CAUSAL REPRESENTATION LEARNING AND INFERENCE VIA MIXTURE-BASED PRIORS

Avinash Kori [†,*] **Carles Balsells-Rodas** [†, *] **Ben Glocker**[†]

Yingzhen Li [†] **Francesco Locatello** [‡]

[†] Imperial College London
[‡] Institute of Science and Technology Austria
a.kori21@imperial.ac.uk

ABSTRACT

Causal Representation Learning (CRL) aims to uncover causal symmetries in the data-generating process with minimal assumptions and data requirements. The challenge lies in identifying the causal factors and learning their relationships, which is an inherently ill-posed problemEnsuring unique solutions, known as *identifiability*, is crucial but often requires strong assumptions or acccess to interventional or counterfactual data. In this work, we propose a novel approach that partitions the latent space: one component captures causal factorsusing diffeomorphic flows to model causal mechanisms, while the other accounts for exogenous noise. This structured decomposition enables our model to scale effectively to high-dimensional data and deep architectures. We establish theoretical guarantees for CRL by proving the identifiability of both causal factors and exogenous noise. Empirical results across multiple datasets validate our theoretical findings.

1 INTRODUCTION

Learning meaningful representations from unlabelled data is a fundamental challenge in deep learning (Bengio et al., 2013). The goal is to extract useful features or abstractions that capture the underlying structure of the data without relying on labelled examples. Early approaches on representation learning primarily focused on identifying statistically independent latent variables. Methods such as β-VAE (Higgins et al., 2017), InfoGAN (Chen et al., 2016), and other disentanglement techniques (Träuble et al., 2021; Liu et al., 2022), enforce independence constraints on the latent space, demonstrating success in controlled synthetic environments where the factors of variation are well-defined and manipulable (Locatello et al., 2019). A key challenge in learning meaningful and disentangled representations is ensuring *identifiability*—i.e., the ability to recover the true underlying generative process distribution (up to a simple transformation). Identifiability establishes an injective (one-to-one) mapping onto the observed data distribution, and guarantees that the learned representations correspond to intrinsic properties of the data, which preserve important invariances and reflect the true data generative process (Yao et al., 2024a). This concept is particularly relevant in the context of non-linear Independent Component Analysis (ICA), where recovering independent latent under arbitrary non-linear transformations is generally ill-defined. This intractability also extends to disentangled representation learning (Locatello et al., 2019).

However, in real-world scenarios, the relationships among variables are complex and structured, making disentangled representations insufficient for robust generalization. In contrast, learning the underlying causal representation offers improved generalisation, as causal structures remain invariant across changing environments (Schölkopf et al., 2021; Ahuja et al., 2022). Additionally, causal representations enhance interpretability by explaining why certain factors influence specific outcomes – an essential property in applications such as healthcare and autonomous systems (Pearl, 2009).

[*]Equal Contribution

Table 1: Non exhaustive list of literature, with their assumption on Structural Causal Models (SCM), mixing function, and data requirement, along with their identifiability criterion. Comparison of assumptions for identifiability.

METHOD	SCM	TRANSFORM	IDENTI.
Khemakhem et al. (2020a;b)	Independent	Non-linear	Permutation + Scaling
Falck et al. (2021); Kivva et al. (2022),	Independent	Piecewise-linear	Permutation + Scaling
Brehmer et al. (2022); Lippe et al. (2022b)	No restrictions	Non-linear	-
Ahuja et al. (2022)	No restrictions	Polynomial	Affine
Yao et al. (2024a;b); Von Kügelgen et al. (2021)	No restrictions	Non-linear	Block-identifiability
Komanduri et al. (2024); An et al. (2023), Yang et al. (2021)	ANM	Non-linear	-
Proposed	Diffeomorphic	Piecewise-linear	Permutation + Scaling

Existing work in causal representation learning (CRL) has largely focused on leveraging invariances and data symmetries to achieve identifiable representations from observational data (Yao et al., 2024a; Khemakhem et al., 2020a;b; Willetts & Paige, 2021; Hyvärinen et al., 2023). Discovering the dependency structure in the latent space is at the core of causal representation learning (CRL) (Schölkopf et al., 2021). However, the majority of previous works in CRL rely on interventional (Ahuja et al., 2022; Varici et al., 2023) or counterfactual (Locatello et al., 2020; Brehmer et al., 2022; Lippe et al., 2022b) data to achieve identifiability. For instance, Komanduri et al. (2024) introduces a causal auto-encoder framework that utilizes a known directed acyclic graph (DAG) for counterfactual generation, learning the both generating process and the associated causal mechanisms.

Despite these advances, most CRL methods are limited to low-dimensional settings, where the number of latent variables is relatively small. This facilitates tractable learning of complex feature dependencies. Generative models for high-dimensional data, such as hierarchical VAEs or diffusion models (Kingma et al., 2021), often require thousands of latent variables to generate high-fidelity images. Treating all latent variables as causal factors in such models is impractical, highlighting the need to separate causal factors from other latent variables, which could instead capture independent style variations. In this work, we extend the identifiability guarantees established in Khemakhem et al. (2020a) and Kivva et al. (2022), to settings with a partitioned latent space. Specifically, we propose a framework that learns both the causal factors and their dependency structure, while preserving key identifiability properties. Our approach requires only mild model and distributional assumptions and is designed to scale effectively to high-dimensional data.

In Table 1, we summarize existing CRL methods and their corresponding constraints on the mixing function. We propose *Causal Representation Learning and Inference with Mixture Based priors* (CLIMB), which relaxes the restrictions on structural causal models (SCM). Our method extends beyond the identification of data symmetries by enforcing *diffeomorphic* mappings between estimated causal factors and demonstrating their utility through counterfactual inference. Figure 1 provides an overview of our approach. Our main contributions are as follows: (i) *Latent Space Partitioning*: We introduce a novel latent structure that models causal factors (using diffeomorphic flows) and exogenous noise, improving scalability; (ii) *Identifiability Guarantees*: We provide theoretical guarantees for identifiable representations, even when partitioning the latent space; and (iii) *Empirical Validation*: We demonstrate the theoretical and practical advantages of our approach through extensive experiments in causal representation learning.

2 RELATED WORKS

Identifiable Representation Learning. Identifiability in representation learning originates from early work on Independent Component Analysis (ICA) (Hyvärinen & Pajunen, 1999; Hyvarinen & Oja, 2000), and has recently gained renewed interest (Hyvarinen & Morioka, 2016; Hyvarinen et al., 2019a; Locatello et al., 2019; Khemakhem et al., 2020a; Von Kügelgen et al., 2021; Lachapelle et al., 2024; Yao et al., 2024b). Several strategies have been developed to address this challenge: (i) restricting the class of mixing functions; (ii) leveraging non-i.i.d., interventional, or counterfac-

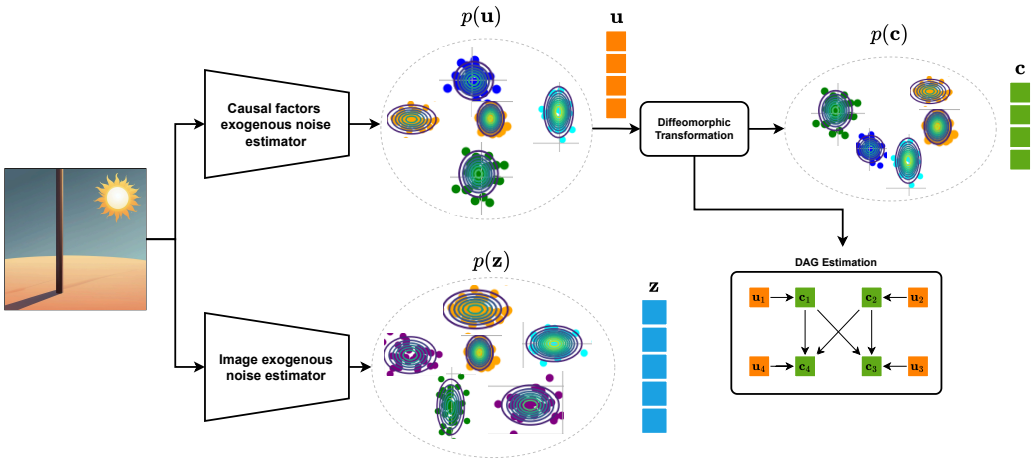


Figure 1: Proposed algorithm: an input, $\mathbf{x} \in \mathcal{X}$, is processed through two distinct estimators—one for estimating the exogenous noise $\mathbf{z}$ of an input and another for estimating the exogenous noise $\mathbf{u}$ of the causal factors. The estimated causal exogenous noise is then transformed via diffeomorphic functions, producing the causal factors $\mathbf{c}$. Additionally, the Jacobian of the diffeomorphic function aids in estimating the underlying directed acyclic graph (DAG) $\boldsymbol{A}$. Finally, the mixing function combines sampled values of $\mathbf{c}$ and $\mathbf{z}$ from the posterior to reconstruct the given input.

tual data; and (iii) imposing structure on the latent space through distributional assumptions. For (i), restricting mixing functions to conformal maps (Buchholz et al., 2022) or volume-preserving transformations (Yang et al., 2022) has been found to enable identifiability. For (ii), contrastive learning approaches utilize paired observations (Zimmermann et al., 2021; Locatello et al., 2020; Brehmer et al., 2022; Ahuja et al., 2022; Von Kügelgen et al., 2021), obtained via data augmentation, interventions, or approximate counterfactual inference, to disentangle latent factors. Regarding (iii), latent space structure can be enforced either by introducing auxiliary variables to induce conditional independence among latent variables (Hyvarinen et al., 2019b; Khemakhem et al., 2020a;b) or by imposing prior distributional constraints, such as a mixture priors in VAEs (Dilokthanakul et al., 2016; Willetts & Paige, 2021; Kivva et al., 2022).

