TECHNICAL REPORT
R-110
First version: May, 2024
Last version: Aug, 2024

Disentangled Representation Learning in Non-Markovian Causal Systems

Adam Li* and **Yushu Pan*** and **Elias Bareinboim**
Causal Artificial Intelligence Lab
Columbia University
{adam.li, yushupan, eb}@cs.columbia.edu

Abstract

Considering various data modalities, such as images, videos, and text, humans perform causal reasoning using high-level causal variables, as opposed to operating at the low, pixel level from which the data comes. In practice, most causal reasoning methods assume that the data is described as granular as the underlying causal generative factors, which is often not the case in various AI applications. In this paper, we acknowledge this issue and study the problem of causal disentangled representation learning from a combination of data gathered from various heterogeneous domains and assumptions in the form of a latent causal graph. To the best of our knowledge, the proposed work is the first to consider i) non-Markovian causal settings, where there may be unobserved confounding, ii) arbitrary distributions that arise from multiple domains, and iii) a relaxed version of user-chosen disentanglement. Specifically, we introduce graphical criteria that allow for disentanglement under various conditions. Building on these results, we develop an algorithm that returns a causal disentanglement map, highlighting which latent variables can be disentangled given the combination of data and assumptions. The theory is corroborated by experiments.

1 Introduction

Causality is fundamental throughout various aspects of human cognition, including understanding, planning, decision-making. The ability to perform causal reasoning is considered one of the hallmarks of human intelligence [1–3]. In the context of AI, the capability of reasoning with cause-and-effect relationships plays a critical role in various challenging tasks, including explainability, fairness, decision-making, robustness, and generalizability. One key assumption of most methods currently available in the literature is that the set of (endogenous) variables is at the right level of granularity. However, this is not the case in many AI applications, where various modalities, such as images, and text, come into play [4]. For example, images may capture a natural scene within a park, where the causal variables are the objects within the scene. The pixels themselves are not the causal variables, and thus AI must disentangle the latent causal variables given the pixels in order to represent the true underlying causal relationships in the image. Similarly, given written text describing the natural scene within the park, the characters in the text are not the causal variables, but rather a low-level mixture of the underlying causal variables.

In machine learning, the representation learning literature is concerned with finding useful representations from data [5]. One important line of work traces back to linear ICA (independent component analysis) [6], where one attempts to disentangle latent variables assuming a linear mixing function. The literature has also considered settings where the mixing function is nonlinear [7, 8]. It has been

*These authors contributed equally to this work.

Work	Input					Output
	Assumptions		Data			Identifiability Goal
	Non-Markovian	Non-parametric	Interventions	Multiple Domains	Distr. Reqs.	
[6, 11–14]	✗	✗	✓	✗	1 per node	Scaling, Mixture or Affine Transformation
[7, 8, 10, 15, 16]	✗	✗	✗/✓	✗/✓	$2\|\mathbf{V}\|+1$	Scaling
[17, 18]	✗	✓	✓	✗	1 per node	Scaling
[19, 20]	✗	✓	✗/✓	✗/✓	1 per node	Scaling
[21]	✗	✓	✓	✗	1 per node	Scaling or Ancestral Mixture
[22]	✗	✓	✗	✓	$2\|\mathbf{V}\|+\|M_G\|+1$	Scaling or Mixture
[23]	✗	✗	✗/✓	✗/✓	TBD	Scaling & Affine Transformation
[24]	✗	✓	✗/✓	✗/✓	TBD	Functional Dependency Map[1]
This work	✓	✓	✓	✓	General	Causal Disentanglement Map

Table 1: A non-exhaustive list of identifiability results given knowledge of the latent graph.

understood that nonlinear-ICA is, in general, not identifiable (ID) given only observational data [9]. Different routes have been taken to circumvent such impossibilities. For instance, one might assume parametric families (e.g., exponential), and auxiliary variables as input, which can be thought of as non-stationary times-series implying certain new invariances that can be exploited [7, 8, 10].

Interestingly, the machinery developed in this context can be applied to causal settings with multimodal data, where there is a mismatch between the causal variables and the granularity at which they are represented in the data. The key observation that links these two worlds is that an underlying causal system generates the data at such granularity (images, videos, and texts). Acknowledging this connection leads to various possibilities regarding disentangling (learning) the causal variables from data, similar to the initial ICA-like literature [6]. First, the assumption that the features underlying a signal are independent needs to be relaxed since it is arguably too stringent, *a priori* ruling out almost any interesting causal system. So, we should consider different assumptions regarding the structure of the underlying generative model. One initial relaxation is that this model is Markovian, where the features need not be independent, and causal relationships are allowed across features. In the context of computer vision, for example, one might assume a specific structure on the latent variables where the style and content of the images are separated and augmented data is leveraged to disentangle these two components [18]. Generalizing this idea to more relaxed causal settings, one can show ID up to certain indeterminancies given observational across multiple domains, or interventional data [21, 22]. Another approach allows for certain parametric mixing functions, which could lead to new ID results [11, 14]. These results have been applied and advanced across various downstream tasks [25–31].

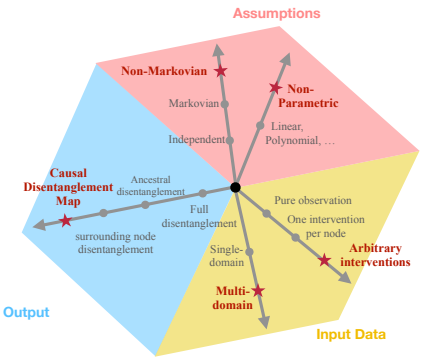


Figure 1: Dimensions of identifiability in causal disentanglement representation learning.

To represent such varied elements, we will study three axes within the different types of input and expected outputs of the causal disentanglement representation learning task, which we summarize in Table 1 and Fig. 1. What we refer to as the input can be partitioned into qualitative and quantitative components. In terms of its qualitative aspect component, we consider different **assumptions** about the underlying generative processes, including non-parametric, in contrast to, for example, linear or Gaussian. As alluded to earlier, we will also account for systems with richer causal topologies than ICA (independent features) while generalizing the Markovian setting. In particular, we will not rule out *a priori* the existence of unobserved confounding among features, which is a challenge pervasive throughout the causal inference literature. Regarding the quantitative component of the input, we will consider **data** gathered from arbitrary combinations of interventions and domains. We follow the discussion of the recent literature, which acknowledges key differences of this distinction [32–37]. Other prior literature often assumes that data comes from different interventions in the same domain or from various (observational) distributions from different domains. In fact, it is feasible that data spawns various interventions and domains in a less well-structured manner. In terms of the expected **output**, we will consider both full disentanglement as well as a more relaxed type of disentanglement, known as the causal disentanglement map.

[1]We recently were made aware of the work in [24], where their definition of a functional dependency map is what we define as a causal disentanglement map. However, the disentanglement map that they can achieve is different from ours as discussed in Section E.

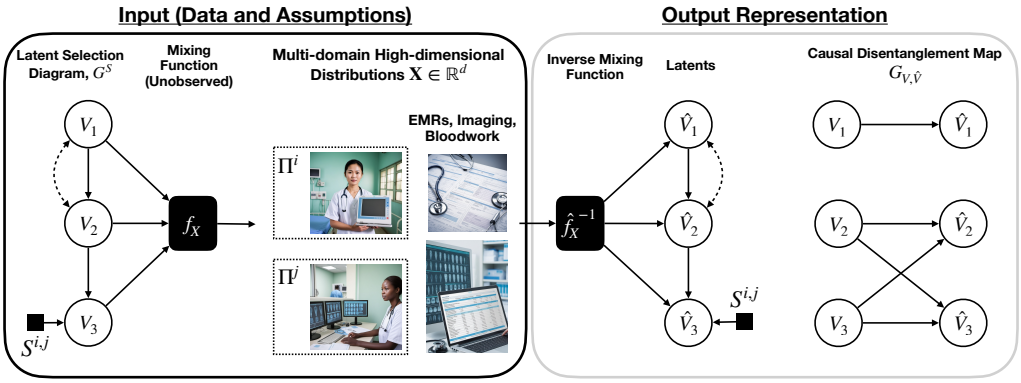


Figure 2: Data generating model and the goal of learning disentangled causal latent representations.

For concreteness, consider a hypothetical latent graph depicted in Fig. 2 in the context of epilepsy research [38–46]. In terms of **assumptions**, hospitals in different countries Π^i and Π^j will differ in the type of seizures (V_1) patients get (represented by the S-node $S^{i,j} \to V_3$). Now suppose sleep (V_1) affects the efficacy of the drug treatment (V_2), and the drug helps epilepsy patients control their seizures (V_3). The quality of sleep and the type of drug treatment are confounded by socioeconomic factors ($V_1 \dashleftarrow\dashrightarrow V_2$). Clinicians are then given electroencephalogram (EEG) **data** from each hospital where they know different drug treatments were administered. The EEG $\mathbf{X}$ is a nonlinear (nonparametric) transformation of latent $\mathbf{V} = \{V_1, V_2, V_3\}$ via f_X. Their goal is to generate realistic EEG data to understand how different drugs affect EEG patterns. This requires a general **output** representation that disentangles sleep from drug as it is understood that sleep affects EEG [47]; It is not required to disentangle the drug treatment and outcomes since it is known the drug treatments will impact the outcomes. One could leverage state-of-the-art generative modeling techniques and train a self-supervised learning model to learn a representation of the EEG that they then perturb to generate new instances of EEG [48–50]. However, there are no guarantees that the representation, or interventions in the latent space will generate realistic EEG. In this case, drug and sleep might remain entangled in the learned representation, which is potentially harmful since it may lead to unrealistic EEG data that contains visual differences due to sleep rather than the drug. More formally, given an input set of distributions and knowledge of the latent causal structure, the goal is to learn the unmixing function $\widehat{f}_X^{-1}$ and a representation $\widehat{\mathbf{V}} = \{\widehat{V_1}, \widehat{V_2}, \widehat{V_3}\}$, where V_2 is disentangled from V_1 [2].

Consider a marketing company creating faces for a female product. The relevant latent factors are Gender $\leftrightarrow$ Age $\rightarrow$ Hair Color (see Appendix D.1 for details). If Gender and Age are entangled, changing Age might also alter Gender, which is undesirable. The company needs a model where Age is disentangled from Gender, while correlation with Hair Color is allowed. Our paper addresses the problem of determining whether a given set of input data and assumptions in the form of a LSD is sufficient to learn such a disentangled representation.

In this work, we introduce graphical and algorithmic machinery to determine whether (and how) causal representations can be disentangled from heterogeneous data and assumptions about the underlying causal system, which might help improve various downstream tasks. Our contributions are as follows:

1. **Graphical criteria for determining the disentangleability of causal factors.** We formalize a general version of the causal representation learning problem and develop methods to determine if a pair of (user-chosen) variables are disentangled in a non-Markovian setting with arbitrary distributions from multiple heterogeneous domains (Props. 3,4, and 5)[3].
2. **An algorithm to learn the causal disentanglement map.** Leveraging these new conditions, we develop an algorithmic called **CRID**, which systematically determines whether two sets of latent variables are disentangleable given their selection diagram and a collection of intervention targets (Thm. 1). The theoretical findings are corroborated with simulations.

[2]We separate the tasks of disentanglement and structural learning, and consider the latent causal graph as input of our task. Still, there are works in the literature that study both tasks simultaneously [13, 22, 51–56].

[3]All proofs are provided in Appendix C.

Preliminaries. We introduce basic definitions used throughout the paper. Uppercase letters (X) represent random variables, lowercase letters (x) signify assignments, and bold letters ($\mathbf{X}$) indicate sets. For a set $\mathbf{X}$, $|\mathbf{X}|$ denotes its dimension. Denote $P(\mathbf{X})$ as a probability distribution over $\mathbf{X}$ and $p(\mathbf{x})$ as its density function. The basic semantic framework of our analysis rests on structural causal models (SCMs) [1, Ch. 7]. An SCM is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where (1) $\mathbf{U}$ is a set of background variables, also called exogenous variables, that are determined by factors outside the model; (2) $\mathbf{V} = \{V_1, V_2, \ldots, V_d\}$ is the set of endogenous variables that are determined by other variables in the model; (3) $\mathcal{F}$ is the set of functions $\{f_{V_1}, f_{V_2} \ldots, f_{V_d}\}$ mapping $\mathbf{U}_{V_j} \cup \mathbf{Pa}_{V_j}$ to V_j, where $\mathbf{U}_{V_j} \subseteq \mathbf{U}$ and $\mathbf{Pa}_{V_j} \subseteq \mathbf{V} \backslash V_j$; (4) $P(\mathbf{U})$ is a probability function over the domain of $\mathbf{U}$.

Each SCM induces a causal diagram G, which is a directed acyclic graph where every V_j is a vertex. There is a directed arrow from V_j to V_k if $V_j \in \mathbf{Pa}_{V_k}$. There is a bidirected arrow between V_j and V_k if $\mathbf{U}_{V_j}$ and $\mathbf{U}_{V_k}$ are not independent [3]. Variables $\mathbf{V}$ can be partitioned into subsets called *c-components* [57]. The c-component of X, denoted as $\mathbf{C}(X)$, is a set of variables connected to X by bidirected paths. The c-component of a set $\mathbf{X}$, denoted as $\mathbf{C}(\mathbf{X})$, is defined as the union of the c-component of every $X \in \mathbf{X}$. We will use $\mathbf{Pa}(X)$ or $\mathbf{Pa}_X$ to denote parents of X in G. Let $\overline{\mathbf{Pa}}(X) = \mathbf{Pa}(X) \cup X$, which includes X itself. Let $\mathbf{Anc}(X)$, or $\mathbf{Anc}_X$ denote ancestors of X in G, and $\overline{\mathbf{Anc}}(X) = \mathbf{Anc}(X) \cup \{X\}$, which includes X itself. A subgraph over $\mathbf{X} \subseteq \mathbf{V}$ in G is denoted as $G(\mathbf{X})$ and $G_{\overline{\mathbf{X}}}$ denotes the subgraph by removing arrows coming into nodes in $\mathbf{X}$. A path p is said to be active given (or conditioned on) Z if every vertex on p is active relative to Z. Otherwise, p is said to be inactive. Given a graph G, X and Y are d-separated by Z if every path between X and Y is inactive given Z. We denote this d-separation by $(X \perp\!\!\!\perp Y | Z)_G$.

A *soft intervention* on a variable X, denoted σ_X, replaces f_X with a new function f'_X of $\mathbf{Pa}' \subset \mathbf{V}$ and variables $\mathbf{U}'_X$ [58, 59]. For interventions on a set of variables $\mathbf{X} \subseteq \mathbf{V}$, let $\sigma_{\mathbf{X}} = \{\sigma_{\mathbf{X}}\}_{X \in \mathbf{X}}$, that is, the result of applying one intervention after the other. Given an SCM $\mathcal{M}$, let $\mathcal{M}_{\sigma_{\mathbf{X}}}$ be a submodel of $\mathcal{M}$ induced by intervention $\mathcal{M}_{\sigma_{\mathbf{X}}}$. The observational distribution can be thought of as the result of a special class of soft interventions, called *idle* intervention. Specifically, an idle intervention leaves the function as it is, which means $\sigma_{\mathbf{X}} = \{\}$. Another special class of soft interventions, called *hard* (or perfect) interventions [21, 51] and denoted as $do(\mathbf{X})$, such that $\mathbf{Pa}(\mathbf{X}) = \emptyset$ and $\mathbf{U}'_X \cap \mathbf{U} = \emptyset$. This implies that the modified diagram induced by $\mathcal{M}_{\sigma_{\mathbf{X}}}$ is $G_{\overline{\mathbf{X}}}$. We assume soft interventions that are not hard do not change the structure of the graph [4]. Namely, the diagram induced by $\mathcal{M}_{\sigma_{\mathbf{X}}}$ is the same with G.

2 Modeling Disentangled Representation Learning (General Case)

In this section, we formalize the disentangled representation learning task in causal language. We leverage Augmented SCMs to model the generative process over *latent* causal variables $\mathbf{V}$.

Definition 2.1 (Augmented Structure Causal Model)**.** An Augmented Structure Causal Model (**ASCM**) over a generative level SCM $\mathcal{M}_0 = \langle \{\mathbf{U}_0, \mathbf{V}_0, \mathcal{F}_0, P^0(\mathbf{U}_0)\} \rangle$ is a tuple $\mathcal{M} = \langle \mathbf{U}, \{\mathbf{V}, \mathbf{X}\}, \mathcal{F}, P(\mathbf{U}) \rangle$ such that
(1) exogenous variables $\mathbf{U} = \mathbf{U}_0$;
(2) $\mathbf{V} = \mathbf{V}_0 = \{V_1, \ldots, V_d\}$ are d latent endogenous variables; $\mathbf{X}$ is an m dimensional mixture variable;
(3) $\mathcal{F} = \{\mathcal{F}_0, f_{\mathbf{X}}\}$, where $f_{\mathbf{X}} : \mathbb{R}^d \to \mathbb{R}^m$ is a diffeomorphic [5] function that maps from (the respective domains of) $\mathbf{V}$ to $\mathbf{X}$. $\exists \quad h = f_{\mathbf{X}}^{-1}$ such that $\mathbf{V} = h(\mathbf{X})$; and
(4) $P(\mathbf{U}_0) = P^0(\mathbf{U}_0)$. □

In words, an ASCM $\mathcal{M}$ describes a two-stage generative process involving latent generative factors $\mathbf{V}$ and high-dimensional mixture $\mathbf{X}$ (e.g., images, text). First, the latent generative factors $\mathbf{V} \in \mathbb{R}^d$ are generated by an underlying SCM. The causal diagram induced by $\mathcal{M}_0$ over $\mathbf{V}$ will be called *a latent causal graph* (LCG), denoted as G. Next, a nonparametric diffeomorphism $f_{\mathbf{X}}$ mixes $\mathbf{V}$ to get the high-dimensional mixture $\mathbf{X} \in \mathbb{R}^m$. An important aspect of $f_{\mathbf{X}}$ is that it is invertible regarding $\mathbf{V}$, which implies that the generative factors $\mathbf{V}$ are recognized in a given $\mathbf{X}$ [6]. The function

[4] In general, soft interventions can arbitrarily change the graph by adding or removing edges. We do not consider this setting, and refer the readers to [1, 58, 59] for a general discussion on soft interventions.

[5] A diffeomorphism is a bijective function $f_{\mathbf{X}}$ such that both $f_{\mathbf{X}}$ and $f_{\mathbf{X}}^{-1}$ are continuously differentiable [27]

[6] Further discussion on the invertibility and non-parametric assumption is provided in Appendix A.2.

of the underlying mechanisms $\mathcal{F}_0$, or the mixing function $f_{\mathbf{X}}$ are not restricted, and we focus on the non-parametric setting. This assumption is commonly in the non-linear ICA and representation learning literature [9, 10, 17, 60].

The initial disentangled representation learning setting can be traced back to linear/nonlinear ICA [7–9], which assumes that no directed edges in G ($\mathbf{V}$ are independent of each other) and Markovianity of G (no bidirected edges). More recently, allowing latent variables to have edges in the LCG was studied, under the Markovian assumption [11, 13, 14, 21, 22, 51, 55]. We discuss a relaxation of this assumption and allow for unobserved confounding to exist between $\mathbf{V}$, which is called non-Markovianity[7].

Domains. We address the general setting of distributions that arise from multiple domains. Following [32–35, 63, 64], we define the so-called *latent selection diagram* that will represent a collection of ASCMs and formally model the multi-domain setting. Selection diagrams enable us to compactly represent causal structure and cross-domain invariances [8].

Definition 2.2 (Latent Selection Diagrams)**.** Let $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_N \rangle$ be a collection of ASCMs relative to N domains $\boldsymbol{\Pi} = \langle \Pi_1, \Pi_2, ..., \Pi_N \rangle$, sharing mixing function $f_{\mathbf{X}}$ and LCG, G. $\boldsymbol{\mathcal{M}}$ defines a latent selection diagram (**LSD**) G^S, constructed as follows: (1) every edge in G is also an edge in G^S; (2) G^S contains an extra node $S^{i,j}$ and corresponding edge $S^{i,j} \to V_k$ whenever there exists a discrepancy $f^i_{V_k} \neq f^j_{V_k}$, or $P^i(U_k) \neq P^j(U_k)$ between $\mathcal{M}_i$ and $\mathcal{M}_j$. □

S-nodes indicate possible differences over $\mathbf{V}$ due to changes in the underlying mechanism or exogenous distributions across domains. For example, consider the LSD in Fig. 2. The S-node $S^{i,j}$ implies that V_1 possibly changes from domain Π^i to Π^j, while the mechanisms of V_2 and V_3 are assumed to be invariant. Note no S-node points to $\mathbf{X}$ since $f_{\mathbf{X}}$ is shared across $\boldsymbol{\mathcal{M}}$.

Interventions. A set of interventions $\boldsymbol{\Sigma} = \{\sigma^{(k)}\}_{k=1}^K$ are applied across domains $\boldsymbol{\Pi}$, where k is an index from 1 to K. The corresponding domains that $\boldsymbol{\Sigma}$ are intervened on is denoted as $\boldsymbol{\Pi}^{\boldsymbol{\Sigma}} = \{\Pi^{(k)}\}_{k=1}^K$ (the domains associated with each $\sigma^{(k)} \in \boldsymbol{\Sigma}$). We study a general setting where each intervention can be applied to any subset of nodes and in any domain, which can be seen as a generalization of the more restricted settings in prior work (see Appendix E).

The intervention targets collection of these K interventions $\{\sigma^{(k)}\}_{k=1}^K$ is denoted as $\boldsymbol{\Psi} = \{\mathbf{I}^{(k)}\}_{k=1}^K$. Each intervention target $\mathbf{I}^{(k)}$ is given in the form of $\{V_i^{\Pi^{(k)},\{b\},t}, V_j^{\Pi^{(k)},\{b'\},t'}, \dots\}$, which indicates the intervention $\sigma^{(k)}$ changes the mechanism of $\{V_i, V_j, \dots\}$ in domain $\Pi^{(k)}$. The superscript $\{b\}$ indicates the mechanism of the intervention on the same node. The mechanisms of $V_i^{\{1\}}$ and $V_i^{\{2\}}$ are different while the mechanisms of different nodes ($V_i^{\{1\}}$ and $V_j^{\{1\}}$) is default different; the superscript $t = \text{do}$ indicates the intervention is hard. When $\{b\}$ or t is omitted, the intervention is assumed to be different mechanisms, or not hard, respectively. When $\mathbf{I}^{(k)}$ is an idle intervention in Π_n (i.e., observational), it is denoted as $\{\}^n$. The set $do[\mathbf{I}^{(k)}]$ is a set of variables with hard interventions in $\sigma^{(k)}$. $\boldsymbol{\Psi}_{\mathbf{T}}$ is a subset of $\boldsymbol{\Psi}$ such that $\mathbf{T} \subseteq do[\mathbf{I}^{(j)}]$ for every $\mathbf{I}^{(j)} \in \boldsymbol{\Psi}_{\mathbf{T}}$, which implies $\mathbf{I}^{(j)}$ contains hard interventions on $\mathbf{T}$; see Fig. S1 and Ex. 1 for an illustration of the notation.

Example 1 (**Notation illustration**)**.** Let an intervention target collection be

$$\boldsymbol{\Psi} = \{\mathbf{I}^{(1)} = \{\{\}^{\Pi_1}\}, \mathbf{I}^{(2)} = \{V_1^{\Pi_1,\{1\}}\}, \mathbf{I}^{(3)} = \{V_1^{\Pi_2,\{2\}}, V_2^{\Pi_2,\{1\},\text{do}}\}, \mathbf{I}^{(4)} = \{V_1^{\Pi_2,\{1\}}, V_2^{\Pi_2,\text{do}}\}\} \tag{1}$$

In words, $\boldsymbol{\Psi}$ represents 4 different interventions $\Sigma = \{\sigma^{(k)}\}_{k=1}^4$:
$\sigma^{(1)}$: an idle intervention is applied resulting in an observational distribution in the domain Π^1.
$\sigma^{(2)}$: a soft intervention with mechanism $\{1\}$ is applied to V_1 in domain Π^1.
$\sigma^{(3)}$: an intervention is applied to V_1 and V_2 in domain Π^2, where the mechanism of V_1 is different from $\sigma^{(2)}$ and the intervention on V_2 is hard.
$\sigma^{(4)}$: an intervention is applied to V_1 and V_2 in domain Π^2, where the mechanism of V_1 is the same with $\sigma^{(2)}$ and the mechanism of V_2 is different from $\sigma^{(3)}$.

[7] To our knowledge, this is the first work in disentangled causal representation learning to relax Markovianity, which we believe is important since a significant challenge in causal inference stems from the existence of confounding bias traced back to Rubin [61], Pearl [1, 62], and more recently data fusion [36].

[8] See [32, 33] and Appendix Sec. A.3 for a more detailed discussion on the fundamental differences between interventions and domains, and why modeling their distinction is fundamental for this task.

Domain	Observational	Interventional			
Π^1	$P^1_{\{\}}(\mathbf{X})$	$P^1_{v_i}(\mathbf{X})$	$P^1_{v_j}(\mathbf{X})$	$P^1_{v_i,v_j}(\mathbf{X})$	...
Π^2	$P^2_{\{\}}(\mathbf{X})$	$P^2_{v_i}(\mathbf{X})$	$P^2_{v_k}(\mathbf{X})$	$P^2_{v_i,v_k}(\mathbf{X})$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Π^N	$P^N_{\{\}}(\mathbf{X})$	$P^N_{v_l}(\mathbf{X})$	$P^N_{v_m}(\mathbf{X})$	$P^N_{v_l,v_j}(\mathbf{X})$	...

Table 2: **Possible distributions observed for any given causal representation learning task** - Each domain $\mathbf{\Pi} = \{\Pi^1, \Pi^2, ..., \Pi^N\}$ may contain observational and interventional distributions over latent variables $\mathbf{V}$, which are mixed via $f_{\mathbf{X}}$ to generate $\mathbf{X} \in \mathbb{R}^m$. The first row and column are interchangeable under the multi-domain intervention exchangeability assumption [32]. Prior work also requires distributions across the entire column (i.e. many domains must be observed), or entire row (i.e. an intervention per latent variable). This paper discusses a more general disentangled representation learning setting when an arbitrary combination of distributions from interventions and domains can be input (i.e. any combination of cells in yellow, and green).

$do[\mathbf{I}^{(3)}] = \{V_2\}$: $\sigma^{(3)}$ perfectly intervenes on $\{V_2\}$.
$\mathbf{\Psi}_{V_2} = \{\mathbf{I}^{(3)}, \mathbf{I}^{(4)}\}$; the interventions targets that contain hard interventions on V_2. $\mathbf{\Psi}_{\{\}} = \mathbf{\Psi}$. □

Domains vs Interventions. In previous studies, there has been a tendency to conflate the notions of interventions and domain shifts [65–69]. However, it is essential to recognize their distinctiveness, particularly when considering various real-world examples spanning different scientific domains that utilize observational and interventional data. The differentiation between interventions and domains is not only conceptually significant but has implications for causal inference and the characterization of corresponding causal structures as discussed in depth by [32]. Moreover, it is crucial to avoid conflating these qualitatively distinct concepts of interventions and domains, as highlighted in transportability analysis [63]. Pearl and Bareinboim have introduced clear semantics for (S) nodes (environments), presenting a unified representation in the form of selection diagrams [33, 35, 36].

By recognizing these differences, this work leverages any combination of observational and/or interventional data arising from multiple domains and develops a general approach to disentanglement learning compared to the literature. Specifically, the set of distributions used throughout multiple domains is shown in Table 2). We note that prior work generally considered either interventions in a single domain (top row in Π^1), where there must be an intervention per latent variable [14, 21], or observational distributions from many domains $\Pi^1, \Pi^2, ..., \Pi^N$ (first column). However, we examine a general setting, where an arbitrary collection of interventional, or observational data from an arbitrary combination of domains is available, which includes both yellow and green in Table 2.

Observed Distributions. The interventions $\mathbf{\Sigma} = \{\sigma^{(k)}\}_{k=1}^K$ induce distributions $\mathcal{P} = \{P^{(k)}\}_{k=1}^K$ in various domains, where $P^{(k)} = P^{\Pi^{(k)}}(\mathbf{X}; \sigma^{(k)})$. Considering an arbitrary pair of distributions $P^{(j)}, P^{(k)} \in \mathcal{P}$, we assume $P^{(j)}$ is sufficiently different from $P^{(k)}$, unless explicitly stated otherwise [9].

Problem Statement Suppose the underlying true model $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \ldots, \mathcal{M}_n \rangle$ over generative factors $\mathbf{V}$ induces the LSD G^S and a collection of distributions $\mathcal{P}$ over $\mathbf{X}$ resulting from the corresponding collection of interventions Σ. In disentangled representation learning works, a proxy generative models $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \ldots, \widehat{\mathcal{M}}_n \rangle$ is commonly used to approximate the true model $\boldsymbol{\mathcal{M}}$ and obtain $\widehat{h}(\mathbf{X})$, which is a mapping from the high-dimension variable $\mathbf{X}$, as the representation of the true $\mathbf{V}$. Specifically, after the proxy model $\widehat{\boldsymbol{\mathcal{M}}}$ is matched to G^S and the given $\mathcal{P}$, $\widehat{\mathbf{V}} = \widehat{f_{\mathbf{X}}^{-1}}(\mathbf{X})$ is representative of $\mathbf{V}$. The goal of this work is to evaluate the disentanglement relationships of the learned representations $\widehat{\mathbf{V}}$ and the true generative factors $\mathbf{V}$.

[9] A formal version of "sufficiently different" (Assumption 6), as well as other technical assumptions, are stated and discussed in Appendix A.2.

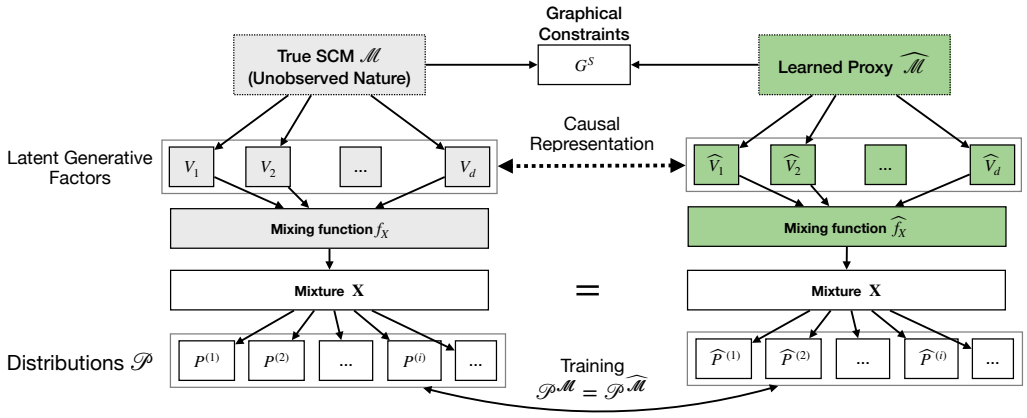


Figure 3: The relationship between the proxy model $\widehat{\mathcal{M}}$ and the true underlying model $\mathcal{M}$ in disentangled representation learning tasks. The distributions $\mathcal{P}$, and G^S are observed (white), while a proxy model, a causal representation, and (un)mixing function are learned (green) such that the decoded distributions match the observed distributions (i.e. $\mathcal{P}^{\mathcal{M}} = \mathcal{P}^{\widehat{\mathcal{M}}}$).

In the literature, the representation $\widehat{V}_i$ for $V_i \in \mathbf{V}$ is typically required to be disentangled from all other variables [7, 21] or some special subset (such as the non-ancestors of V_i) [21, 22]. In practice, sometimes only the target variables ($\mathbf{V}^{tar} \subseteq \mathbf{V}$) need to be disentangled from some user-chosen entangled variables ($\mathbf{V}^{en}$), as illustrated earlier in Fig. 2. We formally define this type of general indeterminacy next, and the formal version of our identification task.

Definition 2.3 (General Identifiability/Disentangleability (ID)). Consider a collection of ASCMs, $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \ldots, \mathcal{M}_n \rangle$ that induces an LSD G^S, and a set of distributions $\mathcal{P} = \{P^{(k)}\}_{k=1}^K$ resulting from K intervention sets Σ, and consider target variables $\mathbf{V}^{tar} \in \mathbf{V}$ and $\mathbf{V}^{en} \subseteq \mathbf{V} \backslash \mathbf{V}^{tar}$. The set $\mathbf{V}^{tar}$ is said to be identifiable (disentangled) with respect to (from) $\mathbf{V}^{en}$ if there exists a function $\boldsymbol{\tau}$ such that $\widehat{\mathbf{V}}^{tar} = \boldsymbol{\tau}(\mathbf{V} \backslash \mathbf{V}^{en})$ for any collection of ASCMs, $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \ldots, \widehat{\mathcal{M}}_n \rangle$, compatible with G^S and $\mathcal{P}^{\widehat{\mathcal{M}}} = \mathcal{P}$. For short, $\mathbf{V}^{tar}$ is said to be ID w.r.t. $\mathbf{V}^{en}$ from G^S and $\mathcal{P}$. □

To illustrate, consider a target variable $\mathbf{V}^{tar}$ such that one aims to obtain a representation that is disentangled from another subset of variables $\mathbf{V}^{en}$. The above definition states that $\mathbf{V}^{tar}$ is disentangled from $\mathbf{V}^{en}$ (or is ID w.r.t. $\mathbf{V}^{en}$) if the learned representations $\widehat{\mathbf{V}}^{tar}$ in $\widehat{\boldsymbol{\mathcal{M}}}$ is only a function of $\mathbf{V} \backslash \mathbf{V}^{en}$ for *any* $\widehat{\boldsymbol{\mathcal{M}}}$ that matches with the LSD G^S and distributions $\mathcal{P}$[10].

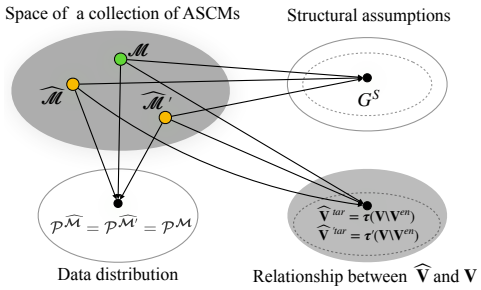


Figure 4: General ID and disentangleability (Def. 2.3).

An illustration of how the elements in this definition are related is shown in Fig. 4. Following the example of Fig. 2, suppose the user wants V_3 to be disentangled from V_1 while considering the entanglement between V_2 and V_3 acceptable. If $\widehat{V}_3 = \tau(V_2, V_3)$ for any ASCM $\widehat{\mathcal{M}}$ that matches the observed distributions and the LSD, V_3 is said to be ID w.r.t. V_1. This definition of disentanglement also allows for full disentanglement (i.e., any V_i is disentangled from other variables), as well as partial disentanglement given user-chosen target $\mathbf{V}^{tar}$ and $\mathbf{V}^{en}$. For concreteness, the next examples illustrate the difference between full disentanglement in prior work and partial disentanglement in general distributional settings (this work). We first examine Def. 2.3 in terms of the well-known nonlinear ICA setting.

[10] In general, this definition is considered after a permutation of variables; for details, refer to Sec. A.4.

Assumptions (Informal) and Modeling Concepts Before discussing the main theoretical contributions, we informally state our assumptions and provide remarks to ground the ASCM model [11].

Assumption 1 (Soft interventions without altering the causal structure)**.** *Interventions do not change the causal diagram. Hard interventions cut all incoming parent edges, and soft interventions preserve them [59]. However, more general interventions may arbitrarily change the parent set for any given node [59]. We do not consider such interventions, and leave this general case for future work.*

Assumption 2 (Known-target interventions)**.** *All interventions occur with known targets, reducing permutation indeterminacy for intervened variables.*

Assumption 3 (Sufficiently different distributions)**.** *Each pair of distributions $P^{(j)}$ and $P^{(k)} \in \mathcal{P}$ are sufficiently different, unless stated otherwise. This is naturally satisfied if ASCMs and interventions are randomly chosen [51]. Similar assumptions include the "genericity" [51], "interventional discrepancy" [21], and "sufficient changes" assumptions [10, 22].*

Remark 1 (Mixing is invertible)**.** As a consequence of Def. 2.1, the mixing function f_{*X} is invertible since it is defined as a diffeomorphism, ensuring that latent variables are uniquely learnable [9, 10, 17, 60].

Remark 2 (Confounders are not part of the mixing function)**.** According to Def. 2.1, latent exogenous variables $\mathbf{U}$ influence the high-dimensional mixture $\mathbf{X}$ only through latent causal variables $\mathbf{V}$, so unobserved confounding does not directly affect the mixing function. For instance, in EEG data, sleep quality and drug treatment may influence EEG appearance, while socioeconomic status may confound sleep and drug treatment but not directly affect EEG. This idea is also present in prior work, such as nonlinear ICA, where independent exogenous variables U_i each point to a single V_i. [7].

Remark 3 (Mixing function is shared across all domains)**.** According to Def. 2.1, the mixing function $f_{\mathbf{X}}$ is the same for all ASCMs $\mathcal{M}^i \in \boldsymbol{\mathcal{M}}$, enabling cross-domain analysis. If the mixing function varied across distributions, the latent representations would not be identifiable from iid data alone [9, 51].

Remark 4 (Shared causal structure)**.** As a consequence of Def. 2.2, each environment's ASCM shares the same latent causal graph, with no structural changes among latent variables [12].

Example 2 (**Disentanglement in ICA**)**.** In this example, we will formalize classic ICA problems with causal language introduced in this section. This example helps extend ICA tasks to general causal disentangled representation learning tasks later. We first introduce the true and unobserved generative model and some of its properties (I) and then the proxy model (II).

I. True model. Consider the true underlying ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \dots, \mathcal{M}_N \rangle$ in domains $\langle \Pi_1, \dots, \Pi_N \rangle$ over latent generative factors $\{V_1, V_2, V_3\}$ and mixture $\mathbf{X} = \{X_1, X_2, X_3\}$. In each domain Π_i, the SCM $\mathcal{M}_i$ can be written as:

$$\langle \mathbf{U} = \{U_1, U_2, U_3\}, \mathbf{V} = \{V_1, V_2, V_3\}, \mathbf{X} = \{X_1, X_2, X_3\}, \mathcal{F} = \{f_{V_1}, f_{V_2}, f_{V_3}, f_{\mathbf{X}}\}, P_i(\mathbf{U}) \rangle. \tag{2}$$

The low-level data $\mathbf{X}$ is generated through a two-step process in each domain. At the generative step, the mechanism of each generative factor $V_j (j = 1, 2, 3)$ is an identity mapping from the exogenous variable $U_j (i = 1, 2, 3)$, namely;

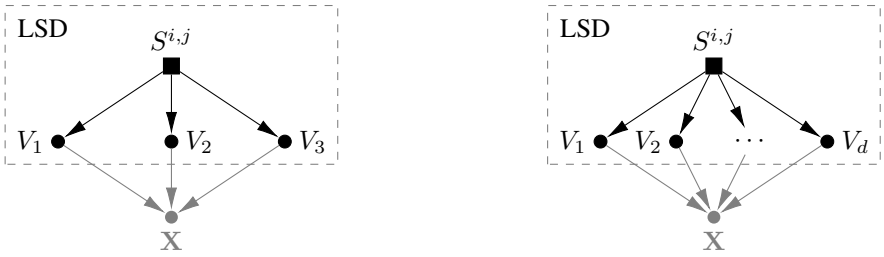
$$\begin{aligned} V_1 &\leftarrow f_{V_1}(U_1) = U_1 \\ V_2 &\leftarrow f_{V_2}(U_2) = U_2 \\ V_3 &\leftarrow f_{V_3}(U_3) = U_3 \end{aligned} \tag{3}$$

where the exogenous U_1, U_2, and U_3 are independent and

$$P_i(\mathbf{U}) = \begin{cases} U_1 \sim \mathcal{N}(0, i) \\ U_2 \sim \mathcal{N}(0, 1 + i) \\ U_3 \sim \mathcal{N}(0, 2 + i) \end{cases} \tag{4}$$

[11] For a formal discussion on the assumptions, we refer the readers to Appendix A.2)

[12] The assumption that there are no structural changes between domains can be relaxed and is considered in the context of inference when causal variables are fully observed, as discussed in [33]. This is an interesting topic for future explorations, and we do not consider this avenue here.



(a) (b)

Figure 5: (a) The LSD in the setting of nonlinear ICA with three generative factors in Example 2. (b) The LSD in the general setting of nonlinear ICA with d factors.

This implies that (1) all generative factors $\mathbf{V}$ are independent of each other in each domain; (2) the variance of each exogenous variable increases from domain Π_1 to Π_N.

At the second stage, the mixture function $f_{\mathbf{X}}$ maps latent V_1, V_2, and V_3 to the observed $\mathbf{X}$, where

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \leftarrow f_{\mathbf{X}}(\mathbf{V}) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \tag{5}$$

and $f_{\mathbf{X}}$ is shared across all the domains (see Def. 2.2 for details on this construction). Note that $f_{\mathbf{X}}$ is invertible, so there exists a mapping from $\mathbf{X}$ to $\mathbf{V}$, namely,

$$\mathbf{V} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \leftarrow f_{\mathbf{X}}^{-1}(\mathbf{X}) = \begin{pmatrix} 0.5 & 0.5 & 0 \\ 0.5 & -0.5 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \tag{6}$$

Putting all of these together, the true ASCM $\mathcal{M}_i$ in domain Π_i can be written as:

$$\mathcal{M}_i = \begin{cases} \mathbf{U} = \{U_1, U_2, U_3\} \\ \mathbf{V} = \{V_1, V_2, V_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} V_1 \leftarrow U_1 \\ V_2 \leftarrow U_2 \\ V_3 \leftarrow U_3 \\ \mathbf{X} = \begin{cases} X_1 \leftarrow V_1 + V_2 \\ X_2 \leftarrow V_1 - V_2 \\ X_3 \leftarrow V_1 + V_3 \end{cases} \end{cases} \\ P_i(\mathbf{U}) : \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} i & 0 & 0 \\ 0 & i+1 & 0 \\ 0 & 0 & i+2 \end{pmatrix}) \end{cases} \tag{7}$$

Latent structure of the true model. The LSD G^S induced by $\boldsymbol{\mathcal{M}}$ is shown in Fig. 5a since each V_j is independent of each other in each domain and all U_j changes across domains. The S-node here can be though of as auxiliary information in the ICA literature [7, 10] such that all generative factors are independent of each other given the auxiliary information about which domain each distribution comes from.

Input distribution induced by the true model. In each domain, the observational distribution over $\mathbf{X}$ is given, which is also known as the idle intervention. According to the notations introduced earlier in this section, the intervention targets $\Psi = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, \ldots, \{\}^{\Pi_N}\}$ and the corresponding given collection of distribution induced by $\boldsymbol{\mathcal{M}}$ is $\mathcal{P} = \langle P^{(1)}, P^{(2)}, \ldots, P^{(N)} \rangle$, where $P^{(k)} = P^{\Pi_k}(\mathbf{X})$. Since $\mathbf{X}$ is a linear mixture of independent Gaussian variables $\mathbf{V}$, $P^{(k)}$ also follows a Gaussian distribution:

$$P^{(k)} = P^{\Pi_k}(\mathbf{X}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 2k+1 & -1 & k \\ -1 & 2k+1 & k \\ k & k & 2k+2 \end{pmatrix}), k = 1, 2, \ldots, 5, \tag{8}$$

where

$$\begin{aligned} Var(X_1) &= Var(V_1) + Var(V_2) = 2k+1 \\ Var(X_2) &= Var(V_1) + Var(V_2) = 2k+1 \\ Var(X_3) &= Var(V_1) + Var(V_3) = 2k+2 \\ Cov(X_1, X_2) &= Var(V_1) - Var(V_2) = -1 \\ Cov(X_1, X_3) &= Var(V_1) = k \\ Cov(X_2, X_3) &= Var(V_1) = k \end{aligned} \tag{9}$$

Goal. The ICA tasks aim to find the unmixing function $f_{\mathbf{X}}^{-1}$ and recover the true generative factors $\{V_1, V_2, V_3\}$. However, it is hard to obtain such learn $f_{\mathbf{X}}^{-1}$ precisely since all $\mathbf{V}$ are unobserved and only sampled of $\mathbf{X}$ are given. Thus, the formal goal of the ICA task is to learn a relaxed approximation $\widehat{f}_{\mathbf{X}}^{-1} \approx f_{\mathbf{X}}^{-1}$ and the corresponding representations $\widehat{\mathbf{V}} = \{\widehat{V}_1, \widehat{V}_2, \widehat{V}_3\} = \widehat{f}_{\mathbf{X}}^{-1}(\mathbf{X})$ such that there exists a transformation $\boldsymbol{\tau} = \{\tau_1, \tau_2, \tau_3\}$ from the true latent generative factor V_i to $\widehat{V}_i$, namely:

$$\begin{aligned} \widehat{V}_1 &= \tau_1(V_1) \\ \widehat{V}_2 &= \tau_2(V_2) \\ \widehat{V}_3 &= \tau_3(V_3) \end{aligned} \tag{10}$$

To illustrate, functions τ_i must only be a transformation from one single true latent factor V_i to one single representation. In other words, after unmixing $\mathbf{X}$ by $\widehat{f_{\mathbf{X}}}^{-1}(\mathbf{X}) = \widehat{\mathbf{V}}$, each representation $\widehat{V}_i$ are disentangled from other factors and represent V_i distinctly.

II. Learned model. As mentioned earlier, the approach taken in the current literature can be thought of as learning a proxy model $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \ldots, \widehat{\mathcal{M}}_N \rangle$ to obtain such a representation. Specifically, existing work would construct a model to generate $\mathbf{X}$ with 3 independent underlying variables and then fit the distribution $\mathbf{X}$ with the given input $\mathcal{P}$. This can be regarded as the process of i) constraining $\widehat{\boldsymbol{\mathcal{M}}}$ with the LSD G^S and ii) fitting $\widehat{\boldsymbol{\mathcal{M}}}$ with the given distributions $\mathcal{P}$ to generate the same $\mathbf{X}$ as the true model (going back to [70]). We will use proxy models to describe the learning process in this work since it is smooth to generalize to settings where richer causal structures exist among generative factors $\mathbf{V}$, other than full independence. For concreteness, consider the learned $\widehat{\mathcal{M}}_i$ by such process as follows:

$$\widehat{\mathcal{M}}_i = \begin{cases} \widehat{\mathbf{U}} = \{\widehat{U}_1, \widehat{U}_2, \widehat{U}_3\} \\ \widehat{\mathbf{V}} = \{\widehat{V}_1, \widehat{V}_2, \widehat{V}_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} \widehat{V_1} \leftarrow 0.5\sqrt{i}\,\widehat{U}_1 \\ \widehat{V_2} \leftarrow 2\sqrt{1+i}\,\widehat{U}_2 \\ \widehat{V_3} \leftarrow 2\sqrt{2+i}\,\widehat{U}_3 \\ \mathbf{X} = \begin{cases} X_1 \leftarrow 2\widehat{V}_1 + 0.5\widehat{V}_2 \\ X_2 \leftarrow 2\widehat{V}_1 - 0.5\widehat{V}_2 \\ X_3 \leftarrow 2\widehat{V}_1 + 0.5\widehat{V}_3 \end{cases} \end{cases} \\ P^i(\widehat{\mathbf{U}}) = \begin{pmatrix} \widehat{U}_1 \\ \widehat{U}_2 \\ \widehat{U}_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}) \end{cases} \tag{11}$$

In fact, the $\widehat{\boldsymbol{\mathcal{M}}}$ above is compatible with the LSD in Fig. 5a since $\mathbf{Pa}(\widehat{V}_i)$ is the empty set, the exogenous variables U_i are independent of each other, and each latent variable distribution changes

over the N domains. Also, it is verifiable that the observational distribution $\widehat{P}^{\Pi_k}(\mathbf{X})$ in domain Π_k describes the following Gaussian distribution

$$\widehat{P}^{\Pi_k}(\mathbf{X}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 2k+1 & -1 & k \\ -1 & 2k+1 & k \\ k & k & 2k+2 \end{pmatrix}) \tag{12}$$

since

$$\begin{aligned} Var(X_1) &= 4Var(\widehat{V_1}) + 0.25Var(\widehat{V_2}) = 4.25k + 0.25 \times 4(k+1) = 2k+1 \\ Var(X_2) &= 4Var(\widehat{V_1}) + 0.25Var(\widehat{V_2}) = 4.25k + 0.25 \times 4(k+1) = 2k+1 \\ Var(X_3) &= 4Var(\widehat{V_1}) + 0.25Var(\widehat{V_2}) = 4.25k + 0.25 \times 4(k+2) = 2k+2 \\ Cov(X_1, X_2) &= 4Var(\widehat{V_1}) - 0.25Var(\widehat{V_2}) = 4.25k - 0.25 \times 4(k+1) = -1 \\ Cov(X_1, X_3) &= 4Var(\widehat{V_1}) = 4 \times 0.25k = k \\ Cov(X_2, X_3) &= 4Var(\widehat{V_1}) = 4 \times 0.25 = k \end{aligned} \tag{13}$$

Thus, the collection of observational distributions $\mathcal{P}^{\widehat{\mathcal{M}}} = \{\widehat{P}^{(1)}, \ldots, \widehat{P}^{(N)}\}$ matches with the given distribution $\mathcal{P}$ induced by the true generative model, $\mathbf{M}$ since $\widehat{P}^{(k)} = P^{(k)}$ (Eq.(8) and (12)).

