WEAKLY SUPERVISED CAUSAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning high-level causal representations together with a causal model from unstructured low-level data such as pixels is impossible from observational data alone. We prove under mild assumptions that this representation is identifiable in a weakly supervised setting. This requires a dataset with paired samples before and after random, unknown interventions, but no further labels. Finally, we show that we can learn the representation and causal graph reliably in a simple synthetic domain using a variational autoencoder with a structural causal model as prior.

1 INTRODUCTION

The dynamics of many systems can be described in terms of some high-level variables and causal relations between them. Often, these causal variables are not known but only observed in some unstructured, low-level representation, such as the pixels of a camera feed. Learning the causal representations together with the causal structure between them is a challenging problem and important for instance for applications in robotics and autonomous driving (Schölkopf et al., 2021). Without prior assumptions on the data-generating process or supervision, it is impossible to identify the causal variables and their causal structure uniquely (Eberhardt, 2016; Locatello et al., 2019).

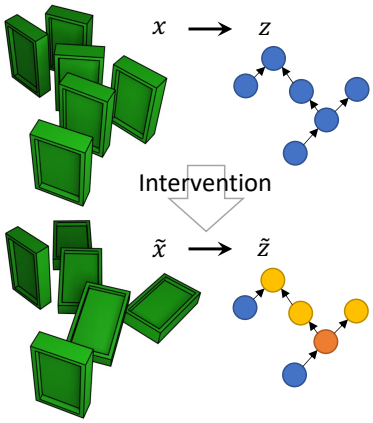

Figure 1: We learn to represent pixels x as causal variables z. The bottom shows the effect of intervening on the orange variable. We prove that variables and causal model can be identified from samples $(x, \tilde{x})$.

In this work, we show that a weak form of supervision is sufficient to identify both the causal representations and the structural causal model between them. We consider a setting in which we have access to data pairs, representing the system before and after a randomly chosen unknown intervention. Neither labels on the intervention targets nor active control of the interventions are necessary for our identifiability theorem, making this setting useful for offline learning. We prove that with this form of weak supervision, latent causal models (LCMs)—structural causal models (SCMs) together with a decoder from the causal factors to the data space—are identifiable up to a relabelling and elementwise reparameterizations of the causal variables.

We then introduce a practical implementation for LCM inference by using an SCM as a prior in a variational autoencoder (VAE). In a range of experiments, we show that LCMs can learn the true causal variables and the causal structure from unstructured data.

Related work Our work builds on the approach of Locatello et al. (2020) to *disentangled representation learning*. The authors introduce a similar weakly supervised setting where observations are collected before and after unknown interventions. In contrast to our work, however, they focus on disentangled representations, i. e. independent factors of variation with a trivial causal graph, which our work subsumes as a special case. Other relevant works on disentangled representation learning and (nonlinear) independent component analysis include Hyvärinen & Oja (2000); Shu et al. (2020); Khemakhem et al. (2020); Hälvä et al. (2021) and Lachapelle et al. (2022).

The problem of *causal representation learning* has been gaining attention lately, for a recent review see Schölkopf et al. (2021). Lu et al. (2021) learn causal representations by observing similar causal models in different environments. Von Kügelgen et al. (2021) use the weakly supervised setting to study self-supervised learning, using a known but non-trivial causal graph between content and style factors. Lippe et al. (2022b) learn causal representations from time-series data from labelled interventions, assuming that causal effects are not instantaneous but can be temporally resolved. Yang et al. (2021) propose to train a VAE with an SCM prior, but require the true causal variables as labels. To the best of our knowledge, our work is the first to provide identifiability guarantees for arbitrary, unknown causal graphs in this weakly supervised setting.

2 Identifiability of latent causal models from weak supervision

In this section, we show theoretically that causal variables and causal mechanisms are identifiable from weak supervision. In Sec. 3 we will then demonstrate how we can learn causal models in practice by training a causally structured VAE.

Setup We begin by defining latent causal models and the weakly supervised setting. Here, we only provide informal definitions and assume familiarity with common concepts from causality; we refer the reader to Appendix A.2 for a complete and precise treatment. In Appendix A.4, we discuss limitations of the setup and possible generalizations.

We describe the causal structure between latent variables as an SCM:

Definition 1 (Structural causal model (SCM))**.** *An SCM is a tuple $\mathcal{C} = \langle \mathcal{Z}, \mathcal{E}, F, p_{\mathcal{E}} \rangle$ consisting of the following:*

- *domains $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$ of causal (endogenous) variables $z_1, \ldots, z_n$;*
- *domains $\mathcal{E} = \mathcal{E}_1 \times \cdots \times \mathcal{E}_n$ of noise (exogeneous) variables $\epsilon_1, \ldots, \epsilon_n$;*
- *a directed acyclic graph $\mathcal{G}(\mathcal{C})$, whose nodes are the causal variables and edges represent causal relations between the variables;*
- *causal mechanisms $F = \{f_1, \ldots, f_n\}$ with $f_i : \mathcal{E}_i \times \prod_{j \in \mathbf{pa}_i} \mathcal{Z}_j \to \mathcal{Z}_i$; and*
- *a probability measure $p_{\mathcal{E}}(\epsilon) = p_{\mathcal{E}_1}(\epsilon_1)\, p_{\mathcal{E}_2}(\epsilon_2) \ldots p_{\mathcal{E}_n}(\epsilon_n)$ with full support that admits a continuous density.*

Additionally, we assume that $\forall i, \forall z_{\mathbf{pa}_i}, f_i(\cdot, z_{\mathbf{pa}_i}) : \mathcal{E}_i \to \mathcal{Z}_i$ is a diffeomorphism.

An SCM entails a unique solution $s : \mathcal{E} \to \mathcal{Z}$ defined by successively applying the structure functions. Because we require the structural equations to be pointwise diffeomorphic, s is also diffeomorphic. It also entails an observational distribution $p_{\mathcal{C}}(z)$ (Markov with respect to the graph of the SCM), which is the pushforward of $p_{\mathcal{E}}$ through the solution.

A perfect intervention $(I, (\tilde{f}_i)_{i \in I})$ modifies an SCM by replacing for a subset of the causal variables, called the intervention target set $I \subset \{1, ..., n\}$, the causal mechanism f_i with a new mechanism $\tilde{f}_i : \mathcal{E}_i \to \mathcal{Z}_i$, which does not depend on the parents. The intervened SCM has a new solution $\tilde{s}_I : \mathcal{E} \to \mathcal{Z}$. We define interventions to be atomic if the number of targeted variables is one or zero.

We will reason about generative models in a data space $\mathcal{X}$, in which the causal structure is latent. Also including a distribution of interventions, we define LCMs:

Definition 2 (Latent causal model (LCM))**.** *A latent causal model (LCM) is a tuple $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ of*

- *an acyclic SCM $\mathcal{C} = \langle \mathcal{Z}, \mathcal{E}, F, p_{\mathcal{E}} \rangle$ that admits a faithful distribution,*
- *an observation space $\mathcal{X}$,*
- *a decoder $g : \mathcal{Z} \to \mathcal{X}$ that is diffeomorphic onto its image,*
- *a set $\mathcal{I}$ of perfect interventions on $\mathcal{C}$, and*
- *a probability measure $p_{\mathcal{I}}$ over $\mathcal{I}$.*

We define two LCMs as equivalent if all of their components are equal up to a permutation of the causal variables and elementwise diffeomorphic reparameterizations of each variable.

Definition 3 (LCM isomorphism (informal))**.** *Let $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be two LCMs with identical observation space. An LCM isomorphism between*

them is a graph isomorphism $\psi : \mathcal{G}(\mathcal{C}) \to \mathcal{G}(\mathcal{C}')$ together with elementwise diffeomorphisms for noise and causal variables that tell us how to reparameterize them, such that the structure functions, noise distributions, decoder, intervention set, and intervention distribution of $\mathcal{M}'$ are compatible with the corresponding elements of $\mathcal{M}$ reparameterized through the graph isomorphism and elementwise diffeomorphisms. $\mathcal{M}$ and $\mathcal{M}'$ are equivalent, $\mathcal{M} \sim \mathcal{M}'$, if and only if there is an LCM isomorphism between them.

Following Locatello et al. (2020), we define a generative process of pre and post interventional data:[1]

Definition 4 (Weakly supervised generative process)**.** *Consider an LCM $\mathcal{M}$ where the SCM $\mathcal{C} = \langle \mathcal{Z}, \mathcal{E}, F, p_{\mathcal{E}} \rangle$ has continuous noise spaces $\mathcal{E}_i$, independent probabilities $p_{\mathcal{E}_i}$, and admits a solution s. We define the weakly supervised generative process of data pairs $(x, \tilde{x}) \sim p^{\mathcal{X}}_{\mathcal{M}}(x, \tilde{x})$ as follows:*

$$\begin{aligned} & & \epsilon \sim p_{\mathcal{E}}\,, & & z = s(\epsilon)\,, & & x = g(z)\,, \\ I \sim p_{\mathcal{I}}\,, & & \tilde{\epsilon} \sim \tilde{p}_{\tilde{\mathcal{E}}}(\tilde{\epsilon} \mid \epsilon, I)\,, & & \tilde{z} = \tilde{s}_I(\tilde{\epsilon})\,, & & \tilde{x} = g(\tilde{z})\,. \end{aligned} \tag{1}$$

Here we parameterize stochastic interventions on I by $\tilde{\epsilon}_i \sim p_{\mathcal{E}_i}$ for $i \in I$ and $\tilde{\epsilon}_i = \epsilon_i$ for $i \notin I$.

