Unifying Causal Representation Learning with the Invariance Principle

Dingling Yao **Dario Rancati**

Riccardo Cadei **Marco Fumero** **Francesco Locatello**

Institute of Science and Technology Austria

Abstract

Causal representation learning aims at recovering latent causal variables from high-dimensional observations to solve causal downstream tasks, such as predicting the effect of new interventions or more robust classification. A plethora of methods have been developed, each tackling carefully crafted problem settings that lead to different types of identifiability. The folklore is that these different settings are important, as they are often linked to different rungs of Pearl's causal hierarchy, although not all neatly fit. Our main contribution is to show that *many existing causal representation learning approaches methodologically align the representation to known data symmetries*. Identification of the variables is guided by equivalence classes across different "data pockets" that are not necessarily causal. This result suggests important implications, allowing us to unify many existing approaches in a single method that can mix and match different assumptions, including non-causal ones, based on the invariances relevant to our application. It also significantly benefits applicability, which we demonstrate by improving treatment effect estimation on real-world high-dimensional ecological data. Overall, this paper clarifies the role of causality assumptions in the discovery of causal variables and shifts the focus to preserving data symmetries.

1 Introduction

Causal representation learning [1] posits that many real-world high-dimensional perceptual data can be described through a simplified latent structure specified by a few interpretable low-dimensional causally-related variables. Discovering hidden causal structures from data has been a long-standing goal across many scientific disciplines, spanning neuroscience [2, 3], communication theory [4, 5], economics [6] and social science [7]. From the machine learning perspective, algorithms and models integrated with causal structure are often proven to be more robust at distribution shift [8–10], providing better out-of-distribution generalization results and reliable agent planning [11–13]. The general goal of causal representation learning approaches is formulated as to provably identify ground-truth latent causal variables and their causal relations (up to certain ambiguities). Many existing approaches in causal representation learning carefully formulate their problem settings to guarantee identifiability and justify the assumptions within the framework of Pearl's causal hierarchy, such as "observational, interventional, or counterfactual CRL" [14–19]. However, some causal representation learning works may not perfectly fit within this causal language framework; for instance, the problem setting of temporal CRL works [20–23] does not always align straightforwardly with existing categories. They often assume that an individual trajectory is "intervened" upon, but this is not an intervention in the traditional sense, as noise variables are not resampled. It is also not a counterfactual, as the value of non-intervened variables can change due to default dynamics. Similarly, domain generalization [24, 25, 8] and certain multi-task learning approaches [26, 11] are sometimes framed as informally related to causal representation learning. However, the precise relation to

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

causality is not always clearly articulated. This has resulted in a variety of methods and findings, some of which rely on assumptions that might be too narrowly tailored for practical, real-world applications. For example, Cadei et al. [27] collected a data set for estimating treatment effects from high-dimensional observations in real-world ecology experiments. Despite the clear causal focus of the benchmark, they note that even having access to multiple views and being able to perform some interventions, neither existing multi-view nor interventional causal representation learning methods are directly applicable due to mismatching assumptions.

This paper contributes a unified rephrasing of many existing nonparametric CRL works through the lens of invariance [28–30]. We observe that *many existing causal representation approaches share methodological similarities, particularly in aligning the representation with known data symmetries*, while differing primarily in how the invariance principle is invoked. This invariance principle is usually formulated *implicitly* in the assumed data-generating process. Instead, we make this explicit and show that latent causal variable identification broadly originates from multiple data pockets with certain underlying equivalence relations. These are not necessarily causal and (with minor exceptions) have to be known *apriori*. Unifying causal representation learning approaches using the invariance principle brings several potential benefits: First, it helps clarify the alignment between seemingly different categories of CRL methods, contributing to a more coherent and accessible framework for understanding causal representation learning. This perspective may also allow for the integration of multiple invariance relations in latent variable identification, which could improve the flexibility of these methods in certain practical settings. Additionally, our theory underscores a gap in the necessary causal assumptions for graph learning, which is essential for generalizing to unseen interventions and helps distinguish it from the problem of identifying causal variables. These invariances can be expressed in causal terms, such as interventions, but do not always need to be. Last but not least, this formulation of invariance relation links causal representation learning to many existing representation learning areas outside of causality, including invariant training [30, 8], domain adaptation [24, 25], and geometric deep learning [31–33].

We highlight our contributions as follows:

- We propose a unified rephrasing for existing nonparametric causal representation learning approaches leveraging the invariance principles and proving latent variable identifiability in this general setting (§ 3). We show that 30 existing identification results can be seen as special cases directly implied by our framework (Tab. 3). This approach also enables us to derive new results, including latent variable identifiability from one imperfect intervention per node in the nonparametric setting (Cor. D.1).
- In addition to employing different methods, many works in the CRL literature use varying definitions of "identifiability." We formalize these definitions at different levels of granularity, highlight their connections, and demonstrate how various definitions can be addressed within our framework by leveraging different invariance principles (App. C.1).
- Upon the identifiability of the latent variables, we discuss the necessary causal assumptions for graph identification and the possibility of partial graph identification using the language of causal consistency. With this, we draw a distinction between the causal assumptions necessary for graph discovery and those that may not be required for variable discovery (App. C.2).
- Our framework is broadly applicable across a range of settings. We observe improved results on real-world experimental ecology data using the causal inference benchmark from high-dimensional observations provided by Cadei et al. [27] (§ 4.1). Additionally, we present a synthetic ablation to demonstrate that existing methods, which assume access to interventions, actually only require a form of distributional invariance to identify variables. This invariance does not necessarily need to correspond to a valid causal intervention (§ 4.2).

2 Problem Setting

Notation. $[N]$ is used as a shorthand for $\{1, \dots, N\}$. We use bold lower-case $\mathbf{z}$ for random vectors and normal lower-case z for their realizations. A vector $\mathbf{z}$ can be indexed either by a single index $i \in [\dim(\mathbf{z})]$ via $\mathbf{z}_i$ or a index subset $A \subseteq [\dim(\mathbf{z})]$ with $\mathbf{z}_A := \{\mathbf{z}_i : i \in A\}$. $P_{\mathbf{z}}$ denotes the probability distribution of the random vector $\mathbf{z}$ and $p_{\mathbf{z}}(z)$ denotes the associated probability density function. By default, a "measurable" function is *measurable* w.r.t. the Borel sigma algebras and defined w.r.t. the Lebesgue measure. A more comprehensive summary of notations is provided in App. A.

This section defines our problem setting using standard CRL concepts and assumptions (Formal definitions are deferred to App. B). While prior works in CRL typically categorize their settings using established causal language (such as "counterfactual," "interventional," or "observational"), our approach introduces a more general invariance principle that aims to unify diverse problem settings. We introduce the following concepts as mathematical tools to describe our data generating process.

Definition 2.1 (Invariance property)**.** Let $A \subseteq [N]$ be an index subset of the Euclidean space $\mathbb{R}^N$ and let $\sim_\iota$ be an equivalence relationship on $\mathbb{R}^{|A|}$, with A of known dimension. Let $\mathcal{M} := \mathbb{R}^{|A|} / \sim_\iota$ be the quotient of $\mathbb{R}^{|A|}$ under this equivalence relationship; $\mathcal{M}$ is a topological space equipped with the quotient topology. Let $\iota : \mathbb{R}^{|A|} \to \mathcal{M}$ be the projection onto the quotient induced by the equivalence relationship $\sim_\iota$. This projection ι is termed the *invariant property* of this equivalence relation. Two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{|A|}$ are invariant under ι if and only if they belong to the same $\sim_\iota$ equivalence class, i.e.:

$$\iota(\mathbf{a}) = \iota(\mathbf{b}) \Leftrightarrow \mathbf{a} \sim_\iota \mathbf{b}.$$

Extending this definition to the whole latent space $\mathbb{R}^N$, a pair of latents $\mathbf{z}, \tilde{\mathbf{z}} \in \mathbb{R}^N$ are ***non-trivially*** *invariant on a subset* $A \subseteq [N]$ *under the property* ι only if

(i) the invariance property ι holds on the indices $A \subseteq [N]$ in the sense that $\iota(\mathbf{z}_A) = \iota(\tilde{\mathbf{z}}_A)$;

(ii) for any smooth function $h_1, h_2 : \mathbb{R}^N \to \mathbb{R}^{|A|}$, the invariance property between $\mathbf{z}, \tilde{\mathbf{z}}$ *breaks* under the h_1, h_2 transformations if h_1 or h_2 *directly* depends on some other component $\mathbf{z}_q$ with $q \in [N] \setminus A$. Taking h_1 and $\mathbf{z}$ as an example, we have:

$$\exists q \in [N] \setminus A, \mathbf{z}^* \in \mathbb{R}^N, \quad s.t. \ \frac{\partial h_1}{\partial \mathbf{z}_q}(\mathbf{z}^*) \text{ exists and is non zero} \quad \Rightarrow \quad \iota(h_1(\mathbf{z})) \neq \iota(h_2(\tilde{\mathbf{z}}))$$

which means: given that the partial derivative of h_1 w.r.t. some latent variable $\mathbf{z}_q \in \mathbf{z}_{[N]\setminus A}$ exists and is non-zero at some point $\mathbf{z}^* \in \mathbb{R}^N$, $h_1(\mathbf{z}), h_2(\mathbf{z})$ violates invariance principle in the sense that $\iota(h_1(\mathbf{z})) \neq \iota(h_2(\tilde{\mathbf{z}}))$.

