Towards the Reusability and Compositionality of Causal Representations

Davide Talon[1, 2] **Phillip Lippe**[3] **Stuart James**[1, 4] **Alessio Del Bue**[1] **Sara Magliacane**[5, 6]

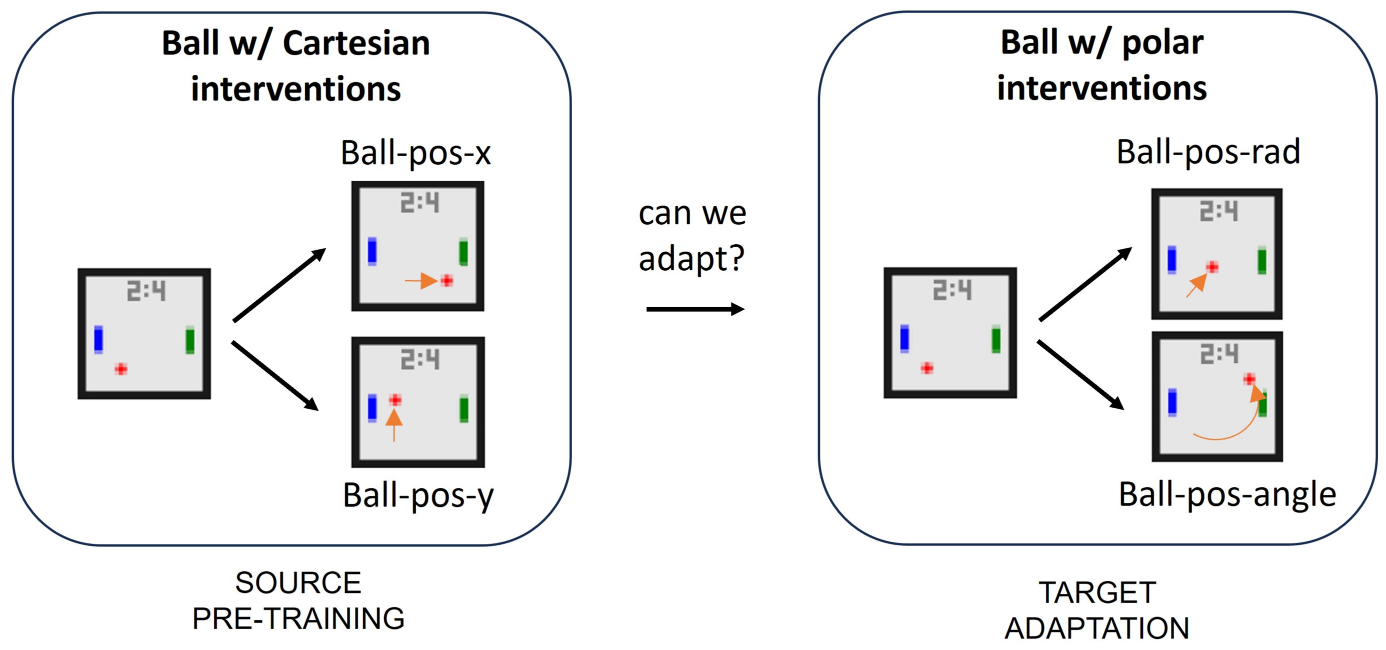
[1]PAVIS, Istituto Italiano di Tecnologia (IIT) [2]University of Genova [3]QUVA Lab, University of Amsterdam [4]Durham University
[5]AMLab, University of Amsterdam [6]MIT-IBM Watson AI Lab

TLDR;

Assuming temporal sequences with intervention targets, we detect changing causal factors across source and target environments and adapt or compose the representation for the unseen target.

Problem setting

- Causal representation learning is challenging with limited data
- Multiple environments where part of the factors are shared
- Example: only a subset of factors change, e.g., the ball position from Cartesian to polar coordinates