Identifiability result The main theoretical result of this paper is that an LCM $\mathcal{M}$ can be identified from $p(x, \tilde{x})$ up to a relabeling and elementwise transformations of the causal variables:

Theorem 1 (Identifiability of $\mathbb{R}$-valued LCMs from weak supervision)**.** *Let $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be LCMs with the following properties:*

- *The LCMs have an identical observation space $\mathcal{X}$.*
- *The SCMs $\mathcal{C}$ and $\mathcal{C}'$ both consist of n real-valued endogeneous causal variables and corresponding exogenous noise variables, i. e. $\mathcal{E}_i = \mathcal{Z}_i = \mathcal{Z}'_i = \mathcal{E}'_i = \mathbb{R}$.*
- *The intervention sets $\mathcal{I}$ and $\mathcal{I}'$ consist of all atomic, perfect interventions, $\mathcal{I} = \{\emptyset, \{z_0\}, \dots, \{z_n\}\}$ and similar for $\mathcal{I}'$.*
- *The intervention distribution $p_{\mathcal{I}}$ and $p'_{\mathcal{I}'}$ have full support.*

Then the following two statements are equivalent:

1. *The weakly supervised distributions entailed by the LCMs are equal, $p^{\mathcal{X}}_{\mathcal{M}}(x, \tilde{x}) = p^{\mathcal{X}}_{\mathcal{M}'}(x, \tilde{x})$.*
2. *The LCMs are equivalent, $\mathcal{M} \sim \mathcal{M}'$.*

Let us summarize the key steps of our proof, which we provide in its entirety in Appendix A.3. The direction $2 \Rightarrow 1$ follows from the definition of equivalence. The direction $1 \Rightarrow 2$ is proven constructively along the following steps:

1. We begin by defining the noise encoding of an LCM, $\sigma : \mathcal{X} \to \mathcal{E}, x \mapsto \sigma(x) = s^{-1} \circ g^{-1}(x)$. We show that if two LCMs have identical weakly supervised distributions, then their noise encodings are equal up to a permutation ψ and elementwise diffeomorphisms ϕ_i.[2]
   Under interventions, noise encodings almost surely change only the components corresponding to the intervention targets: $\sigma(x)_i \neq \sigma(\tilde{x})_i \Leftrightarrow i \in I$. For each pair $(x, \tilde{x})$, the intervention targets are therefore identifiable, up to the permutation ψ.
2. Next, we show that the independence patterns in $p^{\mathcal{X}}_{\mathcal{M}}(x, \tilde{x})$ carry a signal of the causal descendance structure in $\mathcal{C}$. In particular, under an intervention $I = \{z_j\}$, we show that $\sigma(x)_i \not\perp\!\!\!\perp \sigma(\tilde{x})_j \mid I \Leftrightarrow i \in \mathbf{desc}_j$. This implies that two LCMs with identical weakly supervised distributions must have the same causal descendance structure, up to the permutation ψ. Note that this step crucially relies on the interventions being perfect, i. e. not depending on the pre-intervention value of any variable.
3. Finally, we show that the causal variables of $\mathcal{M}$ and $\mathcal{M}'$ need to be equal up to permutations and elementwise transformations. We show that the variable $z'_{i'}$ in $\mathcal{M}'$ is unchanged when intervening on a non-ancestor of z_i and thus must be constant in the non-ancestor. Also, it is constant in all ancestors of z_i, because intervening on i makes $\tilde{z}_{i'}$ and the ancestors independent. Thus $z'_{i'}$ can only depend on z_i and $\mathcal{M} \sim \mathcal{M}'$.

[1] This construction is closely related to twinned SCMs (Bongers et al., 2021, Def. 2.17), typically used to compute counterfactual queries $p(\tilde{z}_{\setminus I} | z, \tilde{z}_I)$. We instead focus on the joint distribution of pre-intervention and post-intervention data.

[2] This part of our proof also follows from the identifiability theorem for disentangled representations in Locatello et al. (2020, Theorem 1).

3 EXPERIMENTS

The identifiability result in Sec. 2 suggests that one can identify an LCM by finding the maximum-likelihood solution of $p(x, \tilde{x})$. We now investigate whether we can indeed learn LCMs in practice. The datasets, models, and training setup are described in more detail in Appendix B.

Relaxation of the weakly supervised setting In realistic systems, the "before" and "after" state of a system are recorded some time apart. We do not expect that the causal variables unaffected by an intervention remain *perfectly* invariant during this time. Similarly, we want to allow for causal descendants of the intervention targets to slightly change from the exact value predicted by the weakly supervised process in Eq. (1). We model this by applying a Gaussian convolution with small variance σ^2 to $p(z, \tilde{z} \mid I)$ on all dimensions of $\tilde{z}$ that are not intervened on, resulting in a continuous density. This relaxation means that there is a gap between the requirements of our identifiability theorem is and the experimental setting. In particular, relaxing the exact manifold in the weakly supervised data space to a "fuzzy" one renders our argument for the identifiability of noise encodings and intervention targets invalid; see Brehmer & Cranmer (2020) for a related discussion. Nevertheless, we empirically find that we can still reliably identify LCMs in this more realistic setup.

Table 1: Experiment results, comparing LCMs to unstructured β-VAEs and disentanglement VAEs (dVAE), best results in bold. For Causal3DIdent, we average over five datasets with different graphs, see Appendix B. We measure disentanglement with the DCI score D. $D \approx 1$ implies disentanglement succeeds only with our method. LCMs learn the correct causal graph most of the time, visible as a structural Hamming distance SHD ≈ 0. The quality of intervention inference is evaluated with the intervention negative log posterior ($-\log p_I$).

Method	D	SHD	$-\log p_I$
2D toy data			
LCM	**0.99**	**0.0**	**0.28**
dVAE	0.35	n/a	0.33
β-VAE	0.00	n/a	n/a
Causal3DIdent			
LCM	**0.98**	**0.2**	**0.21**
dVAE	0.62	n/a	1.82
β-VAE	0.44	n/a	n/a

LCM implementation In practice, we implement LCMs as VAEs (Kingma & Welling, 2014). In the simplest version, the causal graph $\mathcal{G}$ is fixed. The SCM and interventional distributions define the relaxed prior $p(z, \tilde{z}|I)$. Rather than with a diffeomorphism g, we map causal latents to observed variables through a stochastic encoder $q(z|x)$ and decoder $p(x|z)$. We then maximize the following lower bound on the weakly supervised log likelihood:

$$\log p(x, \tilde{x}) \geq \mathop{\mathbb{E}}_{\substack{z \sim q(z|x) \\ \tilde{z} \sim q(\tilde{z}|\tilde{x})}} \left[\log \mathop{\mathbb{E}}_{I \sim p_{\mathcal{I}}(I)} [p(z, \tilde{z}|I)] + \log p(x|z) + \log p(\tilde{x}|\tilde{z}) - \log q(z|x) - \log q(\tilde{z}|\tilde{x}) \right]. \quad (2)$$

For our experiments with small graphs, we instantiate one such LCM per graph structure, train them on the VAE loss in Eq. (2), and select the model with the lowest validation loss. To scale to larger graphs, it will be beneficial to learn a belief over graphs and to sample adjacency matrices and intervention targets from suitable priors, for instance using Gumbel-Softmax methods (Jang et al., 2017; Brouillard et al., 2020) or other approaches (Lippe et al., 2022a); we will explore this in future work.

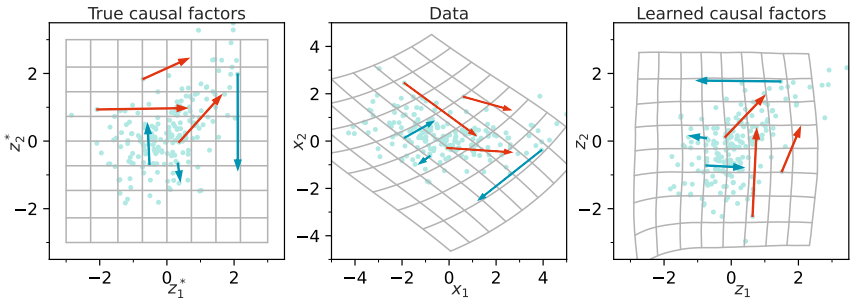

Figure 2: 2D toy data with graph $z_1^* \to z_2^*$. The grey grids show the map between true causal factors, data, and latent causal factors learned by the LCM, with graph $z_2 \to z_1$. The mint dots indicate the observational data distribution, the arrows from z to $\tilde{z}$ show interventions targeting z_1^* (red) or z_2^* (blue). The fact that axis-aligned lines in the true latent space are mapped to axis-aligned lines in the learned latent space, implies that the disentanglement succeeded.

Following common practice in causal discovery (Zheng et al., 2018; Brouillard et al., 2020; Lippe et al., 2022a), we incentivize learning the sparsest graph compatible with the data distribution by adding a regularization term proportional to the number of edges in the graph to the loss.

Baselines We compare LCMs to an unstructured β-VAE that treats x and $\tilde{x}$ as i. i. d. and uses a standard Gaussian prior. We also compare to a disentanglement VAE, which models our weakly supervised process for a trivial causal graph (i. e. independent factors of variation), similar to the method proposed by Locatello et al. (2020).

2D toy experiment We first test LCMs in a toy experiment with $\mathcal{X} = \mathcal{Z} = \mathbb{R}^2$. Training data is generated from a nonlinear SCM with the graph $z_1 \to z_2$ and mapped to the data space through a randomly initialized normalizing flow.

An LCM trained in the weakly supervised setting is able to reconstruct the causal factors and the causal graph accurately up to a permutation of the two variables, as shown in the Fig. 2. It fits the weakly supervised data distribution with a better log likelihood than the acausal baselines. In Tbl. 1 we quantify the quality of the learned representations with the DCI disentanglement score D (Eastwood & Williams, 2018). We find that our LCM is able to disentangle the causal factors almost perfectly, while the baselines, which assume independent factors of variation, fail as expected. Finally, we test whether the learned LCM correctly infers the interventions by computing the intervention posterior $p(I|x, \tilde{x})$ and evaluating it for the true intervention I^*, again finding better results for the causal LCM than for the baselines.

Causal3DIdent Next, we test LCMs on an adaptation of the Causal3DIdent dataset (von Kügelgen et al., 2021), which contains images of three-dimensional objects under variable positions and lighting conditions. We consider three latent variables representing object hue, the spotlight hue, and the position of the spotlight. We consider five versions of this dataset, each with a different causal graph, randomly initialized nonlinear structure functions, and heteroskedastic noise. These are mapped to images with a resolution of 64×64, see Fig. 3 for examples.

An LCM with convolutional encoder and decoder is again able to learn the causal variables and the causal structure between them, finding the correct causal graph in all but one settings. The results in Tbl. 1 and Fig. 3 show that the learned representations are more disentangled than those learned by methods that do not account for causal structure and that the LCM can infer interventions more reliably than the baselines.

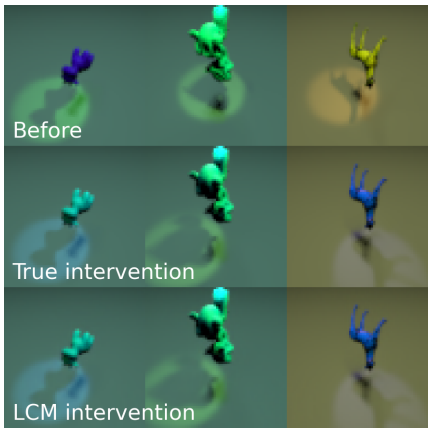

Figure 3: Causal3DIdent samples before (top row) and after (middle row) interventions, and post-intervention samples generated from the LCM under the intervention inferred from the data, indicating we correctly learned to intervene.

