TECHNICAL REPORT
R-110
May, 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].

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 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].

[*] 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
This work	✓	✓	✓	✓	General	Causal Disentanglement Map

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

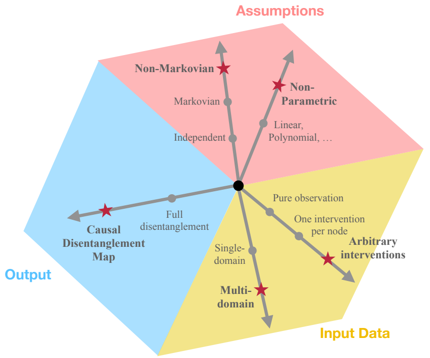


Figure 1: Dimensions of causal disentanglement representation learning tasks.

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, texts). Acknowledging this connection leads to various possibilities regarding learning, or disentangling the causal variables from data, similar to the initial ICA-like literature. 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 [23–29].

Considering this background, we study three axes within the different types of input and expected outputs of the causal disentanglement representation learning task, as summarized in Table 1. What we refer to as the input can be partitioned into qualitative and quantitative components. In terms of the qualitative aspect of the input, 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 also account for systems with richer causal topologies than ICA (independent features) while generalizing the Markovian setting. In particular, we do not rule out *a priori* the existence of unobserved confounding among features, which is a challenge pervasive throughout causal inference in the empirical sciences. Regarding the quantitative part of the output, we consider **data** gathered from arbitrary combinations of interventions and domains. The recent literature on this distinction acknowledges key differences [30–35], while the 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. We discuss the nuances of interventions vs domains in Section Domains vs Interventions. In terms of the expected **output**, current methods often aim for a full disentanglement, while we consider more relaxed types of disentanglement.

For concreteness, consider a hypothetical latent graph depicted in Fig. 2 in the context of epilepsy research [36–44]. In terms of **assumptions**, hospitals in different countries Π^i and Π^j will differ in the amount of sleep (V_1) patients get (represented by the S-node $S^{i,j} \rightarrow V_1$). 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 [45]; we are not as interested in disentangling the drug treatment and outcomes because it is known the

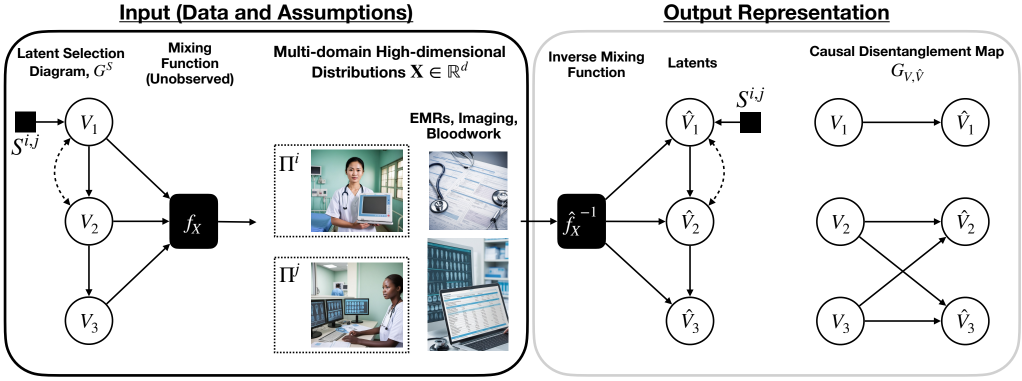

Input (Data and Assumptions)
Output Representation
Latent Selection Diagram, G^S
Mixing Function (Unobserved)
Multi-domain High-dimensional Distributions $\mathbf{X} \in \mathbb{R}^d$
EMRs, Imaging, Bloodwork
$S^{i,j}$
V_1
V_2
V_3
f_X
Π^i
Π^j
Inverse Mixing Function
Latents
Causal Disentanglement Map $G_{V,\hat{V}}$
$\hat{f}_X^{-1}$
$\hat{V}_1$
$\hat{V}_2$
$\hat{V}_3$

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

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 [46–48]. 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 variable causal structure, the goal is to learn the inverse of the mixing 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 [1].

In this paper, we develop 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)[2].
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.

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* [55]. 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. 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 $\mathbf{X}$.

[1] 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, 49–54].

[2] All proofs are provided in Appendix C.

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 $\mathbf{U}'_X$ [56, 57]. 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, 49] 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 [3]. Namely, the diagram induced by $\mathcal{M}_{\sigma_{\mathbf{X}}}$ is the same with G.

2 Modeling Disentanglement 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) $\mathbf{V} = \mathbf{V}_0 = \{V_1, \dots, V_d\}$ 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 [4] 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 $\mathbf{V}$ is 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 $\mathbf{X}$ [5]. We do not restrict the function of the underlying mechanisms $\mathcal{F}_0$ in SCMs, or the mixing function $f_{\mathbf{X}}$, and focus on the non-parametric setting. This assumption is commonly in the non-linear ICA and representation learning literature [9, 10, 17, 58].

The initial disentangled representation learning setting can be traced back at least to linear/nonlinear ICA [7–9], where G is assumed to have no edges ($\mathbf{V}$ are independent of each other) and Markovian (no bidirected edges in the LCG). More recently, allowing latent variables to have edges in the LCG was studied, albeit still under the Markovian assumption [11, 13, 14, 21, 22, 49, 53]. We relax this assumption and allow unoberserved confounding to exist between $\mathbf{V}$, which we call non-Markovianity[6].

Domains. We address the general setting of distributions that arise from multiple domains. Following [30–33, 61, 62], we define the so-called *latent selection diagram* that represents a collection of ASCMs to formally model the multi-domain setting. Selection diagrams enable us to compactly represent causal structure and cross-domain invariances [7].

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 $\mathbf{\Pi} = \langle \Pi_1, \Pi_2, ..., \Pi_N\rangle$, 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 $\mathcal{M}_j$. □

[3] 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, 56, 57] for a general discussion on soft interventions.

[4] A diffeomorphism is a bijective function $f_{\mathbf{X}}$ such that both $f_{\mathbf{X}}$ and $f_{\mathbf{X}}^{-1}$ are continuously differentiable [25]

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

[6] 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 [59], Pearl [1, 60], and more recently data fusion [34].

[7] See [30, 31] 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	$P^1_{\{\}}(\mathbf{X})$	$P^1_{v_i}(\mathbf{X})$	$P^1_{v_j}(\mathbf{X})$	$P^1_{v_i,v_j}(\mathbf{X})$	$\dots$
Π^2	$P^2_{\{\}}(\mathbf{X})$	$P^2_{v_i}(\mathbf{X})$	$P^2_{v_k}(\mathbf{X})$	$P^2_{v_i,v_k}(\mathbf{X})$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Π^N	$P^N_{\{\}}(\mathbf{X})$	$P^N_{v_l}(\mathbf{X})$	$P^N_{v_m}(\mathbf{X})$	$P^N_{v_l,v_j}(\mathbf{X})$	$\dots$

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 studied in the existing literature under the lens of the multi-domain intervention exchangeability assumption [30]. 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).

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 $\mathbf{\Sigma} = \{\sigma^{(k)}\}_{k=1}^K$ are applied across domains $\mathbf{\Pi}$, where k is an index from 1 to K. The corresponding domains that $\mathbf{\Sigma}$ are intervened in is denoted as $\mathbf{\Pi}^{\mathbf{\Sigma}} = \{\Pi^{(k)}\}_{k=1}^K$ (the domains associated with each $\sigma^{(k)} \in \mathbf{\Sigma}$). 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 $\mathbf{\Psi} = \{\mathbf{I}^{(k)}\}_{k=1}^K$. 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 mechanism on 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 $\sigma^{(k)}$. $\mathbf{\Psi}_{\mathbf{T}}$ is a subset of $\mathbf{\Psi}$ such that $\mathbf{T} \subseteq do[\mathbf{I}^{(j)}]$ for every $\mathbf{I}^{(j)} \in \mathbf{\Psi}_{\mathbf{T}}$, 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. Let an intervention target collection be

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

In words, $\mathbf{\Psi}$ indicates 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 $\sigma^{(3)}$.
$do[\mathbf{I}^{(3)}] = \{V_2\}$: $\sigma^{(3)}$ 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 [63–67]. 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 have implications for causal inference and the characterization of corresponding causal structures as discussed in depth by [30]. Moreover, it is crucial to avoid conflating these qualitatively distinct concepts of interventions and domains, as highlighted in transportability analysis [61]. Pearl and Bareinboim have introduced clear semantics for (S) nodes (environments), presenting a unified representation in the form of selection diagrams [31, 33, 34].

By recognizing these differences, this work leverages any combination of observational and/or interventional data arising from multiple domains to present a general approach to disentanglement learning compared to the literature (see Table 2). Specifically, 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 under "Observational"). However, we examine the a general setting, where an arbitrary collection of interventions, or observations from any combination of domains is available (green section).

Observed Distributions. The interventions $\mathbf{\Sigma} = \{\sigma^{(k)}\}_{k=1}^{K}$, induce distributions $\mathcal{P} = \{P^{(k)}\}_{k=1}^{K}$ in multi-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 $P^{(k)}$ (formally defined in Assumption. 7), unless explicitly stated otherwise [8].

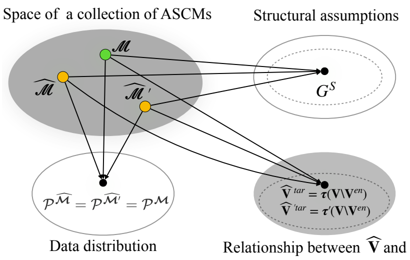

Space of a collection of ASCMs
Structural assumptions
$\boldsymbol{\mathcal{M}}$
$\widehat{\boldsymbol{\mathcal{M}}}$
$\widehat{\boldsymbol{\mathcal{M}}}'$
G^S
$\mathcal{P}^{\widehat{\mathcal{M}}} = \mathcal{P}^{\widehat{\mathcal{M}}'} = \mathcal{P}^{\mathcal{M}}$
$\widehat{\mathbf{V}}^{tar} = \boldsymbol{\tau}(\mathbf{V}\backslash\mathbf{V}^{en})$
$\widehat{\mathbf{V}}'^{tar} = \boldsymbol{\tau}'(\mathbf{V}\backslash\mathbf{V}^{en})$
Data distribution
Relationship between $\widehat{\mathbf{V}}$ and $\mathbf{V}$

Figure 3: General ID/disentangleability (Def. 2.3).

Suppose the underlying true model $\boldsymbol{\mathcal{M}}$ induces the LSD G^S and a collection of distributions $\mathcal{P}$ over $\mathbf{X}$ is given according to a corresponding collection of interventions Σ. The goal of this paper is to learn a disentangled representation $\widehat{\mathbf{V}}$ of the latent generative factors $\mathbf{V}$ in $\boldsymbol{\mathcal{M}}$. In other words, our goal is to estimate the inverse of the true mixing function $f_{\mathbf{X}}$ and determine the latent variables $\mathbf{V}$ up to indeterminacies. In the literature, every variable $V_i \in \mathbf{V}$ is required to be disentangled from all other variables [7, 21] or some special subset (e.g. non-ancestors of V_i) [21, 22]. However, as illustrated in Fig. 2, sometimes only the target variables ($\mathbf{V}^{tar} \subseteq \mathbf{V}$) is needed to be disentangled from some user-chosen entangled variables ($\mathbf{V}^{en}$). We formally define this type of general indeterminacy next as well as the formal version of our ID task.

Problem Statement

Definition 2.3 (General Identifiability/Disentangleability (ID))**.** Let 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. 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, $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \ldots, \widehat{\mathcal{M}}_n \rangle$, that is compatible with G^S and $\mathcal{P}^{\widehat{\boldsymbol{\mathcal{M}}}} = \mathcal{P}$. For short, $\mathbf{V}^{tar}$ is said to be ID w.r.t. $\mathbf{V}^{en}$. □

To illustrate, consider a target variable $\mathbf{V}^{tar}$ such that one aims to obtain a representation that is disentangled from another subset 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 distribution $\mathcal{P}$[9]. Def 2.3 is illustrated in Fig. 3. Following the example illustrated in 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 $\widehat{\boldsymbol{\mathcal{M}}}$ that matches the distributions and LSD, V_3 is ID w.r.t. V_1. Def. 2.3 is more relaxed than the full disentanglement (which means any V_i is disentangled from other variables) since any target $\mathbf{V}^{tar}$ and $\mathbf{V}^{en}$ can be chosen. It can be reduced to existing identifiability definitions (as discussed in Appendix E.2).

[8] A formal version of "sufficiently different" (Assumption 7), as well as other technical assumptions are stated and discussed in Appendix A.2.

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

Example 2 (Formal ID task)**.** Suppose the pair of underlying ASCMs $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ induces the LSG G^S in Fig. 2 and distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}; \sigma_{V_3}), P^{\Pi_2}(\mathbf{X}; \sigma_{V_4})\}$ from interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)}\} = \{\{\}, \{\}, \sigma_{V_3}, do(V_2)\}$. Given intervention targets $\mathbf{\Psi} = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, V_3^{\Pi_2}, V_2^{\Pi_1, \text{do}}\}$ 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 $\{V_2, V_3\}$. The answer will be provided in the next two sections. □

3 Graphical Criterion for Causal Disentanglement

In this section, we study identifiability criteria given general assumptions and 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 connection (Eq. 11), we introduce in Sec. 3.2 three graphical criteria (Prop. 3, 4 and 5) to check the identifiability of a target representation.

3.1 Latent variable factorization and invariances

First, we revisit the factorization of distributions induced by non-Markovian models [3, Def. 15]. Specifically, 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, $P_{\mathbf{T}}(\mathbf{V})$ can be factorized as follows:

$$P_{\mathbf{T}}(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} P_{\mathbf{T}}(V_i | \mathbf{Pa}_i^{\mathbf{T}+}), \tag{2}$$

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 $G_{\overline{\mathbf{T}}}$.

For example, the factorization of $P(\mathbf{V})$ according to the causal graph shown in Fig. 2 is

$$P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2) \tag{3}$$

with the order $V_1 < V_2 < V_3$ and $\mathbf{T} = \{\}$. 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 $G_{\overline{\mathbf{T}}}$. The following example illustrates such differences.

