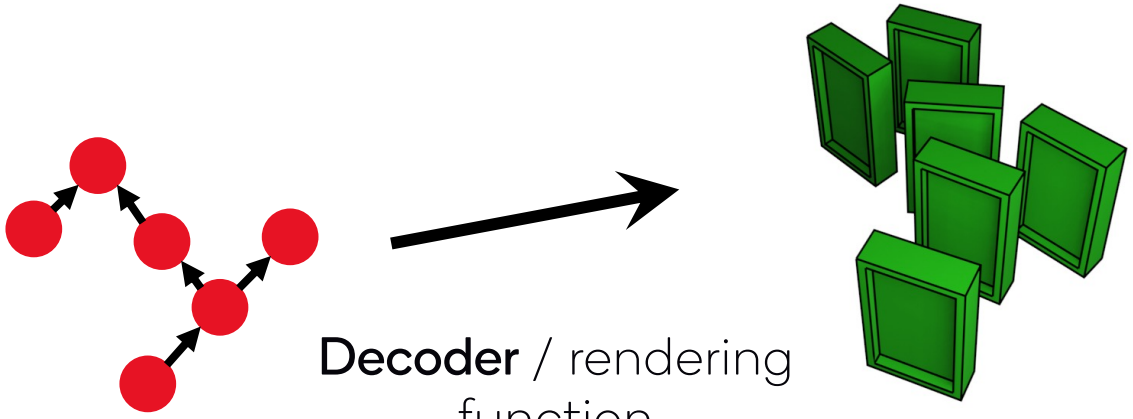
Can we **learn causal variables & causal structure from pixels**, without labels?