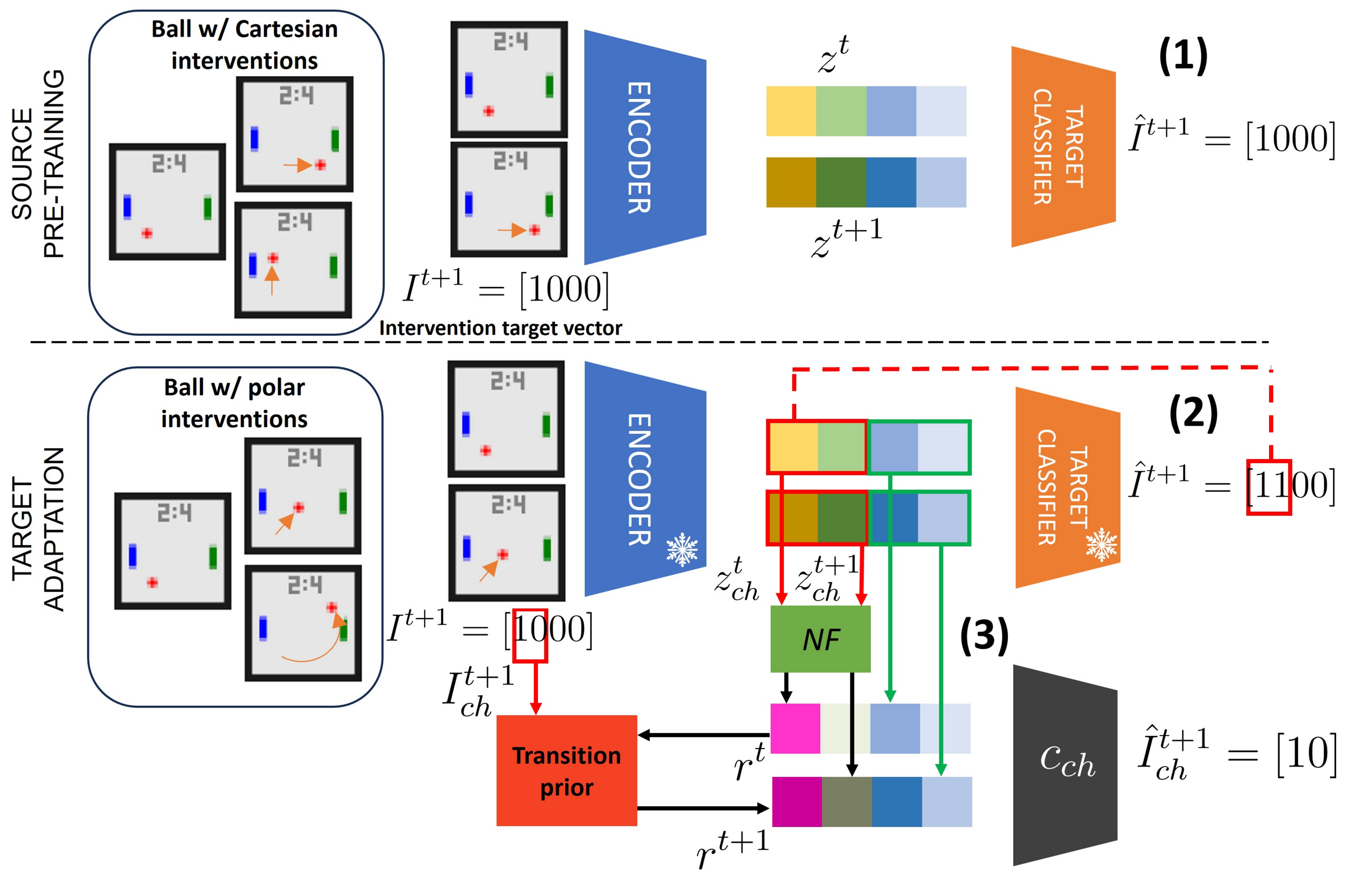


- Key assumption: we consider a temporal setting with access to intervention targets [1]

Method: DECAF

Build on available modular causal representations:

1. Pre-train on large source data
2. Detect changed factors as missclassified intervention targets
3. Adapt by learning a map from source to target representation for changed factors, re-use other factors