Example 3. Consider a collection of ASCMs, $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \ldots, \mathcal{M}_n \rangle$ induces the LSD shown in Fig. 4(c). Given order A: $V_1 < V_2 < V_3 < V_4$, $P(\mathbf{V})$ can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2, V_1)P(V_4 \mid V_3) \tag{4}$$

Notice that the conditioning part of V_3 includes $\{V_2, V_1\}$, which are not parents of V_3. Choosing order B: $V_1 < V_3 < V_2 < V_4$, $P(\mathbf{V})$ can be factorized as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1, V_3)P(V_3)P(V_4 \mid V_3) \tag{5}$$

The conditioning part of V_2 and V_3 are different from Eq. (4) and (5). 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 then the observation distribution would factorize as:

$$P(\mathbf{V}) = P(V_1)P(V_2 \mid V_1)P(V_3 \mid V_2)P(V_4 \mid V_3) \tag{6}$$

which is clearly different from both Eqs. (4) and (5). □

Armed with this new factorization, the representation $\widehat{V}$ in $\widehat{\boldsymbol{\mathcal{M}}}$ and the true underlying variables $\mathbf{V}$ in $\boldsymbol{\mathcal{M}}$ can be related by comparing 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 collection distribution $\mathcal{P}$ 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 mechanism on $\mathbf{T}$. 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 LSG G^S, then*

$$\sum_i^d \log p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = \sum_i^d \log p_{\mathbf{T}}^{(j)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}), \tag{7}$$

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

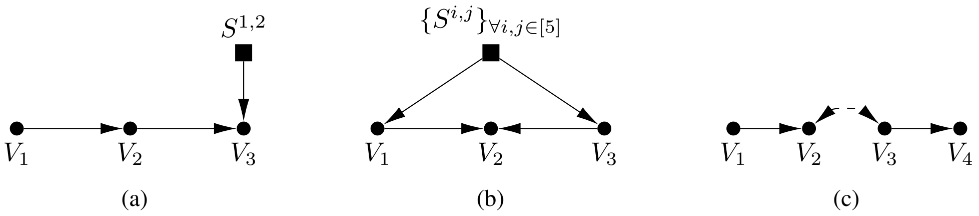


Figure 4: LSDs in Ex. and Exps. (a) chain, (b) collider and (c) non-markovian graphs.

To illustrate, Prop.1 shows that if $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ agree the given distributions over observed $\mathbf{X}$ and the LSG, then the connection between unobserved $\mathbf{V}$ and $\widehat{\mathbf{V}}$ can be built upon Eq. (7). More specifically, the left (right) side of Eq. (7) 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 $\Pi^{(j)}$. In addition, both $P(\mathbf{V})$ and $P(\widehat{\mathbf{V}})$ can be factorized through Eq. (2) since $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ are compatible with G.

Example 4. (Example 2 continued.) Consider $\widehat{\boldsymbol{\mathcal{M}}}$ that agrees on $\mathcal{P}$ and G^S introduced in Example 2. According to Prop. 1, comparing $P^{(2)}$ and $P^{(1)}$ under $\mathbf{T} = \{\}$, we can write

$$\begin{aligned}&p^{(2)}(v_1) - p^{(1)}(v_1) + p^{(2)}(v_2 \mid v_1) - p^{(1)}(v_2 \mid v_1) + p^{(2)}(v_3 \mid v_2) - p^{(1)}(v_3 \mid v_2)\\ &= p^{(2)}(\widehat{v}_1) - p^{(1)}(\widehat{v}_1) + p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1) - p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1) + p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) - p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)\end{aligned} \tag{8}$$

To illustrate, first, $p^{(2)}(\mathbf{v}) - p^{(1)}(\mathbf{v})$ are equal to $p^{(2)}(\widehat{\mathbf{v}}) - p^{(1)}(\widehat{\mathbf{v}})$ since $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ agree on $\mathcal{P}$. Second, $p(\mathbf{v})$ and $p(\widehat{\mathbf{v}})$ can be both factorized with Eq. (3) since $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ agree on G^S. Finally, this connection is built with the change of factors $p(v_1), p(v_2 \mid v_1)$, and $p(v_3 \mid v_2)$. □

Example 5. (Example 3 continued.) Consider the pair $\boldsymbol{\mathcal{M}}$ and $\widehat{\boldsymbol{\mathcal{M}}}$ that induces the diagram in Fig. 4(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. (4),

$$\begin{aligned}&p^{(2)-(1)}(v_1) + p^{(2)-(1)}(v_2 \mid v_1) + p^{(2)-(1)}(v_3 \mid v_1, v_2) + p^{(2)-(1)}(v_4 \mid v_3)\\ &= p^{(2)-(1)}(\widehat{v}_1) + p^{(2)-(1)}(\widehat{v}_2 \mid \widehat{v}_1) + p^{(2)-(1)}(\widehat{v}_3 \mid \widehat{v}_1, \widehat{v}_2) + p^{(2)-(1)}(\widehat{v}_4 \mid \widehat{v}_3)\end{aligned} \tag{9}$$

where $p^{(2)-(1)}(\cdot)$ denotes $p^{(2)}(\cdot) - p^{(1)}(\cdot)$. The connection between $\mathbf{V}$ and $\widehat{\mathbf{V}}$ is built with factor $p(v_1), p(v_2 \mid v_1), p(v_3 \mid v_1, v_2)$ and $p(v_4 \mid v_3)$. With another order and factorization of Eq. (5), we have

$$\begin{aligned}&p^{(2)-(1)}(v_1) + p^{(2)-(1)}(v_2 \mid v_1, v_3) + p^{(2)-(1)}(v_3) + p^{(2)-(1)}(v_4 \mid v_3)\\ &= p^{(2)-(1)}(\widehat{v}_1) + p^{(2)-(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) + p^{(2)-(1)}(\widehat{v}_3) + p^{(2)-(1)}(\widehat{v}_4 \mid \widehat{v}_3)\end{aligned} \tag{10}$$

The connection $\mathbf{V}$ and $\widehat{\mathbf{V}}$ is built with factor $p(v_1), p(v_2 \mid v_1, v_3), p(v_3)$, and $p(v_4 \mid v_3)$. □

However, not all factors necessarily contribute to Eq. (7). 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. Other invariant factors will be canceled out in Eq. (7). The following result generalizes finding invariant factors 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 not needed) with target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ as follows. Add variable V_l to $\Delta\mathbf{V}$ if,*

1. *$V_l \in \Delta\mathbf{V}$ 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. *$V_l \in \Delta\mathbf{V}$ 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 $V_i \in \mathbf{V}\backslash\mathbf{C}_{\geq}(\Delta\mathbf{V})$, then $p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+})$, which will be denoted as invariant factors, where $\mathbf{C}_{\geq}(\Delta\mathbf{V})$ are variables in the same C-Component with $\Delta\mathbf{V}$ and not before $\Delta\mathbf{V}$ in the topological order for factorization. □

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 the $\mathbf{C}_{\geq}(\Delta\mathbf{V})$, which are variables in the same C-Component with changed variable set $\Delta\mathbf{V}$ and not before $\Delta\mathbf{V}$ in the order. $\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$ contains the variables that are intervened differently in $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ as well as the variables pointed by S-node, $S^{j,k}$ [10].

Example 6. (Ex. 2 and 4 continued.) 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 $V_3 \in \mathbf{I}^{(3)}$ while $V_3 \notin \mathbf{I}^{(1)}$. 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 $p(v_3 \mid v_2)$ changes. □

Example 7. (Ex. 3 and 5 continued.) Consider the diagram in Fig. 4(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 $V_2 \in \mathbf{I}^{(2)}$, $V_2 \notin \mathbf{I}^{(1)}$, and $\mathbf{C}(V_2) = \{V_2, V_3\}$. Thus, comparing $P^{(2)}$ with $P^{(1)}$ following topological order A ($V_1 < V_2 < V_3 < V_4$) in Ex. 3, factors $p(v_1), p(v_4 \mid v_2, v_1)$ are invariant, whereas $p(v_2 \mid v_1), p(v_3 \mid v_2, v_1)$ change. Following another topological order B ($V_1 < V_3 < V_1 < V_4$) in Ex. 3, the invariant factors are $p(v_1), p(v_3), p(v_4 \mid v_2, v_1)$ while $p(v_2 \mid v_1, v_3)$ changes. □

With Prop. 2, Eq. (7) naturally keeps factors only in the $\mathbf{C}_{\geq}(\Delta\mathbf{V})$, i.e.,

$$\sum_{V_i \in \tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = \sum_{V_i \in \tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(j)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) \tag{11}$$

where $\tilde{\mathbf{V}} = \mathbf{C}_{\geq}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S])$. For example, according to Ex. 6, Eq. (8) can be simplified as:

$$p^{(2)}(v_3 \mid v_2) - p^{(1)}(v_3 \mid v_2) = p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - p^{(1)}(\widehat{v}_3 \mid v_2) \tag{12}$$

Similarly, according to Ex. 5, Eq. (9) and (10) are simplified as:

$$p^{(2)-(1)}(v_2 \mid v_1) + p^{(2)-(1)}(v_3 \mid v_1, v_2) = p^{(2)-(1)}(\widehat{v}_2 \mid v_1) + p^{(2)-(1)}(\widehat{v}_3 \mid \widehat{v}_1, \widehat{v}_2) \tag{13}$$

$$p^{(2)-(1)}(v_2 \mid v_1, v_3) = p^{(2)-(1)}(\widehat{v}_3 \mid \widehat{v}_1, \widehat{v}_2) \tag{14}$$

These equality constraints follow from the non-Markovian factorization hint that the learned representation $\widehat{\mathbf{V}}$ (r.h.s of Eq. (11)) is a function of variables that appear on the l.h.s. The next definition formalizes the possible change variables between r.h.s and l.h.s of Eq. 11.

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 index not needed) of the target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ is the remaining variables after comparison (i.e. Eq. 11),

$$\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S] = \tilde{\mathbf{V}} \cup \mathbf{Pa}^{\mathbf{T}+}(\tilde{\mathbf{V}}), \tag{15}$$

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

To illustrate, the $\Delta\mathbf{Q}$ set encompasses all variables in l.h.s of Eq. (11), including $\tilde{\mathbf{V}}$ and its extended parents. These variables come from factors that possibly change and are kept in Eq. (11). The set $\mathbf{V} \backslash \Delta\mathbf{Q}$ is called *canceled variables* since the invariant factors are canceled from the comparison. Continuing Example 6, $\Delta\mathbf{Q} = \{V_3, V_2\}$ since the changed factors is $p(v_3 \mid v_2)$ and invariant factors are canceled out. Continuing Ex. 7, $\Delta\mathbf{Q} = \{V_1, V_2, V_3\}$ given either topological order [11].

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

Example 8 (Observational data in two homogenous domains)**.** Consider the LSD over two domains Π_1, Π_2 shown in Fig. 4(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 domains are completely 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. 3), 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. □

[10] Notice that the same intervention 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.

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

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

Example 9 (Observational data in two completely heterogenous domains)**.** From Fig. 4(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 $P^{(1)}$ (order A in Ex. 3), all factors $p(v_1), p(v_2|v_1), p(v_3|v_2, v_1), p(v_4|v_3)$ change, which means nothing is invariant across domains. □

Our approach to disentangled representation learning leverages comparisons of distributions. The two example above illustrate that nothing is disentanglable in a fully non-parametric setting when all latent variables are invariant factors (no discrepancies exist), or when all latent variables change between two distributions (no commonalities exist). There is a distinct trade-off in leveraging distributions with both invariant factors and changed variable sets. We will explore this interplay between invariant and changed factors across distributions to achieve identifiability and disentangled representation in the next section.

3.2 Graphical Criteria for Disentanglement

First, to motivate the disentanglement results, we derive a novel disentanglement result in a simple setting with three latent variables.

Example 10 (Deriving disentanglement with three distributions)**.** Consider the identification task in Ex. 2, where G^S is shown in Fig. 2 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. We demonstrate that these distributions allow the disentanglement of $\{V_2, V_3\}$ with respect to V_1, which has not been acknowledged in the literature.

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. (12). Next, we can 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. The r.h.s contains dependencies, 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:

$$\begin{aligned}
0 &= \frac{\partial \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_2}\frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_3}\frac{\partial \widehat{v}_3}{\partial v_1} \\
0 &= \frac{\partial \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_2}\frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_3}\frac{\partial \widehat{v}_3}{\partial v_1}
\end{aligned} \tag{16}$$

We note that Eq. (16) is a linear of equations with unknown $\frac{\partial \widehat{v}_3}{v_1}$ and $\frac{\partial \widehat{v}_2}{v_1}$. According to Assumption 7, the coefficients matrix A is full rank, where

$$A = \begin{pmatrix} \frac{\partial \log p^{(2)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_2} & \frac{\partial \log p^{(2)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_3} \\ \frac{\partial \log p^{(3)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_2} & \frac{\partial \log p^{(3)}(\widehat{v}_3|\widehat{v}_2) - \log p^{(1)}(\widehat{v}_3|\widehat{v}_2)}{\partial \widehat{v}_3} \end{pmatrix} \tag{17}$$

Then we have

$$\frac{\partial \widehat{v}_2}{\partial v_1} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0 \tag{18}$$

Then $\widehat{V_2} = \tau_2(V_2, V_3)$ and $\widehat{V_3} = \tau_3(V_2, V_3)$ meaning $\{V_2, V_3\}$ is ID w.r.t V_1. □

This example demonstrates some of the principles behind disentanglement, which we will now formalize.

Leveraging the comparisons among distributions in $\mathcal{P}$ (Eq. 11), we next develop three criterion 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. (11)) is irrelevant to canceled variables (l.h.s of Eq. (11)).

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 $\mathbf{T}$) $\forall\ l \in [L]$, $\mathbf{T} = do[\mathbf{I}^{(a_0)}] \subseteq do[\mathbf{I}^{(a_l)}]$ [12].

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

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

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

Prop. 3 disentangles target variables $\mathbf{V}^{tar}$ constructed by $\Delta\mathbf{Q}$ sets from canceled variables according to Eq. (11). To illustrate, it considers to find a collection of L distribution $\{P^{(a_1)}, \dots, P^{(a_L)}\}$ to compare with the baseline $P^{(a_0)}$ such that (1) the hard intervention variables set of $\{\mathbf{I}^{(a_1)}, \dots, \mathbf{I}^{(a_L)}\}$ 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 wrt $\mathbf{V} \backslash \mathbf{V}^{tar}$.