Finally, we can now check whether the learned representation $\widehat{V}$ of the proxy model $\widehat{\mathcal{M}}$ is capable of distinctly representing $\mathbf{V}$ in the underlying true $\mathcal{M}$. Specifically, we now check whether $\widehat{V}$ satisfies the condition in Eq. 10. For a $\mathbf{X}$, we can obtain it by mixing $\mathbf{V}$ with $f_{\mathbf{X}}$, and by mixing $\widehat{\mathbf{V}}$ through $\widehat{f}_{\mathbf{X}}$, i.e.

$$\mathbf{X} = f_{\mathbf{X}}(\mathbf{V}) = \widehat{f}_{\mathbf{X}}(\widehat{\mathbf{V}}) \tag{14}$$

then we have

$$\widehat{\mathbf{V}} = \begin{pmatrix} \widehat{V}_1 \\ \widehat{V}_2 \\ \widehat{V}_3 \end{pmatrix} = \widehat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}(\mathbf{V}) = \underbrace{\begin{pmatrix} 2 & 0.5 & 0 \\ 2 & -0.5 & 0 \\ 2 & 0 & 0.5 \end{pmatrix}}_{\text{Proxy Model (Eq. 11)}}^{-1} \left[\underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{\text{True Unknown Model (Eq. 7)}} \underbrace{\begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix}}_{\text{Observed Distribution } \mathbf{X}} \right] \tag{15}$$

$$= \underbrace{\begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}}_{\text{Mapping } \mathbf{V} \text{ to } \widehat{\mathbf{V}}} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \tag{16}$$

This implies that the transformation from V_i to $\widehat{V}_i$ exists, i.e.,

$$\begin{aligned} \widehat{V}_1 &= \tau_1(V_1) = 0.5V_1 \\ \widehat{V}_2 &= \tau_2(V_2) = 2V_2 \\ \widehat{V}_3 &= \tau_3(V_3) = 2V_3 \end{aligned} \tag{17}$$

□

The above example demonstrates a learned proxy model $\widehat{\mathcal{M}}$ (Eq. 11) compatible with the input LSD G^S and capable of inducing distributions $\mathcal{P}$ can provide disentangled representations. However, in general, it is not guaranteed that a learned $\widehat{\mathcal{M}}$ will be a valid proxy model given above; for instance, what if one learns a model $\widehat{\mathcal{M}}'$ that is aligned with G^S and $\mathcal{P}$ but does not satisfy the condition in Eq. 10? The answer to this question is related to the problem of identifiability/disentangleability, as formalized in Def. 2.3. The full identifiability in an ICA setting says that for any $\widehat{\mathcal{M}}$ that matches with $\mathcal{P}$ and G^S, there exists a one-to-one mapping (subject to nonlinear scaling and suitable permutation) from V_i to $\widehat{V_{i'}}$. That is $\widehat{V_{i'}} = \tau_i(V_i)$ for all i (Eq. 10). Once full identifiability is achieved, one can obtain disentangled representations from any $\widehat{\mathcal{M}}$ by enforcing the structure G^S and fitting the given distributions $\mathcal{P}$.

Using proper causal language, the full identifiability task (Eq. 10) can be seen as a special case of the work proposed here. First, the input structure is restricted to the case where each V_i is mutually

independent in each domain. Second, only the observational distribution is given within each domain and interventional distributions are not considered. Third, the full identifiability/disentangleability in ICA is a special case of general identifiability (Def. 2.3) where $\mathbf{V}^{tar}$ is set as V_j for $(j = 1, \ldots, d)$ and $\mathbf{V}^{en}$ is set as $\mathbf{V} \backslash \{V_j\}$, which means V_j is expected to be disentangled from all other variables. [13]

Formally, we can write the typical ICA identifiability task in causal notation as follows:

Task 1 (**Full Disentanglement in ICA**). *Supose the true underlying ASCMs $\boldsymbol{\mathcal{M}}$ induces independent $\mathbf{V}$ and the observational distributions from N different domains are given. In other words, the true underlying ASCMs $\boldsymbol{\mathcal{M}}$ induces the LSD G^S in Fig. 5b and observational distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, \ldots, P^{(N)}\}$ in domains $\{\Pi_1, \Pi_2, \ldots, \Pi_N\}$. Given intervention targets*

$$\boldsymbol{\Psi} = \{\mathbf{I}^{(1)}, \ldots, \mathbf{I}^{(N)}\} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, \ldots, \{\}^{\Pi_N}\} \tag{18}$$

and G^S, the task is to determine whether V_i is ID w.r.t. $\{V_1, V_2, ..., V_{i-1}, V_{i+1}, ..., V_d\}$ for all $V_i \in \mathbf{V}$. □

It can be shown that if the number of domains N is no less than $2d + 1$, where d is the number of generative factors, full disentanglement can be achieved in the ICA setting [7]. In the specific example above, the number of generative factors is 3 ($d = 3$). We next show a derivation based on Example 2 to demonstrate how this result holds true.

Example 3 (**Proxy model for ICA and full disentanglement**). Consider an arbitrary proxy model $\widehat{\boldsymbol{\mathcal{M}}}$ compatible with G^S and $\mathcal{P}$. Note that this $\widehat{\boldsymbol{\mathcal{M}}}$ is not necessarily the one listed above (Eq.(11)). Denote $\boldsymbol{\phi} = \widehat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}$, where $\widehat{f}_{\mathbf{X}}^{-1}$ is the unmixing function of the proxy model and $f_{\mathbf{X}}$ is the mixture function of the true model. According to Eq. 14,

$$\widehat{\mathbf{V}} = \begin{pmatrix} \widehat{V}_1 \\ \widehat{V}_2 \\ \widehat{V}_3 \end{pmatrix} = \boldsymbol{\phi}(\mathbf{V}) = \begin{pmatrix} \phi_1(V_1, V_2, V_3) \\ \phi_2(V_1, V_2, V_3) \\ \phi_3(V_1, V_2, V_3) \end{pmatrix} \tag{19}$$

This implies $\widehat{V}_i$ is naturally a mixture of all generative factors $\{V_1, V_2, V_3\}$ after fitting the proxy model with given distribution $\mathcal{P}$. That is, without further information, the representations are fully entangled. To check whether $\mathcal{M}_i$ satisfies the full disentanglement conditions (Eq. 10), one can check whether each ϕ_i is only a function of V_i and is not a function of the other variables. For example, if

$$\frac{\partial \widehat{V}_1}{\partial V_2} = \frac{\partial \phi_1}{\partial V_2} = 0, \frac{\partial \widehat{V}_1}{\partial V_3} = \frac{\partial \phi_1}{\partial V_3} = 0, \tag{20}$$

then there exists a transformation from only V_1 to $\widehat{V}_1$, i.e.,

$$\widehat{V}_1 = \phi_1(V_1, \cdot, \cdot) = \tau(V_1), \tag{21}$$

where $\phi_1(V_1, \cdot, \cdot)$ is the function derived from V_1 that does not depend on V_2 and V_3. We will show this by *comparing distributions*. The well-known change-of-variable formula allows one to compare distributions under bijective transformations [71, 72].

Here, we apply the change-of-variable formula to the density function $p^{(k)}(\mathbf{v})$ over the true factors $\mathbf{V}$ in domain Π_k with Eq. 19,

$$p^{(k)}(\mathbf{v}) = p^{(k)}(\widehat{\mathbf{v}}) |\det J_{\boldsymbol{\phi}}| \tag{22}$$

where

$$J_{\boldsymbol{\phi}} = \begin{pmatrix} \frac{\partial \widehat{V}_1}{\partial V_1} & \frac{\partial \widehat{V}_1}{\partial V_2} & \frac{\partial \widehat{V}_1}{\partial V_3} \\ \frac{\partial \widehat{V}_2}{\partial V_1} & \frac{\partial \widehat{V}_2}{\partial V_2} & \frac{\partial \widehat{V}_2}{\partial V_3} \\ \frac{\partial \widehat{V}_3}{\partial V_1} & \frac{\partial \widehat{V}_3}{\partial V_2} & \frac{\partial \widehat{V}_3}{\partial V_3} \end{pmatrix} \tag{23}$$

is the Jacobian matrix. Note that this matrix is full rank ($|\det J_{\boldsymbol{\phi}}| \neq 0$) since $f_{\mathbf{X}}$ and $\widehat{f_{\mathbf{X}}^{-1}}$ are invertible. According to the fact that each V_j is independent of each other in domain Π_k,

$$p^{(k)}(v_1) p^{(k)}(v_2) p^{(k)}(v_3) = p^{(k)}(\widehat{v}_1) p^{(k)}(\widehat{v}_2) p^{(k)}(\widehat{v}_3) |\det J_{\boldsymbol{\phi}}| \tag{24}$$

[13] Notice that the original identifiability definition in ICA literature allows a permutation among generative factors. To illustrate, $\widehat{V}_1, \widehat{V}_2, \ldots, \widehat{V}_d$ does not necessarily represent $V_1, V_2, \ldots, V_d$ respectively but can represent $V_{\pi(1)}, V_{\pi(2)}, \ldots, V_{\pi(d)}$ where $\pi(1), \pi(2), \ldots, \pi(d)$ is a permutation over $\{1, 2, \ldots, d\}$. Def. 2.3 applies after such a permutation of variables; for details, please refer to Sec. A.4.

Taking the logarithm of the above equation,

$$\log p^{(k)}(\mathbf{v}) = \sum_{i=1}^{3} \log p^{(k)}(v_i) = \sum_{i=1}^{3} \log p^{(k)}(\widehat{v}_i) + \log |\det J_{\boldsymbol{\phi}}| \tag{25}$$

In each domain, from Eq. 25, we can see the discrepancy of density function between $\mathbf{V}$ and $\widehat{\mathbf{V}}$ is the determinant of the Jacobian matrix. Note that the Jacobian is the same in each domain comparison since both $f_{\mathbf{X}}$ and $\widehat{f}_{\mathbf{X}}$ are shared across domains by Def. 2.1. Then, by comparing the distributions across domains $P^{(k)}(\mathbf{V})$ for $(k = 2, \ldots, N)$, and $P^{(1)}(\mathbf{V})$ (subtracting $\log p^{(1)}(\mathbf{v})$ from $\log p^{(k)}(\mathbf{v})$), we can write

$$\sum_{i=1}^{3} \log p^{(k)}(v_i) - \log p^{(1)}(v_i) = \sum_{i=1}^{3} \log p^{(k)}(\widehat{v}_i) - \log p^{(1)}(\widehat{v}_i) \tag{26}$$

Note that the Jacobian matrix factor is canceled out. Next, taking derivatives of both sides of Eq.(26) w.r.t. v_1,

$$\frac{\partial \log p^{(k)-(1)}(v_1)}{\partial v_1} = \frac{\partial \log p^{(k)-(1)}(\widehat{v}_1)}{\partial \widehat{v}_1}\frac{\partial \widehat{v}_1}{\partial v_1} + \frac{\partial \log p^{(k)-(1)}(\widehat{v}_2)}{\partial \widehat{v}_2}\frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(k)-(1)}(\widehat{v}_3)}{\partial \widehat{v}_3}\frac{\partial \widehat{v}_3}{\partial v_1} \tag{27}$$

where $p^{(k)-(1)}(v) = p^{(k)}(v) - p^{(1)}(v)$. Then taking the derivatives of both sides of Eq. 27 w.r.t. v_2,

$$0 = a'_{k1}\frac{\partial^2 \widehat{v}_1}{\partial v_1 \partial v_2} + a''_{k2}\frac{\partial^2 \widehat{v}_2}{\partial v_1 \partial v_2} + a''_{k3}\frac{\partial^2 \widehat{v}_3}{\partial v_1 \partial v_2} + a''_{k1}\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_2} + a''_{k2}\frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_2} + a''_{k3}\frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_2} \tag{28}$$

where

$$\begin{aligned} a'_{kj} &= \frac{\partial \log p^{(k)-(1)}(\widehat{v}_j)}{\partial \widehat{v}_j}, \quad \text{for } j = 1, 2, 3 \\ a''_{kj} &= \frac{\partial^2 \log p^{(k)-(1)}(\widehat{v}_j)}{\partial \widehat{v}_j^2}, \quad \text{for } j = 1, 2, 3 \end{aligned} \tag{29}$$

Compose Eq. 28 from $k = 2$ to N, a linear system with unknowns can be built as follows:

$$\begin{pmatrix} a'_{21} & a'_{22} & a'_{23} & a''_{21} & a''_{22} & a''_{23} \\ a'_{31} & a'_{32} & a'_{33} & a''_{31} & a''_{32} & a''_{33} \\ \vdots & & \vdots & \vdots & & \vdots \\ a'_{N1} & a'_{N2} & a'_{N3} & a''_{N1} & a''_{N2} & a''_{N3} \end{pmatrix} \mathbf{y} = 0 \tag{30}$$

where the unknowns are:

$$\mathbf{y} = \begin{pmatrix} \frac{\partial^2 \widehat{v}_1}{\partial v_1 \partial v_2} & \frac{\partial^2 \widehat{v}_2}{\partial v_1 \partial v_2} & \frac{\partial^2 \widehat{v}_3}{\partial v_1 \partial v_2} & \frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_2} & \frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_2} & \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_2} \end{pmatrix}^{\mathrm{T}} \tag{31}$$

When $N \geq 7$ $(7 = 2 \times 3 + 1)$ and the rows of the coefficient matrix are linearly independent, this system only has the zero solution, i.e., $\mathbf{y} = 0$. Thus, we have

$$\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_2} = 0, \frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_2} = 0, \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_2} = 0. \tag{32}$$

Following the same procedure as above and taking derivatives of both sides of Eq. 26 w.r.t. v_1 and v_3, we have

$$\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3} = 0. \tag{33}$$

And taking derivatives of both sides of Eq. 26 w.r.t. v_2 and v_3, we have

$$\frac{\partial \widehat{v}_1}{\partial v_2}\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_2}{\partial v_2}\frac{\partial \widehat{v}_2}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_2}\frac{\partial \widehat{v}_3}{\partial v_3} = 0 \tag{34}$$

Now, consider the first part of Eq. 32, 33, and 34, namely the three constraints:

$$\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_2} = 0, \frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_1}{\partial v_2}\frac{\partial \widehat{v}_1}{\partial v_3} = 0. \tag{35}$$

This implies that $\partial \widehat{v}_1/\partial v_j = 0$ and $\partial \widehat{v}_1/\partial v_k = 0$ for $\{j, k\} \subseteq \{1, 2, 3\}$. Note that a permutation π can be arbitrarily applied to the ordering of $\widehat{v}_i$, such that the constraints still hold. For example, if $\widehat{v}_1, \widehat{v}_2, \widehat{v}_3$ is associated with v_2, v_3, v_1 in that order, then: $\partial \widehat{v}_1/\partial v_3 = 0$ and $\partial \widehat{v}_1/\partial v_1 = 0$. Thus w.l.o.g. [14], let $\widehat{v}_i$ be permuted such that:

$$\frac{\partial \widehat{v}_1}{\partial v_2} = 0, \frac{\partial \widehat{v}_1}{\partial v_3} = 0 \tag{36}$$

This implies V_1 is disentangled from V_2 and V_3 according to Eq. 21. In practice, the permutation can be examined over the learned representations to match each representation $\widehat{v}_i$ with a latent generative factor v_j by considering the correlation between $\widehat{v}_i$ and v_j over all combinations of i, j. Similarly, the second part and the third part of Eq. 32, 33, and 34 imply that

$$\begin{aligned} \frac{\partial \widehat{v}_2}{\partial v_1} = 0, \frac{\partial \widehat{v}_2}{\partial v_3} = 0, \\ \frac{\partial \widehat{v}_3}{\partial v_1} = 0, \frac{\partial \widehat{v}_3}{\partial v_2} = 0 \end{aligned} \tag{37}$$

after permutation otherwise the Jacobian matrix (Eq. 23) would be singular. This implies that V_2 is ID w.r.t V_1, and V_3 and V_3 is ID w.r.t V_1 and V_2.

□

As the example above suggests, the typical ICA problem aims to find a fully disentangled representation where latent generative factors are independent of each other. Current literature studies this case with various parametric assumptions and input data, and provides different conditions for the full disentanglement problem [7, 8, 10, 17, 73]. More recently, specific causal relationships are allowed among $\mathbf{V}$, generalizing the original ICA problem, as illustrated next.

Full disentanglement. We now consider the full disentanglement representation introduced in Task 1 for the Markovian ASCM setting. The formal ID task can be written as follows:

Task 2 (Full Disentanglement in Markovian ASCM). *Suppose the true underlying ASCMs $\boldsymbol{\mathcal{M}} = \mathcal{M}$ induces the LSD G^S in Fig. 6, a causal chain, and observational distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ in a single domain Π_1. Given intervention targets*

$$\Psi = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, V_1^{\Pi_1}, V_2^{\Pi_1}, V_3^{\Pi_1}\} \tag{38}$$

and G^S, the goal of the task is to determine whether V_1 is ID w.r.t. $\{V_2, V_3\}$, V_2 is ID w.r.t $\{V_1, V_3\}$, and V_3 is ID w.r.t $\{V_1, V_2\}$. □

Observe that the task in the Markovian case is not exactly the same as the ICA setting, and now includes causal (and probabilistic) relationships among the latent factors encoded by the LSD G^S.

Example 4 (Disentanglement in Markovian ASCM). Consider a similar setting as Ex. 2, where the true underlying $\boldsymbol{\mathcal{M}}$ describes the generation process of factors $\{V_1, V_2, V_3\}$ and the mixture $\mathbf{X} = \{X_1, X_2, X_3\}$. Unlike the domain discrepancy in Ex. 2, in this example, the true $\boldsymbol{\mathcal{M}}$ exists in a single domain Π_1, i.e., $\boldsymbol{\mathcal{M}} = \mathcal{M}$, where

$$\mathcal{M} = \langle \mathbf{U} = \{U_1, U_2, U_3\}, \mathbf{V} = \{V_1, V_2, V_3\}, \mathbf{X} = \{X_1, X_2, X_3\}, \mathcal{F} = \{f_{V_1}, f_{V_2}, f_{V_3}, f_{\mathbf{X}}\}, P(\mathbf{U})\rangle, \tag{39}$$

In contrast to the scenario in Ex. 2 where all generative factors are independent, V_i now have causal relationships encoded via directed edges in the graph. Specifically, the mechanisms of generative factors are as follows:

$$\begin{aligned} V_1 &\leftarrow f_{V_1}(U_1) = U_1 \\ V_2 &\leftarrow f_{V_2}(V_1, U_2) = V_1 + U_2 \\ V_3 &\leftarrow f_{V_3}(V_2, U_3) = V_2 + U_3 \end{aligned} \tag{40}$$

where the exogenous U_1, U_2, and U_3 are independent of each other and given by:

$$P_i(\mathbf{U}) = \begin{cases} U_1 \sim \mathcal{N}(0, 1) \\ U_2 \sim \mathcal{N}(0, 2) \\ U_3 \sim \mathcal{N}(0, 3) \end{cases} \tag{41}$$

[14] If $j = 1, k = 2$ or $j = 1, k = 3$, a permutation π can be applied to $1, 2, 3$ to let $\pi(j) = 2$ and $\pi(j) = 3$

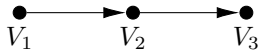


Figure 6: Example LSD for a Markovian ASCM.

The mechanisms of the generative factors implies that V_1 has a direct effect on V_2, and V_2 has a direct effect on V_3. The exogenous variable distributions imply there is no latent confounding among the generative factors. The LSD G^S induced by $\boldsymbol{\mathcal{M}}$ is shown in Fig. 6. To illustrate, an arrow points from V_1 to V_2 since f_{V_2} takes V_2 as input. Similarly, an arrow points from V_2 to V_3 since f_{V_3} takes V_2 as input. Also, there is no bidirected arrow between nodes since exogenous variables U_1, U_2, and U_3 are independent.

Similar to Ex. 2, $\{V_1, V_2, V_3\}$ are mapped to generate $\mathbf{X} = \{X_1, X_2, X_3\}$ by the same invertible mixing function $f_{\mathbf{X}}$ (Eq. 5). Formally, the ASCM $\mathcal{M}$ can be written as:

$$\mathcal{M} = \begin{cases} \mathbf{U} = \{U_1, U_2, U_3\} \\ \mathbf{V} = \{V_1, V_2, V_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} V_1 \leftarrow U_1 \\ V_2 \leftarrow V_1 + U_2 \\ V_3 \leftarrow V_2 + U_3 \end{cases} \\ \mathbf{X} = \begin{cases} X_1 \leftarrow V_1 + V_2 \\ X_2 \leftarrow V_1 - V_2 \\ X_3 \leftarrow V_1 + V_3 \end{cases} \\ P(\mathbf{U}) : \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}) \end{cases} \tag{42}$$

Suppose a collection of 4 interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)}\}$ are applied to $\mathcal{M}$ and induces a collection of distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ over $\mathbf{X}$, where

$$\begin{aligned} \sigma^{(1)} &= \{\} \\ \sigma^{(2)} &= \sigma_{V_1} : V_1 \leftarrow U_1', U_1' \sim \mathcal{N}(0, 0.5) \\ \sigma^{(3)} &= \sigma_{V_2} : V_2 \leftarrow V_1 + U_2', U_2' \sim \mathcal{N}(0, 0.5) \\ \sigma^{(4)} &= \sigma_{V_3} : V_3 \leftarrow V_2 + U_3', U_3' \sim \mathcal{N}(0, 0.5) \end{aligned} \tag{43}$$

and $U_1, U_1', U_2, U_2', U_3, U_3'$ are independent of each other. Then, the corresponding interventional targets are

$$\Psi = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, V_1^{\Pi_1}, V_2^{\Pi_1}, V_3^{\Pi_1}\} \tag{44}$$

indicating that the first distribution is observational, the second contains an intervention on V_1, the third contains an intervention on V_2, the fourth contains an intervention on V_3, all in domain Π_1. Note these interventions are all soft, and preserve the structure of the LSD.

The first intervention $\sigma^{(1)}$ is an idle intervention and leads to an observational distribution $P(\mathbf{X})$, where

$$P^{(1)} = P(\mathbf{X}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 6 & -2 & 6 \\ -2 & 2 & -2 \\ 6 & -2 & 9 \end{pmatrix}) \tag{45}$$

since

$$\begin{aligned} Var(X_1) &= Var(V_1) + Var(V_2) + 2Cov(V_1, V_2) = 6 \\ Var(X_2) &= Var(V_1) + Var(V_2) - 2Cov(V_1, V_2) = 2 \\ Var(X_3) &= Var(V_1) + Var(V_3) + 2Cov(V_1, V_3) = 9 \\ Cov(X_1, X_2) &= Var(V_1) - Var(V_2) = -2 \\ Cov(X_1, X_3) &= Var(V_1) + Cov(V_1, V_2) + Cov(V_1, V_3) + Cov(V_2, V_3) = 6 \\ Cov(X_2, X_3) &= Var(V_1) - Cov(V_1, V_2) + Cov(V_1, V_3) - Cov(V_2, V_3) = -2 \end{aligned} \tag{46}$$

Other interventions $\sigma^{(2)}, \sigma^{(3)}$, and $\sigma^{(4)}$ are soft interventions applied to V_1, V_2, V_3, respectively, and lead to interventional distributions $P(\mathbf{X}; \sigma_{V_1})$, $P(\mathbf{X}; \sigma_{V_2})$, and $P(\mathbf{X}; \sigma_{V_3})$. For example, consider $\sigma^{(3)}$ corresponding to an intervention on V_2 in domain Π_1. After $\sigma^{(3)}$ is applied to $\mathcal{M}$, the mechanism f_{V_2} of V_2 is replaced by $\sigma^{(3)}$. In other words, the soft intervention induces a submodel $\mathcal{M}_{\sigma^{(3)}}$ from $\mathcal{M}$, where the factors in blue are changed from the original ASCM due to the intervention.

$$\mathcal{M}_{\sigma^{(3)}} = \begin{cases} \mathbf{U} = \{U_1, U_2, U_3, U_2'\} \\ \mathbf{V} = \{V_1, V_2, V_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} V_1 \leftarrow U_1 \\ V_2 \leftarrow V_1 + U_2' \\ V_3 \leftarrow V_2 + U_3 \\ \mathbf{X} = \begin{cases} X_1 \leftarrow V_1 + V_2 \\ X_2 \leftarrow V_1 - V_2 \\ X_3 \leftarrow V_1 + V_3 \end{cases} \end{cases} \\ P(\mathbf{U}) : \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_2' \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0.5 \end{pmatrix}) \end{cases} \tag{47}$$

Based on the submodel, $\mathcal{M}_{\sigma^{(3)}}$, one can derive the observed distribution $P^{(3)}$,

$$P^{(3)} = P(\mathbf{X}; \sigma^{(3)}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 4.5 & -0.5 & 4.5 \\ -0.5 & 0.5 & -0.5 \\ 4.5 & -0.5 & 7.5 \end{pmatrix}) \tag{48}$$

Similarly, $P^{(2)}$ and $P^{(3)}$ can be derived as

$$P^{(2)} = P(\mathbf{X}; \sigma^{(2)}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 4 & -2 & 4 \\ -2 & 2 & -2 \\ 4 & -2 & 7 \end{pmatrix}), \tag{49}$$

$$P^{(4)} = P(\mathbf{X}; \sigma^{(4)}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 4.5 & -0.5 & 4.5 \\ -0.5 & 0.5 & -0.5 \\ 4.5 & -0.5 & 6.5 \end{pmatrix}) \tag{50}$$

Suppose one learns the proxy model $\widehat{\boldsymbol{\mathcal{M}}} = \widehat{\mathcal{M}}$ compatible with the G^S and $\mathcal{P}$ to obtain representations $\widehat{\mathbf{V}}$, where the proxy ASCM $\widehat{\mathcal{M}}$ is

$$\langle \widehat{\mathbf{U}} = \{\widehat{U}_1, \widehat{U}_2, \widehat{U}_3\}, \mathbf{V} = \{\widehat{V}_1, \widehat{V}_2, \widehat{V}_3\}, \mathbf{X} = \{X_1, X_2, X_3\}, \mathcal{F} = \{\widehat{f}_{V_1}, \widehat{f}_{V_2}, \widehat{f}_{V_3}, \widehat{f}_{\mathbf{X}}\}, P(\widehat{\mathbf{U}}) \rangle, \tag{51}$$

To illustrate full disentanglement, as discussed in Ex. 2, the task is to check whether for any $\widehat{\boldsymbol{\mathcal{M}}}$ compatible with G^S and $\mathcal{P}$, $\widehat{V}_i$ is always a mapping from V_i, i.e.,

$$\begin{aligned} \widehat{V}_1 &= \tau_1(V_1) \\ \widehat{V}_2 &= \tau_2(V_2) \\ \widehat{V}_3 &= \tau_3(V_3) \end{aligned} \tag{52}$$

However, full disentanglement is not achieved in this setting. To witness, we provide a counterexample following a proxy model $\widehat{\boldsymbol{\mathcal{M}}}$

$$\widehat{\mathcal{M}} = \begin{cases} \widehat{\mathbf{U}} = \{\widehat{U}_1, \widehat{U}_2, \widehat{U}_3\} \\ \widehat{\mathbf{V}} = \{\widehat{V}_1, \widehat{V}_2, \widehat{V}_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} \widehat{V_1} \leftarrow \widehat{U}_1 \\ \widehat{V_2} \leftarrow \widehat{V}_1 - \sqrt{2}\widehat{U}_2 \\ \widehat{V_3} \leftarrow \widehat{V}_2 + \sqrt{3}\,\widehat{U}_3 \end{cases} \\ \mathbf{X} = \begin{cases} X_1 \leftarrow 3\widehat{V}_1 - \widehat{V}_2 \\ X_2 \leftarrow -\widehat{V}_1 + \widehat{V}_2 \\ X_3 \leftarrow 3\widehat{V}_1 - 2\widehat{V}_2 + \widehat{V}_3 \end{cases} \\ P^i(\widehat{\mathbf{U}}) = \begin{pmatrix} \widehat{U}_1 \\ \widehat{U}_2 \\ \widehat{U}_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}) \end{cases} \tag{53}$$

First, we verify that $\widehat{\boldsymbol{\mathcal{M}}}$ is compatible with the G^S in Fig. 6 since $\widehat{f}_{V_2}$ takes V_1 as input and $\widehat{f}_{V_3}$ takes V_2 as input and all exogenous are independent with each other. Then, we consider the collection of interventions $\widehat{\Sigma} = \{\widehat{\sigma}^{(1)}, \widehat{\sigma}^{(2)}, \widehat{\sigma}^{(3)}, \widehat{\sigma}^{(4)}\}$ applied to $\widehat{\boldsymbol{\mathcal{M}}}$:

$$\begin{aligned} \widehat{\sigma}^{(1)} &= \{\} \\ \widehat{\sigma}^{(2)} &= \sigma_{\widehat{V}_1} : \widehat{V}_1 \leftarrow \widehat{U}_1', \widehat{U}_1' \sim \mathcal{N}(0, 0.5) \\ \widehat{\sigma}^{(3)} &= \sigma_{\widehat{V}_2} : \widehat{V}_2 \leftarrow \widehat{V}_1 - \widehat{U}_2', \widehat{U}_2' \sim \mathcal{N}(0, 0.5) \\ \widehat{\sigma}^{(4)} &= \sigma_{\widehat{V}_3} : \widehat{V}_3 \leftarrow \widehat{V}_2 + \widehat{U}_3', \widehat{U}_3' \sim \mathcal{N}(0, 0.5) \end{aligned} \tag{54}$$

It is verifiable that

$$\widehat{P}^{(1)} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 6 & -2 & 6 \\ -2 & 2 & -2 \\ 6 & -2 & 9 \end{pmatrix}) \tag{55}$$

which is equivalent to $P^{(1)}$ (Eq. 45) since

$$\begin{aligned} Var(X_1) &= 9Var(\widehat{V}_1) + Var(\widehat{V}_2) - 6Cov(\widehat{V}_1, \widehat{V}_2) = 4Var(\widehat{U}_1) + 2Var(\widehat{U}_2) = 6 \\ Var(X_2) &= Var(\widehat{V}_1) + Var(\widehat{V}_2) - 2Cov(\widehat{V}_1, \widehat{V}_2) = 2Var(\widehat{U}_2) = 2 \\ Var(X_3) &= 9Var(\widehat{V}_1) + 4Var(\widehat{V}_2) + Var(\widehat{V}_3) - 12Cov(\widehat{V}_1, \widehat{V}_2) + 6Cov(\widehat{V}_1, \widehat{V}_3) - 4Cov(\widehat{V}_2, \widehat{V}_3) \\ &= 4Var(\widehat{U}_1) + 2Var(\widehat{U}_2) + 3Var(\widehat{U}_3) = 9 \\ Cov(X_1, X_2) &= -3Var(\widehat{V}_1) - Var(\widehat{V}_2) + 4Cov(\widehat{V}_1, \widehat{V}_2) \\ &= -2Var(\widehat{U}_2) = -2 \\ Cov(X_1, X_3) &= 9Var(\widehat{V}_1) + 2Var(\widehat{V}_2) - 9Cov(\widehat{V}_1, \widehat{V}_2) + 3Cov(\widehat{V}_1, \widehat{V}_3) - Cov(\widehat{V}_2, \widehat{V}_3) \\ &= 4Var(\widehat{U}_1) + 2Var(\widehat{U}_2) = 6 \\ Cov(X_2, X_3) &= -3Var(\widehat{V}_1) - 2Var(\widehat{V}_2) + 5Cov(\widehat{V}_1, \widehat{V}_2) - Cov(\widehat{V}_1, \widehat{V}_3) + Cov(\widehat{V}_2, \widehat{V}_3) \\ &= -2Var(\widehat{U}_2) = -2 \end{aligned} \tag{56}$$

and similarly,

$$\widehat{P}^{(2)} = P^{(2)}, \widehat{P}^{(3)} = P^{(3)}, \widehat{P}^{(4)} = P^{(4)}. \tag{57}$$

Now, consider the mapping from $\widehat{\mathbf{V}}$ to $\mathbf{V}$ through the mixing functions $\widehat{f}_{\mathbf{X}}$ and $f_{\mathbf{X}}$,

$$\widehat{\mathbf{V}} = \begin{pmatrix} \widehat{V}_1 \\ \widehat{V}_2 \\ \widehat{V}_3 \end{pmatrix} = \widehat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}(\mathbf{V}) = \begin{pmatrix} 3 & -1 & 0 \\ 1 & 1 & 0 \\ 3 & -2 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \tag{58}$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 2 & -1 & 0 \\ 2 & -2 & 1 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \tag{59}$$

This implies the learned representation $\widehat{V}_1$ is a function of V_1, $\widehat{V}_2$ is a mixture of V_2 and V_3, and $\widehat{V}_3$ is a mixture of V_2 and V_3, i.e.,

$$\widehat{V}_1 = V_1, \quad \widehat{V}_2 = 2V_1 - V_2, \quad \widehat{V}_3 = 2V_1 - 2V_2 + V_3. \tag{60}$$

Thus, there does not exist τ_1, τ_2, τ_3 such that Eq. 52 holds. In the context of Figure 4, we see that identifiability is not achieved because the constraints imposed by the distributions and the LSD are not strong enough to fully disentangle the representation; there exists a model compatible with the constraints in which disentanglement is not achieved. □

In the previous example, a proxy model was described that matches the latent causal graph, and the observed distributions, but does not exhibit full disentanglement as described in Task 2.

Ancestral Disentanglement. Apart from full disentanglement, some special types of partial disentanglement are also studied in the literature. For example, *ancestral disentanglement* allows each V_i to be entangled with its ancestors in the LSD [21]. Specifically, ancestral disentanglement states that for any $\widehat{\mathcal{M}}$ that induces G^S and $\mathcal{P}$, there always exists a mapping from $\overline{(Anc)}(V_j)$ [15] to the representation $\widehat{V}_i$.

Using the causal machinery developed here, the ancestral disentanglement task can also be seen as a special case of our work. First, the input LSD is restricted to the Markovian setting (as opposed to general non-Markovian) where unobserved confounders are assumed away apriori. Second, interventions per node are applied to the true model, and an observational distribution is assumed to be given as input, as opposed to considering any observation, or interventional distributions that arise among heterogenous domains. Third, ancestral disentanglement is a special case of general identifiability (Def. 2.3), where $\mathbf{V}^{tar}$ is set as V_j for $(j = 1, \ldots, d)$ and $\mathbf{V}^{en}$ is set as $\mathbf{V} \backslash \overline{\mathbf{Anc}(V_j)}$. This means that V_j is expected to be disentangled from all other variables except its ancestors.

Formally, the ancestral identifiability task for Ex. 4 can be described as follows:

Task 3 (Ancestral Disentanglement in Markovian ASCM). *Suppose the true underlying ASCMs $\mathcal{M} = \mathcal{M}$ induces the LSD G^S in Fig. 6 and distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ in a single domain Π_1. Given intervention targets*

$$\Psi = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, V_1^{\Pi_1}, V_2^{\Pi_1}, V_3^{\Pi_1}\} \tag{61}$$

and G^S, the task is to determine whether V_1 is ID w.r.t. $\{V_2, V_3\}$, V_2 is ID w.r.t $\{V_3\}$[16]. Ancestral disentanglement is achievable in Ex. 4 [21]. □

Remark 5 (Ancestral disentanglement in Ex. 4). It is verifiable that $\mathcal{M}$ in Eq. 42, and $\widehat{\mathcal{M}}$ in Eq. 53 match with $\mathcal{P}$ and G^S. Similarly, we can find a transformation τ_1, τ_2, τ_3 from the set of true generative factors $\overline{Anc}(V_i)$ to $\widehat{V}_i$ such that:

$$\begin{aligned} \widehat{V}_1 &= \tau_1(V_1) = V_1 \\ \widehat{V}_2 &= \tau_2(V_2) = 2V_1 - V_2 \\ \widehat{V}_3 &= \tau_3(V_3) = 2V_1 - 2V_2 + V_3 \end{aligned} \quad . \tag{62}$$

According to the definition of identifiability (Def. 2.3), ancestral disentanglement means that, for all $\widehat{\mathcal{M}}$ matching with $\mathcal{P}$ and G^S (such as the above $\widehat{\mathcal{M}}$), there exists a transformation τ_i that maps $\overline{\mathbf{Anc}}(V_i)$ to $\widehat{V}_i$. The full algebraic derivation of ancestral disentanglement is provided in Appendix Section D.4. □

[15] $\overline{\mathbf{Anc}}(V_j)$ is the set of ancestors of V_i in the diagram plus V_i itself.

[16] $\mathbf{V} \backslash \overline{\mathbf{Anc}}(V_3)$ is empty so no need to check for V_3.

The previous examples illustrate full disentanglement and a special type of disentanglement (ancestral disentanglement) when the underlying ASCMs are assumed to be Markovian and interventions are given per node. However, our work studies a more relaxed setting with the non-Markovian assumption, arbitrary input data from multiple domains, and aims for more general identification. Rather than aiming for full disentanglement, or a specific disentanglement (such as ancestral disentanglement), we allow any arbitrary combination of variables to be entangled, disentangled. This results in a collection of ID tasks given the input set of distributions, LSD, or latent factors. The following example illustrates this novel setting.

Example 5 (General disentanglement in non-Markovian ASCMs). Consider a similar setting provided in Ex. 4. The true underlying $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ describes the generation process of factors $\{V_1, V_2, V_3\}$ and the mixture $\mathbf{X} = \{X_1, X_2, X_3\}$, where each $\mathcal{M}_i$ is given by:

$$\langle \mathbf{U} = \{U_{12}, U_1, U_2, U_3\}, \mathbf{V} = \{V_1, V_2, V_3\}, \mathbf{X} = \{X_1, X_2, X_3\}, \mathcal{F} = \{f_{V_1}, f_{V_2}, f_{V_3}, f_{\mathbf{X}}\}, P_i(\mathbf{U}) \rangle, \tag{63}$$

Compared to the Markovian ASCM settings, V_1 and V_2 are now confounded by an unobserved factor U_{12} [17]. Specifically, the mechanisms of generative factors are as follows:

$$\begin{aligned} V_1 &\leftarrow f_{V_1}(U_1, U_{12}) = U_1 + U_{12} \\ V_2 &\leftarrow f_{V_2}(V_1, U_{12}, U_2) = V_1 + U_2 + U_{12} \\ V_3 &\leftarrow f_{V_3}(V_2, U_3) = V_2 + U_3 \end{aligned} \tag{64}$$

where the exogenous U_{12}, U_1, U_2, U_3 are mutually independent of each other in each domain. Also U_{12}, U_1, U_2 do not change across domains, while U_3 does change as follows:

$$P_i(\mathbf{U}) = \begin{cases} U_{12} \sim \mathcal{N}(0, 0.5) \\ U_1 \sim \mathcal{N}(0, 1) \\ U_2 \sim \mathcal{N}(0, 2) \\ U_3 \sim \mathcal{N}(0, 2 + i) \end{cases} \tag{65}$$

Similar to the previous example, $\{V_1, V_2, V_3\}$ are mapped to generate $\mathbf{X} = \{X_1, X_2, X_3\}$ by the same invertible mixing function $f_{\mathbf{X}}$ (Eq. 5). Formally, each ASCM $\mathcal{M}_i$ can be written as follows:

$$\mathcal{M}_i = \begin{cases} \mathbf{U} = \{U_{12}, U_1, U_2, U_3\} \\ \mathbf{V} = \{V_1, V_2, V_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} V_1 \leftarrow U_1 + U_{12} \\ V_2 \leftarrow V_1 + U_{12} + U_2 \\ V_3 \leftarrow V_2 + U_3 \end{cases} \\ \mathbf{X} = \begin{cases} X_1 \leftarrow V_1 + V_2 \\ X_2 \leftarrow V_1 - V_2 \\ X_3 \leftarrow V_1 + V_3 \end{cases} \\ P(\mathbf{U}) : \begin{pmatrix} U_{12} \\ U_1 \\ U_2 \\ U_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 0.5 & 0 & 0 & 0 \\ 0 & i & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix}) \end{cases} \tag{66}$$

The LSD G^S induced by $\boldsymbol{\mathcal{M}}$ is shown in Fig. 7. To illustrate, V_1 and V_2 are connected by a bidirected arrow since they are confounded by U_{12}. The S-node only points to V_3 since only U_3 is different between Π_1 and Π_2, and the other factors are assumed invariant for convenience.

[17] We slightly abuse the word "unobserved" here since all generative factors are themselves unobserved. In the context of causal representation learning, unobserved confounding means there exists a confounding causal variable among the latent generative factors that is unaccounted for.

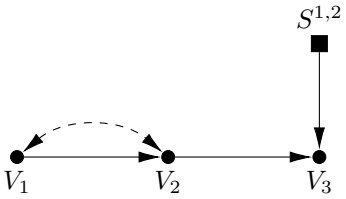


Figure 7: Example LSD for a non-Markovian ASCM (re-plot of Fig. 2).