4 Discussion

We have presented a method for causal representation learning in a weakly supervised setting, in which data consist of a system before and after a random intervention. We have proven that in this setting both the causal variables and the causal structure between them is identifiable up to permutations and elementwise reparameterizations of the causal variables. This extends the results by Locatello et al. (2020) from independent factors of variation (trivial causal graphs) to arbitrary causal structures.

This identifiability result relies on a few key assumptions, including that interventions are perfect, that all atomic interventions may be observed, and that the causal variables are real-valued. We discuss these requirements and their potential relaxation in Appendix A.4.

In practice, LCMs can be implemented in a variational autoencoder. We demonstrated that such models succesfully learn the causal variables and nonlinear causal structure from unstructured data, both in two-dimensional toy experiments and in rendered image data.

REFERENCES

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2021. http://www.blender.org.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M. Mooij. Foundations of Structural Causal Models with Cycles and Latent Variables. *Annals of Statistics*, 49(5):2885–2915, 2021. doi: 10.1214/21-AOS2064.

Johann Brehmer and Kyle Cranmer. Flows for simultaneous manifold learning and density estimation. In *Advances in Neural Information Processing Systems*, volume 33, pp. 442–453, 2020.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable Causal Discovery from Interventional Data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Yuri Burda, Roger B. Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

Keenan Crane. Keenan's 3D Model Repository. https://www.cs.cmu.edu/~kmcrane/Projects/ModelRepository/, 2021.

Brian Curless and Marc Levoy. A Volumetric Method for Building Complex Models from Range Images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pp. 303–312, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237269.

Cian Eastwood and Christopher K I Williams. A framework for the quantitative evaluation of disentangled representations. *International Conference on Learning Representations*, February 2018.

Frederick Eberhardt. Green and grue causal variables. *Synthese*, 193(4):1029–1046, 2016.

Gerald B. Folland. *Real analysis : modern techniques and their applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, New York, 1999. ISBN 978-0-471-31716-6.

Hermanni Hälvä, Sylvain Le Corff, Luc Lehéricy, Jonathan So, Yongjie Zhu, Elisabeth Gassiat, and Aapo Hyvarinen. Disentangling Identifiable Features from Noisy Data with Structured Nonlinear ICA. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13:411–430, 2000.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016.

Venkat Krishnamurthy and Marc Levoy. Fitting Smooth Surfaces to Dense Polygon Meshes. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '96, pp. 313–324, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917464. doi: 10.1145/237170.237270.

Sebastien Lachapelle, Pau Rodriguez, Rémi Le, Yash Sharma, Katie E Everett, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient Neural Causal Discovery without Acyclicity Constraints. In *International Conference on Learning Representations*, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences. *arXiv preprint arXiv:2202.03169*, 2022b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

Martin Edward Newell. *The Utilization of Procedure Models in Digital Image Synthesis.* PhD thesis, The University of Utah, 1975. AAI7529894.

Judea Pearl. *Causality : models, reasoning, and inference*. Cambridge University Press, Cambridge, U.K. New York, 2000. ISBN 978-0521895606.

Emil Praun, Adam Finkelstein, and Hugues Hoppe. Lapped Textures. In *Proceedings of ACM SIGGRAPH 2000*, pp. 465–470, July 2000.

Emily Riehl. *Category Theory in Context*. Aurora: Dover Modern Math Originals. Dover Publications, 2017. ISBN 9780486820804.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly Supervised Disentanglement with Guarantees. In *International Conference on Learning Representations*, 2020.

Greg Turk and Marc Levoy. Zippered Polygon Meshes from Range Images. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '94, pp. 311–318, New York, NY, USA, 1994. Association for Computing Machinery. ISBN 0897916670. doi: 10.1145/192161.192241.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9593–9602, 2021.

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous Optimization for Structure Learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9492–9503, 2018.

A Identifiability result

A.1 Technical lemmata

We first prove some technical lemmata that we will use in Appendix A.3 to prove Thm. 1. The first formalizes the idea that if two random variables are independent and at least one is absolutely continuous, they almost surely do not take on the same value.

Lemma 1 (Independent variables are almost surely unequal)**.** *Let $X = \mathbb{R}^d$ be a d-dimensional Euclidean space. Define measures P, Q on X, such that P is absolutely continuous with respect to the Lebesgue measure. Define closed, thus measurable, subset $\Delta = \{(x, x) \mid x \in X\} \subset X \times X$. Then $(P \times Q)(\Delta) = 0$.*

Proof. As P is absolutely continous w. r. t. the Legesgue measure, it assigns zero measure to a singleton set. Then applying the Fubini-Tonelli theorem (Folland, 1999, Thms 7.26, 7.27), we get

$$(P \times Q)(\Delta) = \int_X P(\{x\}) dQ(x) = 0 \qquad \square$$

The next result links statistical and functional independence for real-valued variables.

Lemma 2 (Statistical and functional independence)**.** *Consider a smooth map $h : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}$ such that for any $a \in \mathbb{R}^n, z \in \mathbb{R}^m$, $h(a, z, \cdot) =: h_{az}(\cdot)$ is a diffeomorphism onto its image $U_{az} \subseteq \mathbb{R}$. We choose symbols $c = h(a, z, b)$. Write $b = g_{az}(c) = h_{az}^{-1}(c)$. Consider measure P_{AZ} on (a, z) and smooth cumulative density function P_B on b, independent of (a, z), with density p_B. Let $p(c \mid a, z) = (h_{az})_* p_B$ be the push-forward distribution. Then the following statements are equivalent:*

1. *h is constant in a*
2. *$a \perp\!\!\!\perp c \mid z$*
3. *If additionally $a \perp\!\!\!\perp z$, then $a \perp\!\!\!\perp c$*

Proof. *"$(2) \Leftrightarrow (3)$":* is a standard result of conditional independencies.

"$(1) \Rightarrow (2)$": If h is constant in a, then trivially a and c are independent given z.

"$(2) \Rightarrow (1)$": The push-forward CDF and density are (assuming g_{az} is strictly monotonically increasing, otherwise a similar argument holds):

$$P(c \mid a, z) = P_B(g_{az}(c)) \tag{3}$$

$$p(c \mid a, z) = \frac{\partial P(c \mid a, z)}{\partial c} = \frac{\partial P_B(g_{az}(c))}{\partial c} \tag{4}$$

A necessary, but not sufficient, requirement for conditional independence is that locally this density is constant in a. Thus the gradient w. r. t. a is zero. Writing $d(a, z, c) = P_B(g_{az}(c)) \in [0, 1]$, we obtain a set of partial differential equations:

$$0 = \frac{\partial p(c \mid a, z)}{\partial a_i} = \frac{\partial^2 d(a, z, c)}{\partial a_i \, \partial c} \quad \forall i \in \{1, ..., n\} \tag{5}$$

The general solution of the PDE is:

$$d(a, z, c) = u(c, z) - v(a, z) = P_B(b) \tag{6}$$

for smooth functions $u : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}$ and $v : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$. This establishes a diffeomorphism

$$h_{az} : \mathbb{R} \to U_{az} : b \mapsto c = u(\cdot, z)^{-1}(P_B(b) + v(a, z)) \tag{7}$$

if and only if $u(\cdot, z)_{|U_{az}} : U_{az} \to [v(a, z), v(a, z) + 1]$ is a diffeomorphism.

Equation (7) only satisfies the independence requirement if $v(a, z)$ is constant in a. To see this, first consider the case where $u(\cdot, z)$ is the identity. Then the density $p(c \mid a, z)$ is a shifted uniform distribution on range $(v(a, z), v(a, z) + 1)$, where locally, the density does not depend on a. However, unless v is a constant, there exist values a, a' for which there exists a c such that $p(c \mid a, z) = 0, p(c \mid a', z) = 1$ violating independence. Hence, for a and c to be independent given z, v and thus h must be constant in a.

When u is not the identity, the value of c is related by a diffeomorphism to the above case, which preserves statistical independence to a given z. □

A.2 Definitions

Here we define objects and relations that were not formally defined in the main body of the paper, but are necessary to make Thm. 1 precise and to prove it.

We use the following notation:

- $[n] = \{1, ..., n\}$
- $\mathbf{pa}_i^{\mathcal{C}} \subseteq [n]$ the set of parent nodes of node i in graph $\mathcal{G}(\mathcal{C})$.
- $\mathbf{desc}_i^{\mathcal{C}} \subseteq [n]$ the set of descendant nodes of node i in graph $\mathcal{G}(\mathcal{C})$, excluding i itself.
- $\mathbf{anc}_i^{\mathcal{C}} \subseteq [n]$ the set of ancestor nodes of node i in graph $\mathcal{G}(\mathcal{C})$, excluding i itself.
- $\mathbf{nonanc}_i^{\mathcal{C}} = [n] \setminus (\mathbf{anc}_i^{\mathcal{C}} \cup \{i\})$ the set of non-ancestor nodes of node i in graph $\mathcal{G}(\mathcal{C})$, excluding i itself.
- Given measure p on space A and measurable function $f : A \to B$, $f_* p$ is the push-forward measure on B.

We will need to reason about vectors being "equal up to permutation and elementwise reparameterizations". We formalize this in the following definition:

Definition 5 (ψ-diagonal)**.** *Let $\psi : [n] \to [n]$ be a bijection (that is, a permutation). Let $\varphi : \prod_{i=1}^n X_i \to \prod_{i=1}^n Y_i$ be a function between product spaces. Then φ is ψ-diagonal if there exist functions, called components, $\varphi_i : X_i \to Y_{\psi(i)}$ such that $\forall i, \forall x, \varphi(x_1, ..., x_i, ..., x_n)_{\psi(i)} = \varphi_i(x_i)$.*

This lets us define isomorphisms between SCMs:

Definition 6 (Isomorphism of SCMs)**.** *Let $\mathcal{C} = \langle \mathcal{Z}, \mathcal{E}, F, p_{\mathcal{E}} \rangle$ and $\mathcal{C}' = \langle \mathcal{Z}', \mathcal{E}', F', p'_{\mathcal{E}} \rangle$ be SCMs. An isomorphism $\varphi : \mathcal{C} \to \mathcal{C}'$ consists of*

1. *a graph isomorphism $\psi : \mathcal{G}(\mathcal{C}) \to \mathcal{G}(\mathcal{C}')$ that tells us how to identify corresponding variables in the two models and which preserves parents: $\mathbf{pa}_{\psi(i)}^{\mathcal{C}'} = \psi(\mathbf{pa}_i^{\mathcal{C}})$ and*
2. *ψ-diagonal diffeomorphisms for noise and endogenous variables that tell us how to reparameterize them $\varphi_{\mathcal{E}} : \mathcal{E} \to \mathcal{E}'$ and $\varphi_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Z}'$, where $\varphi_{\mathcal{E}}$ must be measure preserving $p_{\mathcal{E}'} = \varphi_{\mathcal{E}*} p_{\mathcal{E}}$.*

The elementwise diffeomorphisms are required to make the following diagrams commute $\forall i, i' = \psi(i)$:

$$\begin{tikzcd} \mathcal{Z}_{\mathbf{pa}_i} \times \mathcal{E}_i \arrow[r, "(\varphi_{\mathcal{Z},\mathbf{pa}_i}, \varphi_{\mathcal{E},i})"] \arrow[d, "f_i"] & \mathcal{Z}'_{\mathbf{pa}'_{i'}} \times \mathcal{E}'_{i'} \arrow[d, "f'_{i'}"] \\ \mathcal{Z}_i \arrow[r, "\varphi_{\mathcal{Z},i}"] & \mathcal{Z}'_{i'} \end{tikzcd} \tag{8}$$

Intuitively, this says that if we apply a causal mechanism f_i and then reparameterize the causal variable i using $\varphi_{\mathcal{Z},i}$, we get the same thing as first reparameterizing the parents and noise variable of variable i, and then applying the causal mechanism $f'_{i'}$.