Causal Representation Learning. Learning causal representations is particularly feasible when interventional or non-i.i.d. data is available. Ahuja et al. (2022) employ an injective polynomial decoder trained on both observational and interventional data. Locatello et al. (2020) utilize conterfactual data by capturing observations before and after unknown interventions, while Brehmer et al. (2022) extend this approach to more complex causal graphs. For non-i.i.d. settings, Lippe et al. (2022b) extract causal factors from spatio-temporal data through interventions over time. Some methods assume partial supervision, leveraging ground-truth causal factors, while others, such as Yang et al. (2021), explicitly model exogenous noise and map it to causal latent variables via a linear SCM. Recently, Yao et al. (2024a) proposed a unified framework that preserves known invariances in the data, such as those arising from multi-view, temporal, or counterfactual settings. However, one missing component in existing methods is structural assumptions governing the interplay between latent distribution and data-generating functions. To address this, we assume a Gaussian Mixture Model (GMM) prior in the latent space, enabling the estimation of causal structure from purely observational data with minimal assumptions. Our approach integrates constraints inspired by causal discovery techniques (Glymour et al., 2019; Zhang et al., 2024), relaxing the reliance on interventional or counterfactual data.

3 Formalism

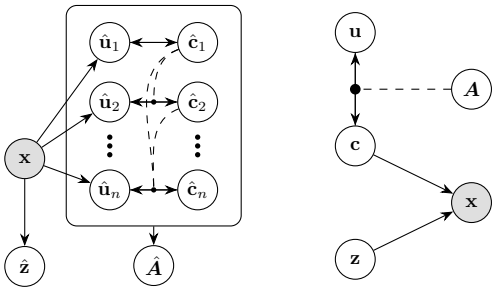
Our approach focuses on estimating the exogenous noise ($\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$) for an observed variable ($\mathbf{x} \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^{H \times W \times C}$) alongside learning the causal factors. These causal factors consist on both endogenous ($\mathbf{c}$) and exogenous ($\mathbf{u}$) components, where $\mathbf{c} \in \mathcal{C}$ and $\mathbf{u} \in \mathcal{U}$ with $\mathcal{C}, \mathcal{U} \subseteq \mathbb{R}^n$. We

formalize the relationship between $\mathbf{x}, \mathbf{z}, \mathbf{c}$, and $\mathbf{u}$ by defining the following mappings: $\phi_z : \mathcal{X} \to \mathcal{Z}$ maps an observed input $\mathbf{x}$ to its corresponding exogenous noise vector $\mathbf{z}$. $\phi_u : \mathcal{X} \to \mathcal{U}$ extracts exogenous noise of the parent variables in the causal model, $\mathbf{u}$. Finally,$\psi_x : \mathcal{C} \times \mathcal{Z} \to \mathcal{X}$ denotes the mixing function that generates $\mathbf{x}$ from the causal factors $\mathbf{c}$ and exogenous noise $\mathbf{z}$. A comprehensive list of notations is provided in appendix A. Given the generative model $\mathbf{x} = \psi_x(\mathbf{c}, \mathbf{z})$, we denote $\mathbf{c} = \psi_x^{-1}(\mathbf{x}; \mathbf{z})$ as the inverse mapping for a fixed $\mathbf{z}$. We provide further modelling details in the following sections.

Assumption 1. (n-causal factors) We assume access to the number of causal factors n.

3.1 Generative model

We introduce structured latent space partitioning; in contrast to existing methods that learn a single latent space while observed causal signals as instrumental or conditioning variables (Komanduri et al., 2024; Khemakhem et al., 2020a;b). Our approach separates latent variables into causal factors ($\mathbf{c}$), which help identifying the DAG that captures their relationships, and independent latent variables $\mathbf{z}$, which account for other generative factors. In high-dimensional data, generative models must capture not only the causal factors, but also additional elements required for reconstruction even if they lack causal dependencies. To address this, we decompose the latent space into $\mathcal{C}$ and $\mathcal{Z}$. For instance, consider a pendulum dataset with image observations. The pendulum's motion is governed by few causal factros (e.g. angle, angular velocity), but representing the full image requires additional latent variables. Expanding the latent space in traditional methods would make SCM/DAG estimation computationally infeasible. Our method mitigates this by explicitly disentangling causal and non-causal variables. The probabilistic graphical model is illustrated in Figure 2(b). The generative process starts with independent exogenous variables: where $\mathbf{z}$ denotes noise directly for observed variables, and $\mathbf{u}$ denotes noise associated to causal factors. Next, a DAG $\boldsymbol{A}$ with prior $p(\boldsymbol{A})$, generates the causal factors $\mathbf{c}$ based on $\mathbf{u}$. Finally, combining $\mathbf{c}$ with $\mathbf{z}$ generate the observed variable $\mathbf{x}$, ensuring both causal and generative aspects of the data are captured. Here, we relax $p(\boldsymbol{A})$ to follow a continuous distribution, modelled using a `RelaxedBernoulli`. The full generative model is described as follows:



(**a**) Inference Model (**b**) Generation

Figure 2: **Graphical Model for CLIMB**. (a) shows the inference model, where variables $\mathbf{z}$ and $\mathbf{u}$ are estimated from $\mathbf{x}$, and $\mathbf{u}$ is transformed through flow-based functions to yield $\mathbf{c}$. (b) illustrates the generative process: sampled values of $\mathbf{u}$ and the graph structure $\boldsymbol{A}$ are transformed to produce $\mathbf{c}$, which is then combined with sampled $\mathbf{z}$ to generate $\mathbf{x}$. The dashed lines in both figures illustrate conditioning variables.

$$p(\mathbf{x}) = \iiint p(\mathbf{x} \mid \mathbf{c}, \mathbf{z}) p(\mathbf{z}) p(\mathbf{c}|\boldsymbol{A}) p(\boldsymbol{A}) \, d\mathbf{c} \, d\mathbf{z} \, d\boldsymbol{A}. \tag{1}$$

3.2 Causal factors estimation

Causal factors are estimated as exogenous noise transformations, conditioned on the information encoded in the DAG $\boldsymbol{A}$. Specifically, we define each causal factor as $\mathbf{c}_i = g_i(f_i(\mathbf{pa}_i^{\boldsymbol{A}}), \mathbf{u}_i)$, where $\mathbf{pa}_i^{\boldsymbol{A}}$ denotes the parents of $\mathbf{c}_i$ with respect to $\boldsymbol{A}$ ($\mathbf{pa}_i \subseteq \{\mathbf{c}_0, \dots \mathbf{c}_n\} \setminus \mathbf{c}_i$). The functions f_i and g_i model the conditional dependencies and transformations, respectively.

Prior distribution Since we learn causal variables implicitly, our prior should account for their complexity under different conditions. To achieve this, we model the base distribution $p(\mathbf{u})$ as a Gaussian Mixture Model (GMM):

$$p(\mathbf{u}) = \sum_{k=0}^{K} \pi_k \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2). \tag{2}$$

Given assumption 3, we transform $\mathbf{u}_i$ into $\mathbf{c}_i$ using a multi-layered diffeomorphic flow: $g_i = g_i^1 \circ \cdots \circ g_i^L$.

$$\mathbf{c}_i := g_i(f_i(\mathbf{pa}_i), \mathbf{u}_i), \quad p(\mathbf{c}_i) = p(\mathbf{u}_i) \left| \det \mathbf{J}_{g_i^{-1}}(\mathbf{c}_i) \right| \tag{3}$$

where $\mathbf{J}_{g_i^{-1}}$ is the Jacobian of the inverse transformation g_i^{-1}. If g_i consists of affine flows, $p(\mathbf{c})$ remains GMM-distributed with transformed Gaussian components. For more expressive transformations, such as spline or any non-linear flows, the resulting $p(\mathbf{c})$ is still a mixture distribution.