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

Latent causal model

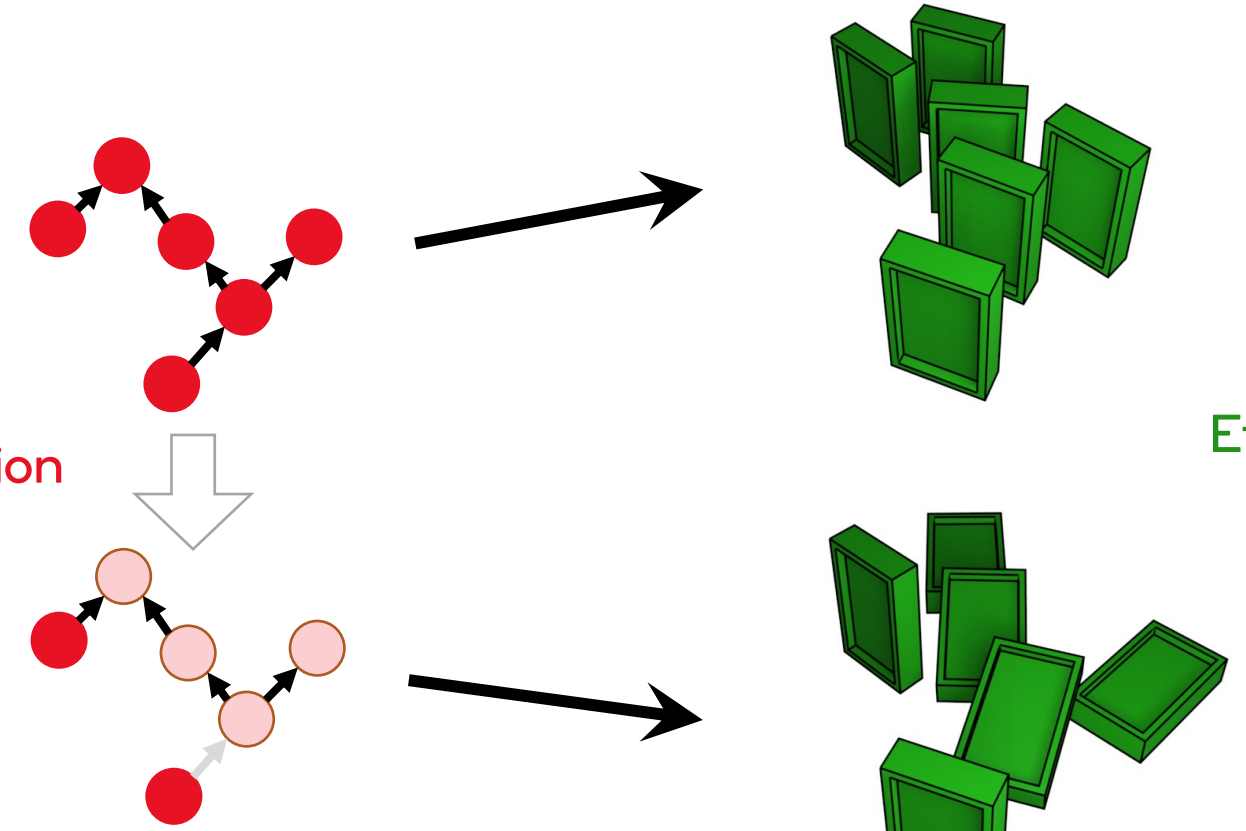


Decoder / rendering function

High-level variables with a structural causal model between them

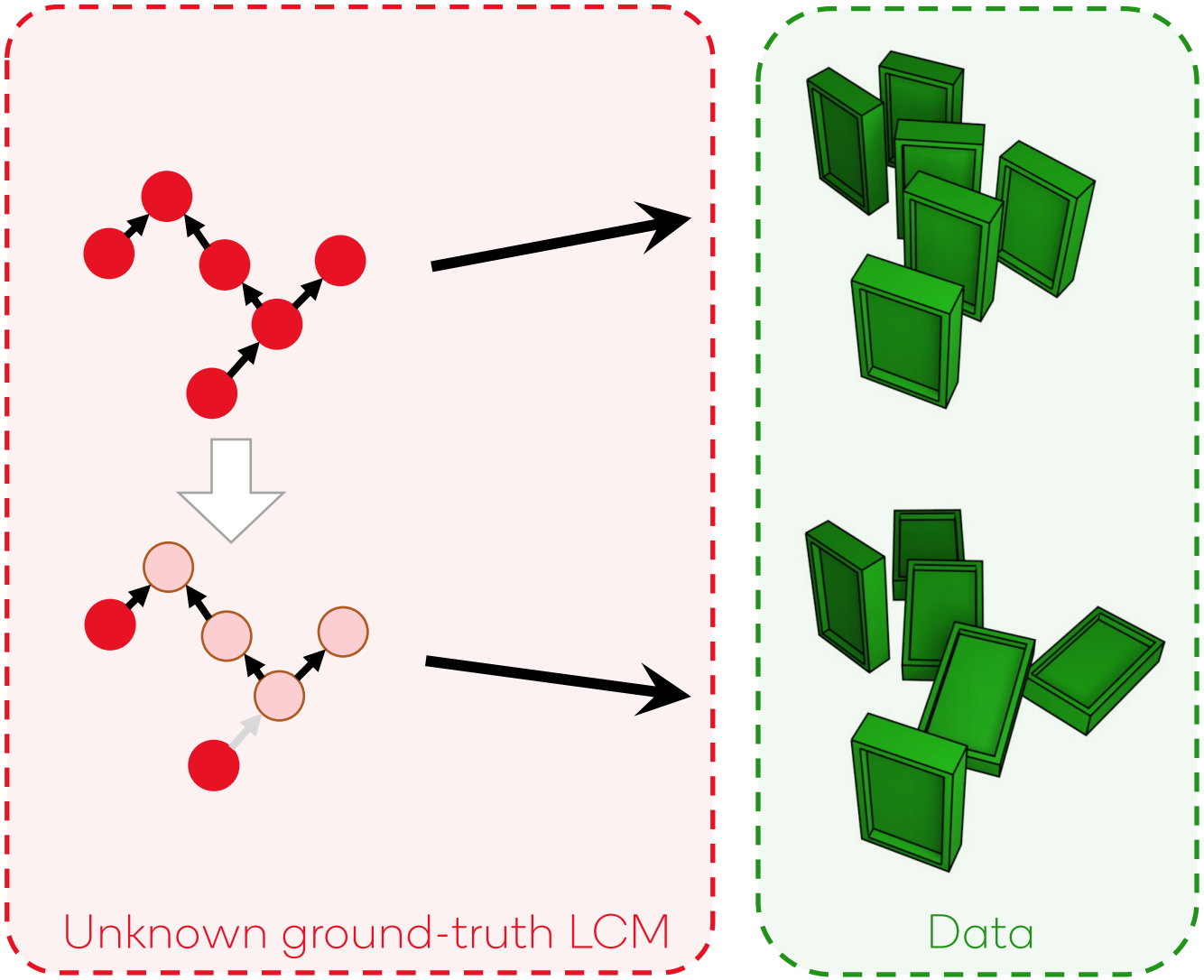
Low-level data (pixels)