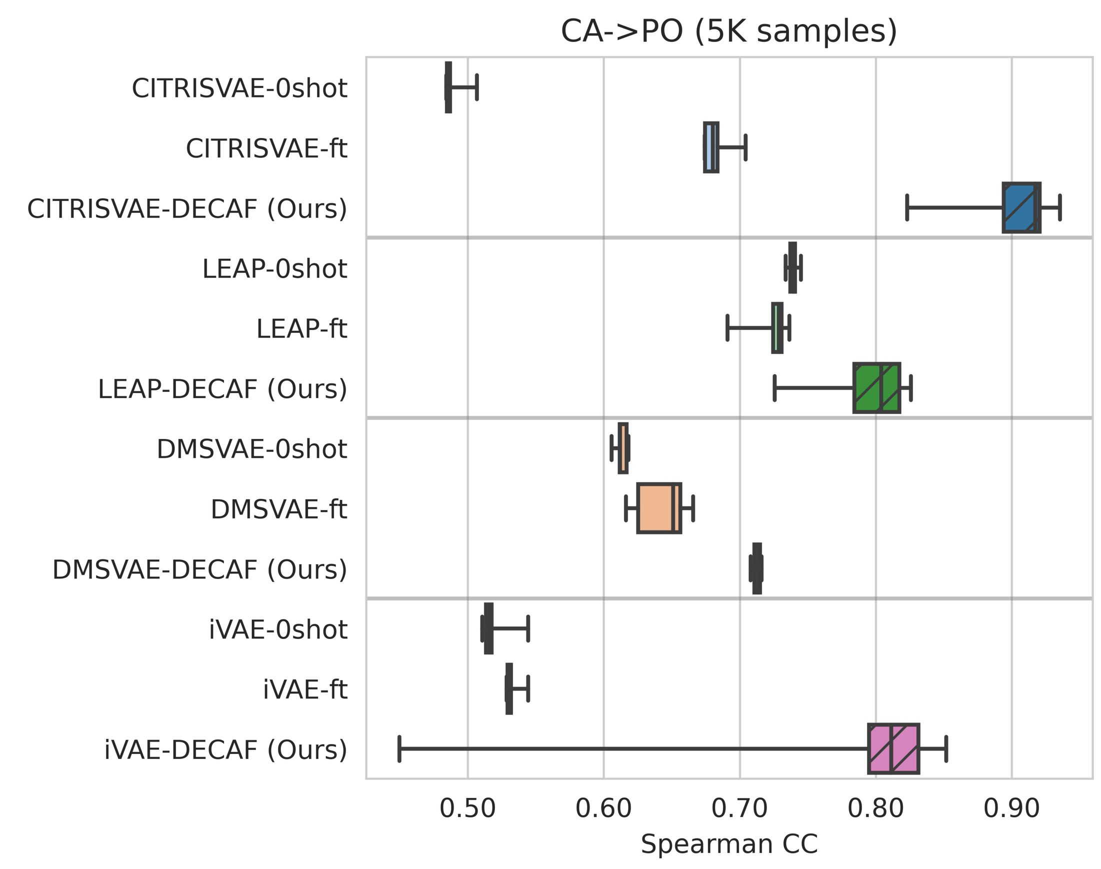
- Similarly, we can compose the representation of invariant factors coming from multiple environments

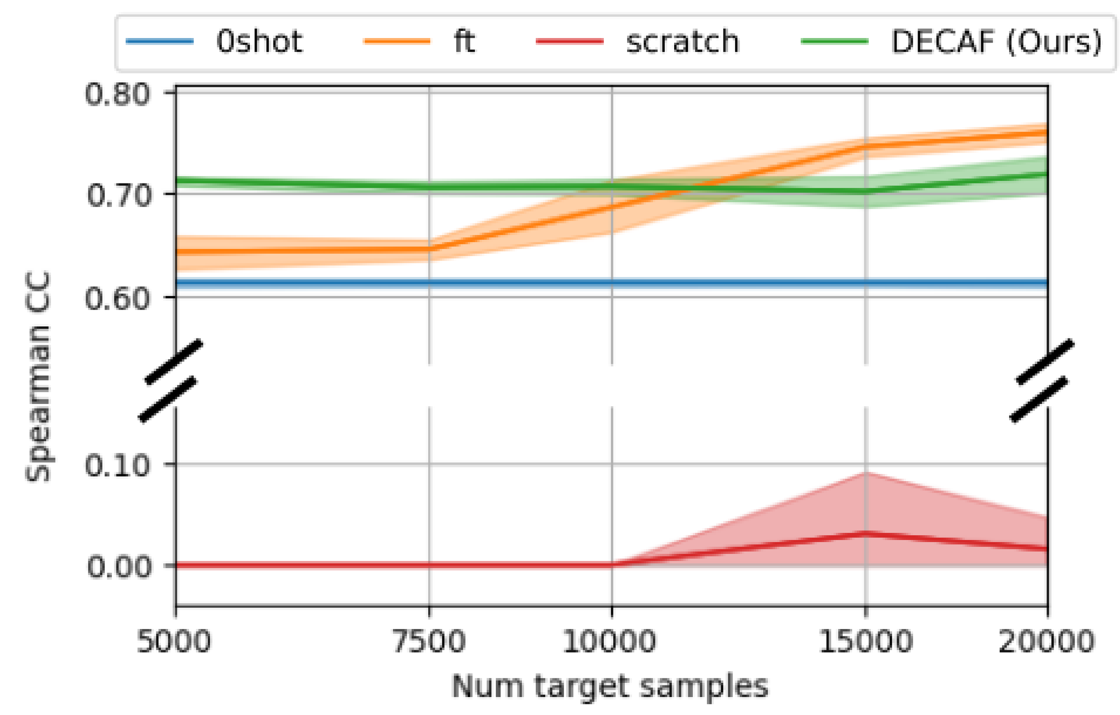
SCAN ME

[1] Lippe, Phillip, et al. "Citris: Causal identifiability from temporal intervened sequences." *International Conference on Machine Learning*, 2022.

