A Causal Ordering Prior for Unsupervised Representation Learning

A Causal Ordering Prior for Unsupervised Representation Learning

Avinash Kori *,‡
a.kori21@ic.ac.uk

Pedro Sanchez *,†
pedro.sanchez@ed.ac.uk

Konstantinos Vilouras †
konstantinos.vilouras@ed.ac.uk

Ben Glocker ‡
b.glocker@ic.ac.uk

Sotirios A. Tsaftaris †
s.tsaftaris@ed.ac.uk

* **Joint first authors**
† School of Engineering, University of Edinburgh
‡ Department of Computing, Imperial College London

Abstract

Unsupervised representation learning with variational inference relies heavily on independence assumptions over latent variables. Causal representation learning (CRL), however, argues that factors of variation in a dataset are, in fact, causally related. Allowing latent variables to be correlated, as a consequence of causal relationships, is more realistic and generalisable. So far, provably identifiable methods rely on: auxiliary information, weak labels, and interventional or even counterfactual data. Inspired by causal discovery with functional causal models, we propose a fully unsupervised representation learning method that considers a data generation process with a latent additive noise model (ANM). We encourage the latent space to follow a causal ordering via loss function based on the Hessian of the latent distribution.

1 Introduction

The objective of extracting meaningful representations from unlabelled data is a longstanding pursuit in the field of deep learning [2]. Conventionally, methods of unsupervised representation learning have concentrated on unveiling statistically independent latent variables [9, 5, 41, 27, 10], demonstrating appreciable success in synthetic benchmarks and datasets where generation parameters can be carefully manipulated [28]. However, it is essential to acknowledge the differences between controlled environments and real-world scenarios. In the latter, the factors contributing to data variation are often intertwined within causal relationships. Therefore, it is not merely advantageous but imperative to integrate causal understanding into the process of learning representations [39], which can improve the models from a generalisation, and interpretability, viewpoint.

The main challenge in learning meaningful and disentangled latent representations is identifiability, i.e. ensuring the true distribution of a data generation process can be learned (up to a simple transformation, given the inherent limitation that we can never observe the hidden latent factors from observational data alone), implying the model to be injective (one-to-one mapping) onto the observed distribution. Identifiability ensures that if an estimation method perfectly fits the data distribution, the learned parameters will correspond to the true generative model. For example, discovering independent sources of variation which are observed via a nonlinear mixing function is impossible [13]. This established result from the nonlinear ICA literature has been replicated for disentangled representation learning with variational autoencoders [28].

Preprint. Under review.

Representation learning becomes identifiable when non-i.i.d. (independent and identically distributed) samples from a given data generation process are considered [19, 14]. For instance, temporal contrastive learning [12] and iVAE [19] can provably ensure identifiability by utilising knowledge of auxiliary information. Indeed, [19] develops a comprehensive proof that generative models become identifiable when variables in the latent space are conditionally independent, given the auxiliary information. Conditional independence given external information allows variables to be dependent (or correlated) [20], which is more realistic. Further reinforcing the notion of dependence between latent variables, the identifiability of unsupervised representations can be proven by assuming a latent space to follow a Gaussian Mixture Model (GMM) and an injective decoder [23]. Any distribution can be approximated by a mixture model with sufficiently many components, including distributions following a causal model. In fact, [23] assumes that latent variables are conditionally independent, given a component of the mixture model. The mixture component can correspond to using a "learned" auxiliary variable [44], bridging the gap with [19].

These works [12, 19, 20, 44, 14] on identifiable representation learning from observational data do not consider latent causal structure. They build up, however, a theory around identifiable representation learning which allows arbitrary distribution encoding statistical dependencies in latent variables. Discovering the dependency structure in the latent space is at the core of causal representation learning (CRL) [39] via the *common cause principle*[1] [36]. Learning causally related variables enable (i) robustness to distribution shifts via the independent causal mechanism (ICM) principle; (ii) better generalisation, e.g. in transfer learning settings; (iii) answering causal queries, i.e. estimation of interventional and counterfactual distributions. Previous work on CRL, however, utilise data from interventional [1, 42] or counterfactual (pre- and post-intervention) [29, 3, 26] distributions for learning identifiable causal representations.

In this work, we bridge the gap between identifiable representation learning from observational data and CRL by using functional constraints (which are very common in the causal discovery [33] literature). We propose the first (to the best of our knowledge) method for unsupervised CRL under some data and model assumptions. This can be done by assuming a data generation process in which the latent space adheres to an additive noise model (ANM) and applies an injective nonlinear mapping to generate observational data. The main **contributions** in this work include (i) Based on the universal approximation capabilities of GMMs, we show that models with a latent ANM prior are identifiable to block diagonal transformation; and (ii) We propose an estimation method that encourages the latent space to follow an ANM by leveraging asymmetries in the learned latent distribution. More specifically, the latent distribution's second-order derivatives (Hessian) can be incorporated into a loss function that promotes latent ordering. We term models trained with the proposed estimation method as COVAE (causally ordered Variational AutoEncoders).

2 Related Works

Disentangled Representation Learning. Early efforts on unsupervised representation learning focused on the Variational Autoencoder framework [22]. β-VAE [9] and extensions [21, 6, 30] rely on independence assumptions between latent variables to learn disentangled representations [27, 10]. Despite showing some success, there is a lack of theory around the identifiability of independent representations. In fact, learning independent (disentangled) representations from i.i.d. data in an unsupervised manner is provably impossible [13, 28].

Representation Learning with Auxiliary Information. A line of work based on nonlinear ICA leverages auxiliary information to learn identifiable models. [19] derive a more general proof of identifiability using the concept of conditional independence given auxiliary variables. An extension of nonlinear ICA, called Independently Modulated Component Analysis (IMCA) was proposed in [20], where the components are allowed to be dependent. On the contrary, [23] prove that identifiability of deep generative models can also be achieved without auxiliary information by considering a GMM prior in the latent space. In the same line, empirical results in [44] show that the GMM prior assumption is as efficient as utilising auxiliary information in terms of learning stability (latents learned for different training seeds are correlated).

[1] "If two observables X and Y are statistically dependent, then there exists a variable Z that causally influences both and explains all the dependence in the sense of making them independent when conditioned on Z. As a special case, Z can coincide with X or Y."

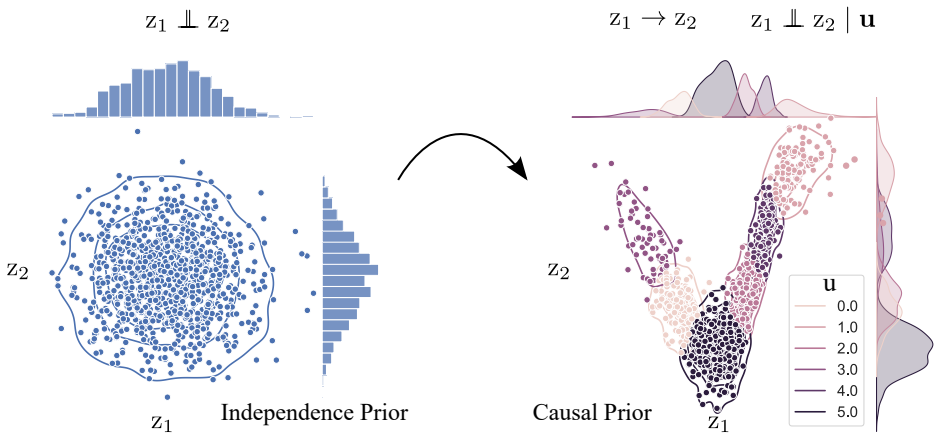


Figure 1: [Left] Independence assumption used in previous work for disentangled representations such as $beta$-VAE and extensions. [Right] We propose to model causally related latent variables. CRL is made possible by using a mixture model in the latent space which approximates an structural causal model (SCM) with functional constraints. z_1, z_2 are latent variables, and $\mathbf{u}$ correspond to mixture components.

Causal Representation Learning. Following the *common cause principle* [36], causal relationships between variables also imply statistical dependencies. Recent works have shown that it is possible to model causal relationships given access to either interventional or non-i.i.d. data. To this end, the method in [1] uses an injective polynomial decoder and the overall model is trained on both observational and interventional data. Similarly, [42] consider the case of an injective linear decoder and directly optimize the score function of the distribution (in both the latent and observation space). In [29] a setting where observations are collected before and after unknown interventions (i.e. counterfactual data) is introduced, while [3] extends this idea to causal graphs of higher complexity. Under the non-iid scenario, [26] focuses on extracting causal factors from spatiotemporal data by performing interventions across different time steps. There also exist works that assume some level of supervision, i.e. having access to ground-truth causal factors. [40] propose a method based on the GAN framework where the prior follows a nonlinear Structural Causal Model (SCM). Others [45] instead model exogenous noise directly, which is then mapped to causal latent variables via a linear SCM. Table 1 describes data and latent space assumptions of previously existing models in comparison to the proposed method.