Interventions



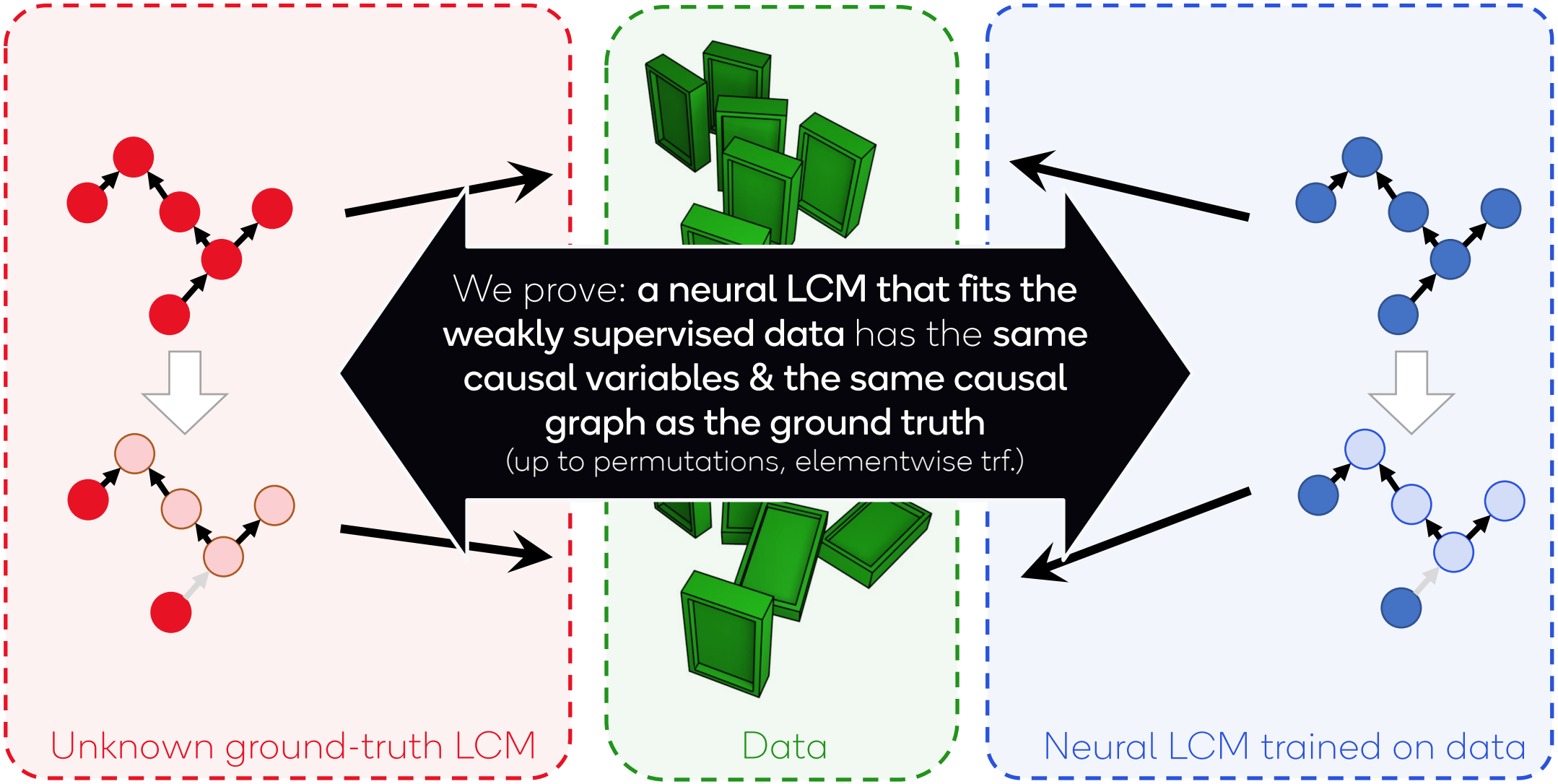
Effect of intervention
in data space

Weakly supervised data setting



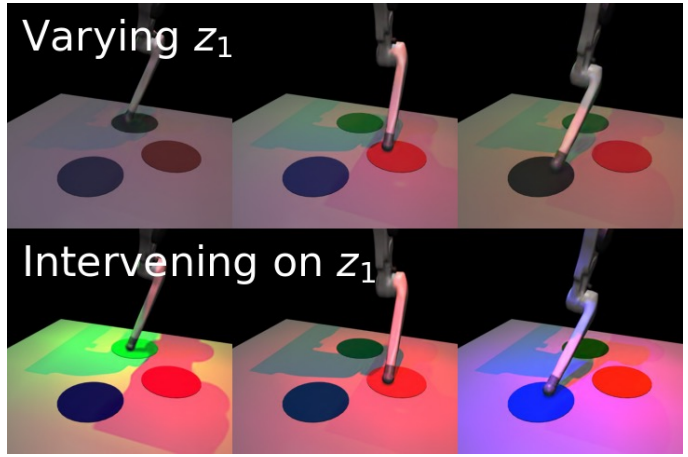
- We assume access to **data pairs of the system before and after interventions**
- Otherwise, **no labels**