We denote by $\mathcal{S}_\mathbf{z} := \{\mathbf{z}^1, \dots, \mathbf{z}^K\}$ the set of latent random vectors with $\mathbf{z}^k \in \mathbb{R}^N$ and write its joint distribution as $P_{\mathcal{S}_\mathbf{z}}$. The joint distribution $P_{\mathcal{S}_\mathbf{z}}$ has a probability density $p_{\mathcal{S}_\mathbf{z}}(z^1, \dots, z^K)$. Each individual random vector $\mathbf{z}^k \in \mathcal{S}_\mathbf{z}$ follows the marginal density $p_{\mathbf{z}^k}$ with the non-degenerate support $\mathcal{Z}^k \subseteq \mathbb{R}^N$, whose interion is a non-empty open set of $\mathbb{R}^N$.

Definition 2.2 (Observable of a set of latent random vectors)**.** Consider a set of random vectors $\mathcal{S}_\mathbf{z} := \{\mathbf{z}^1, \dots, \mathbf{z}^K\}$ with $\mathbf{z}^k \in \mathbb{R}^N$, the corresponding set of observables $\mathcal{S}_\mathbf{x} := \{\mathbf{x}^1, \dots, \mathbf{x}^K\}$ is generated by $\mathcal{S}_\mathbf{x} = F(\mathcal{S}_\mathbf{z})$, where the map F defines a push-forward measure $F_\#(P_{\mathcal{S}_\mathbf{z}})$ on the image of F as:

$$F_\#(P_{\mathcal{S}_\mathbf{z}})(\mathbf{x}^1, \dots, \mathbf{x}^K) = P_{\mathcal{S}_\mathbf{z}}(f_1^{-1}(\mathbf{x}^1), \dots, f_K^{-1}(\mathbf{x}^K)) \tag{2.1}$$

with the support $\mathcal{X} := \text{Im}(f) \subseteq \mathbb{R}^{K \times D}$. Note that F satisfies the diffeomorphism assumption (Asm. B.1) as each f_k is a diffeomorphism onto its image according to Asm. B.1.

In the following, we denote by $\mathfrak{I} := \{\iota_i : \mathbb{R}^{|A_i|} \to \mathcal{M}_i\}$ a finite set of invariance properties with their respective invariant subsets $A_i \subseteq [N]$ and their equivalence relationships $\sim_{\iota_i}$, each inducing as a projection onto its quotient and invariant property ι_i (Defn. 2.1). For a set of observables $\mathcal{S}_\mathbf{x} := \{\mathbf{x}^1, \dots, \mathbf{x}^K\} \in \mathcal{X}$ generated from the data generating process described in § 2, we assume:

Assumption 2.1. For each $\iota_i \in \mathfrak{I}$, there exists a ***unique known*** index subset $V_i \subseteq [K]$ with at least two elements (i.e., $|V_i| \geq 2$) s.t. $\mathbf{x}_{V_i} = F([\mathbf{z}]_{\sim_{\iota_i}})$ forms the set of observables generated from an equivalence class $[\mathbf{z}]_{\sim_{\iota_i}} := \{\tilde{\mathbf{z}} \in \mathbb{R}^N : \mathbf{z}_{A_i} \sim_{\iota_i} \tilde{\mathbf{z}}_{A_i}\}$, as given by Defn. 2.2. In particular, if $\mathfrak{I} = \{\iota\}$ consists of a single invariance property $\iota : \mathbb{R}^{|A|} \to \mathcal{M}$, we have $\mathcal{S}_\mathbf{x} = F([\mathbf{z}]_{\sim_\iota})$.

Remark: While $\mathfrak{I}$ does not need to be fully described, which observables should belong to the same equivalence class is known (denoted as $V_i \subseteq [K]$ for the invariance property $\iota_i \in \mathfrak{I}$). This is a standard assumption and is equivalent to knowing, e.g., two views are generated from partially overlapped latents [34].

Problem setting. Given a set of observables $\mathcal{S}_\mathbf{x} \in \mathcal{X}$ satisfying Asm. 2.1, we show that we can simultaneously identify multiple invariant latent blocks A_i under a set of weak assumptions. In the best case, if each individual latent component is represented as a single invariant block through individual invariance property $\iota_i \in \mathfrak{I}$, we can learn a fully disentangled representation and further identify the latent causal graph by additional technical assumptions.

3 Identifiability Theory via the Invariance Principle

High-level overview. This section presents a general theory for latent variable identification that brings together many identifiability results from existing CRL works, including multiview, interventional, temporal, and multi-task CRL. Our theory of latent variable identifiability, based on the invariance principle, consists of two key components: (1) ensuring the encoder's sufficiency, thereby obtaining an adequate representation of the original input for the desired task; (2) guaranteeing the learned representation to preserve known data symmetries as invariance properties. The sufficiency is often enforced by minimizing the reconstruction loss [35, 36, 22, 21, 20] in auto-encoder based architecture, maximizing the log likelihood in normalizing flows or maximizing entropy [37–39, 34] in self-supervised approaches. The invariance property in the learned representations is often enforced by minimizing some equivalence relation-induced regularizer [38, 34, 22, 18] or by some iterative algorithm that provably ensures the invariance property on the output [40, 41]. As a result, all invariant blocks $A_i, i \in [n_{\mathfrak{I}}]$ can be identified up to a mixing within the blocks while being disentangled from the rest. This type of identifiability is defined as *block-identifiability* [38] which we restate as follows:

Definition 3.1 (Block-identifiability [38])**.** A subset of latent variable $\mathbf{z}_A := \{\mathbf{z}_j\}_{j\in A}$ with $A \subseteq [N]$ is block-identified by an encoder $g : \mathbb{R}^D \to \mathbb{R}^N$ on the invariant subset A if the learned representation $\hat{\mathbf{z}}_{\hat{A}} := [g(\mathbf{x})]_{\hat{A}}$ with $\hat{A} \subseteq [N], |A| = |\hat{A}|$ contains all and only information about the ground truth $\mathbf{z}_A$, i.e. $\hat{\mathbf{z}}_{\hat{A}} = h(\mathbf{z}_A)$ for some diffeomorphism $h : \mathbb{R}^{|A|} \to \mathbb{R}^{|A|}$.

Definition 3.2 (Encoders)**.** The encoders $G := \{g_k : \mathcal{X}^k \to \mathcal{Z}^k\}_{k\in[K]}$ consist of smooth functions mapping from the observational support $\mathcal{X}^k$ to the corresponding latent support $\mathcal{Z}^k$ (§ 2).

Definition 3.3 (Selection [34])**.** A selection $\oslash$ operates between two vectors $a \in \{0,1\}^d, b \in \mathbb{R}^d$ s.t. $a \oslash b := [b_j : a_j = 1, j \in [d]]$.

Definition 3.4 (Invariant block selectors)**.** The invariant block selectors $\Phi := \{\phi^{(i,k)}\}_{i\in[n_{\mathfrak{I}}],k\in V_i}$ with $\phi^{(i,k)} \in \{0,1\}^N$ perform selection (Defn. 3.3) on the encoded information: for any invariance property $\iota_i \in \mathfrak{I}$, any observable $\mathbf{x}^k, k \in V_i$ we have the selected representation:

$$\phi^{(i,k)} \oslash \hat{\mathbf{z}}^k = \phi^{(i,k)} \oslash g_k(\mathbf{x}^k) = \left[[g_k(\mathbf{x}^k)]_j : \phi_j^{(i,k)} = 1, j \in [N]\right], \tag{3.1}$$

with $\left\|\phi^{(i,k)}\right\|_0 = \|\phi^{(i,k')}\|_0 = |A_i|$ for all $\iota_i \in \mathfrak{I}, k, k' \in V_i$.

Constraint 3.1 (Invariance constraint)**.** *For any invariance property $\iota_i \in \mathfrak{I}, i \in [n_{\mathfrak{I}}]$, the **selected** representations $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k), k \in V_i$ must be ι_i-invariant across the observables from the subset $V_i \subseteq [K]$:*

$$\iota_i(\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)) = \iota_i(\phi^{(i,k')} \oslash g_{k'}(\mathbf{x}^{k'})) \quad \forall i \in [n_{\mathfrak{I}}]\ \forall k, k' \in V_i \tag{3.2}$$

Constraint 3.2 (Sufficiency constraint)**.** *For any $\iota_i \in \mathfrak{I}, i \in [n_{\mathfrak{I}}]$, the **selected** representation $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k), k \in V_i$ must preserve all information of the invariant partition $\mathbf{z}_{A_i}$ that we aim to identify, i.e., $I(\mathbf{z}_{A_i}, \phi^{(i,k)} \oslash g_k(\mathbf{x}^k)) = H(\mathbf{z}_{A_i})\ \forall i \in [n_{\mathfrak{I}}], k \in V_i$.*

Theorem 3.1 (Identifiability of multiple invariant blocks)**.** *Consider a set of observables $\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K\} \in \mathcal{X}$ generated from § 2 satisfying Asm. 2.1. Let G, Φ be the set of smooth encoders (Defn. 3.2) and selectors (Defn. 3.4) that satisfy Constraints 3.1 and 3.2, then the invariant component $\mathbf{z}^k_{A_i}$ is block-identified (Defn. 3.1) by $\phi^{(i,k)} \oslash g_k$ for all $\iota_i \in \mathfrak{I}, k \in [K]$.*

What about the variant latents? Intuitively, the variant latents are not identifiable, as the invariance constraint (Constraint 3.1) is applied only to the selected invariant encodings, leaving the variant part without any weak supervision [42]. This result is formalized as follows:

Proposition 3.2 (General non-identifiability of variant latent variables)**.** *Consider the setup in Thm. 3.1, let $A := \bigcup_{i\in[n_{\mathfrak{I}}]} A_i$ denote the union of block-identified latent indices and $A^{\mathrm{c}} := [N] \setminus A$ the complementary set where no ι-invariance $\iota \in \mathfrak{I}$ applies, then the variant latents $\mathbf{z}_{A^{\mathrm{c}}}$ cannot be identified.*

Although variant latent variables are generally non-identifiable, they can be identified under certain conditions. The following demonstrates that variant latent variables can be identified under invertible encoders when the variant and invariant partitions are mutually independent.