Table 1: Comparison of assumptions for identifiability proofs. We classify methods by type of training data: observational (*obs*), interventional (*int*) or counterfactual (*count*); and latent assumptions: independent (*ind*), conditionally independent (*cond ind*), with auxiliary information (*aux*) or structural causal model (*SCM*).

Method	Data	Latents
ADA-GVAE [29]	count	ind
IVAE [19]	obs + aux	cond ind $\mid$ aux
VADE [16, 44], MFC-VAE [7, 23]	obs	cond ind $\mid$ learned aux
CAUSALVAE [45], DEAR [40]	obs + aux	SCM
[1], [42]	int	SCM
ILCM [3], CITRIS [26]	count	SCM
Ours (COVAE)	obs	SCM (ANM)

3 Identifiability of Latent Additive Noise Models

A key challenge in unsupervised representation learning is identifiability. The intuition is that if two parameters result in an identical distribution of observations, then they must be equivalent in order to ensure model identifiability. Note that identifiability is the property of the data generation process,

and not of the estimation method. Model identifiability is important because it gives theoretical guarantees that an estimation method is capable of learning the true variables that generated the observed data. Therefore, we first define our model assumptions, show identifiability results and leave the description of the estimation method for the next section. In this section, we define *and* distinguish between the different forms of identifiability and theoretically show that stronger forms of identifiability can be guaranteed when the latent variables are causally ordered.

3.1 Preliminaries

We assume the data generation process maps a latent space $\mathbf{z}$, following a structural causal model (SCM), to an observational space $\mathbf{x}$ as

$$\mathbf{x} = \mathbf{f}_o(\mathbf{z}) + \epsilon_x, \qquad \mathbb{P}(\mathbf{z}) = \prod_i \mathbb{P}(\mathrm{z}_i \mid \mathbf{pa}(\mathrm{z}_i)). \tag{1}$$

$\mathbf{f}_o : \mathbb{R}^d \to \mathbb{R}^o$ is a non-linear injective mapping (or mixing function), d is the number of latent variables and $o = |\mathcal{O}| \geq d$. $\mathbb{P}(\mathbf{z})$ is a distribution entailed by a SCM following a directed acyclic graph (DAG) $\mathcal{G}$, containing d nodes, which describes the true causal structure of the latent. $\mathbf{pa}(\mathrm{z}_i)$ are the parents of z_i in $\mathcal{G}$.

Additive Noise Models. We assume that the latent SCM consists of a collection of assignments following an additive noise model (ANM) $\mathrm{z}_i \coloneqq f_i(\mathbf{pa}(\mathrm{z}_i)) + \epsilon_i$. ϵ_i is a noise term independent of x_i, also called exogenous noise. ϵ_i are i.i.d. from a smooth distribution $\mathbb{P}^\epsilon$. When using an ANM assumption over $\mathbf{z}$, the latent distribution in 1 becomes

$$\mathbb{P}(\mathrm{z}) = \prod_i \mathbb{P}(\mathrm{z}_i \mid \mathbf{pa}(\mathrm{z}_i)) = \prod_i \mathbb{P}^\epsilon(\mathrm{z}_i - f_i(\mathbf{pa}(\mathrm{z}_i))). \tag{2}$$

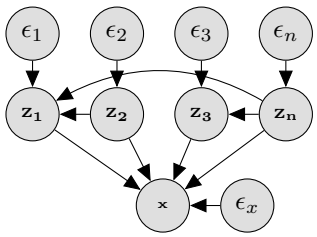


Figure 2: Data generation process with a latent SCM (endogenous and exogenous variables) causing an observation space.

This assumption is particularly important to demonstrate guarantees on stronger forms of identifiability. Assuming a functional form for the causal mechanism between variables, such as ANMs [11, 34], is an established method for identifying causal relationships [33, 8] due to asymmetries in the joint distribution. Moreover, the ANM assumption has been shown to perform well on real benchmarks from various domains such as meteorology, biology, medicine, engineering and economy [31], for the task of causal discovery.

Causal Ordering. Since we assume $\mathcal{G}$ to be a DAG, there is a non-unique permutation τ of d nodes such that a given node always appears first in the list compared to its descendants. Formally, $\tau_i < \tau_j \iff j \in \mathbf{de}(\mathrm{z}_i)$ where $\mathbf{de}(\mathrm{z}_i)$ are the descendants of z_i in $\mathcal{G}$ (Appendix B in [33]).

3.2 Identifiability Equivalence

The exact definition of model identifiability can be too restrictive. In reality, identifying a representation up to a simple transformation is enough. Therefore, we now formally define identifiability 1 and its weaker forms, which guarantee identifiability up to affine transformation 2, permutation and scaling 3, and block diagonal and scaling transformations 4. In the case of an ANM data generating process, [35] demonstrates the identifiability of models with only observational data; further, [37] discuss the identifiability of these models under data *score* functions. However, they do not discuss the identifiability of latent ANM models.

In this section, we define and make a distinction between different forms of identifiabilities and theoretically show that stronger forms of identifiability can be guaranteed when latent variables are causally ordered.

Definition 1. (Strong Identifiability) For parameter domain Θ and equivalence relation $\sim$ on Θ, the considered model is $\sim$-identifiable if equation 3 is satisfied.

$$\mathbb{P}_{\theta_1}(\mathbf{x}) = \mathbb{P}_{\theta_2}(\mathbf{x}), \Rightarrow \theta_1 \sim \theta_2. \tag{3}$$

Remark 1. According to [19], strong model identifiability makes the latent space $\mathbb{P}(\mathbf{z})$ identifiable.

Definition 2. (Affine Equivalence, $\sim_A$) For $\theta = \{\mathbf{f}, \mathbf{p}\}$ a set of parameters corresponding to the mixing function and prior, the affine equivalence relation $\sim_A$ on Θ is defined as:

$$(\mathbf{f}, \mathbf{p}) \sim_A (\tilde{\mathbf{f}}, \tilde{\mathbf{p}}) \iff \exists \quad \mathbf{A} \in \mathbb{R}^{n \times n}, \mathbf{c} \in \mathbb{R}^n \quad s.t. \quad \mathbf{f}^{-1}(\mathbf{x}) = \mathbf{A}\tilde{\mathbf{f}}^{-1}(\mathbf{x}) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{O}. \quad (4)$$

where $\mathbf{A}$ is an invertible matrix and $\mathcal{O}$ is an observational data space.

Remark 2. $\sim_A$ states that the images of $\mathbf{f}^{-1}$ and $\tilde{\mathbf{f}}^{-1}$ are related by an affine transformation.

Definition 3. (Permutation Equivalence, $\sim_P$) For $\theta = \{\mathbf{f}, \mathbf{p}\}$ a set of parameters corresponding to the mixing function and prior, the permutation equivalence relation $\sim_P$ on Θ is defined as by:

$$(\mathbf{f}, \mathbf{p}) \sim_P (\tilde{\mathbf{f}}, \tilde{\mathbf{p}}) \iff \exists \quad \mathbf{P} \in \mathbb{R}^{n \times n}, \mathbf{c} \in \mathbb{R}^n \quad s.t. \quad \mathbf{f}^{-1}(\mathbf{x}) = \mathbf{P}\tilde{\mathbf{f}}^{-1}(\mathbf{x}) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{O}. \quad (5)$$

where $\mathbf{P}$ is a block permutation matrix and $\mathcal{O}$ is an observational data space.

Remark 3. $\sim_P$ states that the images of $\mathbf{f}^{-1}$ and $\tilde{\mathbf{f}}^{-1}$ are related by rotation, scaling, and translation.

Definition 4. (Block Diagonal Equivalence, $\sim_D$) For $\theta = \{\mathbf{f}, \mathbf{p}\}$ a set of parameters corresponding to the mixing function and prior, the identity equivalence relation $\sim_D$ on Θ is defined as by:

$$(\mathbf{f}, \mathbf{p}) \sim_D (\tilde{\mathbf{f}}, \tilde{\mathbf{p}}) \iff \exists \quad \mathbf{D}, \mathbf{c} \quad s.t. \quad \mathbf{f}^{-1}(\mathbf{x}) = \mathbf{D}\tilde{\mathbf{f}}^{-1}(\mathbf{x}) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{O}. \quad (6)$$

where $\mathbf{D}$ is a block diagonal matrix, $\mathbf{c} \in \mathbb{R}^d$ is a shift vector, and $\mathcal{O}$ is an observational data space.