Identifiability theorem



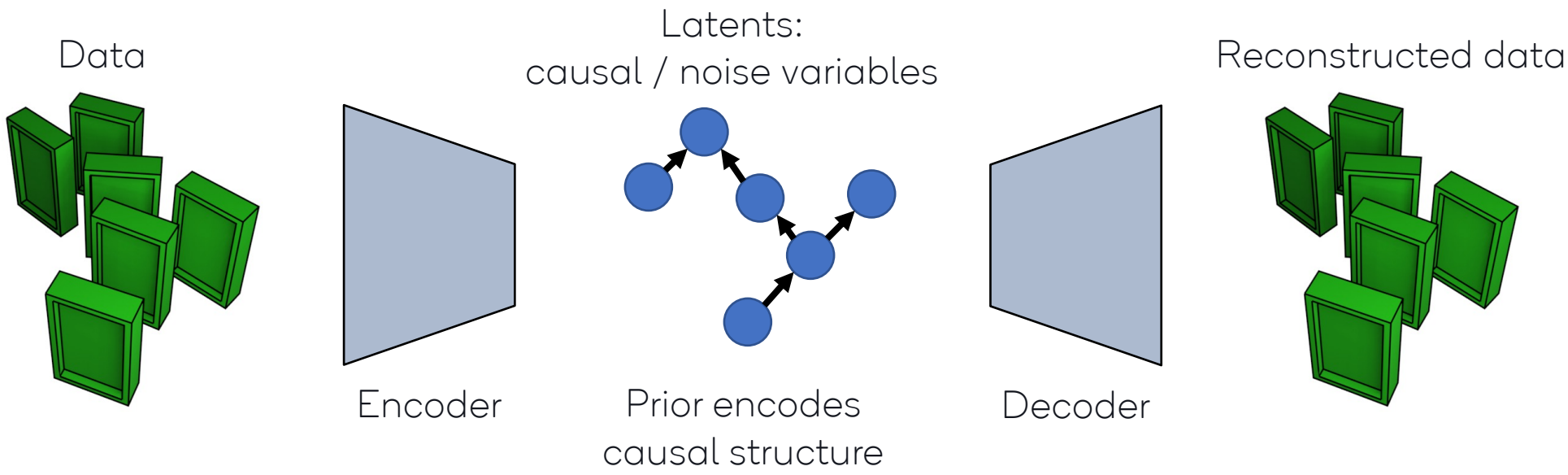
Can we **learn causal variables & causal structure from pixels**, without labels?