Example 11. (Example 6 continued.) Consider $\mathbf{V}^{tar} = \{V_2, V_3\}$, $\mathbf{V}^{en} = \mathbf{V} \backslash \{V_2, V_3\} = \{V_1\}$. When comparing $\{P^{(2)}, P^{(3)}\}$ with the baseline $P^{(1)}$, $\mathbf{T} = do[\mathbf{I}^{(1)}] = \{\}$, and then

$$\Delta\mathbf{Q}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, \mathbf{T}, G^S] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(1)}, \mathbf{T}, G^S] = \{V_2, V_3\} \tag{19}$$

$\{P^{(2)}, P^{(3)}\}$ satisfies condition [1] in Prop. 3, since the hard intervention variable set is $\{\}$. They also satisfies condition [2], since $\Delta\mathbf{Q}^{(2),(1)} = \Delta\mathbf{Q}^{(3),(1)} = \mathbf{V}^{tar}$. Condition [3] are satisfied since $V_2 \in \Delta\mathbf{Q}^{(2),(1)}$ and $V_3 \in \Delta\mathbf{Q}^{(3),(1)}$ Thus, $\mathbf{V}^{tar}$ is ID w.r.t $\mathbf{V}^{en}$ by Prop. 3. This demonstrates that a variable V_2 can be disentangled from another variable that is in the C-component (V_1). See Appendix Ex. 22 for a detailed derivation. □

Example 12. (Ex. 7 continued.) Suppose $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ with intervention targets

$$\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{V_2\}^{\Pi_1}, \mathbf{I}^{(3)} = \{V_3\}^{\Pi_1}, \mathbf{I}^{(4)} = \{V_1\}^{\Pi_1} \tag{20}$$

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:

$$\Delta\mathbf{Q}^{(2),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(3),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(4),(1)} = \{V_1\}. \tag{21}$$

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. See App. Ex. 23 for a detailed derivation. □

The second result disentangles variables within $\Delta\mathbf{Q}$ sets leveraging conditional independence among variables within the $\Delta\mathbf{Q}$ set.

Proposition 4 (ID of variables within ΔQ sets). *Consider the 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 $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, 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) there exists $\{a'_1, \dots, a'_{2d'+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \dots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}, V_i^{tar} \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], V_i^{tar} \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$ and for all $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2d'+j}),(a_0)}$, where $d' = |\mathbf{V}^{tar}|$ and

$$\begin{aligned}\boldsymbol{\mathcal{E}} =&\{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid i)\ \exists a_l, \{V_k, V_r\} \in \Delta\mathbf{Q}^{(a_l),(a_0)}; \\ &V_k \textit{ is connected to } V_r \textit{ conditioning } \mathbf{V}^{tar} \backslash \{V_k, V_r\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\}\end{aligned} \tag{22}$$

, then V_i is ID w.r.t V_j. □

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

Prop. 4 disentangles target variables V_i and V_j both in $\Delta\mathbf{Q}$ sets. To illustrate, it considers a set of distributions that satisfies conditions [1] and [2] as Prop. 3 and a new condition [4]. Condition (4) states that each V_i^{tar} can be covered twice by first $2d'$ different $\Delta\mathbf{Q}$s, and $\{V_k, V_r\} \in \boldsymbol{\mathcal{E}}$ can be covered by other different $\Delta\mathbf{Q}$s.

Example 13. Suppose LSD given G^S is the graph shown in (Fig. 2). Suppose the 9 given intervention targets are

$$\boldsymbol{\Psi} = \{\{\}^{\Pi_1}, \{V_1^{\Pi_1}\} \times 4, \{V_2^{\Pi_1}, V_3^{\Pi_1}\} \times 4\} \tag{23}$$

, which 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. Consider $\mathbf{T} = \{\}$. Let $\mathbf{V}^{tar} = \{V_1, V_2, V_3\}$. Then $V_1 \perp\!\!\!\perp V_3 \mid \{V_2\}$ in G. Choose $\mathbf{I}^{(1)} = \{\}^{\Pi_1}$ as the baseline. Based on Def. 3.1,

$$\begin{aligned} \Delta\mathbf{Q}^{(2),(1)} &= \cdots = \Delta\mathbf{Q}^{(5),(1)} = \{V_1, V_2\} \\ \Delta\mathbf{Q}^{(6),(1)} &= \cdots = \Delta\mathbf{Q}^{(9),(1)} = \{V_1, V_2, V_3\}, \end{aligned} \tag{24}$$

and we have

$$\boldsymbol{\mathcal{E}} = \{\{V_1, V_2\}, \{V_2, V_3\}\} \tag{25}$$

Now we check if condition [4] is satisfied. From

$$\begin{aligned} & V_1 \in \Delta\mathbf{Q}^{(2),(1)}, V_1 \in \Delta\mathbf{Q}^{(3),(1)}, \\ & V_2 \in \Delta\mathbf{Q}^{(4),(1)}, V_2 \in \Delta\mathbf{Q}^{(6),(1)}, \\ & V_3 \in \Delta\mathbf{Q}^{(7),(1)}, V_3 \in \Delta\mathbf{Q}^{(8),(1)} \\ & \{V_1, V_2\} \subseteq \Delta\mathbf{Q}^{(5),(1)} \\ & \{V_2, V_3\} \subseteq \Delta\mathbf{Q}^{(9),(1)} \end{aligned} \tag{26}$$

, we know the condition (4) is satisfied. Then 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

$$\boldsymbol{\Psi} = \{\{\}^{\Pi_1}, \{V_1^{\Pi_1}\} \times 7, \{V_2^{\Pi_1}, V_3^{\Pi_1}\} \times 2\}, \tag{27}$$

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

We establish the following corollary that can be more commonly used according to Prop. 3.

Corollary 1 (ID of variables within $\Delta\mathbf{Q}$ sets). *Consider the variables $\mathbf{V}^{tar} \subseteq \mathbf{V}$, $\mathcal{P}_{\mathbf{T}}$ 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 $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$.* □

To illustrate, if one can find a set of distributions that satisfies conditions (1) as Prop. 3 and let all $\Delta\mathbf{Q}$ are equivalent to $\mathbf{V}^{tar}$ and $V_i, V_j \in \mathbf{V}^{tar}$ are conditionally independent given all other variables in $\mathbf{V}^{tar}$, then V_i can be disentangled from V_j. Here, the condition (2) and (4) in Prop. 4 reduce to this equivalence condition.

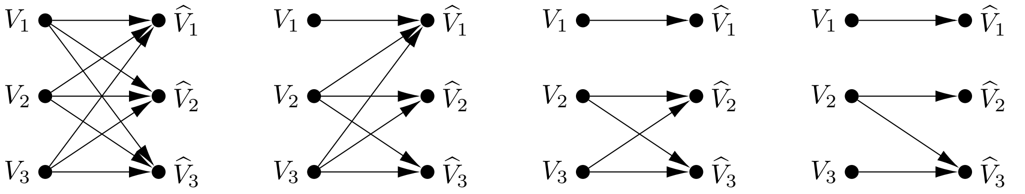
Example 14. Suppose LSD given G^S is a collider graph (Fig. 4(b)). Suppose the given intervention targets are $\boldsymbol{\Psi} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, \{\}^{\Pi_3}, \{\}^{\Pi_4}, \{\}^{\Pi_5}\}$, which means that observational distributions are available in each domain. Consider $\mathbf{T} = \{\}$. 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 App. Ex. 25 for a derivation. □

With these existing disentanglements from Props. 3 and 4, the following Proposition considers an inverse direction, which identifies canceled variables w.r.t. $\Delta\mathbf{Q}$ sets [13].