Remark 4. $\sim_D$ states that the images of $\mathbf{f}^{-1}$ and $\tilde{\mathbf{f}}^{-1}$ are related just by translation and scaling.

3.3 Identifiability of Latent ANMs

Universal Approximation of GMMs. Assuming the data generating process is an affine or piece-wise affine function, GMMs with a sufficient amount of components can model any densities in the limiting case [32], which in turn breaks the symmetry in the latent space behaving like auxiliary information in iVAE [44, 23]. In light of this, we model our latent distribution $\mathbb{P}(\mathbf{z}) = \prod_i \mathbb{P}^{\epsilon}(\mathbf{z}_i - f_i(\mathbf{pa}(\mathbf{z}_i))) = \sum_{j=1}^{J} \pi_j \mathcal{N}(\mu_j, \Sigma_j)$ as a mixture of densities.

Theorem 1. *(Identifiability of $\mathbf{z}$ under $\mathcal{G}$) Let $\mathbf{f}_o, \tilde{\mathbf{f}}_o$, satisfying injectivity assumption with $y \sim \mathbb{P}(\mathbf{z}), y' \sim \tilde{\mathbb{P}}(\mathbf{z})$, where $\mathbb{P}, \tilde{\mathbb{P}}$ follow the same causal graph $\mathcal{G}$. Suppose $\mathbf{f}_o(y)$ and $\tilde{\mathbf{f}}_o(y')$ are equally distributed, then, $\mathbb{P}(\mathbf{z}) \sim \tilde{\mathbb{P}}(\mathbf{z})$.*

Remark 5. This theorem is similar to, but goes beyond, Theorem E.1 in [23]. We show equivalence up to $\sim$ rather than $\sim_P$, given that the latent variables are constrained with respect to some causal graph (with all conditional independencies).

The proof is detailed in the appendix. The main outline of this proof includes showing that, under the constrain that the latent distribution respects the same causal graph $\mathcal{G}$, the block permutation matrix (in Theorem E.1 of [23]) can be reduced to a diagonal matrix. Similar to [23] we approximate the posterior distribution using GMMs.

Lemma 1. *(Identifiability of $\mathbf{z}$ under causal ordering) In the case when only causal ordering is known, the strong identifiability in theorem 1 reduces to block diagonal identifiability ($\sim_D$).*

Remark 6. Given the fact that constraining latent variables based on the complete causal graph may not be feasible, the lemma relaxes this constraint to enforce causal ordering, which guarantees $\sim_D$ identifiability. In section 4, we show how to achieve causal ordering in the latent space.

Theorem 2. *(Model Identifiability) Let $\mathbf{f}_o, \tilde{\mathbf{f}}_o$, satisfy the injectivity assumption with $y \sim \mathbb{P}(\mathbf{z}), y' \sim \tilde{\mathbb{P}}(\mathbf{z})$, where $\mathbb{P}, \tilde{\mathbb{P}}$ follow the same causal graph $\mathcal{G}$ and let $\mathcal{D} \subseteq \mathbb{R}^o$, where $o = |\mathcal{O}|$ such that $\mathbf{f}_o, \tilde{\mathbf{f}}_o$ are injective on to $\mathcal{D}$. Suppose $\mathbf{f}_o(y)$ and $\tilde{\mathbf{f}}_o(y')$ are equally distributed, then, $\mathbf{f}_o(\mathbf{z}) = \tilde{\mathbf{f}}_o(\mathbf{z})$.*

Remark 7. This theorem is similar to, but goes beyond Theorem D.4 in [23]. We show equivalence up to $\sim$ rather than $\sim_A$, given that the latent variables are constrained with respect to some causal graph (with all conditional independencies).

We detail the proof in the appendix. Similar to the proof of Theorem 1, we use GMMs to model our posterior distribution. The main component of the proof is to reduce affine transformation in Theorem D.4 [23] to an identity transformation.

Lemma 2. *(Model identifiability under causal ordering) In the case when latent variables follow a particular causal ordering τ rather than the entire causal graph $\mathcal{G}$, there exists a block diagonal transformation $\mathbf{D}$ such that $\mathbf{f}_o(\mathbf{z}) = (\tilde{\mathbf{f}}_o \circ \mathbf{D})(\mathbf{z})$.*

4 Estimation

We now derive an estimation procedure for learning the data generation process in equation 1. The findings of the previous section show that a data generation process with an ANM in the latent space is identifiable if the causal graph (or causal ordering) is known. Therefore, we proceed to define a loss function that will ensure that the latent space is causally ordered. Then, we describe a variational inference estimation method which models latent variables using a GMM.

4.1 Causal Ordering Loss

In causal representation learning, the goal is to learn causal variables from observations without information about the causal structure. However, there is always a causal ordering associated with a DAG. It is well known in the causal discovery literature that a complete causal graph is not identifiable from observational data without extra assumptions. If the functional form of the causal mechanism is assumed to be an ANM, causal directions become identifiable due to asymmetries. Interestingly, previous works on causal discovery [37, 38] explore a property of the distribution of ANMs to find a causal ordering. Here, we use the same property to enforce causal ordering instead of discovering it.

Enforcing causal ordering allows us to approximate the assumption of known causal ordering from Lemma 1. We use this property as a loss function for learning the latent representations. The property is based on the Jacobian of an ANM distribution's score function. Firstly, let the latent distribution be $\mathbb{P}(\mathbf{z})$ which follows an ANM and $\mathbb{P}^{\epsilon}$ be any quadratic exponential noise prior (e.g. Gaussian-like) [37, 38]. We can express its score function as

$$\nabla_{\mathrm{z}_i} \log \mathbb{P}(\mathbf{z}) = \frac{\partial \log \mathbb{P}^{\epsilon}(\mathrm{z}_i - f_i(\mathbf{pa}(\mathrm{z}_i)))}{\partial \mathrm{z}_i} - \sum_{j \in \mathbf{ch}(\mathrm{z}_i)} \frac{\partial f_j}{\partial \mathrm{z}_i} \frac{\partial \log \mathbb{P}^{\epsilon}(\mathrm{z}_j - f_j(\mathbf{pa}(\mathrm{z}_j)))}{\partial \mathrm{z}_i}. \tag{7}$$

Based on the above formalism it can be derived that $\nabla^2_{\mathbf{z}_i} \log \mathbb{P}(\mathbf{z}) = a \iff \mathbf{z}_i$ is a leaf node, where a is some constant and $\nabla^2_{\mathbf{z}_i} \log \mathbb{P}(\mathbf{z})$ is i^{th} diagonal element of the distribution's Hessian.

Proposition 1. *Assuming that $\mathbb{P}(\mathbf{z})$ follows an ANM and let $H^i_{var}(\mathbf{z}) = \mathrm{var}\Big(\nabla^2_{\mathrm{z}_i} \log \mathbb{P}(\mathbf{z})\Big)$. The latent space $\mathbf{z}$ can be causally ordered by minimising the causal ordering loss defined as*

$$\mathcal{L}_{order} = -\sum_{i}^{d-1} \log \frac{H^i_{var}(\mathrm{z}_i, \ldots, \mathrm{z}_d)^{-1}}{\sum_{j=i}^{d} H^j_{var}(\mathrm{z}_i, \ldots, \mathrm{z}_d)^{-1}} \tag{8}$$

Proof. The proof directly extends from analysing equation 7. As described in [37], the minimum variance in the latent log-likelihood's hessian corresponds to a leaf node. The loss term $\mathcal{L}_{order}$ is minimum if, and only if, the nodes at position i are leaves. We show this by contradiction; without loss of generality, consider the random latent order τ, s.t. $\tau_i \neq i$, then $H^0_{var}(\mathbf{z}) \geq \epsilon \Rightarrow \mathcal{L}_{order} > 0$. Based on the above expression $\mathcal{L}_{order} \to 0, \iff \tau_i = i$, where τ_i correspond to true causal order. It is important to note that as the representations are learned end-to-end, enforcing this loss would organise the latent order to follow the sorted true causal ordering. □

Hessian Estimation. To compute $H^i_{var}(\mathbf{z})$, we approximate the score's Jacobian (Hessian) with Stein kernel estimators [25] as described in [37]:

$$\mathbf{J}^{Stein} = -\mathrm{diag}(\mathbf{G}^{Stein}(\mathbf{G}^{Stein})^T) + (\mathbf{K} + \eta\mathbf{I})^{-1}\langle \nabla^2_{diag}, \mathbf{K} \rangle \tag{9}$$

Where $\mathbf{G}^{Stein} = -(\mathbf{K} + \eta\mathbf{I})^{-1}\langle \nabla, \mathbf{K} \rangle$ is the Stein gradient estimator [25], $\mathbf{K}$ is the median kernel, $\mathbf{I}$ is the Identity matrix, and $\langle a, b \rangle$ correspond to applying operation a on b element-wise. The final algorithm for computing $\mathcal{L}_{order}$ is described in Alg. 1.