Assumption 2. (f_i-function type) We assume piece-wise linear models for estimating conditioning signals using parents of node i.

Assumption 3. (g_i-function type) We assume g_i to be diffeomorphic functions, transforming the exogenous variable $\mathbf{u}$ into the endogenous variable $\mathbf{c}$.

Posterior distribution Given the DAG is estimated using the causal factors, the posterior model for estimating $\mathbf{c}$ is slightly different from the prior defined above. Similar to Lippe et al. (2022a); Kyono et al. (2020), we model the posterior of feature $\mathbf{c}_i$ using all features except $\mathbf{c}_i$, which we represent using $\mathbf{c}_{-i} = \{\mathbf{c}_0, \dots \mathbf{c}_n\} \setminus \mathbf{c}_i$, resulting in posterior as equation 4, where q_u is considered to be GMM. A key distinction in this posterior model is the use of alternative sets of flow and conditioning functions, denoted by $\hat{g}$ and $\hat{f}$, respectively, which differ from those used in the prior, allowing greater flexibility. In our ablations we consider different class of function for both $\hat{f}$ and $\hat{g}$. Notably, the posterior $q(\mathbf{c})$ structure closely resembles the pseudo-likelihood approach, due to potentially over-counting the involved variables. In terms of sampling, we sample individual variable $\mathbf{c}_i$ at a time, creating a Gibbs sweep.

$$q(\mathbf{c}) = \prod_{i=0}^{n} q(\mathbf{c}_i \mid \mathbf{c}_{-i}), \quad q_c(\mathbf{c}_i \mid \mathbf{c}_{-i}) = q_u(\hat{g}_i^{-1}(\hat{f}_i(\mathbf{c}_{-1}), \mathbf{c}_i)) \left| \det \mathbf{J}_{\hat{g}_i^{-1}}(\mathbf{c}_i) \right| \tag{4}$$

3.3 DAG ESTIMATION

For graph prior $p(\boldsymbol{A})$ we consider `RelaxedBernoulli` distribution with linearly decaying temperature, which can be descrbed as a `sigmoid` transformation of samples from `Gumbel(0,1)` distribution. Additionally, simialr to Zheng et al. (2018); Geffner et al. (2022) we include an acyclicity factor defined by equation 5 providing an inductive bias for DAG generation, which is non-negative and zero only when the graph is DAG.

$$h(\boldsymbol{A}) = \operatorname{tr}(\exp(\boldsymbol{A} \odot \boldsymbol{A})) - n \tag{5}$$

Posterior For graph posterior $q(\boldsymbol{A} \mid \mathbf{c})$, we learn logits with a network mapping causal factors to the parameters of the distribution for graph edges. Similar to the graph prior we use `RelaxedBernoulli` distribution to facilitate re-parametrised sampling for end to end gradient propagation. To minimise variance in the posterior we consider the batched average of all logits as parameters for `RelaxedBernoulli`.

3.4 INPUT-EXOGENOUS NOISE ESTIMATION

We model $\mathbf{z}$ as an exogenous noise variable for a given input $\mathbf{x}$. For this, similar to Kivva et al. (2022); Willetts & Paige (2021) we consider the prior distribution $p(\mathbf{z})$ to be GMMs with M components, providing us with the identifiability guarantees on exogenous noise variables. We consider the *local* input conditioned posterior $q(\mathbf{z} \mid \mathbf{x})$ to be modelled using Gaussian with learned mean and diagonal covariance. This model learns mutually independent variables, unlike $p(\mathbf{c})$, where the interdependencies are also captured.

3.5 MIXING FUNCTION AND TRAINING OBJECTIVE

We consider the mixing function $\psi_x : \mathcal{C} \times \mathcal{Z} \to \mathcal{X}$ mapping causal factor conditioned exogenous noise to create an observational data. In this work, we perform conditioning as concatenation of $\mathbf{z} \oplus \mathbf{c}$ which is passed through piece-wise linear mixing functions as in definition 1.

Definition 1 (Piece-wise linear functions)**.** Let $\mathbf{c}, \mathbf{z}$ denote vectors sampled from $p(\mathbf{z})$ and $p(\mathbf{c})$ respectively. Let $\sigma : \mathbb{R} \to \mathbb{R}$ denote the leaky-ReLU activation function, and let $H(n_1, n_2)$ denote the set of full-rank affine functions $h_i : \mathbb{R}^{n_i} \to \mathbb{R}^{n_j}$. We consider piece-wise functions mapping pair of $\mathbf{c} \in \mathcal{C}, \mathbf{z} \in \mathcal{Z}$ to an input $\mathbf{x} \in \mathcal{X} \in \mathbb{R}^m$ in the output space, $\mathcal{F}_\sigma^{nk \to m} : \mathcal{C} \times \mathcal{Z} \to \mathcal{X}$, of the form below:

$$\mathcal{F}_\sigma^{n_0,\dots,n_t} = \Big\{ h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \cdots \sigma \circ h_1 \mid h_i \in H(n_{i-1}, n_i) \Big\}. \tag{6}$$

Probabilistically, the resulting generative model can be described by a graphical model in Figure 2(b), this results in the likelihood as expressed in equation 7. To train our model in an end-to-end fashion, we maximise the log-likelihood, resulting in the evidence lower bound (ELBO), equation 9. Here, we consider the distribution $p(\mathbf{x} \mid \mathbf{c}, \mathbf{z})$ to be Gaussian with learnable mean and isotropic covariance.

$$\log p(\mathbf{x}) = \log \iiint p(\mathbf{x} \mid \mathbf{c}, \mathbf{z}) p(\mathbf{z}) p(\mathbf{c}|\boldsymbol{A}) p(\boldsymbol{A}) \, d\mathbf{c} \, d\mathbf{z} \, d\boldsymbol{A} \tag{7}$$

$$\geq \iiint q(\mathbf{c} \mid \mathbf{x}) q(\mathbf{z} \mid \mathbf{x}) q(\boldsymbol{A} \mid \mathbf{c}) \log \frac{p(\mathbf{x} \mid \mathbf{c}, \mathbf{z}) p(\mathbf{z}) p(\mathbf{c} \mid \boldsymbol{A}) p(\boldsymbol{A})}{q(\mathbf{c} \mid \mathbf{x}) q(\mathbf{z} \mid \mathbf{x}) q(\boldsymbol{A} \mid \mathbf{c})} \tag{8}$$

$$= \int q(\mathbf{c}, \mathbf{z} \mid \mathbf{x}) \log p(\mathbf{x} \mid \mathbf{c}, \mathbf{z}) - \mathrm{KL}\left(q(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right) - \mathrm{KL}\left(q(\boldsymbol{A} \mid \mathbf{c}) \| p(\boldsymbol{A})\right)$$

$$- \mathbb{E}_{q(\boldsymbol{A}|\mathbf{c})} \mathrm{KL}\left(q(\mathbf{c} \mid \mathbf{x}) \| p(\mathbf{c} \mid \boldsymbol{A})\right) \tag{9}$$

Remark 1. We use variational posterior for DAGs for faster training; during inference and for proofs, we rely on the DAG generated wrt trained SCMs.

3.6 Counterfactual generation

Conterfactual generation refers to the generation of retrospective hypothetical scenarios, something like "*what would have happened, if had I done this instead of that?*". Methodically, counterfactual generation can be seen as three-step process Pearl (2009): (i) *Abduction:* inferring exogenous noise of input and all the involved causal factors, in our framework, it corresponds to the inference of $\mathbf{z}$ and $\mathbf{u}$; (ii) *Action:* intervene on a causal factor of interest ($\mathbf{c}_i \to \bar{\mathbf{c}}_i$); (iii) *Prediction:* this involves the propagation of effects of changes on the intervened causal factor. To quantitively evaluate the quality of generated counterfactual, we consider two properties, Composition and Reversibility, as proposed in Monteiro et al. (2023); here, we mainly perform a qualitative evaluation of these properties, due to which we do not consider Effectiveness, in-depth analysis on these properties on CLIMB is left as a future work.

Composition: measures the divergence in the generated image with respect to the original input when the model is intervened on variables to have the value it would otherwise have without the intervention. This basically measures the model behaviour under null interventions.

Reversibility: This measure of the divergence of the model being truly invertible; in an ideal scenario, the considered model must be cycle-consistent. This is measured by calculating the distance between the original and cycled-back observation.

4 Theoretical Analysis

In this section, we leverage on the properties discussed in section 3 to theoretically demonstrate the how the proposed model preserves symmetries in the dataset and also learn causal relationships among these symmetries. Here, we demonstrate identifiability of the exogenous noise distribution of causal factors $p(\mathbf{u})$ in Theorem 1 and demonstrate the identifiability of $\boldsymbol{A}$ in Theorem 4 under *faithfulness* assumption.