Proposition 3.3 (Identifiability of variant latent under independence)**.** *Consider an optimal encoder $g \in G^*$ and optimal selector $\phi \in \Phi^*$ from Thm. 3.1 that jointly identify an invariant block $\mathbf{z}_A$ (we omit subscriptions k, i for simplicity), then $\mathbf{z}_{A^c}$ ($A^c := [N] \setminus A$) can be identified by the complementary encoding partition $(1 - \phi) \oslash g$ only if*

(i) g is invertible in the sense that $I(\mathbf{x}, g(\mathbf{x})) = H(\mathbf{x})$;

(ii) $\mathbf{z}_{A^c}$ is independent on $\mathbf{z}_A$.

4 Experiments

This section demonstrates the real-world applicability of causal representation learning under the invariance principle, evidenced by superior treatment effect estimation performance on the high-dimensional causal inference benchmark [27] using a regularizer for the domain generalization literature that utilizes the invariance principle [25] (§ 4.1). Additionally, we provide ablation studies on existing interventional causal representation learning methods [43, 15, 14], showcasing that non-trivial distributional invariance is needed for latent variable identification. This distributional invariance could, but does not have to, arise from a valid intervention in the sense of causality (§ 4.2).

4.1 Case Study: ISTAnt

This experiment focuses on ISTAnt [27], a recent real-world ecological benchmark designed for treatment effect estimation. ISTAnt consists of video recordings of ants triplets with occasional grooming behavior. The goal is to extract a per-frame representation for supervised behavior classification (grooming or not) to estimate the Average Treatment Effect of an intervention (exposure to a chemical substance). Further details about the problem setting are provided in App. F.1.

Experiment settings. Different videos in ISTAnt are considered different *experiments* as the experiment settings and treatments vary. We consider hard annotation sampling criteria (more non-annotated than annotated) for both experiments (videos) and positions, as described by [27]. For the training, we adopt a domain generalization objective that utilizes the invariance principle [25], which is restated as follows:

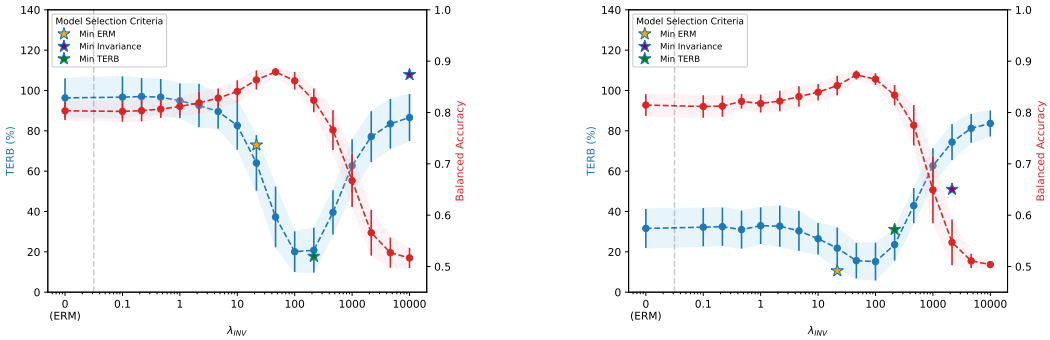
$$\mathcal{R}_{\text{V-REx}}(\mathbf{w} \circ g) = \underbrace{\lambda_{\text{INV}} \operatorname{Var}(\{\mathcal{R}_1(\mathbf{w} \circ g), \ldots, \mathcal{R}_K(\mathbf{w} \circ g)\})}_{\text{invariance}} + \underbrace{\textstyle\sum_{k \in [K]} \mathcal{R}_k(\mathbf{w} \circ g)}_{\text{sufficiency}}, \tag{4.1}$$

we provide a detailed derivation in (f) showing the invariance term above is indeed enforcing risk invariance. We vary the strength of the invariant component in eq. (4.1) by setting the regularization multiplier λ_{INV} from 0 (ERM) to 10 000. We repeat 20 independent runs for each λ_{INV} to estimate the statistical error. All other implementational details follow Cadei et al. [27]. We evaluate the performance with both *balanced accuracy* and *Treatment Effect Relative Bias* (TERB). TERB is defined by Cadei et al. [27] as the ratio between the bias in the predictions across treatment groups and the true average treatment effect estimated with ground-truth annotations over the whole trial.

Results. Fig. 1 depicts the model performance regarding varying invariance regularization strength λ_{INV}. As expected, the balanced accuracy initially increases with the λ_{INV}, as adequate invariance enforces identifying task-related latents, thus benefiting the prediction problem. At a later point, the performance decreases because the sufficiency component is not correctly balanced with the invariance. Similarly, the TERB improves positively, weighting the invariance component until a certain threshold. On average, with $\lambda_{\text{INV}} = 100$ the TERB decreases to 20% (from 100% using ERM) with experiment subsampling. In agreement with [27], a naive estimate of the TEB on a small validation set is a reasonable (albeit not perfect) model selection criterion. Although it performs slightly worse than model selection based on ERM loss in the position sampling case, it shows more reliability overall. This experiment underscores the advantages of flexibly enforcing known invariances in the data, corroborating our identifiability theory (§ 3).

4.2 Synthetic Ablation with "Ninterventions"

This subsection presents identifiability results under controversial (non-causal) conditions using simulated data. We consider a simple graph of three causal variables as $\mathbf{z}_1 \to \mathbf{z}_2 \to \mathbf{z}_3$. The corresponding joint density has the form of $p_{\mathbf{z}}(z_1, z_2, z_3) = p(z_3 \mid z_2)p(z_2 \mid z_1)p(z_1)$. This experiment aims to demonstrate that existing methods for interventional CRL rely primarily on distributional invariance, regardless of whether this invariance arises from a well-defined intervention or some other arbitrary



(a) Experiment Sampling

(b) Position Sampling

Figure 1: TERB and Balanced Accuracy with standard deviation over 20 different seeds varying the invariance weight λ_{INV} of V-REx [25] on ISTAnt dataset [27]. Stars represent the selected best models based on a small but heterogeneous validation set.

transformation. To illustrate this, we introduce the concept of a "nintervention," which has a similar distributional effect to a regular intervention, maintaining certain conditionals invariant while altering others, but without a causal interpretation.

Definition 4.1 (Nintervention)**.** We define a "*nintervention*" on a causal conditional as the process of changing its distribution but cutting all incoming and outgoing edges. Child nodes condition on the old, pre-intervention, random variable. Formally, we consider the latent SCM as defined in Defn. B.1, an *nintervention* on a node $j \in [N]$ gives rise to the following conditional factorization

$$\tilde{p}_{\mathbf{z}}(z) = \tilde{p}(z_j) \prod_{i \in [N] \setminus \{j\}} p(z_i \mid z^{\text{old}}_{\text{pa}(i)})$$

Note that the marginal distribution of all non-nintervened nodes $P_{\mathbf{z}_{[N]\setminus j}}$ remain invariant after nintervention. In previous example, we perform a nintervention by replacing the conditional density $p(z_2 \mid z_1)$ using a sufficiently different marginal distribution $p(\tilde{z}_2)$ that satisfies Defn. 2.1 (ii), which gives rise to the following new factorization $\tilde{p}_{\mathbf{z}}(z_1, z_2, z_3) = p(z_3 \mid z_2^{\text{old}})\tilde{p}(z_2)p(z_1)$. Note that $\mathbf{z}_3$ conditions on the random variable $\mathbf{z}_2$ before nintervention, whose realization is denoted as z_2^{old}. Differing from a causal *intervention*, we cut both the incoming and outgoing links of $\mathbf{z}_2$ and keep the marginal distribution of $\mathbf{z}_3$ the same. Clearly, this is a non-sensical intervention from the causal perspective because we eliminate the causal effect from $\mathbf{z}_2$ to its descendants.

Experiment settings. As a proof of concept, we choose a linear Gaussian additive noise model and a nonlinear mixing function implemented as a 3-layer invertible MLP. We average the results over three independently sampled *ninterventional* densities $\tilde{p}(z_2)$ while guaranteeing all *ninterventional* distributions satisfy Defn. 2.1 (ii). As the marginal distribution of both $\mathbf{z}_1, \mathbf{z}_3$ remains the same after a *nintervention*, we expect $\mathbf{z}_1, \mathbf{z}_3$ to be block-identified (Defn. 3.1) according to Thm. 3.1. In practice, we enforce the marginal invariance constraint (Constraint 3.1) by minimizing the MMD loss, as implemented by the interventional CRL works [18, 44] and train an auto-encoder for a sufficient representation (Constraint 3.2). Further details are included in App. F.

Results. To validate block-identifiability, we perform Kernel-Ridge Regression between the estimated block $[\hat{\mathbf{z}}1, \hat{\mathbf{z}}3]$ and the ground truth latents $\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$. Both $\mathbf{z}_1$ and $\mathbf{z}_3$ are block-identified with high R^2 scores of 0.863 ± 0.031 and 0.872 ± 0.035. In contrast, $\mathbf{z}_2$ is not identified, with a low R^2 of 0.065 ± 0.017, indicating identification is driven by the underlying distributional invariance.

5 Conclusions

In this paper, we take a closer look at the wide range of causal representation learning methods. Interestingly, we find many CRL approaches share methodological similarities in aligning the representation to known data symmetries. We identified two components involved in identifiability results: preserving information of the data and a set of known invariances (§ 3). Our results help clarify the role of causal assumptions in causal variable identification, shifting the focus from a characterization of specific assumptions for identifiability, which are not necessarily satisfied in real-world scenarios, to a general recipe that allows practitioners to specify known invariances in their problem and learn representations that align with them. Following the general recipe, we successfully exemplified the real-world applicability of CRL on ecological data, as shown in § 4.1.