4.2 Variational Inference

We are now interested in modelling a latent space with an arbitrarily complex distribution based on an ANM using the deep variational framework. That is, learning a posterior distribution that can approximate the ANM prior $\mathbb{P}(\mathbf{z})$ given a sample from the observational distribution. A multivariate

Algorithm 1 Compute topological loss ($\mathcal{L}_{order}$)

Initialize: $\mathcal{L}_{order} = 0$
Given: $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$
for $i = 0, \dots, d-1$
  $\tilde{\mathbf{z}} = \mathbf{z}[i:]$
  $\mathbf{v} = \text{var}\left(\mathbf{J}^{Stein}(\tilde{\mathbf{z}})\right)$ ▷ Compute variance of a Jacobian of a score
  $\tilde{\mathbf{v}} = \text{softmax}(-\log \mathbf{v})$ ▷ Smallest variance $\rightarrow$ highest $\tilde{\mathbf{v}}$
  $\mathcal{L}_{order} += \text{BCE}(\tilde{\mathbf{v}}, [1, 0 \dots 0])$ ▷ First element should have smallest variance
return $\mathcal{L}_{order}$

diagonal Gaussian prior cannot model these distributions. Therefore, we consider a prior following a GMM, following established literature [15, 17, 7], which is proven to be identifiable and have universal approximation capabilities [23].

In particular, we utilise the framework from MFC-VAE [7].We consider the generative model to be $\mathbb{P}(\mathbf{x}, \mathbf{z}, \mathbf{u}) = \mathbb{P}(\mathbf{x} \mid \mathbf{z})\mathbb{P}(\mathbf{z} \mid \mathbf{u})\mathbb{P}(\mathbf{u})$. MFC-VAE choose a posterior $\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x}) = \mathbb{Q}(\mathbf{u} \mid \mathbf{x})\mathbb{Q}(\mathbf{z} \mid \mathbf{x})$, where $\mathbb{Q}(\mathbf{z} \mid \mathbf{x})$ is a multivariate Gaussian with diagonal covariance and $\mathbb{Q}(\mathbf{u} \mid \mathbf{x})$ a categorical distribution over GMM components. Similar to MFC-VAE [7], we consider our inference model as described above, where the mixture components are inferred via prior (as $\mathbb{Q}_{\mathbf{f}}(\mathbf{u} \mid \mathbf{x}) \propto \exp(\mathbb{E}_{\mathbb{Q}_{\mathbf{f}}(\mathbf{z}|\mathbf{x})} \log \mathbb{P}_{\mathbf{p}}(\mathbf{u} \mid \mathbf{z}))$). In this case, the posterior $\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})$ is a GMM and can approximate the prior $\mathbb{P}(\mathbf{z})$ following a ANM. The ELBO for this model is described in Eqn. 10, where $\mathbb{E}$ is over $\mathbb{Q}(\mathbf{z} \mid \mathbf{x})$ distribution.

$$\mathcal{L}_{ELBO} = -\mathbb{E} \log \mathbb{P}(\mathbf{x} \mid \mathbf{z}) + \text{KL}\Big(\mathbb{Q}(\mathbf{z} \mid \mathbf{x}) \mid\mid \mathbb{P}(\mathbf{z})\Big) + \mathbb{E}\ \text{KL}\Big(\mathbb{Q}(\mathbf{c} \mid \mathbf{x}) \mid\mid \mathbb{P}(\mathbf{c} \mid \mathbf{z})\Big) \tag{10}$$

Lemma 3. *(Training Objective) Based on the proposition 1 and lemmas 1 and 2, models trained with the following objective: $\mathcal{L}_{total} = \mathcal{L}_{ELBO} + \alpha\mathcal{L}_{order}$, where will converge at true latents with $\sim_D$ equivalence.*

4.3 Neural Network Constraints

Injective Decoder. It is common to assume an injective decoder for proving the identifiability of a data generation process [23]. When implementing a deep generative model in practice, some constraints in the decoder are necessary to ensure that neural networks are modelling injective functions. We follow similar modelling assumptions of ICE-BeeM [20]: (i) Monotonicity: The latent dimension of the decoder is monotonically increasing, *i.e.,* $d_{l+1} \geq d_l \quad \forall l \in \{0, \dots, L-1\}$, where d_l corresponds to the feature dimension at layer l and L is the total number of layers in the decoder. (ii) Activation: The activation function after every layer corresponds to LeakyReLU ($\max(0, x) + \alpha \min(0, x), \alpha \in (0, 1)$). (iii) Full rank: All weight matrices $\mathbf{f}_l$ are full row ranked, as the number of rows is greater than or equal to the number of columns. (iv) Invertible sub-matrix: All weight sub-matrices $\mathbf{f}'_l$ of size $d_l \times d_l$ are invertible.

Discussion: Proposition 1 shows that, given sufficient data and compute, under the non-linear ANM assumption, latent representations are organised with respect to evidential ordering. Additionally, given the organised latent representations, the causal relationships among the representations can be estimated using conditional independencies similar to [37, 38, 18]. We later discuss how latent causal discovery can be achieved. As previously discussed in equation 1, it is important to note that we consider all features in $\mathbf{z}$ to be direct parents of $\mathbf{x}$, thus any indirect cause $y \rightarrow (\mathbf{z}_i \in \mathbf{z}) \rightarrow \mathbf{x}$ cannot be recovered by our approach.

5 Experiments

Here, we demonstrate the effectiveness of latent ANM models with topological constraints on both tabular (including a synthetic data generating process) and image (MorphoMNIST and Causal3DIdent) datasets. We compare the proposed model against two baseline methods β-VAE and MFC-VAE with a single facet on mean correlation coefficient (MCC) and causal ordering divergence (COD).

5.1 Metrics

We compute different variants of MCC: (i) across multiple random seeds (MCC-R): measures the stability of the training process given the model; (ii) with respect to ground truth variables (MCC-GT): measures the faithfulness of the estimated latent variables to true latent variables [20]; and (iii) subset MCC (MCC-SG): in the case when all parents of $\mathbf{X}$ are not observed, we measure the faithfulness by considering a subset of latent variables. All three variants are formally described in definition 5. As these MCC measures are permutation invariant by nature, to capture the perceived order among latent variables, we also calculate COD, which measures the divergence of the topological order in an estimated causal graph from the causal order, formally defined in equation 13. In addition, to quantify the injectiveness of the model we compute MIC and RRO defined in 7.

Definition 5. (Mean Correlation Coefficient) We compute the mean correlation coefficient with respect to ground truth (MCC-G) as described in [20]. MCC-SG and MCC-R are based on MCC-G and are described as:

$$\text{MCC-SG}(\hat{\mathbf{z}}, \mathbf{z}) = \max \Big\{ \text{MCC-G}(\hat{\mathbf{z}}[S_j], \mathbf{z}), \quad \forall j = \{1, \dots, |\mathcal{S}|\}, \quad S = \binom{|\hat{\mathbf{z}}|}{|\mathbf{z}|} \Big\} \tag{11}$$

$$\text{MCC-R}(\{\hat{\mathrm{z}}_0, \dots, \hat{\mathrm{z}}_K\}) = \frac{1}{K-1} \sum_k \text{MCC-G}(\hat{\mathrm{z}}_k, \hat{\mathrm{z}}_0), \tag{12}$$

where $\hat{\mathbf{z}}_k = \mathbf{f}_k^{-1}(\mathbf{X})$, S is the set of all the partition indices of $\hat{z}$ with the size of $|\mathbf{z}|$, $\mathbf{z}$ corresponds to the ground truth latent features and K total number of experimental runs.

Definition 6. (Causal Order Divergence, COD) Similar to divergence metric in [37, 38], we define COD as:

$$\text{COD}(\tau, A) = \sum_{i=0}^{d} \sum_{j>i}^{d} A_{ij} \tag{13}$$

where $\tau = \{0, \dots, d\}$ is the expected order and A is an estimated adjacency graph predicted using the resulting latent space after training.