To reason about interventions, we equip SCMs with intervention distributions in the following definition.

Definition 7 (Intervention structural causal model (ISCM))**.** *An intervention structural causal model (ISCM) is a tuple $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ of*

1. *an acyclic SCM $\mathcal{C} = \langle \mathcal{Z}, \mathcal{E}, F, p_{\mathcal{E}} \rangle$ that admits a faithful distribution, meaning that conditional independence of causal variables z implies d-separation (Pearl, 2000).*
2. *a set $\mathcal{I}$ of interventions on $\mathcal{C}$, where each intervention $(I, (\tilde{f}_i)_{i \in I}) \in \mathcal{I}$ consist of*
   (a) *a subset $I \subset \{1, ..., n\}$ of the causal variables, called the intervention target set, and*
   (b) *for each $i \in I$, a new causal mechanism $\tilde{f}_i : \mathcal{E}_i \to \mathcal{Z}_i$ which replaces the original mechanism and which does not depend on the parents.*
   We define intervention set $\mathcal{I}$ to be atomic if the number of targeted variables is one or zero.
3. *a probability measure $p_{\mathcal{I}}$ over $\mathcal{I}$.*

We can extend the notion of isomorphism from SCMs to ISCMs.

Definition 8 (Isomorphism of ISCMs)**.** *Let $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{D}' = \langle \mathcal{C}', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be ISCMs. An ISCM isomorphism is an SCM isomorphism $\varphi : \mathcal{C} \to \mathcal{C}'$ with underlying graph isomorphism $\psi : \mathcal{G}(\mathcal{C}) \to \mathcal{G}(\mathcal{C}')$ and a ψ-diagonal diffeomorphism $\tilde{\varphi}_{\mathcal{E}} : \mathcal{E} \to \mathcal{E}$ such that*

- *the graph isomorphism ψ induces a bijection of intervention sets*

$$\psi_{\mathcal{I}} : \mathcal{I} \to \mathcal{I}' : (I, (\tilde{f}_i)_{i \in I}) \mapsto (\psi(I), (\tilde{f}'_{i'})_{i' \in \psi(I)})$$

- *for each intervention $(I, (\tilde{f}_i)_{i \in I}) \in \mathcal{I}$, and each intervened on variable $i \in I$, the following diagram commutes:*

$$\begin{tikzcd} \mathcal{E}_i \arrow[r, "\tilde{\varphi}_{\mathcal{E},i}"] \arrow[d, "\tilde{f}_i"] & \mathcal{E}'_{\psi(i)} \arrow[d, "\tilde{f}'_{\psi(i)}"] \\ \mathcal{Z}_i \arrow[r, "\varphi_{\mathcal{Z},i}"] & \mathcal{Z}'_{\psi(i)} \end{tikzcd} \tag{9}$$

- *$\tilde{\varphi}_{\mathcal{E}}$ is measure preserving, i. e. $p_{\mathcal{E}'} = (\tilde{\varphi}_{\mathcal{E}})_* p_{\mathcal{E}}$.*
- *the bijection $\psi_{\mathcal{I}} : \mathcal{I} \to \mathcal{I}'$ preserves the distribution over interventions: $\psi_* p_{\mathcal{I}} = p'_{\mathcal{I}'}$.*

Latent Causal Models (LCMs), defined in Def. 2, add a map to the data space to an ILCM. We can lift ISCM isomorphisms to LCM isomorphisms by requiring that these decoders must respect the ISCM isomorphism.

Definition 9 (Isomorphism of LCMs)**.** *Let $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be LCMs with identical observation space $\mathcal{X} = \mathcal{X}'$. An LCM isomorphism of LCM is an ISCM isomorphism $\varphi : \mathcal{D} \to \mathcal{D}'$ such that the decoders respect the SCM isomorphism, so this diagram must commute:*

$$\begin{tikzcd} \mathcal{Z} \arrow[rr, "\varphi_{\mathcal{Z}}"] \arrow[dr, "g"'] & & \mathcal{Z}' \arrow[dl, "g'"] \\ & \mathcal{X} & \end{tikzcd} \tag{10}$$

Remark 1. By defining objects and isomorphisms, we have defined a groupoid of SCMs, a groupoid of ISCMs and a groupoid of LCMs, as the isomorphisms are composed and inverted in an obvious way.

Definition 10 (Equivalence)**.** *We call two SCMs, ISCMs, or LCMs equivalent if an isomorphism exists between them.*

Informally, two SCMs, ISCMs, or LCMs are equivalent if there is a ψ-diagonal map between their causal variables (i. e. the causal variables are equal up to permutation and elementwise diffeomorphisms), there is a ψ-diagonal map between their noise encodings, and all other structure (decoders, intervention sets, intervention distributions) is compatible with these reparameterizations.

Next, we define the solution function of an SCM or ISCM, which maps from noise variables to causal variables by repeatedly applying the causal mechanisms.

Definition 11 (Solution)**.** *Given an ISCM $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$, the solution function $s : \mathcal{E} \to \mathcal{Z}$ is the*

unique function such that for all $i \in [n]$, the following diagram commutes (Bongers et al., 2021)

$$\begin{tikzcd} & \mathcal{E} \arrow[dl, "(s_{\mathbf{pa}_i}, \mathrm{id}_{\mathcal{E}_i})"'] \arrow[dr, "s_i"] & \\ \mathcal{Z}_{\mathbf{pa}_i} \times \mathcal{E}_i \arrow[rr, "f_i"'] & & \mathcal{Z}_i \end{tikzcd}$$

Or equations, we have that $s(\epsilon)_i = f(\epsilon_i; s(\epsilon_{\mathbf{pa}_i}))$. Similarly, intervention $(I, (\tilde{f}_i)_{i \in I}) \in \mathcal{I}$ yields a solution function $\tilde{s}_I : \mathcal{E} \to \mathcal{Z}$ with the modified causal mechanisms.

For example, with two variables with $z_1 \to z_2$, the solution is given by:

$$s : \mathcal{E} \to \mathcal{Z} : \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \end{pmatrix} \mapsto \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} f_1(\epsilon_1) \\ f_2(\epsilon_2, f_1(\epsilon_1)) \end{pmatrix}.$$

Since we require causal mechanisms to be pointwise diffeomorphic, the solution function is a diffeomorphism as well.

Definition 12 (Noise encoding)**.** *We define the noise-encoding to be the inverse of the (unintervened) solution: $e = s^{-1}(z)$.*

In general, we have that

$$e_i = s^{-1}(z)_i = f_i(\cdot, z_{\mathbf{pa}_i})^{-1}(z_i).$$

Perhaps surprisingly, computing s via f_i requires traversing the causal graph, but computing s^{-1} requires only a single step. The same is observed in autoregressive flows (Kingma et al., 2016).

Pushing the noise distribution of an SCM through the solution function finally gives us the (observable) distribution entailed by an SCM or ISCM. In an ISCM or LCM we can define several other (observable or interventional) distributions.

Definition 13 (Distributions)**.** *Given an LCM $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$, we have the following generative process:*

$$\begin{aligned} & & \epsilon \sim p_{\mathcal{E}}\,, & & z = s(\epsilon)\,, & & x = g(z)\,, & & e = s^{-1}(z) \\ I \sim p_{\mathcal{I}}\,, & & \tilde{\epsilon} \sim \tilde{p}_{\tilde{\mathcal{E}}}(\tilde{\epsilon} \mid \epsilon, I)\,, & & \tilde{z} = \tilde{s}_I(\tilde{\epsilon})\,, & & \tilde{x} = g(\tilde{z})\,, & & \tilde{e} = s^{-1}(\tilde{z})\,. \end{aligned} \tag{11}$$

where $p(\tilde{\epsilon}_i \mid \epsilon_i, i \in I) = p_{\mathcal{E}_i}(\tilde{\epsilon}_i)$ and $p(\tilde{\epsilon}_i \mid \epsilon_i, i \notin I) = \delta(\tilde{\epsilon}_i \mid \epsilon_i)$ is the Dirac measure.

Then we define the following weakly supervised distributions:

- *The weakly supervised noise distribution with interventions: $p_{\mathcal{C}}^{\mathcal{E},\mathcal{I}}(\epsilon, \tilde{\epsilon}, I)$.*
- *The weakly supervised causal distribution with interventions: $p_{\mathcal{C}}^{\mathcal{Z},\mathcal{I}}(z, \tilde{z}, I)$.*
- *The weakly supervised noise encoding distribution with interventions: $p_{\mathcal{C}}^{e,\mathcal{I}}(e, \tilde{e}, I)$.*
- *The weakly supervised observational distribution with interventions: $p_{\mathcal{M}}^{\mathcal{X},\mathcal{I}}(x, \tilde{x}, I)$.*

These distributions are given by appropriate pushforwards of the noise distributions through the transformations in Eq. (11).

By marginalizing over I, we get $p_{\mathcal{C}}^{\mathcal{E}}, p_{\mathcal{C}}^{\mathcal{Z}}, p_{\mathcal{C}}^{e}, p_{\mathcal{M}}^{\mathcal{X}}$ respectively.