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

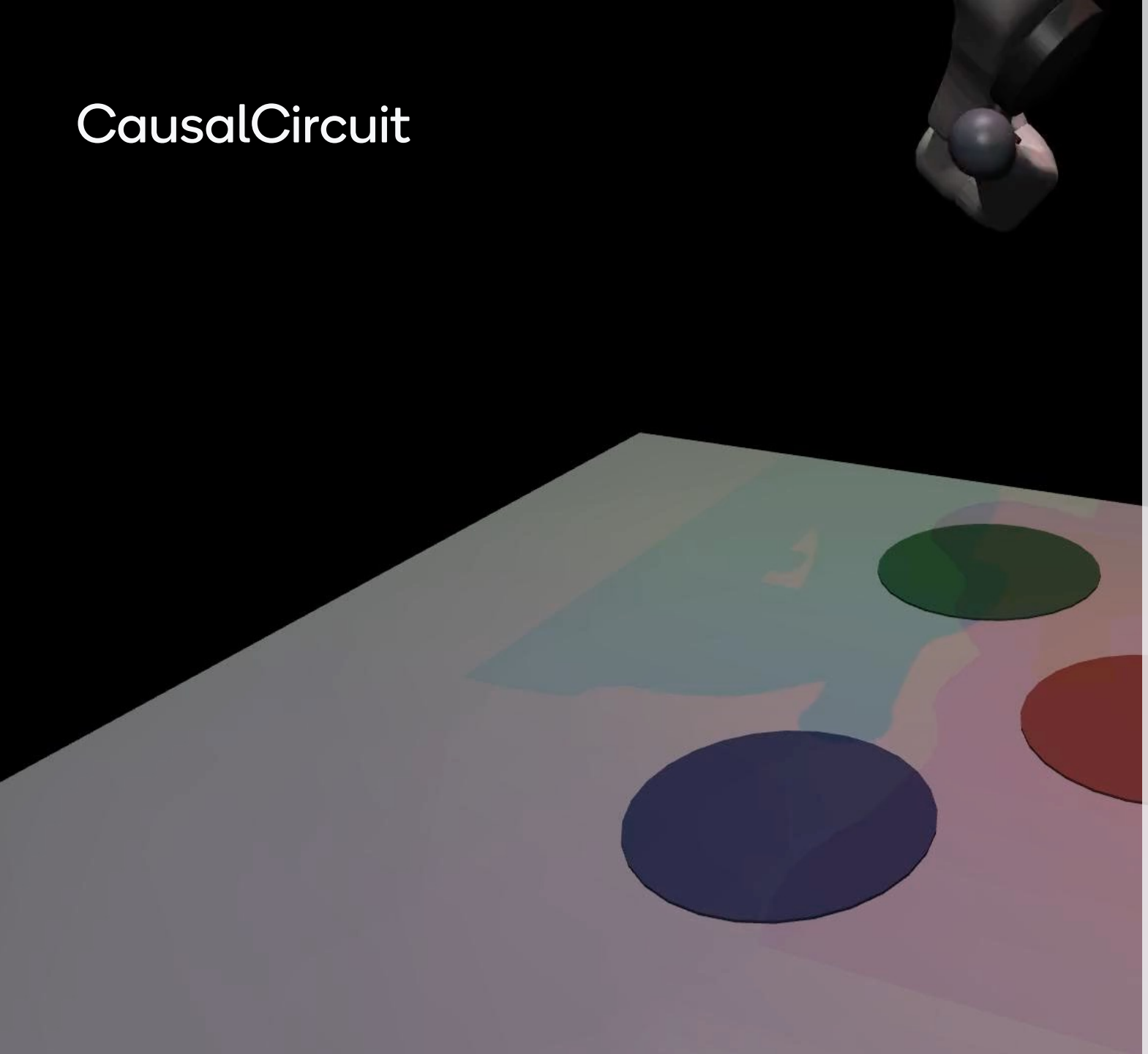


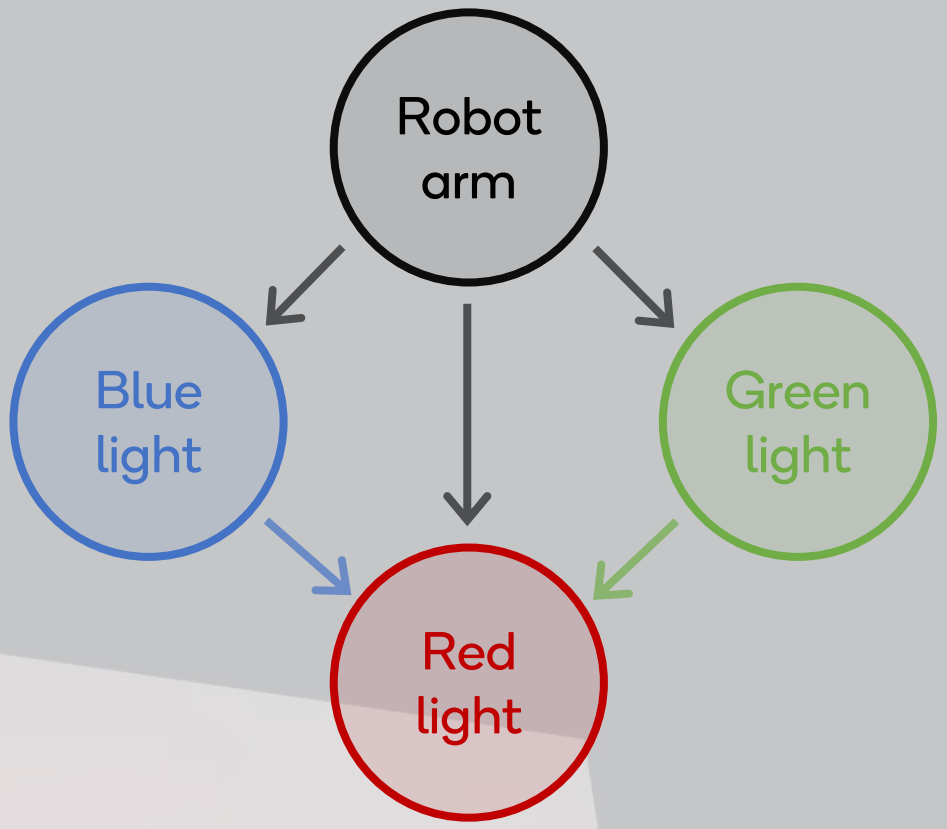
In practice, **implicit latent causal models** can identify the causal structure in image datasets