Acknowledgements

We thank Jiaqi Zhang, Francesco Montagna, David Lopez-Paz, Kartik Ahuja, Thomas Kipf, Sara Magliacane, Julius von Kügelgen, Kun Zhang, and Bernhard Schölkopf for extremely helpful discussion. Riccardo Cadei was supported by a Google Research Scholar Award to Francesco Locatello. We acknowledge the Third Bellairs Workshop on Causal Representation Learning held at the Bellairs Research Institute, February 9/16, 2024, and a debate on the difference between interventions and counterfactuals in disentanglement and CRL that took place during Dhanya Sridhar's lecture, which motivated us to significantly broaden the scope of the paper. We thank Dhanya and all participants of the workshop.

References

[1] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 1, 24

[2] Ricardo Vigário, Veikko Jousmäki, Matti Hämäläinen, Riitta Hari, and Erkki Oja. Independent component analysis for identification of artifacts in magnetoencephalographic recordings. *Advances in neural information processing systems*, 10, 1997. 1

[3] Glen D Brown, Satoshi Yamada, and Terrence J Sejnowski. Independent component analysis at the neural cocktail party. *Trends in neurosciences*, 24(1):54–63, 2001. 1

[4] Tapani Ristaniemi. On the performance of blind source separation in cdma downlink. In *Proceedings of the International Workshop on Independent Component Analysis and Signal Separation (ICA'99)*, pages 437–441, 1999. 1

[5] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006. 1

[6] Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009. 1

[7] J Antonakis and R Lalive. Counterfactuals and causal inference: Methods and principles for social research. *Structural Equation Modeling*, 18(1):152–159, 2011. 1

[8] Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization, 2022. 1, 2, 23, 39

[9] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[10] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018. 1

[11] Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 22, 24, 38

[12] Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22905–22918, 2021.

[13] Núria Armengol Urpí, Marco Bagatella, Marin Vlastelica, and Georg Martius. Causal action influence aware counterfactual data augmentation. *Forty-first International Conference on Machine Learning*, 2024. 1

[14] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 5, 15, 16, 17, 20, 21, 24, 25, 36

[15] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023. 5, 16, 20, 25, 28

[16] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35:38319–38331, 2022. 17, 18, 37

[17] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *Advances in Neural Information Processing Systems*, 36, 2024. 15, 19, 20, 23, 34

[18] Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 6, 16, 17, 19, 20, 21, 23, 28, 36

[19] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024. 1, 16, 17, 19, 20, 21, 24, 25, 35

[20] Sébastien Lachapelle, Rodriguez Lopez, Pau, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022. 1, 4, 15, 16, 21, 22, 24

[21] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*, 2022. 4, 21, 22, 38

[22] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022. 4, 21, 22, 25, 37

[23] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. In *Uncertainty in Artificial Intelligence*, pages 1263–1273. PMLR, 2023. 1, 21, 22, 38

[24] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 1, 2, 22, 23, 24, 38

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021. 1, 2, 5, 6, 22, 23, 24, 30, 32, 39

[26] Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, pages 18171–18206. PMLR, 2023. 1, 22, 24, 38

[27] Riccardo Cadei, Lukas Lindorfer, Sylvia Cremer, Cordelia Schmid, and Francesco Locatello. Smoke and mirrors in causal downstream tasks. *Advances in Neural Information Processing Systems*, 37, 2024. 2, 5, 6, 29, 30

[28] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014. 2, 32

[29] Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2018.

[30] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 2, 22, 23, 39

[31] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016. 2, 32

[32] Michael M Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017. 32

[33] Michael M Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021. 2, 32

[34] Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2023. 3, 4, 16, 18, 19, 37

[35] Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 13–18 Jul 2020. 4, 18, 19, 37

[36] Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35: 15516–15528, 2022. 4, 18, 19, 37

[37] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021. 4

[38] Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34: 16451–16467, 2021. 4, 16, 18, 19, 20, 29, 36

[39] Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. 4, 18, 19, 37

[40] Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, volume 202, pages 32540–32560. PMLR, 2023. 4, 17, 19, 20, 23, 34

[41] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. *arXiv preprint arXiv:2406.05937*, 2024. 4, 16, 19, 20, 24, 25, 35, 36

[42] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 4

[43] Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 24, 25, 36

[44] Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *International Conference on Artificial Intelligence and Statistics*, pages 865–873. PMLR, 2024. 6, 19, 20, 21, 34

[45] Judea Pearl. *Causality*. Cambridge university press, 2009. 17

[46] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008. 17

[47] Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*, pages 157–164. PMLR, 2010. 17

[48] K Zhang and A Hyvärinen. On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press, 2009. 17, 18

[49] P Rubenstein, S Weichwald, S Bongers, J Mooij, D Janzing, M Grosse-Wentrup, and B Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017. 18

[50] Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685, 2019. 18

[51] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 18

[52] Yuejiang Liu, Alexandre Alahi, Chris Russell, Max Horn, Dominik Zietlow, Bernhard Schölkopf, and Francesco Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *Conference on Causal Learning and Reasoning*, pages 553–573. PMLR, 2023. 18

[53] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023. 19, 20, 21, 23, 25, 29, 35

[54] Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally disentangled representation learning. *Advances in Neural Information Processing Systems*, 35:26492–26503, 2022. 21, 22

[55] Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *International Conference on Learning Representations*, 2022. 21, 22

[56] Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024. 21, 22, 24

[57] Zijian Li, Ruichu Cai, Zhenhui Yang, Haiqin Huang, Guangyi Chen, Yifan Shen, Zhengming Chen, Xiangchen Song, Zhifeng Hao, and Kun Zhang. When and how: Learning identifiable latent states for nonstationary time series forecasting. *arXiv preprint arXiv:2402.12767*, 2024.

[58] Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Zhifeng Hao, Zhengmao Zhu, Guangyi Chen, and Kun Zhang. On the identification of temporally causal representation with instantaneous dependence. *arXiv preprint arXiv:2405.15325*, 2024. 21

[59] Thomas Dean and Keiji Kanazawa. A model for reasoning about persistance and causation. In *Computational Intelligence*, page 5, 1989. 21

[60] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002. 21

[61] David A Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. 21, 22

[62] Ilyes Khemakhem, Ricardo Monti, Diederik Kingma, and Aapo Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *Advances in Neural Information Processing Systems*, 33:12768–12778, 2020. 21, 25

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

[64] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. 22

[65] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997. 22

[66] Yu Zhang and Qiang Yang. An overview of multi-task learning. *National Science Review*, 5 (1):30–43, 2018. 22

[67] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. *Advances in Neural Information Processing Systems*, 35:2664–2678, 2022. 22

[68] Jinze Bai, Rui Men, Hao Yang, Xuancheng Ren, Kai Dang, Yichang Zhang, Xiaohuan Zhou, Peng Wang, Sinan Tan, An Yang, et al. Ofasys: A multi-modal multi-task learning system for building generalist models. *arXiv preprint arXiv:2212.04408*, 2022. 22

[69] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 22, 23

[70] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016. 22, 23

[71] Chen Fang, Ye Xu, and Daniel N Rockmore. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1657–1664, 2013. 23

[72] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 23

[73] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 23

[74] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018. 23

[75] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 23

[76] Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. *Advances in neural information processing systems*, 24, 2011. 23

[77] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4396–4415, 2022. 23

[78] Yuli Zhang, Huaiyu Wu, and Lei Cheng. Some new deformation formulas about variance and covariance. In *2012 proceedings of international conference on modelling, identification and control*, pages 987–992. IEEE, 2012. 24

[79] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022. 24

[80] Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *Forty-first International Conference on Machine Learning*, 2024. 24

[81] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15687–15701. Curran Associates, Inc., 2022. 25

[82] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *Internatinal Conference on Machine Learning*, 2024. 25

[83] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019. 25

[84] James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000. 29

[85] Jonathan M Samet, Francesca Dominici, Frank C Curriero, Ivan Coursac, and Scott L Zeger. Fine particulate air pollution and mortality in 20 us cities, 1987–1994. *New England journal of medicine*, 343(24):1742–1749, 2000.

[86] Egbert H Van Nes, Marten Scheffer, Victor Brovkin, Timothy M Lenton, Hao Ye, Ethan Deyle, and George Sugihara. Causal feedbacks in climate change. *Nature Climate Change*, 5 (5):445–448, 2015.

[87] Jakob Runge. Modern causal inference approaches to investigate biodiversity-ecosystem functioning relationships. *nature communications*, 14(1):1917, 2023. 29

[88] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 30

[89] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 30

[90] Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *Advances in Neural Information Processing Systems*, 37, 2024. 31

[91] Dominik Janzing and Sergio Hernan Garrido Mejia. A phenomenological account for causality in terms of elementary actions. *Journal of Causal Inference*, 12(1):20220076, 2024. 32

[92] Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319. 32

[93] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? *arXiv preprint arXiv:1511.07543*, 2015. 32

[94] Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. *International Conference on Learning Representations*, 2022.

[95] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

[96] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024. 32

[97] Elliot Creager, Joern-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2189–2200. PMLR, 18–24 Jul 2021. 32

[98] Md Rifat Arefin, Yan Zhang, Aristide Baratin, Francesco Locatello, Irina Rish, Dianbo Liu, and Kenji Kawaguchi. Unsupervised concept discovery mitigates spurious correlations. In *Forty-first International Conference on Machine Learning*, 2024.