Now the distinction between noise variables ϵ and noise *encodings* e becomes apparent. The former are the noise variables used when sampling data from the SCM or ISCM distribution. In particular, $\tilde{\epsilon}$ are the noise variables associated with the changed mechanism after an intervention. On the other hand, the noise encodings e are defined as the inverse solution function s^{-1} applied to causal variables. It is perhaps instructive to think of them as the result of an inference task rather than as part of the generative process. Pre-intervention we have $e = \epsilon$, but post-intervention we have in general $\tilde{e} \equiv s^{-1}(\tilde{z}) \neq \tilde{\epsilon} \equiv \tilde{s}_I^{-1}(\tilde{z})$.

The relationships between all the morphisms can be found in Figure 4.

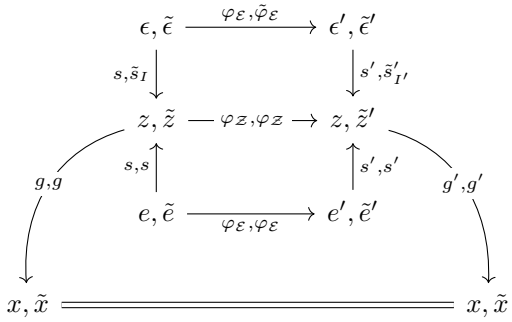

Figure 4: An illustration of the spaces and maps in our definitions and proof. When LCMs $\mathcal{M}, \mathcal{M}'$ are isomorphic, all squares in the diagram should commute. Additionally, all morphisms should preserve the weakly supervised distributions on the variables and all horizontal morphisms should be ψ-diagonal. Note that the latent variables (ϵ, z, e) can differ up to a diffeomorphism, but the x variables are actually observed, so must be identically equal. From that equality, all other horizontal maps are uniquely defined.

A.3 Identifiability proof

We now prove a few lemmata that characterize LCMs and the weakly supervised setting, before finally proving our identifiability result.

Lemma 3 (Behaviour of noise encodings under interventions)**.** *Let $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ be an ISCM with n real-valued causal variables. Let $(e, \tilde{e}, I) \sim p_{\mathcal{C}}^{e,\mathcal{I}}$, then $\forall i, i \in I \implies e_i \perp\!\!\!\perp \tilde{e}_i \mid I$, and almost surely*

$$e_i = \tilde{e}_i \iff i \notin I\,. \tag{12}$$

Proof. If $i \notin I$, then $\tilde{e}_i = f_i(\cdot, \tilde{z}_{\mathbf{pa}_i})^{-1}(\tilde{z}_i) = f_i(\cdot, \tilde{z}_{\mathbf{pa}_i})^{-1}(f_i(\tilde{\epsilon}_i, \tilde{z}_{\mathbf{pa}_i})) = \tilde{\epsilon}_i = \epsilon_i = e_i$.

Otherwise, if $i \in I$, we have that $\epsilon_i \perp\!\!\!\perp \tilde{\epsilon}_i \mid I$ and by pairwise independence of noise variables also $\epsilon_i \perp\!\!\!\perp \tilde{\epsilon} \mid I$. Then

$$\begin{aligned} &\tilde{z}_i = \tilde{f}_i(\tilde{\epsilon}_i) = f_i(\tilde{z}_{\mathbf{pa}_i}, \tilde{e}_i) \\ \implies &\tilde{e}_i = f_i(\tilde{z}_{\mathbf{pa}_i}, \cdot)^{-1}(\tilde{f}_i(\tilde{\epsilon}_i)) = f_i(\tilde{s}_I(\tilde{e})_{\mathbf{pa}_i}, \cdot)^{-1}(\tilde{f}_i(\tilde{\epsilon}_i)). \end{aligned}$$

Therefore $\epsilon_i \perp\!\!\!\perp \tilde{e}_i \mid I$, and as $e_i = \epsilon_i$, also $e_i \perp\!\!\!\perp \tilde{e}_i \mid I$. Because $p_{\mathcal{E},i}(e_i)$ is absolutely continuous, we use Lemma 1 and show that almost surely $e_i \neq \tilde{e}_i$. □

Lemma 4 (Identifiability of noise variables and interventions from weak supervision)**.** *Let $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{D}' = \langle \mathcal{C}', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be ISCMs with n real-valued causal variables, atomic interventions and intervention distributions with full support. Let $\varphi_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Z}'$ be a diffeomorphism such that $p_{\mathcal{C}'}^{\mathcal{Z}'} = (\varphi_{\mathcal{Z}}, \varphi_{\mathcal{Z}})_* p_{\mathcal{C}}^{\mathcal{Z}}$. Then there exists a permutation $\psi : [n] \to [n]$ and a diffeomorphism $\varphi_{\mathcal{E}} : \mathcal{E} \to \mathcal{E}'$ such that:*

- *the weakly supervised noise encoding distribution with intervention is preserved $p_{\mathcal{C}'}^{e',\mathcal{I}'} = (\varphi_{\mathcal{E}}, \varphi_{\mathcal{E}}, \psi)_* p_{\mathcal{C}}^{e,\mathcal{I}}$, and*
- *$\varphi_{\mathcal{E}}$ is a ψ-diagonal diffeomorphism.*

Proof. We prove this similar to Locatello et al. (2020). Define the diffeomorphism $\varphi_{\mathcal{E}} = s'^{-1} \circ \varphi_{\mathcal{Z}} \circ s : \mathcal{E} \to \mathcal{E}'$. Then, as the noise encoding is defined by pushing the causal variables through s^{-1}, we have $p_{\mathcal{C}'}^{e'} = (\varphi_{\mathcal{E}}, \varphi_{\mathcal{E}})_* p_{\mathcal{C}}^{e}$.

Now, for any $(e, \tilde{e}, I) \sim p_{\mathcal{C}}^{e,\mathcal{I}}$, if $I = \{i\}$ for some i, using Lemma 3, there is almost surely exactly one $i' \in [n]$ for which $\varphi_{\mathcal{E}}(e)_{i'} \neq \varphi_{\mathcal{E}}(\tilde{e})_{i'}$. Let $\psi(i) = i'$. As $p_{\mathcal{I}}$ has full support, this fully defines function $\psi : [n] \to [n]$. This is surjective (thus bijective), as otherwise there exists a $j' \in [n] \setminus \psi([n])$ for which almost surely $\varphi_{\mathcal{E}}(e)_{j'} = \varphi_{\mathcal{E}}(\tilde{e})_{j'}$, contradicting the full support of $p_{\mathcal{I}'}$. Since the interventions i (in $\mathcal{D}$) and i' (in $\mathcal{D}'$) must be equally likely, $p_{\mathcal{I}'} = \psi_* p_{\mathcal{I}}$. This proves the first statement.

For the second statement, let $i \neq j \in [n]$ and write $i' = \psi(i)$. Consider $(e, \tilde{e}) \sim p_{\mathcal{C}}^{e,\mathcal{I}}(e, \tilde{e} \mid I = \{j\})$. Then $\varphi_{\mathcal{E}}(e)_{i'} = \varphi_{\mathcal{E}}(\tilde{e})_{i'} = \varphi_{\mathcal{E}}(e_1, ..., e_{j-1}, \tilde{e}_j, e_{j+1}, .., e_n)_{i'}$, with $e_j \neq \tilde{e}_j$, again using Lemma 3. The distribution $(e, \tilde{e}_j)$ has full support, hence $\varphi_{\mathcal{E}}(e)_{i'}$ must be constant in e_j. This holds for all $j \neq i$, so $\varphi_{\mathcal{E}}$ is ψ-diagonal. □

Lemma 5 (Descendancy structure and noise encoding independencies)**.** *Let $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ be an ISCM with n real-valued causal variables, atomic interventions and intervention distributions with full support. Then*

$$z_j \in \mathbf{desc}_i \iff \tilde{e}_j \not\perp\!\!\!\perp e_i \mid I = \{j\}\,. \tag{13}$$

Proof. We can write the solution for variable j as $s_j : \mathcal{E}_j \times \prod_{k \in \mathbf{anc}_j} \mathcal{E}_k \to \mathcal{Z}_j$, which is a diffeomorphism in the first argument for any value of the other arguments. We have that:

$$z_j \in \mathbf{desc}_i \tag{14}$$

$$\overset{1}{\iff} z_j = s_j(e_j; e_{\mathbf{anc}_j}) \text{ not constant in } e_i \tag{15}$$

$$\overset{2}{\iff} \tilde{e}_j = s_j^{-1}(\tilde{z}_j; e_{\mathbf{anc}_j}) \text{ not constant in } e_i \tag{16}$$

$$\overset{3}{\iff} \tilde{e}_j \not\perp\!\!\!\perp e_i \mid I = \{j\}\,. \tag{17}$$

These steps follow from the following arguments:

1. If (14) while not (15), we have a faithfulness violation $z_j \perp\!\!\!\perp z_i \mid z_{\mathbf{pa}_i}$. So assuming faithfulness, (14) $\implies$ (15). The converse, (15) $\implies$ (14), is trivial.
2. Assuming (15), there exist $e_i \neq \bar{e}_i$ for which $s_j(\cdot; e_i, ...) \neq s_j(\cdot; \bar{e}_i, ...) : \mathcal{E}_j \xrightarrow{\sim} \mathcal{Z}_j$ are different diffeomorphisms. A function can not be the inverse of two different function, so the map $s_j^{-1}(\cdot; e_i, ...) : \mathcal{Z}_j \xrightarrow{\sim} \mathcal{E}_j$ can not be constant in e_i, hence (16) holds. The converse is shown similarly.
3. This last step follows from Lemma 2, as $\tilde{z}_j \perp\!\!\!\perp e \mid I = \{j\}$ and the noise variables $e = \epsilon$ are pairwise independent. □

Lemma 6 (Causal variable map ψ-diagonal when descendancy matches)**.** *Let $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{D}' = \langle \mathcal{C}', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be ISCMs with n real-valued causal variables, atomic interventions and intervention distributions with full support. Let $\psi : [n] \to [n]$ be a permutation and $\varphi_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Z}'$ be a diffeomorphism such that $p_{\mathcal{C}'}^{\mathcal{Z}',\mathcal{I}'} = (\varphi_{\mathcal{Z}}, \varphi_{\mathcal{Z}}, \psi)_* p_{\mathcal{C}}^{\mathcal{Z},\mathcal{I}}$ and ψ preserves descendancy: $i \in \mathbf{desc}_j^{\mathcal{C}} \iff \psi(i) \in \mathbf{desc}_{\psi(j)}^{\mathcal{C}'}$. Then $\varphi_{\mathcal{Z}}$ is ψ-diagonal.*

Proof. Let $i \in [n], i' = \psi(i)$. Let $z = (z_{\mathbf{anc}_i}, z_i, z_{\mathbf{nonanc}_i}) \in \mathcal{Z}$. We first prove that $\varphi_{\mathcal{Z}}(z)_{i'}$ is constant in $z_{\mathbf{nonanc}_i}$. For any $j \in \mathbf{nonanc}_i$, consider $\tilde{z} \sim p_{\mathcal{C}}^{\mathcal{Z},\mathcal{I}}(\tilde{z} \mid z, I = \{j\})$. Note that because of acyclicity, $\tilde{z} = (z_{\mathbf{anc}_i}, z_i, \tilde{z}_{\mathbf{nonanc}_i})$. Furthermore, the distribution on $\tilde{z}_{\mathbf{nonanc}_i}$ is supported on a curve, because we re-sample one noise variable $\tilde{e}_j$, and when the curve is projected to the $\tilde{z}_j$-axis, it is surjective, because any value of $\tilde{z}_j$ can be sampled. When the curve is projected to the $\tilde{z}_k$-axis, for a $k \notin \mathbf{desc}_j$, the curve is constant. Also, because the weakly-supervised distribution with intervention and the descendancy structure is preserved, for $z' = \varphi_{\mathcal{Z}}(z), \tilde{z}' = \varphi_{\mathcal{Z}}(\tilde{z})$, we have that $\tilde{z}'_{i'} = z'_{i'}$. Thus, moving $\tilde{z}$ over the curve leaves $\varphi_{\mathcal{Z}}(\tilde{z})_{i'}$ invariant.