Suppose the collection of given distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ resulting from 4 interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)}\}$ in domain $\{\Pi_1, \Pi_2, \Pi_2, \Pi_1\}$, respectively, are:

$$\begin{aligned}
\sigma^{(1)} &= \{\} \\
\sigma^{(2)} &= \{\} \\
\sigma^{(3)} &= \sigma_{V_3} : V_3 \leftarrow V_2 + U_3', U_3' \sim \mathcal{N}(0, 0.5) \\
\sigma^{(4)} &= (V_2) : V_2 \leftarrow U_2', U_2' \sim \mathcal{N}(0, 0.5)
\end{aligned} \tag{67}$$

and $U_{12}, U_1, U_2, U_2', U_3.U_3'$ are independent. Then, the interventional targets are

$$\Psi = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, V_3^{\Pi_2}, V_2^{\Pi_1, \text{do}}\} \tag{68}$$

In words, the intervention $\sigma^{(1)}$ and $\sigma^{(2)}$ are idle interventions in domains Π_1 and Π_2, $\sigma^{(3)}$ is a soft intervention on V_3 in Π_2, and $\sigma^{(4)}$ is a hard intervention on V_2 in Π_1. The interventions $\sigma^{(1)}$ and $\sigma^{(2)}$ lead to observational distributions $P^{(1)}$ and $P^{(2)}$, where

$$P^{(1)} = P^{\Pi_1}(\mathbf{X}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 10.5 & -3.5 & 10.5 \\ -3.5 & 2.5 & -3.5 \\ 10.5 & -3.5 & 13.5 \end{pmatrix}) \tag{69}$$

since

$$\begin{aligned}
Var(X_1) &= Var(V_1) + Var(V_2) + 2Cov(V_1, V_2) \\
&= 4Var(U_1) + 9Var(U_{12}) + Var(U_2) = 10.5 \\
Var(X_2) &= Var(V_1) + Var(V_2) - 2Cov(V_1, V_2) \\
&= Var(U_{12}) + Var(U_2) = 2.5 \\
Var(X_3) &= Var(V_1) + Var(V_3) + 2Cov(V_1, V_3) \\
&= 4Var(U_1) + 9Var(U_{12}) + Var(U_2) + Var(U_3) = 13.5 \\
Cov(X_1, X_2) &= Var(V_1) - Var(V_2) = -3Var(U_12) - Var(V_2) = -3.5 \\
Cov(X_1, X_3) &= Var(V_1) + Cov(V_1, V_2) + Cov(V_1, V_3) + Cov(V_2, V_3) \\
&= 4Var(U_1) + 9Var(U_{12}) + Var(U_2) = 10.5 \\
Cov(X_2, X_3) &= Var(V_1) - Cov(V_1, V_2) + Cov(V_1, V_3) - Cov(V_2, V_3) \\
&= -3Var(U_12) - Var(V_2) = -3.5
\end{aligned} \tag{70}$$

And similarly,

$$P^{(2)} = P^{\Pi_2}(\mathbf{X}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 10.5 & -3.5 & 10.5 \\ -3.5 & 2.5 & -3.5 \\ 10.5 & -3.5 & 14.5 \end{pmatrix}) \tag{71}$$

The third intervention $\sigma^{(3)}$ is a soft intervention applied to V_3 in domain Π_2 and

$$P^{(3)} = P^{\Pi_2}(\mathbf{X}; \sigma_{V_3}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 10.5 & -3.5 & 10.5 \\ -3.5 & 2.5 & -3.5 \\ 10.5 & -3.5 & 11 \end{pmatrix}) \tag{72}$$

The forth intervention $\sigma^{(4)}$ is a hard intervention applied to V_2 in domain Π_2, which means V_2 in the submodel induced by $\sigma^{(4)}$ does not rely on any endogenous variables or endogenous variables in $\boldsymbol{\mathcal{M}}$ (i.e. all incoming edges to V_2 are cut). Then, we have

$$P^{(4)} = P^{\Pi_2}(\mathbf{X}; do_{V_2}) = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 2.5 & -0.5 & 2.5 \\ -0.5 & 2.5 & -0.5 \\ 2.5 & -0.5 & 6.5 \end{pmatrix}) \tag{73}$$

Suppose one learns the proxy model $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2 \rangle$ compatible with the G^S and $\mathcal{P}$, where $\widehat{\mathcal{M}}_i$ is

$$\langle \widehat{\mathbf{U}} = \{\widehat{U}_{12}, \widehat{U}_1, \widehat{U}_2, \widehat{U}_3\}, \mathbf{V} = \{\widehat{V}_1, \widehat{V}_2, \widehat{V}_3\}, \mathbf{X} = \{X_1, X_2, X_3\}, \mathcal{F} = \{\widehat{f}_{V_1}, \widehat{f}_{V_2}, \widehat{f}_{V_3}, \widehat{f}_{\mathbf{X}}\}, P_i(\widehat{\mathbf{U}}) \rangle, \tag{74}$$

The decision of what entanglements and disentanglements are desired within the proxy model's representation may be guided, for example, by domain expertise. It will typically depend on the downstream task and intended usage of the representations $\widehat{\mathbf{V}}$. As discussed in the introduction, consider the setting where one allows entanglement between V_2 and V_3, and aims for disentanglement of V_1 from $\{V_2, V_3\}$ and the disentanglement of $\{V_2, V_3\}$ from V_1. Specifically, it is expected that there exists a mapping from $\{V_2, V_3\}$ to $\widehat{V}_2$, a mapping from $\{V_2, V_3\}$ to $\widehat{V}_3$ and a mapping from V_1 to $\widehat{V}_1$, i.e.,

$$\begin{aligned} \widehat{V}_1 &= \tau_1(V_1) \\ \widehat{V}_2 &= \tau_2(V_2, V_3) \\ \widehat{V}_3 &= \tau_3(V_2, V_3) \end{aligned} \tag{75}$$

One valid $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \widehat{\mathcal{M}}_2 \rangle$ that matches G^S and $\mathcal{P}$, and also satisfies the user-chosen disentanglement goals (Eq. 75) is given as follows:

$$\widehat{\mathcal{M}}_i = \begin{cases} \widehat{\mathbf{U}} = \{\widehat{U}_{12}, \widehat{U}_2, \widehat{U}_3\} \\ \widehat{\mathbf{V}} = \{\widehat{V}_1, \widehat{V}_2, \widehat{V}_3\}, \\ \mathbf{X} = \{X_1, X_2, X_3\} \\ \mathcal{F} = \begin{cases} \widehat{V_1} \leftarrow \sqrt{6}\widehat{U}_{12} \\ \widehat{V_2} \leftarrow \widehat{V}_1 - \dfrac{\sqrt{6}}{3}\widehat{U}_{12} + \dfrac{\sqrt{21}}{3}\widehat{U}_2 \\ \widehat{V_3} \leftarrow 2\widehat{V}_2 + \sqrt{i+2}\widehat{U}_3 \\ \mathbf{X} = \begin{cases} X_1 \leftarrow 0.5\widehat{V}_1 + \widehat{V}_2 \\ X_2 \leftarrow 0.5\widehat{V}_1 - \widehat{V}_2 \\ X_3 \leftarrow 0.5\widehat{V}_1 - \widehat{V}_2 + \widehat{V}_3 \end{cases} \end{cases} \\ P^i(\widehat{\mathbf{U}}) = \begin{pmatrix} \widehat{U}_{12} \\ \widehat{U}_2 \\ \widehat{U}_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}) \end{cases} \tag{76}$$

First, we verify that $\widehat{\boldsymbol{\mathcal{M}}}$ induces the same G^S as in Fig. 7. Since $\widehat{V}_1$ and $\widehat{V}_2$ are confounded by $\widehat{U}_{12}$ and the mechanism of V_3 is different between domain Π_1 and Π_2, and the other causal relationships are evident. Next, consider the collection of interventions $\widehat{\Sigma} = \{\widehat{\sigma}^{(1)}, \widehat{\sigma}^{(2)}, \widehat{\sigma}^{(3)}, \widehat{\sigma}^{(4)}\}$ applied to $\widehat{\boldsymbol{\mathcal{M}}}$ in domain $\{\Pi_1, \Pi_2, \Pi_2, \Pi_1\}$, respectively:

$$\begin{aligned} \widehat{\sigma}^{(1)} &= \{\} \\ \widehat{\sigma}^{(2)} &= \{\} \\ \widehat{\sigma}^{(3)} &= \sigma_{\widehat{V}_3} : \widehat{V}_3 \leftarrow 2\widehat{V}_2 + \widehat{U}_3', \widehat{U}_3' \sim \mathcal{N}(0, 0.5) \\ \widehat{\sigma}^{(4)} &= \sigma_{\widehat{V}_2} : \widehat{V}_2 \leftarrow \widehat{U}_2', \widehat{U}_2' \sim \mathcal{N}(0, 0.5) \end{aligned} \tag{77}$$

It is verifiable that

$$\widehat{P}^{(1)} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} 10.5 & -3.5 & 10.5 \\ -3.5 & 2.5 & -3.5 \\ 10.5 & -3.5 & 13.5 \end{pmatrix}) \tag{78}$$

which is equivalent to $P^{(1)}$ (Eq. 78) since

$$
\begin{aligned}
Var(X_1) &= 0.25Var(\widehat{V}_1) + Var(\widehat{V}_2) + Cov(\widehat{V}_1, \widehat{V}_2) = \frac{49}{6}Var(\widehat{U}_{12}) + \frac{7}{3}Var(\widehat{U}_2) = 10.5 \\
Var(X_2) &= 0.25Var(\widehat{V}_1) + Var(\widehat{V}_2) - Cov(\widehat{V}_1, \widehat{V}_2) = \frac{1}{6}Var(\widehat{U}_{12}) + \frac{7}{3}Var(\widehat{U}_2) = 2.5 \\
Var(X_3) &= 0.25Var(\widehat{V}_1) + Var(\widehat{V}_2) + Var(\widehat{V}_3) - Cov(\widehat{V}_1, \widehat{V}_2) + Cov(\widehat{V}_1, \widehat{V}_3) - 2Cov(\widehat{V}_2, \widehat{V}_3) \\
&= \frac{49}{6}Var(\widehat{U}_{12}) + \frac{7}{3}Var(\widehat{U}_2) + 3Var(\widehat{U}_3) = 13.5 \\
Cov(X_1, X_2) &= 0.25Var(\widehat{V}_1) - Var(\widehat{V}_2) \\
&= -\frac{7}{6}Var(\widehat{U}_{12}) - \frac{7}{3}Var(\widehat{U}_2) = -3.5 \\
Cov(X_1, X_3) &= 0.25Var(\widehat{V}_1) - Var(\widehat{V}_2) + 0.5Cov(\widehat{V}_1, \widehat{V}_3) + Cov(\widehat{V}_2, \widehat{V}_3) \\
&= \frac{49}{6}Var(\widehat{U}_{12}) + \frac{7}{3}Var(\widehat{U}_2) = 10.5 \\
Cov(X_2, X_3) &= 0.25Var(\widehat{V}_1) + Var(\widehat{V}_2) - 2Cov(V_1, V_2) + 0.5Cov(\widehat{V}_1, \widehat{V}_3) + Cov(\widehat{V}_2, \widehat{V}_3) \\
&= -\frac{7}{6}Var(\widehat{U}_{12}) - \frac{7}{3}Var(\widehat{U}_2) = -3.5
\end{aligned}
\tag{79}
$$

and similarly,

$$
\widehat{P}^{(2)} = P^{(2)}, \widehat{P}^{(3)} = P^{(3)}, \widehat{P}^{(4)} = P^{(4)} \tag{80}
$$

Now, consider the mapping from $\widehat{\mathbf{V}}$ to $\mathbf{V}$ through the mixing functions $\widehat{f}_{\mathbf{X}}$ and $f_{\mathbf{X}}$,

$$
\widehat{\mathbf{V}} = \begin{pmatrix} \widehat{V}_1 \\ \widehat{V}_2 \\ \widehat{V}_3 \end{pmatrix} = \widehat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}(\mathbf{V}) = \underbrace{\begin{pmatrix} 0.5 & 1 & 0 \\ 0.5 & -1 & 0 \\ 0.5 & -1 & 1 \end{pmatrix}^{-1}}_{\text{Learned } \widehat{\mathcal{M}} \text{ (Eq. 76)}} \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{\text{True } \mathcal{M} \text{ (Eq. 66)}} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \tag{81}
$$

$$
= \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} \tag{82}
$$

This implies $\widehat{\mathbf{V}}$ satisfies Eq. 75, i.e.,

$$
\begin{aligned}
\widehat{V}_1 &= \tau_1(V_1) = 2V_1 \\
\widehat{V}_2 &= \tau_2(V_2, V_3) = V_2 \\
\widehat{V}_3 &= \tau_3(V_2, V_3) = V_2 + V_3
\end{aligned}
\tag{83}
$$

□

The partial identifiability result from the above derivation says that for any $\widehat{\mathcal{M}}$ that matches with $\mathcal{P}$ and G^S, Eq. 75 always hold. In the context of Def. 2.3, this is a partial disentanglement goal where (1) V^{tar} is set as V_1 and V^{en} is set as $\{V_2, V_3\}$; (2) V^{tar} is set as V_2 and V^{en} is set as V_1; (3) V^{tar} is set as V_3 and V^{en} is set as V_1.

We conclude by formally stating the partial disentanglement instance considered in this example.

Task 4 (Partial disentanglement in a general setting). *Suppose the true underlying ASCMs $\mathcal{M}$ induces the LSD G^S in Fig. 7 and distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ from interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)}\} = \{\{\}, \{\}, \sigma_{V_3}, do(V_2)\}$ in domains $\{\Pi_1, \Pi_2, \Pi_2, \Pi_1\}$. Given intervention targets*

$$
\mathbf{\Psi} = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, V_3^{\Pi_2}, V_2^{\Pi_1, \text{do}}\} \tag{84}
$$

and G^S, the task is to determine whether (and how) $\{V_2, V_3\}$ is ID w.r.t. V_1, and V_1 is ID w.r.t $\{V_2, V_3\}$. □

This task is relevant since it relaxes the requirements expected from a learned representation even when not all factors are disentangled, the important factors are. This possibility may allow for for settings that are present in real data, but not solved in the previous two settings of nonlinear ICA (Ex. 2, Task 1) and Markovian ASCM disentanglement (Ex. 4, Tasks 2 and 3). Firstly, unobserved confoundering can be commonly found in datasets and neither the ICA nor Markovian settings allows for such a pervasive phenomenon. Second, the input data is not restricted as shown in Table 2. For example, the observational data is not always assumed in this general task and not all variables are intervened on. The $\mathbf{V}^{tar}$ and $\mathbf{V}^{en}$ set of variables in the general disentanglement setting modeled by Def. 2.3 is user-chosen. As such, full disentanglement or ancestral disentanglement are valid special cases, depending on the user's goal. How we solve the general disentanglement task will be answered in the next section.

To summarize, the task presented in Ex. 5 is a generalization of Ex. 2 and Ex. 4, where the goal is partial disentanglement considering an arbitrary, more relaxed combinations of distributions arising from multiple domains in a non-Markovian ASCM setting. Def. 2.3 allows us to characterize disentangled causal representations given an arbitrary combinations of distributions and assumptions. Furthermore, existing identifiability definitions that have been discussed in the literature can be seen as a special case of Def. 2.3 (as shown in Appendix E.3). The next section will develop sufficient graphical criteria that enables one to determine when partial (or full) disentanglement is possible even when confounding among latent variables is present.

3 Graphical Criterion for Causal Disentanglement

In this section, we study identifiability criteria given general assumptions and arbitrary input distributions. More specifically, we connect latent variables $\mathbf{V}$ and a representation $\widehat{\mathbf{V}}$ by comparing distributions in Sec. 3.1. Leveraging this understanding, we introduce in Sec. 3.2 three graphical criteria to evaluate the identifiability of a target representation.

3.1 Latent variable factorization and invariances

We revisit the factorization of distributions induced by non-Markovian models [3, Def. 15]. First, consider $P_{\mathbf{T}}(\mathbf{V})$ induced by an ASCM $\mathcal{M}$ after a hard intervention on $\mathbf{T}$. Then, given a topological order $<$ of G, the joint interventional distribution can be factorized as follows:

$$P_{\mathbf{T}}(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} P_{\mathbf{T}}(V_i | \mathbf{Pa}_i^{\mathbf{T}+}), \tag{85}$$

where $\mathbf{Pa}_i^{\mathbf{T}+} = \overline{\mathbf{Pa}}(\{V \in \mathbf{C}(V_i) : V \leq V_i\}) \setminus \{V_i\}$ is the extended parents set of V_i in $G_{\overline{\mathbf{T}}}$.

Unlike the standard factorization in Markovian settings, the factorization takes the extended parents ($\mathbf{Pa}_i^{\mathbf{T}+}$) as the conditioning part, not $\mathbf{Pa}_i$ in $G_{\overline{\mathbf{T}}}$, which plays a key role in the disentanglement discussions. The following example illustrates such differences.

Example 6 (Factorization and a unique topological order continued with Task 4**).** Consider a collection of ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ that induces the LSD in Example 5 (also shown in Fig. 11(a)). There is a single topological order: $V_1 < V_2 < V_3$. The observational distribution can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2) \tag{86}$$

The LSD induces a single unique topological order, and similarly, only a single factorization. □

Example 7 (Factorization and different topological orders in Markovian ASCM). Consider a collection of ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \ldots, \mathcal{M}_n \rangle$ that induces the LSD shown in Fig. 11(b). Given an order A; $V_1 < V_3 < V_2$, the observational distribution can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1, V_3)P(V_3) \tag{87}$$

Choosing another order B: $V_3 < V_1 < V_2$, $P(\mathbf{V})$ can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1, V_3)P(V_3) \tag{88}$$

The conditioning part of V_2 is consistent in both factorizations since $\mathbf{Pa}^{+}(V_2) = \{V_1, V_3\}$ and the c-components of every variable is the singleton $\mathbf{C}(V_i) = \{V_i\}$ for all $V_i \in \mathbf{V}$. □

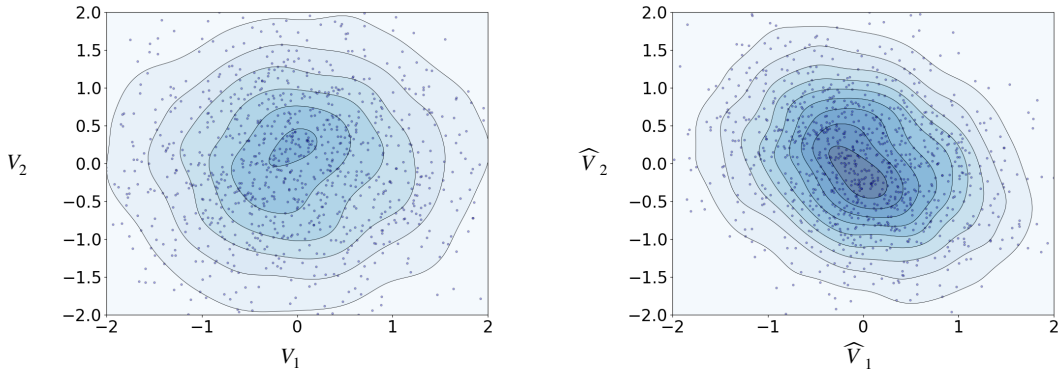


Figure 8: Illustration of change-of-variable formula in Ex. 9: density function $p(v_1, v_2)$ (left) and $p(\widehat{v}_1, \widehat{v}_2)$ (right). The intensity of the blue shade represents the density.

Example 8 (Factorization and different topological orders in non-Markovian ASCMs). Consider a collection of ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \dots, \mathcal{M}_n \rangle$ that induces the LSD shown in Fig. 11(c). Given an order A; $V_1 < V_2 < V_3 < V_4$, the observational distribution can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2, V_1)P(V_4 \mid V_3) \tag{89}$$

Notice that the conditioning part of V_3 includes $\{V_2, V_1\}$ even though they are not parents of V_3 in a typical sense. Choosing another order B: $V_1 < V_3 < V_2 < V_4$, $P(\mathbf{V})$ can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1, V_3)P(V_3)P(V_4 \mid V_3) \tag{90}$$

The conditioning parts of V_2 and V_3 are different from Eq. 89. Note that if the bidirected arrow $V_2 \dashleftarrow\!\dashrightarrow V_3$ is replaced with a directed arrow $V_2 \to V_3$, the model would be Markovian, and the observational distribution would factorize as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2)P(V_4 \mid V_3) \tag{91}$$

which is clearly different from both Eqs. 89 and 90. □

It is worth noting at this point that the factorization implied by Eq. 85 is unique in the Markovian setting. That is, given any topological ordering of the causal graph, the factors relative to the same families are involved. This occurs in part because $\mathbf{Pa}_i^{T+}$ is simply the set of nodes that have a directed edge into V_i. In the non-Markovian setting, the $\mathbf{Pa}_i^{T+}$ set is dependent on the chosen topological order, which results in possibly different factorizations. This non-uniqueness will play a fundamental role in disentanglement in the non-Markovian setting as we will see in the following sections.

Armed with this factorization, the representation $\widehat{V}$ in $\widehat{\boldsymbol{\mathcal{M}}}$ and the true underlying variables $\mathbf{V}$ in $\boldsymbol{\mathcal{M}}$ can be related via the change-of-variable formula with $\widehat{\mathbf{V}} = \boldsymbol{\phi}(\mathbf{V})$:

$$p(\mathbf{V}) = p(\boldsymbol{\phi}(\mathbf{V}))|\det J_{\boldsymbol{\phi}}| = p(\widehat{\mathbf{V}})|\det J_{\boldsymbol{\phi}}| \tag{92}$$

where $\boldsymbol{\phi} = \widehat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}$ and $\mathbf{J}_{\boldsymbol{\phi}}$ is the Jacobian matrix of $\boldsymbol{\phi}$.

Example 9 (Change-of-variable between latent variables and learned representations). Consider the following ASCM $\mathcal{M}$:

$$\mathcal{M} = \begin{cases} \mathbf{U} = \{U_1, U_2\} \\ \mathbf{V} = \{V_1, V_2\}, \\ \mathcal{F} = \begin{cases} V_1 \leftarrow U_1 \\ V_2 \leftarrow U_2 \\ \mathbf{X} = f_{\mathbf{X}}(V_1, V_2) \end{cases} \\ P(\mathbf{U}) : \begin{pmatrix} U_1 \sim \mathcal{N}(0,1) \\ U_2 \sim \mathcal{N}(0,1) \end{pmatrix} \end{cases} \tag{93}$$

The random latent variables V_1, V_2 are represented in the observations $\mathbf{X} = f_{\mathbf{X}}(V_1, V_2)$; the form of f_X is not important for the discussion of this example.

Consider a proxy ASCM $\widehat{\mathcal{M}}$ compatible with the distribution of $\mathbf{X}$ and G^S. The observed high-dimensional $\mathbf{X}$ can be constructed from the true underlying factors $\mathbf{V}$ in the true $\mathcal{M}$ and from representations $\widehat{\mathbf{V}}$ in the proxy $\widehat{\mathcal{M}}$, i.e.,

$$\mathbf{X} = f_{\mathbf{X}}(\mathbf{V}) = \widehat{f}_{\mathbf{X}}(\widehat{\mathbf{V}}) \tag{94}$$

Due to the invertibility of the mixing functions, there always exists a mapping $\boldsymbol{\phi} : \mathbf{V} \to \widehat{\mathbf{V}} = f_X \circ \widehat{f_X^{-1}}$ from the true factors $\mathbf{V}$ to representations $\widehat{\mathbf{V}}$, namely,

$$\widehat{\mathbf{V}} = \begin{pmatrix} \widehat{V}_1 \\ \widehat{V}_2 \end{pmatrix} = \boldsymbol{\phi}(\mathbf{V}) = \begin{pmatrix} \phi_1(V_1, V_2) \\ \phi_2(V_1, V_2) \end{pmatrix} \tag{95}$$

Consider the following specific mapping

$$\begin{aligned} \widehat{V}_1 &= \phi_1(V_1, V_2) = V_1, \\ \widehat{V}_2 &= \phi_2(V_1, V_2) = -0.2V_1 - 0.8V_2 \end{aligned} \tag{96}$$

Then, we can calculate the absolute value of the determinant of the Jacobian as

$$|\det J_{\boldsymbol{\phi}}| = \det \begin{vmatrix} \frac{\partial \widehat{V_1}}{\partial V_1} & \frac{\partial \widehat{V_1}}{\partial V_2} \\ \frac{\partial \widehat{V_2}}{\partial V_1} & \frac{\partial \widehat{V_2}}{\partial V_2} \end{vmatrix} = \det \begin{vmatrix} 1 & 0 \\ -0.2 & -0.8 \end{vmatrix} = -0.8. \tag{97}$$

To illustrate, $|\det J_{\boldsymbol{\phi}}|$ here determines the volume change when going from the space of (V_1, V_2) to $(\widehat{V_1}, \widehat{V_2})$.

Next, we can compare the representations $\widehat{\mathbf{V}} = \{\widehat{V_1}, \widehat{V_2}\}$ with the latent variables $\mathbf{V} = \{V_1, V_2\}$.The distribution of the latent variables and the new representation $\widehat{\mathbf{V}}$ is shown in Fig. 8. It is clear that the joint distributions $P(V_1, V_2)$ are different from $P(\widehat{V_1}, \widehat{V_2})$. However, each distribution is exactly mappable to the other via f_X, or $\widehat{f_X^{-1}}$, and thus $\widehat{\mathbf{V}}$ may serve as a representation of our latent space. In the context of disentanglement, since the mapping ϕ_1 are mapped from only V_1, then there exists

$$\widehat{V}_1 = \tau_1(V_1) = \phi_1(V_1, v_2) = V_1, \tag{98}$$

where v_2 can be an arbitrary value. This implies $\widehat{\mathcal{M}}$ satisfies the identification requirement in the identification definition (Def. 2.3). In contrast, since $\widehat{V}_2$ is mapped from V_1 and V_2 together, these does not exist $\widehat{V}_1 = \tau_2(V_2)$. Thus, V_2 is not ID (V_2 is entangled with V_1). □

Since f_X is assumed to be deterministic and shared across domains (Def. 2.2), ϕ is invariant to interventions and changes in domains in the latent space $\mathbf{V}$. Specifically, the shared ϕ indicates that the way generative factors are mixed to create high-dimensional observations, $\mathbf{X}$, are the same across different interventional regimes and domains. For example, consider photographers taking photos to produce images: objects and scenes within the image are the observed low-level latent generative factors ($\mathbf{V}$), and the image comprised of pixels is the high-level feature ($\mathbf{X}$) generated from an unknown mixing function f_X. The mixing function can be thought of as the process in which a camera and photographer setup turns light into an image. The generative process to construct the pixels is invariant even if the generative factors change due to interventions or domain changes.

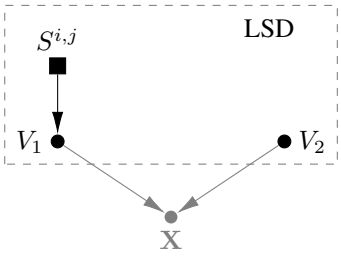


Figure 9: The G^S induced by $\mathcal{M}$ in Example 10.

The relationship between the true latent $\mathbf{V}$ and the representation $\widehat{\mathbf{V}}$ is complicated by the Jacobian term. It is non-trivial to compute this term if the latent distribution is unobserved, but if the value is known, then one can relate a representation with the true latent variables using Eq. 92. Prior work

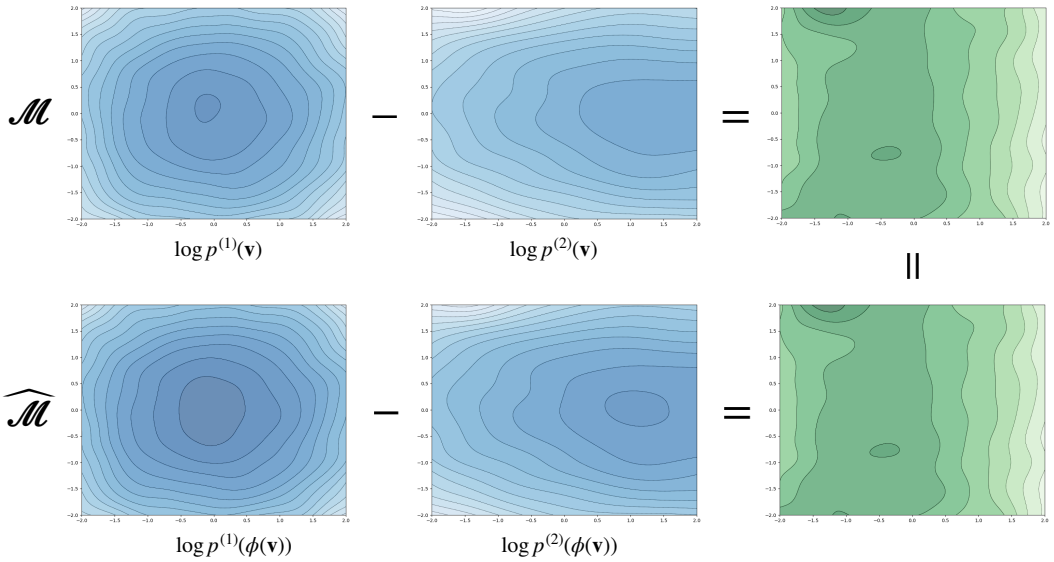


Figure 10: Illustration of log densities of different distributions (Eq. 100) in Ex. 10. The top row shows the difference of densities induced by the true model $\mathcal{M}$, and the bottom row shows the difference of densities induced by the proxy model $\widehat{\mathcal{M}}$.

has described various strategies for handling the Jacobian term. One such approach assumes that the mixing function is volume-preserving, which makes $|\det J_{\phi}| = 1$ [74]. However, this is a rather restrictive assumption.

Another method commonly taken in nonlinear ICA is to compare distributions, where the $|\det J_{\phi}|$ ends up being canceled out. To illustrate, suppose two distributions $P^{(1)}, P^{(2)}$ arising from different domains or interventions are given in the input. Applying the change-of-variable formula (Eq. 92) to both density function $p^{(1)}(\mathbf{v})$ and $p^{(2)}(\mathbf{v})$, we have

$$\begin{aligned} \log p^{(1)}(\mathbf{v}) - \log p^{(2)}(\mathbf{v}) &= \log p^{(1)}(\phi(\mathbf{v})) + \log|\det J_{\phi}| \\ &\quad - \log p^{(2)}(\phi(\mathbf{v})) - \log|\det J_{\phi}| \qquad (99) \\ \Rightarrow \log p^{(1)}(\mathbf{v}) - \log p^{(2)}(\mathbf{v}) &= \log p^{(1)}(\phi(\mathbf{v})) - \log p^{(2)}(\phi(\mathbf{v})) \qquad (100) \end{aligned}$$

See the following example for explaining the cancelation of the Jacobian matrix.

Example 10 (Comparing distributions and the change-of-variable formula). Consider the ASCM in Example 9 (Eq. 93). Suppose V_1 is perturbed due to the domain change, and the resulting ASCM $\mathcal{M}'$ is shown as follows:

$$\mathcal{M}' = \begin{cases} \mathbf{U} = \{U_1, U_2\} \\ \mathbf{V} = \{V_1, V_2\}, \\ \mathcal{F} = \begin{cases} V_1 \leftarrow U_1 \\ V_2 \leftarrow U_2 \\ \mathbf{X} = f_{\mathbf{X}}(V_1, V_2) \end{cases} \\ P(\mathbf{U}) : \begin{pmatrix} U_1 \sim \mathcal{N}(1.5, 2) \\ U_2 \sim \mathcal{N}(0, 1) \end{pmatrix} \end{cases} \qquad (101)$$

Then, ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}, \mathcal{M}' \rangle$ induces LSD G^S in Fig. 9 and distributions $\mathcal{P} = \{P^{(1)} = P^{\Pi_1}(\mathbf{X}), P^{(2)} = P^{\Pi_2}(\mathbf{X})\}$. Further, consider a proxy model $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}, \widehat{\mathcal{M}}' \rangle$ is learned that matches $\mathcal{P}$ and G^S, where $\widehat{\mathcal{M}}$ is the proxy model introduced in Example 9. Then, the log densities

can be compared via the change-of-variable formula both in the true model and the proxy model:

$$\begin{aligned}\log p^{(1)}(v_1,v_2)-\log p^{(2)}(v_1,v_2) &= \log p^{(1)}(\phi_1(v_1,v_2),\phi_2(v_1,v_2)))+\log|\det J_{\boldsymbol{\phi}}| \\ &\quad -\log p^{(2)}(\phi_1(v_1,v_2),\phi_2(v_1,v_2))-\log|\det J_{\boldsymbol{\phi}}| \quad (102)\\ \Rightarrow \log p^{(1)}(v_1,v_2)-\log p^{(2)}(v_1,v_2) &= \log p^{(1)}(\phi_1(v_1,v_2),\phi_2(v_1,v_2))) \\ &\quad -\log p^{(2)}(\phi_1(v_1,v_2),\phi_2(v_1,v_2))) \quad (103)\end{aligned}$$

To illustrate, Eq. 103 implies that the difference of density function in the true model and the proxy model is the same. This fact is illustrated in Fig. 10. □

Note that by the subtraction of two distributions with the change-of-variable formula (Eq. 100), the distribution over $\mathbf{V}$ is related to the distribution under the transformation $\phi(\mathbf{V})$. We call this subtraction process "comparing distributions." The next proposition leverages the factorization in Eq. 85 to compare two distributions as follows.

Proposition 1 (Distribution Comparison). *Consider a collection of ASCMs $\boldsymbol{\mathcal{M}}=\langle\mathcal{M}_1,\ldots,\mathcal{M}_n\rangle$ that induces a collection of distributions $\mathcal{P}$ with interventions Σ and LSD G^S. Consider comparing two distributions $P^{\Pi^{(j)}}(\mathbf{X};\sigma^{(j)}),P^{\Pi^{(k)}}(\mathbf{X};\sigma^{(k)})\in\mathcal{P}$ with intervention targets $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$. Suppose $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$ both contain a hard intervention on $\mathbf{T}$ ($\mathbf{T}$ can be an empty set, which means no hard intervention is involved). If another collection of ASCMs, $\widehat{\boldsymbol{\mathcal{M}}}=\langle\widehat{\mathcal{M}}_1,\ldots,\widehat{\mathcal{M}}_n\rangle$, matches with distribution $\mathcal{P}$ and LSD G^S, then*

$$\underbrace{\sum_i^d \log p^{(j)}_{\mathbf{T}}(v_i\mid\mathbf{pa}_i^{\mathbf{T}+})-\log p^{(k)}_{\mathbf{T}}(v_i\mid\mathbf{pa}_i^{\mathbf{T}+})}_{\boldsymbol{\mathcal{M}}}=\underbrace{\sum_i^d \log p^{(j)}_{\mathbf{T}}(\widehat{v}_i\mid\widehat{\mathbf{pa}}_i^{\mathbf{T}+})-\log p^{(k)}_{\mathbf{T}}(\widehat{v}_i\mid\widehat{\mathbf{pa}}_i^{\mathbf{T}+})}_{\widehat{\boldsymbol{\mathcal{M}}}}, \quad (104)$$

where $p^{(j)}_{\mathbf{T}}(\cdot),p^{(k)}_{\mathbf{T}}(\cdot)$ are density functions. □

Prop. 1 shows that if $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ agree on the given distributions over observed $\mathbf{X}$ and the LSD, then unobserved $\mathbf{V}$ and $\widehat{\mathbf{V}}$ can be related via Eq. 104. More specifically, the left (right) side of Eq. 104 is the difference of $P(\mathbf{V})$ ($P(\widehat{\mathbf{V}})$) when the intervention and domain changes from $\sigma^{(k)}$ to $\sigma^{(j)}$ and $\Pi^{(k)}$ to $\Pi^{(j)}$. The jacobian matrix is canceled through this comparison process. In addition, $P(\mathbf{V})$ and $P(\widehat{\mathbf{V}})$ factorize accordingly to Eq. 85 since $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ are both compatible with G. The factorization involves decomposing the joint distributions $p(\mathbf{v})$ into components $p(v_i\mid\mathbf{pa}_i^{\mathbf{T}+})$, comprised of factors of variables and their conditioning set of variables, which are chosen based on the topological order. We will explain later (in the context of Prop. 2) that some components $p(v_i\mid\mathbf{pa}_i^{\mathbf{T}+})$ are invariant with the index (j) and (k), i.e.,

$$p^{(j)}(v_i\mid\mathbf{pa}_i^{\mathbf{T}+})=p^{(k)}(v_i\mid\mathbf{pa}_i^{\mathbf{T}+}), \quad (105)$$

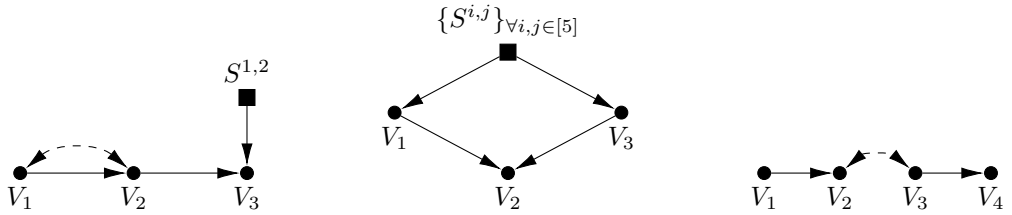
Thus, Eq. 85 can be simplified by canceling these invariant components and then an advanced connection between $\mathbf{V}$ and $\widehat{\mathbf{V}}$, which is helpful for the disentanglement.

The following examples illustrate the result of Prop. 1 and the comparison of distributions with different topological orders.

Example 11 (Task 4 and Example 6 continued). Consider $\widehat{\boldsymbol{\mathcal{M}}}$ that agrees on $\mathcal{P}$ and G^S shown in Fig. 11(a). Compare $P^{(2)}$ and $P^{(1)}$ following Prop. 1 under $\mathbf{T}=\{\}$:

$$\begin{aligned}&\log p^{(2)}(v_1)-\log p^{(1)}(v_1)+\log p^{(2)}(v_2\mid v_1)-\log p^{(1)}(v_2\mid v_1)+\log p^{(2)}(v_3\mid v_2)-\log p^{(1)}(v_3\mid v_2)\\ &=\log p^{(2)}(\widehat{v}_1)-\log p^{(1)}(\widehat{v}_1)+\log p^{(2)}(\widehat{v}_2\mid\widehat{v}_1)-\log p^{(1)}(\widehat{v}_2\mid\widehat{v}_1)+\log p^{(2)}(\widehat{v}_3\mid\widehat{v}_2)-\log p^{(1)}(\widehat{v}_3\mid\widehat{v}_2).\end{aligned} \quad (106)$$

In words, the difference $\log p^{(2)}(\mathbf{v})-\log p^{(1)}(\mathbf{v})$ is equal to $\log p^{(2)}(\widehat{\mathbf{v}})-\log p^{(1)}(\widehat{\mathbf{v}})$ since $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ agree on $\mathcal{P}$. Also, $\log p(\mathbf{v})$ and $\log p(\widehat{\mathbf{v}})$ can be both factorized as in Eq. 86 since $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ agree on G^S. We will show that some components such as $p(v_1)$ and $p(\widehat{v}_1)$ can be canceled further from the above equation in Example 15. Then Eq. 106 after cancelation will be used directly for disentanglement in Example 20. □



(a) (b) (c)

Figure 11: LSDs in Examples. (a) the re-plot of Fig 2, (b) collider, and (c) non-markovian graph.

Example 12 (Example 8 continued with multiple topological orders). Consider the pair $\mathcal{M}$ and $\widehat{\mathcal{M}}$ that induces the diagram in Fig. 11(c) and agrees on two distributions $P^{(1)}, P^{(2)} \in \mathcal{P}$ with intervention targets $\mathbf{I}^{(1)} = \{\}^{\Pi_1}$ and $\mathbf{I}^{(2)} = \{V_2^{\Pi_1}\}$. According to Prop. 1 and Eq. 89,

$$\begin{aligned}&\log p^{(2)-(1)}(v_1) + \log p^{(2)-(1)}(v_2 \mid v_1) + \log p^{(2)-(1)}(v_3 \mid v_1, v_2) + \log p^{(2)-(1)}(v_4 \mid v_3)\\ &= \log p^{(2)-(1)}(\widehat{v}_1) + \log p^{(2)-(1)}(\widehat{v}_2 \mid \widehat{v}_1) + \log p^{(2)-(1)}(\widehat{v}_3 \mid \widehat{v}_1, \widehat{v}_2) + \log p^{(2)-(1)}(\widehat{v}_4 \mid \widehat{v}_3)\end{aligned} \tag{107}$$

where $\log p^{(2)-(1)}(\cdot)$ denotes the difference $\log p^{(2)}(\cdot) - \log p^{(1)}(\cdot)$. $\mathbf{V}$ and $\widehat{\mathbf{V}}$ are related via the factors $p(v_1), p(v_2 \mid v_1), p(v_3 \mid v_1, v_2)$, and $p(v_4 \mid v_3)$. With another order and factorization of Eq. 90, a different comparison could be written:

$$\begin{aligned}&p^{(2)-(1)}(v_1) + p^{(2)-(1)}(v_2 \mid v_1, v_3) + p^{(2)-(1)}(v_3) + p^{(2)-(1)}(v_4 \mid v_3)\\ &= p^{(2)-(1)}(\widehat{v}_1) + p^{(2)-(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) + p^{(2)-(1)}(\widehat{v}_3) + p^{(2)-(1)}(\widehat{v}_4 \mid \widehat{v}_3)\end{aligned} \tag{108}$$

Here, $\mathbf{V}$ and $\widehat{\mathbf{V}}$ are related via the factors $p(v_1), p(v_2 \mid v_1, v_3), p(v_3)$, and $p(v_4 \mid v_3)$. □

From this example, we observe that the topological order may determine the factorization of a distribution. Not all factors are necessarily involved in Eq. 104 when comparing distributions. For example, in the Markovian setting, only one factor $p_{\mathbf{T}}(v_i \mid \mathbf{pa}_i)$ will possibly change when comparing the observational to a singleton interventional distribution in the same domain, while other invariant factors may remain invariant, and therefore canceled out upon subtraction. The following result leverages the invariant factors in $\mathbf{V}$ and $\widehat{\mathbf{V}}$ when comparing distributions from different domains and interventions in non-Markovian settings.

Proposition 2 (Invariant Factors). *Consider two distributions $P^{(j)}, P^{(k)} \in \mathcal{P}$ with intervention targets $\sigma^{(j)}$ and $\sigma^{(k)}$ containing $do(\mathbf{T})$. Construct the changed variable set $\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$ (for short $\Delta\mathbf{V}^{(j),(k)}$ or $\Delta\mathbf{V}$ if index clear from the context) with target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ as follows. Add a variable V_l to $\Delta\mathbf{V}$ whenever one of the conditions is satisfied:*

1. *(Interventions) $V_l \in \Delta\mathbf{V}$ if $V_l^{\pi_l,\{b_l\},t_l} \in \mathbf{I}^{(j)}$ but $V_l^{\pi'_l,\{b_l\},t'_l} \notin \mathbf{I}^{(k)}$, and vice versa;*
2. *(Domains) $V_l \in \Delta\mathbf{V}$ if (i) $S^{\Pi^{(j)},\Pi^{(k)}}$ point to V_l, (ii) $V_l^{\pi_l,\{b_l\},t_l} \notin \mathbf{I}^{(j)}$, (iii) $V_l^{\pi_l,\{b_l\},t_l} \notin \mathbf{I}^{(j)}$.*

If $V_i \in \mathbf{V}\backslash\mathbf{C}(\Delta\mathbf{V})$, then

$$p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}), \text{ and } p_{\mathbf{T}}^{(j)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+})$$

which will be denoted as invariant factors, where $\mathbf{C}(\Delta\mathbf{V}) = \bigcup_{Z\in\Delta\mathbf{V}}\{V \in \mathbf{C}(Z)\}$. The factors that are not invariant will be termed non-invariant. □

Prop. 2 states that factors $p_{\mathbf{T}}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+})$ are guaranteed to be invariant if V_i is not in $\mathbf{C}(\Delta\mathbf{V})$, which is the set of variables in the same c-component as $\Delta\mathbf{V}$. The changed variable set, $\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$, contains the variables that are intervened differently in $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ as well as the variables pointed by S-node, $S^{j,k}$ [18].

[18] Notice that a change in the interventional mechanism will dominate the domain changes, which means when the intervened mechanism of V_l is the same between $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$, the discrepancy of V_l due to the change of domain between $\Pi^{(j)}$ and $\Pi^{(k)}$ will be canceled. See Appendix A.3 for an example.

(a) (b)

Figure 12: Example graph used in Ex. 13 (a), and in Ex. 14 (b).

Non-invariant factors can change when comparing two distributions, accounting for any possible topological order from G. The following two examples demonstrate how invariant factors arise in Markovian and non-Markovian ASCMs.

Example 13 (Invariant Factors in Markovian ASCM). Consider the LSD, G^S in Fig. 12(a). There are two possible topological orders compatible with G^S: i) $V_2 < V_1 < V_3$, or ii) $V_2 < V_3 < V_1$.

Given two distribution $P^{(1)}, P^{(2)}$ according to $\Sigma = \{\{\}, \sigma_{V_3}\}$, we observe that $P^{(1)} - P^{(2)}$ results in the same factorization for different orders:

$$P(\mathbf{V}) = P(V_3|V_2)P(V_1|V_2)P(V_2) \tag{109}$$

More specifically, we observe that the comparison of the two distributions yields:

$$\log p(\mathbf{v}) - \log p(\mathbf{v}; \sigma_{V_3}) = \log p^{(1)}(v_3|v_2) - \log p^{(2)}(v_3|v_2) \tag{110}$$

According to Prop. 2, $\Delta\mathbf{V}$ can be constructed as follows:

1. $V_3 \in \Delta\mathbf{V}$ because $V_3 \in \mathbf{I}^{(2)}$ and $V_3 \notin \mathbf{I}^{(1)}$, where $\mathbf{I}^{(1)}, \mathbf{I}^{(2)} = \{\{\}, \sigma_{v_3}\}$, and
2. There are no S-nodes in LSD G^S in Fig. 12(a).

Then, each factor $p(v_i \mid \mathbf{pa}^{\mathbf{T}+})$ is invariant if

$$V_i \in \mathbf{V}\backslash\mathbf{C}(\Delta\mathbf{V}) = \mathbf{V}\backslash\mathbf{C}(V_3) = \{V_1, V_2\}, \tag{111}$$

Thus, the invariant factors are $p^{(t)}(v_1|v_2), p^{(t)}(v_2)$ for $t = \{1, 2\}$. □

Example 14 (Invariant Factors in non-Markovian ASCM). Consider the LCG, G^S in Fig. 12(b), where V_2 is marginalized out and now "unobserved". There are two possible topological orders: i) $V_1 < V_3$, or ii) $V_3 < V_1$. Given two distributions $P^{(1)}, P^{(2)}$ according to $\Sigma = \{\{\}, \sigma_{V_3}\}$, we observe that $P^{(1)} - P^{(2)}$ results in different factorizations:

$$\begin{aligned} P(\mathbf{V}) &= P(V_3|V_1)P(V_1) \\ &= P(V_1|V_3)P(V_3) \end{aligned} \tag{112}$$

Given two distributions $P^{(1)}, P^{(2)}$ resulting from $\Sigma = \{\{\}, \sigma_{V_3}\}$, we can compare $P^{(1)}$ and $P^{(2)}$ considering the first factorization of $P(\mathbf{V})$, $P(V_3|V_1)P(V_1)$:

$$\log p^{(2)}(v_3|v_1; \sigma_{V_3}) - \log p^{(1)}(v_3|v_1; \sigma_{V_3}) = \log p^{(2)}(\widehat{v}_3|\widehat{v}_1; \sigma_{\widehat{V}_3}) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_1; \sigma_{\widehat{V}_3}) \tag{113}$$

When comparing the two distributions with another factorization of $P(\mathbf{V})$ $P(V_1|V_3)P(V_3)$, the difference between the two distributions can be written as:

$$\begin{aligned} &\log p^{(2)}(\mathbf{v}; \sigma_{V_3}) - \log p^{(1)}(\mathbf{v}) \\ = \log p^{(2)}(v_3; \sigma_{V_3}) - \log p^{(1)}(v_3; \sigma_{V_3}) &+ \log p^{(2)}(v_1|v_3) - \log p^{(1)}(v_1|v_3; \sigma_{V_3}) \end{aligned} \tag{114}$$

That is, we cannot say that $p^{(1)}(v_1|v_3) = p^{(2)}(v_1|v_3; \sigma_{v_3})$ because conditioning on V_3 makes $P(V_1 \mid V_3)$ dependent on the intervention on V_3. Leveraging the σ-calculus presented in [58], one may observe that none of the rules can be applied and $V_1 \perp\!\!\!\perp V_3$ cannot be obtained. More formally, explicitly annotating the independent exogenous variables $\mathbf{U} = \{U_1, U_{13}, U_3\}$ where

$$\begin{aligned} V_1 &\leftarrow f_{V_1}(U_1, U_{13}), \\ V_3 &\leftarrow f_{V_3}(U_3, U_{13}), \end{aligned} \tag{115}$$

we have that:

$$P(V_1 \mid V_3) = \frac{\sum_{u_1,u_{13},u_3} P(V_1 \mid u_1, u_{13})P(V_3 \mid u_{13}, u_3)P(u_1)P(u_{13})P(u_3)}{\sum_{u_{13},u_3} P(V_3 \mid u_{13}, u_3)P(u_{13})P(u_3)} \tag{116}$$

for the observational distribution. After applying the intervention σ_{V_3}, the mechanisms are replaced as

$$\begin{aligned} V_1 &\leftarrow f_{V_1}(U_1, U_{13}), \\ V_3 &\leftarrow f'_{V_3}(U_{13}, U'_3) \end{aligned} \tag{117}$$

in $\mathcal{M}_{\sigma_{V_3}}$. Then the interventional distribution can be expressed as

$$P(V_1 \mid V_3; \sigma_{V_3}) = \frac{\sum_{u_1,u_{13},u'_3} P(V_1 \mid u_1, u_{13})P(V_3 \mid u_{13}, u'_3; \sigma_{V_3})P(u_1)P(u_{13})P(u'_3)}{\sum_{u_{13},u'_3} P(V_3 \mid u_{13}, u'_3; \sigma_{V_3})P(u_{13})P(u'_3)} \tag{118}$$

where $P(V_3 \mid u_{13}, u'_3; \sigma_{V_3})$ is the probability resulting from the mechanism $f'(U_{13}, U'_3)$, not originally in the ASCMs $\boldsymbol{\mathcal{M}}$.

Written in this form, it is clear that $p^{(1)}(v_1|v_3) \neq p^{(2)}(v_1|v_3; \sigma_{V_3})$ since $\{P(V_3 \mid u_{13}, u_3), P(u_3)\}$ and $\{P(V_3 \mid u_{13}, u'_3; \sigma_{V_3}), P(u'_3)\}$ can be arbitrarily different. According to Prop. 2, $p(v_i \mid \mathbf{pa}^{\mathbf{T}+})$ are invariant if $V_i \in \mathbf{V} \backslash \mathbf{C}(\Delta\mathbf{V}) = \mathbf{V} \backslash \mathbf{C}(V_3) = \{\}$, and since there are no invariant factors for all orders. □

We also illustrate the notion of invariant factors in the context of the previous example.

Example 15 (Task 4 and 11 continued). Consider the diagram in Fig. 11(a). The changed variable set $\Delta\mathbf{V}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, G^S] = \{V_3\}$ since the S-node points to V_3 in G^S and $\Delta\mathbf{V}[\mathbf{I}^{(3)}, \mathbf{I}^{(1)}, G^S] = \{V_3\}$ since $V_3 \in \mathbf{I}^{(3)}$, and $V_3 \notin \mathbf{I}^{(1)}$. Since V_3 is not bidirected connected to other, $\mathbf{V} \backslash \mathbf{C}(V_3) = \mathbf{V} \backslash \{V_3\} = \{V_1, V_2\}$. Thus, comparing $P^{(2)}$ and $P^{(3)}$ with the baseline $P^{(1)}$, $p(v_2 \mid v_1)$ and $p(v_1)$ are invariant factors while the factor $p(v_3 \mid v_2)$ changes, or is non-invariant. □

Example 16 (Example 8 and 12 continued). Consider the diagram in Fig. 11(c) and two distributions $P^{(1)}, P^{(2)} \in \mathcal{P}$ with intervention targets $\mathbf{I}^{(1)} = \{\}^{\Pi_1}$, and $\mathbf{I}^{(2)} = \{V_2^{\Pi_1}\}$. The changed variable set $\Delta\mathbf{V}^{(2),(1)} = \{V_2, V_3\}$ since $V_2 \in \mathbf{I}^{(2)}$, $V_2 \notin \mathbf{I}^{(1)}$. The invariant factors by Prop. 1 are $p(v_4 \mid v_2)$ and $p(v_1)$. □

With Prop. 2, Eq. 104 contains only the factors in $\mathbf{C}(\Delta\mathbf{V})$, i.e.,

$$\sum_{V_i \in \tilde{\mathbf{V}}} \log p^{(j)}_{\mathbf{T}}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - \log p^{(k)}_{\mathbf{T}}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = \sum_{V_i \in \tilde{\mathbf{V}}} \log p^{(j)}_{\mathbf{T}}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p^{(k)}_{\mathbf{T}}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) \tag{119}$$

where $\tilde{\mathbf{V}} = \mathbf{C}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S])$. For example, according to Ex. 15, Eq. 106 can be simplified to:

$$p^{(2)}(v_3 \mid v_2) - p^{(1)}(v_3 \mid v_2) = p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) \tag{120}$$

Similarly, according to Ex. 12, Eq. 107 and 108 are simplified to, respectively:

$$p^{(2)-(1)}(v_2 \mid v_1) + p^{(2)-(1)}(v_3 \mid v_1, v_2) = p^{(2)-(1)}(\widehat{v}_2 \mid v_1) + p^{(2)-(1)}(\widehat{v}_3 \mid \widehat{v}_1, \widehat{v}_2) \tag{121}$$

$$p^{(2)-(1)}(v_2 \mid v_1, v_3) = p^{(2)-(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) \tag{122}$$

These equality constraints following from the non-Markovian factorization show that the learned representation $\widehat{\mathbf{V}}$ (r.h.s of Eq. 119, or 120, or 121) is a function of variables that appear on the l.h.s.