[13] Recall that ID is one-way. ID of V_i wrt V_j does not imply V_j is ID wrt 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:     $\mathbf{HARD} = \{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_s \mid do(T_i) \in \mathbf{\Psi}$    ▷ Get hard intervention variables sets.
4:     for all $\mathbf{T} \in \mathbf{DO}$ do
5:         $\mathbf{\Psi}_{\mathbf{T}} \leftarrow \mathbf{\Psi}$    ▷ Collect intervention targets that contain hard intervention variables $\mathbf{T}$
6:         for all $\mathbf{I} \in \mathbf{\Psi}_{\mathbf{T}}$ such that $do[\mathbf{I}] = \mathbf{T}$ do    ▷ Iterate intervention targets as the baseline
7:             $\mathcal{Q} = \{\mathbf{Q}_1, \ldots, \mathbf{Q}_{|\mathbf{\Psi}_{\mathbf{T}}\backslash\mathbf{I}|}\}$, where $Q_k \leftarrow \Delta\mathbf{Q}[\mathbf{J}^{(k)}, \mathbf{I}, \mathbf{T}, G^S]$    ▷ Construct $\Delta\mathbf{Q}$ sets.z
8:             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:                 $G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ Dis$\Delta$QFromCancel$(\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}}, \Psi_{\mathbf{T}}, \mathbf{I}, \mathcal{Q})$ ▷ Alg. F.3 and Prop. 3
10:                $G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ DisWithin$\Delta$Q$(\mathbf{Q}, G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}}, \Psi_{\mathbf{T}}, \mathbf{I}, \mathcal{Q})$    ▷ Alg. F.6 and Prop. 4
11:    for all $\mathbf{T} \in \mathbf{HARD}$ do
12:        $G_{\mathbf{V},\widehat{\mathbf{V}}} \leftarrow$ DisCancelFrom$\Delta$Q$(G_{\mathbf{V},\widehat{\mathbf{V}}}, G_{\overline{\mathbf{T}}})$    ▷ Alg. F.8 and Prop. 5
13: return $G_{\mathbf{V},\widehat{\mathbf{V}}}$
```



(a) Initialization.

(b) 1st epoch, Step 11.

(c) 1st epoch, Step 14.

(d) 2nd epoch, Step 11.

Figure 5: Process of removing edges from CDM $G_{\mathbf{V},\widehat{\mathbf{V}}}$ using Alg. 1 in Ex. 16. (d) is the final output.

Proposition 5 (ID of canceled variables w.r.t. $\Delta\mathbf{Q}$ sets). *Suppose $\mathbf{\Psi}$ contains $do(\mathbf{T})$. 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: if $\mathbf{V}\backslash\{V^{tar}\}$ is already disentangled from V^{tar}, then V^{tar} can ID wrt $\mathbf{V}\backslash\{V^{tar}\}$ 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. See the following example.

Example 15. (Example 11 (continued).) Consider $\mathbf{V}^{tar} = \{V_1\}$. From Ex. 6, 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 with existing work in Appendix Section E.

4 Algorithmic Disentanglement of Causal Representations

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 $\mathbf{\Psi}$.

The whole algorithm **CausalRepresentationID** (**CRID**, for short) 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)). 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 proceeds by first constructing the 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 enumerated. 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). After the union of $\Delta\mathbf{Q}$ sets (denoted as $\mathbf{Q}$) is chosen (Step 8) iteratively, Props. 3 and 4 are leveraged in two procedures (Step 9 and 10) 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 (see details in Alg. F.3 and F.6). 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). The CRID algorithm iterates through the application of each of Props. 3, 4, and 5

Example 16. (Ex. 2 continued.) Consider the selection diagram (Fig. 2) and the intervention target sets in Ex. 2 The hard intervention variable sets are the empty set $\{\}$ and $\{V_3\}$. First, $\mathbf{T}$ is chosen as $\{\}$ and then $\mathbf{\Psi_T} = \mathbf{\Psi}$. 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. 11 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 iter, with the fact that $\{V_2, V_3\}$ is ID wrt 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 $\widehat{V}_1$. See Ex. 15 for details.

In the second iteration, the algorithm repeats the choice of $\mathbf{T}$ and the baseline. At this iteration, 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. At the end of this epoch no further can be removed by Alg. F.8. In the third epoch, $G_{\mathbf{V},\widehat{\mathbf{V}}}$ is not updated and the process of CDM returned is shown in Fig. 5. □

After obtaining $G_{\mathbf{V},\widehat{\mathbf{V}}}$ from **CRID**, the identifiability of target variables $\mathbf{V}^{tar}$ wrt $\mathbf{V}^{en}$ can be inferred through the absence of edges in $G_{\mathbf{V},\widehat{\mathbf{V}}}$. The following theorem indicates 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

We corroborate the theoretical findings through simulations. We consider LSDs shown in Fig. 4 with different collection of distributions $\mathcal{P} = \{P^{(k)}(\mathbf{X}; \sigma^{(k)})\}_{k=1}^{K}$ and the results are presented in Fig. 6. For the evaluation, we follow a standard evaluation protocol in prior work [18], where we take the latent representations $\widehat{\mathbf{V}}$ and compute their mean correlation coefficient (MCC) wrt the latent $\mathbf{V}$. We compare MCC with what is expected from **CRID**.

5.1 Synthetic data-generating process

We generate data according to latent causal diagrams shown in Fig. 4. 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.

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:

$$V_k := \epsilon'_k, \quad \text{with } \epsilon'_k \sim \mathcal{N}(\mu_k, \sigma_k), \ \forall V_k \in \mathbf{I} \tag{28}$$

such that μ_k is not within $+/-1$ of any other distribution for variable $V_k \in \mathbf{V}$. This ensures the Assumption of Generalized Distribution Change (Assump. 7). With a soft intervention $\mathbf{J}$ that is not hard:

$$V_k := \sum_{j \in Pa_k} \alpha_{i,j} V_k + \epsilon'_k, \quad \text{with } \epsilon'_k \sim \mathcal{N}(\mu_k, \sigma_k), \ \forall V_k \in \mathbf{J} \tag{29}$$

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.

We modify the mean and the variance to ensure that the Assumption of distribution change is met (Assump. 7).

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, 49], to generate an invertible mixing function, we will use a multilayer perceptron $\mathbf{f_X} = \sigma \circ \mathbf{A}_M \circ ... \circ \sigma \mathbf{A}_1$, 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 [68]:

$$\sigma(x) = tanh(x) + 0.1x \tag{30}$$

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 Assump. 4, and then $\mathcal{P}$ is drawn according to all ASCMs instantiated.

5.2 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 [69] 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:

$$\mathbf{X} = \mathbf{g}_\theta(\mathbf{V}) \tag{31}$$

Specifically, g_θ will comprise of Neural Spline Flows [70] 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:

$$\hat{p}^k_\theta(v_i | \mathbf{Pa_i}) = \mathcal{N}\left(\sum_{j \in Pa_i} \hat{\alpha}_{i,j} v_j, \hat{\sigma}_i \right) \tag{32}$$

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:

$$\hat{p}^k_\theta(v_i) = \mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i) \tag{33}$$

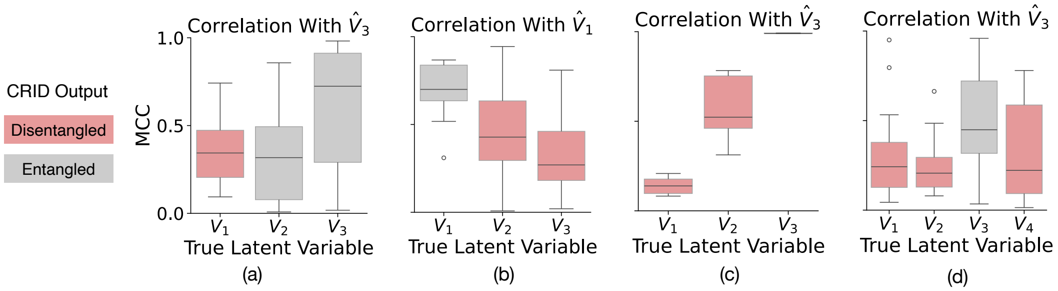


Figure 6: Correlation of learned latent representations with true latent variables from Fig. 4 are analyzed with: the chain graph (a) with $\Sigma = \{\sigma_{\{\}}, \sigma_3, \sigma_3\}$ chain graph, (b) an extra hard intervention $\Sigma = \{\sigma_{\{\}}, \sigma_3, \sigma_3, do(V_2)\}$, the collider graph (c) with $\Sigma = \{\sigma_{\{\}}, \sigma_{1,3}, \sigma_{1,3}, \sigma_{1,3}, \sigma_{1,3}\}$, and (d) the non-Markovian graph with $\Sigma = \{do(V_3), do(V_3)\}$.

5.3 Training details

We use the ADAM optimizer [71].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:

$$\theta^* = \arg\max_\theta \sum_{k=0}^{N} \Big(\frac{1}{n_k} \sum_{n=1}^{n_k} \log p_\theta^k(\mathbf{X}^{(k)})\Big) \tag{34}$$

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.

5.4 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.

5.5 Results

Chain Graph Fig. 4(a). Fig. 6(a) shows ID of V_3 wrt $\{V_1\}$ using input distributions $\mathcal{P}$ with interventions $\Sigma = \{\sigma_{\{\}}, \sigma_3^{\{1\}}, \sigma_3^{\{2\}}\}$ because $MCC(\widehat{V}_3, V_1)$ is relatively low compared to $MCC(\widehat{V}_3, V_3)$, which is consistent with CRID. The ID results of [21] states V_3 would still be entangled with V_1 because $V_1 \in \overline{Anc}(V_3)$. Fig. 6(b) shows ID of V_1 wrt $\{V_2, V_3\}$ by adding an extra hard intervention on V_2. Interestingly, we do not even have to intervene on V_1 to obtain full disentanglement.

Collider Graph Fig. 4(b). Fig. 6(c) shows V_1 and V_3 are ID wrt V_2 and each other because $MCC(\widehat{V}_3, V_3) > MCC(\widehat{V}_3, V_i)$ and $MCC(\widehat{V}_2, V_2) > MCC(\widehat{V}_2, V_i)$, which is consistent with CRID. There are distributions from four domains that have a change-in-mechanism on $\{V_1, V_3\}$ (represented by the S-node). 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 Fig. 4(c). Fig. 6(d) shows V_3 is ID wrt $\{V_1, V_2, V_4\}$ with interventions $\Sigma = \{do^{\{1\}}(V_3), do^{\{2\}}(V_3)\}$, which is consistent with CRID. No prior results achieve disentanglement with confounding among $\mathbf{V}$.

6 Conclusions

This work introduces theory and a practical ID algorithm for determining which latent variables are disentangleable from a given set of assumptions in the form of a LSD, and input distributions from heterogenous 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] 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.

[24] 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).

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

[38] 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).

[39] 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.

[40] 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.

[41] 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).

[42] 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.

[43] 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).

[44] 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).

[45] 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.

[46] 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.

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

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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).

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

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

[63] 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.

[64] 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.

[65] 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).

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

[67] 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.

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

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

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

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

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

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

[74] 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.

[75] K. Rajamanickam. “A Mini Review on Different Methods of Functional-MRI Data Analysis Citation: Karunanithi Rajamanickam. A Mini Review on Different Methods of Functional-MRI Data Analysis.” In: *Archives of Internal Medicine Research* 03 (2020), pp. 44–060.

[76] Nuzillard, D. and Bijaoui, A. “Blind source separation and analysis of multispectral astronomical images.” In: *Astron. Astrophys. Suppl. Ser.* 147.1 (2000), pp. 129–138.

[77] E. Bingham and A. Hyvarinen. “ICA of complex valued signals: a fast and robust deflationary algorithm.” In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*. Vol. 3. 2000, 357–362 vol.3.

[78] A. D. Back and A. S. Weigend. “A First Application of Independent Component Analysis to Extracting Structure from Stock Returns.” In: *Econometrics: Applied Econometrics & Modeling eJournal* (1997).

[79] E. Bingham, J. Kuusisto, and K. Lagus. “ICA and SOM in text document analysis.” In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. 2002, pp. 361–362.

[80] 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.

[81] 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.

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

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

[84] 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).

[85] 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).

[86] 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.

[87] 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.

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

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

[90] 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.

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

[92] 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.

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

[94] 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
$[d]$	$\{1, 2, \dots, d\}$
G	Latent Causal Graph (LCG) over $\mathbf{V}$ induced by an $\mathcal{M}$
$\mathcal{M}$	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 (LSG, 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. (2).
$\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]$	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.(7).
$\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S]$	$\Delta\mathbf{Q}$ set defined in Def. 3.1. Variables in $\Delta\mathbf{Q}$ set remains from Eq.(7) to Eq.(11). 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 $\mathbf{V} \backslash \Delta\mathbf{Q}$.
L	Number of distributions used to compare in Proposition 3 and 4.

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, 30, 49]. We discuss those assumptions and their implications here.

Assumptions for ASCMs Here we discuss more about the invertibility assumption for the mixing function $f_{\mathbf{X}}$ in an ASCM.

Assumption 1 (The invertibility of mixing function)**.** *Assume that the mixing function $f_{\mathbf{X}}$ is invertible.* □

The mapping from generative factors $\mathbf{V}$ to high dimensional mixture $\mathbf{X}$ is a one-to-one mapping. Consider images. In one direction, $\mathbf{V}$ constructs the image through a mixing tool $f_{\mathbf{X}}$ (such as a camera lens). In the reverse direction, these generative factors $\mathbf{V}$ can be uniquely labeled through $f_{\mathbf{X}}^{-1}$. We take images example in Sec. D.1 as an example. The generative factors $Gender$, Age and $Haircolor$ are directly expressed through pixels in images. Given an image, the values of these generative factors are uniquely determined. This assumption is commonly used in non-linear ICA and representation learning literature [9, 10, 17, 58].

Assumption 2 (Unobserved confounders are not part of the high-dimensional mixing function)**.** *Assume that for all unobserved confounders $\mathbf{U}$ in the ASCM M, $f_{\mathbf{X}}$ is not a function of $\mathbf{U}$. That is the unobserved confounders do not show up in the function signature of the mixing function g_X.* □

This assumption is a technical one, which assumes the unobserved confounders in the LCG (represented as bidirected edges), do not directly influence the high-dimensional mixture $\mathbf{X}$, but only through the latent causal variables $\mathbf{V}$. An example of when this can occur in the real-wordl is when modeling high-dimensional T1 MRI scans. Let the LCG comprise of Drug Treatment $\rightarrow$ Outcome, but they are confounded by socioeconomic status (Drug Treatment $\leftrightarrow$ Outcome). The drug treatment and outcome are assumed to be visually discernable on the MRI. However, the socioeconomic status does not directly impact how the MRI appears, except through how it impacts the drug treatment efficacy or outcome.

Assumptions for Domain Changes We discuss assumptions related to domain changes here. The next assumption simplifies the effect of S-nodes when considering the selection diagram.

Assumption 3 (Shared causal structure)**.** *Assume that each environment's ASCM shares the same latent causal graph. That is, the S-nodes do not change the underlying structure of the causal diagram among the latent variables.* □

This means that the S-nodes will not represent structural changes such as when V_i has a different parent set across domains [14].

Assumption 4 (Mixing function is shared across all domains)**.** *Assume that $f_{\mathbf{X}}$ is shared for all ASCMs $\mathcal{M}^i \in \boldsymbol{\mathcal{M}}$. That is, there is no S-node that points to $\mathbf{X}$ such that the mixing function is different across any two domains $\Pi_i \neq \Pi_j \in \mathbf{\Pi}$.* □

This assumption characterizes the generative model that we consider. Sharing of the mixing function is needed for the multi-domain setting because if everything may change across environments, the domains can only be analysed in isolation, and thus unable to leverage the changes (and similarities) across domains.

Assumptions for Interventions We discuss assumptions related to interventions here.

Assumption 5 (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.* □

[14] The assumption that there are no structural changes between domains can be relaxed and is considered in the context of inference, as discussed in [31]. This is an interesting topic for future explorations, and we do not consider this avenue here.

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 [57]. We do not consider such interventions, and leave this general case for future work. Note Assumption 5 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 [36, 37, 39]. 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 6 (Known-target assumption)**.** *Assume for any $\mathbf{I}^{(k)} \in \boldsymbol{\Psi}$, 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.

Assumptions for Distributions In Sec. 2, we discuss that each distribution resulting from an intervention is sufficiently distinct from another distribution Assumption 7. Here we formally define and illustrate what is "change sufficiently".

Assumption 7 (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 LSG 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 $\{a'_1, \dots, a'_{|Q|}\} \subseteq \{a_1, \dots, a_L\}$ such that for all $V_q \in \mathbf{Q}, V_q \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_q)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$, 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*

$$\boldsymbol{\omega}_1(\mathbf{v}, a_l) = \left(\oplus \left(\frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q}\right)_{V_q \in \mathbf{Q}} \right) \tag{35}$$

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,*

$$\begin{aligned} \boldsymbol{\mathcal{E}} = \{\boldsymbol{\epsilon}_j = \{V_p, V_q\} \,| i)\ &\exists a_l, \{V_p, V_q\} \in \Delta\mathbf{Q}^{(a_l),(a_0)}; \\ &ii)\ V_p \not\perp\!\!\!\perp V_q \mid \mathbf{V}^{tar} \backslash \{V_p, V_q\} \text{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\} \end{aligned} \tag{36}$$

If there exists $\{a'_1, \dots, a'_{2|Q|+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \dots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}, V_i^{tar} \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], V_i^{tar} \in \Delta\mathbf{Q}^{(a'_{|Q|+i}),(a_0)}$ and for all $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2|Q|+j}),(a_0)}$, 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*

$$\begin{aligned}\boldsymbol{\omega}_2(\mathbf{v}, a_l) = \Bigg(& \oplus (\frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q})_{V_q \in \mathbf{Q}}, \\ & \oplus (\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_q^2})_{V_q \in \mathbf{Q}}, \\ & \oplus (\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(a_0)}(\mathbf{v})}{\partial v_p v_q})_{(V_p, V_q) \in \mathcal{E}(G_{\overline{T}}(\mathbf{Q}))} \Bigg)\end{aligned} \tag{37}$$

□

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 [49]. This kind of assumption is generally included in the causal representation learning literature, such as the "genericity" assumption [49], 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 first order condition:

Example 17. Consider $\Delta\mathbf{Q}$ obtained after comparisons as

$$\Delta\mathbf{Q}^{(1),(0)} = \{V_1\}, \Delta\mathbf{Q}^{(2),(0)} = \{V_1\}, \Delta\mathbf{Q}^{(1),(0)} = \{V_1, V_2, V_3\}, \tag{38}$$

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

$$\frac{\log p_{\mathbf{T}}^{(1)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_2} = 0 \tag{39}$$

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

$$\begin{aligned}&\boldsymbol{\omega}_1(v_1, v_2, v_3, 1) = (\frac{\partial \log p_{\mathbf{T}}^{(1)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_1}, 0, 0) \\ &\boldsymbol{\omega}_1(v_1, v_2, v_3, 2) = (\frac{\partial \log p_{\mathbf{T}}^{(2)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_1}, 0, 0) \\ &\boldsymbol{\omega}_1(v_1, v_2, v_3, 3) = (\frac{\partial \log p_{\mathbf{T}}^{(3)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_1}, \frac{\partial \log p_{\mathbf{T}}^{(3)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_2}, \frac{\partial \log p_{\mathbf{T}}^{(3)}(\mathbf{v}) - \log p_{\mathbf{T}}^{(0)}(\mathbf{v})}{\partial v_3})\end{aligned} \tag{40}$$

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 [49]. Here we give an example of how this assumption can be violated.

Example 18 (Distributions do not change sufficiently)**.** Consider intervention targets

$$\boldsymbol{\Psi} = \{\mathbf{I}^{(1)} = \{\{\}^{\Pi_1}\}, \mathbf{I}^{(2)} = \{V_1^{\Pi_1, \{1\}}\}, \mathbf{I}^{(3)} = \{V_2^{\Pi_1, \{2\}}\}, \mathbf{I}^{(4)} = \{V_1^{\Pi_1, \{1\}}, V_2^{\Pi_1, \{2\}}\}\} \tag{41}$$

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

$$\boldsymbol{\omega}_1(\mathbf{v}, 2) + \boldsymbol{\omega}_1(\mathbf{v}, 3) = \boldsymbol{\omega}_1(\mathbf{v}, 4) \tag{42}$$

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 7 formally.

Lemma 1. *Assumption 7 almost surely holds.* □

A.3 Domains vs Interventions

Example 19 (Example illustrating CRID with domains)**.** Consider the LSD shown in Fig. 4(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 20 (Example illustrating CRID with interventions across domains with different mechanisms)**.** Consider the LSD shown in Fig. 4(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 21 (Example illustrating CRID with interventions across domains with the same mechanisms)**.** Consider the LSD shown in Fig. 4(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, 49]. 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 25).

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: $G_{\mathbf{V},\widehat{\mathbf{V}}}$
Initialize an empty graph $G_{\mathbf{V},\widehat{\mathbf{V}}}$
for $V_i$ in $\mathbf{V}$ do
    for $V_j$ in $\hat{\mathbf{V}}$ do
        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 [3'] 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 condition [3] is also satisfied:*

> *[3'] there exists $\{a'_1, \dots, a'_{|\mathbf{V}^{tar}|}\} \subseteq \{a_1, \dots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}\backslash\mathbf{Mem}, V_i^{tar} \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_i)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$.*

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.

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

```
Input: $G_{V,\hat{V}}$, $Z$, $\mathbf{Q}$
Output: $\mathbf{Mem}$
$\mathbf{Mem} \leftarrow \{\}$
for all $\widehat{V} \in \mathbf{Q}$ do
    if $Z \to \widehat{V} \notin G_{\mathbf{V},\hat{\mathbf{V}}}$ then
        $\mathbf{Mem}.append(V)$
return $\mathbf{Mem}$
```

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 $\mathbf{Q}$. $G_{\mathbf{V},\hat{\mathbf{V}}}$ 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: $\mathbf{Q}, G_{\mathbf{V},\hat{\mathbf{V}}}, G_{\overline{\mathbf{T}}}, \mathbf{\Psi_T}, \mathbf{I}, \mathcal{Q}$
Output: $G_{\mathbf{V},\hat{\mathbf{V}}}$
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
        $\mathbf{Mem}_i \leftarrow CheckMemoize(G_{\mathbf{V},\hat{\mathbf{V}}}, V_i, \mathbf{Q})$    ▷ Variables in $\mathbf{Q}$ is ID w.r.t $V_i$ already.
        $\mathbf{Mem}_j \leftarrow CheckMemoize(G_{\mathbf{V},\hat{\mathbf{V}}}, V_j, \mathbf{Q})$    ▷ 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 $G_{\mathbf{V},\hat{\mathbf{V}}}$
return $G_{\mathbf{V},\hat{\mathbf{V}}}$
```

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 [4] in Prop. 4 can be reduced to a weaker condition [4'] leveraging existing disentanglements in CDM.