Definition 7. (Mean Injectivity Coefficient, MIC) Based on the network constraints described in section 4.3, we compute the MIC to measure the *injectivity* of the model. MIC is formally described as:

$$\text{MIC}(\mathbf{f}) = \min \Big\{ \frac{1}{|\mathcal{C}|} \sum_j \frac{Rank(\mathbf{f}_i(\mathcal{C}_j)^T)}{ri} \quad \forall i \in \{0, \dots, |\mathbf{f}|\} \Big\} \tag{14}$$

where, ci, ri correspond to number of columns and rows of $\mathbf{f}_i$, with abuse of notation, we use $\mathcal{C} = \binom{ci}{ri}$ as a set of all partitions of column indices with size ri, and $|S|$ is the cardinality of set S.

Remark 8. We measure the average row rank ratio $\text{RRO} = \Big(\frac{1}{L} \sum_l \frac{Rank(f_l)}{d_l} \Big)$ and MIC (ref. definition 14) to measure the injectivity of the decoder.

5.2 Data Generation

Simulation Data: To generate the synthetic dataset we first randomly generate a latent causal DAG with n nodes and e edges using [46]. We randomly select all the involved structural causal models f_i with an *injective* mapping from $\mathbf{pa}(\mathrm{z}_i)$ to z_i. Finally, we select an injective random transformation function $\mathbf{f}_o$ mapping from latent space $\mathbf{z}$ to observational data $\mathbf{X}$. In our experiments we generate 2,000 datapoints from SYN-2, SYN-15, and SYN-50 processes, where SYN-K correspond to the above data-generating process with latent variable $\mathbf{z} \in \mathbb{R}^k$ and observational data $\mathbf{X} \in \mathbb{R}^{2k}$.

Image Datasets: We further extend our method on imaging datasets, which include MorphoMNIST [4] variants and Causal3DIdent [43]. In the case of MorphoMNIST, we use MorphoMNIST-IT, MorphoMNIST-TI, MorphoMNIST-TS, and MorphoMNIST-TSWI variants where I, T, S, and W correspond to latent variables $\mathbf{z}$ with the semantics of intensity, thickness, slant, and width respectively. We detail all the data-generating processes in Appendix. All the MorphMNIST variants have 60,000 training images and 10,000 testing images. Similarly, Causal3DIdent includes 252,000 training

samples and 25,200 test samples that were generated using a fixed causal graph with 10 nodes (more details about this dataset can be found in [43], Appendix B).

5.3 Results

In each of our experiments, we adopt a model adhering to the properties delineated in Section 4.3. Observations pertaining to MIC and RRO measures suggest that the injectivity of the decoder is predominantly influenced by choice of architecture and the dataset under consideration.

For instance, the MIC for the SYN-2, SYN-15, and SYN-50 datasets are recorded as 1.0, 0.68, and 1.0, respectively, while the corresponding RRO values are 0.88, 0.93, and 0.95. To gauge the effectiveness in terms of stability and faithfulness, we tabulated the results concerning MCC-R and MCC-GT metrics for synthetic and image datasets in Table 2. Here, we employed five random seeds to compute the MCC-R and report the mean and standard deviation across these five runs for COD and MCC-G. These results, illustrate that given additive noise models in latent space, the proposed loss enforces evidential structure as COD goes to 0 and achieves stronger identifiability which can be inferred from MCC-R and MCC-G values.

Similarly, in the case of imaging datasets for both MorphoMNIST-IT and MorphoMNIST-TSWI we observed MIC of 1.0 and RRO of 0.85, and the resulting MCC-SG (as previously described, in the case of image datasets, all the parents are not observed) and COD measures are described in Table 2. In all our experiments, we observed that topological ordering with respect to the evidential graph is better enforced in COVAE and even in terms of stability and faithfulness of the latent representations, COVAE outperforms VAE and MFC-VAE. Additional experiments on other variants of the MorphoMNIST dataset and Causal3DIdent are detailed in the Appendix.

Table 2: MCC and COD results on synthetic datasets with 2, 15, and 50 nodes in the latent space along with imaging datasts MorphoMNIST-IT and MorphoMNIST-TSWI.

METHODS(↓), METRICS(→)	SYN-2		
	COD (↓)	MCC-R(↑)	MCC-G(↑)
VAE	0.13 ± 0.08	0.11	0.26± 0.03
MFC-VAE	0.17 ± 0.09	0.14	0.35 ± 0.06
COVAE	**0.00** ± 0.01	**0.62**	**0.52** ± 0.07
	SYN-15		
VAE	1.68 ± 0.22	0.21	0.22 ± 0.02
MFC-VAE	1.43 ± 0.24	0.26	0.26 ± 0.03
COVAE	**0.03** ± 0.01	**0.42**	**0.34** ± 0.03
	SYN-50		
VAE	5.53 ± 0.81	0.23	0.28 ± 0.24
MFC-VAE	5.17 ± 0.62	0.31	0.26 ± 0.01
COVAE	**0.78** ± 0.46	**0.39**	**0.34** ± 0.02
	MORPHOMNIST-IT		
	COD (↓)	MCC-R(↑)	MCC-SG(↑)
VAE	1.61 ± 0.44	0.29	0.23 ± 0.11
MFC-VAE	1.04 ± 0.46	0.36	0.34 ± 0.09
COVAE	**0.0**	**0.59**	**0.47** ± 0.08
	MORPHOMNIST-TSWI		
VAE	0.81 ± 0.26	0.47	0.21 ± 0.00
MFC-VAE	1.35 ± 0.24	0.52	0.28 ± 0.04
COVAE	**0.0**	**0.61**	**0.31** ± 0.04

6 Conclusion

In this work, we propose the first fully unsupervised causal representation learning method for data adhering to ANM by imposing a topological ordering on the latent space that corresponds to the underlying causal graph. We present a multitude of results pertaining to the identifiability of latent representations, demonstrating these outcomes both empirically and experimentally. Evaluations on synthetic and image datasets corroborate the efficacy of the proposed estimation method, which in practice exhibits superior identifiability. Possible future works would be to investigate sample efficiency and robustness of the models trained with the proposed estimation method. Additionally, extending the proposed approach from ANM to post-ANM and simplifying modelling assumptions would be of particular interest. Although modelling assumptions are standard and widely used in practice, formulating a model and estimation methods without these assumptions would be ideal.

References

[1] Kartik Ahuja, Yixin Wang, Divyat Mahajan, and Yoshua Bengio. Interventional causal representation learning. *arXiv preprint arXiv:2209.11924*, 2022.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[3] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022.

[4] Daniel C Castro, Jeremy Tan, Bernhard Kainz, Ender Konukoglu, and Ben Glocker. Morpho-mnist: Quantitative assessment and diagnostics for representation learning. *Journal of Machine Learning Research*, 20(178):1–29, 2019.

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2016.

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

[7] Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris C Holmes. Multi-facet clustering variational autoencoders. *Advances in Neural Information Processing Systems*, 34:8676–8690, 2021.

[8] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10, 2019.

[9] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[10] Irina Higgins, Sébastien Racanière, and Danilo Rezende. Symmetry-based representations for artificial and biological general intelligence. *Frontiers in Computational Neuroscience*, 16, 2022.

[11] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

[12] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[13] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

[14] Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear, 2023.

[15] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

[16] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017.

[17] Matthew James Johnson, David Duvenaud, Alexander B. Wiltschko, Sandeep R. Datta, and Ryan P. Adams. Composing graphical models with neural networks for structured representations and fast inference. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016.

[18] Markus Kalisch and Peter Bühlman. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(3), 2007.

[19] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[20] Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[21] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2018.

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

[24] Avinash Kori, Ben Glocker, and Francesca Toni. Glance: Global to local architecture-neutral concept-based explanations. *arXiv preprint arXiv:2207.01917*, 2022.

[25] Yingzhen Li and Richard E Turner. Gradient estimators for implicit models. *arXiv preprint arXiv:1705.07107*, 2017.

[26] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[27] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Learning disentangled representations in the imaging domain. *Medical Image Analysis*, 80, 2022.

[28] 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 *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

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

[30] Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[31] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

[32] Hien D Nguyen and Geoffrey McLachlan. On approximations via convolution-defined mixture models. *Communications in Statistics-Theory and Methods*, 48(16):3945–3955, 2019.

[33] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

[34] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15, 2014.

[35] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

[36] Hans Reichenbach. The direction time. *Univ. of California Press*, 1956.