Now, given any $\bar{z}_{\mathbf{nonanc}_i} \in \mathcal{Z}_{\mathbf{nonanc}_i}, \bar{z} = (z_{\mathbf{anc}_i}, z_i, \bar{z}_{\mathbf{nonanc}_i})$, we can move $\tilde{z}$ over such curves from z to $\bar{z}$ leaving $\varphi_{\mathcal{Z}}(\tilde{z})_{i'}$ invariant. We do this in a topological ordering of $\mathbf{nonanc}_i$. By the above properties of the intervention curves, when intervening on $k \in \mathbf{nonanc}_i$, we do not affect the previous variables, while able to reach any value of $\tilde{z}_k$, making such a trajectory always possible. Hence, $\varphi_{\mathcal{Z}}(z)_{i'}$ is constant in $z_{\mathbf{nonanc}_i}$.

Next, we prove that $\varphi_{\mathcal{Z}}(z)_{i'}$ is constant in $z_{\mathbf{anc}_i}$. Following the above argument, we know that $\varphi_{\mathcal{Z}}$ has a triangular structure, in that it factorizes into diffeomorphism $f : \mathcal{Z}_{\mathbf{anc}_i} \xrightarrow{\sim} \mathcal{Z}'_{\mathbf{anc}_{i'}}$, and conditional diffeomorphisms $g : \mathcal{Z}_{\mathbf{anc}_i} \to (\mathcal{Z}_i \xrightarrow{\sim} \mathcal{Z}'_{i'})$ and $\mathcal{Z}_{\mathbf{anc}_i} \times \mathcal{Z}_i \to (\mathcal{Z}_{\mathbf{nonanc}_i} \xrightarrow{\sim} \mathcal{Z}'_{\mathbf{nonanc}_{i'}})$. When

intervening on i', we have:

$$
\begin{aligned}
& z' \perp\!\!\!\perp \tilde{z}'_{i'} \mid I' = \{i'\} \\
\implies & z'_{\mathbf{anc}_{i'}} \perp\!\!\!\perp \tilde{z}'_{i'} \mid I' = \{i'\} \\
\overset{1}{\implies} & f(z_{\mathbf{anc}_i}) \perp\!\!\!\perp g(z_{\mathbf{anc}_i}, \tilde{z}_i) \mid I = \{i\} \\
\overset{2}{\implies} & z_{\mathbf{anc}_i} \perp\!\!\!\perp g(z_{\mathbf{anc}_i}, \tilde{z}_i) \mid I = \{i\} \\
\overset{3}{\implies} & g \text{ constant in } z_{\mathbf{anc}_i}
\end{aligned}
$$

where we:

1. pull back the independency through measure preserving diffeomorphism $(\varphi_{\mathcal{Z}}, \varphi_{\mathcal{Z}}, \psi)$
2. use that independency is preserved by diffeomorphisms
3. use Lemma 2, using independence $z_{\mathbf{anc}_i} \perp\!\!\!\perp \tilde{z}_i \mid I = \{i\}$

Thus $\varphi_{\mathcal{Z}}(z)_{i'}$ is also constant in $z_{\mathbf{anc}_i}$, proving the lemma. □

Lemma 7 (ψ-diagonal causal variable map implies equivalence)**.** *Let $\mathcal{D} = \langle \mathcal{C}, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{D}' = \langle \mathcal{C}', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be ISCMs with n real-valued causal variables, atomic interventions and intervention distributions with full support. Let $\psi : [n] \to [n]$ be a permutation and $\varphi_{\mathcal{Z}} : \mathcal{Z} \to \mathcal{Z}'$ a ψ-diagonal diffeomorphism such that $p_{\mathcal{C}'}^{\mathcal{Z}',\mathcal{I}'} = (\varphi_{\mathcal{Z}}, \varphi_{\mathcal{Z}}, \psi)_* p_{\mathcal{C}}^{\mathcal{Z},\mathcal{I}}$ and ψ preserves descendency. Then $\mathcal{D} \sim \mathcal{D}'$.*

Proof. The ψ-diagonal map $\varphi_{\mathcal{Z}}$ forms the map between causal variables, and the map $\varphi_{\mathcal{E}} = s'^{-1} \circ \varphi_{\mathcal{Z}} \circ s$ between noise variables has been shown in Lemma 4 to be ψ-diagonal and preserve the noise distribution $p_{\mathcal{E}}$.

Consider for variable $i, i' = \psi(i)$ the following commuting diagram. Note that we write the causal mechanism f_i as a function of all ancestors, not just the parents, so it is constant in the non-parents. Also, it is non-constant in the parents, because of faithfulness. Because ψ preserves ancestors, $f'_{i'}$ is well-typed.

$$
\begin{tikzcd}
\mathcal{E} \arrow[r, "\varphi_{\mathcal{E}}"] \arrow[d, "(s_{\mathbf{anc}_i}, \mathrm{id}_{\mathcal{E}_i})"] & \mathcal{E}' \arrow[d, "(s'_{\mathbf{anc}_{i'}}, \mathrm{id}_{\mathcal{E}'_{i'}})"] \\
\mathcal{Z}_{\mathbf{anc}_i} \times \mathcal{E}_i \arrow[r, "(\varphi_{\mathcal{Z},\mathbf{anc}_i}, \varphi_{\mathcal{E},i})"] \arrow[d, "f_i"] & \mathcal{Z}'_{\mathbf{anc}_{i'}} \times \mathcal{E}'_{i'} \arrow[d, "f'_{i'}"] \\
\mathcal{Z}_i \arrow[r, "\varphi_{\mathcal{Z},i}"] & \mathcal{Z}'_{i'}
\end{tikzcd}
$$

Because the composition of left and right the vertical morphisms equal s_i and $s'_{i'}$ respectively and because of the definition of $\varphi_{\mathcal{E}}$, the outer and the top square commute. Then, because $(s_{\mathbf{anc}_i}, \mathrm{id}_{\mathcal{E}_i})$ is surjective, the bottom square also commutes (Riehl, 2017, Lemma 1.6.21). Then for $z_j \in \mathbf{anc}_i$, we have that

$$
z_j \in \mathbf{pa}_i^{\mathcal{C}} \iff f_i \text{ not constant in } z_j \iff f'_{i'} \text{ not constant in } z'_{j'} \iff z'_{j'} \in \mathbf{pa}_{i'}^{\mathcal{C}'} .
$$

And thus ψ not only preserves descendancy, but also parenthood and is thus is a graph isomorphism $\psi : \mathcal{G}(\mathcal{C}) \to \mathcal{G}(\mathcal{C}')$. Diagram (8) commutes and we have established SCM isomorphism $\varphi : \mathcal{C} \to \mathcal{C}'$. To have this also be an ISCM isomorphism, we need diagram (9) to commute and the distribution over interventions to be preserved. For the first, use the fact that all morphisms in (9) are isomorphisms to simply define $\tilde{\varphi}_{\mathcal{E}}$ so that the diagram commutes. The second follows directly from the assumptions. Hence $\varphi : \mathcal{D} \to \mathcal{D}$ is an ISCM isomorphism and thus $\mathcal{D} \sim \mathcal{D}'$. □

We now have all the pieces to prove Thm. 1, following the three steps laid out at the end of Sec. 2.

Theorem 1 (Identifiability of $\mathbb{R}$-valued LCMs from weak supervision)**.** *Let $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ and $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$ be LCMs with the following properties:*

- *The SCMs $\mathcal{C}$ and $\mathcal{C}'$ both consist of n real-valued endogeneous variables, i. e. $\mathcal{E}_i = \mathcal{Z}_i = \mathcal{Z}'_i = \mathcal{E}'_i = \mathbb{R}$.*
- *The intervention sets $\mathcal{I}$ and $\mathcal{I}'$ consist of the empty intervention and all atomic interventions, $\mathcal{I} = \{\emptyset, \{z_0\}, \dots, \{z_n\}\}$ and similar for $\mathcal{I}'$.*

- *The intervention distribution $p_{\mathcal{I}}$ and $p'_{\mathcal{I}'}$ have full support.*

Then the following two statements are equivalent:

1. *The weakly supervised distributions entailed by the LCMs are equal, $p_{\mathcal{M}}(x, \tilde{x}) = p_{\mathcal{M}'}(x, \tilde{x})$.*
2. *The LCMs are equivalent, $\mathcal{M} \sim \mathcal{M}'$.*

Proof. "(2) $\Rightarrow$ (1)": If the LCMs are equivalent, then the fact that $\varphi_{\mathcal{E}}$ and $\tilde{\varphi}_{\mathcal{E}}$ are measure preserving and that diagrams (8) and (9) commute, implies that $p^{\mathcal{Z}'}_{\mathcal{C}'} = (\varphi_{\mathcal{Z}}, \varphi_{\mathcal{Z}})_* p^{\mathcal{Z}}_{\mathcal{C}}$. Then because diagram (10) commutes, the weakly supervised distributions coincide $p^{\mathcal{X}}_{\mathcal{M}'} = p^{\mathcal{X}}_{\mathcal{M}}$.