Operationalizing latent causal models

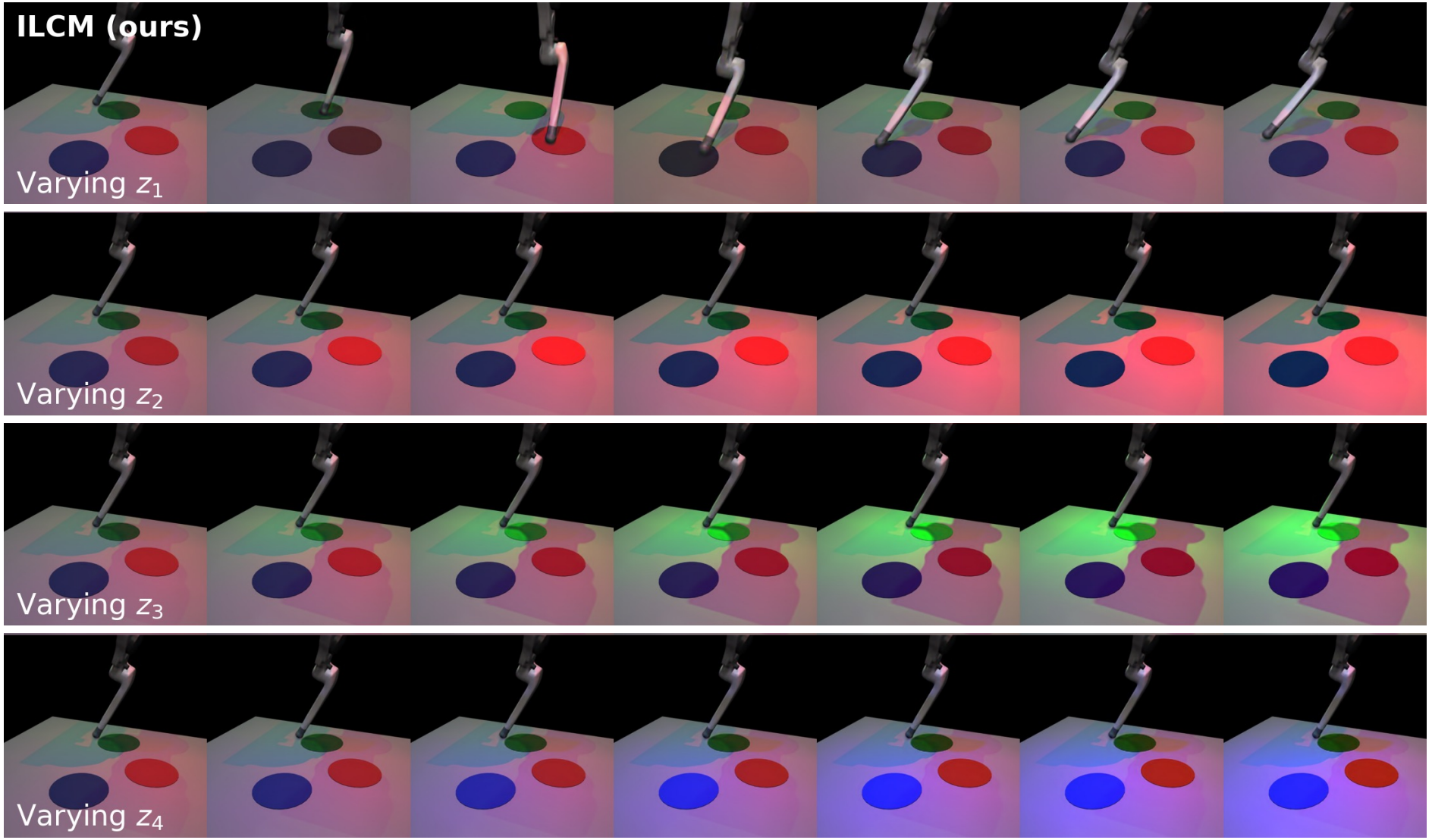


CausalCircuit

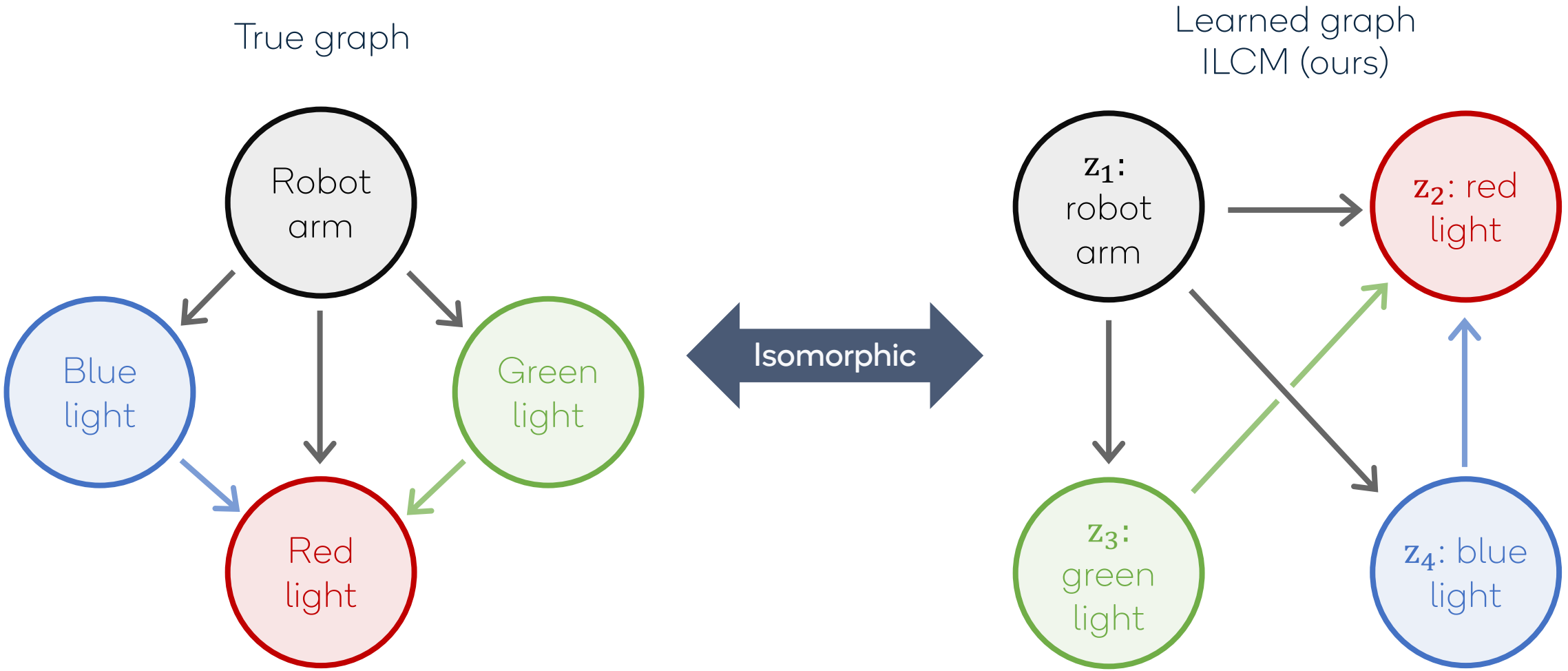




LCMs **disentangle** the causal variables

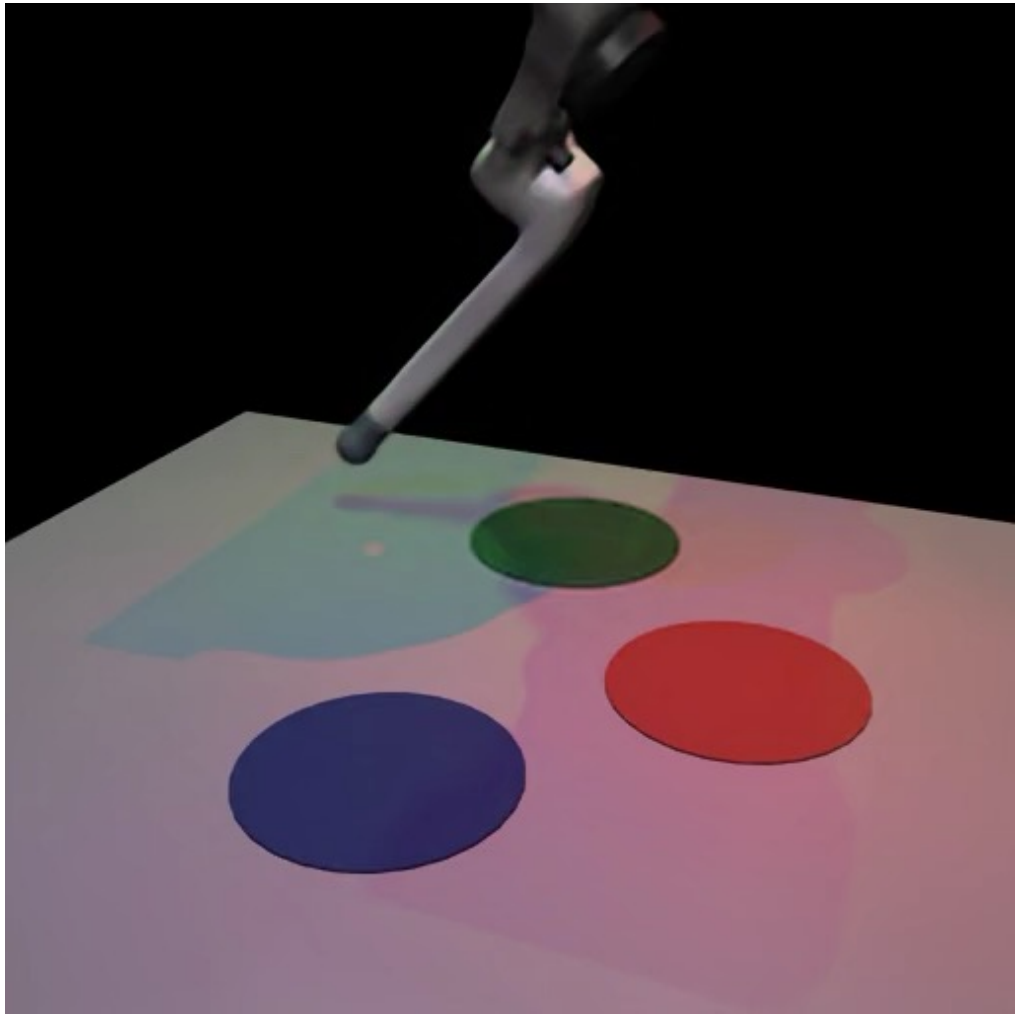


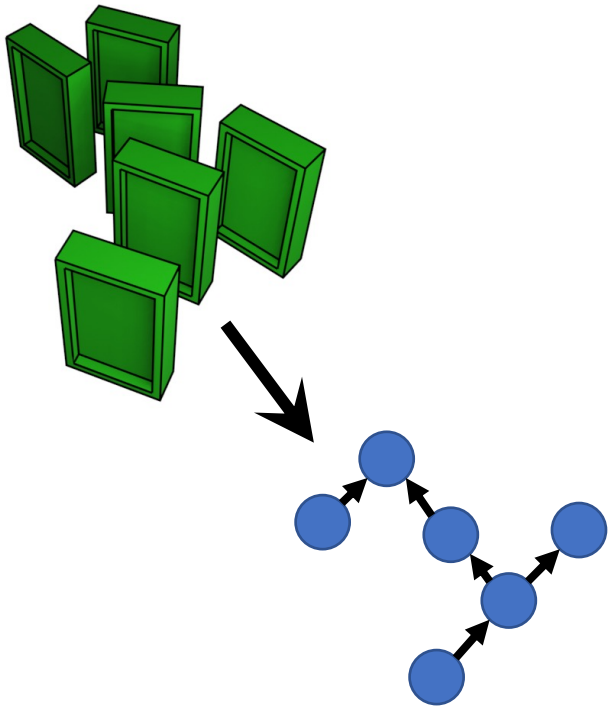