Assumption 4. (Sufficient variability) We assume to have an access to $K = n + 1$ environments, such that all $K - 1$ vectors of type $\mathbf{v}_k^2 - \mathbf{v}_1^2$ and $\mathbf{v}_k^1 - \mathbf{v}_1^1, k \in \{2, \dots, K\}$ are linearly independent respectively, where $\mathbf{v}_k^2 := \left(\frac{\partial^2 \eta_{0k}}{\partial u_0^2}, \dots, \frac{\partial^2 \eta_{nk}}{\partial u_n^2} \right), \mathbf{v}_k^1 := \left(\frac{\partial \eta_{0k}}{\partial u_0}, \dots, \frac{\partial \eta_{nk}}{\partial u_n} \right)$, with $\eta_{ik} = \log p(u_i \mid k)$.

Table 2: Identifiability results for both $\mathbf{u}$ and $\mathbf{z} \oplus \mathbf{c}$ factors across five runs

METHOD	PENDULUM		MORPHOMNIST		COLOR-MORPHOMNIST	
	$\mathbf{u}$-MCC ↑	$\mathbf{z} \oplus \mathbf{c}$-MCC ↑	$\mathbf{u}$-MCC ↑	$\mathbf{z} \oplus \mathbf{c}$-MCC ↑	$\mathbf{u}$-MCC ↑	$\mathbf{z} \oplus \mathbf{c}$-MCC ↑
$\beta-$VAE	-	$0.36 \pm .03$	-	$0.32 \pm .02$	-	$0.34 \pm .03$
VADE	-	$0.40 \pm .03$	-	$0.38 \pm .03$	-	$0.39 \pm .02$
IVAE	-	$0.36 \pm .03$	-	$0.46 \pm .04$	-	$0.48 \pm .06$
Ours:						
CLIMB	0.76 ± 0.03	0.53 ± 0.02	0.88 ± 0.03	$0.76 \pm .02$	0.92 ± 0.03	$0.84 \pm .01$
CLIMB-GL	$\mathbf{0.78 \pm 0.05}$	$\mathbf{0.55 \pm .04}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.84 \pm .01}$	$\mathbf{0.92 \pm 0.01}$	$\mathbf{0.86 \pm .02}$

Remark 2. The linear independence criterion is almost always satisfied as long as the model parameters are randomly generated - in that case it almost surely holds as singular solutions will lie in a submanifold (Hälvä & Hyvarinen, 2020; Khemakhem et al., 2020a).

Definition 2. ($\sim_\tau$ − Translation-scaling equivalence) For $\theta = \{\psi_x, \mathbf{p}\}$ a set of parameters of the mixing function and prior, the equivalence $\sim_\tau$ on θ is defined as:

$$(\psi_x, \mathbf{p}) \sim_\tau (\tilde{\psi}_x, \tilde{\mathbf{p}}) \iff \exists \quad s_i, b_i \in \mathbb{R}\ \forall i \in [n]$$
$$s.t. \quad \psi_x^{-1}(\mathbf{x}; \mathbf{z})_i = s_{\tau(i)} \tilde{\psi}_x^{-1}(\mathbf{x}; \mathbf{z})_{\tau(i)}) + b_{\tau(i)}, \forall \mathbf{x} \in \mathcal{X}, \tag{10}$$

where $s_{\tau(i)}, b_{\tau(i)}$ are element wise scaling and translation terms with permutation function τ.

Theorem 1. *($\mathbf{u}$-identifiability) Assuming the data generation process follow equation 1, with an invertible demixing function $\phi_{uz}(\mathbf{x})$, such that $(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = (\phi_u(\mathbf{x}), \phi_z(\mathbf{x}))$, and a prior distribution $p(\mathbf{u})$ is modelled as a non-degenerate GMM. Given assumptions 1 and 4, the causal exogenous noise $\mathbf{u}$ is identifiable up to $\sim_\tau$ equivalence, as defined in 2.*

Remark 3. It is important to note that, the following theory holds for any mixture which is analytic and closed under affine transformations (Kivva et al., 2022). Alternatively, the sufficient variability assumption requires a modification, where $K = 2n + 1$ and the vectors $(\mathbf{v}_k^2 - \mathbf{v}_1^2, \mathbf{v}_k^1 - \mathbf{v}_1^1) \in \mathbb{R}^{2n}$, with $k \in \{2, \ldots, K\}$, are linearly independent (Yao et al., 2022)

Remark 4. The equivalence relation can be further strengthened by enforcing autoregressive flows in SCMs, removing the permutation requirement, however we do not consider them.

Proof Sketch. To prove this, we proceed in following steps: (i) Identifiability of $p(\mathbf{x}|k)$ given $p(\mathbf{x})$; (ii) We derive the disentanglement result between $\mathbf{u}$ and $\mathbf{z}$; (iii) We derive the Jacobian structure for the considered mixing function; (iv) We derive the second order partial derivative of log-likelihood wrt two independent elements of $\mathbf{u}$; and (v) We analyse the resulting expression to demonstrate when the variables are identifiable. □

Theorem 2 ($\boldsymbol{A}-$identifiability)**.** *Given $p(\mathbf{u})$ is element-wise identifiable from theorem 1, and diffeomorphic transformation from $\mathbf{u}$ to $\mathbf{c}$. Then with an assumption 6 we can recover DAG up to $\sim_{node}$ equivalence, definition 6.*

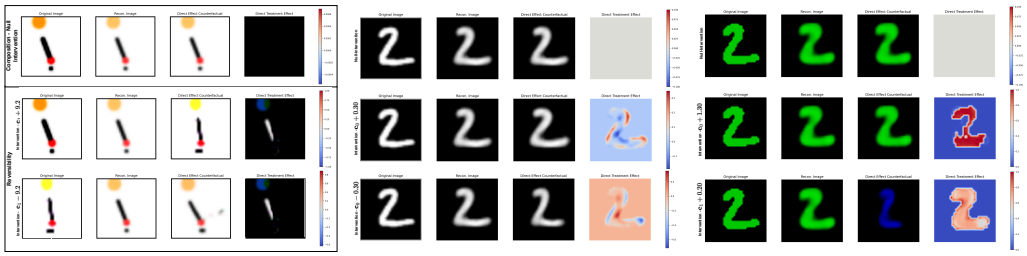
Proof Sketch. To prove this result, we proceed with the following steps: (i) we establish the relations between $\mathbf{u}, \hat{\mathbf{u}}$ and $\mathbf{c}, \hat{\mathbf{c}}$; (ii) Find the Jacobian structure; and finally (iii) analyse the Jacobians to establish $\boldsymbol{A}-$ identifiability. □

5 EMPIRICAL EVALUATION

Given the work's theoretical focus, our experimental aim is to provide strong empirical evidence supporting our identifiability claims. For that, we conduct experiments on standard imaging benchmarks, including the PENDULUM, MORPHOMNIST, and COLOR-MORPHOMNIST datasets, with image resolution of $64 \times 64, 32 \times 32$ and 32×32 respectively. In our evaluation, we perform both qualitative and quantitative assessments. For baselines, we consider several existing models, including β-VAE (Kingma & Welling, 2013), VADE (Jiang et al., 2016), and IVAE (Khemakhem et al., 2020a). We propose two variants: one with linear diffeomorphic flows using affine flows, and another with non-linear diffeomorphic flows using spline flows.

Quantitative Evaluation. For evaluating the necessary conditions of CRL, we primarily focus on measuring self-consistency, by computing MCC across multiple runs. For this, to evaluate against baselines, we consider the concatenated features $\mathbf{z} \oplus \mathbf{c}$, we make sure that the dimensions of latent features in baselines match with the dimension of these concatenated features. Table 2, illustrates all the results, based on the results we can observe significant improvement wrt baselines, indicating models ability to better align with the data symmetries present in the data generating process. We can associate this behaviour to structural dependency learning present in CLIMB. Additionally, to evaluate CRL sufficiency, we measure $\mathbf{u}$-MCC with respect to the ground-truth causal factors and compute the structural Hamming distance (SHD) of an estimated DAG. We tabulate all our findings in Table 1, which indicates that in most cases we can identify the correct DAG, reflecting the true data generating process.

Qualitative Evaluation. To showcase model behaviour in generating counterfactuals, we evaluate its composition and reversibility properties. In Figure 3(a) we illustrate reversibility behaviour of the model on PENDULUM dataset, where we can the models capabilities to recover the original image with negative intervention on given image. For visual illustration, we plot treatment effect graphs with null interventions and cycled-back transformations, demonstrating the model's ability to capture these causal dynamics, this can be observed in Figure 3(b,c).