Remark 6. Comparing Ex. 13 and Ex. 14, note that the presence of an unobserved confounder $V_1 \dashleftarrow\dashrightarrow V_3$ induces some degree of freedom in how the factorization is carried out. In the Markovian setting, even with different topological orders, the conditioning set is consistent (i.e. the parents set pa_i is stable). In the non-Markovian setting, the conditioning set may change with different topological orders, since the order in which one considers variables in the same c-component may change. Thus, considering one topological order is not sufficient to characterize invariant factors. After applying Prop. 1, the remaining variables in Eq. 119 are those in $\tilde{\mathbf{V}}$, as well as their extended parents.

The next definition formalizes the possible change variables between r.h.s and l.h.s of Eq. 119.

Definition 3.1 (**$\Delta\mathbf{Q}$ Set**). Given two distributions $P^{(j)}, P^{(k)}$ with interventions targets $\sigma^{(j)}$ and $\sigma^{(k)}$ containing $do(\mathbf{T})$, the $\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S]$ set (for short: $\Delta\mathbf{Q}^{(j),(k)}$, or $\Delta\mathbf{Q}$ if the index ic clear from the context) of the target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ is the remaining variables after comparison (i.e. Eq. 119),

$$\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S] = \tilde{\mathbf{V}} \cup \mathbf{Pa}^{\mathbf{T}+}(\tilde{\mathbf{V}}), \tag{123}$$

where $\tilde{\mathbf{V}} = \mathbf{C}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S])$. □

The $\Delta\mathbf{Q}$ set encompasses all variables in the l.h.s of Eq. 119, including $\tilde{\mathbf{V}}$ and its extended parents. The set $\mathbf{V}\backslash\Delta\mathbf{Q}$ is called *canceled variables* since the invariant factors are canceled from the comparison. comment for camera ready Continuing Example 15, $\Delta\mathbf{Q} = \{V_3, V_2\}$ since the changed factors is $p(v_3 \mid v_2)$ and invariant factors are canceled out. Continuing Ex. 16, $\Delta\mathbf{Q} = \{V_1, V_2, V_3\}$ given any of the topological orders [19].

The next examples illustrate how this factorization and $\Delta\mathbf{Q}$-set plays a role in disentangling latent variable representations.

Example 17 (**Observational data in two homogenous domains**). Consider the LSD over two domains Π_1, Π_2 shown in Fig. 11(c), where there is no S-node edge and two distributions $P^{(1)}, P^{(2)} \in \mathcal{P}$ with intervention targets $\mathbf{I}^{(1)} = \{\}^{\Pi_1}$ and $\mathbf{I}^{(2)} = \{\}^{\Pi_2}$. The distributions in each domain are invariant with respect to each other. The changed variable set $\Delta\mathbf{V}^{(2),(1)} = \{\}$ since $\Delta\mathbf{V} = \{\}$, and $\mathbf{C}(\{\}) = \{\}$. Thus, comparing $P^{(2)}$ with $P^{(1)}$ (following topological order A in Ex. 8), all factors $p(v_1), p(v_2|v_1), p(v_3|v_2, v_1), p(v_4|v_3)$ are invariant across domains. □

Observational data in two homogeneous domains is similar to the setting where only observational data in one domain is given. As acknowledged in nonlinear ICA cases, ID is in general not feasible in such cases [9].

Example 18 (**Observational data in two completely heterogenous domains**). From Fig. 11(c), consider a modified LSD, G^S over two domains Π_1, Π_2 with an S-node pointing to each variable V_1, V_2, V_3, V_4, and two distributions $P^{(1)}, P^{(2)} \in \mathcal{P}$ with intervention targets $\mathbf{I}^{(1)} = \{\}^{\Pi_1}$ and $\mathbf{I}^{(2)} = \{\}^{\Pi_2}$. The changed variable set $\Delta\mathbf{V}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, G^S] = \{V_1, V_2, V_3, V_4\}$ since the S-node points to all variables in G^S. Thus $\Delta\mathbf{Q} = \{V_1, V_2, V_3, V_4\}$. Comparing $P^{(2)}$ with $P^{(1)}$, all factors are simultaneously varying, which means nothing is invariant across domains. □

Our approach to disentangled representation learning leverages comparisons of distributions. The two examples above illustrate two extremes in a spectrum, where nothing is invariant, or nothing changes across distributions. There is an intrinsic trade-off in leveraging distributions with both invariant factors and changed variable sets. In the next section, we will explore the interplay between invariant and changed factors across distributions to achieve disentangled representations.

3.2 Graphical Criteria for Disentanglement

First, to motivate our new criteria for partial disentanglement, we start by showing a novel disentanglement result in the simple setting of Ex. 5.

Example 19 (**Deriving disentanglement with three distributions**). Consider the setup in Ex. 5, and also 11 and 15, where G^S is shown in Fig. 11(a) and the distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}\}$ with intervention targets $\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{\}^{\Pi_2}, \mathbf{I}^{(3)} = \{V_3\}^{\Pi_2}$ are available. Our goal is to determine whether V_1 is ID w.r.t. $\{V_2, V_3\}$, and $\{V_2, V_3\}$ is ID w.r.t. V_1. We demonstrate that these distributions and assumptions allow for the disentanglement of $\{V_2, V_3\}$ with respect to V_1, which has not been acknowledged in the literature before.

First, let us take the difference between the three distributions, treating $P^{(1)}$ as a "baseline". The connection between $\mathbf{V}$ and $\widehat{\mathbf{V}}$ is built in the form of Eq. 120, namely,

$$\begin{aligned} p^{(2)}(v_3 \mid v_2) - p^{(1)}(v_3 \mid v_2) &= p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) \\ p^{(3)}(v_3 \mid v_2) - p^{(1)}(v_3 \mid v_2) &= p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) \end{aligned} \tag{124}$$

[19] The $\Delta\mathbf{Q}$ sets resulting from different topological order are guaranteed to be the same, as elaborated in D.3.

Next, we take the first order partial derivative w.r.t. V_3. The l.h.s will go to 0, as there is no dependency on V_3, while the r.h.s does not as each as each representation $\widehat{V}_i$ $(i = 2, 3)$ may be a function of V_j $(j = 1, 2, 3)$. After applying the multivariate chain-rule, namely:

$$\begin{aligned} 0 &= \frac{\partial \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_2} \frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_3} \frac{\partial \widehat{v}_3}{\partial v_1} \\ 0 &= \frac{\partial \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_2} \frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_3} \frac{\partial \widehat{v}_3}{\partial v_1} \end{aligned} \tag{125}$$

We note that Eq. 125 is a linear system of equations with unknowns $\frac{\partial \widehat{v}_2}{v_1}$, and $\frac{\partial \widehat{v}_3}{v_1}$. According to Assumption 6, the coefficients of matrix A is full rank, where

$$A = \begin{pmatrix} \frac{\partial \log p^{(2)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_2} & \frac{\partial \log p^{(2)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_3} \\ \frac{\partial \log p^{(3)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_2} & \frac{\partial \log p^{(3)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_3} \end{pmatrix} \tag{126}$$

Then we have

$$\frac{\partial \widehat{v}_2}{\partial v_1} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0, \tag{127}$$

and

$$\widehat{V_2} = \tau_2(V_2, V_3), \widehat{V_3} = \tau_3(V_2, V_3) \tag{128}$$

meaning $\{V_2, V_3\}$ is ID w.r.t. V_1 and the representations of V_2 and V_3 are disentangled from V_1. □

This example demonstrates that disentanglement is related to the invariances of the factorizations of the log densities, and the resulting partial derivatives. We will now formalize this connection.

Leveraging the comparisons among distributions in $\mathcal{P}$ (Eq. 119), we next develop three criteria for disentanglement within the set $\mathbf{V}$. First, we can disentangle canceled variables from $\Delta\mathbf{Q}$ set since the difference of density over representations $\widehat{\mathbf{V}}$ in the $\Delta\mathbf{Q}$ set (r.h.s of Eq. 119) is irrelevant to canceled variables (l.h.s of Eq. 119).

Proposition 3 (ID the ΔQ set w.r.t. Canceled Variables). *Consider variables $\mathbf{V}^{tar} = \{V_1^{tar}, V_2^{tar}, \ldots, V_{d'}^{tar}\} \subseteq \mathbf{V}$. If there exists a subset of $\mathcal{P}$, $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \ldots, P^{(a_L)}\} \subseteq \mathcal{P}$ with intervention target sets $\Psi_{\mathbf{T}} = \{\mathbf{I}^{(a_0)}, \mathbf{I}^{(a_1)}, \ldots, \mathbf{I}^{(a_L)}\}$ such that*

(1) All distributions contain hard intervention on $\mathbf{T}$, i.e., $\forall\, l \in [L]$, $\mathbf{T} = do[\mathbf{I}^{(a_0)}] \subseteq do[\mathbf{I}^{(a_l)}]$ [20].

(2) The union of all ΔQ sets is $\mathbf{V}^{tar}$, i.e., $\bigcup_{l\in[L]} \Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S] = \mathbf{V}^{tar}$.

(3) Each V_i^{tar} changes once, i.e., there exists $\{a'_1, \ldots, a'_{d'}\} \subseteq \{a_1, \ldots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}$, $V_i^{tar} \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_i)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, where $d' = |\mathbf{V}^{tar}|$.

then $\mathbf{V}^{tar}$ is ID w.r.t. $\mathbf{V}\backslash\mathbf{V}^{tar}$. □

Prop. 3 disentangles target variables $\mathbf{V}^{tar}$ constructed by $\Delta\mathbf{Q}$ sets from canceled variables according to Eq. 119. To illustrate, it considers a collection of L distribution $\{P^{(a_1)}, \ldots, P^{(a_L)}\}$ to compare with the baseline $P^{(a_0)}$ such that (1) the hard intervention variables set of $\{\mathbf{I}^{(a_1)}, \ldots, \mathbf{I}^{(a_L)}\}$ contains the hard intervention variables set of the baseline $\mathbf{I}^{(a_0)}$, (2) the union of $\Delta\mathbf{Q}$ induced by comparison is equivalent to $\mathbf{V}^{tar}$, and (3) each V_i^{tar} can be covered by different $\Delta\mathbf{Q}$s. Then, if such a collection exists, then $\mathbf{V}^{tar}$ can be ID w.r.t. $\mathbf{V}\backslash\mathbf{V}^{tar}$.

Example 20 (Task 4 and Ex. 15 continued). Recall the model in Fig. 11(a), and consider $\mathbf{V}^{tar} = \{V_2, V_3\}$ and $\mathbf{V}\backslash\mathbf{V}^{tar} = \{V_1\}$. After comparing $\{P^{(2)}, P^{(3)}\}$ with the baseline $P^{(1)}$ choosing $\mathbf{T} = do[\mathbf{I}^{(1)}] = \{\}$,

$$\Delta\mathbf{Q}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, \mathbf{T}, G^S] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(1)}, \mathbf{T}, G^S] = \{V_2, V_3\} \tag{129}$$

Then, we check the conditions of Prop. 3 one by one.

[20] Recall we use the notation $do[\mathbf{I}]$ to denote that all variables that perfectly interventions on in $\mathbf{I}$.

(1) $\{P^{(2)}, P^{(3)}\}$ satisfies the first condition since the hard intervention variable set is an empty set.

(2) The $\Delta\mathbf{Q}$ sets from comparisons are equal to $\mathbf{V}^{tar}$, namely, $\Delta\mathbf{Q}^{(2),(1)} = \Delta\mathbf{Q}^{(3),(1)} = \mathbf{V}^{tar}$. Thus the second condition is naturally satisfied.

(3) Variables in $\mathbf{V}^{tar}$ are covered by different $\Delta\mathbf{Q}$ sets, namely, $V_2 \in \Delta\mathbf{Q}^{(2),(1)}$ and $V_3 \in \Delta\mathbf{Q}^{(3),(1)}$. Thus the third condition is satisfied.

By Prop. 3, $\mathbf{V}^{tar}$ is ID w.r.t. $\{V_1\}$. This demonstrates that a variable V_2 can be disentangled from another variable in the C-component (V_1). The derivation details are provided in Ex. 19. □

Example 21 (Ex. 16 continued). Suppose $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ with intervention targets

$$\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{V_2\}^{\Pi_1}, \mathbf{I}^{(3)} = \{V_3\}^{\Pi_1}, \mathbf{I}^{(4)} = \{V_1\}^{\Pi_1} \tag{130}$$

Consider $\mathbf{V}^{tar} = \{V_1, V_2, V_3\}$, and comparing $\{\mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\}$ with baseline $\mathbf{I}^{(1)}$, we have:

$$\Delta\mathbf{Q}^{(2),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(3),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(4),(1)} = \{V_1\}. \tag{131}$$

Condition (1) and (2) in Prop. 3 are satisfied. Condition (3) is also satisfied since $V_1 \in \Delta\mathbf{Q}^{(4),(1)}, V_2 \in \Delta\mathbf{Q}^{(2),(1)}$, and $V_3 \in \Delta\mathbf{Q}^{(3),(1)}$. Thus, $\mathbf{V}^{tar}$ is ID w.r.t. V_3 by Prop. 3. See Ex. A33 for a detailed derivation. □

The first result identifies $\Delta\mathbf{Q}$ sets w.r.t. canceled variables through Eq. 119. Next, we define a graphical object that represents conditional dependencies among random variables using an undirected graph.

Definition 3.2 (Markov Network)**.** Let M_V be the Markov network over variables $\mathbf{V}$ with vertices $\{V_i\}_{i=1}^n$ and $\mathcal{E}(M_V)$ denote the set of edges. An edge (V_i, V_j) is added to $\mathcal{E}(M_V)$ if $V_i \not\perp\!\!\!\perp V_j | \mathbf{V} \backslash \{V_i, V_j\}$. □

Now, we present a second result to disentangle variables within $\Delta\mathbf{Q}$ sets leveraging conditional independence among variables.

Proposition 4 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider the variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$. Define $\boldsymbol{\mathcal{E}}$ as the set of edges within the Markov Network of $G_{\overline{T}}(\mathbf{V}^{tar})$ that are contained within a $\Delta\mathbf{Q}$ set.*

$$\begin{aligned}\boldsymbol{\mathcal{E}} = \{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \Big| \\ &\textit{(i)}\ \exists a_l, \{V_k, V_r\} \subseteq \Delta\mathbf{Q}^{(a_l),(a_0)}; \\ &\textit{(ii)}\ V_k \textit{ is d-connected to } V_r \textit{ conditioned on } \mathbf{V}^{tar} \backslash \{V_k, V_r\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\},\end{aligned} \tag{132}$$

For any pair of $V_i, V_j \in \mathbf{V}^{tar}$ such that $V_i \perp\!\!\!\perp V_j \mid \mathbf{V}^{tar} \backslash \{V_i, V_j\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, if there exists $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \ldots, P^{(a_L)}\} \subseteq \mathcal{P}$ that satisfies conditions (1-2) in Prop. 3 and the following condition (3').

(3') Enough changes occur across distributions, i.e., Formally, there exists $\{a'_1, \ldots, a'_{2d'+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \ldots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}$, i) $V_i^{tar} \in \Delta\mathbf{Q}^{(a'_i),(a_0)}$, ii) $V_i^{tar} \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$, and iii) for all $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2d'+j}),(a_0)}$, where $d' = |\mathbf{V}^{tar}|$

then, V_i is ID w.r.t. V_j. □

Prop. 4 disentangles target variables V_i and V_j both in possibly the same $\Delta\mathbf{Q}$ sets. It considers a set of distributions that satisfies conditions (1) and (2) from Prop. 3 and the new condition (3').

Condition (3') conceptually requires that enough changes across all variables in $\mathbf{V}^{tar}$ occur in L distributions. Each variable in $\mathbf{V}^{tar}$ must change a sufficient number of times across different distributions in order to satisfy condition (3'). The first and second part of (3') states that i) $V_i^{tar} \in \Delta\mathbf{Q}^{(a'_i),(a_0)}$, and ii) $V_i^{tar} \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$, which means that each variable in $\mathbf{V}^{tar}$ is present within two different $\Delta\mathbf{Q}$ sets. The third part of (3') considers pairs of variables, (V_k, V_r), that are d-connected conditioned on all other variables [75]. If they are not conditionally independent, then there are $\Delta\mathbf{Q}$ sets that contains the pair of variables.

Example 22 (Disentangle variables within same $\Delta\mathbf{Q}$ sets). Suppose LSD given G^S is the graph shown in Fig. 11(a). Suppose the 9 given intervention targets are

$$\mathbf{\Psi} = \{\{\}^{\Pi_1}, \{V_1^{\Pi_1}\} \times 4, \{V_2^{\Pi_1}, V_3^{\Pi_1}\} \times 4\}, \tag{133}$$

where $\times 4$ indicates multiple distributions with the same intervention targets but different mechanisms. The intervention targets $\mathbf{\Psi}$ implies that all interventions are applied in the same domain Π_1. More specifically, the first intervention $\sigma^{(1)}$ is an idle intervention ($P^{(1)}$ is an observational distribution). $\sigma^{(2)}$ to $\sigma^{(5)}$ is an intervention on V_1. $\sigma^{(6)}$ to $\sigma^{(9)}$ is an intervention on V_2 and V_3. Choose $\mathbf{I}^{(1)} = \{\}^{\Pi_1}$ as the baseline. Then the hard intervention set is empty, namely, $\mathbf{T} = do[\mathbf{I}^{(1)}] = \{\}$. Let $\mathbf{V}^{tar} = \{V_1, V_2, V_3\}$. We have $V_1 \perp\!\!\!\perp V_3 \mid \{V_2\}$ in $G_{\overline{\mathbf{T}}}$. Comparing $\{P^{(2)}, \ldots, P^{(9)}\}$ to the baseline $P^{(1)}$,

$$\begin{aligned} \Delta\mathbf{Q}^{(2),(1)} = \cdots = \Delta\mathbf{Q}^{(5),(1)} = \{V_1, V_2\} \\ \Delta\mathbf{Q}^{(6),(1)} = \cdots = \Delta\mathbf{Q}^{(9),(1)} = \{V_1, V_2, V_3\}, \end{aligned} \tag{134}$$

Since $\{V_1, V_2\}, \{V_2, V_3\} \subseteq \Delta\mathbf{Q}^{(9),(1)}$, V_1 and V_2 are connected conditioning on V_1, and V_2 and V_3 are connected conditioning on V_1, we have

$$\boldsymbol{\mathcal{E}} = \{\{V_1, V_2\}, \{V_2, V_3\}\} \tag{135}$$

Conditions (1-2) are satisfied straightforwardly. Since

$$\begin{aligned} &V_1 \in \Delta\mathbf{Q}^{(2),(1)}, V_1 \in \Delta\mathbf{Q}^{(3),(1)}, \\ &V_2 \in \Delta\mathbf{Q}^{(4),(1)}, V_2 \in \Delta\mathbf{Q}^{(6),(1)}, \\ &V_3 \in \Delta\mathbf{Q}^{(7),(1)}, V_3 \in \Delta\mathbf{Q}^{(8),(1)}, \\ &\{V_1, V_2\} \subseteq \Delta\mathbf{Q}^{(5),(1)} \\ &\{V_2, V_3\} \subseteq \Delta\mathbf{Q}^{(9),(1)} \end{aligned} \tag{136}$$

we know the $V_i^{tar} \in \mathbf{V}^{tar}$ are covered twice and $\boldsymbol{\epsilon} \in \mathcal{E}$ are also covered by $\Delta\mathbf{Q}$ sets. Then condition (4) is satisfied. Thus, V_1 is ID w.r.t. V_3 and V_3 is ID w.r.t. V_1 according to Prop 4. Using the previous Prop. 3, we can only get V_1 is ID w.r.t. V_3 while now we also have V_3 is ID w.r.t. V_1. In contrast, if the 9 given intervention targets are

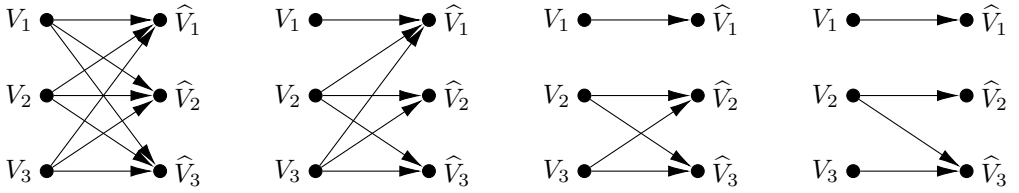
$$\mathbf{\Psi} = \{\{\}^{\Pi_1}, \{V_1^{\Pi_1}\} \times 7, \{V_2^{\Pi_1}, V_3^{\Pi_1}\} \times 2\}, \tag{137}$$

we cannot use Prop. 4 since condition (3') will not be satisfied. The reason is that there will be only two $\Delta\mathbf{Q}$ containing V_3, and then we cannot find three $\Delta\mathbf{Q}$ to cover V_3 two times and cover $\{V_2, V_3\}$ at the same time. □

Prop. 4 implies a search for a collection of distributions to disentangle two conditionally independent variables V_i and V_j in $\mathbf{V}^{tar}$. This result is powerful since the intervention target sets do not always need to involve V_i and V_j. For example, V_1 and V_3 are not simultaneously intervened in any given distributions in Ex. 22. However, condition (3') in Prop. 4 is able to disentangle V_1 and V_3. The complexity of condition (3') arises because it requires that a subset of $\Delta\mathbf{Q}$ sets uniquely covers variables in $\mathbf{V}^{tar}$ twice and elements in $\boldsymbol{\mathcal{E}}$. The following result provides a less general conditionj, but more straightforward condition than Prop. 4.

Corollary 1 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider the variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$ and distributions $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \ldots, P^{(a_L)}\} \subseteq \mathcal{P}$ that satisfies conditions (1) in Prop. 3 and $\Delta\mathbf{Q}^{(a_l),(a_0)} = \mathbf{V}^{tar}$, for $l \in [L]$. For any pair of $V_i, V_j \in \mathbf{V}^{tar}$ such that $V_i \perp\!\!\!\perp V_j \mid \mathbf{V}^{tar} \backslash \{V_i, V_j\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, V_i is ID w.r.t. V_j if $L \geq 2|\mathbf{V}^{tar}| + \delta_{\not\perp}$, where $\delta_{\not\perp}$ is the number of pair $V_k, V_r \in \mathbf{V}^{tar}$ such that V_k and V_r are connected given $\mathbf{V}^{tar} \backslash \{V_k, V_r\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$.* □

To illustrate, if one can find a set of distributions such that all $\Delta\mathbf{Q}$ sets are equivalent to $\mathbf{V}^{tar}$ and the number of distributions ($|\mathcal{P}_{\mathbf{T}}|$) is not smaller than $2|\mathbf{V}^{tar}| + \delta_{\not\perp}$, then V_i and V_j are disentangled from each other. Specifically, Corol. 1 requires all $\Delta\mathbf{Q}$ sets to be equal to $\mathbf{V}^{tar}$. Corol. 1 is a special case of Prop. 4 where $\mathcal{P}_{\mathbf{T}}$ simply induces $\Delta\mathbf{Q}$ sets that are exactly $\mathbf{V}^{tar}$.



(a) Initialization. (b) 1st epoch, Step 11. (c) 1st epoch, Step 14. (d) 2nd epoch, Step 11.

Figure 13: Process of removing edges from CDM $G_{\mathbf{V},\widehat{\mathbf{V}}}$ using Alg. 1 in Ex. 25. (d) is the final output.

Example 23 (Disentangle variables with repeated distribution changes). Consider a LSD G^S that is a collider graph (Fig. 11(b)). Suppose the given intervention targets are $\mathbf{\Psi} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, \{\}^{\Pi_3}, \{\}^{\Pi_4}, \{\}^{\Pi_5}\}$, which means that observational distributions are available in each domain. Consider $\mathbf{T} = \{\}$. Let $\mathbf{V}^{tar} = \{V_1, V_3\}$. Then, $V_1 \perp\!\!\!\perp V_3$ in $G(V_1, V_3)$. Based on Def. 3.1, $\Delta\mathbf{Q}[\mathbf{I}^j, \mathbf{I}^1, \mathbf{T}, G^S] = \{V_1, V_3\}$ for $j = 2, 3, 4, 5$. Then the number of distributions used for comparing (i.e., four) is not smaller than the required $(2 \times 2 + 0)$, which means V_1 is ID w.r.t. V_3 and V_3 is ID w.r.t. V_1 by Corol. 4. See Ex. A35 for a derivation. □

With these existing disentanglements from Props. 3 and 4, the following result considers the inverse direction, which identifies canceled variables w.r.t. $\Delta\mathbf{Q}$ sets [21].

Proposition 5 (ID of canceled variables w.r.t. $\Delta\mathbf{Q}$ sets). *Suppose there exists $\mathbf{I}^{(k)} \in \mathbf{\Psi}$ such that $do(\mathbf{I}^{(k)}) = \mathbf{T}$. Given $\mathbf{V}\backslash V^{tar}$ is ID w.r.t. a single variable V^{tar}, V^{tar} is ID w.r.t. $\mathbf{V}\backslash V^{tar}$ if $V^{tar} \perp\!\!\!\perp \mathbf{V}\backslash V^{tar}$ in $G_{\overline{\mathbf{T}}}$.* □

To illustrate, Prop. 5 states that if $\mathbf{V}\backslash\{V^{tar}\}$ is already disentangled from V^{tar}, then V^{tar} can ID w.r.t. $\mathbf{V}\backslash\{V^{tar}\}$ if a hard intervention on $\mathbf{T}$ exists to separate V^{tar} and $\mathbf{V}\backslash\{V^{tar}\}$ in $G_{\overline{\mathbf{T}}}$. Prop. 5 does not compare distributions but relies on existing disentanglement and independence.

Example 24 (Task 4 and **Ex. 20 continued**). Consider $\mathbf{V}^{tar} = \{V_1\}$. From Ex. 15, we have $\{V_2, V_3\}$ is ID w.r.t. V_1 by comparing $\{P^{(2)}, P^{(3)}\}$ with $P^{(1)}$ according to Prop. 3. Now we will leverage $P^{(4)}$ with intervention target $\mathbf{I}^{(4)} = \{V_2^{\Pi_1, do}\}$. Consider $\mathbf{T} = \{V_2\}$ (from $\mathbf{I}^{(4)}$). Since $V_1 \perp\!\!\!\perp \{V_2, V_3\}$ in $G_{\overline{V_2}}$, then V_1 is ID w.r.t. $\{V_2, V_3\}$ according to Prop. 5. □

We illustrate and provide additional comparisons between these three graphical criteria with existing work in Appendix E.

4 Algorithmic Causal Disentanglement

In this section, we develop an algorithmic procedure for determining whether any $\mathbf{V}^{tar}$ and $\mathbf{V}^{en}$ are disentangleable given the LSD G^S and interventions sets $\mathbf{\Psi}$ leveraging Props. 3, 4 and 5.

The procedure is called **CausalRepresentationID** (**CRID**, for short), and is described in Alg. 1. We start by introducing a bipartite graph $G_{\mathbf{V},\widehat{\mathbf{V}}}$, called *Causal Disentanglement Map (CDM)* (which was informally shown in Fig 2 (right)).

Definition 4.1 (Causal disentanglement map)**.** Let $\mathcal{M}$ be an ASCM, G^S be the induced LSD, and $\mathcal{P}$ a collection of observed distributions. Let $\widehat{\mathbf{V}}$ be the learned representation from a proxy model that is compatible with G^S and $\mathcal{P}$. $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is the causal disentanglement map. If $V_i \not\to \widehat{V}_j$, then we say V_j is ID w.r.t. V_i. □

In words, the absence of the edge $V_i \not\to \widehat{V}_j$ implies V_j is ID w.r.t V_i. If each $\widehat{V}_i$ is only pointed by V_i, then we have full disentanglement of $\mathbf{V}$. If $\widehat{V}_i$ is pointed by $\mathbf{V} \subset \widehat{V}_i$, then we have partial disentanglement of V_i.

CRID begins with a fully connected CDM in Step 1. In each iteration, the hard intervention set $\mathbf{T}$ and the baseline intervention target set $\mathbf{I}$ (Steps 4 and 6) are considered. For each $\mathbf{T}$ and baseline, all $\Delta\mathbf{Q}$ sets are constructed based on Def. 3.1 and put into a collection $\mathcal{Q}$ (Steps 7). Next, $\mathbf{Q}$, the union of all $\Delta\mathbf{Q}$ sets, is constructed (Step 8). Props. 3 and 4 are leveraged in two procedures (Step 9 and 10)

[21] Recall that ID is one-way, since the ID of V_i w.r.t. V_j does not imply V_j is ID w.r.t. V_i.

Algorithm 1 CRID: Algorithm for determining causal representation identifiability - G^S is the LSD; $\mathbf{\Psi}$ is the intervention target sets; $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is the output bipartite graph (i.e. CDM).

Input: $G_{\mathbf{V}}$, and intervention target sets $\mathbf{\Psi}$.
Output: CDM $G_{\mathbf{V},\widehat{\mathbf{V}}}$

1: $G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ **FullyConnectedBipartiteGraph**$(\mathbf{V}, \widehat{\mathbf{V}})$ ▷ Initialize $G_{\mathbf{V},\widehat{\mathbf{V}}}$ with Alg. F.2
2: **while** $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is updated in the last epoch **do**
3: $\quad \mathbf{HARD} = \{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_s \mid do(T_i) \in \mathbf{\Psi}$ ▷ Get hard intervention variables sets.
4: $\quad$ **for** all $\mathbf{T} \in \mathbf{DO}$ **do**
5: $\quad\quad \mathbf{\Psi}_{\mathbf{T}} \leftarrow \mathbf{\Psi}$ ▷ Collect intervention targets that contain hard intervention variables $\mathbf{T}$
6: $\quad\quad$ **for** all $\mathbf{I} \in \mathbf{\Psi}_{\mathbf{T}}$ such that $do[\mathbf{I}] = \mathbf{T}$ **do** ▷ Iterate intervention targets as the baseline
7: $\quad\quad\quad \mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_{|\mathbf{\Psi}_{\mathbf{T}} \setminus \mathbf{I}|}\}$, where $Q_k \leftarrow \Delta\mathbf{Q}[\mathbf{J}^{(k)}, \mathbf{I}, \mathbf{T}, G^S]$ ▷ Construct $\Delta\mathbf{Q}$ sets.z
8: $\quad\quad\quad$ **for** all $\mathbf{Q}$ such that $\mathbf{Q} = \bigcup_{\mathbf{Q}_l \in \mathcal{Q}} \mathbf{Q}_l \subset \mathbf{V}$ **do** ▷ Iterate the union of $\Delta\mathbf{Q}$ factorsx
9: $\quad\quad\quad\quad G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ **DisΔQFromCancel**$(\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}}, \Psi_{\mathbf{T}}, \mathbf{I}, \mathcal{Q})$ ▷ Alg. F.3 and Prop. 3
10: $\quad\quad\quad\quad G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ **DisWithinΔQ**$(\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}}, \Psi_{\mathbf{T}}, \mathbf{I}, \mathcal{Q})$ ▷ Alg. F.6 and Prop. 4
11: $\quad$ **for** all $\mathbf{T} \in \mathbf{HARD}$ **do**
12: $\quad\quad G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ **DisCancelFromΔQ**$(G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}})$ ▷ Alg. F.8 and Prop. 5
13: **return** $G_{\mathbf{V},\widehat{\mathbf{V}}}$

to check the identification of $\mathbf{Q}$ w.r.t. $\mathbf{V}\backslash\mathbf{Q}$ and the identification within $\mathbf{Q}$. The disentanglements in CDM at the current stage are leveraged to reduce the required number of distributions. At the end of the iteration, Prop. 5 is used for identifying $\mathbf{V}\backslash\mathbf{Q}$ from $\mathbf{Q}$ leveraging current disentanglement in CDM (Step 11-12).

Example 25 (**Task 4 continued**). Consider the selection diagram (Fig. 11(a)) and the setup in Ex. 5. The hard intervention variable sets are the empty set $\{\}$ and $\{V_2\}$. First, $\mathbf{T}$ is chosen as $\{\}$. Choosing the baseline $\mathbf{I} = \mathbf{I}^{(1)}$, the $\Delta\mathbf{Q}$ collection: $\mathcal{Q} = \{\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{Q}_3\} = \{\{V_2, V_3\}, \{V_2, V_3\}, \{V_1, V_2\}\}$. We consider the $\mathbf{Q}$ as $\{V_2, V_3\}$ and $\{V_1, V_2\}$. For $\mathbf{Q} = \{V_2, V_3\}$, leveraging Step 9 (Prop. 3), the edges from V_1 to $\{\widehat{V}_2, \widehat{V}_3\}$ are removed (See Ex. 20 for details) and Step 10 (Prop. 4) does not remove further edges. However, for $\mathbf{Q} = \{V_1, V_2\}$, no edge can be removed, since it at least needs two comparisons for claiming disentanglement.

Choosing $\{\}^{\Pi_2}$ or $V_3^{\Pi_2}$ or $V_2^{\Pi_1, \text{do}}$ as the baseline, no new $\mathbf{Q}$ can be constructed. Thus no further edges are removed. When $\mathbf{T}$ is chosen as $\{V_2\}$, the comparison does not work since only one distribution is available. At the end of this iteration, with the fact that $\{V_2, V_3\}$ is ID w.r.t. V_1 and $V_1 \perp\!\!\!\perp \{V_2, V_3\}$ in $G_{\overline{V_2}}$, Step 12 (Prop. 5) removes edges from V_2 to $\widehat{V}_1$ and V_3 to $\widehat{V}_1$.

In the second iteration, the algorithm repeats the choice of $\mathbf{T}$ and the baseline. Now, for $\mathbf{Q} = \{V_1, V_2\}$, the edge from V_3 to $\widehat{V}_2$ is removed since V_3 to $\widehat{V}_1$ has already been removed in CDM and only 1 comparison is needed now (see details in Ex. A31). At the end of this iteration no more edges can be removed by Alg. F.8. In the last iteration, $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is not updated and the CDM is returned (Fig. 13). □

After obtaining $G_{\mathbf{V},\widehat{\mathbf{V}}}$ from **CRID**, the identifiability of target variables $\mathbf{V}^{tar}$ w.r.t. $\mathbf{V}^{en}$ can be inferred through the absence of edges in $G_{\mathbf{V},\widehat{\mathbf{V}}}$. The following result considers the soundness of CRID.

Theorem 1 (**Soundness of CRID**). *Consider a LSD G^S and intervention targets $\mathbf{\Psi}$. Consider the target variables $\mathbf{V}^{tar}$ and $\mathbf{V}^{en} \subseteq \mathbf{V}\backslash\mathbf{V}^{tar}$. If no edges from $\mathbf{V}^{tar}$ points to $\widehat{\mathbf{V}}^{en}$ in the output causal disentanglement map (CDM) from **CRID**, $G_{V,\widehat{V}}$, then $\mathbf{V}^{tar}$ is ID w.r.t $\mathbf{V}^{en}$.* □

5 Experiments

In this section, we provide empirical evaluations to corroborate with our theoretical contributions and the newly proposed CRID algorithm. We start with a synthetic dataset (Sec. 5.1) and then move to a modified Colored MNIST dataset (Sec. 5.2).

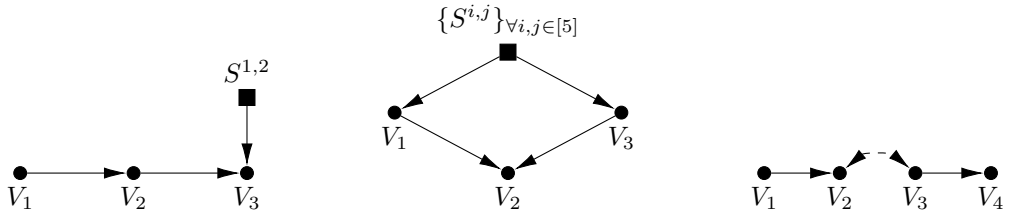


Figure 14: LSDs in Experiments. (a) chain, (b) collider, and (c) non-markovian graph.

5.1 Synthetic Experiments

We consider three settings following the LSDs shown in Fig. 14 and different interventional datasets as input. Specifically, the settings are as follows:

1. **Chain setting**. The input graph is shown in Fig. 14(a). To illustrate, generative factors V_1, V_2, V_3 forms a "chain" relationship. V_3 is perturbed due to the domain change from Π_1 to Π_2. We consider two groups of input distributions in this setting. The first group includes observational distributions from Π_1 and Π_2, and a soft interventional distribution on V_3 in Π_1,
$$\mathcal{P} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_1}(\mathbf{X}), P^{\Pi_1}(\mathbf{X}; \sigma_{V_3})\} \tag{138}$$
The second group of distributions is the first group plus a hard interventional distribution on V_2 in Π_1,
$$\mathcal{P} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}), P^{\Pi_1}(\mathbf{X}; \sigma_{V_3}), P^{\Pi_1}(\mathbf{X}; do(V_2))\} \tag{139}$$
2. **Collider setting**. The input graph is shown in Fig. 14(b). To illustrate, generative factors V_1, V_2, V_3 forms a "collider" relationship. V_1 and V_3 are perturbed due to the domain change from Π_1 to Π_5. The input distribution includes observational distribution in each domain,
$$\mathcal{P} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}), P^{\Pi_3}(\mathbf{X}), P^{\Pi_4}(\mathbf{X}), P^{\Pi_5}(\mathbf{X})\}. \tag{140}$$
3. **Non-Markovain setting**. The input graph is shown in Fig. 14(c). To illustrate, generative factors V_1, V_2, V_3, V_4 forms a non-Markovian model. The input distribution includes the observational distribution and two different hard interventions on V_3,
$$\mathcal{P} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}; do(V_3)), P^{\Pi_3}(\mathbf{X}; do(V_3))\} \tag{141}$$

In each setting, the underlying true collection of ASCMs $\boldsymbol{\mathcal{M}}$ induces the corresponding LSD G^S, and data sampled from $\mathcal{P}$ are collected, which is shown in Sec. 5.1.1. Taking these as input, **CRID** will output a CDM to illustrate what latent variables are expected to exhibit disentanglement, and which will still likely be entangled. The goal of experiments here is to empirically corroborate with the disentanglement relationships outputted by CRID.

We first train neural proxy models $\widehat{\boldsymbol{\mathcal{M}}}$ that are compatible with the given G^S to match the collected data for obtaining representations $\widehat{\mathbf{V}}$ (see model details in Sec. 5.1.2). Then, disentanglement between learned representations and the ground truth underlying generative factors can be empirically evaluated by the mean correlation coefficient (MCC) of $\widehat{\mathbf{V}}$ w.r.t. $\mathbf{V}$, which is a standard protocol used in prior work [18] (Sec. 5.1.3). We expect the theoretical disentanglement outputted by CRID can be reflected by these MCCs computed with ground truth, as shown in Sec. 5.1.4.

5.1.1 Data-generating Process

In this section, we illustrate the data generation process of the ground truth ASCMs $\boldsymbol{\mathcal{M}}^\dagger = \{\mathcal{M}_1, \mathcal{M}_2, \dots\}$ in each setting. We start with the ASCM $\mathcal{M}_1^\dagger = \langle \mathbf{U}^\dagger, \{\mathbf{V}, \mathbf{X}\}, \mathcal{F} = \{\mathcal{F}_0, f_{\mathbf{X}}\}, P(\mathbf{U}^\dagger)\rangle$ in the first domain and then move to other domains.

Exogenous variables. The exogenous variables are set as
$$\mathbf{U}^\dagger = \{U_{\mathbf{C}}^\dagger : \mathbf{C} \in \mathcal{C}(G)\} \cup \{U_i\}_{i=1}^d, \tag{142}$$
where $\mathcal{C}(G)$ is the set of all maximal cliques over bidirected edges of LCG G. All $U \in \mathbf{U}^\dagger$ follows an independent standard normal distribution.

Mechanism of generative factors $\mathbf{V}$. Mechanisms in generative SCMs for $\mathbf{V}$ are set as linear functions:

$$V_i \leftarrow \sum_{V_j \in Pa_i} \alpha_{i,j} V_j + \sum_{U \in \mathbf{U}^\dagger_{V_i}} U + U_i \tag{143}$$

where $\mathbf{U}^\dagger_{V_i} = \{U^\dagger_{\mathbf{C}} : U^\dagger_{\mathbf{C}} \in \mathbf{U}^\dagger \text{ s.t. } V_i \in \mathbf{C}\}$; linear parameters $\alpha_{i,j}$ are drawn from a uniform distribution, i.e., $\alpha_{i,j} \sim \text{Uniform}[-3, 3]$.

Generating multiple domains. To generate data in a new domain, where $S^{j,k} \to V_i$ indicates a change of V_i due to the change in ASCMs between $\mathcal{M}_j$ and $\mathcal{M}_k$, we start from the first ASCM generated, and then modify the distribution of the exogenous variable U_i with a mean shift and variance scaling.

Generating interventions within each domain. We set each hard intervention as a Gaussian stochastic intervention. To illustrate, suppose the intervention target set $\mathbf{I}^{(k)}$ will be applied to generate distribution $P^{(k)}$ and $V_i^{\text{do}} \in \mathbf{I}^{(k)}$, we set:

$$V_i \leftarrow \epsilon_i^{(k)}, \quad \text{with } \epsilon_i^{(k)} \sim \mathcal{N}(\mu_i^{(k)}, \sigma_i^{(k)}) \tag{144}$$

such that $\mu_i^{(k)}$ is uniformly drawn from $[-3, -1) \cup (1, 3]$ and $\sigma_i^{(k)}$ is drawn from $[2, 10]$. This ensures the given distribution changes sufficiently.

In terms of soft intervention, suppose the intervention target set $\mathbf{I}^{(k)}$ will be applied to generate distribution $P^{(k)}$ and $V_i \in \mathbf{I}^{(k)}$, we set:

$$V_i \leftarrow \sum_{V_j \in Pa_k} \alpha'_{i,j} V_j + \epsilon_i^{(k)}, \quad \text{with } \epsilon_i^{(k)} \sim \mathcal{N}(\mu_i^{(k)}, \sigma_i^{(k)}) \tag{145}$$

where parameters $\alpha'_{i,j}$ are drawn from a uniformly from $[-a, a]$, $\mu_i^{(k)}$ is uniformly drawn from $[-3, -1) \cup (1, 3]$, and $\sigma_i^{(k)}$ is drawn from $[2, 10]$.

For each distribution over $\mathbf{V} \in \mathbb{R}^d$, we generate 200,000 data points resulting in $d \times 200,000$ data points in total for N total distributions.

Mixing function. In order to generate the low-level data $\mathbf{X}$, we will apply a mixing function $f_{\mathbf{X}}$ to the generated latent variables $\mathbf{V}$. Following [21, 51], we will use a multilayer perceptron $\mathbf{f}_{\mathbf{X}} = \sigma \circ \mathbf{A}_M \circ ... \circ \sigma \mathbf{A}_1$, where $\mathbf{A}_M \in \mathbb{R}^{d \times d}$ for $m \in [1, M]$ denotes invertible linear matrices and σ is an element-wise invertible nonlinear function. We then use the tanh function as done in [76]:

$$\sigma(x) = tanh(x) + 0.1x \tag{146}$$

In a rejection-like procedure, each sampled matrix $\mathbf{A}_i$ is re-drawn if $|\det \mathbf{A}_i| < 0.1$. This ensures that the linear maps are not ill-conditioned and close to being singular. Once the mixing function is drawn for a given simulation, it is fixed across all domains and interventions according to Def. 2.1, and then $\mathcal{P}$ is drawn according to all ASCMs instantiated.

5.1.2 Proxy Model

We train invertible MLPs with normalizing flows. The parameters of the causal mechanisms are learned while the causal graph is assumed to be known. We leverage the implementation in [21], and extend it for our experiments. The encoder is trained with the following objective that estimates the inverse f^{-1}, and the latent densities $P(\mathbf{V})$ reproducing the ground-truth up to certain mixture ambiguities (c.f. Lemmas 3, 7). The encoder parameters are estimated by maximizing the likelihood.

Normalizing flows. We use a normalizing flows architecture [77] to learn an encoder $\mathbf{g}_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Therefore, the observations $\mathbf{X}$ will be the result of an invertible and differentiable transformation:

$$\mathbf{X} = \mathbf{g}_\theta(\mathbf{V}) \tag{147}$$

Specifically, g_θ will comprise of Neural Spline Flows [78] with a 3-layer feedforward neural network with hidden dimension 128 and a permutation in each flow layer.

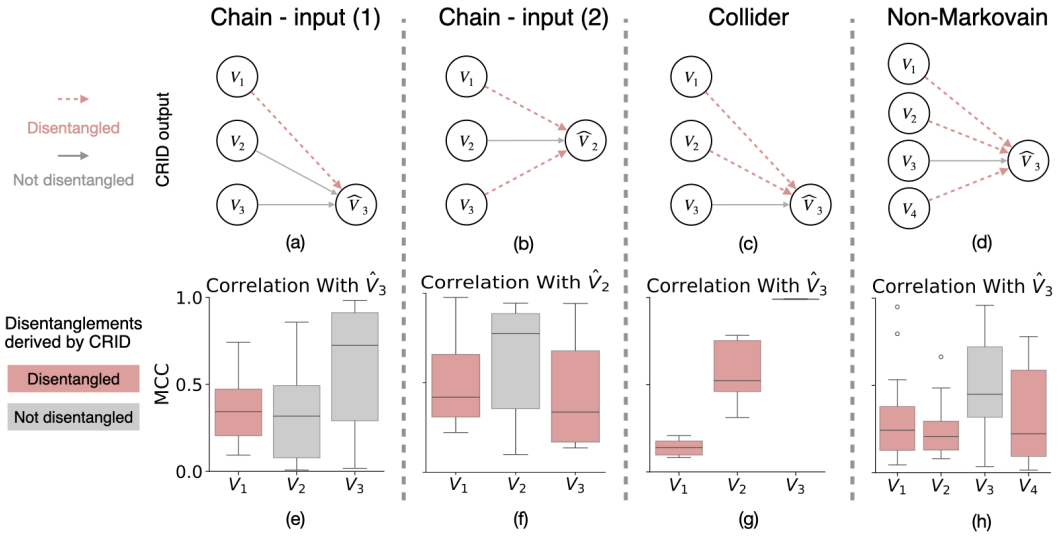


Figure 15: CRID outputs ((a)-(d)) and MCC ((e)-(d)) of learned latent representations with true latent variables in synthetic data experiments. (a) and (e): the chain graph with $\Sigma = \{\sigma_{\{\}}, \sigma_3, \sigma_3\}$; (b) and (f) the chain graph with $\Sigma = \{\sigma_{\{\}}, \sigma_3, \sigma_3, do(V_2)\}$; (c) and (g): the collider graph with $\Sigma = \{\sigma_{\{\}}, \sigma_{1,3}, \sigma_{1,3}, \sigma_{1,3}, \sigma_{1,3}\}$; (d) and (h): the non-Markovian graph with $\Sigma = \{do(V_3), do(V_3)\}$.

Base distributions. Normalizing flows require a base distribution. We leverage one baseline distribution per sampled dataset, $(\hat{p}_\theta^k)_{k\in[d]}$ over the base noise variables $\mathbf{V}$. The conditional density of any variable is given by:

$$\hat{p}_\theta^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{Pa}_\mathbf{i}}^+) = \mathcal{N}\Big(\mu_{i,\theta}(\widehat{\mathbf{pa}}^+), \sigma_{i,\theta}(\widehat{\mathbf{pa}}^+)\Big) \tag{148}$$

where $\mu_{i,\theta}$ and $\sigma_{i,\theta}$ are also flow networks. If a change-in-domain or an intervention is applied, the corresponding mean and variance are replaced by another transformation. When a hard intervention is applied, we let

$$\hat{p}_\theta^k(v_i) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i) \tag{149}$$

where $\hat{\mu}_i, \hat{\sigma}_i$ are randomly sampled from uniform distribution.

To fit the proxy normalizing flow with given data, we maximize data likelihood. The learning objective is expressed as:

$$\theta^* = \arg\max_\theta \sum_{k=0}^{N} \Big(\frac{1}{n_k} \sum_{n=1}^{n_k} \log \hat{p}_\theta^{(k)}(\mathbf{x}_n)\Big) \tag{150}$$

where n_k represents the size of the dataset P^k, which is 200,000 in our simulations. We perform 10 training runs over different seeds for each experiment and show the distributions of the mean-correlation coefficient (MCC). ADAM optimizer [79] is applied for optimization and the learning rate is set as 1e-4. We train the model for 200 epochs with a batch size of 4096. We use NVIDIA H100 GPUs to train the neural network models.