[99] Mohammad Pezeshki, Diane Bouchacourt, Mark Ibrahim, Nicolas Ballas, Pascal Vincent, and David Lopez-Paz. Discovering environments with xrm. In *Forty-first International Conference on Machine Learning*, 2024. 32

[100] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018. 32

[101] Marco Fumero, Luca Cosmo, Simone Melzi, and Emanuele Rodolà. Learning disentangled representations via product manifold projection. In *International conference on machine learning*, pages 3530–3540. PMLR, 2021. 32

[102] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 32

[103] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021. 33

Appendix

Table of Contents

A Notation and Terminology **14**

B Preliminaries **15**

C Identifiability Theory **15**
C.1 On the granularity of identification . . . 15
C.2 Identifying the causal graph . . . 17

D Related Works **18**
D.1 Multiview Causal Representation Learning . . . 18
D.2 Multi-environment Causal Representation Learning . . . 19
D.3 Temporal Causal Representation Learning . . . 21
D.4 Multi-task Causal Representation Learning . . . 22
D.5 Domain Generalization . . . 22
D.6 Further Explanations for Tab. 3 . . . 23
D.7 Notable Cases Not Directly Covered by the Theory . . . 24

E Proofs **25**
E.1 Assumption Justification . . . 25
E.2 Proof for Thm. 3.1 . . . 26
E.3 Proofs for Generalization of Variant Latents . . . 27
E.4 Proofs for Granularity of Latent Variable Identification . . . 28
E.5 Proof for Cor. D.1 . . . 29

F Implementation Details **29**
F.1 Case Study: ISTAnt . . . 29
F.2 Synthetic Ablation with “Ninterventions" . . . 31

G Further Discussions and Connections to Other Fields **31**
G.1 Representational Alignment and Platonic Representation . . . 32
G.2 Environment Discovery . . . 32
G.3 Geometric Deep Learning . . . 32

A Notation and Terminology

f	Mixing function
g	Smooth encoder
$\mathcal{G}$	Ground truth causal graph
$\mathbf{x}$	Entangled observables
$\mathbf{z}$	Ground truth latent variables
D	Dimensionality of observable $\mathbf{x}$
N	Dimensionality of latents $\mathbf{z}$
A	Subset of latent indices with invariance properties ($A \subseteq [N]$)
ι	Projector which maps the latents to the space where the invariance property holds
$\sim_\iota$	The latent equivalence relation
$\mathfrak{I}$	A set of invariance properties

$\mathcal{X}$	Support of a set of observables $\mathcal{S}_{\mathbf{x}}$
$\mathcal{Z}$	Support of a set of latent vectors $\mathcal{S}_{\mathbf{z}}$
G	A set of smooth encoders
Φ	A set of selectors
TC	Transitive closure

B Preliminaries

In this subsection, we revisit the common definitions and assumptions in identifiability works from causal representation learning. We begin with the definition of a latent structural causal model:

Definition B.1 (Latent SCM [14])**.** Let $\mathbf{z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$ denote a set of causal "endogenous" variables with each $\mathbf{z}_i$ taking values in $\mathbb{R}$, and let $\mathbf{u} = \{\mathbf{u}_1, \dots, \mathbf{u}_N\}$ denotes a set of mutually independent "exogenous" random variables. The latent SCM consists of a set of structural equations

$$\{\mathbf{z}_i := m_i(\mathbf{z}_{\text{pa}(i)}), \mathbf{u}_i\}_{i=1}^N, \tag{B.1}$$

where $\mathbf{z}_{\text{pa}(i)}$ are the causal parents of $\mathbf{z}_i$ and m_i are the deterministic functions that are termed "causal mechanisms". We indicate with $P_{\mathbf{u}}$ the joint distribution of the exogenous random variables, which, due to the independence hypothesis, is the product of the probability measures of the individual variables. The associated causal diagram $\mathcal{G}$ is a directed graph with vertices $\mathbf{z}$ and edges $\mathbf{z}_i \to \mathbf{z}_j$ iff. $\mathbf{z}_i \in \mathbf{z}_{\text{pa}(j)}$; we assume the graph $\mathcal{G}$ to be acyclic.

The latent SCM induces a unique distribution $P_{\mathbf{z}}$ over the endogenous variables $\mathbf{z}$ as a pushforward of $P_{\mathbf{u}}$ via eq. (B.1). Its density $p_{\mathbf{z}}$ follows the causal Markov factorization:

$$p_{\mathbf{z}}(z) = \prod_{i=1}^{N} p_i(z_i \mid z_{\text{pa}(i)}). \tag{B.2}$$

Instead of directly observing the endogenous and exogenous variables $\mathbf{z}$ and $\mathbf{u}$, we only have access to some "entangled" measurements $\mathbf{x}$ of $\mathbf{z}$ generated through a nonlinear mixing function:

Definition B.2 (Mixing function)**.** A deterministic smooth function $f : \mathbb{R}^N \to \mathbb{R}^D$ mapping the latent vector $\mathbf{z} \in \mathbb{R}^N$ to its observable $\mathbf{x} \in \mathbb{R}^D$, where $D \geq N$ denotes the dimensionality of the observational space.

Assumption B.1 (Diffeomorphism)**.** The mixing function f is diffeomorphic onto its image, i.e. f is C^∞, f is injective and $f^{-1}|_{\mathcal{I}(f)} : \mathcal{I}(f) \to \mathbb{R}^D$ is also C^∞.

Remark: Settings with noisy observations ($\mathbf{x} = f(\mathbf{z}) + \epsilon$, $\mathbf{z} \perp \epsilon$) can be easily reduced to our denoised version by applying a standard deconvolution argument as a pre-processing step, as indicated by Lachapelle et al. [20], Buchholz et al. [17].

C Identifiability Theory

In addition to the general results for latent variable identification presented in § 3, we compare in App. C.1 different granularity of latent variable identification and show their transitions through certain assumptions on the causal model or mixing function. Afterward, App. C.2 discusses the identification level of a causal graph depending on the granularity of latent variable identification under certain structural assumptions. Proofs are deferred to App. E.

C.1 On the granularity of identification

Different levels of identification can be achieved depending on the degree of underlying invariance and data symmetry. Below, we present three standard identifiability definitions from the CRL literature, each providing a stronger identification result than block-identifiability (Defn. 3.1).

Definition C.1 (Block affine-identifiability)**.** Let $\hat{\mathbf{z}}$ be the learned representation, for a subset $A \subseteq [N]$ it satisfies that:

$$\hat{\mathbf{z}}_{\pi(A)} = D \cdot \mathbf{z}_A + \mathbf{b}, \tag{C.1}$$

where $D \in \mathbb{R}^{|A| \times |A|}$ is an invertible matrix, $\pi(A)$ denotes the index permutation of A, then $\mathbf{z}_A$ is block affine-identified by $\hat{\mathbf{z}}_{\pi(A)}$.

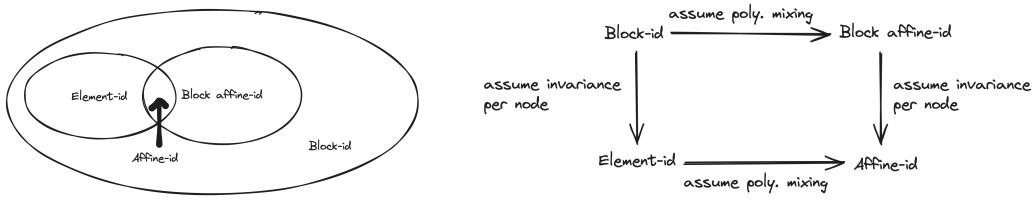


Figure 2: Relations between different identification classes (Defns. 3.1 and C.1 to C.3). Some CRL works proposed a more fine-grained classification of identifiability concepts with slightly different terminology, which we omit here for readability.

Definition C.2 (Element-identifiability)**.** The learned representation $\hat{\mathbf{z}} \in \mathbb{R}^N$ satisfies that:

$$\hat{\mathbf{z}} = \mathbf{P}_\pi \cdot h(\mathbf{z}), \tag{C.2}$$

where $\mathbf{P}_\pi \in \mathbb{R}^{N \times N}$ is a permutation matrix, $h(\mathbf{z}) := (h_1(\mathbf{z}_1), \dots h_N(\mathbf{z}_N)) \in \mathbb{R}^N$ is a is an element-wise diffeomorphism.

Definition C.3 (Affine-identifiability)**.** The learned representation $\hat{\mathbf{z}} \in \mathbb{R}^N$ satisfies that:

$$\hat{\mathbf{z}} = \Lambda \cdot \mathbf{P}_\pi \cdot \mathbf{z} + \mathbf{b}, \tag{C.3}$$

where $\mathbf{P}_\pi \in \mathbb{R}^{N \times N}$ is a permutation matrix, $\Lambda \in \mathbb{R}^{N \times N}$ is a diagonal matrix with nonzero diagonal entries.

Remark: Block affine-identifiability (Defn. C.1) is defined by Ahuja et al. [15], stating that the learned representation $\hat{\mathbf{z}}$ is related to the ground truth latents $\mathbf{z}$ through some sparse matrix with zero blocks. Defn. C.2 indicates element-wise identification of latent variables up to individual diffeomorphisms. Element-identifiability for the latent variable identification together with the graph identifiability (Defn. C.4) is defined as $\sim_{\text{CRL}}$-identifiability [14, Defn. 2.6], perfect identifiability [19, Defn. 3]. Affine identifiability (Defn. C.3) describes when the ground truth latent variables are identified up to permutation, shift, and linear scaling. In many CRL works, affine identifiability (Defn. C.3) is also termed as follows: perfect identifiability under linear transformation [41, Defn. 1], CD-equivalence [18, Defn. 1], disentanglement [20, Defn. 3].

Proposition C.1 (Granularity of identification)**.** *Affine-identifiability (Defn. C.3) implies element-identifiability (Defn. C.2) and block affine-identifiability (Defn. C.1) while element-identifiability and block affine-identifiability implies block-identifiability (Defn. 3.1).*

Proposition C.2 (Transition between identification levels)**.** *The transition between different levels of latent variable identification (Fig. 2) can be summarized as follows:*

(i) Element-level identifiability (Defns. C.2 and C.3) can be obtained from block-wise identifiability (Defns. 3.1 and C.1) when each individual latent constitutes an invariant block;

(ii) Identifiability up to an affine transformation (Defns. C.1 and C.3) can be obtained from general identifiability on arbitrary diffeomorphism (Defns. 3.1 and C.2) by additionally assuming that both the ground truth mixing function and decoder are finite degree polynomials of the same degree.