Lemma 3 (ID of variables within ΔQ 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*

$$\begin{aligned}\boldsymbol{\mathcal{E}}_{ij} =& \{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid i)\ \exists a_l, \{V_k, V_r\} \in \Delta\mathbf{Q}^{(a_l),(a_0)};\\ & ii)\ V_k \text{ is connected to} V_r \text{ conditioning } \mathbf{V}^{tar} \backslash \{V_k, V_r\} \text{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\\ & iii)\ V_k, V_r \notin \mathbf{Mem}_i \cup \mathbf{Mem}_j\}\end{aligned} \tag{43}$$

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) $\{V_k, V_r\} \not\subseteq \mathbf{Mem}_i \cup \mathbf{Mem}_j$ 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 the detailed proofs of theoretical results in the main paper.

C.1 Distribution comparison - Proof of Proposition 1

Proposition 1 (Distribution Comparison). *Consider a collection of ASCMs $\boldsymbol{\mathcal{M}} = \langle \mathcal{M}_1, \dots, \mathcal{M}_n \rangle$ that induces collection distribution $\mathcal{P}$ 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 mechanism on $\mathbf{T}$. If another collection of ASCMs, $\widehat{\boldsymbol{\mathcal{M}}} = \langle \widehat{\mathcal{M}}_1, \dots, \widehat{\mathcal{M}}_n \rangle$, matches with distribution $\mathcal{P}$ and LSG G^S, then*

$$\sum_i^d \log p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = \sum_i^d \log p_{\mathbf{T}}^{(j)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(k)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}), \tag{7}$$

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:

$$\widehat{\mathbf{V}} = \widehat{f}_{\mathbf{X}}^{-1}(\mathbf{X}) = \widehat{f}_{\mathbf{X}}^{-1}(f_{\mathbf{X}}(\mathbf{V})) \tag{44}$$

Then based on the change variable formula, we have

$$p(\mathbf{v}) = p(\widehat{\mathbf{v}})|\mathbf{J}_{\boldsymbol{\phi}}| \tag{45}$$

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. 2 and taking log of the above equation,

$$\sum_{i=1}^d \log p_{\mathbf{T}}(v_i \mid \mathbf{pa}^{\mathbf{T}+}) = \sum_{i=1}^d \log p_{\mathbf{T}}(\widehat{v}_i \mid \widehat{\mathbf{pa}}^{\mathbf{T}+}) + \log|\mathbf{J}_{\boldsymbol{\phi}}| \tag{46}$$

Subtract the above factorization of density function induced by $\mathbf{I}^{(j)}$ and $\mathbf{I}^{(k)}$, and we have Eq.(7). □

Eq 7 naturally gives a connection from $\mathbf{V}$ to $\widehat{\mathbf{V}}$. Comparing two factorization for Fig. 4(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.2 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 not needed) with target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ as follows. Add variable V_l to $\Delta\mathbf{V}$ if,*

1. *$V_l \in \Delta\mathbf{V}$ 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. *$V_l \in \Delta\mathbf{V}$ 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 $V_i \in \mathbf{V} \backslash \mathbf{C}_{\geq}(\Delta\mathbf{V})$, then $p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+})$, which will be denoted as invariant factors, where $\mathbf{C}_{\geq}(\Delta\mathbf{V})$ are variables in the same C-Component with $\Delta\mathbf{V}$ and not before $\Delta\mathbf{V}$ in the topological order for factorization. □

Proof. Consider an arbitrary $V_i \in \backslash \mathbf{C}_{>}(\Delta\mathbf{V})$. First, 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.

Second, consider variables $\mathbf{Z}$ that and are in the same C-component with V_i and also are before V_i in the topological order. Then we have, $\mathbf{Z} \cap \mathbf{C}_{\geq}(\Delta\mathbf{V}) = \emptyset$. The reason is that (1) if $V_i \in \mathbf{C}(\Delta\mathbf{V})$

but V_i has a strictly lower order than $\Delta\mathbf{V}$, then $\mathbf{Z}$ also have strictly lower order than $\Delta\mathbf{V}$); (2) if $V_i \notin \mathbf{C}(\Delta\mathbf{V})$, $\mathbf{C}(V_i) \cap \mathbf{C}(\Delta\mathbf{V}) = \emptyset$. According to the definition of $\mathbf{Pa}_i^{\mathbf{T}^+}$, we know $\mathbf{Pa}_i^{\mathbf{T}^+} \backslash \mathbf{Z} = \mathbf{Pa}(\{V_i\} \cup \mathbf{Z})$. We have

$$P_{\mathbf{T}}^{\Pi^{(k)}}(V_i \mid \mathbf{Pa}_i^{\mathbf{T}+}; \sigma^{(k)}) = P_{\mathbf{T}}^{\Pi^{(k)}}(V_i, \mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)}) / P_{\mathbf{T}}^{\Pi^{(k)}}(\mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)}) \tag{47}$$

Since the mechanism and exogenous variables will not change, the numerator $P_{\mathbf{T}}^{\Pi^{(k)}}(V_i, \mathbf{Z} \mid \mathbf{Pa}_i(\{V_i\} \cup \mathbf{Z}); \sigma^{(k)})$ also does not change. This is because $Z = f_Z(\mathbf{Pa}_Z, U_Z)$ and $U_Z \mid \mathbf{Pa}_Z$ for any $Z \in \{V_i\} \cup \mathbf{Z}$. The denominator is an integral of the numerator thus the denominator also does not change. Then the corresponding density function will remain the same, namely,

$$p_{\mathbf{T}}^{(j)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) = p_{\mathbf{T}}^{(k)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) \tag{48}$$

□

C.3 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 $\mathbf{V}^{tar} = \{V_1^{tar}, V_2^{tar}, \ldots, V_{d'}^{tar}\} \subseteq \mathbf{V}$. 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 $\mathbf{T}$) $\forall\ l \in [L]$, $\mathbf{T} = do[\mathbf{I}^{(a_0)}] \subseteq do[\mathbf{I}^{(a_l)}]$ [15].

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

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

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

Proof. **We denote $\mathbf{V}^{tar}$ as Q for convenience.** Notice that the Assumption 7 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

$$\begin{aligned}\sum_{V_i \in \tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - p_{\mathbf{T}}^{a_0)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) &= \sum_{V_i \in \tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) \\ &= \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})\end{aligned} \tag{49}$$

from Eq. (11).

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 $Z \in \mathbf{V}\backslash\mathbf{Q}$,

$$\forall l \in [L], \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{z}} = 0 \tag{50}$$

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

$$\begin{aligned}\forall l \in [L], 0 &= \sum_{v_i \in \mathbf{V}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_i} \frac{\partial \widehat{v}_i}{\partial z} && \text{(Chain Rule)} \qquad (51)\\ &= \sum_{v_q \in \mathbf{Q}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_q} \frac{\partial \widehat{v}_q}{\partial z} && \text{(Eq.(50))} \qquad (52)\end{aligned}$$

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

Eq. (52) is a linear system for unknowns $\{\partial\widehat{V}_q/\partial Z\}_{V_q\in\mathbf{Q}}$. When distribution changes sufficiently, namely under Assumption 7, 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,

$$\forall\, V_q \in \mathbf{Q}, \frac{\partial\widehat{v}_q}{\partial z} = 0 \tag{53}$$

Recall that $V_q = \phi_{V_q}(\mathbf{V})$. For any $Z \in \mathbf{V}\backslash\mathbf{Q}$, Eq.(53) 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 condition [3] is also satisfied:*

> *[3'] there exists $\{a'_1, \dots, a'_{|\mathbf{V}^{tar}|}\} \subseteq \{a_1, \dots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}\backslash\mathbf{Mem}, V_i^{tar} \in \Delta\mathbf{Q}[\mathbf{I}^{(a'_i)}, \mathbf{I}^{(a_0)}, \mathbf{T}, G^S]$.*

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.(52) exclude $\frac{\partial v_m}{\partial z}$. Then, when [3'] holds, the system will have zero solutions and Eq.(53) will hold. □

C.4 ID within ΔQ 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. 7).

Proposition 4 (ID of variables within ΔQ sets). *Consider the 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 $G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})$, 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) there exists $\{a'_1, \dots, a'_{2d'+|\boldsymbol{\mathcal{E}}|}\} \in \{a_1, \dots, a_L\}$ such that for all $V_i^{tar} \in \mathbf{V}^{tar}, V_i^{tar} \in \Delta\mathbf{Q}^{(a'_i),(a_0)}], V_i^{tar} \in \Delta\mathbf{Q}^{(a'_{d'+i}),(a_0)}$ and for all $\boldsymbol{\epsilon}_j \in \boldsymbol{\mathcal{E}}, \boldsymbol{\epsilon}_j \subseteq \Delta\mathbf{Q}^{(a'_{2d'+j}),(a_0)}$, where $d' = |\mathbf{V}^{tar}|$ and*
>
> $$\begin{aligned}\boldsymbol{\mathcal{E}} =&\{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid i)\ \exists a_l, \{V_k, V_r\} \in \Delta\mathbf{Q}^{(a_l),(a_0)};\\ &V_k \textit{ is connected to } V_r \textit{ conditioning } \mathbf{V}^{tar}\backslash\{V_k, V_r\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\}\end{aligned} \tag{22}$$

, then V_i is ID w.r.t V_j. □

Proof. **We denote $\mathbf{V}^{tar}$ as Q for convenience.** Notice that Assumption 7 will be used in the proof. From Eq. 11, we have

$$\begin{aligned}\sum_{V_i\in\tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) - p_{\mathbf{T}}^{a_0)}(v_i \mid \mathbf{pa}_i^{\mathbf{T}+}) &= \sum_{V_i\in\tilde{\mathbf{V}}} \log p_{\mathbf{T}}^{(a_l)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{v}_i \mid \widehat{\mathbf{pa}}_i^{\mathbf{T}+})\\ &= \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})\end{aligned} \tag{54}$$

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 $V_i \notin \mathbf{Pa}_j^{\mathbf{T}+}, V_j \notin \mathbf{Pa}_i^{\mathbf{T}+}$ and $V_i, V_j \notin \mathbf{Pa}_m^{\mathbf{T}+}$ 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,

$$\frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(v_m \mid \mathbf{pa}_m^{\mathbf{T}+})}{\partial v_i v_j} = 0 \tag{55}$$

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

$$\begin{aligned} \forall l \in [L], & \sum_{V_m \in \tilde{V}} \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{v}_m \mid \mathbf{pa}_m^{\mathbf{T}+})}{\partial \widehat{v}_k \widehat{v}_r} \\ & = \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} = 0 \end{aligned} \tag{56}$$

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

$$\forall l \in [L], = \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} = 0 \tag{57}$$

since

$$\frac{\partial p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k} = 0 \quad \text{or} \quad \frac{\partial p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_r} = 0 \tag{58}$$

Upon Eq. (55), taking the second partial derivative on both sides of Eq. (54), the left side will be 0, and then $\forall\ l \in [L]$, we have

$$0 = \sum_{V_k, V_r \in \mathbf{Q}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} \frac{\partial \widehat{v}_k}{\partial v_i} \frac{\widehat{v}_r}{\partial v_j} \qquad \text{Chain Rule} \tag{59}$$

$$\begin{aligned} = & \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_i^2} \frac{\partial \widehat{v}_i}{\partial v_i} \frac{\partial \widehat{v}_i}{\partial v_j} + \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_j^2} \frac{\partial \widehat{v}_j}{\partial v_i} \frac{\partial \widehat{v}_j}{\partial v_j} \\ & + \sum_{V_q \in \mathbf{Q}} \frac{\partial^2 \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_q^2} \frac{\partial \widehat{v}_q}{\partial v_i} \frac{\partial \widehat{v}_q}{\partial v_j} \\ & + \sum_{V_q \in \mathbf{Q}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_q} \frac{\partial^2 \widehat{v}_q}{\partial v_i v_j} \\ & + \sum_{(V_k, V_r) \in \boldsymbol{\mathcal{E}}} \frac{\partial \log p_{\mathbf{T}}^{(a_l)}(\widehat{\mathbf{v}}) - \log p_{\mathbf{T}}^{(a_0)}(\widehat{\mathbf{v}})}{\partial \widehat{v}_k \widehat{v}_r} \frac{\partial \widehat{v}_k}{\partial v_i} \frac{\widehat{v}_r}{\partial v_j} \qquad \text{Eq. (56) and (56)} \end{aligned} \tag{60}$$

Eq.(60) is also a linear system. When distribution changes sufficiently, namely under Assumption 7, 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,

$$\frac{\partial \widehat{v}_i}{\partial v_i} \frac{\partial \widehat{v}_i}{\partial v_j} = 0, \frac{\partial \widehat{v}_j}{\partial v_i} \frac{\partial \widehat{v}_j}{\partial v_j} = 0 \tag{61}$$

Then we have

$$\frac{\partial \widehat{v}_i}{\partial v_j} = 0, \frac{\partial \widehat{v}_i}{\partial v_j} = 0 \tag{62}$$

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*

$$\begin{aligned}\boldsymbol{\mathcal{E}}_{ij} =&\{\boldsymbol{\epsilon}_j = \{V_k, V_r\} \mid \textit{i) } \exists a_l, \{V_k, V_r\} \in \Delta\mathbf{Q}^{(a_l),(a_0)};\\ &\textit{ii) } V_k \textit{ is connected to} V_r \textit{ conditioning } \mathbf{V}^{tar} \backslash \{V_k, V_r\} \textit{ in } G_{\overline{\mathbf{T}}}(\mathbf{V}^{tar})\\ &\textit{iii) } V_k, V_r \notin \mathbf{Mem}_i \cup \mathbf{Mem}_j\}\end{aligned} \tag{43}$$

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

$$\frac{\partial \widehat{v}_q}{\partial v_i} \frac{\partial \widehat{v}_q}{\partial v_j} = 0, \tag{63}$$

if V_p is ID w.r.t V_i or V_q is ID w.r.t V_j.