[37] Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *International Conference on Machine Learning*, pages 18741–18753. PMLR, 2022.

[38] Pedro Sanchez, Xiao Liu, Alison Q O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *The Eleventh International Conference on Learning Representations*, 2023.

[39] 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, 2021.

[40] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23:1–55, 2022.

[41] Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[42] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

[43] 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. *Advances in neural information processing systems*, 34:16451–16467, 2021.

[44] Matthew Willetts and Brooks Paige. I don't need u: Identifiable non-linear ica without side information. *arXiv preprint arXiv:2106.05238*, 2021.

[45] 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*, pages 9593–9602, 2021.

[46] Keli Zhang, Shengyu Zhu, Marcus Kalander, Ignavier Ng, Junjian Ye, Zhitang Chen, and Lujia Pan. gcastle: A python toolbox for causal discovery, 2021.

A Proofs

Theorem 3. *(Identifiability of* $\mathbf{z}$ *under* $\mathcal{G}$*) Let* $\mathbf{f}_o, \tilde{\mathbf{f}}_o$*, satisfying injectivity assumption with* $y \sim \mathbb{P}(\mathbf{z}), y' \sim \tilde{\mathbb{P}}(\mathbf{z})$*, where* $\mathbb{P}, \tilde{\mathbb{P}}$ *follow the same causal graph* $\mathcal{G}$*. Suppose* $\mathbf{f}_o(y)$ *and* $\tilde{\mathbf{f}}_o(y')$ *are equally distributed, then,* $\mathbb{P}(\mathbf{z}) \sim \tilde{\mathbb{P}}(\mathbf{z})$.

Remark 9. This theorem is similar to, but goes beyond, Theorem E.1 in [23]. We show equivalence up to $\sim$ rather than $\sim_P$, given that the latent variables are constrained with respect to some causal graph (with all conditional independencies).

Proof. The proof is detailed in the Appendix. The main outline of this proof includes showing that, under the constrain that the latent distribution respects the same causal graph $\mathcal{G}$, the block permutation matrix (in Theorem E.1 of [23]) can be reduced to a diagonal matrix. Similar to [23], we approximate the posterior distribution using GMMs.

Based on our formulation, we consider the following:

$$y \sim \mathbb{P}(\mathbf{z}) = \prod_i \mathbb{P}_{\mathcal{G}}(\mathrm{z}_i \mid \mathbf{pa}(\mathrm{z}_i)) = \sum_{j=0}^{J} \pi(j) \mathcal{N}(\mu_j, \Sigma_j)$$

$$y' \sim \tilde{\mathbb{P}}(\mathbf{z}) = \prod_i \tilde{\mathbb{P}}_{\mathcal{G}}(\mathrm{z}_i \mid \mathbf{pa}(\mathrm{z}_i)) = \sum_{j=0}^{J} \tilde{\pi}(j) \mathcal{N}(\tilde{\mu}_j, \tilde{\Sigma}_j)$$

Now, we consider $y' = Ay + b$; we show that when the latent causal distribution is known, A is the identity matrix.

Without loss of generality, we consider y' to belong to component $k \in \{0, \ldots, J\}$. Given that $y' = Ay + b$, a linear transformation of a Gaussian random variable results in:

$$\tilde{\Sigma}_k = A \Sigma_k A^T$$

As both matrices are diagonal and positive semi-definite (PSD), spectral decomposition using singular value decomposition (SVD) results in $\tilde{\Sigma}_k = V_k V_k^T = {V'}_k {V'}_k^T$, where $V_k, {V'}_k$ are PSD matrices and are unique up to orthogonal transformation $\Rightarrow V_k = R_k V'_k$ for some unitary matrix R_k for each and every $k \in \{0, \ldots, J\}$, resulting in:

$$\tilde{\Sigma}_k^{1/2} = V_k R_k = A \Sigma_k^{1/2}$$

Without loss of generality, let's consider two components $k = 1$ and $k = 2$,

$$\tilde{\Sigma}_1^{-1/2} \Sigma_1^{-1/2} = \tilde{\Sigma}_2^{-1/2} \Sigma_2^{-1/2} \Rightarrow V_1 R_1 \Sigma_1^{-1/2} = V_2 R_2 \Sigma_2^{-1/2}$$

By rearranging terms, we get:

$$R_1 (\Sigma_1^{-1/2} \Sigma_2^{1/2}) R_2^{-1} = V_1^{-1} V_2$$

As R_1, R_2 are unitary, Σ_1, Σ_2 are diagonal and PSD, and y, y' follow the same causal structure $\mathcal{G}$, the SVD decomposition of $V_1^{-1} V_2$ results in R' such that:

$$V_1 R'_1 = A \Sigma_1^{1/2}, \text{ for } A' \coloneqq V_1 R_1, \quad \text{we have } (A')^{-1} A = \Sigma_1^{-1/2} \Rightarrow A = I$$

This concludes the proof. □

Description: The theorem mainly focuses on showing identifiability results when latent distribution follows the same factorization with respect to a known causal graph and has an injective and a perfect mixing function. Here, we show that based on the fact that GMMs are universal approximators of any arbitrary latent density.

Lemma 4. *(Identifiability of* $\mathbf{z}$ *under causal ordering) In the case when only causal ordering is known, the strong identifiability in theorem 1 reduces to block diagonal identifiability (*$\sim_D$*).*

Proof. Similar to the previous theorem, we consider here the case of latent variables that can be factorized w.r.t. two different causal graphs $\mathcal{G}, \mathcal{G}'$ with the same causal order τ.

$$y \sim \mathbb{P}(\mathbf{z}) = \prod_i \mathbb{P}_{\mathcal{G}}(\mathrm{z}_i \mid \mathbf{pa}(\mathrm{z}_i)) = \sum_{j=0}^{J} \pi(j)\mathcal{N}(\mu_j, \Sigma_j)$$

$$y' \sim \tilde{\mathbb{P}}(\mathbf{z}) = \prod_i \tilde{\mathbb{P}}_{\mathcal{G}'}(\mathrm{z}_i \mid \mathbf{pa}(\mathrm{z}_i)) = \sum_{j=0}^{J} \tilde{\pi}(j)\mathcal{N}(\tilde{\mu}_j, \tilde{\Sigma}_j)$$

Now, we consider $y' = Ay + b$; we show that when the causal ordering is known, A is a block diagonal matrix.

Without loss of generality, we consider y' to belong to component $k \in \{0, \ldots, J\}$. Given that $y' = Ay + b$, a linear transformation of a Gaussian random variable results in:

$$\tilde{\Sigma}_k = A\Sigma_k A^T$$

As both matrices are diagonal and positive semi-definite (PSD), spectral decomposition using singular value decomposition (SVD) results in $\tilde{\Sigma}_k = V_k V_k^T = {V'}_k {V'}_k^T$, where $V_k, {V'}_k$ are PSD matrices and are unique up to orthogonal transformation $\Rightarrow V_k = R_k V_k'$ for some unitary matrix R_k for each and every $k \in \{0, \ldots, J\}$, resulting in:

$$\tilde{\Sigma}_k^{1/2} = V_k R_k = A\Sigma_k^{1/2}$$

Without loss of generality, let's consider two components $k = 1$ and $k = 2$,

$$\tilde{\Sigma}_1^{-1/2}\Sigma_1^{-1/2} = \tilde{\Sigma}_2^{-1/2}\Sigma_2^{-1/2} \Rightarrow V_1 R_1 \Sigma_1^{-1/2} = V_2 R_2 \Sigma_2^{-1/2}$$

By rearranging terms, we get:

$$R_1(\tilde{\Sigma}_2^{1/2}\tilde{\Sigma}_1^{1/2})R_2^{-1} = R_1(\Sigma_1^{-1/2}\Sigma_2^{1/2})R_2^{-1} = V_1^{-1}V_2$$

As R_1, R_2 are unitary and Σ_1, Σ_2 are diagonal and PSD with all distinct entries, and y and y' follow same causal order τ, the SVD decomposition of $V_1^{-1}V_2$ results in R', with transformation matrix $\mathbf{D}$ such that:

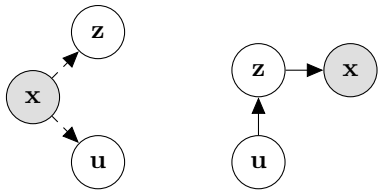
$$V_1 R_1' \mathbf{D} = A\Sigma_1^{1/2}, \text{ for } A' \coloneqq V_1 R_1, \quad \text{we have } (A')^{-1}A = \mathbf{D}\Sigma_1^{-1/2} \Rightarrow A = \mathbf{D}$$

As the entries of Σ_1, Σ_2 are distinct and y, y' follow the same causal order τ, the resulting transformation matrix is either diagonal or block diagonal in nature.