Discussion. We note that the granularity of identifiability results is primarily determined by the strength of invariance and parametric assumptions (such as those on mixing functions or causal models) rather than by the specific algorithmic choice. For example, for settings that can achieve element-identifiability [14], affine-identifiability results can be obtained by additionally assuming *finite degree polynomial* mixing function (proof see App. E). Similarly, one reaches element-identifiability from block-identifiability by enforcing invariance properties on each latent component [34, Thm. 3.8] instead of having only *one* multivariate invariant block [38]. Tab. 3 provides an overview of recent identifiability results along with their corresponding invariance and parametric assumptions, illustrating the direct relationship between these assumptions and the level of identifiability they achieve.

C.2 Identifying the causal graph

In addition to latent variable identification, another goal of causal representation learning is to infer the underlying latent dependency, namely the causal graph structure. Hence, we restate the standard definition of graph identifiability in causal representation learning.

Definition C.4 (Graph-identfiability)**.** The estimated graph $\hat{\mathcal{G}}$ is isomorphic to the ground truth $\mathcal{G}$ through a bijection $h : V(\mathcal{G}) \to V(\hat{\mathcal{G}})$ in the sense that two vertices $\mathbf{z}_i, \mathbf{z}_j \in V(\mathcal{G})$ are adjacent in $\mathcal{G}$ if and only if $h(\mathbf{z}_i), h(\mathbf{z}_j) \in V(\hat{\mathcal{G}})$ are adjacent in $\hat{\mathcal{G}}$.

We remark that the "faithfulness" assumption [45, Defn. 2.4.1] is a standard assumption in the CRL literature, commonly required for graph discovery. We restate it as follows:

Assumption C.1 (Faithfulness (or Stability))**.** $P_{\mathbf{z}}$ is a faithful distribution induced by the latent SCM (Defn. B.1) in the sense that $P_{\mathbf{z}}$ contains no extraneous conditional independence; in other words, the only conditional independence relations satisfied by $P_{\mathbf{z}}$ are those given by $\{\mathbf{z}_i \perp \mathbf{z}_{\text{nd}(i)} \mid \mathbf{z}_{\text{pa}(i)}\}$ where $\mathbf{z}_{\text{nd}(i)}$ denotes the non-descends of $\mathbf{z}_i$.

As indicated by Defn. C.4, the preliminary condition of identifying the causal graph is to have an element-wise correspondence between the vertices in the ground truth graph $\mathcal{G}$ (i.e., the ground truth latents) and the vertices of the estimated graph. Therefore, the following assumes that the learned encoders G (Defn. 3.2) achieve element-identifiability (Defn. C.2), that is, for each $\mathbf{z}_i \in \mathbf{z}$, we have a differmorphism $h_i : \mathbb{R} \to \mathbb{R}$ such that $\hat{\mathbf{z}}_i = h_i(\mathbf{z}_i)$. However, additional assumptions are needed to identify the graph structure: either on the source of invariance or on the parametric form of the latent causal model.

Graph identification via interventions. Under the element-identifiability (Defn. C.2) of the latent variables $\mathbf{z}$, the causal graph structure $\mathcal{G}$ can be identified up to its isomorphism (Defn. C.4), given multi-domain data from *paired perfect* interventions per-node [14, 19]. Using data generated from *imperfect* interventions is generally insufficient to identify the direct edges in the causal graph. It can only identify the ancestral relations, i.e., up to the transitive closure of $\mathcal{G}$ [16, 18]. Unfortunately, even imposing the linear assumption on the latent SCM does not provide a solution [40]. Nevertheless, by adding sparsity assumptions on the causal graph $\mathcal{G}$ and polynomial assumption on the mixing function f, Zhang et al. [18] has shown isomorphic graph identifiability (Defn. C.4) under *imperfect* intervention per node. In general, access to the interventions is necessary for graph identification if one is uncomfortable making other parametric assumptions about the graph structure. Conveniently, in this setting, the graph identifiability is linked with that of the variables since the latter leverages the invariance induced by the intervention.

Graph identification via parametric assumptions. It is well known in causal discovery that the additive noise model [46] is identifiable under certain mild assumptions [47, 48]. In the following, we assume an additive exogenous noise in the latent SCM (Defn. B.1):

Assumption C.2 (Additive noise)**.** The endogenous variable $\mathbf{z}_i \in \mathbb{R}$ in the previously defined latent SCM (Defn. B.1) relates to the corresponding exogenous noise variable $\mathbf{u}_i \in \mathbb{R}$ through additivity. Namely, the causal mechanism (eq. (B.1)) can be rewritten as:

$$\{\mathbf{z}_i = m_i(\mathbf{z}_{\text{pa}(i)}) + \mathbf{u}_i\}. \tag{C.4}$$

As a generalization of the additive noise model, the post-nonlinear acyclic causal model [47, Sec. 2] allows extra nonlinearity on the top of the additive causal mechanism, providing additional flexibility on the latent model assumption:

Definition C.5 (Post-nonlinear acyclic causal model)**.** The following causal mechanism describes a post-nonlinear acyclic causal model:

$$\mathbf{z}_i = h_i(m_i(\mathbf{z}_{\text{pa}(i)}) + \mathbf{u}_i), \tag{C.5}$$

where $h_i : \mathbb{R} \to \mathbb{R}$ is a diffeomorphism and m_i is a non-constant function.

Assume the latent variable $\mathbf{z}_i$ is element-wise identified through a bijective mapping $h_i : \mathbb{R} \to \mathbb{R}$ for all $i \in [N]$, define the estimated causal parents $\hat{\mathbf{z}}_{\text{pa}(i)} := \{h_j(\mathbf{z}_j) : \mathbf{z}_j \in \mathbf{z}_{\text{pa}(i)}\}$, then the latent

SCM (Defn. B.1) is translated to a post-nonlinear acyclic causal model (Defn. C.5) because

$$\begin{aligned}\hat{\mathbf{z}}_i &= h_i(\mathbf{z}_i) = h_i(m_i(\mathbf{z}_{\text{pa}(i)}) + \mathbf{u}_i) \\ &= h_i(m_i(\{h_j^{-1}(\hat{\mathbf{z}}_j) : \mathbf{z}_j \in \mathbf{z}_{\text{pa}(i)}\}) + \mathbf{u}_i) \\ &= h_i(\tilde{m}_i(\hat{\mathbf{z}}_{\text{pa}(i)}) + \mathbf{u}_i),\end{aligned} \tag{C.6}$$

where

$$\tilde{m}_i(\hat{\mathbf{z}}_{\text{pa}(i)}) := m_i(\{h_j^{-1}(\hat{\mathbf{z}}_j) : \mathbf{z}_j \in \mathbf{z}_{\text{pa}(i)}\}).$$

Thus, the underlying causal graph $\mathcal{G}$ can be identified up to an isomorphism (Defn. C.4) following the approach given by Zhang and Hyvärinen [48, Sec. 4]

What happens if variables are identified in blocks? Consider the case where the latent variables cannot be identified up to element-wise diffeomorphism; instead, one can only obtain a coarse-grained version of the variables (e.g., as a mixing of a block of variables (Defn. 3.1)). Nevertheless, certain causal links between these coarse-grained block variables are of interest. These block variables and their causal relations in between form a "macro" level of the original latent SCM, which is shown to be causally consistent under mild structural assumptions [49, Thm. 11]. In particular, the macro-level model can be obtained from the micro-level model through an *exact transformation* [50, Defn. 3.4] and thus produces the same causal effect as the original micro-level model under the same type of interventions, providing useful knowledge for downstream causal analysis. More formal connections are beyond the scope of this paper. Still, we see this concept of coarse-grained identification on both causal variables and graphs as an interesting avenue for future research.

D Related Works

This section reviews related causal representation learning works and frames them as specific instances of our theory (§ 3). These works were initially categorized into various causal representation learning types (multiview, multi-domain, multi-task, and temporal CRL) based on the level of invariance in the data-generating process, leading to varying degrees of identifiability results (App. C.1). While the implementation of individual works may vary, the *methodological principle of aligning representation with known data symmetries* remain consistent, as shown in § 3. We begin with revisiting the data-generating process of each category and explain how they can be viewed as specific cases of the proposed invariance framework (§ 2). We then present individual identification algorithms from the CRL literature as particular applications of our theorems based on the implementation choices needed to satisfy the invariance and sufficiency constraints (Constraints 3.1 and 3.2). A more detailed overview of the individual works is provided in Tab. 3.

D.1 Multiview Causal Representation Learning

High-level overview. The multiview setting in causal representation learning [39, 34] considers multiple views that are *concurrently* generated by an overlapping subset of latent variables, and thus having *non-independently* distributed data. Multiview scenarios are often found in a partially observable setup. For example, multiple devices on a robot measure different modalities, jointly monitoring the environment through these real-time measurements. While each device measures a distinct subset of latent variables, these subsets probably still overlap as they are measuring the same system at the same time. In addition to partial observability, another way to obtain multiple views is to perform an "intervention/perturbation" [35, 38, 36, 16] and collect both pre-action and post-action views on the same sample. This setting is often improperly termed "counterfactual"[1] in the CRL literature, and this type of data is termed "paired data". From another perspective, the paired setting can be cast in the partial observability scenario by considering the same latent before and after an action (mathematically modeled as an intervention) as two separate latent nodes in the causal graph, as shown by von Kügelgen et al. [38, Fig. 1]. Thus, both pre-action and post-action views are partial because neither of them can observe pre-action and post-action latents simultaneously. These works assume the latents that are not affected by the action remain constant, an assumption that is relaxed in temporal CRL works. See App. D.3 for more discussion in this regard.