(**a**) PENDULUM (**b**) MORPHOMNIST (**c**) COLOR-MORPHOMNIST

Figure 3: Illustration of counterfactual generation on inferred causal factors: (a) demonstrates the reversibility and composition on PENDULUM dataset, (b) and (c) describes intervention on multiple causal factors on MORPHOMNIST and COLOR-MORPHOMNIST datasets.

6 DISCUSSION

Relation to invariance principles. Yao et al. (2024a) introduces a unified framework for understanding CRL through invariance principles, connecting various existing approaches. Their formulation of invariance is described by $\mathbf{c}_1 := \bar{\phi}_1(\mathbf{x}_1)_B$ and $\mathbf{c}_2 := \bar{\phi}_2(\mathbf{x}_2)_B$, ensuring that for any $z \notin \mathbf{c}_1, \mathbf{c}_2$, we have $\frac{\partial h_1(\mathbf{c}_1)}{\partial z} = \frac{\partial h_2(\mathbf{c}_2)}{\partial z} = 0$, where $\bar{\phi}_i$ and h_i denote input-specific encoders and smooth transformations, respectively, while $\mathbf{c}_i$'s correspond to invariant subset of features given by B. Where the invariance results relies on feature similarity across distinct inputs to maintain this invariance. In contrast, by partitioning our latent space into $\mathcal{C}$ and $\mathcal{Z}$, we inherently achieve $\frac{\partial \mathbf{c}}{\partial \mathbf{z}} = 0$, as demonstrated in the proof of Theorem 1, where we leverage on a mixture-based pivoting strategy. Moreover, with the assumption of mixture distribution based priors, we further ensure identifiability up to permutation and scaling.

Limitations: (i) The piecewise linear dependency may be limiting in certain scenarios, impacting performance in more complex settings; (ii) an extensive empirical evaluation on large-scale datasets and with deeper models is essential to fully understand the method's applicability and robustness from causal inference and counterfactual generation scenarios; (iii) extending identifiability proof of $\mathbf{c}$ in the case of general transformation; and finally, (iv) this work does not account for discrete causal factors, which could be a valuable direction for future research to explore assumptions and adaptations needed for discrete variables.

In conclusion, we introduce a novel algorithm through latent space partitioning for CRL. By structuring the latent space to independently capture causal factors and separate exogenous noise, enhancing

scalability, which we empirically demonstrated to an extent on imaging datasets. We also establish identifiability guarantees for causal factors, exogenous noise, and the underlying DAG using observational data alone in a mixture prior setting.

REFERENCES

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

SeungHwan An, Kyungwoo Song, and Jong-June Jeon. Causally disentangled generative variational autoencoder. In *ECAI 2023*, pp. 93–100. IOS Press, 2023.

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.

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.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

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.

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.

Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

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.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

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

Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pp. 939–948. PMLR, 2020.

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.

A Hyvarinen and E Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

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

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019a.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 859–868. PMLR, 2019b.

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

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.

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*, pp. 2207–2217. PMLR, 2020a.

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, 2020b. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/962e56a8a0b0420d87272a682bfd1e53-Paper.pdf.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

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

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.

Aneesh Komanduri, Chen Zhao, Feng Chen, and Xintao Wu. Causal diffusion autoencoders: Toward counterfactual generation via diffusion probabilistic models. *arXiv preprint arXiv:2404.17735*, 2024.

Avinash Kori, Pedro Sanchez, Konstantinos Vilouras, Ben Glocker, and Sotirios A Tsaftaris. A causal ordering prior for unsupervised representation learning. *arXiv preprint arXiv:2307.05704*, 2023.

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=eYciPrLuUhG.

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*, pp. 13557–13603. PMLR, 2022b.

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.

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*, 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.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. *arXiv preprint arXiv:2303.01274*, 2023.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Patrik Reizinger, Yash Sharma, Matthias Bethge, Bernhard Schölkopf, Ferenc Huszár, and Wieland Brendel. Jacobian-based causal discovery with nonlinear ica. *Transactions on Machine Learning Research*, 2023.

Patrik Reizinger, Siyuan Guo, Ferenc Huszár, Bernhard Schölkopf, and Wieland Brendel. Identifiable exchangeable mechanisms for causal structure and representation learning. *arXiv preprint arXiv:2406.14302*, 2024.

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.

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.

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

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.

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

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

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.

Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, and Junchi Yan. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. *arXiv preprint arXiv:2409.02772*, 2024a.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. 2024b.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Vi-sZWNA_Ue.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

A NOTATIONS

- $\mathbf{u} \in \mathcal{U} = \mathcal{U}_0 \times \cdots \times \mathcal{U}_n \subseteq \mathbb{R}^n$ prior noise factors, with distribution $p(\mathbf{u}) = \sum_j \pi_j \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}_j, \boldsymbol{\sigma}_j^2)$
- $\hat{\mathbf{u}} \in \mathbb{R}^n$ inferred noise for causal factors, with distribution $q(\mathbf{u} \mid \mathbf{x}) = \mathcal{N}(\mathbf{u}; \boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\sigma}^2(\mathbf{x}))$
- $\boldsymbol{A} \in \mathcal{A} \subseteq [0,1]^{n \times n}$, prior adjacency matrix with probability distribution $p(\boldsymbol{A}) = \text{RelaxedBernoulli}(\boldsymbol{A})$
- $\hat{\boldsymbol{A}} \in \mathcal{A} \subseteq [0,1]^{n \times n}$, inferred adjacency matrix with probability distribution $q(\boldsymbol{A} \mid \mathbf{c}, \mathbf{u})$
- $\mathbf{c} \in \mathcal{C} = \mathcal{C}_0 \times \cdots \times \mathcal{C}_n \subseteq \mathbb{R}^n$ prior causal factors, with distributions $p(\mathbf{c} \mid \mathbf{u}, \boldsymbol{A})$.
- $\hat{\mathbf{c}} \in \mathbb{R}^n$ inferred causal factors, with distributions $q(\mathbf{c} \mid \mathbf{u})$.
- $\mathbf{z} \in \mathbb{R}^{d_z}$ prior exogenous noise for an image, $p(\mathbf{z})$
- $\hat{\mathbf{z}} \in \mathbb{R}^{d_z}$ inferred exogenous noise for a given image, $q(\hat{\mathbf{z}} \mid \mathbf{x})$
- $g_i, \hat{g}_i : \mathcal{C}_i \to \mathcal{U}_i$ is a diffeomorphic function
- $f_i, \hat{f}_i : \bar{\mathcal{C}} \to \mathcal{C}_i$, conditioning network, where conditioning is applied wrt DAG $\boldsymbol{A}$ and $\bar{\mathcal{C}} \subset \mathcal{C}$
- $\mathbf{T}, \hat{\mathbf{T}} : \mathcal{U} \to \mathcal{C}$ both correspond to prior and posterior flows, reflecting the joint operation of $\{g_i \;\; \forall\, i \in [n]\}$
- $\phi_z : \mathcal{X} \to \mathcal{Z}, \phi_u : \mathcal{X} \to \mathcal{U}$, functionally $\hat{\mathbf{z}} = \phi_z(\mathbf{x}), \hat{\mathbf{c}} = \hat{\mathbf{T}}(\phi_c(\mathbf{x}))$
- $\psi_x : \mathcal{C} \times \mathcal{Z} \to \mathcal{X}$, functionally $\mathbf{c} = \mathbf{T}(\mathbf{u}), \mathbf{x} = \psi_x(\mathbf{c}, \mathbf{z})$

B ADDITIONAL DEFINITIONS

Definition 3. (Diffeomorphic functions) A function f is said to be a diffeomorphism onto its image, *i.e.,* f is C^∞, if f is injective and its inverse f^{-1} is C^∞ as well.

Definition 4. (Data symmetries) For a given model $\mathbf{z} = \psi^{-1}(\mathbf{x})$, the model ψ is said to capture data symmetries if the recovered factors $\mathbf{z}$, are invariant under a transformation T such that:

$$\psi^{-1}(\mathbf{x}) = T(\psi^{-1}(\mathbf{x})),$$

where T represents a symmetry-preserving transformation in the data-generating process. In this sense, ψ can identify factors in $\mathbf{z}$ up to the transformation T, meaning that the decomposition achieved by ψ respects inherent relations in the data-generating process.

Definition 5. (Causal symmetries) Similar to the definition 4, the transformation T can be categories to capture causal symmetries when it can be decomposed wrt individual element as:

$$\psi^{-1}(\mathbf{x})_i = T_i(\psi^{-1}(\mathbf{x})_{\tau(i)}, \mathbf{pa}^{\boldsymbol{A}}_{\tau(i)}) \; \forall i \in [|\psi^{-1}(\mathbf{x})|],$$

where τ is a permutation, T_i is a symmetry transformation specific to each factor i that preserves the causal structure, and $\boldsymbol{A}$ correspond to underlying causal dependencies in the data-generating process.