$$\frac{\partial^2 \widehat{v}_q}{\partial v_i v_j} = 0 \tag{64}$$

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 ΔQ sets). *Consider the variables* $\mathbf{V}^{tar} \subseteq \mathbf{V}$, $\mathcal{P}_{\mathbf{T}}$ *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.5 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. ΔQ sets). *Suppose* $\boldsymbol{\Psi}$ *contains* $do(\mathbf{T})$. *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

$$P_{\mathbf{T}}(\mathbf{V}) = P_{\mathbf{T}}(V^{tar}) P_{\mathbf{T}}(\mathbf{Z}) \tag{65}$$

With the change of variable formulation and taking log:

$$\log p_{\mathbf{T}}(\mathbf{v}^{tar}) + \log p_{\mathbf{T}}(\mathbf{z}) = \log p_{\mathbf{T}}(\widehat{\mathbf{v}}^{tar}) + \log p_{\mathbf{T}}(\widehat{\mathbf{z}}) + \log |\mathbf{J}_{\boldsymbol{\phi}}| \tag{66}$$

Since $\mathbf{Z}$ is ID w.r.t V^{tar}, $\partial \widehat{\mathbf{Z}} / \partial \mathbf{V}^{tar} = 0$. In other words, the elements $\partial \phi_Z / \partial \mathbf{V}^{tar} = 0$ 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

$$\log |\mathbf{J}_{\boldsymbol{\phi}}| = \log |\mathbf{J}_{\mathbf{Z}}| + \log |\mathbf{J}_{\mathbf{V}^{tar}}| \tag{67}$$

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,

$$\log p_{\mathbf{T}}(\mathbf{z}) = \log p_{\mathbf{T}}(\widehat{\mathbf{z}}) + \log|\mathbf{J}_{\mathbf{Z}}| \tag{68}$$

Subtracting this to Eq. (66)

$$\log p_{\mathbf{T}}(v^{tar}) = \log p_{\mathbf{T}}(\widehat{v}^{tar}) + \log|\mathbf{J}_{\mathbf{V}^{tar}}| \tag{69}$$

Denote $\boldsymbol{\phi}_{\mathbf{V}^{tar}}(\mathbf{z}, \cdot)$ as $\boldsymbol{\phi}^{\mathbf{z}}_{\mathbf{V}^{tar}}(\cdot)$, which is the function $\boldsymbol{\phi}_{\mathbf{V}^{tar}}$ fixing value $\mathbf{Z} = \mathbf{z}$ mapping from $\mathbf{V}^{tar}$ to $\widehat{\mathbf{V}}$. This suggests for every $\mathbf{z}$,

$$P_{\mathbf{T}}(\widehat{V}^{tar}) = P_{\mathbf{T}}(\boldsymbol{\phi}^{\mathbf{z}}_{V^{tar}}(V^{tar})) \tag{70}$$

Apply Lemma 2 of [20], $\boldsymbol{\phi}_{V^{tar}}$ should be a constant regarding $\mathbf{Z}$. Thus,

$$\forall\, Z \in \mathbf{Z}, \frac{\partial V^{tar}}{\partial Z} = 0 \tag{71}$$

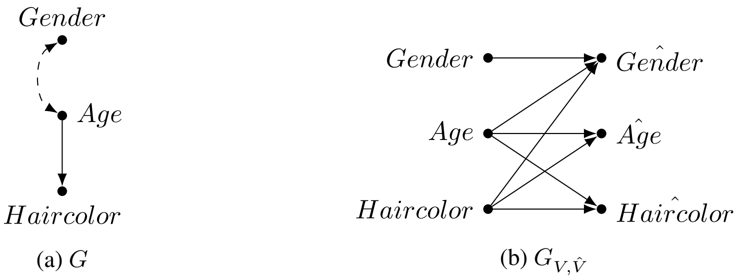
□

C.6 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**, $G_{V,\hat{V}}$, 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 [25]. 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.

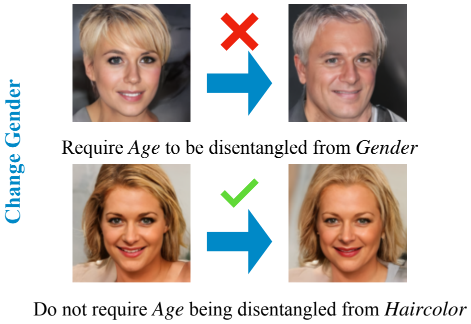

Change Gender
Require Age to be disentangled from Gender
Do not require Age being disentangled from Haircolor

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. (2). First, the concept C-component is formally defined as follows:

Definition 6.1 (Confounded Component)**.** Let $\{\mathbf{C_1}, \mathbf{C_2}, \dots, \mathbf{C_k}\}$ 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

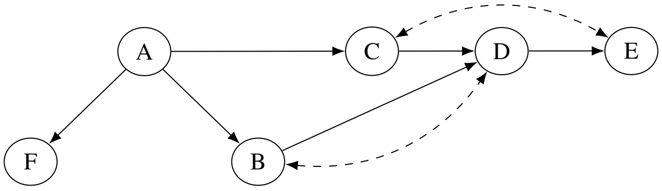


Figure S4: Causal graph with four C-components.

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 $V_1, \dots, V_n$ 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. (2) 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 [72, 73].

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 [74])**.** 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 (Δ**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 index not needed) of the target sets $\mathbf{I}^{(j)}, \mathbf{I}^{(k)}$ is the remaining variables after comparison (i.e. Eq. 11),

$$\Delta\mathbf{Q}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, \mathbf{T}, G^S] = \tilde{\mathbf{V}} \cup \mathbf{Pa}^{\mathbf{T}+}(\tilde{\mathbf{V}}), \tag{15}$$

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

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. 7, $\Delta\mathbf{Q}[\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{T}, G^S)]$ 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. Let the $\Delta\mathbf{Q}_A$ be the result derived from order A and $\Delta\mathbf{Q}_B$ be the result derived from order B. More specifically,

$$\begin{aligned}\Delta\mathbf{Q}_A &= \mathbf{C}_{\geq_A}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]) \cup \mathbf{Pa}_A^{\mathbf{T}+}(\mathbf{C}_{\geq_A}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]))\\ \Delta\mathbf{Q}_B &= \mathbf{C}_{\geq_B}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]) \cup \mathbf{Pa}_B^{\mathbf{T}+}(\mathbf{C}_{\geq_B}(\Delta\mathbf{V}[\mathbf{I}^{(j)}, \mathbf{I}^{(k)}, G^S]))\end{aligned} \tag{72}$$

Notice that the order only has influence on the extended parents set and $\mathbf{C}_{\geq}$. Recall the extended parents $\mathbf{Pa}_i^{\mathbf{T}+} = \overline{\mathbf{Pa}}(\{V \in \mathbf{C}(V_i) : V \leq V_i\}) \setminus \{V_i\}$. We will discuss the following two cases.

First Case. $\mathbf{C}_{\geq_A}(\Delta\mathbf{V}) = \mathbf{C}_{\geq_B}(\Delta\mathbf{V}) = \mathbf{Z}$. Then, $\bigcup_{Z \in \mathbf{Z}}\{V \in \mathbf{C}(Z) : V \leq_A Z\}$ is equivalent to $\bigcup_{Z \in \mathbf{Z}}\{V \in \mathbf{C}(Z) : V \leq_B Z\}$. Thus, $\mathbf{P}a_A^{\mathbf{T}+}(\mathbf{Z}) = \mathbf{P}a_B^{\mathbf{T}+}(\mathbf{Z})$.

Second Case. $\mathbf{C}_{\geq_A}(\Delta\mathbf{V}) \neq \mathbf{C}_{\geq_B}(\Delta\mathbf{V})$. W.L.O.G, assume $W \in \mathbf{C}_{\geq_A}(\Delta\mathbf{V})$ but $W \notin \mathbf{C}_{\geq_B}(\Delta\mathbf{V})$. This implies W is before $\Delta\mathbf{V}$. Then W must be in $\mathbf{P}a_B^{\mathbf{T}+}(\mathbf{C}_{\geq_B}(\Delta\mathbf{V}))$. Thus, $\Delta\mathbf{Q}_A = \Delta\mathbf{Q}_B$. □

This lemma 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 detailed examples of Proposition 3 and 4

Proposition 3 and 4 disentangle variables through comparing distributions. With enough distributions, one can build a linear system (illustrated in Appendix C.3 and C.4).

Example 22. (details for Example 11).

Suppose the pair of underlying ASCMs $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ induces the LSG G^S in Fig. 2 and distributions $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\} = \{P^{\Pi_1}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}), P^{\Pi_2}(\mathbf{X}; \sigma_{V_3}), P^{\Pi_2}(\mathbf{X}; \sigma_{V_4})\}$ from interventions $\Sigma = \{\sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)}\} = \{\{\}, \{\}, \sigma_{V_3}, do(V_2)\}$. Suppose we are given intervention targets $\mathbf{\Psi} = \{\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{I}^{(3)}, \mathbf{I}^{(4)}\} = \{\{\}^{\Pi_1}, \{\}^{\Pi_2}, V_3^{\Pi_2}, V_2^{\Pi_1, \text{do}}\}$ and G^S.

Consider $\mathbf{V}^{tar} = \{V_2, V_3\}$, $\mathbf{V}^{en} = \mathbf{V} \backslash \{V_2, V_3\} = \{V_1\}$. When comparing $\{P^{(2)}, P^{(3)}\}$ with the baseline $P^{(1)}$, $\mathbf{T} = do[\mathbf{I}^{(1)}] = \{\}$, and then

$$\Delta\mathbf{Q}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, \mathbf{T}, G^S] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(1)}, \mathbf{T}, G^S] = \{V_2, V_3\} \tag{73}$$

$\{P^{(2)}, P^{(3)}\}$ satisfies condition [1] in Prop. 3, since the hard intervention variable set is $\{\}$. They also satisfies condition [2], since $\Delta\mathbf{Q}^{(2),(1)} = \Delta\mathbf{Q}^{(3),(1)} = \mathbf{V}^{tar}$. Condition [3] are satisfied since $V_2 \in \Delta\mathbf{Q}^{(2),(1)}$ and $V_3 \in \Delta\mathbf{Q}^{(3),(1)}$ Thus, $\mathbf{V}^{tar}$ is ID w.r.t $\mathbf{V}^{en}$ by Prop. 3. This demonstrates that a variable V_2 can be disentangled from another variable that is in the C-component (V_1).

By comparing distribution resulting from $\sigma^{(2)}$ and $\sigma^{(3)}$ with the baseline $\sigma^{(1)}$,

$$\log p^{(2)}(v_3 \mid v_2) - \log p^{(1)}(v_3 \mid v_2) = \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) \tag{74}$$

$$\log p^{(3)}(v_3 \mid v_2) - \log p^{(1)}(v_3 \mid v_2) = \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) \tag{75}$$

Taking the first order partial derivative w.r.t. V_1:

$$\begin{aligned}0 &= \frac{\partial \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_2}\frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_3}\frac{\partial \widehat{v}_3}{\partial v_1}\\ 0 &= \frac{\partial \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_2}\frac{\partial \widehat{v}_2}{\partial v_1} + \frac{\partial \log p^{(3)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2)}{\partial \widehat{v}_3}\frac{\partial \widehat{v}_3}{\partial v_1}\end{aligned} \tag{76}$$

In this system, notice that

$$\log p^{(2)}(\widehat{v}_3 \mid \widehat{v}_2) - \log p^{(1)}(\widehat{v}_3 \mid \widehat{v}_2) = \log p^{(2)}(\widehat{v}_1, \widehat{v}_2, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_1, \widehat{v}_2, \widehat{v}_3) \tag{77}$$

Then since the coefficient is linear independent assumed in Assumption 7, we have

$$\frac{\partial \widehat{v}_2}{\partial v_1} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0 \tag{78}$$

Then $V_2 = \tau_2(V_2, V_3)$, and $V_3 = \tau_3(V_2, V_3)$.

First, this example shows we can disentangle two variables in the same C-component (V_1, V_2). Second, compared with the baseline, one can disentangle variable $V_i \in \mathbf{V}$ with its descendants when soft interventions are given per node, and V_i is considered to be still entangled with its ancestors (see Sec. E.6). The above result shows that it is possible to disentangle variables from their ancestors using only soft interventions. More interestingly, no intervention is performed on V_2 while we disentangle V_2 from V_1. Compared with [22], one can disentangle V_1 and V_3 using 10 distributions and we demonstrate 3 distributions are enough. □

Example 23. (details for Example 12).

Consider the diagram in Fig. 4(c).