5.1.3 Evaluation Metrics

The output of our trained model is the learned representations $\hat{\mathbf{V}} = g_\theta(\mathbf{X})$. We compute the mean correlation coefficients (MCC) between such learned representations and ground-truth latent generative factors $\widehat{\mathbf{V}}$ as the measurement of the ground truth of disentanglement.

To illustrate, according to the disentanglement definition (Def. 2.3), if there exists a function τ such that $\widehat{V}_j = \tau(\mathbf{V}^{en})$ and $V_k \notin \mathbf{V}^{en}$, then V_j is disentangled from V_k. The lower MCC of $\widehat{V}_i$ w.r.t. V_j indicates that $\mathbf{V}^{en}$ does not need to include V_k to be mapped to V_j, which exactly implies that V_j is disentangled from V_k. Similarly, the relatively higher MCC indicates the entanglement

between V_i and V_j. We expect there to be an overall lower MCC for variables that are predicted to be disentangled by CRID relative to variables that are not deemed disentangled.

Note that our algorithm is shown to be sufficient (but not necessary), so there may be variables that are disentangled at the end of our training process that are not captured by the output of CRID. Characterizing when this occurs and coming up with a complete theoretical characterization of disentanglement is a line for future work.

5.1.4 Evaluation Results

In this section, we show and illustrate our empirical verification results. In each setting, we illustrate the CRID outputs and check whether the outputs are aligned with empirical MCCs. Two baselines [21, 22] who also work in non-parametric settings are also chosen here to compare with CRID.

Chain Graph. Taking the input graph assumption Fig. 14(a) and intervention targets $\Psi = \{\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{\}^{\Pi_2}, \mathbf{I}^{(3)} = \{V_3^{\Pi_1}\}$, CRID states that V_3 is disentangled from V_1 (Fig. 15(a)). The MCC reported in Fig. 15(e) indicates this disentanglement because $MCC(\widehat{V}_3, V_1)$ is relatively lower compared to $MCC(\widehat{V}_3, V_3)$ No prior work can achieve this disentanglement results with such input. To disentangle V_3 from V_1, [21] needs interventions on V_1 and V_2 while CRID does not. [22] needs observational distributions from at least 5 domains.

Second, after adding an additional hard intervention on V_2 ($\{\mathbf{I}^{(1)} = \{V_2^{\Pi_1, do}\}$), CRID outputs disentanglements in Fig. 15(b). To illustrate, V_2 is ID w.r.t. V_1 and V_3. Fig. 15(f) shows the identification of V_2 w.r.t. $\{V_1, V_3\}$ from MCCs, which is consistent with the CRID outputs. Also, this result is not achieved in previous works. [21] needs interventions on V_1 to get V_2 to be disentangled from V_1. And [22] does not claim any result between $\{V_1, V_3\}$ and V_2.

Collider Graph. In this setting, taking the input graph assumption Fig. 14(b) and intervention targets $\Psi = \{\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{\}^{\Pi_2}, \mathbf{I}^{(3)} = \{\{\}^{\Pi_3}, \mathbf{I}^{(4)} = \{\{\}^{\Pi_4}, \mathbf{I}^{(5)} = \{\{\}^{\Pi_5}\}$, CRID suggests V_1 and V_3 are ID w.r.t. V_2 as shown in Fig. 15(c). Fig. 15(g) verifies this disentanglement since $MCC(\widehat{V}_3, V_3) > MCC(\widehat{V}_3, V_i)$ and $MCC(\widehat{V}_2, V_2) > MCC(\widehat{V}_2, V_i)$. No prior achieved this result. [21] needs a single node intervention on V_3. According to [22], since V_1 and V_3 are adjacent in the Markov Network, V_1 and V_3 are not disentangleable.

Non-Markovian Graph. In this setting, taking the input graph assumption Fig. 14(c) and intervention targets $\Psi = \{\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{V_3^{\Pi_1, do}\}, \mathbf{I}^{(3)} = \{V_3^{\Pi_1, do}\}$, CRID outputs that V_3 is ID w.r.t $\{V_1, V_2, V_4\}$ as shown in Fig. 15(d). Fig. 15(h) verifies this result since the $MCC(\widehat{V}_i, V_3)$ $(i = 1, 2, 4)$ is lower than $MCC(\widehat{V}_3, V_3)$. Compared to existing works, no prior results achieve this disentanglement since there are confounding among $\mathbf{V}$.

5.2 Semi-synthetic: Handwritten digits (ColorMNISTBar)

Next, we analyze the identifiability results of CRID on an image dataset consisting of handwritten digits and bars. This dataset is a modification of the popular MNIST dataset [80]. To illustrate, we artificially added a bar to the top of every image that is 4 pixels in height and consider the generative factors: digit V_1, color of digit V_2, the color of bar V_3, and writing style V_4. The full generative process is described in Sec. 5.2.1.

In order to obtain the representation $\widehat{\mathbf{V}}$ from the proxy model for this image dataset, we train a normalizing flows model that is compatible with the true LSD to fit the given data distribution leveraging VAE embedding. The model details are shown in Sec. 5.2.2. Then, we compute the MCC between learned $\widehat{\mathbf{V}}$ and $\mathbf{V}$, and consider a downstream image editing task to verify the disentanglement output from the proposed CRID. The experimental results are shown in Sec. 5.2.3.

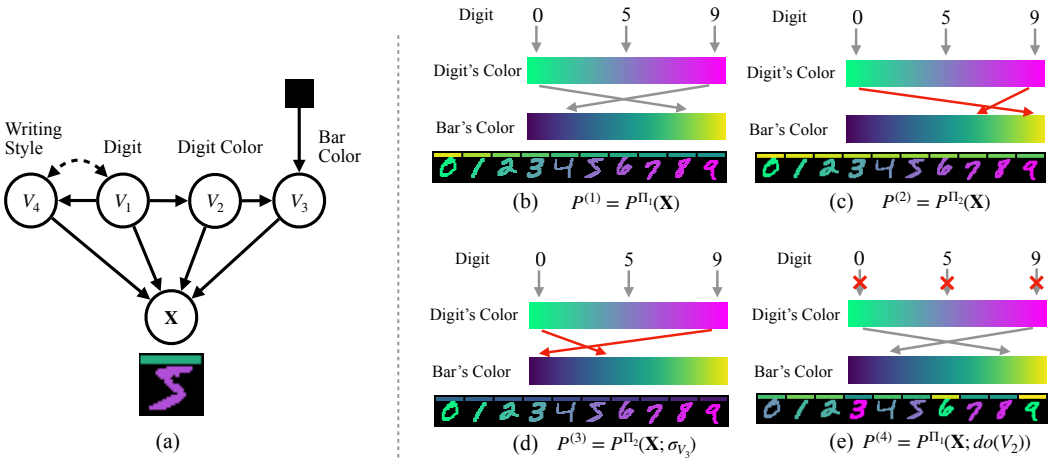


Figure 17: Color MNIST with bar data generation. The ground-truth LSD is shown in (a). Four distributions are generated with observations in domain Π_1 (b), observations in domain Π_2 (c), soft intervention on the color-bar (V_3) in Π_1 (d), and a hard intervention on the color-digit (V_2) in Π_1.

5.2.1 ColorMNISTBar data-generating process

The generation process of this original gray-scale digit dataset could be described as the following ASCM $\mathcal{M}^{\text{mnist}}$ over the generative factors digit V_1 and writing style V_4:

$$\mathcal{M}^{\text{mnist}} = \begin{cases} \mathbf{U} = \{U_{14}, U_4\} \\ \mathbf{V} = \{V_1, V_4\}, \\ \mathbf{X}: \quad \text{a gray-scale digit image} \\ \mathcal{F} = \begin{cases} V_1 \leftarrow f_1(U_{14}) \\ V_4 \leftarrow f_4(V_1, U_4, U_{14}) \\ \mathbf{X} \leftarrow f_{\mathbf{X}}^{\text{mnist}}(\mathbf{V}) \end{cases} \\ P(U_{14}, U_4) \end{cases}, \tag{151}$$

Fig. 16 illustrates that the digits in MNIST exhibit different writing styles, which are confounded by the setting in which the original digit was written.

Figure 16: Example of different samples of the digit "4" in MNIST dataset.

We modify this dataset by artificially adding a bar to the top of every image that is 4 pixels in height. This bar will vary in color as described below. Then, we impose the following causal generative factors: digit (V_1), color of the digit (V_2), and color of the bar (V_3) are generated according to a causal chain $V_1 \to V_2 \to V_3$. And V_3 will be perturbed due to the domain change. The final LSD we consider is the one shown in Fig. 17(a).

Specifically, the ColorMNISTBar data generating process follows the underlying true ASCM $\boldsymbol{\mathcal{M}}^{\text{c-b-mnist}} = \langle \mathcal{M}_1^{\text{c-b-mnist}}, \mathcal{M}_2^{\text{c-b-mnist}} \rangle$ in two domains over the generative latent factors V_1, V_2, V_3, V_4. $\boldsymbol{\mathcal{M}}^{\text{c-b-mnist}}$ is shown as follows:

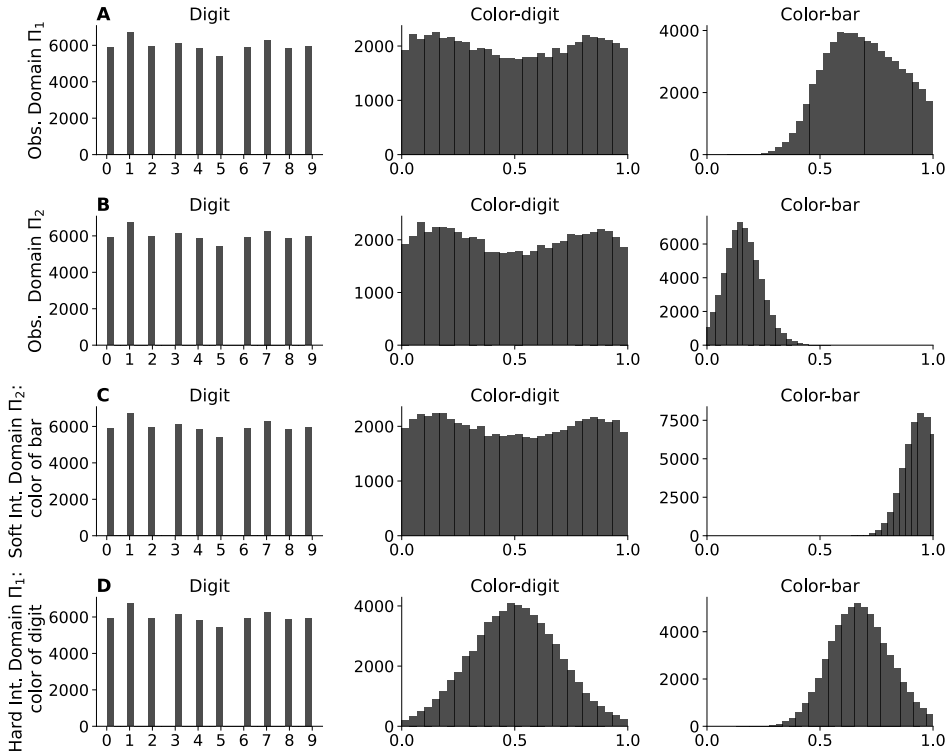


Figure 18: Distributions of the ground-truth causal latent generative factors in the ColorMNISTBar experiment. (a) shows the observational distribution in domain Π_1, (b) the observational distribution in domain Π_2, (c) the soft intervention on color-digit (V_3) in Π_2, and (d) the hard intervention on color-digit (V_2) in Π_1.

$$\mathcal{M}_i = \begin{cases} \mathbf{U} = \{U_{14}, U_2, U_3, U_4\} \\ \mathbf{V} = \{V_1, V_2, V_3, V_4\}, \\ \mathbf{X}: \text{ a colored digit image with colored bar} \\ \mathcal{F} = \begin{cases} V_1 \leftarrow f_1(U_{14}) \\ V_2 \leftarrow f_2(V_1, U_2) \\ V_3 \leftarrow f_3(V_2, U_3^i) \\ V_4 \leftarrow f_4(V_1, U_4, U_{14}) \\ \mathbf{X} \leftarrow f_{\mathbf{X}}^{\text{c-b-mnist}}(\mathbf{V}) \end{cases} \\ P_i(U_{14}, U_1, U_2, U_3^i, U_4) \end{cases}, \tag{152}$$

We illustrate the generative process $\mathcal{F}$ with $P_i(\mathbf{U})$ one by one.

$f_1(\cdot), f_4(\cdot), U_1, U_4$ constructs digit V_1 and its writing style V_2. These two processes follow the original MNIST dataset and they are unobserved.

$f_2(\cdot), U_2$ constructs the $P(V_2 \mid v_1)$ follows a truncated normal distribution with means according to each digit equally spaced between 0 and 1 (e.g. mean of color-digit for digit 5 is 0.5). The standard deviation is 0.15 for all digits. The values are truncated between 0 and 1 to determine the final color of the sample given the digit.

$f_3(\cdot), U_3$ constructs $P(V_3 \mid v_2)$ that also follows a truncated normal distribution where the mean of each sample is $1/(v_2^i + 1)$ and the standard deviation is 0.1. The final value is truncated between 0 and 1. U_3 is perturbed when the domain changes from Π_1 to Π_2. In the domain Π_2, $P(V_3 \mid v_2)$ is a skewed normal distribution with a mean of 0.1 and a standard deviation of $1/(v_2^i + 0.5)$.

The mixing function $f_{\mathbf{X}}^{\text{c-b-mnist}}$ is aggregated by the $f_{\mathbf{X}}^{\text{mnist}}$ from the original MNIST generation process and our modification. $f_{\mathbf{X}}^{\text{c-b-mnist}}$ map the digit and writing style exactly the same as $f_{\mathbf{X}}^{\text{c-b-mnist}}$ does. The

value v_2 is mapped through to a RGB value where $R = v_2^i$, $G = 1 - v_2^i$, and $B = v_2^i + (1 - v_2^i)/2$. And $f_{\mathbf{X}}^{\text{c-b-mnist}}$ dye the digit using this RGB value. The value value v_3^i is mapped to an RGB value using the "viridis" colormap from the python matplotlib package [81]. And then $f_{\mathbf{X}}^{\text{c-b-mnist}}$ dye the bar using this RGB value.

Two interventions are applied Π_1. The first intervention is a soft intervention on V_3 in domain Π_2. This intervention leads $P(V_3 \mid v_2)$ to follow a skewed normal distribution where the mean changes to 0.9. For the hard intervention on V_2 in domain Π_1, the color-digit distribution $P(V_2)$ is generated according to a truncated normal distribution with a mean of 0.5 and a standard deviation of 0.2 truncated between 0 and 1.

To summarize, the input data consists of 2 observational distributions from domain Π_1 and Π_2, a soft intervention from Π_2 and a hard intervention from Π_1. Namely,

$$\mathcal{P} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}; \sigma_{V_3}), P^{\Pi_1}(\mathbf{X}; do(V_2))\} \tag{153}$$

This results in a new dataset where the images follow the distributions in Fig. 17(b-e). To illustrate, the gray and red arrows demonstrate the designed mechanism and exogenous variables for V_2 and V_3. The red arrows denote the difference compared to (b). We do not show the mechanism for V_1 and V_4 since they are unobserved (following the original MNIST dataset). The distributions of the ground-truth latent variables for V_1, V_2, V_3 are shown in Fig. 18.

5.2.2 Proxy Model

We mostly follow the modeling, training, and evaluation discussed in Section 5 with the exception of how we handle the large dimensionality differences between the latent generative factors and the input image data. When scaling up disentangled causal representation learning algorithms to images, such as the ColorMNISTBar setting, normalizing flows is not trivial to apply due to the fact that the latent dimensions must equal the input dimensions. For an image in our dataset of (28, 28, 3), this consists of 2352 dimensions, whereas the dimensionality of the latent causal generative factors is 4. We propose two empirical methods for handling the mismatch in dimensionality between the generative latent factors, and the input observed data.

VAE Embedding. A VAE model is used to learn an embedding $\tilde{\mathbf{X}}$ of the input images $\mathbf{X}$ as in Fig. 19 [82]. The VAE helps compress the dimensionality of the input to an embedding with 32 dimensions. This then allows the normalizing flow to learn an invertible flow from 32 dimensions to latent generative factors.

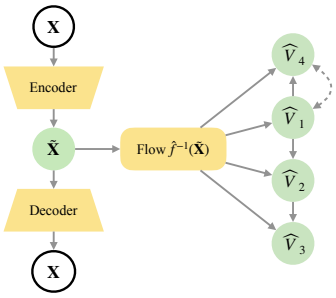
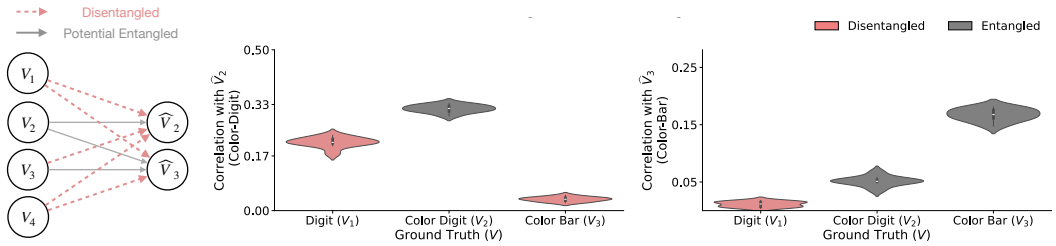


Figure 19: VAE-NF model with input images $\mathbf{X}$. The model learns a representation of the causal latent generative factors ($\widehat{V}_i$), and an embedding of the observed input images ($\tilde{\mathbf{X}}$).

Causal Clustered DAG Base Distribution. A normalizing flow model requires the base distribution dimensionality to be the same as the input [77]. This may present considerable challenges since the causal latent generative factors may usually have considerably lower dimensionality compared to their corresponding image data. Here, we conceptually generalize the notion of a DAG over the latent variables by allowing each latent variable to be represented by more than a single dimension. This is known as a C-DAG, and preserves relevant properties of the causal diagram [83]. Therefore, we specify a causal baseline distribution as Sec. 5.1.2 with the difference that now we assign 8 dimensions to each V_1, V_2, V_3, V_4. We then maximize the log-likelihood on the relevant embeddings per image in our dataset as in Fig. 19.

We propose a multivariate generalization of the causal base distribution compared to the model used in simulations of Section 5.1.2. We leverage one baseline distribution per sampled dataset, $(p_\theta^k)_{k\in[m]}$ over the base noise variables $\widehat{\mathbf{V}} \in \mathbb{R}^d$. In this example, we have $m = 4$ distributions and $d = 32$ dimensions. The conditional density of any variable is given by:



(a) (b) (c)

Figure 20: Correlation of color-digit (a) and color-bar (b) representations w.r.t. ground-truth latent generative factors. Note: the MNIST dataset does not provide access to the ground-truth "style" (V_4) per image, and thus there is no correlation w.r.t. V_4.

$$p_\theta^{(k)}(\widehat{\mathbf{v}}_\mathbf{i}|\widehat{\mathbf{Pa}}_i^+) = \mathcal{N}\left(\boldsymbol{\mu}_\theta(\widehat{\mathbf{Pa}}_i^+), \sigma_\theta(\widehat{\mathbf{Pa}}_i^+) \cdot \mathbf{I}\right) \tag{154}$$

where $\boldsymbol{\mu}_{i,\theta}$ and $\sigma_{i,\theta}$ are flow networks and $\mathbf{I}$ is an identity matrix. If a change-in-domain or an intervention is applied, the corresponding mean and variance are replaced by another transformation. When a hard intervention is applied, we let: where the parameters are replaced by their corresponding counterparts if there is a change-in-domain or an intervention applied. When a hard intervention is applied, we let:

$$p_\theta^k(\widehat{\mathbf{v}}_i) = \mathcal{N}(\hat{\boldsymbol{\mu}}_\mathbf{i}, \hat{\sigma}_i \mathbf{I}) \tag{155}$$

where $\hat{\boldsymbol{\mu}}_i$ and $\hat{\sigma}_i$ are uniformly sampled.

5.2.3 Evaluation Results

Taking the input graph assumption Fig. 17(a) and the interventional target sets, CRID will output the CDM (for $\widehat{V}_2$ and $\widehat{V}_3$) in Fig. 20(a). To illustrate, V_2 is ID w.r.t $\{V_1, V_3, V_4\}$ and V_3 is ID w.r.t $\{V_1, V_4\}$. The MCCs between the learned $\widehat{V}_2, \widehat{V}_3$ in VAE-NF proxy model and the ground truth $\mathbf{V}$ are shown in Fig. 20. To illustrate, the $MCC(\widehat{V}_2, V_1)$, $MCC(\widehat{V}_2, V_3)$ is lower than $MCC(\widehat{V}_2, V_2)$ and $MCC(\widehat{V}_3, V_1)$ is lower than $MCC(\widehat{V}_3, V_2)$. This implies V_2 is ID w.r.t $\{V_1, V_3\}$ and V_3 is ID w.r.t. $\{V_1\}$, which is aligned the CDM output from CRID. No correlation w.r.t. V_4 is shown since ground-truth "style" (V_4) is not provided in the MNIST dataset.

Next, we demonstrate qualitatively that the generalized disentanglement proposed in this work is valuable for downstream tasks, such as counterfactual image editing [27]. Specifically, we use our learned proxy model to generate initial images and perform interventions on learned representations $\widehat{\mathbf{V}}$ to edit images. We generate initial image samples from observational distribution of Π_1, and then perturb the relevant representations with random Gaussian noise to edit the image. This is done for the color of the bar, color of the digit, and the digit representations. If the learned $\widehat{\mathbf{V}}$ satisfy the CRID output disentanglement,

1. editing the color of the digit ($\sigma_{\widehat{V}_2}$) should keep the original digit and writing style but may change the color of the bar since V_2 has a causal effect on V_3.
2. editing the color of the bar ($\sigma_{\widehat{V}_3}$) should keep the original digit and writing style but may change the color of the digit since V_3 is not disentangled with V_2.
3. editing digit ($\sigma_{\widehat{V}_1}$) may change all variables since no disentanglement of V_1 is claimed by CRID.

The editing results are shown in Fig. 21. All editing results are aligned with the CDM output as expected, which are illustrated above. Specifically, Fig. 21(a) shows the learned VAE-NF model can change the color of the bar without arbitrarily changing the digit, or writing style. Fig. 21(b) shows the learned VAE-NF model can change the color of the digit without arbitrarily changing the digit, or writing style. Finally, Fig. 21(c) shows the learned VAE-NF model did not learn a disentangled

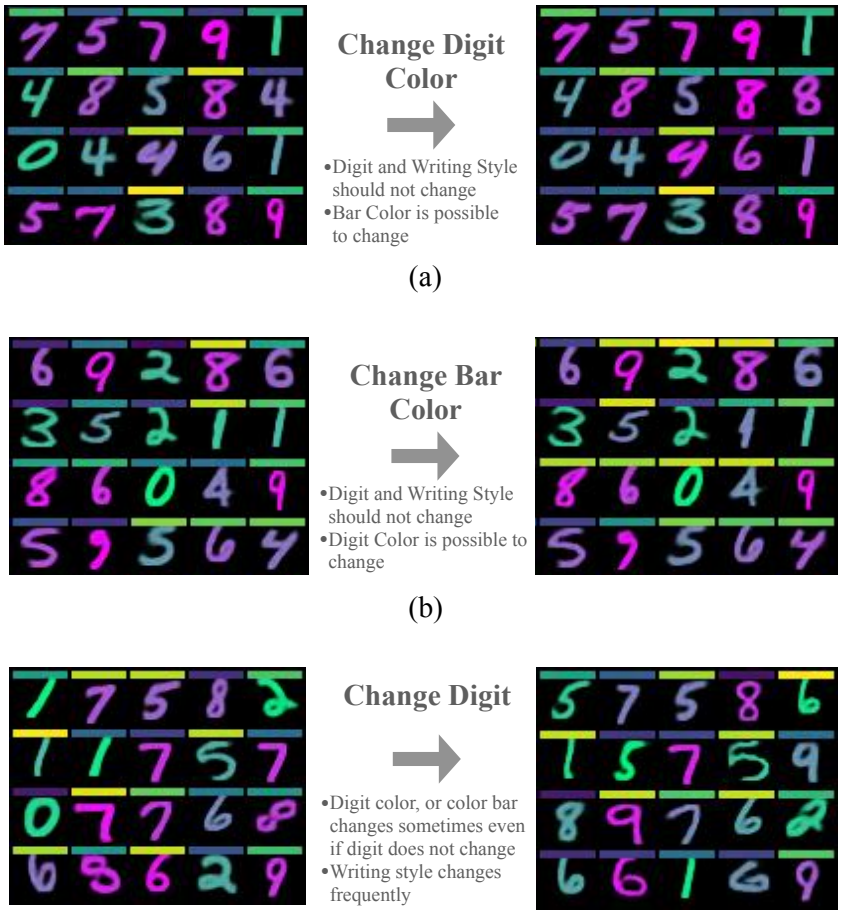


Figure 21: Editing the image using the learned representations. The representation of the color of the digit (a), color of the bar (b), and the digit (c) is perturbed. Only (a) and (b) show robust editing due to the learned representation being relatively disentangled as predicted by CRID.

representation for "digit". When perturbing the representation for digit, sometimes the digit does not change, while the color of the bar, color of the digit, or the writing style changes. This experiment also demonstrates one usage of CRID. Before training a model that is potentially computationally and time-intensive, one can leverage CRID to determine if their input data and input assumptions are sufficient for learning a relevant disentangled representation for their downstream task.

6 Conclusions

This work develops graphical conditions and a practical identification algorithm for determining which latent variables are disentangleable from a given set. The algorithm takes as input assumptions in the form of a LSD and distributions from heterogeneous domains. This brings us one step closer to building robust AI that can causally reason over high-level concepts when only given low-level data, such as images, video, or text.

Acknowledgements

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation. AL was supported by the NSF Computing Innovation Fellowship (#2127309).

References

[1] Judea Pearl. *Causality: Models, reasoning, and inference.* 2nd. Cambridge University Press, 2009.

[2] J. Pearl and D. Mackenzie. *The book of why : the new science of cause and effect*. Pages: 418. 2019.

[3] E. Bareinboim, J. D. Correa, D. Ibeling, and T. Icard. "On Pearl's Hierarchy and the Foundations of Causal Inference." In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. Vol. 36. New York, NY, USA: Association for Computing Machinery, 2022, pp. 507–556.

[4] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. *Towards Causal Representation Learning*. arXiv:2102.11107 [cs]. 2021.

[5] Y. Bengio, A. Courville, and P. Vincent. "Representation Learning: A Review and New Perspectives." In: *Arxiv* (2012).

[6] A. Hyvärinen and E. Oja. "Independent component analysis: algorithms and applications." In: *Neural networks* 13.4-5 (2000), pp. 411–430.

[7] A. Hyvarinen, H. Sasaki, and R. E. Turner. *Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning*. arXiv:1805.08651 [cs, stat]. 2019.

[8] A. Hyvärinen, I. Khemakhem, and H. Morioka. "Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning." In: *Patterns* 4.10 (2023), p. 100844.

[9] A. Hyvärinen and P. Pajunen. "Nonlinear independent component analysis: Existence and uniqueness results." In: *Neural networks* 12.3 (1999), pp. 429–439.

[10] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. "Variational autoencoders and nonlinear ica: A unifying framework." In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2207–2217.

[11] C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. "Linear Causal Disentanglement via Interventions." en. In: *Proceedings of the 40th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, 2023, pp. 32540–32560.

[12] K. Ahuja, J. S. Hartford, and Y. Bengio. "Weakly supervised representation learning with sparse perturbations." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 15516–15528.

[13] B. Varici, E. Acarturk, K. Shanmugam, A. Kumar, and A. Tajer. "Score-based causal representation learning with interventions." In: *arXiv preprint arXiv:2301.08230* (2023).

[14] K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. *Interventional Causal Representation Learning*. 2024.

[15] L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf. "The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica." In: *Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 217–227.

[16] L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. "Independent mechanism analysis, a new concept?" In: *Advances in neural information processing systems* 34 (2021), pp. 28233–28248.

[17] S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. "Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA." In: *First Conference on Causal Learning and Reasoning*. 2021.

[18] J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. *Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style*. 2022.

[19] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. "Weakly-supervised disentanglement without compromises." In: *International conference on machine learning*. PMLR. 2020, pp. 6348–6359.

[20] J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen. "Weakly supervised causal representation learning." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 38319–38331.

[21] L. Wendong, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. *Causal Component Analysis*. arXiv:2305.17225 [cs, stat]. 2023.

[22] K. Zhang, S. Xie, I. Ng, and Y. Zheng. *Causal Representation Learning from Multiple Distributions: A General Setting*. arXiv:2402.05052 [cs, stat]. 2024.

[23] S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. "Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA." In: *Conference on Causal Learning and Reasoning*. PMLR. 2022, pp. 428–484.

[24] S. Lachapelle, P. R. López, Y. Sharma, K. Everett, R. L. Priol, A. Lacoste, and S. Lacoste-Julien. "Nonparametric Partial Disentanglement via Mechanism Sparsity: Sparse Actions, Interventions and Sparse Temporal Dependencies." In: *arXiv preprint arXiv:2401.04890* (2024).

[25] Z. Jin, J. Liu, Z. Lyu, S. Poff, M. Sachan, R. Mihalcea, M. Diab, and B. Schölkopf. *Can Large Language Models Infer Causation from Correlation?* arXiv:2306.05836 [cs]. 2023.

[26] M. Zečević, M. Willig, D. S. Dhami, and K. Kersting. "Causal Parrots: Large Language Models May Talk Causality But Are Not Causal." en. In: *Transactions on Machine Learning Research* (2023).

[27] Y. Pan and E. Bareinboim. "Counterfactual Image Editing." In: *arXiv preprint arXiv:2403.09683* (2024).

[28] P. C. Austin. "An introduction to propensity score methods for reducing the effects of confounding in observational studies." In: *Multivariate Behavioral Research* 46.3 (2011), pp. 399–424.

[29] M. Brookhart, T. Stürmer, R. Glynn, J. Rassen, and S. Schneeweiss. "Confounding control in healthcare database research: challenges and potential approaches." In: *Medical care* 48.6 0 (2010), S114–S120.

[30] C. Wachinger, B. G. Becker, A. Rieckmann, and S. Pölsterl. "Quantifying Confounding Bias in Neuroimaging Datasets with Causal Inference." en. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan. Cham: Springer International Publishing, 2019, pp. 484–492.

[31] F. Mahmood, R. Chen, and N. J. Durr. "Unsupervised Reverse Domain Adaptation for Synthetic Medical Images via Adversarial Training." In: *IEEE Transactions on Medical Imaging* 37.12 (2018), pp. 2572–2581.

[32] A. Li, A. Jaber, and E. Bareinboim. "Causal discovery from observational and interventional data across multiple environments." In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[33] E. Bareinboim and J. Pearl. "Transportability of Causal Effects: Completeness Results." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 26.1 (2012), pp. 698–704.

[34] J. Pearl and E. Bareinboim. "Transportability across studies: A formal approach." In: (2018).

[35] E. Bareinboim and J. Pearl. "Meta-Transportability of Causal Effects: A Formal Approach." en. In: *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*. ISSN: 1938-7228. PMLR, 2013, pp. 135–143.

[36] E. Bareinboim and J. Pearl. "Causal inference and the data-fusion problem." In: *Proceedings of the National Academy of Sciences* 113.27 (2016). Publisher: National Academy of Sciences, pp. 7345–7352.

[37] P. Hünermund and E. Bareinboim. *Causal Inference and Data Fusion in Econometrics*. arXiv:1912.09104 [econ]. 2023.

[38] A. Li, C. Huynh, Z. Fitzgerald, I. Cajigas, D. Brusko, J. Jagid, A. O. Claudio, A. M. Kanner, J. Hopp, S. Chen, J. Haagensen, E. Johnson, W. Anderson, N. Crone, S. Inati, K. A. Zaghloul, J. Bulacio, J. Gonzalez-Martinez, and S. V. Sarma. "Neural fragility as an EEG marker of the seizure onset zone." en. In: *Nature Neuroscience* 24.10 (2021). Number: 10 Publisher: Nature Publishing Group, pp. 1465–1474.

[39] A. Li, P. Myers, N. Warsi, K. M. Gunnarsdottir, S. Kim, V. Jirsa, A. Ochi, H. Otusbo, G. M. Ibrahim, and S. V. Sarma. *Neural Fragility of the Intracranial EEG Network Decreases after Surgical Resection of the Epileptogenic Zone*. en. Pages: 2021.07.07.21259385. 2022.

[40] A. Li, B. Chennuri, S. Subramanian, R. Yaffe, S. Gliske, W. Stacey, R. Norton, A. Jordan, K. Zaghloul, S. Inati, S. Agrawal, J. Haagensen, J. Hopp, C. Atallah, E. Johnson, N. Crone, W. Anderson, Z. Fitzgerald, J. Bulacio, J. Gale, S. Sarma, and J. Gonzalez-Martinez. "Using network analysis to localize the epileptogenic zone from invasive EEG recordings in intractable focal epilepsy." In: *Network Neuroscience* 2.2 (2017).

[41] J. M. Bernabei, A. Li, A. Y. Revell, R. J. Smith, K. M. Gunnarsdottir, I. Z. Ong, K. A. Davis, N. Sinha, S. Sarma, and B. Litt. "Quantitative approaches to guide epilepsy surgery from intracranial EEG." In: *Brain* (2023), awad007.

[42] K. M. Gunnarsdottir, A. Li, R. J. Smith, J.-Y. Kang, A. Korzeniewska, N. E. Crone, A. G. Rouse, J. J. Cheng, M. J. Kinsman, P. Landazuri, U. Uysal, C. M. Ulloa, N. Cameron, I. Cajigas, J. Jagid, A. Kanner, T. Elarjani, M. M. Bicchi, S. Inati, K. A. Zaghloul, V. L. Boerwinkle, S. Wyckoff, N. Barot, J. Gonzalez-Martinez, and S. V. Sarma. "Source-sink connectivity: a novel interictal EEG marker for seizure localization." In: *Brain* 145.11 (2022), pp. 3901–3915.

[43] L. Nobili, B. Frauscher, S. Eriksson, S. Gibbs, H. Peter, I. Lambert, R. Manni, L. Peter-Derex, P. Proserpio, F. Provini, A. Weerd, and L. Parrino. "Sleep and epilepsy: A snapshot of knowledge and future research lines." In: *Journal of Sleep Research* 31 (2022).

[44] A. Bagshaw, J. Jacobs, P. LeVan, F. Dubeau, and J. Gotman. "Effect of sleep stage on interictal high-frequency oscillations recorded from depth macroelectrodes in patients with focal epilepsy." In: *Epilepsia* 50 (2008), pp. 617–28.

[45] S. Gibbs, P. Proserpio, M. Terzaghi, A. Pigorini, S. Sarasso, G. Russo, L. Tassi, and L. Nobili. "Sleep-related epileptic behaviors and non-REM-related parasomnias: Insights from stereo-EEG." In: *Sleep Medicine Reviews* 63 (2015).

[46] P. Greene, A. Li, J. Gonzalez-Martinez, and S. Sarma. "Classification of Stereo-EEG Contacts in White Matter vs. Gray Matter Using Recorded Activity." In: *Frontiers in Neurology* 11 (2021).

[47] A. A. Borbély, F. Baumann, D. Brandeis, I. Strauch, and D. Lehmann. "Sleep deprivation: Effect on sleep stages and EEG power density in man." In: *Electroencephalography and Clinical Neurophysiology* 51.5 (1981), pp. 483–493.

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is All you Need." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[49] R. Bommasani et al. *On the Opportunities and Risks of Foundation Models*. 2022.

[50] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. *Language Models are Few-Shot Learners*. 2020.

[51] J. von Kügelgen, M. Besserve, L. Wendong, L. Gresele, A. Kekić, E. Bareinboim, D. M. Blei, and B. Schölkopf. *Nonparametric Identifiability of Causal Representations from Unknown Interventions*. arXiv:2306.00542 [cs, stat]. 2023.

[52] K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. "Interventional causal representation learning." In: *International conference on machine learning*. PMLR. 2023, pp. 372–407.

[53] D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. *Multi-View Causal Representation Learning with Partial Observability*. 2024.

[54] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. "CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models." In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. 2021, pp. 9588–9597.

[55] J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler. "Identifiability guarantees for causal disentanglement from soft interventions." In: *Advances in Neural Information Processing Systems* 36 (2024).

[56] A. Li, J. Feitelberg, A. P. Saini, R. Höchenberger, and M. Scheltienne. "MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python." In: *Journal of Open Source Software* 7.76 (2022), p. 4484.

[57] J. Tian and J. Pearl. "On the testable implications of causal models with hidden variables." In: *arXiv preprint arXiv:1301.0608* (2012).

[58] J. Correa and E. Bareinboim. "A Calculus for Stochastic Interventions:Causal Effect Identification and Surrogate Experiments." en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.06 (2020). Number: 06, pp. 10093–10100.

[59] J. Correa and E. Bareinboim. "General Transportability of Soft Interventions: Completeness Results." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 10902–10912.

[60] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. "Challenging common assumptions in the unsupervised learning of disentangled representations." In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[61] P. R. Rosenbaum and D. B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." In: *Biometrika* 70.1 (1983), pp. 41–55.

[62] J. Pearl. "Causal Diagrams for Empirical Research." In: *Biometrika* 82.4 (1995). Publisher: [Oxford University Press, Biometrika Trust], pp. 669–688.

[63] J. Pearl and E. Bareinboim. "Transportability of Causal and Statistical Relations: A Formal Approach." en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 25.1 (2011). Number: 1, pp. 247–254.

[64] S. Lee, J. D. Correa, and E. Bareinboim. "General identifiability with arbitrary surrogate experiments." In: 2019.

[65] R. Perry, J. von Kügelgen, and B. Schölkopf. *Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis*. arXiv:2206.02013 [cs, stat]. 2022.

[66] B. Huang, K. Zhang, M. Gong, and C. Glymour. "Causal Discovery and Forecasting in Nonstationary Environments with State-Space Models." en. In: *Proceedings of the 36th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, 2019, pp. 2901–2910.

[67] B. Huang, C. J. H. Low, F. Xie, C. Glymour, and K. Zhang. "Latent hierarchical causal structure discovery with rank constraints." In: *arXiv preprint arXiv:2210.01798* (2022).

[68] J. Peters, P. Bühlmann, and N. Meinshausen. *Causal inference using invariant prediction: identification and confidence intervals*. arXiv:1501.01332 [stat]. 2015.

[69] J. M. Mooij, S. Magliacane, and T. Claassen. "Joint causal inference from multiple contexts." In: *The Journal of Machine Learning Research* 21.1 (2020), 99:3919–99:4026.

[70] K. Xia, K.-Z. L. Lee Bloomberg, Y. Bengio, and E. Bareinboim. "The Causal-Neural Connection: Expressiveness, Learnability, and Inference." In: (2021).

[71] W. Kaplan. *Advanced calculus*. Pearson Education India, 1952.

[72] A. Hyttinen, F. Eberhardt, and M. Järvisalo. "Constraint-based causal discovery: conflict resolution with answer set programming." In: *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*. UAI'14. Arlington, Virginia, USA: AUAI Press, 2014, pp. 340–349.

[73] Y. Zheng, I. Ng, and K. Zhang. "On the identifiability of nonlinear ICA: Sparsity and beyond." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 16411–16422.

[74] X. Yang, Y. Wang, J. Sun, X. Zhang, S. Zhang, Z. Li, and J. Yan. "Nonlinear ICA Using Volume-Preserving Transformations." In: *International Conference on Learning Representations*. 2022.

[75] D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[76] L. Gresele, G. Fissore, A. Javaloy, B. Schölkopf, and A. Hyvärinen. *Relative gradient optimization of the Jacobian term in unsupervised deep learning*. 2020.

[77] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. *Normalizing Flows for Probabilistic Modeling and Inference*. 2021.

[78] C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. *Neural Spline Flows*. 2019.

[79] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017.

[80] Y. Lecun, L. Bottou, Y. Bengio, and P. Ha. "LeNet." In: *Proceedings of the IEEE* November (1998), pp. 1–46.

[81] J. D. Hunter. "Matplotlib: A 2D graphics environment." In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95.

[82] D. P. Kingma and M. Welling. "Auto-encoding variational bayes." In: *arXiv preprint arXiv:1312.6114* (2013).

[83] T. V. Anand, A. H. Ribeiro, J. Tian, and E. Bareinboim. "Causal effect identification in cluster dags." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 10. 2023, pp. 12172–12179.

[84] M. Okamoto. "Distinctness of the eigenvalues of a quadratic form in a multivariate sample." In: *The Annals of Statistics* (1973), pp. 763–765.

[85] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. en. Google-Books-ID: AvNID7LyMusC. Morgan Kaufmann, 1988.

[86] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. "Independence properties of directed markov fields." en. In: *Networks* 20.5 (1990). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/net.3230200503, pp. 491–505.

[87] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. "beta-vae: Learning basic visual concepts with a constrained variational framework." In: *ICLR (Poster)* 3 (2017).

[88] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang. "Weakly Supervised Disentangled Generative Causal Representation Learning." In: *Journal of Machine Learning Research* 23 (2022), pp. 1–55.

[89] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan. "Causal protein-signaling networks derived from multiparameter single-cell data." eng. In: *Science (New York, N.Y.)* 308.5721 (2005), pp. 523–529.

[90] J. M. Robins, M. A. Hernán, and B. Brumback. "Marginal structural models and causal inference in epidemiology." eng. In: *Epidemiology (Cambridge, Mass.)* 11.5 (2000), pp. 550–560.

[91] J. Tian and J. Pearl. "A General Identification Condition for Causal Effects." In: *AAAI* (2002).

[92] I. Shpitser and J. Pearl. "Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models." In: *AAAI-Proceedings* (2006), pp. 1219–1226.

[93] M. Kocaoglu, K. Shanmugam, and E. Bareinboim. "Experimental Design for Learning Causal Graphs with Latent Variables." In: *Advances in Neural Information Processing Systems* 30 (2017).

[94] M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. "Characterization and Learning of Causal Graphs with Latent Variables from Soft Interventions." In: *Advances in Neural Information Processing Systems* 32 (2019).

[95] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. "Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning." In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 2020, pp. 9551–9561.

[96] K. Xia, Y. Pan, and E. Bareinboim. "Neural Causal Models for Counterfactual Identification and Estimation." In: *International Conference on Learning Representations*. 2022.

[97] A. Jaber, A. H. Ribeiro, J. Zhang, and E. Bareinboim. "Causal Identification under Markov equivalence: Calculus, Algorithm, and Completeness." In: *Advances in Neural Information Processing Systems*. 2022.

[98] A. Jaber, J. Zhang, and E. Bareinboim. "Causal Identification under Markov Equivalence: Completeness Results." In: (2019). Publisher: PMLR, pp. 2981–2989.

[99] I. Shpitser and J. Pearl. "Complete Identification Methods for the Causal Hierarchy." In: *Journal of Machine Learning Research* 9 (2008), pp. 1941–1979.

[100] J. Zhang, J. Tian, and E. Bareinboim. "Partial Counterfactual Identification from Observational and Experimental Data." In: (2021).

Appendix

Contents

A Background and Assumptions

A.1 Notations

Symbol	Description
$[d]$	$\{1, 2, \dots, d\}$
$\mathcal{M}$	An ASCM (Def. 2.1) describes the data generation process of d latent variables $\mathbf{V} \in \mathbb{R}^d$ and an observed high-dimensional mixture $\mathbf{X} \in \mathbb{R}^m$.
G	Latent Causal Diagram (LCG) over $\mathbf{V}$ induced by an $\mathcal{M}$
$\overline{\mathbf{Pa}}(\mathbf{V}), \overline{\mathbf{Pa}}_{\mathbf{V}}$	The union of parents of $\mathbf{V}$ and $\mathbf{V}$ itself
$\mathbf{C}(\mathbf{V})$	C-Component of $\mathbf{V}$ (Def.6.1).
$\boldsymbol{\mathcal{M}}$	A set of N ASCMs $\langle \mathcal{M}_1, \dots, \mathcal{M}_N \rangle$ (shared mixing function $f_{\mathbf{X}}$) relative to domains $\mathbf{\Pi} = \langle \Pi_1, \dots, \Pi_N \rangle$
G^S	Latent Selection Diagram (LSD, Def 2.2) induced by $\boldsymbol{\mathcal{M}}$
$\Sigma = \{\sigma^{(k)}\}_{k=1}^K$	A set of $K >= N$ interventions applied to $\boldsymbol{\mathcal{M}}$. Each intervention $\sigma^{(k)}$ can be idle, hard, or other soft interventions that do not alter the structure of G
$\mathbf{\Pi}^{\Sigma} = \{\Pi^{(k)}\}_{k=1}^K$	The corresponding domains of interventions Σ. $\sigma^{(k)}$ is applied in $\Pi^{(k)}$
$\mathbf{\Psi} = \{\mathbf{I}^{(k)}\}_{k=1}^K$	The collection of intervened target sets of the intervention collection Σ.
$\mathbf{I}^{(k)} = \{V_i^{\Pi^{(k)},\{b\},t}, \dots\}$	The intervention target set of $\sigma^{(k)}$
$\{b\}$	The mechanism of intervention. Default as interventions have different mechanisms if b is ignored. Also, the mechanism od different variables are different. The mechanism of $V_1^{\{1\}}$ is not equal to $V_2^{\{1\}}$.
t	Whether an intervention is hard or not. $t = do$ means it is hard. Default as not hard if t is ignored.
$do[\mathbf{I}^{(k)}]$	Variables that are perfectly intervened on in $\sigma^{(k)}$.
$\mathbf{\Psi}_{\mathbf{T}}$	The collection of intervention target sets that contain a hard intervention on $\mathbf{T}$.
$\mathcal{P} = \{P^{(k)}\}_{k=1}^K$	Set of distributions induced by $\boldsymbol{\mathcal{M}}$ resulting from collection of interventions Σ. $P^{(k)} = P^{\Pi^{(k)}}(\mathbf{X}; \sigma^{(k)})$
$\mathbf{Pa}^{\mathbf{T}+}(\mathbf{V}), \mathbf{P}a_{\mathbf{V}}^{\mathbf{T}+}$	Extended parents from factorization Eq. (85).
$\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$	Changed variable sets constructed in Proposition 2. For short, $\Delta\mathbf{V}$ or $\Delta\mathbf{V}^{(j),(k)}$ when index is needed.
$\tilde{\mathbf{V}}$	The C-Component of $\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$. The factor $P(v_i \mid \mathbf{p}a^{\mathbf{T}})$ for $V_i \in \mathbf{V} \backslash \tilde{\mathbf{V}}$ remains invariant in Eq.(104).
$\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S]$	$\Delta\mathbf{Q}$ set defined in Def. 3.1. Variables in $\Delta\mathbf{Q}$ set remains from Eq.(104) to Eq.(119). For short, $\Delta\mathbf{Q}$ or $\Delta\mathbf{Q}^{(j),(k)}$ when index is needed.
Canceled variables	The complement of $\Delta\mathbf{Q}$, which is $\mathbf{V} \backslash \Delta\mathbf{Q}$.

Figure S1: Table of Notations

A.2 Discussion on Assumptions

In this paper, we make a few key assumptions about interventions and the differences in domains. We leverage many similar assumptions to the setting proposed in the literature related to causal

representation learning, and handling of multiple domains and interventions [21, 22, 32, 51]. We discuss those assumptions and their implications here.