"(1) $\Rightarrow$ (2)": Conversely, if the weakly supervised distributions coincide, $p^{\mathcal{X}}_{\mathcal{M}'} = p^{\mathcal{X}}_{\mathcal{M}}$, the images of $g : \mathcal{Z} \to \mathcal{X}, g' : \mathcal{Z}' \to \mathcal{X}$ coincide and the diffeomorphism $\varphi_{\mathcal{Z}} = g'^{-1} \circ g : \mathcal{Z} \to \mathcal{Z}'$ preserves the weakly supervised distribution over causal variables: $p^{\mathcal{Z}'}_{\mathcal{C}'} = (\varphi_{\mathcal{Z}}, \varphi_{\mathcal{Z}})_* p^{\mathcal{Z}}_{\mathcal{C}}$. Then, following Lemma 4, we can find the permutation ψ and the ψ-diagonal noise morphism $\varphi_{\mathcal{E}}$ which preserve the weakly supervised distribution over noise encodings with intervention $p^{e',\mathcal{I}'}_{\mathcal{C}'} = (\varphi_{\mathcal{E}}, \varphi_{\mathcal{E}}, \psi)_* p^{e,\mathcal{I}}_{\mathcal{C}}$. Conditional independencies are preserved under this push-forward, so

$$\tilde{e}_j \not\perp\!\!\!\perp e_i \mid I = \{j\} \iff \tilde{e}'_{\psi(j)} \not\perp\!\!\!\perp e'_{\psi(i)} \mid I = \{\psi(j)\}\,.$$

Then using Lemma 5, ψ preserves descendancy: $i \in \mathbf{desc}^{\mathcal{C}}_j \iff \psi(i) \in \mathbf{desc}^{\mathcal{C}'}_{\psi(j)}$. Using Lemma 6, the diffeomorphism $\varphi_{\mathcal{Z}}$ is ψ-diagonal. Finally, we use Lemma 7 to show that this implies equivalence. □

A.4 Limitations & generalization

Our identifiability result relies on a few assumptions. Here we discuss some key requirements of Thm. 1 and whether they can be relaxed.

Diffeomorphic causal mechanisms In Def. 1, we require causal mechanisms to be pointwise diffeomorphisms from noise variables to causal variables. Under some mild smoothness assumptions, any SCM can be brought into this form by elementwise redefinitions of the variables, without affecting the observational or interventional distributions. (However, such a redefinition may change counterfactual distributions.)

Perfect interventions Our proof of Thm. 1 requires perfect interventions, i. e. intervened-upon mechanisms not depending on any causal variables. This is arguably the biggest mismatch between our asumptions and many real-world systems.

Diffeomorphic decoder Definition 2 and Thm. 1 assume that the map from causal variables to observed data is given by a deterministic, diffeomorphic decoder. However, our practical implementation in a VAE uses a stochastic decoder and allows for noisy data. Our experiments provide empirical evidence for identifiability in this setting. We believe that it may be possible to extend Thm. 1 to stochastic decoders, similarly to Khemakhem et al. (2020). We plan to study this extension in future work.

Real-valued causal variables Theorem 1 assumes real-valued causal and noise variables, $\mathcal{Z}_i = \mathcal{E}_i = \mathbb{R}$. We can easily extend this to intervals $(a, b) \in \mathbb{R}$, as these are isomorphic to $\mathbb{R}$. However, the extension to arbitrary continuous spaces or $\mathbb{R}^n$ is not straightforward. The main reason is that our proof relies on Lemma 2 to link functional independence and statistical independence, which does not generalize.

Let us provide an counterexample for identifiability with S^1-valued causal variables.

Example 1 (S^1-valued non-identifiable LCMs)**.** *Consider an LCM $\mathcal{M} = \langle \mathcal{C}, \mathcal{X}, g, \mathcal{I}, p_{\mathcal{I}} \rangle$ with the following components:*

- *The SCM $\mathcal{C}$ consists of two circle-valued variables $z_1, z_2 \in S^1$ with noise variables $\epsilon_1, \epsilon_2 \in S^1$. We parameterize S^1 as $[0, 2\pi)$ with addition defined modulo 2π.*
- *The causal graph is $z_1 \to z_2$.*
- *The causal mechanisms are $f_1(\epsilon_1) = \epsilon_1$ and $f_2(\epsilon_2; z_1) = \epsilon_2 + z_1$.*

- *The noise variables are distributed as* $\epsilon_1 \sim \text{Uniform}$ *and* $\epsilon_2 \sim q$*, which we require to not be invariant under translations (so in particular not uniform). For example, one can take the von Mises distribution* $\log q(\epsilon_2) = \cos(\epsilon_2) + \text{const}$.
- *The observation space is* $\mathcal{X}$ *and the decoder* $g : S^1 \times S^1 \to \mathcal{X}$ *is diffeomorphic.*
- *The intervention set* $\mathcal{I}$ *consists of the empty intervention, atomic interventions on* z_1 *with* $\tilde{z}_1 \sim \text{Uniform}$*, and atomic interventions on* z_2 *with* $\tilde{z}_2 \sim \text{Uniform}$*. Each of these interventions has probability* $\frac{1}{3}$ *in* $p_{\mathcal{I}}$*.*

Note that the SCM is faithful, as $z_1 \not\perp\!\!\!\perp z_2$ *in the observational distribution, because* q *is not translationally invariant. The LCM entails the weakly supervised causal distribution*

$$p^{\mathcal{Z}}_{\mathcal{C}}(z, \tilde{z}) = \text{Uniform}(z_1)\, q(z_2 - z_1) \left[\frac{1}{3}\, \delta(\tilde{z}_1 - z_1)\, \delta(\tilde{z}_2 - z_2) + \frac{1}{3}\, \text{Uniform}(\tilde{z}_1)\, \delta(\tilde{z}_2 - z_2) + \frac{1}{3}\, \delta(\tilde{z}_1 - z_1)\, \text{Uniform}(\tilde{z}_2)\right] \quad (18)$$

with Dirac delta δ*. The corresponding weakly supervised noise-encoding distribution is*

$$p^{e}_{\mathcal{C}}(e, \tilde{e}) = \text{Uniform}(e_1)\, q(e_2) \left[\frac{1}{3}\, \delta(\tilde{e}_1 - e_1)\, \delta(\tilde{e}_2 - e_2) + \frac{1}{3}\, \text{Uniform}(\tilde{e}_1)\, \delta(\tilde{e}_2 - e_2) + \frac{1}{3}\, \delta(\tilde{e}_1 - e_1)\, \text{Uniform}(\tilde{e}_2)\right], \quad (19)$$

where we have used the fact that the uniform density on the cirlce is invariant to translations, and thus $\text{Uniform}(\tilde{e}_2) = \text{Uniform}(\tilde{e}_1 + \tilde{e}_2)$*. The weakly supervised data distribution is then given by* $p^{\mathcal{X}}_{\mathcal{M}} = (g_*, g_*) \circ (s_*, s_*) p^{\mathcal{E}}_{\mathcal{C}}$*.*

Now consider a second LCM $\mathcal{M}' = \langle \mathcal{C}', \mathcal{X}, g', \mathcal{I}', p'_{\mathcal{I}'} \rangle$*:*

- *The SCM* $\mathcal{C}'$ *consists of two circle-valued variables* $z'_1, z'_2 \in S^1$ *with noise variables* $\epsilon'_1, \epsilon'_2 \in S^1$*.*
- *The causal graph is trivial and the causal mechanisms are given by the identity,* $f'_i(\epsilon'_i) = \epsilon'_i$*.*
- *The noise variables are distributed as* $\epsilon'_1 \sim \text{Uniform}$ *and* $\epsilon'_2 \sim q$*.*
- *The observation space is* $\mathcal{X}$ *and the decoder* $g' : S^1 \times S^1 \to \mathcal{X}$ *is given by the diffeomorphism* $g'(z') = g \circ s(z')$*, where* s *is the solution function of* $\mathcal{C}$*.*
- *The intervention set* $\mathcal{I}'$ *consists of empty interventions, atomic interventions on* z'_1 *with* $\tilde{z}'_1 \sim \text{Uniform}$*, and atomic interventions on* z'_2 *with* $\tilde{z}'_2 \sim \text{Uniform}$*. Each of these interventions has probability* $\frac{1}{3}$ *in* $p_{\mathcal{I}'}$*.*

We find a weakly supervised causal distribution

$$p^{\mathcal{Z}'}_{\mathcal{C}'}(z', \tilde{z}') = \text{Uniform}(z'_1)\, q(z'_2) \left[\frac{1}{3}\, \delta(\tilde{z}'_1 - z'_1)\, \delta(\tilde{z}'_2 - z'_2) + \frac{1}{3}\, \text{Uniform}(\tilde{z}'_1)\, \delta(\tilde{z}'_2 - z'_2) + \frac{1}{3}\, \delta(\tilde{z}'_1 - z'_1)\, \text{Uniform}(\tilde{z}'_2)\right], \quad (20)$$

which is equal to $p^{e}_{\mathcal{C}}$*. For the weakly-supervised data distribution we thus find*

$$p^{\mathcal{X}}_{\mathcal{M}'} = (g'_*, g'_*) p^{\mathcal{Z}'}_{\mathcal{C}'} = (g_*, g_*) \circ (s_*, s_*) p^{e}_{\mathcal{C}} = p^{\mathcal{X}}_{\mathcal{M}}\,. \quad (21)$$

So these two models with their non-isomorphic graph structures have identical weakly-supervised distributions. They therefore provide a counter-example for a straightforward generalization of Thm. 1 to causal variables with arbitrary continuous domains.

Why does identifiability fail in this example? It is because the interventional distributions in $\mathcal{M}$ have an accidental symmetry not expected by the graph structure, namely translational invariance. This makes it possible to fit the weakly supervised distribution with a simpler causal graph. This is related to faithfulness, but the standard definition of faithfulness only concerns observational distributions (and in this sense both $\mathcal{M}$ and $\mathcal{M}'$ are faithful). Because of this accidental symmetry, Lemma 5 does not hold. The lemma fails in step 3, because $\tilde{e}_2 = \tilde{z}_2 - \epsilon_1$, with $\tilde{z}_2, \epsilon_1 \sim \text{Uniform}$, and thus $\tilde{e}_2 \perp\!\!\!\perp e_1 \mid I = \{2\}$.

Table 2: Experiment results. For each experiment, we show the true causal graph underlying the data-generating process. We then show the results from our LCMs and compare to unstructured β-VAE and disentanglement VAE (dVAE) baselines. We show the learned causal graph, the structural Hamming distance SHD between the learned and the true graph, the DCI disentanglement score (D), and the intervention negative log posterior ($-\log p_I$). Best results in bold.