This concludes the proof.

□

Description: Given the fact that constraining latent variables based on the complete causal graph may not be feasible, the lemma relaxes this constraint to enforce causal ordering, which guarantees $\sim_D$ identifiability. Intuitively, the transformation is diagonal in the case when the graph follows a Markov chain structure, and block diagonal when the graph consists of multiple sister nodes at each level.



(a) Inference model (b) Generative model

Figure 3: Variational posterior $\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})$ used during inference on the left and generative model on the right. We do not give a causal interpretation for $\mathbf{c}$ in this case.

Theorem 4. *(Model Identifiability) Let $\mathbf{f}_o, \tilde{\mathbf{f}}_o$, satisfy the injectivity assumption with $y \sim \mathbb{P}(\mathbf{z}), y' \sim \tilde{\mathbb{P}}(\mathbf{z})$, where $\mathbb{P}, \tilde{\mathbb{P}}$ follow the same causal graph $\mathcal{G}$ and let $\mathcal{D} \subseteq \mathbb{R}^o$, where $o = |\mathcal{O}|$ such that $\mathbf{f}_o, \tilde{\mathbf{f}}_o$ are injective on to $\mathcal{D}$. Suppose $\mathbf{f}_o(y)$ and $\tilde{\mathbf{f}}_o(y')$ are equally distributed, then, $\mathbf{f}_o(\mathbf{z}) = \tilde{\mathbf{f}}_o(\mathbf{z})$.*

Remark 10. This theorem is similar to, but goes beyond Theorem D.4 in [23]. We show equivalence up to $\sim$ rather than $\sim_A$, given that the latent variables are constrained with respect to some causal graph (with all conditional independencies).

Proof. We detail the proof in the Appendix. Similar to the proof of Theorem 1, we use GMMs to model our posterior distribution.

Similar to the proof of Theorem C.7 in [23], we assume both $\mathbf{f}_o, \tilde{\mathbf{f}}_o$ are piece-wise affine functions and are invertible on $B(x_0, 2\delta) \cap \mathbf{f}_0(\mathbb{R}^d)$, where B is a ball with radius δ. Given both y, y' are sampled from the same causal latent distribution, $\mathbf{f}_o \sim \tilde{\mathbf{f}}_o$. Since, $\tilde{\mathbf{f}}_o$ is invertible on $\mathcal{D}$, $y = (\tilde{\mathbf{f}}_o^{-1} \circ \mathbf{f}_o)(y)$ on $\mathbf{f}_o^{-1}(D)$. This results in $\mathbf{f}_o(y) = \tilde{\mathbf{f}}_o(y')$ for every $y' \in \tilde{\mathbf{f}}_o^{-1}(\mathcal{D})$.

Given that the latent distribution follows the same causal distribution, the injectivity assumption holds, and the mixing functions share the same pre-image, both mixing functions are identifiable.

□

Lemma 5. *(Model identifiability under causal ordering) In the case when latent variables follow a particular causal ordering τ rather than the entire causal graph $\mathcal{G}$, there exists a block diagonal transformation $\mathbf{D}$ such that $\mathbf{f}_o(\mathbf{z}) = (\tilde{\mathbf{f}}_o \circ \mathbf{D})(\mathbf{z})$.*

Proof. Similar to Lemma 1, we consider the latent space to be factorized with respect $\mathcal{G}, \mathcal{G}'$ with the topological order τ, resulting in block diagonal transformation function $\mathbf{D} : \mathbb{R}^d \to \mathbb{R}^d$ such that $x' = \mathbf{D}x$ for $x \sim \mathcal{G}, x' \sim \mathcal{G}'$.

As described in the previous theorem, we consider both $\mathbf{f}_o, \tilde{\mathbf{f}}_o$ are piece-wise affine functions and are invertible on $B(x_0, 2\delta) \cap \mathbf{f}_0(\mathbb{R}^d)$. Based on the transformation $\mathbf{D}$, $\tilde{\mathbf{f}}_o(y) \sim (\mathbf{f}_o \circ \mathbf{D})(y) \; \forall \; y \in \tilde{\mathbf{f}}_o^{-1}(\mathcal{D})$. Since $\mathbf{f}_o, \tilde{\mathbf{f}}_o$ are invertible, above expression can be rewritten as $y \sim (\tilde{\mathbf{f}}_o^{-1} \circ \mathbf{f}_o \circ \mathbf{D})(y)$ on $\tilde{\mathbf{f}}_o^{-1}(\mathcal{D})$. In some special cases, where $\mathcal{D}$ is invertible (Markov chain structure in latent space), where $\mathcal{D}$ is mostly diagonal matrix $y \sim (\mathbf{D}^{-1} \circ \mathbf{f}_o \circ \mathbf{f}_o^{-1})(y)$ on $\tilde{\mathbf{f}}_o^{-1}(\mathcal{D})$, we can infer that $\mathbf{D}$ is a diagonal transformation on $\tilde{\mathbf{f}}_o - 1(\mathcal{D})$. In general cases, there exists an invertible transformation $\mathbf{D}_1, \mathbf{D}_2$ such that:

$$(\tilde{\mathbf{f}}_o^{-1} \circ \mathbf{f}_o \circ \mathbf{D}) \sim (\mathbf{D}_1^{-1} \circ \mathbf{f}_o \circ \mathbf{f}_o^{-1}) + (\mathbf{D}_2^{-1} \circ \mathbf{f}_o \circ \mathbf{f}_o^{-1})$$

From which we can conclude that $\mathbf{D}_1, \mathbf{D}_2$ are diagonal transformations on $\tilde{\mathbf{f}}_o - 1(\mathcal{D})$, for which we have $\mathbf{f}_o(y) = (\tilde{\mathbf{f}}_o \circ \mathbf{D})(y') \; \forall y \in \tilde{\mathbf{f}}_o^{-1}(\mathcal{D})$.

□

Lemma 6. *(Training Objective) Based on the proposition 1 and lemmas 1, 2, and models trained with the following objective: $\mathcal{L}_{total} = \mathcal{L}_{ELBO} + \alpha \mathcal{L}_{order}$, where will converge at true latents with $\sim_D$ equivalence.*

Proof. ELBO Derivation:

For this, we start with the data distribution as $\mathbb{P}(\mathbf{x})$, and the aim is to maximize the log-likelihood of this distribution:

$$\log \mathbb{P}(\mathbf{x})$$
$$= \log \int_{\mathbf{u}} \int_{\mathbf{z}} \mathbb{P}(x, \mathbf{u}, \mathbf{z}) d\mathbf{z} d\mathbf{u}$$

Let's consider variational distributions $\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})$.

$$= \log \int_{\mathbf{u}} \int_{\mathbf{z}} \mathbb{P}(\mathbf{x}, \mathbf{u}, \mathbf{z}) \frac{\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})}{\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})} d\mathbf{z} d\mathbf{u}$$

$$\geq \mathbb{E}_{\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})} \log \frac{\mathbb{P}(\mathbf{x}, \mathbf{u}, \mathbf{z})}{\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})}$$

Based on modelling assumption described in figure 3, $\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})$ decomposes as $\mathbb{Q}(\mathbf{u} \mid \mathbf{x}) \mathbb{Q}(\mathbf{z} \mid \mathbf{x})$

$$= \mathbb{E}_{\mathbb{Q}(\mathbf{u}, \mathbf{z} \mid \mathbf{x})} \Big[\log \mathbb{P}(\mathbf{x} \mid \mathbf{z}) + \log \frac{\mathbb{P}(\mathbf{u} \mid \mathbf{z})}{\mathbb{Q}(\mathbf{u} \mid \mathbf{x})} + \log \frac{\mathbb{P}(\mathbf{z})}{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \Big]$$

$$= \mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \log \mathbb{P}(\mathbf{x} \mid \mathbf{z}) + \mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \mathbb{E}_{\mathbb{Q}(\mathbf{u} \mid \mathbf{x})} \log \frac{\mathbb{P}(\mathbf{u} \mid \mathbf{z})}{\mathbb{Q}(\mathbf{u} \mid \mathbf{x})} + \mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \log \frac{\mathbb{P}(\mathbf{z})}{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})}$$

$$= \mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \log \mathbb{P}(\mathbf{x} \mid \mathbf{z}) - \mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \mathrm{KL} \Big(\mathbb{P}(\mathbf{u} \mid \mathbf{z}) \| \mathbb{Q}(\mathbf{u} \mid \mathbf{x}) \Big) - \mathrm{KL} \Big(\mathbb{P}(\mathbf{z}) \| \mathbb{Q}(\mathbf{z} \mid \mathbf{x}) \Big)$$

$$\Rightarrow \mathcal{L}_{ELBO} = -\mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \log \mathbb{P}(\mathbf{x} \mid \mathbf{z}) + \mathbb{E}_{\mathbb{Q}(\mathbf{z} \mid \mathbf{x})} \mathrm{KL} \Big(\mathbb{P}(\mathbf{u} \mid \mathbf{z}) \| \mathbb{Q}(\mathbf{u} \mid \mathbf{x}) \Big) + \mathrm{KL} \Big(\mathbb{P}(\mathbf{z}) \| \mathbb{Q}(\mathbf{z} \mid \mathbf{x}) \Big)$$