[1]Traditionally, counterfactual in causality refers to non-observable outcomes that are "counter to the fact" [51]. The works we refer to here represent pre- and post-actions that affect some latent variables but not all. This can be mathematically expressed as a counterfactual in an SCM but is conceptually different as both pre- and post-action outcomes are realized [52]. The "counterfactual" terminology silently implies that this is a strong assumption, but nuance is needed and it can in fact be much weaker than an intervention.

Data generating process. In the following, we introduce the data-generating process of a multiview setting in the flavor of the invariance principle as introduced in § 2. We consider a set of views $\{\mathbf{x}^k\}_{k\in[K]}$ with each view $\mathbf{x}^k \in \mathcal{X}^k$ generated from some latents $\mathbf{z}^k \in \mathcal{Z}^k$. Let $S_k \subseteq [N]$ be the index set of generating factors for the view $\mathbf{x}^k$, we define $\mathbf{z}^k_j = 0$ for all $j \in [N] \setminus S_k$ to represent the uninvolved partition of latents. Each entangled view $\mathbf{x}^k$ is generated by a view-specific mixing function $f_k : \mathcal{Z}^k \to \mathcal{X}^k$:

$$\mathbf{x}^k = f_k(\mathbf{z}^k) \quad \forall k \in [K] \tag{D.1}$$

Define the joint overlapping index set $A := \bigcap_{k\in[K]} S_k$, and assume $A \subseteq [N]$ is a non-empty interior of $[N]$. Then the value of the sharing partition $\mathbf{z}_A$ remain ***invariant*** for all observables $\{\mathbf{x}^k\}_{k\in[K]}$ on a ***sample level***. By considering the joint intersection A, we have *one single* invariance property $\iota : \mathbb{R}^{|A|} \to \mathbb{R}^{|A|}$ in the invariance set $\mathfrak{I}$; and this invariance property ι emerges as the identity map id on $\mathbb{R}^{|A|}$ in the sense that $\mathrm{id}(\mathbf{z}^k_A) = \mathrm{id}(\mathbf{z}^{k'}_A)$ and thus $\mathbf{z}^k_A \sim_\iota \mathbf{z}^{k'}_A$ for all $k, k' \in [K]$. Note that Defn. 2.1 (ii) is satisfied because any transformation h_k that involves other components $\mathbf{z}_q$ with $q \notin A$ violates the equity introduced by the identity map. For a subset of observations $V_i \subseteq [K]$ with at least two elements $|V_i| > 1$, we define the latent intersection as $A_i := \bigcap_{k\in V_i} \subseteq [N]$, then for each non-empty intersection A_i, there is a corresponding invariance property $\iota_i : \mathbb{R}^{|A_i|} \to \mathbb{R}^{|A_i|}$ which is the identity map specified on the subspace $\mathbb{R}^{|A_i|}$. By considering all these subsets $\mathcal{V} := \{V_i \subseteq [K] : |V_i| > 1, |A_i| > 0\}$, we obtain a set of invariance properties $\mathfrak{I} := \{\iota_i : \mathbb{R}^{|A_i|} \to \mathbb{R}^{|A_i|}\}$ that satisfy Asm. 2.1.

Identification algorithms. Many multiview works [38, 39, 34] employ the L_2 loss as a regularizer to enforce ***sample-level*** invariance on the invariant partition, cooperated with some sufficiency regularizer to preserve sufficient information about the observables (Constraint 3.2). Aligned with our theory (Thm. 3.1), these works have shown block-identifiability on the invariant partition of the latents across different views. Following the same principle, there are certain variations in the implementations to enforce the invariance principle, e.g. Locatello et al. [35] directly average the learned representations from paired data $g(\mathbf{x}^1), g(\mathbf{x}^2)$ on the shared coordinates before forwarding them to the decoder; Ahuja et al. [36] enforces L_2 alignment up to a learnable sparse perturbation δ. As each latent component constitutes a single invariant block in the training data, these two works *element-identifies* (Defn. C.2) the latent variables, as explained by Proposition C.2.

D.2 Multi-environment Causal Representation Learning

High-level overview. Multi-environment / interventional CRL considers data generated from multiple environments with respective environment-specific data distributions; hence, the considered data is *independently* but *non-identically distributed*. In the scope of causal representation learning, multi-environment data is often instantiated through interventions on the latent structured causal model [38, 18, 17, 40, 53, 41, 19]. Recently, Ahuja et al. [44] provides a more general identifiability statement where multi-environment data is not necessarily originated from interventions; instead, they can be individual data distributions that preserve certain symmetries such as marginal invariance or support invariance [44].

Data generating process The following presents the data generating process described in most interventional causal representation learning works. Formally, we consider a set of *non-identically* distributed data $\{P_{\mathbf{x}^k}\}_{k\in[K]}$ that are collected from multiple environments (indexed by $k \in [K]$) with a shared mixing function $f : \mathbf{x}^k = f(\mathbf{z}^k)$ (Defn. B.2) satisfying Asm. B.1 and a shared latent SCM (Defn. B.1). Let $k = 0$ denote the non-intervened environment and $\mathcal{I}_k \subseteq [N]$ denotes the set of intervened nodes in k-th environment, the latent distribution $P_{\mathbf{z}^k}$ is associated with the density

$$p_{\mathbf{z}^k}(z^k) = \prod_{j\in\mathcal{I}_k} \tilde{p}(z^k_j \mid z^k_{\mathrm{pa}(j)}) \prod_{j\in[N]\setminus\mathcal{I}_k} p(z^k_j \mid z^k_{\mathrm{pa}(j)}), \tag{D.2}$$

where we denote by p the original density and by $\tilde{p}$ the intervened density. Interventions naturally introduce various distributional invariances that can be utilized for latent variable identification: Under the intervention $\mathcal{I}_k$ in the k-th environment, we observe that both (1) the marginal distribution of $\mathbf{z}_A$ with $A := [N] \setminus \mathrm{TC}(\mathcal{I}_k)$, with TC denoting the transitive closure and (2) the score $[S(\mathbf{z}^k)]_{A'} := \nabla_{\mathbf{z}^k_{A'}} \log p_{\mathbf{z}^k}$ on the subset of latent components $A' := [N] \setminus \overline{\mathrm{pa}}(\mathcal{I}_k)$ with $\overline{\mathrm{pa}}(\mathcal{I}_k) := \{j : j \in \mathcal{I}_k \cup \mathrm{pa}(\mathcal{I}_k)\}$ remain *invariant* across the observational and the k-th interventional environment. Formally, under intervention $\mathcal{I}_k$, we have

- ***Marginal invariance***:

$$p_{\mathbf{z}^0}(z_A^0) = p_{\mathbf{z}^k}(z_A^k) \qquad A := [N] \setminus \mathrm{TC}(\mathcal{I}_k); \tag{D.3}$$

- ***Score invariance***:

$$[S(\mathbf{z}^0)]_{A'} = [S(\mathbf{z}^k)]_{A'} \qquad A' := [N] \setminus \overline{\mathrm{pa}}(\mathcal{I}_k). \tag{D.4}$$

According to our theory Thm. 3.1, we can block-identify both $\mathbf{z}_A, \mathbf{z}'_A$ using these invariance principles (eqs. (D.3) and (D.4)). Since most interventional CRL works assume at least one intervention per node [40, 18, 14, 19, 53, 17, 15], more fine-grained variable identification results, such as element-wise identification (Defn. C.2) or affine-identification (Defn. C.3), can be achieved by combining multiple invariances from these per-node interventions, as we elaborate below.

Identifiability with one intervention per node. By applying Thm. 3.1, we demonstrate that latent causal variables $\mathbf{z}$ can be identified up to element-wise diffeomorphism (Defn. C.2) under single node *imperfect* intervention per node, given the following assumption.

Assumption D.1 (Topologically ordered interventional targets)**.** Specifying Asm. 2.1 in the interventional setting, we assume there are exactly N environments $\{k_1, \dots, k_N\} \subseteq [K]$ where each node $j \in [N]$ undergoes one imperfect intervention in the environment $k_j \in [K]$. The interventional targets $1 \preceq \cdots \preceq N$ preserve the topological order, meaning that $i \preceq j$ only if there is a directed path from node i to node j in the underlying causal graph $\mathcal{G}$.

Remark: Asm. D.1 is directly implied by Asm. 2.1 as we need to know which environments fall into the same equivalence class. We believe that identifying the topological order is another subproblem orthogonal to identifying the latent variables, which is often termed "uncoupled/non-aligned problem" [19, 14]. As described by Zhang et al. [18], the topological order of unknown interventional targets can be recovered from single-node imperfect intervention by iteratively identifying the interventions that target the source nodes. This iterative identification process may require additional assumptions on the mixing functions [18, 15, 53, 41, 40] and the latent structured causal model [17, 40], or on the interventions, such *paired perfect* interventions per node [14, 19].

Corollary D.1 (Identifiability from single node interventions per node [38])**.** *Given N environments $\{k_1, \dots, k_N\} \subseteq [K]$ satisfying Asm. D.1, the ground truth latent variables $\mathbf{z}$ can be identified up to element-wise diffeomorphism (Defn. C.2) by combining both marginal and score invariances (eqs.* (D.3) *and* (D.4)*) under our framework (Thm. 3.1).*

The proof for Cor. D.1 is included in App. E.5. Upon element-wise identification from single-node intervention per node, existing works often provide more fine-grained identifiability results by incorporating other parametric assumptions on the mixing functions [53, 15, 18, 40]. This is explained by Proposition C.2, as element-wise identification can be refined to affine-identification (Defn. C.3) given additional parametric assumptions on the mixing functions. However, note that under the milder setting of *imperfect* intervention per node, the full graph is not identifiable without further assumptions. See [18] for more details.