Dataset	True graph	Method	Learned graph	SHD	D	$-\log p_I$
2D toy data		LCM		**0**	**0.99**	**0.28**
		dVAE	–	n/a	0.35	0.33
		β-VAE	–	n/a	0.00	n/a
Causal3DIdent		LCM		**0**	**1.00**	**0.16**
		dVAE	–	n/a	**1.00**	**0.16**
		β-VAE	–	n/a	0.45	n/a
		LCM		1	**0.96**	0.22
		dVAE	–	n/a	0.84	**0.21**
		β-VAE	–	n/a	0.89	n/a
		LCM				
		dVAE		n/a[3]		
		β-VAE				
		LCM		**0**	**0.98**	**0.29**
		dVAE	–	n/a	0.33	4.42
		β-VAE	–	n/a	0.26	n/a
		LCM		**0**	**0.98**	**0.21**
		dVAE	–	n/a	0.62	0.24
		β-VAE	–	n/a	0.35	n/a
		LCM		**0**	**0.96**	**0.18**
		dVAE	–	n/a	0.32	4.06
		β-VAE	–	n/a	0.24	n/a
	Average	LCM		**0.2**	**0.98**	**0.21**
		dVAE		n/a	0.62	1.82
		β-VAE		n/a	0.44	n/a

We have circumvented such issues in Thm. 1 by requiring that all causal and noise variables are $\mathbb{R}$-valued. In this setting, functional dependence implies statistical dependence, as shown by Lemma 2, and the counterexample does not work. We conjecture that it is possible to generalize Thm. 1 to arbitrary continuous domains under mild additional assumptions, but leave this for future work.

Finally, we believe that such accidental symmetries are unlikely in the sense that under an appropriate measure over LCMs, non-identifiable LCMs have have zero measure. We find it likely that this issue will not occur frequently in practical systems (unless these are finetuned to exhibit exactly this behaviour). Overall, we conjecture that identifiability from weak supervision can be generalized beyond the real-valued case presented in Thm. 1.

B Experiment Details

B.1 2D toy experiment

In our first experiment, we generate latent data in $\mathcal{Z} = \mathbb{R}^2$ from a nonlinear SCM with graph $z_1 \to z_2$. In particular, we have that $z_1 \sim \mathcal{N}(z_1; 0, 1^2)$ and $z_2 \sim \mathcal{N}(z_1; 0.3z_1^2 + 0.6z_1, 0.8^2)$. This latent data

[3] The experiments on this subset did not finish in time for the submission. We will add them in a later version and currently report the average over the other five subsets.

is mapped through the data space with a randomly initialized coupling flow with five affine coupling layers interspersed with random permutations of the dimensions. For the weakly supervised setting we use a uniform intervention prior over $\{\emptyset, \{z_1\}, \{z_2\}\}$. We generate 10^5 training samples, 10^5 additional training samples for the models used to compute the DCI metrics, 10^4 validation samples, and 10^4 evaluation samples.

The learned LCM consists of an SCM prior, an encoder, and a decoder. In the SCM, the graph is fixed (we "learn" the graph by training multiple LCMs with different fixed graphs and then selecting the model with the best validation loss). Each causal mechanism is implemented as an MLP of the parents that outputs the parameters of an affine transformation from a standard normal noise variable to a causal variable. The encoder and decoder are diagonal Gaussians, with mean and standard deviations output by an MLP. For each MLP, we use two hidden layers with 100 units each and ReLU activations.

The disentanglement VAE baseline uses the same setup, except with a trivial graph. The β-VAE uses the same encoder and decoder, but uses a standard normal prior rather than an SCM and treats x and $\tilde{x}$ as two i. i. d. samples from the same distribution.

All models are trained on the VAE loss in Eq. (2) plus a regularization term (0.01 times the number of edges in the graph). We train for 10^5 steps using the Adam optimizer Kingma & Ba (2015) with a batch size of 100. The learning rate is initially 10^{-3} and is annealed with a cosine schedule. We estimate the model log likelihood using importance weighting with importance sampling (a la IWAE Burda et al. (2016)), using 10 samples at validation time and 100 samples at test time. We use a manifold "fuzziness" of $\sigma = 0.1$. For each method and each graph, we train three models with different random seeds and select the model with the best validation log likelihood.

B.2 CAUSAL3DIDENT

In the Causal3DIdent experiments we consider six different datasets, each generated from a different causal graph, SCM, and decoder. The six causal graphs we consider are:

- the trivial graph ,
- single edge ,
- the chain ,
- the fork ,
- the collider , and
- the full graph .

For each of these subsets, we randomly generate a nonlinear SCM with heteroskedastic noise: for each causal mechanism, we randomly initialize an MLP that outputs the scale and shift of an affine transformation as a function of the causal parents. We choose an MLP initialization scheme that emphasizes nontrivial, nonlinear causal effects. We then identify a random permutation of the three causal variables with three high-level concepts in the Causal3DIdent dataset: the object hue, the spotlight hue, and the spotlight position. We use the following causal graphs:

- single edge: object hue $\rightarrow$ spotlight position;
- chain: spotlight position $\rightarrow$ spotlight hue $\rightarrow$ object hue;
- fork: spotlight hue $\rightarrow$ spotlight position, object hue;
- collider: spotlight hue $\rightarrow$ object hue $\leftarrow$ spotlight position;
- full graph: spotlight hue $\rightarrow$ object hue $\rightarrow$ spotlight position, spotlight hue $\rightarrow$ spotlight position.

Since all of these properties are defined on a range $[0, 2\pi)$, we apply an elementwise $\operatorname{arctanh}$ transform and rescaling to our variables such that they populate a subset of $[0, 2\pi)$. This also avoids topological issues. Next, we generate images in 64×64 resolution following the procedure described in von Kügelgen et al. (2021). We use Blender (Blender Online Community, 2021) to generate 3D rendered images based on the previously defined causal variables. To increase diversity of the six datasets, we render each dataset with a different object: Teapot (Newell, 1975), Armadillo (Krishnamurthy & Levoy, 1996), Hare (Turk & Levoy, 1994), Cow (Crane, 2021), Dragon (Curless & Levoy,

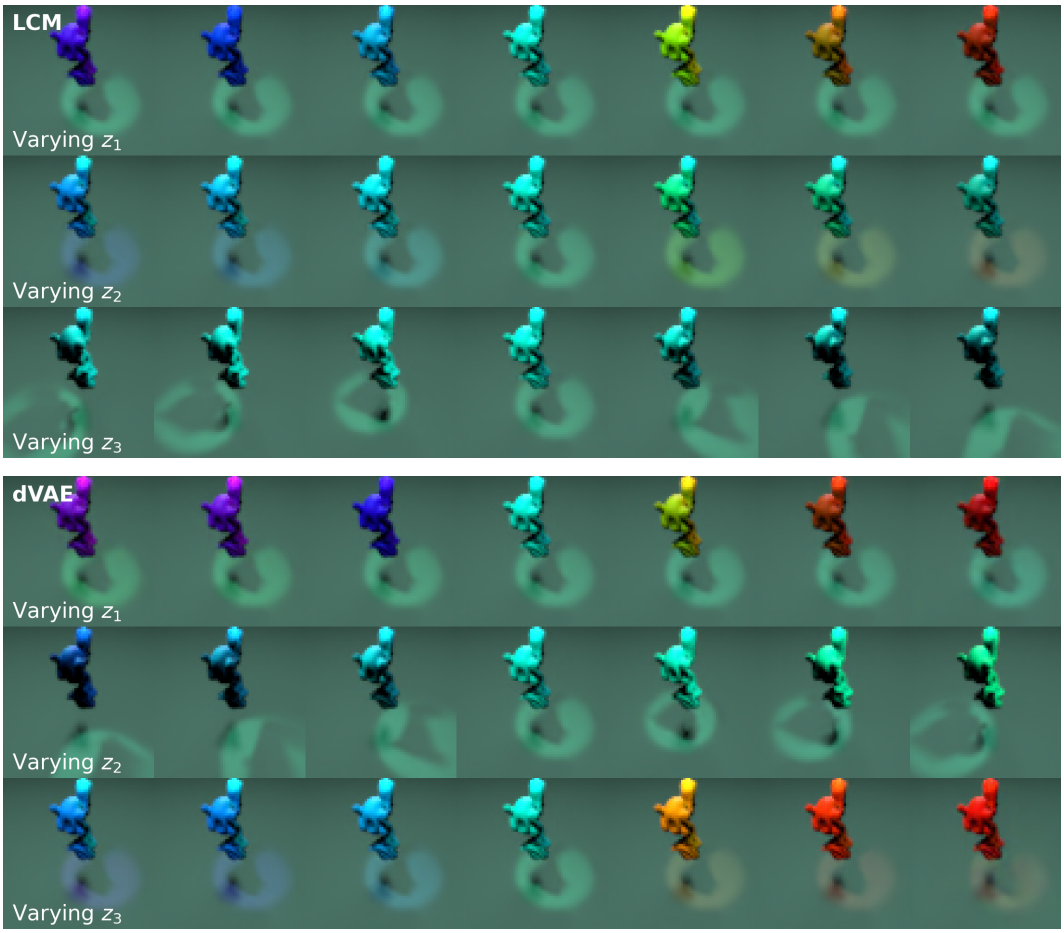

Figure 5: Effect of varying the learned causal factors on the image in the "collider" subset of the Causal3DIdent dataset. We encode a single test images (middle column) into the three learned latents, vary each of these causal factors independently, and show the reconstructed images. The LCM (top) learns a representation that is quite disentangled: z_1 corresponds to the object color, z_2 to the spotlight hue, and z_3 to the spotlight position. In contrast, the acausal dVAE baseline strongly entangles these factors in its learned representation.

1996), and Horse (Praun et al., 2000). We generate 10^5 training samples, 10^4 validation samples, and 10^4 evaluation samples.

The learned LCMs consist again of an SCM prior, which is the same as in the 2D toy experiment, an encoder, and a decoder. For the encoder and decoder we use a convolutional architecture with four residual blocks, using downsampling via average-pooling and bilinear upsampling, respectively. We do not use BatchNorm, as we found that that can lead to practical issues when images in a batch are very similar.

Our training setup is as in the 2D toy experiment, except that we use a batch size of 64, train for $2.3 \cdot 10^5$ steps, and use an initial learning rate of $3 \cdot 10^{-4}$. We find it beneficial to begin training with a lower weight of the KL divergence in the VAE loss, $\beta = 0.01$, and increasing this until the final value of $\beta = 0.1$ during the first half of training. We initialize the manifold "fuzziness" parameter σ to 0.2 and anneal it to 0.01 over the first half of training. For each method and each graph, we train three models with different random seeds and select the model with the best validation log likelihood.[4]

We report our results in Tbl. 2 and visualize the disentanglement properties of the learned representations in Fig. 5.

[4]At the time of submission, for all subsets, for all models, the runs of only one or two of the three random seed have completed and their results used.