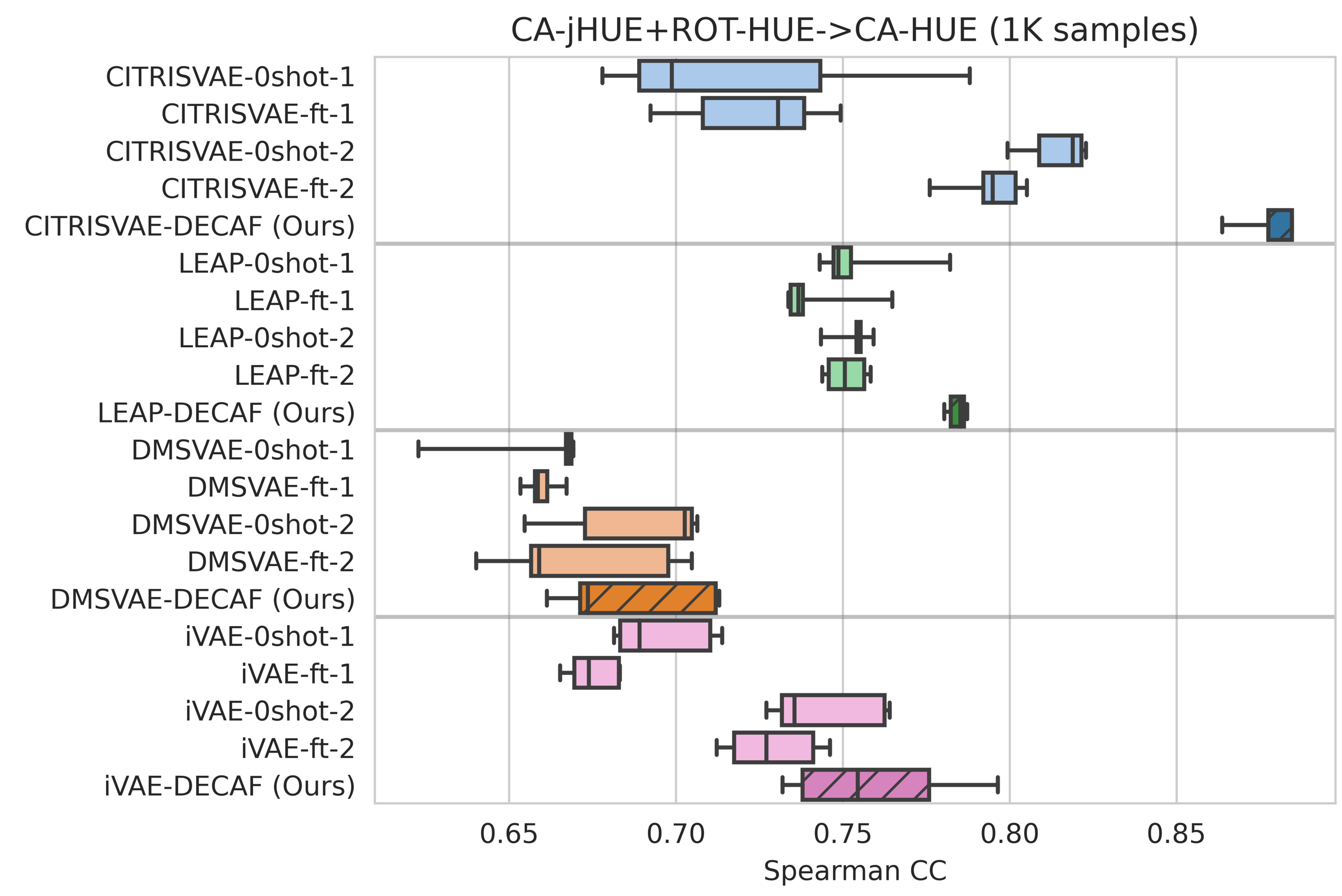
Experiments

- Evaluation metric (↑): combined version of `diag` and `off-diag` correlation w.r.t. generating factors
- InterventionalPong dataset: 6 variables, adapting the ball position from Cartesian to polar coordinates





- Temporal Causal3DIdent: 10 variables on a rendering scene, composing Cartesian position with independent hue colors



- Integrated with several CRL approaches, DECAF aids identification in the low data regime