LCMs learn the **correct graph**



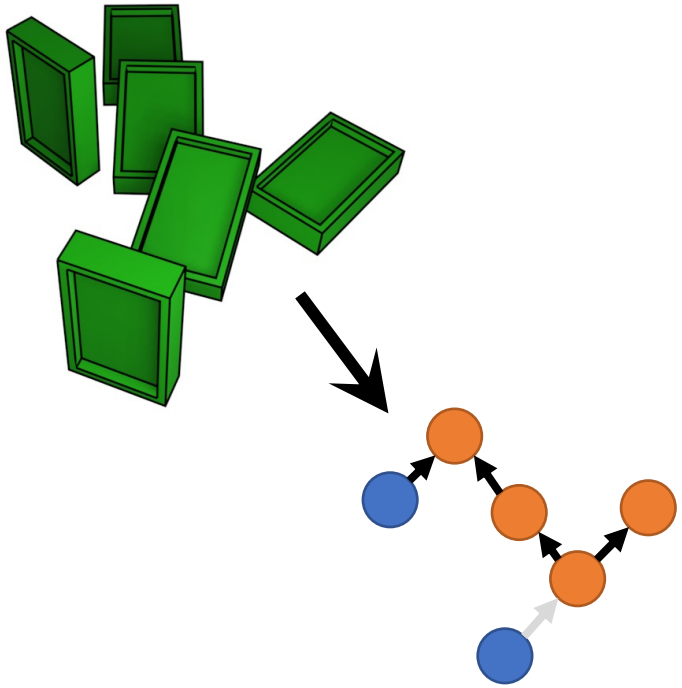
LCMs let us **reason causally**

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

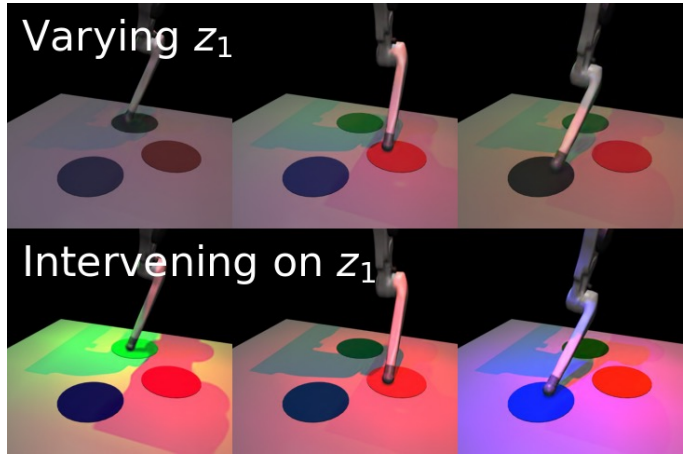




Can we **learn causal variables & causal structure from pixels**, without labels?



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

*equal contribution

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

Qualcomm

Follow us on:

For more information, visit us at:

qualcomm.com & qualcomm.com/blog