Suppose $\mathcal{P} = \{P^{(1)}, P^{(2)}, P^{(3)}, P^{(4)}\}$ with intervention targets

$$\mathbf{I}^{(1)} = \{\}^{\Pi_1}, \mathbf{I}^{(2)} = \{V_2\}^{\Pi_1}, \mathbf{I}^{(3)} = \{V_3\}^{\Pi_1}, \mathbf{I}^{(4)} = \{V_1\}^{\Pi_1} \tag{79}$$

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:

$$\Delta\mathbf{Q}^{(2),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(3),(1)} = \{V_1, V_2, V_3\}, \Delta\mathbf{Q}^{(4),(1)} = \{V_1\}. \tag{80}$$

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 $V_1 < V_3 < V_2 < V_4$.

$$P(\mathbf{V}) = P(V_1)P(V_3)P(V_2 \mid V_1, V_3)P(V_4 \mid V_3) \tag{81}$$

as the factorization. By comparing distribution resulting from $\sigma^{(2)}$ and $\sigma^{(3)}$ with the baseline $\sigma^{(1)}$,

$$\log p^{(2)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3) = \log p^{(2)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) \tag{82}$$

$$\begin{aligned}&\log p^{(3)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3) = \\ &\log p^{(3)}(\widehat{v}_3) - \log p^{(3)}(\widehat{v}_3) + \log p^{(2)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)\end{aligned} \tag{83}$$

$$\log p^{(4)}(v_1) - \log p^{(1)}(v_1) = \log p^{(4)}(\widehat{v}_1) - \log p^{(1)}(\widehat{v}_1) \tag{84}$$

Taking the first order partial derivative w.r.t. V_4:

$$\begin{aligned} 0 &= h_{2,1}\frac{\partial \widehat{v}_1}{\partial v_4} + h_{2,2}\frac{\partial \widehat{v}_2}{\partial v_4} + h_{2,3}\frac{\partial \widehat{v}_3}{\partial v_4} \\ 0 &= h_{3,1}\frac{\partial \widehat{v}_1}{\partial v_4} + h_{3,2}\frac{\partial \widehat{v}_2}{\partial v_4} + h_{3,3}\frac{\partial \widehat{v}_3}{\partial v_4} \\ 0 &= h_{4,1}\frac{\partial \widehat{v}_1}{\partial v_4} \end{aligned} \tag{85}$$

where

$$\begin{aligned} h_{2,i} &= \frac{\partial \log p^{(2)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_i} \quad \text{for } i = 1, 2, 3 \\ h_{3,i} &= \frac{\partial \log p^{(3)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_i} \quad \text{for } i = 1, 2, 3 \\ h_{4,1} &= \frac{\partial p^{(4)}(\widehat{v}_1) - \log p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1} \end{aligned} \tag{86}$$

Then since the coefficient is linear independent assumed in Assumption 7, we have

$$\frac{\partial \widehat{v}_1}{\partial v_4} = 0, \frac{\partial \widehat{v}_2}{\partial v_4} = 0, \frac{\partial \widehat{v}_3}{\partial v_4} = 0 \tag{87}$$

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 24. (Example 13 (continued).) Recall the given LSD G^S is shown in Fig. 2. and the 9 given intervention targets are

$$\mathbf{\Psi} = \{\{\}^{\Pi_1}, \{\{V_1^{\Pi_1}\} \times 4, \{V_2^{\Pi_1}, V_3^{\Pi_1}\} \times 4\} \tag{88}$$

Comparing $P^{(2)}, \dots, P^{(9)}$ with $P^{(1)}$, we have

$$\log p^{(i)-(1)}(v_1) + \log p^{(i)-(1)}(v_2 \mid v_1) = p^{(i)-(1)}(\widehat{v}_1) + \log p^{(i)-(1)}(\widehat{v}_2 \mid \widehat{v}_1) \tag{89}$$

$$\log p^{(j)-(1)}(v_2 \mid v_1) + \log p^{(j)-(1)}(v_3 \mid v_2) = \log p^{(j)-(1)}(\widehat{v}_2 \mid v_1) + \log p^{(j)-(1)}(\widehat{v}_3 \mid \widehat{v}_2) \tag{90}$$

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:

$$\begin{aligned} 0 &= \sum_{p=1,2} h'_{i,p} \frac{\partial^2 \widehat{v}_p}{\partial v_1 \partial v_3} + \sum_{p=1,2} h''_{i,p} \frac{\partial \widehat{v}_p}{\partial v_1} \frac{\partial \widehat{v}_p}{\partial v_3} + h''_{i,1,2}(\frac{\partial \widehat{v}_1}{\partial v_1} \frac{\partial \widehat{v}_2}{\partial v_3} + \frac{\partial \widehat{v}_2}{\partial v_1} \frac{\partial \widehat{v}_1}{\partial v_3}) \\ 0 &= \sum_{p=1,2,3} h'_{j,p} \frac{\partial^2 \widehat{v}_p}{\partial v_1 \partial v_3} + \sum_{p=1,2,3} h''_{i,p} \frac{\partial \widehat{v}_p}{\partial v_1} \frac{\partial \widehat{v}_p}{\partial v_3} + h''_{i,2,3}(\frac{\partial \widehat{v}_2}{\partial v_1} \frac{\partial \widehat{v}_3}{\partial v_3} + \frac{\partial \widehat{v}_3}{\partial v_1} \frac{\partial \widehat{v}_2}{\partial v_3}) \end{aligned} \tag{91}$$

where $i = 2, 3, 4, 5$, $j = 6, 7, 8, 9$, and $k = 10, 11, 12, 13$, and the following are defined accordingly

$$\begin{aligned} h'_{i,p} &= \frac{\partial \log p^{(i)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_p} \quad \text{for } p = 1, 2, 3 \\ h''_{i,p} &= \frac{\partial^2 \log p^{(i)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_p^2} \quad \text{for } p = 1, 2, 3 \\ h''_{i,1,2} &= \frac{\partial^2 \log p^{(i)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3) - \log p^{(1)}(\widehat{v}_2 \mid \widehat{v}_1, \widehat{v}_3)}{\partial \widehat{v}_1 \partial \widehat{v}_2} \\ h'_{j,p} &= \frac{\partial \log p^{(j)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_p} \quad \text{for } p = 1, 2, 3 \\ h''_{j,p} &= \frac{\partial \log p^{(j)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_p^2} \quad \text{for } p = 1, 2, 3 \\ h''_{j,2,3} &= \frac{\partial \log p^{(j)}(v_3) - \log p^{(1)}(\widehat{v}_3) + \log p^{(3)}(v_2 \mid v_1, v_3) - \log p^{(1)}(v_2 \mid v_1, v_3)}{\partial \widehat{v}_2 \partial \widehat{v}_3} \end{aligned} \tag{92}$$

There are 8 unknowns that come from Eqn. 91. We can write that as a vector $\beta \in \mathbb{R}^{12}$.

$$\begin{aligned} \boldsymbol{\beta} = [&\frac{\partial^2 \widehat{v}_1}{\partial v_1 \partial v_3}, \frac{\partial^2 \widehat{v}_2}{\partial v_1 \partial v_3}, \frac{\partial^2 \widehat{v}_3}{\partial v_1 \partial v_3}, \frac{\partial \widehat{v}_1}{\partial v_1} \frac{\partial \widehat{v}_1}{\partial v_3}, \frac{\partial \widehat{v}_2}{\partial v_1} \frac{\partial \widehat{v}_2}{\partial v_3}, \frac{\partial \widehat{v}_3}{\partial v_1} \frac{\partial \widehat{v}_3}{\partial v_3} \\ &(\frac{\partial \widehat{v}_1}{\partial v_1} \frac{\partial \widehat{v}_2}{\partial v_3} + \frac{\partial \widehat{v}_2}{\partial v_1} \frac{\partial \widehat{v}_1}{\partial v_3}), (\frac{\partial \widehat{v}_2}{\partial v_1} \frac{\partial \widehat{v}_3}{\partial v_3} + \frac{\partial \widehat{v}_3}{\partial v_1} \frac{\partial \widehat{v}_2}{\partial v_3})] \end{aligned} \tag{93}$$

Rewriting Eq. (91), we have a linear system

$$\begin{pmatrix} h'_{2,1} & h'_{2,2} & 0 & h''_{2,1} & h''_{2,2} & 0 & h''_{2,1,2} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h'_{6,1} & h'_{6,2} & h'_{6,3} & h''_{6,1} & h''_{6,2} & h''_{6,3} & h''_{6,1,2} & h''_{2,2,3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \boldsymbol{\beta} = 0 \tag{94}$$

The coefficient matrix is assumed with linear independent rows in Assumption 7. Since there are 8 rows, then the matrix is full rank, and we know $\boldsymbol{\beta} = 0$. Then.

$$\frac{\partial \widehat{v}_1}{\partial v_1} \frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} \frac{\partial \widehat{v}_3}{\partial v_3} = 0 \tag{95}$$

Then since $\frac{\partial \widehat{v}_1}{\partial v_1} \neq 0$,

$$\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0 \tag{96}$$

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 25. The factorization based on G^S choosing $\mathbf{T} = \{\}$ is

$$P(\mathbf{V}) = P(V_1)P(V_3)P(V_3 \mid V_1, V_2) \tag{97}$$

By comparing distribution resulting from $\sigma^{(2)}$ and $\sigma^{(3)}$ with the baseline $\sigma^{(1)}$, for $j = 2, 3, 4, 5$

$$\log p^{(j)}(v_1) + \log p^{(j)}(v_3) - \log p^{(1)}(v_1) - \log p^{(1)}(v_3) \tag{98}$$

$$= \log p^{(j)}(\widehat{v}_{\cdot}) + \log p^{(j)}(\widehat{v}_3) - \log p^{(1)}(\widehat{v}_1) - \log p^{(1)}(\widehat{v}_3) \tag{99}$$

Taking the second order partial derivative w.r.t. V_1, V_3:

$$\begin{aligned} 0 = &\frac{\partial^2 \log p^{(j)}(\widehat{v}_1)p^{(j)} - \log p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1^2} \frac{\partial \widehat{v}_1}{\partial v_1} \frac{\partial \widehat{v}_1}{\partial v_3} + \frac{\partial^2 \log p^{(j)}(\widehat{v}_3)p^{(j)} - \log p^{(1)}(\widehat{v}_3)}{\partial \widehat{v}_3^2} \frac{\partial \widehat{v}_3}{\partial v_1} \frac{\partial \widehat{v}_3}{\partial v_3} \\ &+ \frac{\partial \log p^{(j)}(\widehat{v}_1)p^{(j)} - \log p^{(1)}(\widehat{v}_1)}{\partial \widehat{v}_1} \frac{\partial^2 \widehat{v}_1}{\partial v_1 \partial V_3} + \frac{\partial \log p^{(j)}(\widehat{v}_3)p^{(j)} - \log p^{(1)}(\widehat{v}_3)}{\partial \widehat{v}_3} \frac{\partial^2 \widehat{v}_3}{\partial v_1 \partial v_3} \end{aligned} \tag{100}$$

Then since the coefficient is linear independent assumed in Assumption 7, we have

$$\frac{\partial \widehat{v}_1}{\partial v_1} \frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} \frac{\partial \widehat{v}_3}{\partial v_3} = 0 \tag{101}$$

Then after permutation,

$$\frac{\partial \widehat{v}_1}{\partial v_3} = 0, \frac{\partial \widehat{v}_3}{\partial v_1} = 0 \tag{102}$$

which implies that V_3 is ID w.r.t V_1 and V_1 is ID w.r.t V_3. □

E Related Work

Disentangled representation learning aims to obtain approximations $\widehat{\mathbf{V}} = \{\widehat{V}_1, \dots, \widehat{V}_d\}$ that separate the distinct, informative generative factors of variations [5] from the observations of $\mathbf{X}$ and inductive bias of $\mathcal{M}$. In other words, the learning goal is an unmixing function $\widehat{f}_X^{-1}$ that maps from $\mathbf{X}$ to $\widehat{\mathbf{V}}$ (namely $\widehat{\mathbf{V}} = \widehat{f}_X^{-1}(\mathbf{X})$), where $\widehat{V}_i$ is some transformation of $\mathbf{W} \subseteq \mathbf{V}$. The goal of disentangled representation learning is to have $\widehat{V}_i$ be a function only of V_i, i.e. $\mathbf{W} = \{V_i\}$. This is not always possible, and different assumptions, data and relaxed versions of disentanglement may be studied to theoretically ground representation learning. The disentangled representation learning tasks are studied with various assumptions and input. In the following, we discuss related tasks and identifiability results in context of this paper. We also present a few case studies on the nuances between Markovian and non-Markovian ASCM setting.

First, we review the main goal of identifiability in all prior works. It is what is known as scaling identifiability. A special case of our ID definition in Def. 2.3.

Definition 6.4 (Scaling indeterminancy)**.** Consider a collection of ASCM $\boldsymbol{\mathcal{M}}$ 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 $\widehat{\boldsymbol{\mathcal{M}}}$ 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 $\mathbb{R}$. □

E.1 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, 49]. 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.

E.2 Comparisons with other identifiability criterion

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 $\boldsymbol{\mathcal{M}}$ 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. □

Definition 6.5 (Identifiability up to ancestral mixtures [21])**.** 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 ancestral mixtures if for every $\widehat{\boldsymbol{\mathcal{M}}}$ 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 $\{h_1, \dots, h_d\}$ 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.3 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: $\boldsymbol{\Psi} = \langle \sigma_{\{\}}, do(\{V_1\}), do(\{V_2\}) \rangle$. 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:

$$\mathcal{F}^{(1)} = \begin{cases} V_1^{(1)} \leftarrow -U_{1,2}^{(1)} \\ V_2^{(1)} \leftarrow 0.5U_{1,2}^{(1)} + 1.5U_Y \\ \quad X_1 \leftarrow 1/3V_1^{(1)} + 2/3V_2^{(1)}, \\ \quad X_2 \leftarrow 2/3V_1^{(1)} - 2/3V_2^{(1)} \end{cases}$$

$$U_{1,2}^{(1)} \sim \mathcal{N}(0,3), U_{V_2}^{(1)} \sim \mathcal{N}(0,1)$$

$$\sigma_{V_1} = P(\tilde{U_{V_1}}^{(1)}), \tilde{U}_{V_1}^{(1)} \sim \mathcal{N}(0,6)$$

$$\sigma_{V_2} = P(\tilde{U_{V_2}}^{(1)}), \tilde{U}_{V_2}^{(1)} \sim \mathcal{N}(0,7)$$

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. □

E.4 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 $\mathcal{P}_{V_i} = \{P^{(a_0)}, P^{(a_1)}, \dots, P^{(a_L)}\}$ 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 26. 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

$$p_{V_i}^{(2)}(v_i) - p_{V_i}^{(1)}(v_i) = p_{V_i}^{(2)}(\widehat{v}_i) - p_{V_i}^{(1)}(\widehat{v}_i) \tag{103}$$

Taking partial w.r.t. V_j, we have

$$0 = \frac{p_{V_i}^{(2)}(\widehat{v}_i) - p_{V_i}^{(1)}(\widehat{v}_i)}{\widehat{v}_i} \frac{\widehat{v}_i}{v_j} \tag{104}$$

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 16, V_1 and V_2 are disentangled from each other without leveraging two hard interventions.

E.5 Case study on disentangling variables in a Markovian setting

This next example works out the algebraic derivations for analyzing Fig. 4(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 27 (Algebraic derivation of disentanglement in a simple 3-node chain graph)**.** Given the graph shown in Figure 4(a), we can factorize the joint observational distribution of the latent variables

$$P(\mathbf{V}) = P(V_3|V_2)P(V_1|V_2)P(V_2) \tag{105}$$

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.

$$P(\mathbf{V}) = P(\phi_{V_3}(\mathbf{V})|\phi_{V_2}(\mathbf{V}))P(\phi_{V_1}(\mathbf{V})|\phi_{V_2}(\mathbf{V})P(\phi_{V_2}(\mathbf{V}))|detJ_\phi| \tag{106}$$

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 $(V_i \perp\!\!\!\perp V_k|V_j)_{G_{V\overline{V_3}}}$. 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:

$$
\log p(V_3|V_2; \sigma_{V_3^{(u)}}) - \log p(V_3|V_2) = \log p(\phi_3|\phi_2; \sigma_{3^{(u)}}) - \log p(\phi_3|\phi_2) \tag{107}
$$

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. □

E.6 Comparing different identifiability results

In this section, we explicitly compare and discuss our work compared to a non-exhaustive list of related disentangled learning in the setting of causally related latent components. Different from previous literature, we do not make common assumptions such as (1) each intervention is applied to a single node [53]; (2) idle interventions (observational distribution) are present within each domain [21, 22]; (3) *exactly* one intervention is applied per node [13]; (4) *at least* one intervention is applied per node [21, 53].