Based on proposition 1, we can infer that as $\mathcal{L}_{order} \to 0$ the causal order is enforced in latent space:

$$\mathbb{P}(\mathbf{z}) = \prod_i \mathbb{P}(\mathrm{z}_i \mid \{\mathrm{z}_{i+1}, \dots \mathrm{z}_d\})$$

Based on our assumptions, the considered model is injective, and based on lemma 1 and 2 we know that the latent distribution and model converges to a unique solution with $\sim_D$ equivalence given the causal ordering of latent space.

Given infinite training data and compute, as $\mathcal{L}_{total} \to 0$, $\mathcal{L}_{ELBO} \to 0$ converging the obtained unique distribution to true prior distribution. □

B Data Generating Process

B.1 MORPHOMNIST dataset

Here, we synthetic data based on MNIST digits [4]. We define multiple data-generating process with four different variables thickness, width, slant, and intensity, and evaluate our proposed method in terms of MCC's and COD. Here, thickness corresponds to the stroke thickness of a digit, width corresponds to the total width of a written digit, slant corresponds to the shear factor along a horizontal direction, and intensity corresponds to the average intensity of pixels in a digit. Functions $SetIntensity(x; i)$, $SetSlant(x; s)$, $SetWidth(x; w)$, and $SetThickness(x; t)$ refer to the operations applied to the original MNIST digit to generate new image x with desired properties by controlling image morphology. We use the data-generating process similar to the ones described in [24], we formally describe them below.

Morpho-MNIST-TI: In this setting we consider two causal variables thickness and intensity, where thickness causes intensity. Mathematically the functional relationship between variables are defined as described in equation 15.

$$
\begin{aligned}
t &:= f_t \triangleq 0.5 + \epsilon_t \quad \epsilon_t \sim \Gamma(10, 5) \\
i &:= f_i \triangleq 64 + 191 * \sigma(2 * w + 5) + \epsilon_i \quad \epsilon_i \sim \mathcal{N}(0, 1) \\
x &:= f_x = SetIntensity(SetThickness(X; t); i)
\end{aligned} \tag{15}
$$

Morpho-MNIST-IT: In this experiment we inverted a directionality from previous setting resulting in intensity to cause thickness, which is mathematically described in equation 16

$$
\begin{aligned}
i &:= f_i \triangleq \epsilon_i \quad \epsilon_i \sim \mathbb{U}(60, 255) \\
t &:= f_t \triangleq 3 + \sigma(i/255) + \epsilon_s \quad \epsilon_s \sim \mathcal{N}(0, 0.5) \\
x &:= f_x = SetThickness(SetIntensity(X; i); t)
\end{aligned} \tag{16}
$$

Morpho-MNIST-TS: In this setup we use thickness and slant as causal attributes, where thickness causes digit slantness, which is formally described in equation 17

$$
\begin{aligned}
t &:= f_t \triangleq \epsilon_t \quad \epsilon_t \sim \Gamma(0, 5) \\
s &:= f_s \triangleq 10 + 5 * \sigma(2 * t - 5) + \epsilon_s \quad \epsilon_s \sim \mathcal{N}(0, 0.5) \\
x &:= f_x = SetSlant(SetThickness(X; t); s)
\end{aligned} \tag{17}
$$

Morpho-MNIST-TSWI: In this setup we increased a complexity by using intensity, thickness, slant, and digit width as a causal attributes, where thickness causes slant, thickness and slant causes width, and width causes intensity. This data-generating process is formally described in equation 18

$$
\begin{aligned}
t &:= f_t \triangleq \epsilon_t \quad \epsilon_t \sim \Gamma(0, 5) \\
s &:= f_s \triangleq 10 + 20 * t + \epsilon_s \quad \epsilon_s \sim \mathcal{N}(0, 5) \\
w &:= f_w \triangleq 10 + 15 * \sigma(0.5 * t) - 0.25 * s + \epsilon_w \quad \epsilon_w \sim \mathcal{N}(0, 1) \\
i &:= f_i \triangleq 64 + 191 * \sigma(w/25) + \epsilon_i \quad \epsilon_i \sim \mathbb{N}(0, 1) \\
x &:= f_x = SetIntensity(SetWidth(SetSlant(SetThickness(X; t); s); w); i)
\end{aligned} \tag{18}
$$

C Experimental Setup

C.1 Code and Implementation

We use the latent GMM loss from MFC-VAE [7] inspired in the implementation from `https://github.com/FabianFalck/mfcvae`. We also append the code for the model and loss functions used in the paper to the supplemental material.

C.2 Hyperparameters

In Table 3 we detail all the hyper-parameters used in our experiments. We use a fixed decoder standard deviation in the case of CAUSAL3DIDENT and MORPHOMNIST, while in the case of SYN-K dataset it remains learnable (described as σ in the table). It is also worth mentioning that for the VAE method on CAUSAL3DIDENT, we trained a deeper model and also set the KL weight term β equal to 0 to ensure fair comparison with the other two methods and avoid posterior collapse, respectively.

D Results

Table 2 depicts final results on MORPHOMNIST-TI, MORPHOMNIST-TS, and CAUSAL3DIDENT dataset, respectively. For each method, we re-run all experiments and collect metrics across 5 different random seeds for MORPHOMNIST-TI and MORPHOMNIST-TS, and 3 random seeds for CAUSAL3DIDENT. For the latter dataset, we observed that all three metrics exhibit high variance across runs; however, it is clear that both MFC-VAE and COVAE are comparable methods.

Table 3: Experimental details w.r.t models and datasets

DATASETS(↓), METHODS(→)		VAE	MFC-VAE	COVAE
SYN-K	No. Layers		3 if k < 3 else 6	
	Training Steps		15600	
	No. Samples		2000	
	Batch Size		256	
	Optimizer		Adam	
	Learning Rate		5e-4	
	α	-	0.0	1.0
	β	1.0	1.0	1.0
	Decoder σ		σ	
MORPHOMNIST	No. Layers		6	
	Training Steps		6000	
	No. Samples		60000	
	Batch Size		256	
	Optimizer		Adam	
	Learning Rate		1e-4	
	α	-	0.0	1.0
	β	1.0	1.0	1.0
	Decoder σ	0.5	0.5	0.5
CAUSAL3DIDENT	Input resolution		64×64	
	No. Layers	4	3	3
	Training Steps		19687	
	No. Samples		252000	
	Batch Size		128	
	Optimizer		Adam	
	Learning Rate		5e-4	
	Hidden dim		256	
	Latent dim	256	16	16
	α	-	1.0	1.0
	β	0.0	0.01	0.01
	Decoder σ	0.1	0.1	0.1

Table 4: MCC and COD results on MorphoMNIST and Causal3DIdent datasets

METHODS(↓), METRICS(→)	COD (↓)	MCC-R(↑)	MCC-SG(↑)
	MORPHOMNIST-TI		
VAE	1.31 ± 0.28	0.31	0.24 ± 0.06
MFC-VAE	1.33 ± 0.38	0.38	**0.39** ± 0.07
COVAE	**0.0**	**0.58**	0.38 ± 0.06
	MORPHOMNIST-TS		
VAE	1.47 ± 0.65	0.48	0.38 ± 0.05
MFC-VAE	1.75 ± 0.60	0.51	0.36 ± 0.06
COVAE	**0.0**	**0.56**	**0.41** ± 0.05
	CAUSAL3DIDENT		
VAE	22.39 ± 1.49	0.15	0.15 ± 0.0
MFC-VAE	**3.56** ± 0.87	**0.28**	**0.27** ± 0.01
COVAE	3.94 ± 0.86	0.26	0.25 ± 0.02