Definition 6. ($\sim_{node}$ − Node-wise permutation equivalence) For two different adjacency matrix $\boldsymbol{A}, \hat{\boldsymbol{A}}$, the equivalence relation $\sim_{node}$, with permutation matrix P for an associated permutation π; i.e. $P_{i\pi(i)} = 1$ for any $i \in [n]$, and 0 otherwise, is defined as:

$$\boldsymbol{A} \sim_{node} \hat{\boldsymbol{A}} \iff \exists \quad P \in \{0,1\}^{n \times n}; \quad s.t. \quad \hat{\boldsymbol{A}} = P^T \boldsymbol{A} P \tag{11}$$

C DATASETS

PENDULUM The pendulum dataset-generating process consists of four causal factors that will result in the generation of $\mathbf{x}$: the angle of the pendulum, the position of the light source, the position of the shadow, and the length of the shadow. Here, both the angle of the pendulum and the position of the light source make a root node, which causes both shadow position and shadow length. This dataset has images of the resolution 64×64 and has around 20k samples.

MORPHOMNIST MorphoMNIST is a synthetic dataset based on MNIST digits Castro et al. (2019). Here, we consider the data-generating process to be made of two causal factors that will result in the generation of $\mathbf{x}$: thickness and intensity of the digit, where the change in thickness results in changes in intensity. The dataset consists of 60k image samples, with a resolution of 28×28.

COLOR-MORPHOMNIST Color-MorphoMNIST is a coloured version of MorphoMNIST, with three different causal factors in the data-generating process. Colour is an additional factor on top of thickness and intensity, where the change in thickness results in changes in intensity and the resulting change in intensity changes the colour of the digit. The dataset consists of 60k image samples, with a resolution of 28×28.

D NECESSARY AND SUFFICIENT CONDITIONS FOR CRL

Identifiability reflects the models' ability to capture the implicit *symmetries* in the data-generating process Yao et al. (2024a). Usually, these symmetries may or may not correspond to causal representations. When these symmetries are explicitly modelled to map to causal factors, identifiable representations align with causal representations. Identifiability of representations is a necessary condition for CRL but not sufficient; one additional key step involved in CRL is to model the dependencies among the identified representations, which requires additional assumptions on both the data-generating process and the considered model. Assumption 5 states that the underlying data-generating process indeed follows a causal process. While assumption 6, on "Faithfulnes" is a standard assumption in CRL which stems from Pearl (2009) is commonly required for graph discovery, this states that there does not exist any extraneous conditional independence, whose implications are reflected in the assumption.

Assumption 5. (Causal symmetries) We assume that the true data-generating factors have inherent causal symmetries, as defined in 5, which can be exploited during inference.

Assumption 6. (Faithfulness) For any causal factor $\mathbf{c}_i \ni \mathbf{c}_i \notin \mathbf{pa}_j^{\boldsymbol{A}}$, then $\frac{\partial f_j}{\partial \mathbf{u}_i} = 0$.

In most identifiability works based on structural assumptions, *i.e., the methods works with latent distributional assumptions and the properties on mixing function*, the conditional independence is the key assumption (Khemakhem et al., 2020a;b; Kivva et al., 2022; Willetts & Paige, 2021; Falck et al., 2021) and they mostly do not consider graph identification. However, Yang et al. (2021); Komanduri et al. (2024) rely on conditional independence and use ANM modelling assumption to learn causal representations while assuming the dependency structure or DAG used in the data-generating process is known. In this work, we rely on conditional independence assumption on latent variables $\mathbf{u}$; However, we additionally transform these independent variables to encode latent dependencies in latent variables $\mathbf{c}$, we rely on methods similar to Reizinger et al. (2024) to identify the causal graph while diffeomorphic piece-wise linear SCMs, we identify the dependencies up to node wise permutation illustrated in definition 6. Similar to the line of works in causal discovery Reizinger et al. (2024), we use Jacobian of SCMs, allowing us to demonstrate the identifiability of DAG up-to-node permutation. In our case, the Jacobian can be expressed as $\mathbf{J}_{\mathbf{T}_{ij}} = \frac{\partial \mathbf{c}_i}{\partial \mathbf{u}_j} = \frac{\partial g_i}{\partial \bar{f}_i} \frac{\partial \bar{f}_i}{\partial \mathbf{u}_j}$. The inverse of this Jacobian matrix along with faithfulness assumption we get $\boldsymbol{A}-$identifiability as detailed in Theorem 4.

Additionally, as often reflected in evaluations, the mean correlation coefficient (MCC) is usually computed with respect to multiple runs (Kivva et al., 2022; Khemakhem et al., 2020b;a; Kori et al., 2023), this self-consistent behaviour primarily reflects to model's ability to capture unobserved data symmetries, which mostly do not correspond to causal symmetries. While the MCC measured with respect to ground truth reflects the model's ability to uniquely learn the desired properties, in the case

when these desired properties are causal in nature, this measure demonstrates the model's ability to learn causal representations.

E PROOFS

Theorem 1 (**u**-identifiability) Assuming the data generation process follow equation 1, with an invertible demixing function $\phi_{uz}(\mathbf{x})$, such that $(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = (\phi_u(\mathbf{x}), \phi_z(\mathbf{x}))$, and a prior distribution $p(\mathbf{u})$ is modelled as a non-degenerate GMM. Given assumptions 1 and 4, the causal exogenous noise $\mathbf{u}$ is identifiable up to $\sim_e$ equivalence.

Proof. **1. Identifiability of $p(\mathbf{x}|k), k \in \{1, \dots, K\}$ given $p(\mathbf{x})$.**

$p(\mathbf{u})$ and $p(\mathbf{z})$ are non-degenerate Gaussian mixtures with K and M components respectively. The mapping $\psi_x(\mathbf{T}(\mathbf{u}), \mathbf{z})$ is piecewise linear, $p(\mathbf{u}, \mathbf{z})$ is analytic and closed under affine transformations. Define $\hat{\mathbf{T}}(\mathbf{u}, \mathbf{z}) := (\mathbf{T}(\mathbf{u}), \mathbf{z})$. Given $(\psi_x \circ \hat{\mathbf{T}})(\mathbf{u}, \mathbf{z})$ and $\phi_{uz}^{-1}(\hat{\mathbf{u}}, \hat{\mathbf{z}})$ are equally distributed, Lemma C.4 and Corollary C.6 from Kivva et al. (2022) show there exists $\mathbf{x}_0 \in \mathcal{X}$ and $\delta_0 > 0$ such that $\psi_x \circ \hat{\mathbf{T}}$ and ϕ_{uz}^{-1} are invertible on $B(\mathbf{x}_0, \delta_0) \cap \mathcal{X}$. Because the inverse functions are piecewise affine, there exists $\mathbf{x}_1$ and δ_1 with $B(\mathbf{x}_1, \delta_1) \subseteq B(\mathbf{x}_0, \delta_0)$ such that $(\psi_x \circ \hat{\mathbf{T}})^{-1}$ is affine in $B(\mathbf{x}_1, \delta_1) \cap \mathcal{X}$. Following the logic from Kivva et al. (2022), let $L \subseteq \mathcal{X}$ be an affine subspace such that $\dim(L) = n + d_z$ and $B(\mathbf{x}_1, \delta_1) \cap \mathcal{X} = B(\mathbf{x}_1, \delta_1) \cap L$. Then, there exists an invertible affine map $h : \mathbb{R}^{n+d_z} \to L$ such that $h^{-1} = (\psi_x \circ \hat{\mathbf{T}})^{-1}$ on $B(\mathbf{x}_1, \delta_1) \cap L$.

For all $\mathbf{x} \in B(\mathbf{x}_1, \delta_1) \cap \mathcal{X}$, we can view the observed distribution as a pushforward measure of $p(\mathbf{u}, \mathbf{z})$ by h. Considering $p(\mathbf{u})$ is a non-degenerate Gaussian mixture of K components, we denote the family of mixtures under the transformation h.

$$\mathcal{M}_x := \left\{ \sum_k^K c_k p_k(\mathbf{x}), \quad c_k > 0, \quad \sum c_k = 1, \quad p_k \in \mathcal{P}_x, \quad K = \mathbb{R}^+ \right\}, \tag{12}$$

$$\text{where } \mathcal{P}_x := \left\{ (h_{\#} p_a)(\mathbf{x}), \quad h \in \mathcal{H}, \quad p_a \in \mathcal{P} \right\}. \tag{13}$$

$\mathcal{H}$ denotes a family of affine transformations, and $\mathcal{P}$ denotes a family of distributions such that each $p_a(\mathbf{u}, \mathbf{z}) = p_{a_u}(\mathbf{u}) p(\mathbf{z})$, with Gaussian p_{a_u} and fixed p. Yakowitz & Spragins (1968) shows identifiability for mixture models defined on multidimensional CDFs, where linear independence is a sufficient and necessary condition. This result also extends to PDFs, with linear independence on $\mathcal{P}_x$. Given $\mathcal{H}$ is a family of affine transformations, $\mathcal{P}_x$ preserves the linear independence properties of Gaussian families.