Assumption 4 (Soft interventions without altering the causal structure)**.** *Assume that interventions do not alter the causal diagram. That is for each intervention set in the tuple of interventions* $\mathbf{I} \in \mathbf{\Psi}$*, a soft intervention that is not hard does not remove, or add any edges to the graph.* □

This assumption precludes any soft interventions that modify the graphical structure of the causal diagram. This work does allow both hard (can also be called perfect) interventions that cut all incoming parent edges, and soft interventions that preserve all parent edges. However, more general interventions may arbitrarily change the parent set for any given node [59]. We do not consider such interventions, and leave this general case for future work. Note Assumption 4 does not mean that interventions cannot occur with the same mechanism across domains. For example, consider two hospitals Π^1 and Π^2. Treating epilepsy in each of these hospitals can have outcomes differ vastly due to the differences in domains [38, 39, 41]. This is represented graphically in G^S with $S^{1,2} \rightarrow$ outcome. However, if a neurologist that controls every aspect of his treatment procedure treats patients in both hospitals herself for the purposes of an experiment, then the outcomes will not differ in distribution. This is represented graphically as $S^{1,2} \not\rightarrow$ outcome with the S-node being removed from "outcome" variable. Thus if a pair of interventions occurring in different domains are deemed to have the same mechanism, then the S-node (if one is pointing to the intervened variable) is removed when comparing these two distributions.

Another assumption we make is that all interventions have known-target.

Assumption 5 (Known-target assumption)**.** *Assume for any* $\mathbf{I}^{(k)} \in \mathbf{\Psi}$*, all interventions occur with known-target.* □

That is, for each interventional distribution we have, we know the interventions that occurred and at which node(s) they occurred. This assumption allows us to reduce the permutation indeterminacy that would arise if we did not know the intervention targets. In this work, we also are not concerned with permutation indeterminacy for variables we do not necessarily intervene on because we will mostly be concerned with disentanglement wrt the intervened variables (see Appendix Section A.4). It would be interesting for future work to consider unknown intervention targets.

In Sec. 2, we discuss that each distribution resulting from an intervention is sufficiently distinct from another distribution Assumption 6. Here we formally define and illustrate what is "change sufficiently".

Assumption 6 (Changing Sufficiently)**.** *Consider a collection of ASCMs* $\boldsymbol{\mathcal{M}}$ *and a set of distribution* $\mathcal{P}$ *induced by* $\boldsymbol{\mathcal{M}}$ *from a collection of interventions* Σ*. Let the LSD induced by* $\boldsymbol{\mathcal{M}}$ *be* G^S*. Let* $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \dots, P^{(a_L)}\} \subseteq \mathcal{P}$ *be any collection of distributions such that* $\mathbf{T} = do[\mathbf{I}^{(a_0)}] \subseteq do[\mathbf{I}^{(a_l)}]$ *for* $l \in [L]$*, meaning for the baseline distribution all hard interventions must be exactly on* $\mathbf{T}$*, and all other distributions must at least contain* $\mathbf{T}$ *in their hard interventions. Let* $\mathbf{Q} = \bigcup_{l\in[L]} \Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(0)}, \mathbf{T}, G^S]$ *(Def. 3.1). It is assumed:*

1. *The probability density function of* $\mathbf{V}$ *is smooth and positive, i.e.* $p^{(a_l)}_{\mathbf{T}}(\mathbf{v})$ *is smooth and* $p^{(a_l)}_{\mathbf{T}}(\mathbf{v}) > 0$ *almost everywhere.*
2. *First-order discrepancy. If there exists* $\{a'_1, \dots, a'_{|\mathbf{Q}|}\} \subseteq \{a_1, \dots, a_L\}$ *such that* $\forall\ V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_q)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$*, then* $\{\boldsymbol{\omega}_1(\mathbf{v}, a_1), \boldsymbol{\omega}_1(\mathbf{v}, a_2), \dots, \boldsymbol{\omega}_1(\mathbf{v}, a_L)\}$ *are linearly independent, where*
$$\boldsymbol{\omega}_1(\mathbf{v}, a_l) = \left(\oplus \left(\frac{\partial \log p^{(a_l)}_{\mathbf{T}}(\mathbf{v}) - \log p^{(a_0)}_{\mathbf{T}}(\mathbf{v})}{\partial v_q} \right)_{V_q \in \mathbf{Q}} \right) \tag{156}$$
3. *Second-order discrepancy. Let a set* $\boldsymbol{\mathcal{E}}$ *consist of pairs of* (V_p, V_q) *such that* (V_p, V_q) *appears at least in one* $\Delta\mathbf{Q}$ *and* V_p *is connected with* V_q *conditioning on* $\mathbf{V}\backslash\{V_p, V_q\}$ *in* $G_{\overline{T}}(\mathbf{Q})$*. Namely,*
$$\begin{aligned} \boldsymbol{\mathcal{E}} = \{ & \boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid \\ & \textit{(i)}\ \exists a_l, \{V_p, V_q\} \in \Delta\mathbf{Q}^{(a_l),(a_0)}; \\ & \textit{(ii)}\ V_p \textit{ is d-connected to } V_q \textit{ conditioned on } \mathbf{V}^{tar}\backslash\{V_p, V_q\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\}, \end{aligned} \tag{157}$$

If there exists $\{a'_1, \ldots, a'_{2|\mathbf{Q}|+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \ldots, a_L\}$ *such that* $\forall \quad V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], V_q \in \Delta\mathbf{Q}^{(a'_{|Q|+i}),(a_0)}$ *and for all* $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2|\mathbf{Q}|+j}),(a_0)}$*, then* $\{\boldsymbol{\omega}_2(\mathbf{v}, a_1), \boldsymbol{\omega}_2(\mathbf{v}, a_2), \ldots, \boldsymbol{\omega}_2(\mathbf{v}, a_L)\}$ *are linearly independent, where*

$$\begin{aligned}\boldsymbol{\omega}_2(\mathbf{v}, a_l) = \Bigg(&\oplus \left(\frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q}\right)_{V_q \in \mathbf{Q}}, \\ &\oplus \left(\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q^2}\right)_{V_q \in \mathbf{Q}}, \\ &\oplus \left(\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_p v_q}\right)_{(V_p, V_q) \in \mathcal{E}(G_{\overline{\mathcal{T}}}(\mathbf{Q}))} \Bigg)\end{aligned} \tag{158}$$

□

At a high level, this assumption will be naturally satisfied if the ASCMs and interventions are randomly chosen and only will be violated if the probability density of $P^{(j)}$ and $P^{(k)}$ are fine-tuned to each other [51]. This kind of assumption is generally included in the causal representation learning literature, such as the "genericity" assumption [51], the "interventional discrepancy" assumption [21], and the "sufficient changes" assumption [10, 22].

To illustrate, the assumptions contain two linear independence constraints. Specifically, the first-order and second-order partial derivatives of the log discrepancy from $P^{(a_l)}$ to $P^{(a_0)}$ should be independent of each other. Specifically, The two conditions are made because of necessity, since the linear independence constraints can hold only if these conditions hold. The following example illustrates the necessity of the first-order condition:

Example A26. Consider $\Delta\mathbf{Q}$ obtained after comparisons as

$$\Delta\mathbf{Q}^{(1),(0)} = \{V_1\}, \Delta\mathbf{Q}^{(2),(0)} = \{V_1\}, \Delta\mathbf{Q}^{(1),(0)} = \{V_1, V_2, V_3\}, \tag{159}$$

Let $\mathbf{Q} = \{V_1, V_2, V_3\}$. We have

$$\frac{\log p_{\mathbf{T}}^{(1)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_2} = 0 \tag{160}$$

Since $V_2 \notin \Delta\mathbf{Q}^{(1),(0)}$. Similarly, we know

$$\begin{aligned}\boldsymbol{\omega}_1(v_1, v_2, v_3, 1) &= \left(\frac{\partial \log p_{\mathbf{T}}^{(1)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_1}, 0, 0\right) \\ \boldsymbol{\omega}_1(v_1, v_2, v_3, 2) &= \left(\frac{\partial \log p_{\mathbf{T}}^{(2)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_1}, 0, 0\right) \\ \boldsymbol{\omega}_1(v_1, v_2, v_3, 3) &= \left(\frac{\partial \log p_{\mathbf{T}}^{(3)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_1}, \frac{\partial \log p_{\mathbf{T}}^{(3)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_2}, \frac{\partial \log p_{\mathbf{T}}^{(3)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_3}\right)\end{aligned} \tag{161}$$

And this implies $\boldsymbol{\omega}_1(v_1, v_2, v_3, 1), \boldsymbol{\omega}_1(v_1, v_2, v_3, 2), \boldsymbol{\omega}_1(v_1, v_2, v_3, 3)$ are for sure not linearly independent. □

On the other perspective, violating these assumptions is like stating the probability densities are fine-tuned to each other [51]. Here we give an example of how this assumption can be violated.

Example A27 (Distributions do not change sufficiently)**.** Consider intervention targets

$$\boldsymbol{\Psi} = \{\mathbf{I}^{(1)} = \{\{\}^{\Pi_1}\}, \mathbf{I}^{(2)} = \{V_1^{\Pi_1,\{1\}}\}, \mathbf{I}^{(3)} = \{V_2^{\Pi_1,\{2\}}\}, \mathbf{I}^{(4)} = \{V_1^{\Pi_1,\{1\}}, V_2^{\Pi_1,\{2\}}\}\} \tag{162}$$

Choosing $\mathbf{I}^{(1)}$ as the baseline, $\mathbf{T} = \{\}$. The corresponding $\Delta\mathbf{Q}$ sets are $\{\{V_1\}, \{V_2\}, \{V_1, V_2\}\}$. Let $\mathbf{Q}$ be the union of $\Delta\mathbf{Q}$ sets, which is $\{V_1, V_2\}$. One can verify

$$\boldsymbol{\omega}_1(\mathbf{v}, 2) + \boldsymbol{\omega}_1(\mathbf{v}, 3) = \boldsymbol{\omega}_1(\mathbf{v}, 4) \tag{163}$$

since $\mathbf{I}^{(4)}$ is designed as a combination of $\mathbf{I}^{(2)}$ and $\mathbf{I}^{(3)}$. □

We provide the following Lemma to justify Assumption 6 formally.

Lemma 1. *Assumption 6 almost surely holds.* □

A.3 Domains vs Interventions

Example A28 (Example illustrating CRID with domains)**.** Consider the LSD shown in Fig. 11(a). We have the following distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}\} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X})$ from interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}\} = \{\{\}, \{\}\}$. Applying CRID algorithm, we can determine that V_1 is ID wrt V_2 and V_3. □

This example illustrates that observational data in two domains can help disentangle a root variable (V_1) from all its descendants.

Example A29 (Example illustrating CRID with interventions across domains with different mechanisms)**.** Consider the LSD shown in Fig. 11(a). We have the following distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}\} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X})$ from interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}\}$ with targets $\mathbf{\Psi} = \{\{V_2\}^{\Pi_1}, \{\}^{\Pi_2}$. Applying CRID algorithm, we can determine that V_2 and V_1 is ID wrt V_3. □

This example demonstrates that when comparing observational data from domain Π_1 with interventional data from a different domain Π_2, the only invariant factor is $P(V_3|V_2)$, with $\Delta V[\{\{V_2\}^{\Pi_1}, \{\}^{\Pi_2}, G^S] = \{V_1, V_2\}$. The canceled variable is V_3, and thus we achieve our identifiability result.

Example A30 (Example illustrating CRID with interventions across domains with the same mechanisms)**.** Consider the LSD shown in Fig. 11(a). We have the following distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}\}$ from interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}\}$ with targets $\mathbf{\Psi} = \{\{V_1^{[i]}, V_2\}^{\Pi_1}, \{\}^{\Pi_2}, \{V_1^{[i]}\}^{\Pi_2}$. Applying CRID algorithm, we can determine that V_1 is ID wrt $\{V_2, V_3\}$, and V_2 is ID wrt $\{V_3\}$. □

Even with an intervention that changes both V_1, V_2. When comparing the distributions $P^{(1)}$ and $P^{(3)}$, the $P(V_1)$ term becomes an invariant factor because the intervention has the same mechanism. This removes the possible difference encoded by the S-node on V_1 between domains Π^1, Π^2.

These examples further demonstrates the importance of distinguishing domains and interventions because a difference in mechanism is present when comparing all distributions between a pair of domains, $\Pi_i \neq \Pi_j$. This in principle, results in additional variables in the $\Delta\mathbf{Q}$ set. However, interventions may allow us to remove variables from this set by increasing the number of invariant factors.

A.4 Permutation Indeterminancy

In the context of causal representation learning, permutation indeterminacy is a significant challenge that arises when attempting to identify latent variables from observed data. This phenomenon occurs when the ordering of latent variables is not uniquely determined, leading to multiple equivalent representations (i.e. permutations of the latent variables) that can explain the observed data equally well.

In the earliest results of disentangled representation learning, linear ICA was known to be identifiable only up to permutation and scaling indeterminacies [6]. Permutation indeterminacy is still present in nonlinear ICA [7], since the independent components may be permuted arbitrarily.

Interestingly, when generalizing the problem to the Markovian setting where latent variables have causal structure (i.e. edges in a causal graph), permutation indeterminacy can be reduced to a graph isomorphism in certain cases. That is, latent variables are exchangeable with other latent variables that preserve the topological ordering of the latent causal graph (rather than permuted with any arbitrary latent variable) [13, 22, 51]. When the interventions occur with known targets on the latent space, and intervention occurs uniquely on every latent variable, then there is no permutation indeterminacy [21].

In this work, we assume intervention targets are known, but do not necessarily occur on all latent variables, and they may occur on multiple variables at once. For variables that are intervened on uniquely (i.e. one intervention applied on only that variable), there is no permutation ambiguity. For variables that are intervened on in groups, or not intervened on at all, there still exists permutation ambiguity:

1. (Grouped variables) These variables are all intervened on in the same group. In the context of our paper, these variables are consistently in the same $\Delta\mathbf{Q}$ set. For example, consider the following LCG $V_1 \to V_2 \leftarrow V_3$. If we have distributions arising only from interventions on $\{V_1, V_3\}$ and the observational distribution, and assume the learned representation is fully disentangled, then the learned representation still has a permutation indeterminacy wrt $\{V_1, V_3\}$. That is, $\hat{V}_1$ could be the representation for V_1, or V_3 and similarly for $\hat{V}_3$ (See why permutation can hold for details in Example A35).
2. (Non-intervened variables) These variables do not contain any interventions. Then there is still permutation ambiguity among these variables. However, instead of a graph isomorphism ambiguity, these variables form a subgraph isomorphism problem because there may be other variables that change across distributions (i.e. via interventions, or changes in domains), which are not permutable with respect to these invariant variables.

Specifically, the identifiability we talk about (Def. 2.3) is considered after a subgraph isomorphism permutation. For example, in the collider example setting where permutation can happen between V_1 and V_3. The "V_1 is ID w.r.t $\{V_2, V_3\}$" should implies there exists a function τ such that $\pi(\mathbf{V})[V_1] = \tau(\pi(\mathbf{V})[V_1])$, where $\pi(\mathbf{V})[V_i]$ means variable V_i after the permutation on $\mathbf{V}$ and π denotes a permutation only in this text. In our paper, we are primarily concerned with disentanglement and determining if the learned representation is disentangled in some general sense, and the permutation part is out of our scope.

B CRID Algorithm Details

Here, we provide additional pseudocode for the CRID Alg. 1.

First, the following algorithm illustrates how to initialize a fully connected bipartite graph $G_{\mathbf{V},\widehat{\mathbf{V}}}$. In the initial $G_{\mathbf{V},\widehat{\mathbf{V}}}$, the true underlying factors $\mathbf{V}$ points to representations each $\widehat{V}_i \in \widehat{\mathbf{V}}$, which means each variable $V_i \in \mathbf{V}$ is entangled with all other variables.

Algorithm F.2 FullyConnectedBipartiteGraph: Initialization step - Initialize a fully connected bipartite graph.

Input: $\mathbf{V}, \widehat{\mathbf{V}}$
Output: $G_{\mathbf{V},\widehat{\mathbf{V}}}$
1: Initialize an empty graph $G_{\mathbf{V},\widehat{\mathbf{V}}}$
2: **for** V_i in $\mathbf{V}$ **do**
3: **for** V_j in $\widehat{\mathbf{V}}$ **do**
4: Add edge (V_i, V_j) to $G_{\mathbf{V},\widehat{\mathbf{V}}}$

Then, after constructing Q from comparisons of distributions, the Alg. F.3 illustrates the details to check whether $\mathbf{V}\backslash\mathbf{Q}$ can be disentangled from $\mathbf{Q}$ according to Proposition 3. To illustrate, each variable $Z \in \mathbf{V}\backslash\mathbf{Q}$ is checked one by one. The variables that have already been disentangled from Z are collected in the list $\mathbf{Mem}$ through procedure **CheckMemoize**. Next, check if there is a sub-collection of $\mathcal{Q}$ that satisfy the [1-3] conditions in Proposition 3. The checking procedure is shown in Alg.F.5. If conditions are satisfied the edges from Z to $\widehat{\mathbf{Q}}$ are removed to demonstrate disentanglement. Based on the Lemma 2, the condition [3] in Prop. 3 can be reduced to a weaker condition [4] leveraging existing disentanglements in CDM.

Lemma 2. *Consider variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$ and $Z \in \mathbf{V}\backslash\mathbf{V}^{tar}$. Suppose $\mathbf{Mem} = \{V_j \in \mathbf{V}^{tar} \mid V_j$ is ID w.r.t. $Z\}$. Consider, $\mathcal{P}_{\mathbf{T}}$ and its corresponding intervention targets that hold conditions [1-2] in Prop. 3. If the new version of the condition [3] is also satisfied:*

[4] there exists $\{a'_1, \ldots, a'_{|\mathbf{V}^{tar}|}\} \subseteq \{a_1, \ldots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}\backslash\mathbf{Mem}, V_i^{tar} \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_i)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$.

then $\mathbf{V}^{tar}$ is ID w.r.t Z. □

To illustrate, the above lemma indicates not all variables in $\mathbf{V}^{tar}$ needed to be covered uniquely. Variables that have been already disentangled (in $\mathbf{Mem}$) do not need to be considered.

Example A31. Consider the LSD G^S and intervention targets $\mathbf{I}^{(1)} = \{\}$ and $\mathbf{I}^{(4)} = \{V_2^{\Pi_1, do}\}$. Comparing $\mathbf{I}^{(4)}$ and $\mathbf{I}^{(1)}$ taking $\mathbf{T} = \{\}$, $\Delta\mathbf{Q} = \{V_1, V_2\}$. Based on Prop. 3, we cannot get V_2 is ID w.r.t V_3 since to cover V_1 and V_2 separately, at least two $\Delta\mathbf{Q}$ sets are needed.

Now assume it is known that V_1 is ID w.r.t. V_3, namely $\mathbf{Mem} = \{V_1\}$. $\Delta\mathbf{Q}$ sets only need to cover V_2 and does not need to cover V_1 from condition [4] in Lemma 2. Then V_2 is ID w.r.t. V_3. □

Algorithm F.4 CheckMemoize: Memoization step - The variables in $\mathbf{Q}$ is ID w.r.t Z already.

Input: $G_{V,\widehat{V}}$, Z, $\mathbf{Q}$
Output: $\mathbf{Mem}$
1: $\mathbf{Mem} \leftarrow \{\}$
2: **for** all $\widehat{V} \in \mathbf{Q}$ **do**
3: **if** $Z \to \widehat{V} \notin G_{\mathbf{V},\widehat{\mathbf{V}}}$ **then**
4: $\mathbf{Mem}.append(V)$
5: **return** $\mathbf{Mem}$

Algorithm F.3 DisΔQfromCancel - Check whether canceled variables $\mathbf{V}\backslash\mathbf{Q}$ can be disentangled from the LQ factors $\mathbf{Q}$. $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is the current bipartite graph; $G_{\overline{\mathbf{T}}}$ is the LCG after the hard intervention on $\mathbf{T}$; $\mathbf{\Psi_X}$ is the intervened sets that contains hard interventions on $\mathbf{X}$; $\mathbf{I} \in \mathbf{\Psi_T}$ is the chosen baseline distribution; $\mathcal{Q}$ is the collection of $\Delta\mathbf{Q}$ sets after comparing intervention targets $\mathbf{J} \in \mathbf{\Psi_X}\backslash\mathbf{I}$ with the baseline.

Input: $\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{X}}}, \mathbf{\Psi_X}, \mathbf{I}, \mathcal{Q}$
Output: $G_{\mathbf{V},\widehat{\mathbf{V}}}$

for all $Z \in \mathbf{V}\backslash\mathbf{Q}$ **do**
  $\mathbf{Mem} \leftarrow CheckMemoize(G_{\mathbf{V},\widehat{\mathbf{V}}}, Z, \mathbf{Q})$ ▷ Variables in $\mathbf{Q}$ has been already ID w.r.t. Z.
  if $CheckConsition3(\mathcal{Q}, \mathbf{Q}, \mathbf{Mem})$ **then** ▷ Check conditions in Prop. 3 and Lem. 3
    remove edge $Z \to \widehat{\mathbf{Q}}$ in $G_{\mathbf{V},\widehat{\mathbf{V}}}$
return $G_{\mathbf{V},\widehat{\mathbf{V}}}$

Algorithm F.5 CheckCondition3: Check conditions in Proposition 3 and Lemma 2. $\mathcal{Q}$ is the collection of $\Delta\mathbf{Q}$ sets; $\mathbf{Q}$ are target variables;$\mathbf{Mem}$ are variables in $\mathbf{Q}$ have already been disentangled.

Input: $\mathcal{Q}$, $\mathbf{Q}$, $\mathbf{Mem}$
Output: $True$ or $False$

$\mathbf{L} \leftarrow \{\}$
for $\mathbf{Q}_k \in \mathcal{Q}$ **do**
  if $\mathbf{Q}_k \subseteq \mathbf{Q}$ **then**
    $\mathbf{L}.append(\mathbf{Q}_k)$
$\mathbf{Q}^{re} = \{Q_1, \ldots, Q_{d'}\} \leftarrow \mathbf{Q}\backslash\mathbf{Mem}$, $d' \leftarrow |\mathbf{Q}^{re}|$
if $Q_1 \in \mathbf{L}_1, Q_2 \in \mathbf{L}_2, \ldots, Q_{d'} \in \mathbf{L}_{d'}$ after a permutation of $\mathbf{L}$ **then**
  return $True$
return $False$

Algorithm F.6 DisWithinΔQ - Check the disentanglement of variables within $\mathbf{Q}$. $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is the current bipartite graph; $G_{\overline{\mathbf{T}}}$ is the LCG after the hard intervention on $\mathbf{T}$; $\mathbf{\Psi_T}$ is the intervened sets that contains hard interventions on $\mathbf{X}$; $\mathbf{I} \in \mathbf{\Psi_T}$ is the chosen baseline distribution; $\mathcal{Q}$ is the collection of $\Delta\mathbf{Q}$ sets after comparing intervention targets $\mathbf{J} \in \mathbf{\Psi_X}\backslash\mathbf{I}$ with the baseline.

Input: $\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}}, \mathbf{\Psi_T}, \mathbf{I}, \mathcal{Q}$
Output: $G_{\mathbf{V},\widehat{\mathbf{V}}}$

for for all pair $V_i, V_j \in \mathbf{Q}$ **do**
  if $V_i \perp V_j \mid \mathbf{Q}\backslash\{V_i, V_j\}$ **then**
    $\mathbf{Mem}_i \leftarrow CheckMemoize(G_{\mathbf{V},\widehat{\mathbf{V}}}, V_i, \mathbf{Q})$ ▷ Variables in $\mathbf{Q}$ is ID w.r.t V_i already.
    $\mathbf{Mem}_j \leftarrow CheckMemoize(G_{\mathbf{V},\widehat{\mathbf{V}}}, V_j, \mathbf{Q})$ ▷ Variables in $\mathbf{Q}$ is ID w.r.t V_j already.
    if $CheckConsition4(\mathcal{Q}, \mathbf{Q}, \mathbf{Mem}_i, \mathbf{Mem}_j, G_{\overline{\mathbf{T}}})$ **then** ▷ Check conditions in Prop. 4 and Lem. 3
      remove edge $Z \to \widehat{\mathbf{Q}}$ in $G_{\mathbf{V},\widehat{\mathbf{V}}}$
return $G_{\mathbf{V},\widehat{\mathbf{V}}}$

Next, the Alg. F.6 illustrates the details to check whether $V_i, V_j \in \mathbf{Q}$ such that V_i and V_j are independent of each other conditioning on other variables in $\mathbf{Q}$ can be disentangled according to Proposition 4. To illustrate, two lists of variables that have already been disentangled from V_i and V_j are constructed as $\mathbf{Mem}_i$ and $\mathbf{Mem}_j$ respectively through **CheckMemoize**. Next, check if there is a sub-collection of $\mathcal{Q}$ that satisfy the [1-3] conditions in Proposition 3. The checking procedure is shown in Alg.F.7. If conditions are satisfied the edges from Z to $\widehat{\mathbf{Q}}$ are removed to demonstrate disentanglement. Based on the Lem. 3, the condition [3'] in Prop. 4 can be reduced to a weaker condition [4'] leveraging existing disentanglements in CDM.

Lemma 3 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$. For any pair of $V_i, V_j \in \mathbf{V}^{tar}$ such that $V_i \perp\!\!\!\perp V_j | \mathbf{V}^{tar}\backslash\{V_i, V_j\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, let $\mathbf{Mem}_i$ be a list of variables in*

$\mathbf{Q}$ *that have been ID w.r.t.* V_i *and let* $\mathbf{Mem}_j$ *be a list of qvariables in* $\mathbf{Q}$ *that have been ID w.r.t.* V_j. *If there exists* $\mathcal{P}_{\mathbf{T}}$ *that satisfies conditions [1-2] in Prop. 3 and the following condition [4'].*

[4'] (Enough changes occur across distributions) Let $\mathbf{Q}^{re} = \mathbf{V}^{tar} \backslash (\mathbf{Mem}_i \bigcup \mathbf{Mem}_j)$ *and* $d' = |\mathbf{Q}^{re}|$. *And*

$$\begin{aligned}\boldsymbol{\mathcal{E}}_{ij} =& \{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid i)\ \exists a_l, \{V_k, V_r\} \in \Delta\mathbf{Q}^{(a_l),(a_0)};\\ & ii)\ V_k \text{ is connected to} V_r \text{ conditioning } \mathbf{V}^{tar} \backslash \{V_k, V_r\} \text{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\\ & iii)\ V_k, V_r \notin \mathbf{Mem}_i \cup \mathbf{Mem}_j\}\end{aligned} \tag{164}$$

there exists $\{a'_1, \ldots, a'_{2d'+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \ldots, a_L\}$ *such that for all* $Q_i \in \mathbf{Q}^{re}, Q_i \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], Q_i \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$ *and for all* $\boldsymbol{\epsilon}_l \in \boldsymbol{\mathcal{E}}_{ij}, \boldsymbol{\epsilon}_l \subseteq \Delta\mathbf{Q}^{(a'_{2d'+l}),(a_0)}$.

, then V_i *is ID w.r.t* V_j. □

Algorithm F.7 CheckCondition4: Check conditions in Proposition 4 and 3. $\mathcal{Q}$ is the collection of $\Delta\mathbf{Q}$ sets; $\mathbf{Q}$ are target variables;$\mathbf{Mem}_i$ are variables in $\mathbf{Q}$ have already been disentangled with V_i;$\mathbf{Mem}_j$ are variables in $\mathbf{Q}$ have already been disentangled with V_j; $G_{\overline{\mathbf{T}}}$ is the diagram after removing incoming edge to $\mathbf{T}$.

Input: $\mathcal{Q}$, $\mathbf{Q}$, $\mathbf{Mem}_i$, $\mathbf{Mem}_j$, $G_{\overline{\mathbf{T}}}$
Output: $True$ or $False$

$\mathbf{L} \leftarrow \{\}$
for $\mathbf{Q}_k \in \mathcal{Q}$ **do**
  if $\mathbf{Q}_k \subseteq \mathbf{Q}$ **then**
    $\mathbf{L}.append(\mathbf{Q}_k)$
$\boldsymbol{\mathcal{E}} \leftarrow \{\}$
for $\{V_k, V_r\} \subseteq \mathbf{Q}$ **do**
  if (i) $\exists L \in \mathbf{L}$ such that $\{V_k, V_r\} \subseteq L$ (ii) V_k is conditionally connected to V_l (iii) $\{V_k, V_r\} \nsubseteq \mathbf{Mem}_i \cup \mathbf{Mem}_j$ **then**
    $\boldsymbol{\mathcal{E}}.append((V_k, V_r))$ ▷ Construct $\boldsymbol{\mathcal{E}}$ according to Lem. 3
$\mathbf{Q}^{re+} = \{Q_1, \ldots, Q_{d'}\} \leftarrow (\mathbf{Q} \backslash (\mathbf{Mem}_i \cup \mathbf{Mem}_j)) \cup \boldsymbol{\mathcal{E}}$, $d^+ \leftarrow |\mathbf{Q}^{re}|$
if $Q_1 \in \mathbf{L}_1, Q_2 \in \mathbf{L}_2, \ldots, Q_{d'} \in \mathbf{L}_{d'}$ after a permutation of $\mathbf{L}$ **then**
  return $True$
return $False$

Lastly, we leverage the independence and current disentangled results stored in $G_{\mathbf{V},\widehat{\mathbf{V}}}$. Canceled variables with $\mathbf{V} \backslash \mathbf{Q}$ can be disentangled with each other according to Proposition 5. The following algorithm illustrates this step.

Algorithm F.8 DisΔQFromCancel - Disentangle canceled variables from $\Delta\mathbf{Q}$. $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is the current bipartite graph; $G_{\overline{\mathbf{T}}}$ is the LCG after the hard intervention on $\mathbf{T}$.

Input: $\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{X}}}$
Output: $G_{\mathbf{V},\widehat{\mathbf{V}}}$

for for all Z such that $Z \perp \mathbf{V} \backslash Z$ in $G_{\overline{\mathbf{T}}}$ **do**
  if there are no edges from $\mathbf{V} \backslash Z$ to $\mathbf{Z}$ **then**
    remove edges from Z to $\mathbf{V} \backslash Z$
return $G_{\mathbf{V},\widehat{\mathbf{V}}}$

C Proofs

Here, we provide detailed proofs of theoretical results in the main paper.

C.1 "Distribution Change Sufficiently" - Proof of Lemma 1

We assume "distributions changes sufficiently" in Sec. 2. This assumption is formally defined in Assumption 6 and will be used as a technique assumption in the proof of propositions in this work. Lemma 1 provides the justification of this assumption. It suggests Assumption 6 almost surely holds. We first provide proof here.

Assumption 6 (Changing Sufficiently)**.** *Consider a collection of ASCMs $\boldsymbol{\mathcal{M}}$ and a set of distribution $\mathcal{P}$ induced by $\boldsymbol{\mathcal{M}}$ from a collection of interventions Σ. Let the LSD induced by $\boldsymbol{\mathcal{M}}$ be G^S. Let $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \dots, P^{(a_L)}\} \subseteq \mathcal{P}$ be any collection of distributions such that $\mathbf{T} = do[\mathbf{I}^{(a_0)}] \subseteq do[\mathbf{I}^{(a_l)}]$ for $l \in [L]$, meaning for the baseline distribution all hard interventions must be exactly on $\mathbf{T}$, and all other distributions must at least contain $\mathbf{T}$ in their hard interventions. Let $\mathbf{Q} = \bigcup_{l\in[L]} \Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(0)}, \mathbf{T}, G^S]$ (Def. 3.1). It is assumed:*

1. *The probability density function of $\mathbf{V}$ is smooth and positive, i.e. $p_{\mathbf{T}}^{(a_l)}(\mathbf{v})$ is smooth and $p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) > 0$ almost everywhere.*
2. *First-order discrepancy. If there exists $\{a'_1, \dots, a'_{|\mathbf{Q}|}\} \subseteq \{a_1, \dots, a_L\}$ such that $\forall\ V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_q)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, then $\{\boldsymbol{\omega}_1(\mathbf{v}, a_1), \boldsymbol{\omega}_1(\mathbf{v}, a_2), \dots, \boldsymbol{\omega}_1(\mathbf{v}, a_L)\}$ are linearly independent, where*

$$\boldsymbol{\omega}_1(\mathbf{v}, a_l) = \left(\oplus \left(\frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q}\right)_{V_q\in\mathbf{Q}} \right) \tag{156}$$

3. *Second-order discrepancy. Let a set $\boldsymbol{\mathcal{E}}$ consist of pairs of (V_p, V_q) such that (V_p, V_q) appears at least in one $\Delta\mathbf{Q}$ and V_p is connected with V_q conditioning on $\mathbf{V}\backslash\{V_p, V_q\}$ in $G_{\overline{T}}(\mathbf{Q})$. Namely,*

$$\begin{aligned}\boldsymbol{\mathcal{E}} =& \{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid \\ & \textit{(i)}\ \exists a_l, \{V_p, V_q\} \in \Delta\mathbf{Q}^{(a_l),(a_0)}; \\ & \textit{(ii)}\ V_p \textit{ is d-connected to } V_q \textit{ conditioned on } \mathbf{V}^{tar}\backslash\{V_p, V_q\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\},\end{aligned} \tag{157}$$

If there exists $\{a'_1, \dots, a'_{2|\mathbf{Q}|+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \dots, a_L\}$ such that $\forall\ V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], V_q \in \Delta\mathbf{Q}^{(a'_{|Q|+i}),(a_0)}$ and for all $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2|\mathbf{Q}|+j}),(a_0)}$, then $\{\boldsymbol{\omega}_2(\mathbf{v}, a_1), \boldsymbol{\omega}_2(\mathbf{v}, a_2), \dots, \boldsymbol{\omega}_2(\mathbf{v}, a_L)\}$ are linearly independent, where

$$\begin{aligned}\boldsymbol{\omega}_2(\mathbf{v}, a_l) = \Bigg(& \oplus \left(\frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q}\right)_{V_q\in\mathbf{Q}}, \\ & \oplus \left(\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q^2}\right)_{V_q\in\mathbf{Q}}, \\ & \oplus \left(\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_p v_q}\right)_{(V_p,V_q)\in\mathcal{E}(G_{\overline{T}}(\mathbf{Q}))} \Bigg)\end{aligned} \tag{158}$$

□

Lemma 1. *Assumption 6 almost surely holds.* □

Proof. We will prove the first-order discrepancy and second-order discrepancy almost surely hold, which means the situations where first-order discrepancy and second-order discrepancy do not hold have Lebesgue measure 0.

We first consider the first-order discrepancy. Denote $\{\boldsymbol{\omega}_1(\mathbf{v}, a_1), \boldsymbol{\omega}_1(\mathbf{v}, a_2), \dots, \boldsymbol{\omega}_1(\mathbf{v}, a_L)\}$ as $\mathbf{A}$. And every entry in $\mathbf{A}$ is

$$a_{lq} = \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q} \tag{165}$$

According to Eq. (119), we know $\log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})$ is a function of only variables in $\Delta\mathbf{Q}$. Thus, if $V_q \notin \Delta\mathbf{Q}[\mathbf{I}^{(a'_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, $a_{lq} = 0$; if $V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, we assume a_{lq} follows a standard normal distribution, which means the non-zero entries in matrix $\mathbf{A}$ are randomly sampled and are not fine-tuned. Thus, to prove this lemma, it is equivalent to prove if there exists $\{a'_1, \ldots, a'_{|\mathbf{Q}|}\} \subseteq \{a_1, \ldots, a_L\}$ such that $\forall\ V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_q)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, the row of $\mathbf{A}$ are almost surely linear independent. W.O.L.G, we let $\{a'_1 = a_1, \ldots, a'_{\mathbf{Q}} = a_L$. Then, it is equivalent to prove that $\mathbf{A}$ is a full rank matrix.

In order to prove that $\mathbf{A}$ is a full-rank matrix, we prove that the determinant of $\mathbf{A}$ is almost surely non-zero. Since $\forall\ V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a_q)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, there exists $\mathbf{A}$ such that $\det(\mathbf{A})$ is non-zero, and then $\det(\mathbf{A})$ is non-trivial. Based on a simple algebraic lemma in [84], the subset of $\{\mathbf{A} \mid \det(\mathbf{A}) = 0\}$ of the real space has Lebesgue measure 0. Then $\det(\mathbf{A}) = 0$ almost surely holds.

The second-order discrepancy proof is similar. Denote $\{\boldsymbol{\omega}_2(\mathbf{v}, a_1), \boldsymbol{\omega}_2(\mathbf{v}, a_2), \ldots, \boldsymbol{\omega}_2(\mathbf{v}, a_L)\}$ as $\mathbf{B}$. And every entry of $\mathbf{B}$ is

$$
\begin{aligned}
a_{lq} &= \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q}, q \leq |\mathbf{Q}| \\
a_{lq} &= \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q^2}, |\mathbf{Q}| + 1 \leq q \leq 2|\mathbf{Q}| \\
a_{l\epsilon} &= \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_p \partial v_q}, 2|\mathbf{Q}| + 1 \leq \epsilon \leq 2|\mathbf{Q}| + |\boldsymbol{\mathcal{E}}|, \{V_p, V_q\} \in \boldsymbol{\mathcal{E}}
\end{aligned}
\tag{166}
$$

According to Eq. (119), we know $\log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})$ is a function of only variables in $\Delta\mathbf{Q}$. Thus, if $V_q \notin \Delta\mathbf{Q}[\mathbf{I}^{(a'_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, $a_{lq} = 0$; if $V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, we assume a_{lq} follows a standard normal distribution, which means the non-zero entries in matrix $\mathbf{B}$ are randomly sampled and are not fine-tuned. If $\{V_p, V_q\} \not\subseteq \Delta\mathbf{Q}[\mathbf{I}^{(a'_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, $a_{l\epsilon} = 0$; if $\{V_p, V_q\} \subseteq \Delta\mathbf{Q}[\mathbf{I}^{(a'_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, we assume $a_{l\epsilon}$ follows a standard normal distribution, which means the non-zero entries in matrix $\mathbf{B}$ are randomly sampled and are not fine-tuned. Following the same discussion above, the subset of $\{\mathbf{B} \mid \det(\mathbf{B}) = 0\}$ of the real space has Lebesgue measure 0. Then $\det(\mathbf{B}) = 0$ almost surely holds. □

C.2 Distribution comparison - Proof of Proposition 1

Proposition 1 (Distribution Comparison). *Consider a collection of ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \ldots, \mathcal{M}_n \rangle$ that induces a collection of distributions $\mathcal{P}$ with interventions Σ and LSD G^S. Consider comparing two distributions $P^{\Pi^{(j)}}(\mathbf{X}; \sigma^{(j)}), P^{\Pi^{(k)}}(\mathbf{X}; \sigma^{(k)}) \in \mathcal{P}$ with intervention targets $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$. Suppose $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$ both contain a hard intervention on $\mathbf{T}$ ($\mathbf{T}$ can be an empty set, which means no hard intervention is involved). If another collection of ASCMs, $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \ldots, \widehat{\mathcal{M}}_n \rangle$, matches with distribution $\mathcal{P}$ and LSD G^S, then*

$$
\underbrace{\sum_i^d \log p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+})}_{\boldsymbol{\mathcal{M}}} = \underbrace{\sum_i^d \log p_{\mathbf{T}}^{(j)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+})}_{\widehat{\boldsymbol{\mathcal{M}}}},
\tag{104}
$$

where $p_{\mathbf{T}}^{(j)}(\cdot), p_{\mathbf{T}}^{(k)}(\cdot)$ are density functions. □

Proof. According to the ASCM definition Def .2.1, the mapping from $\mathbf{V}$ to $\mathbf{X}$, and the mapping $\mathbf{X}$ to $\widehat{\mathbf{V}}$ can be expressed as:

$$
\widehat{\mathbf{V}} = \widehat{f}_{\mathbf{X}}^{-1}(\mathbf{X}) = \widehat{f}_{\mathbf{X}}^{-1}(f_{\mathbf{X}}(\mathbf{V}))
\tag{167}
$$

Then based on the change variable formula, we have

$$
p(\mathbf{v}) = p(\widehat{\mathbf{v}})|\mathbf{J}_{\boldsymbol{\phi}}|
\tag{168}
$$

where $\boldsymbol{\phi} = \widehat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}$ and $\mathbf{J}_{\boldsymbol{\phi}}$ is the Jacobian matrix of $\boldsymbol{\phi}$. Leveraging the factorization in Eq. 85 and taking log of the above equation,

$$\sum_{i=1}^{d} \log p_{\mathbf{T}}(v_i \mid \mathbf{pa}^{\mathbf{T}+}) = \sum_{i=1}^{d} \log p_{\mathbf{T}}(\widehat{v}_i \mid \widehat{\mathbf{pa}}^{\mathbf{T}+}) + \log |\mathbf{J}_{\boldsymbol{\phi}}| \tag{169}$$

Subtract the above factorization of density function induced by $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$, and we have Eq.(104). □

Eq 104 naturally gives a connection from $\mathbf{V}$ to $\widehat{\mathbf{V}}$. Comparing two factorization for Fig. 11(c), the connection connections are made from $P(v_1), p(v_2 \mid v_1), p(v_3 \mid v_2, v_1), P(v_4 \mid v_3)$ or $P(v_1), p(v_3), p(v_2 \mid v_1, v_3), P(v_4 \mid v_3)$.

C.3 Invariant factors - Proof of Proposition 2

Proposition 2 (Invariant Factors). *Consider two distributions $P^{(j)}, P^{(k)} \in \mathcal{P}$ with intervention targets $\sigma^{(j)}$ and $\sigma^{(k)}$ containing $do(\mathbf{T})$. Construct the changed variable set $\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$ (for short $\Delta\mathbf{V}^{(j),(k)}$ or $\Delta\mathbf{V}$ if index clear from the context) with target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ as follows. Add a variable V_l to $\Delta\mathbf{V}$ whenever one of the conditions is satisfied:*

1. *(Interventions) $V_l \in \Delta\mathbf{V}$ if $V_l^{\pi_l,\{b_l\},t_l} \in \mathbf{I}^{(j)}$ but $V_l^{\pi'_l,\{b_l\},t'_l} \notin \mathbf{I}^{(k)}$, and vice versa;*
2. *(Domains) $V_l \in \Delta\mathbf{V}$ if (i) $S^{\Pi^{(j)},\Pi^{(k)}}$ point to V_l, (ii) $V_l^{\pi_l,\{b_l\},t_l} \notin \mathbf{I}^{(j)}$, (iii) $V_l^{\pi_l,\{b_l\},t_l} \notin \mathbf{I}^{(j)}$.*

If $V_i \in \mathbf{V}\backslash\mathbf{C}(\Delta\mathbf{V})$, then

$$p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}), \text{ and } p_{\mathbf{T}}^{(j)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+})$$

which will be denoted as invariant factors, where $\mathbf{C}(\Delta\mathbf{V}) = \bigcup_{Z\in\Delta\mathbf{V}}\{V \in \mathbf{C}(Z)\}$. The factors that are not invariant will be termed non-invariant. □

Proof. Consider an arbitrary order. Based on the proposition, $\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$ includes all variables that the mechanism f_V or exogenous U possibly change when the intervention changes from $\mathbf{I}^{(k)}$ to $\mathbf{I}^{(j)}$. In other words, for any $V_l \in \mathbf{V}\backslash\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$, f_{V_l} and exogenous U_l are invariant.

Let $V_i \in \mathbf{V}\backslash\mathbf{C}(\Delta\mathbf{V})$. $\mathbf{Z} = \mathbf{C}(V_i) \cup \mathbf{Pa}^{\mathbf{T}+}$. We have $\mathbf{Z} \subseteq \mathbf{V}\backslash\mathbf{C}(\Delta\mathbf{V})$ according to the definition of C-component.

According to the definition of $\mathbf{Pa}_i^{\mathbf{T}^+}$, we know $\mathbf{Pa}_i^{\mathbf{T}^+}\backslash\mathbf{Z} = \mathbf{Pa}(\{V_i\} \cup \mathbf{Z})$. Now reconsider the distribution $P^{\Pi^(j)}(V_i \mid \mathbf{Pa}_i^{\mathbf{T}+}; \sigma^{(j)})$ and $P^{\Pi^(k)}(V_i \mid \mathbf{Pa}_i^{\mathbf{T}+}; \sigma^{(k)})$,

$$P_{\mathbf{T}}^{\Pi^{(j)}}(V_i \mid \mathbf{Pa}_i^{\mathbf{T}+}; \sigma^{(j)}) = P_{\mathbf{T}}^{\Pi^j}(V_i, \mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(j)}) / P_{\mathbf{T}}^{\Pi^{(j)}}(\mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(j)})$$

(170)

$$P_{\mathbf{T}}^{\Pi^{(k)}}(V_i \mid \mathbf{Pa}_i^{\mathbf{T}+}; \sigma^{(k)}) = P_{\mathbf{T}}^{\Pi^{(k)}}(V_i, \mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)}) / P_{\mathbf{T}}^{\Pi^{(k)}}(\mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)}) \tag{171}$$

Since the mechanism and exogenous variables of V_i and $\mathbf{Z}$ are invariant, both the nominators and denominators are the same. Namely,

$$P_{\mathbf{T}}^{\Pi^j}(V_i, \mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(j)}) = P_{\mathbf{T}}^{\Pi^k}(V_i, \mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)}) \tag{172}$$

$$P_{\mathbf{T}}^{\Pi^{(j)}}(\mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(j)}) = P_{\mathbf{T}}^{\Pi^{(k)}}(\mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)}) \tag{173}$$

which implies the density functions are invariant,

$$p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) \tag{174}$$

□

C.4 ID ΔQ w.r.t Canceled Factors - Proof of Proposition 3 and Lemma 2

Proposition 3 (ID the ΔQ set w.r.t. Canceled Variables). *Consider variables* $\mathbf{V}^{tar} = \{V_1^{tar}, V_2^{tar}, \dots, V_{d'}^{tar}\} \subseteq \mathbf{V}$. *If there exists a subset of* $\mathcal{P}$, $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \dots, P^{(a_L)}\} \subseteq \mathcal{P}$ *with intervention target sets* $\Psi_{\mathbf{T}} = \{\mathbf{I}^{(a_0)}, \mathbf{I}^{(a_1)}, \dots, \mathbf{I}^{(a_L)}\}$ *such that*

(1) All distributions contain hard intervention on $\mathbf{T}$, *i.e.,* $\forall\, l \in [L]$, $\mathbf{T} = do[\mathbf{I}^{(a_0)}] \subseteq do[\mathbf{I}^{(a_l)}]$ [22].

(2) The union of all ΔQ *sets is* $\mathbf{V}^{tar}$, *i.e.,* $\bigcup_{l\in[L]} \Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S] = \mathbf{V}^{tar}$.

(3) Each V_i^{tar} *changes once, i.e., there exists* $\{a'_1, \dots, a'_{d'}\} \subseteq \{a_1, \dots, a_L\}$ *such that for all* $V_i^{tar} \in \mathbf{V}^{tar}$, $V_i^{tar} \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_i)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, *where* $d' = |\mathbf{V}^{tar}|$.

then $\mathbf{V}^{tar}$ *is ID w.r.t.* $\mathbf{V}\backslash\mathbf{V}^{tar}$. □

Proof. **We denote $\mathbf{V}^{tar}$ as $\mathbf{Q}$ for convenience.** Notice that the Assumption 6 will be used in the proof.

Comparing $P^{\Pi^{(a_l)}}(\mathbf{V}; \sigma^{(a_l)})$ with $P^{\Pi^{(a_0)}}(\mathbf{V}; \sigma^{(a_0)})$, we have

$$\begin{aligned}\sum_{V_i\in\tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - p_{\mathbf{T}}^{a_0)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) &= \sum_{V_i\in\tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) \\ &= \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})\end{aligned} \tag{175}$$

from Eq. (119).

Notice that the left side only involves variables in $\mathbf{Q} = \bigcup_{l\in[L]} \Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$ based on the Def. 3.1. Thus, for any $Z \in \mathbf{V}\backslash\mathbf{Q}$,

$$\forall l \in [L], \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{z}} = 0 \tag{176}$$

Take partial of the above equation w.r.t. Z, we have:

$$\forall l \in [L], 0 = \sum_{v_i\in\mathbf{V}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_i} \frac{\partial \widehat{v}_i}{\partial z} \qquad \text{(Chain Rule)} \tag{177}$$

$$= \sum_{v_q\in\mathbf{Q}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_q} \frac{\partial \widehat{v}_q}{\partial z} \qquad \text{(Eq.(176))} \tag{178}$$

Eq. (178) is a linear system for unknowns $\{\partial \widehat{V}_q / \partial Z\}_{V_q\in\mathbf{Q}}$. When distribution changes sufficiently, namely under Assumption 6, the row factor of the coefficient matrix of the linear system is linearly independent. When $L \geq |\mathbf{Q}|$ (implied by condition [3]), the matrix is full rank, thus,

$$\forall\, V_q \in \mathbf{Q}, \frac{\partial \widehat{v}_q}{\partial z} = 0 \tag{179}$$

Recall that $V_q = \phi_{V_q}(\mathbf{V})$. For any $Z \in \mathbf{V}\backslash\mathbf{Q}$, Eq.(179) holds. Thus, $\mathbf{Q}$ is enough to be the input of ϕ_{V_q}, which means there exists $V_q = \phi_{V_q}(\mathbf{Q})$.