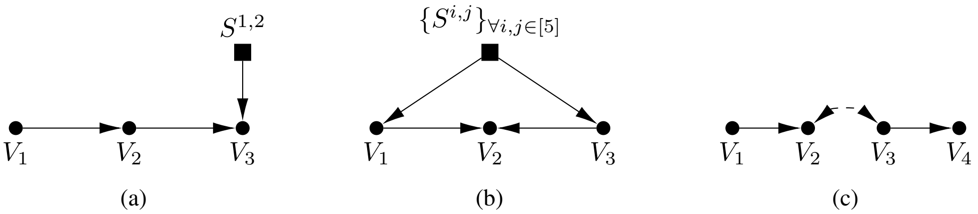


Figure S5: **Reproduced Fig. 4 for convenience.**

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 (Ex. 16 cont.)**.** This example continues off of Ex. 16. 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 LSG does not contain bidirected edges, our results can also guarantee identifiability in this setting.

Example 28 ([21] approach)**.** Given the graph shown in Figure 4(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. 7. 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. 16.

Example (Ex. 16 cont.)**.** This example continues off of Ex. 16. 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.

Linear ICA Linear ICA has been extensively studied over decades, and is applied in magnetic resonance imaging (MRI) [75], astronomy [76], image processing [77], finance [78] and document analysis [79]. In linear ICA settings, the generative factors are assumed to be independent of each other and the mixture function $f_{\mathbf{X})}$ is considered to be an invertible matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. Formally, the mechanism $\mathcal{F}$ and the distribution $P(\mathbf{U})$ of the true ASCM $\mathcal{M}^*$ are written as:

$$\begin{cases} V_j \leftarrow f_j(U_j), \forall j \in [d] \\ \mathbf{X} \leftarrow \mathbf{AV} \\ U_i \perp U_j, \forall i, j \in [d] \end{cases} \tag{108}$$

Notice that $\mathbf{X}$ is d dimensional variable here and $X_i \leftarrow \sum_{j=1}^{d} a_{ij} V_j = \mathbf{a}_i \mathbf{V}, \forall i \in [d]$. Given the observational distribution $P(\mathbf{X})$, the goal of linear tasks is to learn $\widehat{\mathbf{A}}$ such that $\widehat{V}_j$ is a scaling of a true underlying generative factors V_i, where $\widehat{\mathbf{V}} = \widehat{\mathbf{A}}^{-1}\mathbf{X}$. The scaling and permutation identifiability is defined as follows to denote the achievability of linear ICA tasks.

Definition 6.8 (Scaling and Permutation Identifiability)**.** The representation $\widehat{V}$ is said to be identifiable up to scaling and permutation $\mathbf{V}^{(2)} = \mathbf{CPV}^{(1)}$ if for every pair of ASCM $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ such that
(1) $P^{\mathcal{M}^{(1)}}(\mathbf{X}) = P^{\mathcal{M}^{(2)}}(\mathbf{X})$, $P^{\mathcal{M}^{(1)}}(\mathbf{X}; \sigma_{v_k}) = P^{\mathcal{M}^{(2)}}(\mathbf{X}; \sigma_{v_k})$;
(2) $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ are constrained by the modeling process in Eq. 108,
where $\mathbf{C} = \mathrm{diag}(c_1, \ldots, c_d)$ is a scaling diagonal matrix and $\mathbf{P}$ is a permutation matrix. □

Def. 6.8 says that if every pair model $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ in linear ICA settings match the observational distributions, the generative variables can be transformed by permutation and scaling. This implies once one finds a proxy ASCM $\mathcal{M}$ that matches $P(\mathbf{X})$, $\widehat{V}$ is guaranteed to be a scale and permutation representation of the true generative variable if the identifiability is achieved. The next example illustrates ASCMs in linear ICA settings and Def. 6.8.

Example 29 (ICA Identifiability Is Not Achieved)**.** We consider the three augmented generative processes $\mathcal{M}^*$, $\mathcal{M}^{(1)}$ and $\mathcal{M}^{(2)}$ with linear ICA constraints.

$$\begin{cases} V_1 \leftarrow U_1, V_2 \leftarrow U_2 \\ X_1 \leftarrow V_1, X_2 \leftarrow V_2 \end{cases}$$
$$U_1, U_2 \sim \mathcal{N}(0, [1, 0; 0, 1])$$
$$\mathcal{M}^*$$

$$\begin{cases} V_1^{(1)} \leftarrow U_1, V_2^{(1)} \leftarrow U_2 \\ X_1 \leftarrow 2V_1^{(1)}, X_2 \leftarrow 0.5V_2^{(1)} \end{cases}$$
$$U_1, U_2 \sim \mathcal{N}(0, [1/4, 0; 0, 4])$$
$$\mathcal{M}^{(1)}$$

$$\begin{cases} V_1^{(2)} \leftarrow U_1, V_2^{(2)} \leftarrow U_2 \\ X_1 \leftarrow \frac{\sqrt{2}}{2} V_1^{(2)} + \frac{\sqrt{2}}{2} V_2^{(2)} \\ X_2 \leftarrow \frac{\sqrt{2}}{2} V_1^{(2)} - \frac{\sqrt{2}}{2} V_2^{(2)} \end{cases}$$
$$U_1, U_2 \sim \mathcal{N}(0, [1, 0; 0, 1])$$
$$\mathcal{M}^{(2)}$$

It is verifiable that $X_1, X_2 \sim \mathcal{N}(1, 0; 0, 1)$ induced by all three models. The latent generative variables in $\mathcal{M}^{(1)}$ are scaled and permuted representations of the true factors $\mathcal{M}^*$, namely $V_1^{(1)} = 2V_2^{(2)}$ and $V_2^{(1)} = 0.5V_2^{(1)}$. In other words, $V^{(1)}$ and $V^{(2)}$ distinctly represents V_2 and V_1 respectively. However, the representations $V_1^{(2)}$ and $V_2^{(2)}$ in $\mathcal{M}^{(2)}$ are mixture of true generative factors V_1 and V_2, i.e.,

$$\begin{aligned} V_1^{(2)} &= \frac{2}{2} V_1 + \frac{2}{2} V_2 \\ V_2^{(2)} &= \frac{2}{2} V_1 - \frac{2}{2} V_2 \end{aligned} \tag{109}$$

which implies this is not a scaling and permutation transformation. Thus, $\mathcal{M}^{(2)}$ demonstrates that the scaling and permutation identifiability is not achieved in this setting. □

The above example shows a famous result of linear ICA: the representations are not identifiable if generative factors follow a multi-gaussian distribution. This result comes from the symmetricity of gaussian distributions: any white gaussian variables are still white gaussian after an orthogonal transformation. However, orthogonal transformations are not guaranteed to be a scaling or permutation thus a proxy model may have generative factors that are mixtures of the true $\mathbf{V}$ ($\mathbf{V}^{(2)}$ in Example 29). Further, the identifiability result can be concluded as follows with the non-Gaussian assumption.

Nonlinear ICA [7, 10] Compared to linear ICA, nonlinear ICA assumes the mixing function is a nonlinear bijective function (i.e. invertible and differentiable).

In linear ICA settings, the generative factors are assumed to be independent of each other and the mixture function $f_{\mathbf{X})}$ is considered to be an invertible matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$. Formally, the mechanism $\mathcal{F}$ and the distribution $P(\mathbf{U})$ of the true ASCM $\mathcal{M}^*$ are written as:

$$\begin{cases} V_j \leftarrow f_j(U_j), \forall j \in [d] \\ \mathbf{X} \leftarrow \mathbf{f}_X(\mathbf{V}) \\ U_i \perp U_j, \forall i, j \in [d] \end{cases} \tag{110}$$

The traditional approaches for proving identifiability from [7, 8, 10] has the following settings:

- (Assumptions) A parametric exponential family is assumed in [7]. In addition, the causal assumptions of the latent variables is fully disconnected graph, where all variables are mutually independent. Our work assumes a nonparametric mixing model, and only requires the mixing function to be a bijection. In addition, we allow a non-Markovian causal model among the latent variables, which is the first to our knowledge to analyze identifiability in this general setting.
- (Data) Nonlinear ICA assumes that $2d + 1$ number of distributions with mechanism changes of the latent variables such that a version of the Assump. 7 holds. One instantiation of this in real-world data is time-series with non-stationary changes. Our work leverages arbitrary combinations of interventional data arising from multiple domains, and also does not necessarily require observational data.
- (Output) The focus of nonlinear ICA was typically on achieving disentanglement of latent variables up to scaling indeterminancy (Def. 6.4). Our work approaches the goal of identifiability from a more general setting according to Def. 2.3.

Interventional causal representation learning [50] Another potentially promising approach to improving identifiability results lies in assuming a parametric form to the mixing function. [50] considers the setting of having a mixing function that is a composition of polynomial functions (i.e. a polynomial decoder).

Thus, [50] is able to achieve identifiability of latent variables up to an affine transformation:

$$\hat{V} = \mathbf{A}\mathbf{V} + c$$

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $c \in \mathbb{R}^d$ make up an invertible affine transformation of the true latent variables $\mathbf{V}$. In our work, we consider a nonparametric form of the mixing function. However, future work could consider relaxing this assumption in the direction of a parametric mixing function with polynomial functions.

F Experimental Results

F.1 Discussion of Results

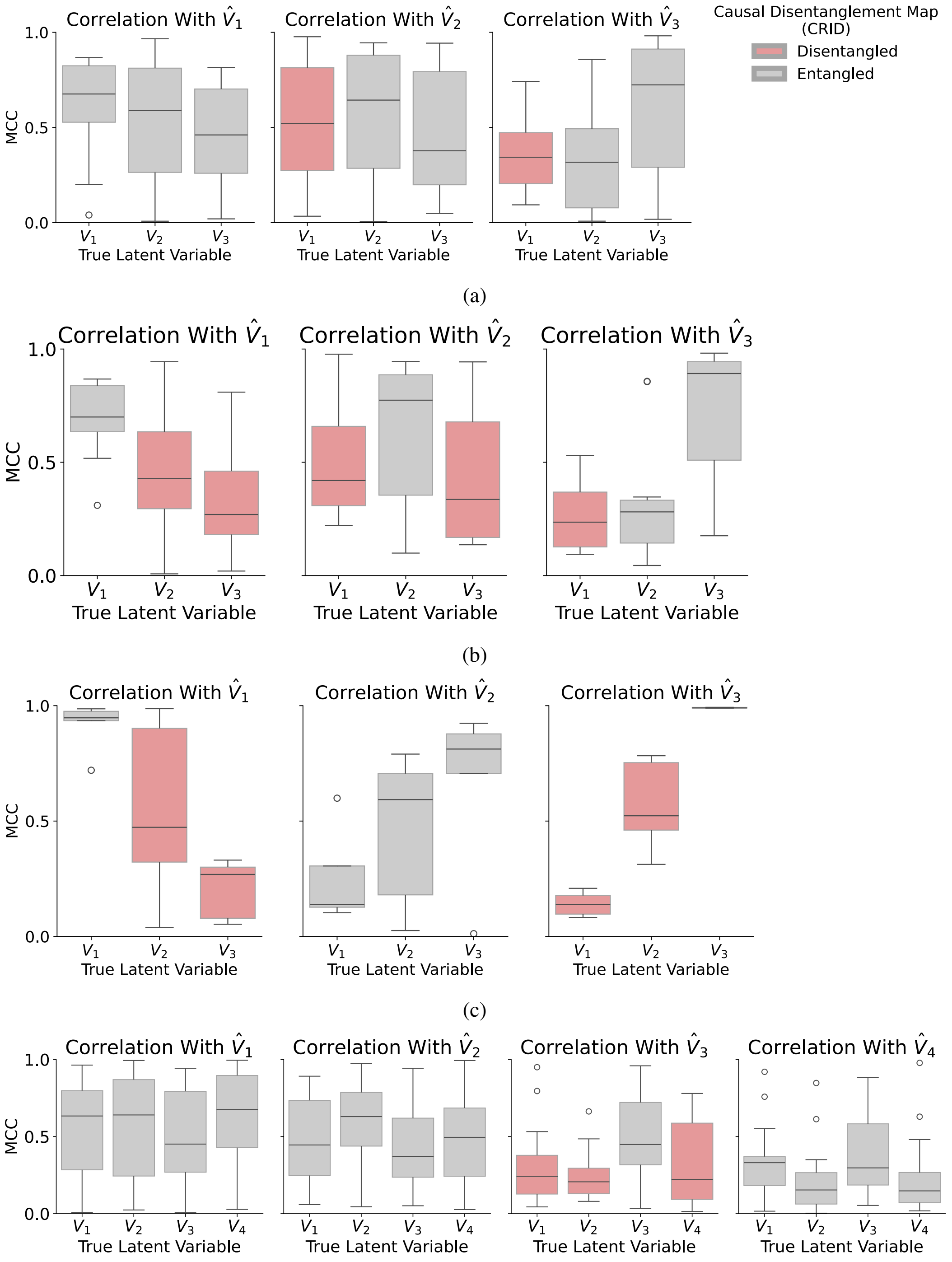
In Fig. S7, we show the MCC values for each learned latent representation $\hat{\mathbf{V}}$ and the corresponding ground-truth latents $\mathbf{V}$ for the three different LSDs shown in Fig. 4. Based on the causal disentanglement map (CDM) output from the CRID algorithm, the disentangled variables are shown in red, while the entangled variables are shown in gray.

In Fig. S7(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. 4(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. S7(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. S7(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 7) 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. S7(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).



(a)

(b)

(c)

(d)

Figure S7: 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. 4: 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 [80], epidemiology [81], economics [35] and neuroscience [36].

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 [82, 83], 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 [30, 84–86]. 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 [25]. 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 $\hat{\mathbf{V}}$ (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, 25, 53, 87–89]. 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 [82, 90–94], 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 $\mathbf{I}^{(1)}$,

$$\Delta\mathbf{Q}[\mathbf{I}^{(2)}, \mathbf{I}^{(1)}, \mathbf{T}] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(1)}, \mathbf{T}] = \{V_1, V_2\}. \tag{111}$$

Comparing $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(3)}$ with the idle intervention $\mathbf{I}^{(2)}$,

$$\Delta\mathbf{Q}[\mathbf{I}^{(1)}, \mathbf{I}^{(2)}, \mathbf{T}] = \Delta\mathbf{Q}[\mathbf{I}^{(3)}, \mathbf{I}^{(2)}, \mathbf{T}] = \{V_2\}. \tag{112}$$

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 [61] 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. [61] 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 → 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 [60]. Naturally, the issue of non-identifiability is much more acute in this setting, due to the existence of unobserved confounding.