Identifiability up to permutations is defined in Yakowitz & Spragins (1968) as follows. Given $p(\mathbf{x}), p'(\mathbf{x})$ such that:

$$p(\mathbf{x}) = \sum_k^K p_k(\mathbf{x}) c_k = \sum_{k'}^{K'} p'_k(\mathbf{x}) c'_k = p'(\mathbf{x});$$

we have $K' = K$ and there exists a permutation $\tau \in S_K$ such that and $p_k(\mathbf{x}) = p'_{\tau(k)}(\mathbf{x})$ and $c_k = c'_{\tau(k)}$, for any $k \in \{1, \dots, K\}$.

2. Disentangling $\hat{\mathbf{u}}$ from $\hat{\mathbf{z}}$

Given $\mathbf{x} \in \mathcal{X}$, we write the correspondence between $\mathbf{x}$ with respect to the true exogenous noise $\mathbf{u}, \mathbf{z}$ and some estimated exogenous noise $\hat{\mathbf{u}}, \hat{\mathbf{z}}$:

$$(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = \big(\phi_u(\mathbf{x}), \phi_z(\mathbf{x})\big) = \phi_{uz}(\mathbf{x}), \qquad \mathbf{x} = \psi_x(\mathbf{T}(\mathbf{u}), \mathbf{z}); \tag{14}$$

where $(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = \phi_{uz}(\psi_x(\mathbf{T}(\mathbf{u}), \mathbf{z}))$. We observe $\hat{\mathbf{u}}$ is both dependent on $\mathbf{u}$ and $\mathbf{z}$, and ϕ_u is not invertible since $\dim(\hat{\mathbf{u}}) < \dim(\mathbf{x})$. Assuming no information on the label assignment $k \in \{1, \dots, K\}$, the log-density $\log p(\hat{\mathbf{u}}, \hat{\mathbf{z}})$ is expressed as:

$$\log p(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = \log p(\mathbf{c}, \mathbf{z}) + \log |\det \mathbf{J}_{(\psi_x^{-1} \circ \phi_{uz}^{-1})}| \tag{15}$$

$$= \log p(\mathbf{u}) + \log p(\mathbf{z}) + \log |\det \mathbf{J}_{\mathbf{T}^{-1}}| + \log |\det \mathbf{J}_{(\psi_x^{-1} \circ \phi_{uz}^{-1})}|. \tag{16}$$

Note that given the non-invertibility of ϕ_u, we cannot write $\frac{\partial \mathbf{u}}{\partial \hat{\mathbf{u}}}$ directly. However, given the identifiability of the emission distributions $p_k(\mathbf{x}), k \in \{1, \dots, K\}$ up to a permutation τ, we can write the distribution mapping from $\mathbf{u}$ to $\hat{\mathbf{u}}$ for any k. Wlog, assume $\tau(k) = k$. Considering $\mathbf{u}, \mathbf{z}$ and $\hat{\mathbf{u}}, \hat{\mathbf{z}}$ are independent, we denote the difference between distributions given two conditionals $k_1, k_2 \in [K]$. From the above equation, we denote de direct dependency of $\mathbf{u}$ and $\hat{\mathbf{u}}$:

$$\log p(\hat{\mathbf{u}}, \hat{\mathbf{z}}|k_1) - \log p(\hat{\mathbf{u}}, \hat{\mathbf{z}}|k_2) = \log p(\mathbf{u} \mid k_1) - \log p(\mathbf{u} \mid k_2), \tag{17}$$

$$\log p(\hat{\mathbf{u}} \mid k_1) - \log p(\hat{\mathbf{u}} \mid k_2) = \log p(\mathbf{u} \mid k_1) - \log p(\mathbf{u} \mid k_2). \tag{18}$$

Where $p(\hat{\mathbf{u}}, \hat{\mathbf{z}}) = p(\hat{\mathbf{u}})p(\hat{\mathbf{z}})$ from our demixing assumptions on ϕ_{uz}. We take the derivative with respect to $\hat{\mathbf{u}}$

$$\frac{\partial}{\partial \hat{\mathbf{u}}} \Big(\log p(\hat{\mathbf{u}} \mid k_1) - \log p(\hat{\mathbf{u}} \mid k_2) \Big) = \frac{\partial}{\partial \mathbf{u}} \Big(\log p(\mathbf{u} \mid k_1) - \log p(\mathbf{u} \mid k_2) \Big) \mathbf{H}, \tag{19}$$

which gives us an identity relating the Jacobian $\frac{\partial \mathbf{u}}{\partial \hat{\mathbf{u}}} \in \mathbb{R}^{n \times n}$, denoted as $\mathbf{H}$. We can create a system of n equations using $K = n+1$ components and $k = 1$ as a pivot. We recall the following definitions for notational simplicity:

$$\eta_{ik} := \log p(u_i|k),\ \mathbf{v}_k^1 := \left(\frac{\partial \eta_{0k}}{\partial u_0}, \dots, \frac{\partial \eta_{nk}}{\partial u_n} \right),\ \hat{\eta}_{ik} := \log p(u_i|k),\ \hat{\mathbf{v}}_k^1 := \left(\frac{\partial \hat{\eta}_{0k}}{\partial \hat{u}_0}, \dots, \frac{\partial \hat{\eta}_{nk}}{\partial \hat{u}_n} \right), \tag{20}$$

introduced in Assumption 4. The system of equations in vector form results as follows

$$\begin{pmatrix} \hat{\mathbf{v}}_2^1 - \hat{\mathbf{v}}_1^1 \\ \vdots \\ \hat{\mathbf{v}}_{n+1}^1 - \hat{\mathbf{v}}_1^1 \end{pmatrix} = \begin{pmatrix} \mathbf{v}_2^1 - \mathbf{v}_1^1 \\ \vdots \\ \mathbf{v}_{n+1}^1 - \mathbf{v}_1^1 \end{pmatrix} \mathbf{H}. \tag{21}$$

where a unique solution for $\mathbf{H}$ exists if the RHS matrix is full rank, i.e. the vectors $\mathbf{v}_k^1 - \mathbf{v}_1^1, k \in \{2, \dots, K\}$ are linearly independent, which is true given sufficient variability (Assumption 4).

3. Identifiability Condition

Note we assume the exogenous variables $\hat{\mathbf{u}}$ are mutually independent given $k \in \{1, \dots K\}$. Following Yao et al. (2022), for any $p \neq q, p, q \in \{1, \dots, n\}$, we have

$$\frac{\partial^2 \log p(\hat{\mathbf{u}}|k)}{\partial \hat{u}_p \partial \hat{u}_q} = 0, \tag{22}$$

with the following closed-form expression:

$$\frac{\partial \log p(\hat{\mathbf{u}} \mid k)}{\partial \hat{u}_p} = \sum_{i=1}^{n} \frac{\partial \eta_{ik}}{\partial u_i} \mathbf{H}_{ip} + \frac{\partial \left(\log p(\mathbf{z}) - \log p(\hat{\mathbf{z}}) + \log |\det \mathbf{J}_{\mathbf{T}^{-1}}| |\det \mathbf{J}_{(\psi_x^{-1} \circ \phi_{uz}^{-1})}| \right)}{\partial \hat{u}_p}, \tag{23}$$

$$\begin{aligned} \frac{\partial^2 \log p(\hat{\mathbf{u}} \mid k)}{\partial \hat{u}_p \partial \hat{u}_q} = \sum_{i=1}^{n} \left(\frac{\partial^2 \eta_{ik}}{\partial u_i^2} \mathbf{H}_{ip} \mathbf{H}_{iq} + \frac{\partial \eta_{ik}}{\partial u_i} \frac{\partial \mathbf{H}_{ip}}{\partial \hat{u}_q} \right) \\ + \frac{\partial^2 \left(\log p(\mathbf{z}) - \log p(\hat{\mathbf{z}}) + \log |\det \mathbf{J}_{\mathbf{T}^{-1}}| |\det \mathbf{J}_{(\psi_x^{-1} \circ \phi_{uz}^{-1})}| \right)}{\partial \hat{u}_p \partial \hat{u}_q} = 0, \end{aligned} \tag{24}$$

where $\mathbf{H}_{ij} = \frac{\partial u_i}{\partial \hat{u}_j}, i, j \in [n]$. Again, we use following definitions for notational simplicity

$$\mathbf{v}_k^2 := \left(\frac{\partial^2 \eta_{0k}}{\partial u_0^2}, \dots, \frac{\partial^2 \eta_{nk}}{\partial u_n^2} \right), \mathbf{v}_k^1 := \left(\frac{\partial \eta_{0k}}{\partial u_0}, \dots, \frac{\partial \eta_{nk}}{\partial u_n} \right), \mathbf{v}_k = \mathbf{v}_k^2 \oplus \mathbf{v}_k^1 \tag{25}$$

$$\mathbf{h}^2 := \left(\mathbf{H}_{0p} \mathbf{H}_{0q}, \dots, \mathbf{H}_{np} \mathbf{H}_{nq} \right), \mathbf{h}^1 = \left(\frac{\partial \mathbf{H}_{0p}}{\partial \hat{u}_q}, \dots, \frac{\partial \mathbf{H}_{np}}{\partial \hat{u}_q} \right), \mathbf{h} = \mathbf{h}^2 \oplus \mathbf{h}^1. \tag{26}$$