□

Lemma 2. *Consider variables* $\mathbf{V}^{tar} \subseteq \mathbf{V}$ *and* $Z \in \mathbf{V}\backslash\mathbf{V}^{tar}$. *Suppose* $\mathbf{Mem} = \{V_j \in \mathbf{V}^{tar} \mid V_j$ *is ID w.r.t.* $Z\}$. *Consider,* $\mathcal{P}_{\mathbf{T}}$ *and its corresponding intervention targets that hold conditions [1-2] in Prop. 3. If the new version of the condition [3] is also satisfied:*

[4] there exists $\{a'_1, \dots, a'_{|\mathbf{V}^{tar}|}\} \subseteq \{a_1, \dots, a_L\}$ *such that for all* $V_i^{tar} \in \mathbf{V}^{tar}\backslash\mathbf{Mem}$, $V_i^{tar} \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_i)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$.

[22] Recall we use the notation $do[\mathbf{I}]$ to denote that all variables that perfectly interventions on in $\mathbf{I}$.

then $\mathbf{V}^{tar}$ *is ID w.r.t* Z. □

Proof. For all $V_m \in \mathbf{Mem}$, $\partial V_m/\partial Z = 0$. Thus, the unknown in Eq.(178) exclude $\frac{\partial v_m}{\partial z}$. Then, when [3'] holds, the system will have zero solutions and Eq.(179) will hold. □

C.5 ID within $\Delta\mathbf{Q}$ set - Proof of Proposition 4 and Lemma 3

The next result provides us an additional way of disentangling latent variables within the same ΔQ-factor leveraging second-order conditions and conditional independence. The assumption made in the result is the assumption of generalized distributional change (see Assump. 6).

Proposition 4 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider the variables* $\mathbf{V}^{tar} \subseteq \mathbf{V}$. *Define* $\boldsymbol{\mathcal{E}}$ *as the set of edges within the Markov Network of* $G_{\overline{T}}(\mathbf{V}^{tar})$ *that are contained within a* $\Delta\mathbf{Q}$ *set.*

$$\begin{aligned}\boldsymbol{\mathcal{E}} =&\{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \Big| \\ &(i)\ \exists a_l, \{V_k, V_r\} \subseteq \Delta\mathbf{Q}^{(a_l),(a_0)}; \\ &(ii)\ V_k \text{ is d-connected to } V_r \text{ conditioned on } \mathbf{V}^{tar}\backslash\{V_k, V_r\} \text{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\},\end{aligned} \tag{132}$$

For any pair of $V_i, V_j \in \mathbf{V}^{tar}$ *such that* $V_i \perp\!\!\!\perp V_j \mid \mathbf{V}^{tar}\backslash\{V_i, V_j\}$ *in* $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, *if there exists* $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \dots, P^{(a_L)}\} \subseteq \mathcal{P}$ *that satisfies conditions (1-2) in Prop. 3 and the following condition (3').*

(3') Enough changes occur across distributions, i.e., Formally, there exists $\{a'_1, \dots, a'_{2d'+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \dots, a_L\}$ *such that for all* $V_i^{tar} \in \mathbf{V}^{tar}$, *i)* $V_i^{tar} \in \Delta\mathbf{Q}^{(a'_i),(a_0)}$, *ii)* $V_i^{tar} \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$, *and iii) for all* $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2d'+j}),(a_0)}$, *where* $d' = |\mathbf{V}^{tar}|$

then, V_i *is ID w.r.t.* V_j. □

Proof. **We denote $\mathbf{V}^{tar}$ as $\mathbf{Q}$ for convenience.** Notice that Assumption 6 will be used in the proof. From Eq. 119, we have

$$\begin{aligned}\sum_{V_i \in \tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - p_{\mathbf{T}}^{a_0)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) &= \sum_{V_i \in \tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) \\ &= \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})\end{aligned} \tag{180}$$

Notice that the left side only involves variables in $\mathbf{Q} = \bigcup_{l\in[L]} \Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$ based on the Def. 3.1.

We first argue that if $V_i \perp\!\!\!\perp V_j | \mathbf{Q}\backslash\{V_i, V_j\}$ in $G_{\overline{\mathbf{T}}}$ then $V_i \notin \mathbf{P}a_j^{\mathbf{T}+}, V_j \notin \mathbf{P}a_i^{\mathbf{T}+}$ and $V_i, V_j \notin \mathbf{P}a_m^{\mathbf{T}+}$ where $V_m \in \mathbf{Q}$.

First, since $V_i \perp\!\!\!\perp V_j | \mathbf{Q}\backslash\{V_i, V_j\}$, V_i and V_j cannot be directly connected by edges in $G_{\overline{\mathbf{T}}}$, which implies $V_i \notin \mathbf{C}(V_j)$ and $V_i \notin \mathbf{Pa}^{\mathbf{T}+}(V_j)$. Also, the outgoing edge from V_i and V_j cannot point to the same C-component. Otherwise, the path is active from V_i and V_j is active when conditioning on other variables (collider structure). Thus, $V_i \notin \mathbf{P}a_j^{\mathbf{T}+}, V_j \notin \mathbf{P}a_i^{\mathbf{T}+}$ and $V_i, V_j \notin \mathbf{P}a_k^{\mathbf{T}+}$ where $V_k \in \mathbf{Q}$. This implies V_i and V_j will not appear to the same factor $p_{\mathbf{T}}^{(a_l)}(v_m \mid \mathbf{pa}_m^{\mathbf{T}+})$ for any $V_m \in \tilde{\mathbf{V}}$. Thus,

$$\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(v_m \mid \mathbf{pa}_m^{\mathbf{T}+})}{\partial v_i v_j} = 0 \tag{181}$$

Thus, for any pair of V_k, V_r such that $V_k \perp\!\!\!\perp V_r | \mathbf{Q}\backslash\{V_k, V_r\}$,

$$\begin{aligned}\forall l \in [L], \sum_{V_m \in \tilde{V}} \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{v}_m \mid \mathbf{pa}_m^{\mathbf{T}+})}{\partial \widehat{v}_k \widehat{v}_r} \\ = \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} = 0\end{aligned} \tag{182}$$

On the other hand, when either V_k or V_r is in $\mathbf{Q}\backslash\Delta\mathbf{Q}^{(a_l),(a_0)}$ for $l \in [L]$,

$$\forall l \in [L], = \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} = 0 \tag{183}$$

since

$$\frac{\partial p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k} = 0 \quad \text{or} \quad \frac{\partial p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_r} = 0 \tag{184}$$

Upon Eq. (181), taking the second partial derivative on both sides of Eq. (180), the left side will be 0, and then $\forall\, l \in [L]$, we have

$$0 = \sum_{V_k, V_r \in \mathbf{Q}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} \frac{\partial \widehat{v}_k}{\partial v_i} \frac{\widehat{v}_r}{\partial v_j} \qquad \text{Chain Rule} \tag{185}$$

$$\begin{aligned}
&= \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_i^2} \frac{\partial \widehat{v}_i}{\partial v_i} \frac{\partial \widehat{v}_i}{\partial v_j} + \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_j^2} \frac{\partial \widehat{v}_j}{\partial v_i} \frac{\partial \widehat{v}_j}{\partial v_j} \\
&+ \sum_{V_q \in \mathbf{Q}} \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_q^2} \frac{\partial \widehat{v}_q}{\partial v_i} \frac{\partial \widehat{v}_q}{\partial v_j} \\
&+ \sum_{V_q \in \mathbf{Q}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_q} \frac{\partial^2 \widehat{v}_q}{\partial v_i v_j} \\
&+ \sum_{(V_k, V_r) \in \boldsymbol{\mathcal{E}}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} \frac{\partial \widehat{v}_k}{\partial v_i} \frac{\widehat{v}_r}{\partial v_j} \qquad \text{Eq. (182) and (182)}
\end{aligned} \tag{186}$$

Eq.(186) is also a linear system. When distribution changes sufficiently, namely under Assumption 6, the row factor of the coefficient matrix of the linear system is linearly independent. When $L \geq 2|\mathbf{Q}| + \delta_{\not\perp}$ (implied by condition 4), the matrix is full rank, thus,

$$\frac{\partial \widehat{v}_i}{\partial v_i} \frac{\partial \widehat{v}_i}{\partial v_j} = 0, \frac{\partial \widehat{v}_j}{\partial v_i} \frac{\partial \widehat{v}_j}{\partial v_j} = 0 \tag{187}$$

Then we have

$$\frac{\partial \widehat{v}_i}{\partial v_j} = 0, \frac{\partial \widehat{v}_i}{\partial v_j} = 0 \tag{188}$$

up to a permutation of V_i and V_j. This implies that V_i is ID w.r.t V_j and V_j is ID w.r.t V_i. □

Lemma 3 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$. For any pair of $V_i, V_j \in \mathbf{V}^{tar}$ such that $V_i \perp\!\!\!\perp V_j | \mathbf{V}^{tar}\backslash\{V_i, V_j\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, let $\mathbf{Mem}_i$ be a list of variables in $\mathbf{Q}$ that have been ID w.r.t. V_i and let $\mathbf{Mem}_j$ be a list of qvariables in $\mathbf{Q}$ that have been ID w.r.t. V_j. If there exists $\mathcal{P}_{\mathbf{T}}$ that satisfies conditions [1-2] in Prop. 3 and the following condition [4'].*

[4'] (Enough changes occur across distributions) Let $\mathbf{Q}^{re} = \mathbf{V}^{tar}\backslash(\mathbf{Mem}_i \bigcup \mathbf{Mem}_j)$ and $d' = |\mathbf{Q}^{re}|$. And

$$\begin{aligned}
\boldsymbol{\mathcal{E}}_{ij} = \{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid &\textit{ i) } \exists a_l, \{V_k, V_r\} \in \Delta\mathbf{Q}^{(a_l),(a_0)}; \\
&\textit{ii) } V_k \textit{ is connected to} V_r \textit{ conditioning } \mathbf{V}^{tar}\backslash\{V_k, V_r\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar}) \\
&\textit{iii) } V_k, V_r \notin \mathbf{Mem}_i \cup \mathbf{Mem}_j\}
\end{aligned} \tag{164}$$

there exists $\{a'_1, \ldots, a'_{2d'+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \ldots, a_L\}$ such that for all $Q_i \in \mathbf{Q}^{re}, Q_i \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], Q_i \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$ and for all $\boldsymbol{\epsilon}_l \in \boldsymbol{\mathcal{E}}_{ij}, \boldsymbol{\epsilon}_l \subseteq \Delta\mathbf{Q}^{(a'_{2d'+l}),(a_0)}$.

, then V_i is ID w.r.t V_j. □

Proof. The unknown in the linear system

$$\frac{\partial \widehat{v}_q}{\partial v_i}\frac{\partial \widehat{v}_q}{\partial v_j} = 0, \tag{189}$$

if V_p is ID w.r.t V_i or V_q is ID w.r.t V_j.

$$\frac{\partial^2 \widehat{v}_q}{\partial v_i v_j} = 0 \tag{190}$$

If V_q is ID w.r.t V_i or V_j. Even these terms are excluded in [4'], the system still has the zero solutions. □

Corollary 1 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider the variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$ and distributions $\mathcal{P}_{\mathbf{T}} = \{P^{(a_0)}, P^{(a_1)}, \dots, P^{(a_L)}\} \subseteq \mathcal{P}$ that satisfies conditions (1) in Prop. 3 and $\Delta\mathbf{Q}^{(a_l),(a_0)} = \mathbf{V}^{tar}$, for $l \in [L]$. For any pair of $V_i, V_j \in \mathbf{V}^{tar}$ such that $V_i \perp\!\!\!\perp V_j | \mathbf{V}^{tar}\backslash\{V_i, V_j\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, V_i is ID w.r.t. V_j if $L \geq 2|\mathbf{V}^{tar}| + \delta_{\not\perp}$, where $\delta_{\not\perp}$ is the number of pair $V_k, V_r \in \mathbf{V}^{tar}$ such that V_k and V_r are connected given $\mathbf{V}^{tar}\backslash\{V_k, V_r\}$ in $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$.* □

Proof. the proof of this Corollary comes directly from Prop. 4. Taking let each condition [2] and [3] are satisfied when $\Delta\mathbf{Q}^{(a_1),(a_0)} = \cdots = \Delta\mathbf{Q}^{(a_L),(a_0)}\mathbf{V}^{tar}$ when $L \geq 2|\mathbf{V}^{tar}| + |\boldsymbol{\mathcal{E}}|$ □

C.6 ID-reverse of existing disentangled variables - Proof of Proposition 5

The next Proposition provides an additional tool to achieve identifiability and leverages the fact that other variables may have previously been disentangled and independence relationships in the factorization.

Proposition 5 (ID of canceled variables w.r.t. $\Delta\mathbf{Q}$ sets). *Suppose there exists $\mathbf{I}^{(k)} \in \boldsymbol{\Psi}$ such that $do(\mathbf{I}^{(k)}) = \mathbf{T}$. Given $\mathbf{V}\backslash V^{tar}$ is ID w.r.t. a single variable V^{tar}, V^{tar} is ID w.r.t. $\mathbf{V}\backslash V^{tar}$ if $V^{tar} \perp\!\!\!\perp \mathbf{V}\backslash V^{tar}$ in $G_{\overline{\mathbf{T}}}$.* □

Proof. We first introduce a lemma for distribution preserving from [20].

Lemma 4 (Lemma 2 of [20]). *Let $A = C = R$ and $B = \mathbb{R}^n$. Let $f : A \times B \to C$ be differentiable. Define differentiable measures P_A on A and P_C on C. Let $\forall b \in B$, $f(\cdot, b) : A \to C$ be measure-preserving. Then f is constant in b.*

Denote $\mathbf{V}\backslash\mathbf{V}^{tar}$ as $\mathbf{Z}$. $V^{tar} \perp\!\!\!\perp \mathbf{Z}$ in $G_{\overline{\mathbf{T}}}$ implies that

$$P_{\mathbf{T}}(\mathbf{V}) = P_{\mathbf{T}}(V^{tar})P_{\mathbf{T}}(\mathbf{Z}) \tag{191}$$

With the change of variable formulation and taking log:

$$\log p_{\mathbf{T}}(\mathbf{v}^{tar}) + \log p_{\mathbf{T}}(\mathbf{z}) = \log p_{\mathbf{T}}(\widehat{\mathbf{v}}^{tar}) + \log p_{\mathbf{T}}(\widehat{\mathbf{z}}) + \log |\mathbf{J}_{\boldsymbol{\phi}}| \tag{192}$$

Since $\mathbf{Z}$ is ID w.r.t V^{tar}, $\partial\widehat{\mathbf{Z}}/\partial\mathbf{V}^{tar} = 0$. In other words, the elements $\partial\boldsymbol{\phi}_Z/\partial\mathbf{V}^{tar} = 0$ for every $Z \in \mathbf{Z}$ in Jacobian matrix are 0, where $\boldsymbol{\phi}_Z$ is a function mapping from $\mathbf{V}$ to $\widehat{Z}$. Then

$$\log |\mathbf{J}_{\boldsymbol{\phi}}| = \log |\mathbf{J}_{\mathbf{Z}}| + \log |\mathbf{J}_{\mathbf{V}^{tar}}| \tag{193}$$

where $|\mathbf{J}_{\mathbf{Z}}| = \begin{bmatrix} \partial\phi_{Z_1}/\partial z_1 & \partial\phi_{Z_1}/\partial z_2 & \dots & \partial\phi_{Z_1}/\partial z_{d-1} \\ \partial\phi_{Z_2}/\partial z_1 & \partial\phi_{Z_2}/\partial z_2 & \dots & \partial\phi_{Z_2}/\partial z_{d-1} \\ \vdots & \vdots & \ddots & \vdots \\ \partial\phi_{Z_{d-1}}/\partial z_1 & \partial\phi_{Z_{d-1}}/\partial z_2 & \dots & \partial\phi_{Z_{d-1}}/\partial z_{d-1} \end{bmatrix}$ and $\log |\mathbf{J}_{\mathbf{V}^{tar}}| = |\partial\phi_{V^{tar}}/\partial v^{tar}|$.

Again, since $\mathbf{Z}$ is ID w.r.t $\mathbf{V}^{tar}$, $\widehat{\mathbf{Z}} = \phi_{\mathbf{Z}}(\mathbf{Z})$. Thus,

$$\log p_{\mathbf{T}}(\mathbf{z}) = \log p_{\mathbf{T}}(\widehat{\mathbf{z}}) + \log |\mathbf{J}_{\mathbf{Z}}| \tag{194}$$

Subtracting this to Eq. (192)

$$\log p_{\mathbf{T}}(v^{tar}) = \log p_{\mathbf{T}}(\widehat{v}^{tar}) + \log |\mathbf{J}_{\mathbf{V}^{tar}}| \tag{195}$$

Denote $\phi_{\mathbf{V}^{tar}}(\mathbf{z}, \cdot)$ as $\phi^{\mathbf{z}}_{\mathbf{V}^{tar}}(\cdot)$, which is the function $\phi_{\mathbf{V}^{tar}}$ fixing value $\mathbf{Z} = \mathbf{z}$ mapping from $\mathbf{V}^{tar}$ to $\widehat{\mathbf{V}}$. This suggests for every $\mathbf{z}$,

$$P_{\mathbf{T}}(\widehat{V}^{tar}) = P_{\mathbf{T}}(\phi^{\mathbf{z}}_{V^{tar}}(V^{tar})) \tag{196}$$

Apply Lemma 2 of [20], $\phi_{V^{tar}}$ should be a constant regarding $\mathbf{Z}$. Thus,

$$\forall\, Z \in \mathbf{Z}, \frac{\partial V^{tar}}{\partial Z} = 0 \tag{197}$$

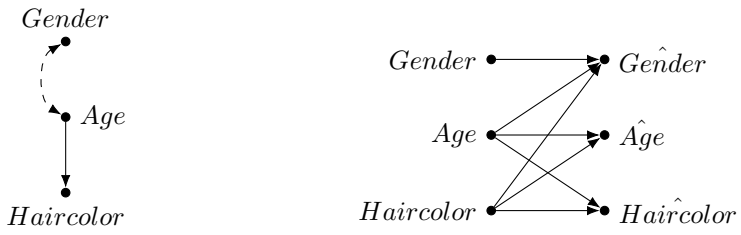
□

C.7 Soundness of LatentID Algorithm - Proof of Thm. 1

The following provides the proof of the soundness of our proposed graphical algorithm for determining whether or not two variables are disentangleable given a collection of distributions from multiple domains and interventions.

Theorem 1 (**Soundness of CRID**). *Consider a LSD G^S and intervention targets $\mathbf{\Psi}$. Consider the target variables $\mathbf{V}^{tar}$ and $\mathbf{V}^{en} \subseteq \mathbf{V} \backslash \mathbf{V}^{tar}$. If no edges from $\mathbf{V}^{tar}$ points to $\widehat{\mathbf{V}}^{en}$ in the output causal disentanglement map (CDM) from **CRID**, $G_{V,\widehat{V}}$, then $\mathbf{V}^{tar}$ is ID w.r.t $\mathbf{V}^{en}$.* □

Proof. In **LatentID**, for each epoch, we iterate to choose $\mathbf{T}$ and the baseline distribution to execute procedure Alg. F.3 and Alg. F.6. Any time an edge is removed, Proposition 3 and/or 4 are applied. At the end of epoch, Alg. F.8 is executed and edges will be removed only if Proposition 5 is applied. Thus the edge removals are all sound. The algorithm will stop when no edge will be removed, and terminate giving the causal disentanglement map $G_{V,\hat{V}}$, which is a valid summary of what is disentangleable. □



(a) G (b) $G_{V,\hat{V}}$

Figure S3: Latent causal graph and the desired causal disentanglement map.

D Discussion and Examples

D.1 Additional Example Illustrating Motivation of Causal Disentangled Learning

In the introduction, we illustrated a medical example for why it is important to learn disentangled representations.

An additional motivating example can be seen through the lens of generating realistic face images [27]. Consider an image dataset of human faces. Based on our understanding of anatomy and facial expressions, we know that both $Gender$ and Age are not causally related, while age does directly affect $HairColor$. There is a strong spurious correlation between age and gender, where there are many old males and young females in the dataset. In addition, let there be face images from both a senior and teen center building. The change in domain (i.e. population center) impacts the age distribution, as senior center faces are older than teen center faces. Given these images and knowledge of the latent causal graph, one would ultimately like to generate realistic face images given perturbations of Age. If the variable representations are entangled, then it is possible for changes in age to also spuriously change gender. This is undesirable, and thus our goal is to achieve disentanglement of age and gender. Note that we do not require Age to be disentangled from $HairColor$ necessarily since changing Age and also simultaneously changing $Haircolor$ would be a realistic image generation. Here, we would seek a causal disentanglement map shown in Fig. S3.

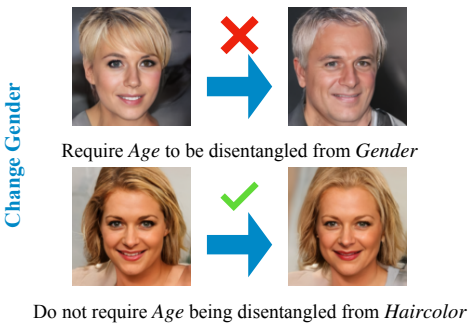


Figure S2: The disentanglement requirements in face examples

If we could get the causal disentanglement map, then we know that when the representations are fully learned, we can intervene on Age without changing the $Gender$ of the face. This motivates the need for a general approach to identifiability, compared to the scaling indeterminacy in Def. 6.4, which requires all variables to be disentangled from each other.

D.2 Examples for non-Markovian Factorization

In this section, we centralize theoretical results in relation to the theory presented in this paper.

Unless specified, we denote the natural log as $\log$.

We first provide more discussion about non-Markovian factorization Eq. (85). First, the concept C-component is formally defined as follows:

Definition 6.1 (Confounded Component)**.** Let $\{\mathbf{C_1}, \mathbf{C_2}, \ldots, \mathbf{C_k}\}$ be a partition over the set of variables $\mathbf{V}$, where $\mathbf{C_i}$ is said to be a confounded component (for short, C-component) of the selection diagram G_V if for every $V_i, V_j \in \mathbf{C_i}$ there exists a path made entirely of bidirected edges between V_i and V_j in G_V, and $\mathbf{C_i}$ is maximal. □

This construct represents clusters of variables that share the same exogenous variations regardless of their directed connections. The selection diagram in Figure 2 has a bidirected edge indicating the

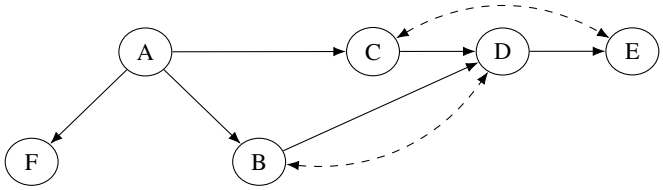


Figure S4: Causal graph with four C-components.

presence of unobserved confounders affecting the pairs (V_1, V_2) and contains two C-components, namely, $\mathbf{C_1} = \{V_1, V_2\}, \mathbf{C_2} = \{V_3\}$.

Akin to parents within a Markovian SCM, the c-components play a fundamental role in factorizing the joint distribution of the observed variables $\mathbf{V}$.

Let $<$ be a topological order $V_1, \ldots, V_n$ of the variables $\mathbf{V}$ in G^S. Then define the $\mathbf{Pa}_i^{\mathbf{T}+} = \mathbf{Pa}(\{V \in \mathbf{C}(V_i) : V \leq V_i\}) \setminus \{V_i\}$. The $\mathbf{Pa}^+(V_i)$ set consists of the nodes in the same c-component that are "$\leq$" in topological order as V_i, their corresponding parents, minus the node V_i itself. For instance, in Fig. S4, $Pa^+(E) = \{D, C, A\}$ and $Pa^+(D) = \{B, C, A\}$.

The general factorization formula Eq. (85) factorizes not only the joint observational distribution related to a causal graph, but also interventional distributions. With a hard intervention on $\mathbf{T}$, the factorization follows the corresponding graph is $G_{\overline{\mathbf{T}}}$, where the incoming arrows towards $\mathbf{T}$ are cut. This factorization encompasses both Markovian and non-Markovian SCM models. When there are no bidirected edges in the diagram, $\mathbf{Pa}_i^{\mathbf{T}+}$ reduce to $\mathbf{Pa}$ in $F_{\mathbf{T}}$.

Next, we introduce the Markov blanket, a fundamental idea in characterizing certain conditional independences in a causal graph [75, 85].

Definition 6.2 (Markov Blanket)**.** Let G be a causal graph over variables $\mathbf{V}$. A Markov blanket of a random variable $Y \in \mathbf{V}$ is any subset $V_1 \subseteq \mathbf{V}$ such that conditioned on V_1, Y is independent of all other variables.

$$Y \perp\!\!\!\perp \mathbf{V} \backslash V_1 | V_1$$

□

The Markov blanket is an important object that captures conditional independences between variables when conditioned on *all other variables* in the graph.

Definition 6.3 ("Global" Markov property of DAGs [86])**.** Consider a joint probability distribution, P over a set of variables $\mathbf{V}$ satisfies the **Markov property** with respect to a graph $G = (V \cup L, E)$ if the following holds for, $(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ disjoint subsets of V:

$$P(y|x, z) = P(y|z) \quad \text{if } Y \perp\!\!\!\perp X|Z \text{ in G (that is Y is d-separated from X given Z)}$$

□

The global Markov property maps graphical structure in causal directed acyclic graphs (DAGs) to conditional independence (CI) statements in the relevant probability distributions from data. The distributions we consider $\mathcal{P}$ are considered Markov wrt the graph, thus mapping d-separations in the graph to conditional independences in the distributions. This allows us to leverage factorizations, such as the one presented in Section 2.

D.3 Discussion about ΔQs resulting from different topological order

Here we revisit the definition of $\Delta\mathbf{Q}$ set.

Definition 3.1 ($\Delta\mathbf{Q}$ **Set**)**.** Given two distributions $P^{(j)}, P^{(k)}$ with interventions targets $\sigma^{(j)}$ and $\sigma^{(k)}$ containing $do(\mathbf{T})$, the $\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S]$ set (for short: $\Delta\mathbf{Q}^{(j),(k)}$, or $\Delta\mathbf{Q}$ if the index ic clear from the context) of the target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ is the remaining variables after comparison (i.e. Eq. 119),

$$\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S] = \tilde{\mathbf{V}} \cup \mathbf{Pa}^{\mathbf{T}+}(\tilde{\mathbf{V}}), \tag{123}$$

where $\tilde{\mathbf{V}} = \mathbf{C}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S])$. □

With the definition of $\Delta\mathbf{Q}$ set (Def. 3.1), we have shown that comparing $\mathbf{I}^{(2)}$ and $\mathbf{I}^{(1)}$ in Ex. 16, $\Delta\mathbf{Q}[\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{T}, G^S)]$ is $\{V_1, V_2, V_3\}$ no matter what topological order is used in the factorization. The following lemma argues that $\Delta\mathbf{Q}$ will not be influenced by the topological order.

Lemma 5 (The invariance of $\Delta\mathbf{Q}$ w.r.t order). *Given two distributions $P^{(j)}, P^{(k)}$ with interventions targets $\sigma^{(j)}$ and $\sigma^{(k)}$ containing $do(\mathbf{T})$. Let A and B be two different orders for factorizing $P(\mathbf{V})$. $\Delta\mathbf{Q}^{(j),(k)}$ are equivalent derived using A and B.* □

Proof. We will proof

$$\Delta\mathbf{Q}(\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S, \mathbf{T}) = \overline{\mathbf{P}a}(\mathbf{C}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S)) \tag{198}$$

where $\overline{\mathbf{Pa}}$ and $\mathbf{C}$ is applied in $G_{\overline{\mathbf{T}}}$. For short, the above equation is written as

$$\Delta\mathbf{Q} = \overline{\mathbf{Pa}}(\mathbf{C}(\Delta\mathbf{V})) \tag{199}$$

If this holds, $\Delta\mathbf{Q}$ does not depend on the topological order since the right side of Eq. (198) does not depend on the order but only the diagram G^S and Ψ.

Based on the definition of $\Delta\mathbf{Q}$ factors,

$$\Delta\mathbf{Q} = \mathbf{C}(\Delta\mathbf{V}) \cup \mathbf{P}a^{\mathbf{T}+}(\mathbf{C}(\Delta\mathbf{V})) \tag{200}$$

Recall the extended parents $\mathbf{Pa}^{\mathbf{T}+}(V_i) = \overline{\mathbf{Pa}}(\{V \in \mathbf{C}(V_i) : V \leq V_i\} \backslash \{V_i\}$. Then

$$\begin{aligned} \Delta\mathbf{Q} &= \mathbf{C}(\Delta\mathbf{V}) \cup \mathbf{Pa}^{\mathbf{T}^+}(\mathbf{C}(\Delta\mathbf{V})) \\ &= \overline{\mathbf{Pa}}(\cup_{W \in \mathbf{C}(\Delta\mathbf{V})} \{V \in \mathbf{C}(W) : V \leq W\}) \end{aligned} \tag{201}$$

Denote $\cup_{W \in \mathbf{C}(\Delta\mathbf{V})} \{V \in \mathbf{C}(W) : V \leq W\}$ as $\mathbf{Z}_{\leq}$.

We will prove

$$\overline{\mathbf{P}a}(\mathbf{C}(\Delta\mathbf{V})) \subseteq \Delta\mathbf{Q}, \tag{202}$$

which means every element in $\overline{\mathbf{P}a}(\mathbf{C}(\Delta\mathbf{V}))$ should be in $\Delta\mathbf{Q}$.

First, let $V_j \in \mathbf{C}(V_i)$. Then, we have

$$V_j \in \mathbf{C}(\Delta\mathbf{V}) \subseteq \cup_{W \in \mathbf{C}(\Delta\mathbf{V})} W = \cup_{W \in \mathbf{C}(\Delta\mathbf{V})} \{V \in \mathbf{C}(W) : V \leq W\} = \mathbf{Z}_{\leq} \tag{203}$$

Thus, for every $V_k \in \overline{\mathbf{P}a}(\mathbf{C}(\Delta\mathbf{V}))$

$$V_k \in \overline{\mathbf{P}a}(\mathbf{Z}_{\leq}) \tag{204}$$

And this implies Eq. (202) holds.

In another direction,

$$\begin{aligned} \Delta\mathbf{Q} &= \mathbf{C}(\Delta\mathbf{V}) \cup \mathbf{P}a^{\mathbf{T}+}(\mathbf{C}(\Delta\mathbf{V})) \\ &\subseteq \mathbf{C}(\Delta\mathbf{V}) \cup \overline{\mathbf{P}a}(\mathbf{C}(\Delta\mathbf{V})) \\ &= \overline{\mathbf{P}a}(\mathbf{C}(\Delta\mathbf{V})) \end{aligned} \tag{205}$$

Combine with Eq. (205) and (202), we know Eq. (198) holds. □

Then it guarantees that our identifiability result does not depend on the factorization order. Otherwise, with the same input Ψ and G^S, the CDM can be different.

D.4 The derivation example of ancestral disentanglement in Example 4.

Here we provide the derivation for ancestral disentanglement.

Example A32. Factorize $P(\mathbf{V})$ with the Markovian graph Fig. 7,

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2) \tag{206}$$

According to this factorization and the change-of-variable formula,

$$p^{(k)}(v_1) + p^{(k)}(v_2 \mid v_1) + p^{(k)}(v_3 \mid v_2) = p^{(k)}(\widehat{v}_1) + p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1) + p^{(k)}(\widehat{v}_3 \mid \widehat{v}_2) + |\det J_\phi| \tag{207}$$

Comparing $P^{(2)}$ and $P^{(1)}$,

$$p^{(2)}(v_1) - p^{(1)}(v_1) = p^{(2)}(\widehat{v}_1) - p^{(1)}(\widehat{v}_1) \tag{208}$$

since the intervention is applied to V_1. Taking the first derivative w.r.t $V_j (j = 2, 3)$,

$$0 = \frac{\partial p^{(k)}(\widehat{v}_1) - p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1} \frac{\partial \widehat{v}_1}{v_j} \tag{209}$$

When the coefficient

$$\frac{\partial p^{(k)}(\widehat{v}_1) - p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1} \neq 0, \tag{210}$$

we have

$$\frac{\partial \widehat{v}_1}{v_j} = 0 \tag{211}$$

which means V_1 is ID w.r.t. $\{V_2, V_3\}$. Comparing $P^{(3)}$ and $P^{(1)}$,

$$p^{(3)}(v_2 \mid v_1) - p^{(3)}(v_2 \mid v_1) = p^{(3)}(\widehat{v}_2 \mid \widehat{v}_1) - p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1) \tag{212}$$

since the intervention is applied to V_2. Taking the first derivative w.r.t V_3,

$$0 = \frac{\partial p^{(k)}(\widehat{v}_2 \mid \widehat{v}_1) - p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1)}{\partial \widehat{v}_1} \frac{\partial \widehat{v}_1}{v_3} + \frac{\partial p^{(k)}(\widehat{v}_2 \mid \widehat{v}_1) - p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1)}{\partial \widehat{v}_2} \frac{\partial \widehat{v}_2}{v_3} \tag{213}$$

Since $\frac{\partial \widehat{v}_1}{v_j} = 0$, the above equation reduce to

$$0 = \frac{\partial p^{(k)}(\widehat{v}_2 \mid \widehat{v}_1) - p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1)}{\partial \widehat{v}_2} \frac{\partial \widehat{v}_2}{v_3} \tag{214}$$

When the coefficient

$$\frac{\partial p^{(k)}(\widehat{v}_2 \mid \widehat{v}_1) - p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1)}{\partial \widehat{v}_2} \neq 0, \tag{215}$$

we have

$$\frac{\partial \widehat{v}_2}{v_3} = 0 \tag{216}$$

which means V_2 is ID w.r.t. V_3. □

D.5 The detailed examples of Proposition 3 and 4

Proposition 3 and 4 disentangle variables through comparing distributions. According to Sec. C.4 and C.5, the identification comes from the zero solutions of linear systems. Here we explicitly show how to build the linear system with concrete examples.

Example A33. (details for Example 21).

Consider the diagram in Fig. 11(c).

Suppose $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ with intervention targets

$$\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{V_2\}^{\Pi_1}, \mathbf{I}^{(3)} = \{V_3\}^{\Pi_1}, \mathbf{I}^{(4)} = \{V_1\}^{\Pi_1} \tag{217}$$

Consider $\mathbf{V}^{tar} = \{V_1, V_2, V_3\}$ and $\mathbf{V}^{en} = \mathbf{V} \backslash \{V_1, V_2, V_3\} = \{V_4\}$. Comparing $\{\mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\}$ with the baseline $\mathbf{I}^{(1)}$, the hard intervention variables are $\mathbf{T} = do[\mathbf{I}^{(1)}] = \{\}$. Then we have $\Delta\mathbf{Q}$ sets:

$$\Delta\mathbf{Q}^{(2),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(3),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(4),(1)} = \{V_1\}. \tag{218}$$

Condition [1] and [2] in Prop. 3 are satisfied straightforwardly. Condition [3] are also satisfied since $V_1 \in \Delta\mathbf{Q}^{(4),(1)}, V_2 \in \Delta\mathbf{Q}^{(2),(1)}$ and $V_3 \in \Delta\mathbf{Q}^{(3),(1)}$ Thus, $\mathbf{V}^{tar}$ is ID w.r.t $\mathbf{V}^{en}$ by Prop. 3.

Choosing order $V_1 < V_3 < V_2 < V_4$.

$$P(\mathbf{V}) = P(V_1)P(V_3)P(V_2 \mid V_1, V_3)P(V_4 \mid V_3) \tag{219}$$

as the factorization. By comparing distribution resulting from $\sigma^{(2)}$ and $\sigma^{(3)}$ with the baseline $\sigma^{(1)}$,

$$\log p^{(2)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3) = \log p^{(2)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) \tag{220}$$

$$\log p^{(3)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3) =$$
$$\log p^{(3)}(\widehat{v}_3) - \log p^{(3)}(\widehat{v}_3) + \log p^{(2)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) \tag{221}$$

$$\log p^{(4)}(v_1) - \log p^{(1)}(v_1) = \log p^{(4)}(\widehat{v}_1) - \log p^{(1)}(\widehat{v}_1) \tag{222}$$

Taking the first order partial derivative w.r.t. V_4:

$$\begin{aligned} 0 &= h_{2,1}\frac{\partial \widehat{v}_1}{\partial v_4} + h_{2,2}\frac{\partial \widehat{v}_2}{\partial v_4} + h_{2,3}\frac{\partial \widehat{v}_3}{\partial v_4} \\ 0 &= h_{3,1}\frac{\partial \widehat{v}_1}{\partial v_4} + h_{3,2}\frac{\partial \widehat{v}_2}{\partial v_4} + h_{3,3}\frac{\partial \widehat{v}_3}{\partial v_4} \\ 0 &= h_{4,1}\frac{\partial \widehat{v}_1}{\partial v_4} \end{aligned} \tag{223}$$

where

$$\begin{aligned} h_{2,i} &= \frac{\partial \log p^{(2)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_i} \quad \text{for } i = 1, 2, 3 \\ h_{3,i} &= \frac{\partial \log p^{(3)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_i} \quad \text{for } i = 1, 2, 3 \\ h_{4,1} &= \frac{\partial p^{(4)}(\widehat{v}_1) - \log p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1} \end{aligned} \tag{224}$$

Then since the coefficient is linear independent assumed in Assumption 6, we have

$$\frac{\partial \widehat{v}_1}{\partial v_4} = 0, \frac{\partial \widehat{v}_2}{\partial v_4} = 0, \frac{\partial \widehat{v}_3}{\partial v_4} = 0 \tag{225}$$

Then $V_1 = \tau_1(V_1, V_2, V_3)$, and $V_2 = \tau_2(V_1, V_2, V_3)$ and $V_3 = \tau_3(V_1, V_2, V_3)$. □

Then, we move to the examples of derivations of Prop. 4. The following example shows how variables within the $\Delta\mathbf{Q}$ set can be disentangled from each other.

Example A34. (Example 22 (continued).) Recall the given LSD G^S is shown in Fig. 2. and the 9 given intervention targets are

$$\mathbf{\Psi} = \{\{\}^{\Pi_1}, \{\{V_1^{\Pi_1}\} \times 4, \{V_2^{\Pi_1}, V_3^{\Pi_1}\} \times 4\} \tag{226}$$

Comparing $P^{(2)}, \ldots, P^{(9)}$ with $P^{(1)}$, we have

$$\log p^{(i)-(1)}(v_1) + \log p^{(i)-(1)}(v_2 \mid v_1) = p^{(i)-(1)}(\widehat{v}_1) + \log p^{(i)-(1)}(\widehat{v}_2 \mid \widehat{v}_1) \tag{227}$$

$$\log p^{(j)-(1)}(v_2 \mid v_1) + \log p^{(j)-(1)}(v_3 \mid v_2) = \log p^{(j)-(1)}(\widehat{v}_2 \mid v_1) + \log p^{(j)-(1)}(\widehat{v}_3 \mid \widehat{v}_2) \tag{228}$$

where $i = 2, 3, 4, 5$, $j = 6, 7, 8, 9$,. Taking the second order partial derivative w.r.t. V_1 and V_3:

$$\begin{aligned} 0 &= \sum_{p=1,2} h'_{i,p}\frac{\partial^2 \widehat{v}_p}{\partial v_1 \partial v_3} + \sum_{p=1,2} h''_{i,p}\frac{\partial \widehat{v}_p}{\partial v_1}\frac{\partial \widehat{v}_p}{\partial v_3} + h''_{i,1,2}\left(\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_3} + \frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3}\right) \\ 0 &= \sum_{p=1,2,3} h'_{j,p}\frac{\partial^2 \widehat{v}_p}{\partial v_1 \partial v_3} + \sum_{p=1,2,3} h''_{i,p}\frac{\partial \widehat{v}_p}{\partial v_1}\frac{\partial \widehat{v}_p}{\partial v_3} + h''_{i,2,3}\left(\frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3} + \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_3}\right) \end{aligned} \tag{229}$$

where $i = 2, 3, 4, 5$, $j = 6, 7, 8, 9$, and $k = 10, 11, 12, 13$, and the following are defined accordingly

$$
\begin{aligned}
h'_{i,p} &= \frac{\partial \log p^{(i)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_p} \quad \text{for } p = 1, 2, 3 \\
h''_{i,p} &= \frac{\partial^2 \log p^{(i)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_p^2} \quad \text{for } p = 1, 2, 3 \\
h''_{i,1,2} &= \frac{\partial^2 \log p^{(i)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_1 \partial \widehat{v}_2} \\
h'_{j,p} &= \frac{\partial \log p^{(j)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_p} \quad \text{for } p = 1, 2, 3 \\
h''_{j,p} &= \frac{\partial \log p^{(j)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_p^2} \quad \text{for } p = 1, 2, 3 \\
h''_{j,2,3} &= \frac{\partial \log p^{(j)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_2 \partial \widehat{v}_3}
\end{aligned}
\tag{230}
$$

There are 8 unknowns that come from Eqn. 229. We can write that as a vector $\beta \in \mathbb{R}^{12}$.

$$
\begin{aligned}
\boldsymbol{\beta} = [&\frac{\partial^2 \widehat{v}_1}{\partial v_1 \partial v_3}, \frac{\partial^2 \widehat{v}_2}{\partial v_1 \partial v_3}, \frac{\partial^2 \widehat{v}_3}{\partial v_1 \partial v_3}, \frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3}, \frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_3}, \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3} \\
&(\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_3} + \frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3}), (\frac{\partial \widehat{v}_2}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3} + \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_2}{\partial v_3})]
\end{aligned}
\tag{231}
$$

Rewriting Eq. (229), we have a linear system

$$
\begin{pmatrix}
h'_{2,1} & h'_{2,2} & 0 & h''_{2,1} & h''_{2,2} & 0 & h''_{2,1,2} & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
h'_{6,1} & h'_{6,2} & h'_{6,3} & h''_{6,1} & h''_{6,2} & h''_{6,3} & h''_{6,1,2} & h''_{2,2,3} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{pmatrix} \boldsymbol{\beta} = 0 \tag{232}
$$

The coefficient matrix is assumed with linear independent rows in Assumption 6. Since there are 8 rows, then the matrix is full rank, and we know $\boldsymbol{\beta} = 0$. Then.

$$
\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3} = 0 \tag{233}
$$

Then since $\frac{\partial \widehat{v}_1}{\partial v_1} \neq 0$,

$$
\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0 \tag{234}
$$

which implies that V_3 is ID w.r.t V_1 and V_1 is ID w.r.t V_3.

The following example illustrates the linear system built in Corol. 1.

Example A35. The factorization based on G^S choosing $\mathbf{T} = \{\}$ is

$$
P(\mathbf{V}) = P(V_1)P(V_3)P(V_3 \mid V_1, V_2) \tag{235}
$$

By comparing distribution resulting from $\sigma^{(2)}$ and $\sigma^{(3)}$ with the baseline $\sigma^{(1)}$, for $j = 2, 3, 4, 5$

$$
\begin{aligned}
&\log p^{(j)}(v_1) + \log p^{(j)}(v_3) - \log p^{(1)}(v_1) - \log p^{(1)}(v_3) && (236) \\
&= \log p^{(j)}(\widehat{v}_\cdot) + \log p^{(j)}(\widehat{v}_3) - \log p^{(1)}(\widehat{v}_1) - \log p^{(1)}(\widehat{v}_3) && (237)
\end{aligned}
$$

Taking the second order partial derivative w.r.t. V_1, V_3:

$$0 = \frac{\partial^2 \log p^{(j)}(\widehat{v}_1)p^{(j)} - \log p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1^2} \frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3} + \frac{\partial^2 \log p^{(j)}(\widehat{v}_3)p^{(j)} - \log p^{(1)}(\widehat{v}_3)}{\partial \widehat{v}_3^2} \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3}$$
$$+ \frac{\partial \log p^{(j)}(\widehat{v}_1)p^{(j)} - \log p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1} \frac{\partial^2 \widehat{v}_1}{\partial v_1 \partial V_3} + \frac{\partial \log p^{(j)}(\widehat{v}_3)p^{(j)} - \log p^{(1)}(\widehat{v}_3)}{\partial \widehat{v}_3} \frac{\partial^2 \widehat{v}_3}{\partial v_1 \partial v_3} \tag{238}$$

Then since the coefficient is linear independent assumed in Assumption 6, we have

$$\frac{\partial \widehat{v}_1}{\partial v_1}\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1}\frac{\partial \widehat{v}_3}{\partial v_3} = 0 \tag{239}$$

Then after permutation,

$$\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0 \tag{240}$$

which implies that V_3 is ID w.r.t V_1 and V_1 is ID w.r.t V_3. □

D.6 Case study on partial disentanglement in a Markovian setting

This next example works out the algebraic derivations for analyzing Fig. 11(a). This derivation is provided to provide additional intuition on the theory presented in Section 3, and how these concepts apply in a simple 3-dimensional latent causal graph.