Similarly as above, for any $k \neq 1$, we can pivot around $k = 1$, and eliminate the dependency over $\mathbf{z}$ and $\hat{\mathbf{z}}$.

$$\frac{\partial^2 \log p(\hat{\mathbf{u}} \mid k)}{\partial \hat{u}_p \partial \hat{u}_q} - \frac{\partial^2 \log p(\hat{\mathbf{u}} \mid 1)}{\partial \hat{u}_p \partial \hat{u}_q} = (\mathbf{v}_k - \mathbf{v}_1)\mathbf{h}^T = 0. \tag{27}$$

With number of components K, we get system of equations as:

$$\begin{pmatrix} \mathbf{v}_2 - \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_K - \mathbf{v}_1 \end{pmatrix} \mathbf{h}^T = \boldsymbol{V}\mathbf{h}^T = 0. \tag{28}$$

4. Result Analysis

From Theorem 3.2 in (Kivva et al., 2022), and considering all the transformations from $(\mathbf{u}, \mathbf{z})$ to $\mathbf{x}$ are injective and piece-wise linear, we know that the latents $(\mathbf{u}, \mathbf{z})$ are identifiable up to affine transformations. Therefore, the Jacobian $\mathbf{H}$ is a linear transformation, and $\frac{\partial \mathbf{H}}{\partial \hat{\mathbf{u}}} = 0$ which implies $\mathbf{h}^1 = 0$. Based the on the sufficient variability assumption (Assumption 4) we know the system admits the only valid solution $\mathbf{h} = 0$ thanks to having $K = n + 1$ linear independent vectors $\mathbf{v}_k^2 - \mathbf{v}_1^2, k \in \{2, \ldots, K\}$. Then, Eq. (28) holds true only if $\mathbf{H}_{ip}\mathbf{H}_{iq} = 0$ for all i and $p \neq q$. In other words, each row in $\mathbf{H}$ only admits one non-zero entry, and therefore $\hat{\mathbf{u}}$ is a component-wise scaling of a permutation of $\mathbf{u}$. This implies $\mathbf{u}$ is element-wise identifiable.

□

Theorem 3 ($\mathbf{z}-$identifiability Kivva et al. (2022))**.** *Given* $\mathbf{z} \sim \sum_{i=1}^{J} \boldsymbol{\pi}_i \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ *and* $\mathbf{z}' \sim \sum_{j=1}^{J'} \boldsymbol{\pi}'_j \mathcal{N}(\mathbf{z}'; \boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)$ *and* $\psi_x(\mathbf{z}, \mathbf{c})$ *and* $\tilde{\psi}_x(\mathbf{z}', \mathbf{c})$ *are equally distributed. Then given* $\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j, \forall i \in [J], j \in [J']$*, there exists two indices* $i_1, i_2 \in [J] \ni ((\boldsymbol{\Sigma}_{i_1})_{tt}/(\boldsymbol{\Sigma}_{i_2})_{tt})$ *are distinct, resulting in an recoverable invertible linear map* $\boldsymbol{H} : \mathbb{R}^{kd_z} \to \mathbb{R}^{kd_z}$ *such that* $\boldsymbol{H} = \boldsymbol{QD}$ *mapping* $\mathbf{z}$ *to* $\mathbf{z}'$*, where* $\boldsymbol{Q}$ *and* $\boldsymbol{D}$ *are permutation and diagonal matrices with positive entries.*

Theorem 4 ($\boldsymbol{A}-$identifiability)**.** *Given* $p(\mathbf{u})$ *is element-wise identifiable from theorem 1, and diffeomorphic transformation from* $\mathbf{u}$ *to* $\mathbf{c}$*. Then with an assumption 6 we can recover DAG up to* $\sim_{node}$ *equivalence, definition 6.*

Proof. We start with Proposition 1 from Reizinger et al. (2023) which states that given $\mathbf{c} = \mathbf{T}(\mathbf{u})$, the inverse jacobian $\mathbf{J}_{\mathbf{T}^{-1}}$ is structurally equivalent to $(I_n - \boldsymbol{A})$ when:

i) The structural equation models (SEM) are given by $c_i = g_i(f_i(\mathbf{pa}_i), u_i) \forall i \in [n]$, such that element-wise identifiability is also enforced on $\mathbf{c}$. g_i denote the components of the vector transformation $\mathbf{T}$;

ii) u_i are independent;

iii) There are no hidden confounders and the jacobians $\mathbf{J}_{\mathbf{T}^{-1}}, \mathbf{J}_{\mathbf{T}}$ are faithful to $\boldsymbol{A}$;

iv) Each g_i is bijective; and

iv) Each c_i depends on u_i.

Our framework is consistent with the above assumptions. The SEM is given by Eq. (3), where the exogenous noise variables $\mathbf{u}$ are mutually independent. Considering element-wise identifiability of $\mathbf{u}$ and $\mathbf{c}$, our setup contains no hidden confounders. Furthermore, Assumption 6 is equivalent to structural faithfulness of $\mathbf{J}_{\mathbf{T}^{-1}}$ and $\mathbf{J}_{\mathbf{T}}$.

Consider element-wise identifiability equivalences of both $\mathbf{u}$ and $\mathbf{c}$

$$\hat{\mathbf{u}} = PD_u\mathbf{u} + \mathbf{b}_u, \quad \hat{\mathbf{c}} = PD_c\mathbf{c} + \mathbf{b}_c. \tag{29}$$

where the permutation matrix P is constant as the variables are aligned, but the scaling and bias could differ. We write the equivalences in terms of $\hat{\mathbf{c}}$ and $\mathbf{c}$ respectively through $\hat{\mathbf{T}}, \mathbf{T}$:

$$\hat{\mathbf{T}}^{-1}(\hat{\mathbf{c}}) = PD_u\mathbf{T}^{-1}(\mathbf{c}) + \mathbf{b}_u, \tag{30}$$

$$\hat{\mathbf{T}}^{-1}(\hat{\mathbf{c}}) = PD_u\mathbf{T}^{-1}(D_c^{-1}P^T(\hat{\mathbf{c}} - \mathbf{b}_c)) + \mathbf{b}_u. \tag{31}$$

Given P is a permutation matrix, we have $P^{-1} = P^T$. We now take the derivative with respect to $\hat{\mathbf{c}}$:

$$\frac{\partial \hat{\mathbf{T}}^{-1}(\hat{\mathbf{c}})}{\partial \hat{\mathbf{c}}} = (PD_u)^T \frac{\partial \mathbf{T}^{-1}(\mathbf{c})}{\partial \mathbf{c}} (D_c^{-1} P^T)^T, \tag{32}$$

$$D_u^{-1} \mathbf{J}_{\hat{\mathbf{T}}^{-1}} D_c = P^T \mathbf{J}_{\mathbf{T}^{-1}} P. \tag{33}$$

As we observe, the jacobian of the inverse transformations $\mathbf{T}^{-1}$ and $\hat{\mathbf{T}}^{-1}$ are equivalent up to a permutation of the rows and columns, and a diagonal scaling. Given that D_u and D_c have non-zero entries, from Proposition 1 in Reizinger et al. (2023) we have:

$$(I_n - \hat{\boldsymbol{A}}) = P^T (I_n - \boldsymbol{A}) P, \tag{34}$$

$$\hat{\boldsymbol{A}} = P^T \boldsymbol{A} P; \tag{35}$$

where $P^T P = I_n$, which implies that A is equivalent to $\hat{A}$ up to a permutation of its rows and columns. I.e. given a permutation π consistent with P, we have $A_{i,j} = A_{\pi(i)\pi(j)}, \forall i, j \in [n]$. □