Example A36 (Algebraic derivation of disentanglement in a simple 3-node chain graph). Given the graph shown in Figure 11(a), we can factorize the joint observational distribution of the latent variables

$$P(\mathbf{V}) = P(V_3|V_2)P(V_1|V_2)P(V_2) \tag{241}$$

By the probability transformation formula, we can similarly write the distribution in terms of its estimated sources via function $\phi = \hat{f}_{\mathbf{X}}^{-1} \circ f_{\mathbf{X}}$ for its distribution Q.

$$P(\mathbf{V}) = P(\phi_{V_3}(\mathbf{V})|\phi_{V_2}(\mathbf{V}))P(\phi_{V_1}(\mathbf{V})|\phi_{V_2}(\mathbf{V})P(\phi_{V_2}(\mathbf{V}))|detJ_\phi| \tag{242}$$

Now, consider the interventional distributions: $P(\mathbf{V}; \sigma_{V_3^{(1)}})$ and $P(\mathbf{V}; \sigma_{V_3^{(2)}})$. Here, we will use shorthand ϕ_i to indicate $\phi_{V_i}(\mathbf{V})$. Similarly, we can factorize the distribution $P(\mathbf{V}; \sigma_{V_3^{(1)}})$:

$$\begin{aligned}
&P(\mathbf{V}; \sigma_{V_3^{(1)}}) \\
&= P(V_3|V_2; \sigma_{V_3^{(1)}})P(V_1|V_2; \sigma_{V_3^{(1)}})P(V_2; \sigma_{V_3^{(1)}}) \quad \text{(Conditional independence)} \\
&= P(\phi_3|\phi_2; \sigma_{3^{(1)}})P(\phi_1|\phi_2; \sigma_{V_3^{(1)}})P(\phi_2; \sigma_{V_3^{(1)}})|detJ_\phi| \quad \text{(Probability transformation formula)}
\end{aligned}$$

Similarly, we can decompose the interventional distribution $P(\mathbf{V}; \sigma_{V_3^{(2)}})$. Now, comparing the log observational distribution with the log intervention $\sigma_{V_3^{(i)}}$, we get:

$$\begin{aligned}
&\log p(\mathbf{V}; \sigma_{V_3^{(i)}}) - \log p(\mathbf{V}) \\
&= \log p(V_3|V_2; \sigma_{V_3^{(i)}}) + \log p(V_1|V_2; \sigma_{V_3^{(i)}}) + \log p(V_2; \sigma_{V_3^{(i)}}) \\
&\quad - \log p(V_3|V_2) - \log p(V_1|V_2) - \log p(V_2) \\
&= \log p(V_3|V_2; \sigma_{V_3^{(i)}}) - \log p(V_3|V_2)
\end{aligned}$$

Where the last line applies the invariance of $P(V_i|V_j; \sigma_{V_k}) = P(V_i|V_j)$ if $(V_i \perp\!\!\!\perp V_k|V_j)_{G_{\overline{V_3}}}$. In the space mapped by ϕ, we similarly get:

$$
\begin{aligned}
&\log p(\phi; \sigma_{V_3^{(i)}}) - \log p(\phi) \\
&= \log p(\phi_3|\phi_2; \sigma_{3^{(i)}}) + \log p(\phi_1|\phi_2; \sigma_{V_3^{(i)}}) + \log p(\phi_2; \sigma_{V_3^{(i)}}) \\
&\quad - \log p(\phi_3|\phi_2) - \log p(\phi_1|\phi_2) - \log p(\phi_2) \\
&= \log p(\phi_3|\phi_2; \sigma_{3^{(i)}}) - \log p(\phi_3|\phi_2)
\end{aligned}
$$

When comparing the distributions of $\widehat{\mathbf{V}}$, interestingly the $\log$ of the determinant of the Jacobian cancels out. Combining the two, we get:

$$
\log p(V_3|V_2; \sigma_{V_3^{(u)}}) - \log p(V_3|V_2) = \log p(\phi_3|\phi_2; \sigma_{3^{(u)}}) - \log p(\phi_3|\phi_2) \tag{243}
$$

Taking the partial derivative now with respect to V_1, we get that the LHS equals 0 and the RHS becomes:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial V_1} \log p(\phi_3|\phi_2; \sigma_{3^{(i)}}) - \log p(\phi_3|\phi_2) \\
&= \frac{\partial \log p(\phi_3|\phi_2; \sigma_{3^{(i)}})}{\partial \phi_3} \frac{\partial \phi_3}{\partial V_1} + \frac{\partial \log p(\phi_3|\phi_2; \sigma_{3^{(i)}})}{\partial \phi_2} \frac{\partial \phi_2}{\partial V_1} \\
&\quad - \frac{\partial \log p(\phi_3|\phi_2)}{\partial \phi_3} \frac{\partial \phi_3}{\partial V_1} - \frac{\partial \log p(\phi_3|\phi_2)}{\partial \phi_2} \frac{\partial \phi_2}{\partial V_1} \quad \text{(Chain rule)} \\
&= \frac{\partial \phi_3}{\partial V_1} \left(\frac{\partial \log p(\phi_3|\phi_2; \sigma_{3^{(i)}})}{\partial \phi_3} - \frac{\partial \log p(\phi_3|\phi_2)}{\partial \phi_3} \right) \\
&\quad + \frac{\partial \phi_2}{\partial V_1} \left(\frac{\partial \log p(\phi_3|\phi_2; \sigma_{3^{(i)}})}{\partial \phi_2} - \frac{\partial \log p(\phi_3|\phi_2)}{\partial \phi_2} \right) \quad \text{(Collect terms)}
\end{aligned}
$$

Thus, we have two unknowns $\frac{\partial \phi_2}{\partial V_1}$ and $\frac{\partial \phi_3}{\partial V_1}$. Given the two interventions with different mechanisms on V_3 compared to the observational distribution, we have two equations that result in a 2-dimensional linear system. We are able to determine that $\frac{\partial \phi_2}{\partial V_1} = \frac{\partial \phi_3}{\partial V_1} = 0$ thus demonstrating that our approach disentangles $\hat{V}_3 = \phi_3(\mathbf{V})$ and $\hat{V}_2 = \phi_2(\mathbf{V})$ from V_1. □

D.7 Challenges for disentanglement in non-Markovian settings

Prior results suggest that in a Markovian setting, given a hard intervention on every node, the latent variables $\mathbf{V}$ are ID up to scaling indeterminancies according to Def. 6.4 [14, 21].

One would suspect that ID may still hold in non-Markovian ASCMs, but the following result states that even with one hard intervention per node, it is not possible to disentangle latent variables within the same c-component.

Lemma 6 (Challenges of identifability in non-Markovian causal models). *Consider the ASCM M that induces the diagram $V_1 \leftrightarrow V_2$. Suppose the intervention set includes an observational distribution, and hard interventions on both V_1 and V_2: $\mathbf{\Psi} = \langle \sigma_{\{\}}, do(\{V_1\}), do(\{V_2\}) \rangle$. Then V_1 is not ID w.r.t V_2 and vice versa.* □

Proof. We prove this by construction of a counter-example.

Consider an ASCM M^* that is constructed as follows:

$$
\mathcal{F}^* = \begin{cases} V_1 \leftarrow U_{1,2} \\ V_2 \leftarrow U_{1,2} + U_{V_2} \\ X_1 \leftarrow V_1, X_2 \leftarrow V_2 \end{cases}
$$
$$
U_{1,2} \sim \mathcal{N}(0,1), U_Y \sim \mathcal{N}(0,3)
$$

$$\sigma_{V_1} = P(\tilde{U_{V_1}}), \tilde{U}_{V_1} \sim \mathcal{N}(0, 2)$$
$$\sigma_{V_2} = P(\tilde{U_{V_2}}), \tilde{U}_{V_1} \sim \mathcal{N}(0, 7)$$

Consider a separate ASCM $M^{(1)}$ that is constructed as follows:

$$\mathcal{F}^{(1)} = \begin{cases} V_1^{(1)} \leftarrow -U_{1,2}^{(1)} \\ V_2^{(1)} \leftarrow 0.5U_{1,2}^{(1)} + 1.5U_Y \\ X_1 \leftarrow 1/3V_1^{(1)} + 2/3V_2^{(1)}, \\ X_2 \leftarrow 2/3V_1^{(1)} - 2/3V_2^{(1)} \end{cases}$$
$$U_{1,2}^{(1)} \sim \mathcal{N}(0, 3), U_{V_2}^{(1)} \sim \mathcal{N}(0, 1)$$
$$\sigma_{V_1} = P(\tilde{U_{V_1}}^{(1)}), \tilde{U}_{V_1}^{(1)} \sim \mathcal{N}(0, 6)$$
$$\sigma_{V_2} = P(\tilde{U_{V_2}}^{(1)}), \tilde{U}_{V_2}^{(1)} \sim \mathcal{N}(0, 7)$$

M^* and $M^{(1)}$ induce the same observational distribution $P(\mathbf{X}) \sim \mathcal{N}(0, \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix})$, and interventional distributions $P(\mathbf{X}; \sigma_{V_1}) \sim \mathcal{N}(0, \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix})$, $P(\mathbf{X}; \sigma_{V_2}) \sim \mathcal{N}(0, \begin{bmatrix} 1 & 0 \\ 0 & 7 \end{bmatrix})$

However, $V_1^{(1)} = V_1 - V_2$, which implies $V_1^{(1)}$ is not ancestral mixture or rescaling of the original V_1. Therefore, V_1 is not identifiable up to ancestral mixtures, or rescaling. □

D.7.1 ID within c-components Lemma 6 shows that even with one hard intervention on each node, it is not possible to disentangle variables within the same c-component. The next lemma provides a means of doing so using two hard interventions on the same node. This provides some intuition for the usefulness of hard interventions in the **CRID** setting.

Lemma 7 (Two hard interventions can disentangle within a c-component). *Let G^S be the LSD induced from a collection of ASCM $\boldsymbol{\mathcal{M}}$. Suppose $V_i, V_j \in \mathbf{V}$ are in the same c-component, and there are $L+1$ hard interventions distributions $\mathcal{P}_{V_i} = \{P^{(a_0)}, P^{(a_1)}, \ldots, P^{(a_L)}\}$ such that $V_i \in do[\mathbf{I}^{(a_l)}]$ and $\Delta\mathbf{Q}[\mathbf{I}^{(a_l)}, \mathbf{I}^{(a_0)}, V_i, G^S]$ are equivalent (denoted as $\mathbf{Q}$) for $l \in [L]$. When $V_j \notin \mathbf{Q}$ and if $L \geq |\mathbf{Q}|$, V_i is identifiable wrt V_j. When $V_j \in \mathbf{Q}$ and if $L \geq 2|\mathbf{Q}| + \delta_{\not\perp}$, V_i is identifiable wrt V_j.*

Proof. The result follows from the application of Proposition 3 and Proposition 4. □

Example A37. In most simple case. Let's have $do[\mathbf{I}^{(j)}] = do[\mathbf{I}^{(k)}] = V_i$ and $\Delta\mathbf{Q} = V_i$. Let $V_i, V_j \in \mathbf{C}_k$ be two arbitrary latent variables in the same c-component. By comparing distributions, we have

$$p_{V_i}^{(2)}(v_i) - p_{V_i}^{(1)}(v_i) = p_{V_i}^{(2)}(\widehat{v}_i) - p_{V_i}^{(1)}(\widehat{v}_i) \tag{244}$$

Taking partial w.r.t. V_j, we have

$$0 = \frac{p_{V_i}^{(2)}(\widehat{v}_i) - p_{V_i}^{(1)}(\widehat{v}_i)}{\widehat{v}_i} \frac{\widehat{v}_i}{v_j} \tag{245}$$

which implies $\frac{\widehat{v}_i}{v_j} = 0$.

Notice that this is not the only way to disentangle to variables in the C-Component. In Example 25, V_1 and V_2 are disentangled from each other without leveraging two hard interventions.

Work	Input					Output
	Assumptions		Data			Identifiability Goal
	Non-Markovian	Non-parametric	Interventions	Multiple Domains	Distr. Reqs.	
[6, 11–14]	✗	✗	✓	✗	1 per node	Scaling, Mixture or Affine Transformation
[7, 8, 10, 15, 16]	✗	✗	✗/✓	✗/✓	$2\|\mathbf{V}\|+1$	Scaling
[17, 18]	✗	✓	✓	✗	1 per node	Scaling
[19, 20]	✗	✓	✗/✓	✗/✓	1 per node	Scaling
[21]	✗	✓	✓	✗	1 per node	Scaling or Ancestral Mixture
[22]	✗	✓	✗	✓	$2\|\mathbf{V}\|+\|M_G\|+1$	Scaling or Mixture
This work	✓	✓	✓	✓	General	Causal Disentanglement Map

Table S1: For convenience, a table similar to Tab. 1 - a non-exhaustive list of identifiability results given knowledge of the latent graph

E Related Work

In Sec. 1 and 2, we have discussed causal disentangled representation learning works have different dimensions: *assumption*, *data*, and *identifiability goal* (as shown in Tab. S1 and Fig. S5). In this section, we will discuss related works in terms of these three dimensions.

E.1 Assumptions: Non-Markovianity and Non-Parametric ASCMs

The assumptions of the true underlying ASCMs can be divided into two parts. The first aspect of our work generalizes to the non-Markovian ASCM setting.

Structural Assumptions - Markovian vs non-Markovian ASCM The initial works related to disentangled representation learning can be traced back to Independent Component Analysis (ICA), where each generative factor is independent of each other (see Ex. 2 for details). Learning representations for independent generative factors is a special case of the setting where the latent causal graph is Markovian (i.e. no bidirected edges).

To our knowledge, all prior works in the literature do not account for the possible presence of confounding among the latent variables [7, 8, 10–12, 15–17, 19, 21–24, 52, 53, 55, 73, 76, 87]. However, the problem of confounding is pervasive in the causal inference literature and the real world. As such, a general understanding of disentangled causal representation learning proposed in this paper is important for implementing robust AI systems in the real world that do not naively ignore the possible presence of confounding among latent variables.

Confounders (represented by bidirected edges) in the latent causal model present complexities for learning disentangled causal representations. In addition, Lemma 6 demonstrates that applying existing results that only apply to Markovian ASCMs will not work in disentangling variables in a non-Markovian setting. Our proposed algorithm is able to handle confounders in the LCG leveraging the graphical criteria introduced in Props. 3, 4 and 5.

Non-parametric vs parametric SCM The second part of the assumption is about assuming parametric form for the SCM, or in the mixing function f_X.

Besides the mixing function, many works have also considered placing parametric assumptions on the SCM itself. In the earliest works of nonlinear ICA, [7, 8, 10, 17, 23, 24] assume an exponential family on the underlying latent variable distributions. [11] study the setting where the underlying SCM among the latent variables is a linear SCM. Our work assumes a fully non-parametric form of the SCM.

Non-parametric vs Parametric Mixing Function Many works have analyzed disentangled representation learning with differing assumptions on the mixing function f_X. [11, 13] show identifiability results in the presence of a linear mixing function. [12, 14] consider the setting where the mixing function is polynomial. [17, 22–24, 73] assume the mixing function is sparse, and leverage this sparse structure to achieve disentanglement. Our work assumes a fully non-parametric mixing function, and inline with the literature simply assumes that it is a diffeomorphism (bijection) to leverage the invertibility of the mixing function. In addition, we do not assume any sparsity structure in the mixing function.

E.2 Input Data: Arbitrary Heterogenous Domains and Interventions

Another axis of comparison with related works is the input distributions. Other works all either consider interventions within a single domain implicitly, or consider observations over different

domains [11–14, 17, 23, 24, 53, 55, 88]. The work proposed in this paper considers a general setting where any combination of distributions from heterogeneous domains may be input. This includes settings where observational data is not given, and settings where the interventions occur with the same mechanism across domains. Moreover, compared to existing work, which requires a certain number, or type of distributions are collected, we do not place explicit conditions on the input distributions. Instead, we provide an actionable algorithm that explicitly determines what will be disentangled under perfect optimization and infinite data. This allows researchers and scientists to explicate what disentangelement they desire within the final learned representations, and see what type of distributions might achieve these desired properties.

Next, we compare two bodies of work that have similarities to the setting we analyze, albeit they do not provide conditions for disentanglement from arbitrary combinations of distributions from multiple domains.

Causal Component Analysis [21] The closest work to ours is [21], which also presupposes knowledge of the latent causal graph and focuses solely on learning the unmixing function and the distributions of the causal variables. In [21], the results emphasized the need for interventions that occur only on a single node in the latent causal graph. However, Lemma 6 demonstrates challenges that are not addressed in the prior work. In addition, in our work, we propose a more general concept of identifiability in Def. 2.3. As a result, Thm. 1 makes significantly weaker assumptions to still achieve identifiability. Exs.2-6 illustrate also the nuances addressed by our work, but not in [21].

Another interesting concept introduced by [21] is the "fat-hand" interventions, which intervene on groups of variables within different groups, and the concept of "block-identifiability".

Here, we illustrate some examples and discussion on how our work compares with that of [21] that also provides sufficient conditions for identifiability given a causal graph over the latent variables. One key difference between our work is that we do not assume Markovianity in the underlying SCM, whereas they do.

Example A38 (Ex. 25 cont.)**.** This example continues off of Ex. 25. Consider the motivating example in healthcare depicted in Fig. 2. In hospitals from different countries Π^i and Π^j, drug treatment (V_1) affect length of ICU stay (V_2), and ultimately whether or not the patient lives or dies (V_3). Our task is to learn representations of the high-level latent variables (V_1, V_2, V_3) that are not collected given a collection of low-level input such as EMRs, imaging and bloodwork data (high-dimensional data $\mathbf{X}$). In existing work [21], there are no guarantees that variables $\{V_2, V_3\}$ are disentangled from their ancestor V_1 from soft interventions per nodes. However, Proposition 3 demonstrates two comparisons are enough to disentangle both V_2 and V_3 from their ancestor V_1. □

Even in the Markovian setting, where the LSD does not contain bidirected edges, our results can also guarantee identifiability in this setting.

Example A39 ([21] approach)**.** Given the graph shown in Figure 11(a), [21] requires an observational, and tuple of intervention sets $\mathbf{\Psi} = \langle\{\}, \{V_1\}, \{V_2\}, \{V_3\}\rangle$. Provided these four distributions, there is still no disentanglement of $\hat{V}_3$ with respect to any variables, $V_i \in \mathbf{V}$. □

Causal Representation Learning from Multiple Distributions: A General Setting [22] Another approach to achieving disentanglement among the latent variables is similar to nonlinear-ICA, but leverages the conditional independence properties within a Markov Network of the causal graph. Then the proof strategy of [22] considers the second order derivative, which leverages the conditional independence constraints.

However, this results in a required $2d + |\mathcal{E}(M_G)| + 1$ number of distributions that satisfy Assump. 6. In addition, this strategy states that in a collider graph $V_1 \to V_2 \leftarrow V_3$, that V_1 is not ID wrt V_2, and V_3 is not ID wrt V_2.

Another example, continues off of Ex. 25.

Example A40 (Ex. 25 cont.)**.** This example continues off of Ex. 25. Consider the motivating example in healthcare depicted in Fig. 2. In hospitals from different countries Π^i and Π^j, drug treatment (V_1) affect length of ICU stay (V_2), and ultimately whether or not the patient lives or dies (V_3). Our task is to learn representations of the high-level latent variables (V_1, V_2, V_3) that are not collected given a collection of low-level input such as EMRs, imaging and bloodwork data (high-dimensional data $\mathbf{X}$). According to [22], 10 distributions can disentangle V_3 from V_1 when $V_3 \perp\!\!\!\perp V_1 \mid V_2$. However,

Proposition 3 demonstrates two comparisons are enough to disentangle both V_2 and V_3 from their ancestor V_1.

E.3 Output Goal: General Disentanglement

Another axis of comparison is the goal of disentanglement. Most works previously propose sufficient conditions for full disentanglement, or the disentanglement up to some structural mixture (e.g. ancestral mixtures, or neighbors within a Markov Network) [7, 8, 10–12, 15–17, 19, 21, 22, 52, 53, 55, 73, 76, 87]. However, our work proposes the causal disentanglement map as a goal of disentanglement. Def. 2.3 generalizes all previous notions of disentanglement, and implies full disentanglement, or disentanglement up to ancestral mixtures as a special case.

Recently, we were made of the line of work from [23, 24]. These are closest work to ours in this axis of generalizing the definition of identifiability. In their paper, partial disentanglement is proposed that allows for entanglement of latent generative factors via a linear affine transformation. In this context, the partial disentanglement proposed in [23, 24]. In addition, [24]'s functional dependency graph is equivalent to our proposed causal disentanglement map. However, our proposed Def. 2.3 is still more general than the partial disentanglement proposed in [23] because we can account for partial disentanglement beyond linear affine transformations. As specified in the above two sections, our work is also general in the axes of the structural assumptions (e.g. non-Markovian vs Markovian, and non-parametric vs parametric SCMs), and the input distributions.

Comparing different definitions of disentanglement First, we review the identifiability/disentanglement definition that is commonly used in much of the literature. It is what is known as scaling identifiability, and is a special case of our ID definition in Def. 2.3.

Definition 6.4 (Scaling indeterminancy). Consider a collection of ASCM $\mathcal{M}$ that induces an LSD G^S and a collection of distribution $\mathcal{P}$. We say $\mathbf{V}$ is identifiable up to scaling indeterminacy if for every $\widehat{\mathcal{M}}$ matches with the G^S and $\mathcal{P}$, there exists functions $\{h_1, \dots, h_d\}$ such that $\widehat{V}_i = \mathbf{h}_i(V_i), i \in [d]$, where h_i is a diffeomorphism in $\mathbb{R}$. □

We also consolidate other definitions of identifiability from the literature using the notion of an ASCM. We have already defined identifiability up to scaling ambiguity in Def. 6.4.

Corollary 2 (Scaling ID is a case in general ID). *Let $\mathcal{M}$ be a collection of ASCM with G^S the LSD over the latent causal variables $\mathbf{V}$. If $\tilde{V} \subseteq \mathbf{V}$ is identifiable up to scaling indeterminacy, then it is identifiable wrt $\mathbf{V} \backslash \tilde{V}$.*

Proof. The proof follows from the application of Def. 2.3 and Def. 6.4. □

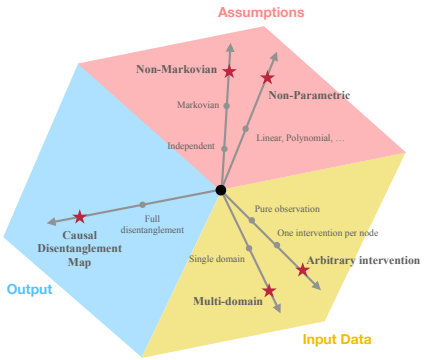


Figure S5: Replication of Fig. 1 for convenience - causal disentanglement representation learning tasks.

Definition 6.5 (Identifiability up to ancestral mixtures [21]). Let $\mathcal{M}$ be a collection of ASCM with G^S the LSD over the latent causal variables $\mathbf{V}$. We say a variable $\tilde{V} \in \mathbf{V}$ is identifiable up to ancestral mixtures if for every $\widehat{\mathcal{M}}$ matches with the G^S and $\mathcal{P}$, there exists functions $\{h_1, \dots, h_d\}$ such that $\widehat{V}_i = \mathbf{h}_i(\overline{\mathbf{Anc}}(V_i)), i \in [d]$. □

Corollary 3 (Ancestral ID is a case in general ID). *Let M be a collection of ASCM with G the LSD over the latent causal variables $\mathbf{V}$. If $\tilde{V} \subseteq \mathbf{V}$ is identifiable up to ancestral mixtures, then it is identifiable wrt $\mathbf{V} \backslash \mathbf{Anc}(\tilde{V})$.*

Proof. The proof follows from the application of Def. 2.3 and Def. 6.5. □

The following definitions are inspired by the identifiability results from [22].

Definition 6.6 (Intimate Neighbor Set). We say $\Psi_{V_i} := \{V_j \mid j \neq i$, but V_j is adjacent to V_i and all other neighbors of V_i in M_G. □

The intimate neighbor set for a variable dictates a set of neighbors that are adjacent to all of that variable's neighbors. It is used in the following definition from [22].

Definition 6.7 (Identfiability up to intimate neighbor set of Markov Network [22]). Let $\boldsymbol{\mathcal{M}}$ be a collection of ASCM with G^S the LSD over the latent causal variables $\mathbf{V}$. We say a variable $\tilde{V} \in \mathbf{V}$ is identifiable up to intimate neighbors in the Markov Network if for every $\widehat{\boldsymbol{\mathcal{M}}}$ matches with the G^S and $\mathcal{P}$, there exists functions $\{h_1, \ldots, h_d\}$ such that $\widehat{V}_i = \mathbf{h}_i(\psi(M_G, V_i)), i \in [d]$, and M_G is the Markov network of G and $\psi(M_G, V_i)$ is the intimate neighbor set of V_i in M_G. □

Corollary 4 (Intimate Neighbor Markov Network ID is a case in general ID). *Let M be a collection of ASCM with G the LSD over the latent causal variables $\mathbf{V}$. If $\tilde{V} \subseteq \mathbf{V}$ is identifiable up to intimate neighbor set of the Markov Network, then it is identifiable wrt $\mathbf{V} \backslash \phi(MN(G); \tilde{V})$.*

Proof. The result follows from the application of Def. 2.3 and Def. 6.7. □

Thus, we showed that each of these identifiability definitions imply a general ID for a non-trivial subset of latent variables $\tilde{\mathbf{V}} \subseteq \mathbf{V}$ with respect to $\mathbf{V}^{en} \subset \mathbf{V}$.

E.4 Future Work: Causal representation learning with unknown latent causal structure

In many prior works, the goal has been not only identifiability of the underlying latent variables, but also the discovery of the causal relationships among the latent variables [4, 22, 51]. That is, the latent causal graph is unknown. The work proposed in this paper is a foundation for the first step of causal representation learning, i.e. identifying the distributions of the latent causal variables. It would be interesting future work to explore how the results proposed in this paper extend to the case when the latent causal graph is unknown.

F Experimental Results

F.1 Synthetic data-generating process

We generate data according to latent causal diagrams shown in Fig. 11. Specifically, we analyze the chain graph $V_1 \to V_2 \to V_3$, and collider graph $V_1 \to V_2 \leftarrow V_3$ with different input distributions.

Each graph is constructed according to an ASCM, where the latent variables are related linearly:

$$V_i := \sum_{j \in Pa_i} \alpha_{i,j} V_j + \epsilon_i$$

where linear parameters are drawn from a uniform distribution $\alpha_{i,j} \sim U(-a, a)$, and the noise is distributed according to the standard normal distribution $\epsilon_i \sim \mathcal{N}(0, 1)$.

Generating Multiple Domains To generate a new domain, where $S^{i,j} \to V_i$ indicates a change in mechanism for V_i due to the change in ASCMs between M^i and M^j, we start from the first ASCM generated, and then we modify the distribution of the noise variable with a mean-shift and variance scaling.

Generating Interventions Within Each Domain To generate interventional datasets within each domain $\Pi^i \in \mathbf{\Pi}$, we modify the $\mathbf{M}^i \in \boldsymbol{M}$ by additionally modifying the SCM, and shifting its mean for a variable. Therefore for distribution k in Π^i, with hard intervention $\mathbf{I}$, we will have:

$$V_k := \epsilon'_k, \quad \text{with } \epsilon'_k \sim \mathcal{N}(\mu_k, \sigma_k),\ \forall V_k \in \mathbf{I}$$

such that μ_k is not within $+/-1$ of any other distribution for variable $V_k \in \mathbf{V}$. This ensures the given distribution changes sufficiently (Assump. 6). With a soft intervention $\mathbf{J}$ that is not hard:

$$V_k := \sum_{j \in Pa_k} \alpha_{i,j} V_k + \epsilon'_k, \quad \text{with } \epsilon'_k \sim \mathcal{N}(\mu_k, \sigma_k),\ \forall V_k \in \mathbf{J}$$

For each distribution over $\mathbf{V} \in \mathbb{R}^d$, we generate 200,000 data points resulting in $d \times 200,000$ data points in total for N total distributions.

Mixing function In order to generate the low-level data $\mathbf{X}$, we will apply a mixing function $f_{\mathbf{X}}$ to the generated latent variables $\mathbf{V}$. Following [21, 51], to generate an invertible mixing function, we will use a multilayer perceptron $\mathbf{f}_{\mathbf{X}} = \sigma \circ \mathbf{A}_M \circ ... \circ \sigma \mathbf{A}_1$, where $\mathbf{A}_M \in \mathbb{R}^{d \times d}$ for $m \in [1, M]$ denotes invertible linear matrices and σ is an element-wise invertible nonlinear function. In our case, we will use the tanh functio as done in [76]:

$$\sigma(x) = tanh(x) + 0.1x$$

In addition, each sampled matrix $\mathbf{A}_i$ is re-drawn if $|\det \mathbf{A}_i| < 0.1$. This ensures that the linear maps are not ill-conditioned and close to being singular. Once the mixing function is drawn for a given simulation, it is fixed across all domains and interventions according to Def. 2.1, and then $\mathcal{P}$ is drawn according to all ASCMs instantiated.

F.2 Model

We train invertible MLPs with normalizing flows. The parameters of the causal mechanisms are learned while the causal graph is assumed to be known. We leverage the implementation in [21], and extend it for our experiments.

The encoder is trained with the following objective that estimates the inverse function f^{-1}, and the latent densities $P(\mathbf{V})$ reproducing the ground-truth up to certain mixture ambiguities (c.f. Lemmas 3, 7). The encoder parameters is estimated by maximizing the likelihood..

Normalizing flows We use a normalizing flows architecture [77] to learn an encoder $\mathbf{g}_\theta : \mathbb{R}^d \to \mathbb{R}^d$. Therefore, the observations $\mathbf{X}$ will be the result of an invertible and differentiable transformation:

$$\mathbf{X} = \mathbf{g}_\theta(\mathbf{V})$$

Specifically, g_θ will comprise of Neural Spline Flows [78] with a 3-layer feedforward neural network with hidden dimension 128 and a permutation in each flow layer.

Base distributions Normalizing flows require a base distribution. We leverage one baseline distribution per sampled dataset, $(\hat{p}^k_\theta)_{k \in [d]}$ over the base noise variables $\mathbf{V}$. The conditional density of any variable is given by:

$$\hat{p}^k_\theta(v_i | \mathbf{Pa_i}) = \mathcal{N}\left(\sum_{j \in Pa_i} \hat{\alpha}_{i,j} v_j, \hat{\sigma}_i \right)$$

where the parameters are replaced by their corresponding counterparts if there is a change-in-domain, or an intervention applied. When a hard intervention is applied, we have that:

$$\hat{p}^k_\theta(v_i) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i \Big)$$

F.3 Training details

We use the ADAM optimizer [79].We start with a learning rate of 1e-4. We train the model for 200 epochs with a batch size of 4096.

The learning objective is expressed as:

$$\theta^* = \arg\max_{\theta} \sum_{k=0}^{N} \left(\frac{1}{n_k} \sum_{n=1}^{n_k} \log p_\theta^k(\mathbf{X}^{(k)})\right)$$

where n_k represents the size of the dataset P^k, which is 200,000 in our simulations. We perform 10 training runs over different seeds for each experiment, and show the distributions of the mean-correlation coefficient (MCC). Using the output of Alg. 1, we compare variables that are expected to be entangled and disentangled. We use NVIDIA H100 GPUs to train the neural network models.

F.4 Evaluation metrics

The output of our trained model is $\hat{\mathbf{V}} = g_\theta(\mathbf{X})$, which is a d-dimensional representation. We will compare this representation with our ground-truth latent variable distributions $\mathbf{V}$ by computing the mean correlation coefficients (MCC) between the learned and ground-truth latents. We expect there to be an overall lower MCC for variables that are predicted to be disentangleable by Alg. 1 relative to variables that are not deemed disentangleable.

Note that our algorithm is not shown to be complete, so there may be variables that are disentangled at the end of our training process that are not captured by the output of Alg. 1. Characterizing when this occurs and coming up with a complete theoretical characterization of disentanglement is a line for future work.

For the evaluation, we follow a standard evaluation protocol taken in prior work [18]. We expect low MCC values when predicting variables that are disentangled, and higher MCC values when predicting variables that are still entangled.

F.5 Limitations

A major limitation of normalizing flows is that the input and output dimensions of the encoder must be the same. This is due to the fact that we wish to constrain the layers to be invertible transformations. It is easy to define invertible transformations for the same input/output dimensions, but it is non-trivial to do so when input/output dimensions vary widely.

Besides the technical limitations of the implementation, it is important to note that our theoretical results are asymptotic results. The theory claims we can achieve ID when the neural network is trained to zero error. However, in practice, this is not always simple to do and may require hyperparameter tuning and a very large sample size.

For example, when we consider Fig. 15, we observe that the disentanglement of (b,c) is significantly better than (a,d). In the experiment involving the collider graph from Fig. 11(b), we sample four distributions each with 200,000 samples, and thus we have almost 2x the data points compared to the settings in Fig. 15(a,c). We illustrate this point to emphasize that there is no correct way to set the sample sizes, hyperparameters, or model architecture as each simulation will be different. We chose a sample size, model architecture, and default hyperparameters based on prior literature [21] instead of biasing our experimental results by tuning significantly for each simulation.

F.6 Discussion of Results

In Fig. S6, we show the MCC values for each learned latent representation $\hat{\mathbf{V}}$ and the corresponding ground-truth latents $\mathbf{V}$ for the three different LSDs shown in Fig. 11. Based on the causal disentanglement map (CDM) output from the CRID algorithm, the disentangled variables are shown in red, while the entangled variables are shown in gray.

In Fig. S6(a), the $MCC(\hat{V}_3, V_1)$ is low relative to the $MCC(\hat{V}_3, V_3)$, which is predicted by the CRID algorithm's CDM output (right plot). This suggests that V_1 is disentangled from V_3. In addition, we

observe that all MCC values wrt $\hat{V}_1$ are relatively similar, which makes sense as we do not obtain any disentanglement wrt V_1 (left plot). CRID also predicts that V_2 is ID wrt V_1 (middle plot). However, we observe quite a large range of MCC values, possibly due to variance, default hyperparameter settings, or insufficient sample size. Importantly, this experiment verifies that two soft interventions on V_3 in the chain graph of Fig. 11(a) can ID V_3 wrt V_1, whereas previous literature suggested that V_3 is not ID wrt V_1 because $V_1 \in \mathbf{Anc}(V_3)$ [21].

In Fig. S6(b), we now have an observational, two soft interventions on V_3, and a hard intervention on V_2. In addition to ID V_2 wrt V_3 (middle plot), we are also able to obtain full disentanglement of V_1 from $\{V_2, V_3\}$ (left plot). Interestingly, we are able to fully disentangle the representation for V_1 without intervening on it. This is the first theoretical (and empirical) result to our knowledge that shows this in a causal representation learning setting.

In Fig. S6(c), we have an observational and four interventional distributions applied on $\{V_1, V_3\}$ all with different mechanisms. We observe that V_1 and V_3 are fully disentangled. $MCC(\hat{V}_3, V_3) > MCC(\hat{V}_3, \{V_1, V_2\})$, and $MCC(\hat{V}_1, V_1) > MCC(\hat{V}_1, \{V_2, V_3\})$. CRID does not predict disentanglement for the V_2 representation (middle plot), yet interestingly we still see some disentanglement. [21] analyzes a similar setup using "fat-hand interventions", and the corresponding theory does predict V_1 and V_3 is ID wrt V_2. However, we also disentangle V_1 and V_3 from each other using many interventions. [22] presents a similar approach by leveraging $2d + |\mathcal{E}(M_G)| + 1$ distributions that "sufficiently change" (i.e. Assumption 6) to disentangle variables. However, the corresponding theory suggests that V_1 and V_3 are still entangled because they are adjacent in the Markov Network of G (M_G). These results demonstrate theoretically (and empirically) that V_1 and V_3 are in fact disentangled from each other in a fundamentally important causal graph (i.e. the collider).

In Fig. S6(d), we consider disentanglement in a non-Markovian LSD. We leverage two hard interventions on V_3 (c.f. Lemma 7), and verify that even without observational distributions and the challenging setting of confounding among the latent variables, we can achieve disentanglement of V_3 wrt all other variables. $MCC(\hat{V}_3, V_3) > MCC(\hat{V}_3, \{V_1, V_2, V_4\})$, which is predicted by the CRID algorithm's CDM output (3rd plot from left). As expected, V_1 and V_2 are still fully entangled with all other variables (1st and 2nd plot from left).

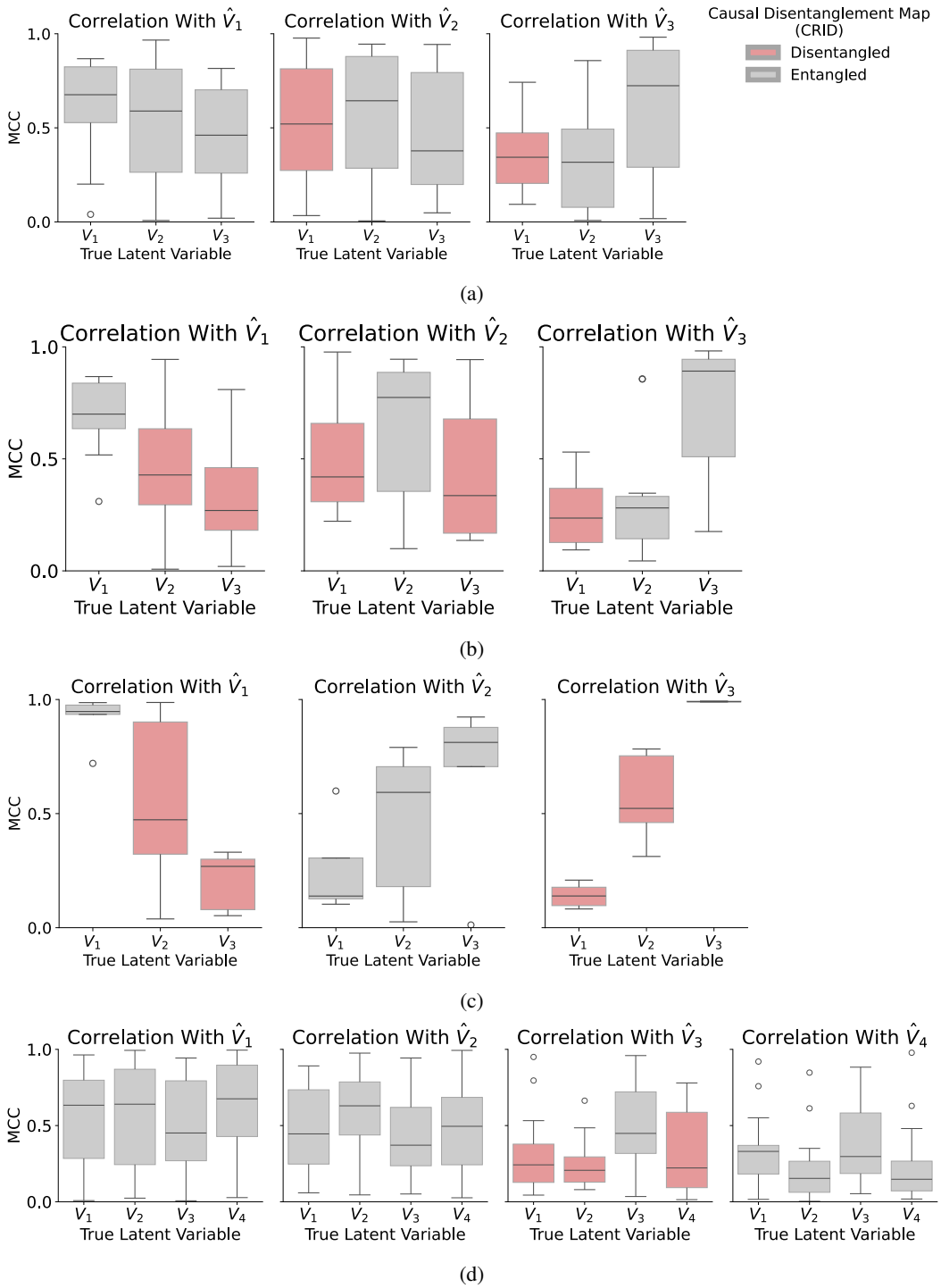


Figure S6: Mean correlation coefficient (MCC) of latent ground truth variables with the learned representation $\hat{\mathbf{V}}$, and expected disentanglement (red) according to the **CRID** algorithm. Each plot corresponds to an experimental setting using the graphs shown in Fig. 11: chain graph with two interventions on V_3 (a). chain graph with two interventions on V_3 and a hard intervention on V_2 (b), collider graph with four interventions on $\{V_1, V_3\}$ (c) and the non-markovian graph with two hard interventions on V_3 (d).

G Broader Impact and Forward-Looking Statements

The development of learning disentangled causal representations has the potential to improve our understanding of complex systems, and to help identify the generative factors for many important problems. By improving our ability to leverage observational and interventional data across multiple domains, this work could ultimately lead to more realistic generative AI. Beyond the machine learning and causal inference community, we expect that our results will enable fundamental contributions in various fields, including biology [89], epidemiology [90], economics [37] and neuroscience [38].

H Frequently Asked Questions

Q1. What's the learning goal of the paper? This work claims to be causal representation learning, but why do we not learn the structure over the latent variables while assuming it as given?

Answer. Causal representation learning may comprise of two parts: i) learning the distributions of the latent variables and ii) learning the causal structure among these latent variables. Learning the distribution over latent variables is a non-trivial problem, especially in the context of non-Markovian ASCMs and the general multi-domain context. For example, consider nonlinear ICA, where the structure of the latent variables is the fully disconnected graph. It was shown to be non-ID with only iid data [9]. Although ID results eventually came about for nonlinear ICA, it was nontrivial to derive. In the same spirit, we seek to analyze the most general setting possible when assuming knowledge of the causal structure. This is analogous to the causal inference task of identification [91, 92], where the goal is to determine if a causal effect over observed variables is estimable given infinite data from some given distributions on the observed variables. Put similarly, our work's goal is to determine if a latent variable $V_i \in \mathbf{V}$ is disentangleable given infinite data from some given distributions over the observed variables $\mathbf{X}$. In traditional causal inference, when the causal graph is unknown, then one is typically interested in causal discovery, or structure learning of the graph over the observed variables given distributions over the observed variables. Future work may assume that even the latent causal structure is unknown, and pursue the structure learning of the LCG given distributions over the observed variables.

Q2. Is it reasonable to expect that the causal diagram is available? How do you get the graph?

Answer. The assumption of the causal diagram is made out of necessity. Even existing methods is able to learn the casual diagram at the same time, however, the setting is more restricted. For example, the SCM should be Markovian and the intervention data per node should be given. In our setting, the underlying SCM can be non-Markovian and the given data can be any observational and interventional data from an arbitrary domain. In the general setting, even when the generative factors are all observed, learning the causal diagram task (structural learning task) is still difficult. Interestingly, recovering the full true diagram is even impossible, and existing works aim to recover an equivalence class of diagrams [32, 93–95]. Thus, in this general setting for causal representation learning, we first provide identification results given a causal diagram and leave structure learning for future work.

We follow closely to the disentangled representation learning works that assume the causal diagram is given. ICA/Nonlinear ICA assumes the diagram G is given and restricts the setting where no edges are in G. Later, [18] assumes focus on disentangling the content variable from the style variable and assumes the knowledge of the diagram is given ($Content$ is the ancestral of $Style$). Recently, [21] focuses on the setting that the given diagram is Markovian. We extend the setting to non-Markovain settings. Notice that our generalization is not only related to diagram assumption but involves more general assumption, data, and output (please see Sec. 1, Tab. 1 and Tab. 2 for details.)

In practice, knowledge of the latent causal graph is typically provided by domain experts, or a modeling assumption. As an example of a realistic setting where the latent causal graph can be assumed, consider generating realistic face images [27]. Here, the latent causal structure comprises of Gender, Age, and Hair Color. Knowledge of the graph is provided due to our understanding of what comprises realistic changes in a face. For a detailed discussion on this, see Appendix Section D.1.

Q3. Why CRID (Alg. 1) only takes intervention targets $\mathbf{\Psi}$ and LSG G^S as input? Do you need distributions $\mathcal{P}$? If not, how do you learn representations?

Answer. CRID leverages the intervention targets $\mathbf{\Psi}$ and the LSG G^S to determine the invariant and changing factors when considering the generalized factorization of probability distributions Markov relative to the provided graph. These invariant and changing factors are what give rise to the theory we develop in Section 3. The CRID algorithm leverages this theory to provide an identifiability algorithm, which answers the question: If we fully learn a representation $\widehat{\mathbf{V}}$ (given the diagram and the distributions), which variables are expected to be disentangled with which variables? This is an asymptotic question and assumes the representation is fully learned.

To fully learn the representations, one can search a proxy model that matches $\mathcal{P}$ and $\mathcal{G}^S$ and the $\widehat{\mathbf{V}}$. Then the proxy model is the learned representation. We do this in the Experiments Section, but note we do not claim that this method of learning the representations is superior to any prior work. Specifically, we implement an approach to train a neural model that is compatible with the diagram to match the given distribution based on normalizing flows. Recently, many graphical constraints proxy neural models have been proposed, and they are trained to fit the given distribution for causal representation learning and downstream tasks [20, 27, 55, 70, 88, 96]. Without our work, one can still try to use these models to learn representations. However, there is no guarantee about how these learned representations is entangled with each other. Our work is the first one to provide general answers for this identification problem. This process can be compared with the identification and estimation problem in classic causal inference. The identification of a specific query given a causal diagram can be answered in symbolic ways [83, 91, 97–100], and then if the query is identifiable, one can take the distribution (or data) as input and use estimation methods to obtain the estimated query. Without the identifiability result, there are no guarantees for the estimation.

Q4. Why not just use observational distributions in each domain as the baseline in the CRID algorithm described in Section 4?

Answer. One may surmise that this is not efficient and propose to choose the observational distribution in each domain alternatively. However, we argue that this enumeration is needed from two perspectives. First, the observational distributions, namely the idle interventions, are not always given. Second, comparing with observational distributions is not guaranteed to offer diverse $\Delta\mathbf{Q}$ sets. For example, consider intervention targets $\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{V_1^{\Pi_1,[1]}, V_2^{\Pi_1,[1]}\}, \mathbf{I}^{(3)} = \{V_1^{\Pi_1,[1]}, V_2^{\Pi_1,[2]}\}$ all applied to the same domain Π_1. Choosing $\mathbf{T} = \{\}$ and comparing $\mathbf{I}^{(2)}$ and $\mathbf{I}^{(3)}$ with the idle intervention $\mathbf{I}^{(1)}$,

$$\Delta\mathbf{Q}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, \mathbf{T}] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(1)}, \mathbf{T}] = \{V_1, V_2\}. \tag{246}$$

Comparing $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(3)}$ with the idle intervention $\mathbf{I}^{(2)}$,

$$\Delta\mathbf{Q}[\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{T}] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(2)}, \mathbf{T}] = \{V_2\}. \tag{247}$$

Then using Proposition 3, it is possible to disentangle V_2 from V_1 with the latter choice. This demonstrates that the observational distribution is not always necessarily the best baseline. Furthermore, consider the challenge of disentangling V_1 from V_2 in the LCG $V_1 \dashleftarrow\dashrightarrow V_2$. As Lemma 7 demonstrates, one can compare two hard intervention distributions on V_1 to achieve ID of V_1 wrt V_2. In this case, one would not even need the observational distribution.

Q5. Why distinguish domains and interventions? Are they not the same thing?

Answer. The literature has typically conflated domains and interventions in the context of causal inference.

Many examples across scientific disciplines demonstrate that the notions of domain/environment and interventions are distinct. For example, when making inferences about humans based on data from bonobos, this distinction becomes clear. The difference between the two species is depicted as the environment/domain in this context. A scientist might perform an intervention on a bonobo's kidney (specifically, what we're representing as Z), and try to determine the effect of medication (X) on fluid equilibrium in the body

(Y). Although we could intervene on Z in bonobos and observe its effect on X and Y, our ultimate goal might be to understand the effect of X on Y in humans. It's generally invalid to conflate these two qualitatively different indices, a point first noted by [63] in the context of transportability analysis. The distinct environments exist regardless of any intervention, such as medication. Also, an intervention on kidney function is different across the two species. [63] formalized this setting, introducing clear semantics for the S-nodes (environments) that essentially offer a combined representation for both environments. With this foundation, we can now address the more general problem of analyzing data generated from interventions across multiple domains in the latent space.

We point the reader to Appendix Section A.3 for a discussion and some examples of how CRID leverages this distinction.

Q6. Is the relaxation of Markovianity important? Since all $\mathbf{V}$ are already latent, can one regard the confounding $\mathbf{U}$ as $\mathbf{V}$ to transfer the model in the non-Markovianity setting to a Markovanity model?

Answer. Yes, the distinction between Markovianity and non-Markovianity is important both qualitatively and quantitatively.

Qualitatively, consider the following example in healthcare, where one has access to high-dimensional T1 MRI scans. Let the LCG comprise of Drug Treatment $\rightarrow$ Outcome, but they are confounded by socioeconomic status (Drug Treatment $\dashleftarrow\dashrightarrow$ Outcome). The drug treatment and outcome are visually discernable on the MRI. However, socioeconomic status does not directly impact how the MRI appears, except through how it impacts the drug treatment efficacy or outcome. The socioeconomic status is therefore an unaccounted confounder in the LCG, and it is important to model this spurious association. If unaccounted for, one may assume that it is possible to disentangle Drug Treatment and Outcome leveraging existing ID results in the literature [11, 13, 14, 21, 22] even if the results do not apply in this setting.

Regarding modeling, an ASCM with confounding cannot be reduced to a Markovian ASCM. Although $\mathbf{U}$ and $\mathbf{V}$ are both latent, every $\mathbf{U}$ is not the direct parents of $\mathbf{X}$, which means $\mathbf{U}$ cannot be uniquely determined by value of $\mathbf{X}$. Take the example where $V_1 \dashleftarrow\dashrightarrow V_2$ is the LCG G. Since U_{12} does not point to $\mathbf{X}$, we cannot let U_{12} be another latent generative factor $\mathbf{V}$.

Regarding results, we point the reader to Lemma 6, where it is shown that even with one hard interventions per node, it is not possible to disentangle variables within the same c-component. This in contrast with results in the Markovian setting, where it is shown in [21] that one hard intervention per latent variable allows us to achieve full identifiability of every latent variable up to scaling indeterminancies.

More broadly, it is noteworthy that transitioning causal reasoning from Markovian to non-Markovian settings was not trivial. For example, it is known that interventional distributions, such as $P(y \mid do(x))$, are always identifiable from the causal graph and observational distribution in Markovian settings in all models. Moving to non-Markovian settings, the celebrated do-calculus is developed primarily to address the decision problem of whether an interventional distribution can be uniquely computed from a combination of causal assumptions (in the form of a causal diagram) and the observational distribution [62]. Naturally, the issue of non-identifiability is much more acute in this setting, due to the existence of unobserved confounding.