Identifiability with two interventions per node Current literature in interventional CRL targeting the general nonparametric setting [19, 14] typically assumed a pair of *sufficiently different* perfect interventions per node. Thus, any latent variable $\mathbf{z}_j, j \in [N]$, as an interventional target, is uniquely shared by a pair of interventional environment $k, k' \in [K]$, forming an invariant partition $A_i = \{j\}$ constituting of individual latent node $j \in [N]$. Note that this invariance property on the interventional target induces the following distributional property:

$$[S(\mathbf{z}^k) - S(\mathbf{z}^{k'})]_j \neq 0 \qquad \text{only if} \qquad \mathcal{I}_k = \mathcal{I}_{k'} = \{j\}. \tag{D.5}$$

According to Thm. 3.1, each latent variable can thus be identified separately, giving rise to element-wise identification, as shown by [19, 14].

Identifiability under multiple distributions. More recently, Ahuja et al. [44] explains previous interventional identifiability results from a general weak distributional invariance perspective. In a nutshell, a set of variables $\mathbf{z}_A$ can be block-identified if certain invariant distributional properties hold: The invariant partition $\mathbf{z}_A$ can be block-identified (Defn. 3.1) from the rest by utilizing the *marginal distributional invariance* or *invariance on the support, mean or variance*. Ahuja et al. [44] additionally assume the mixing function to be finite degree polynomial, which leads to block-affine identification (Defn. C.1), whereas we can also consider a general nonparametric setting; they consider *one* single invariance set, which is a special case of Thm. 3.1 with one joint ι-property.

Identification algorithms. Instead of iteratively enforcing the invariance constraint across the majority of environments as described in Cor. D.1, most single-node interventional works develop equivalent constraints between pairs of environments to optimize. For example, the marginal invariance (eq. (D.3)) implies the marginal of the source node is changed *only if* it is intervened upon, which is utilized by Zhang et al. [18] to identify latent variables and the ancestral relations simultaneously. In practice, Zhang et al. [18] propose a regularized loss that includes Maximum Mean Discrepancy(MMD) between the reconstructed "counterfactual" data distribution and the interventional distribution, enforcing the distributional discrepancy that reveals graphical structure (e.g., detecting the source node). Similarly, by enforcing sparsity on the score change matrix, Varici et al. [53] restricts only score changes from the intervened node and its parents. In the nonparametric case, von Kügelgen et al. [14] optimize for the invariant (aligned) interventional targets through model selection, whereas Varici et al. [19] directly solve the constrained optimization problem formulated using score differences. Considering a more general setup, Ahuja et al. [44] provides various invariance-based regularizers as plug-and-play components for any losses that enforce a sufficient representation (Constraint 3.2).

D.3 Temporal Causal Representation Learning

High-level overview. Temporal CRL [21, 23, 22, 54, 55, 20, 56–58] focuses on retrieving latent causal structures from time series data, where the latent causal structured is typically modeled as a Dynamic Bayesian Network (DBN) [59, 60]. Existing temporal CRL literature has developed identifiability results under varying sets of assumptions. A common overarching assumption is to require the Dynamic Bayesian Network to be first-order Markovian, allowing only causal links from $t-1$ to t, eliminating longer dependencies [22, 23, 21, 55]. While many works assume that there is no instantaneous effect, restricting the latent components of $\mathbf{z}^t$ to be mutually dependent [22, 55, 23], some approaches have lifted this assumption and prove identifiability allowing for instantaneous links among the latent components at the same timestep (Lippe et al. [21]).

Data generating process. We present the data generating process followed by most temporal causal representation works and explain the underlying latent invariance and data symmetries. Let $\mathbf{z}^t \in \mathbb{R}^N$ denotes the latent vector at time t and $\mathbf{x}^t = f(\mathbf{z}^t) \in \mathbb{R}^D$ the corresponding entangled observable with $f : \mathbb{R}^N \to \mathbb{R}^D$ the shared mixing function (Defn. B.2) satisfying Asm. B.1. The actions $\mathbf{a}^t$ with cardinality $|\mathbf{a}^t| = N$ mostly only target a subset of latent variables while keeping the rest untouched, following its default dynamics [22, 23, 20, 56]. Intuitively, these actions $\mathbf{a}^t$ can be interpreted as a component-wise indicator for each latent variable $\mathbf{z}_j^t, j \in [N]$ stating whether $\mathbf{z}_j$ follows the default dynamics $p(\mathbf{z}_j^{t+1} \mid \mathbf{z}^t)$ or the modified dynamics induced by the action $\mathbf{a}_j^t$. From this perspective, the non-intervened causal variables at time t can be considered the invariant partition under our formulation, denoted by $\mathbf{z}_{A_t}^t$ with the index set A_t defined as $A_t := \{j : \mathbf{a}_j = 0\}$. Note that this invariance can be considered as a generalization of the multiview case because the realizations z_j^t, z_j^{t+1} are not exactly identical (as in the multiview case) but are related via a default transition mechanism $p(\mathbf{z}_j^{t+1} \mid \mathbf{z}^t)$. To formalize this intuition, we define $\tilde{\mathbf{z}}^t := \mathbf{z}^t \mid \mathbf{a}^t$ as the conditional random vector conditioning on the action $\mathbf{a}^t$ at time t. For the non-intervened partition $A_t \subseteq [N]$ that follows the default dynamics, the transition model should be invariant:

$$p(\mathbf{z}_{A_t}^t \mid \mathbf{z}^{t-1}) = p(\tilde{\mathbf{z}}_{A_t}^t \mid \mathbf{z}^{t-1}), \tag{D.6}$$

which gives rise to a non-trivial distributional invariance property (Defn. 2.1). Note that the invariance partition A_t could vary across different time steps, providing a set of invariance properties $\mathfrak{I} := \{\iota_t : \mathbb{R}^{|A_t|} \to \mathcal{M}_t\}_{t=1}^T$, indexed by time t. Given by Thm. 3.1, all invariant partitions $\mathbf{z}_{A_t}^t$ can be block-identified; furthermore, as shown in Proposition 3.3, the complementary variant partition can also be identified under an invertible encoder and mutual independence within $\mathbf{z}^t$ (here conditioning on the previous time step $\mathbf{z}^{t-1}$), aligning with the identification results without instantaneous effect, i.e. there is no causal link between variables at the same time step [22, 55, 20, 56]. On the other hand, temporal causal variables with instantaneous effects are shown to be identifiable *only if* "instantaneous parents" (i.e., nodes affecting other nodes instantaneously) are cut by actions [21], reducing to the setting without instantaneous effect where the latent components at t are mutually independent. Upon invariance, more fine-grained latent variable identification results, such as element-wise identifiability, can be obtained by incorporating additional technical assumptions, such as the sparse mechanism shift [20, 56, 58] and parametric latent causal model [55, 61, 62].

Identification algorithms. From a high level, the distributional invariance (eq. (D.6)) indicates full explainability and predictability of $\mathbf{z}_{A_t}^t$ from its previous time step $\mathbf{z}^{t-1}$, regardless of the action $\mathbf{a}^t$.

In principle, this invariance principle can be enforced by directly maximizing the information content of the proposed default transition density between the learned representation $p(\hat{\mathbf{z}}^t_{A_t} \mid \hat{\mathbf{z}}^{t-1})$ [21, 22]. In practice, the invariance regularization is often incorporated together with the predictability of the variant partition conditioning on actions, implemented as a KL divergence between the observational posterior $q(\hat{\mathbf{z}}^t \mid \mathbf{x}^t)$ and the transitional prior $p(\hat{\mathbf{z}}^t \mid \hat{\mathbf{z}}^{t-1}, \mathbf{a}^t)$ [20, 56, 61, 54, 55, 23], estimated using variational Bayes [63] or normalizing flow [64].

D.4 Multi-task Causal Representation Learning

High-level overview. Multi-task causal representation learning aims to identify latent causal variables via external supervision, in this case, the label information of the same instance for various tasks. Previously, multi-task learning [65, 66] has been mostly studied outside the scope of identifiability, mainly focusing on domain adaptation and out-of-distribution generalization. One of the popular ideas that was extensively used in the context of multi-task learning is to leverage interactions between different tasks to construct a generalist model that is capable of solving all classification tasks and potentially better generalizes to unseen tasks [67, 68]. Recently, Lachapelle et al. [26], Fumero et al. [11] systematically studied under which conditions the latent variables can be identified in the multi-task scenario and correspondingly provided identification algorithms.

Data generating process. The multi-task causal representation learning considers a *supervised* setup: Given a latent SCM as defined in Defn. B.1, we generate the observable $\mathbf{x} \in \mathbb{R}^D$ through some mixing function $f : \mathbb{R}^N \to \mathbb{R}^D$ satisfying Asm. B.1. Given a set of task $\mathcal{T} = \{t_1, \dots, t_k\}$, and let $\mathbf{y}^k \in \mathcal{Y}_k$ denote the corresponding task label respect to the task t_k. Each task only *directly* depends on a subset of latent variables $S_k \subseteq [N]$, in the sense that the label $\mathbf{y}^k$ can be expressed as a function that contains all and only information about the latent variable $\mathbf{z}_{S_k}$:

$$\mathbf{y}^k = r_k(\mathbf{z}_{S_k}), \tag{D.7}$$

where $r : \mathbb{R}^{|S_k|} \to \mathcal{Y}_k$ is some deterministic function which maps the latent subspace $\mathbb{R}^{|S_k|}$ to the task-specific label space $\mathcal{Y}_k$, which is often assumed to be linear and implemented using a linear readout in practice [26, 11]. For each task $t_k, k \in [K]$, we observe the associated data distribution $P_{\mathbf{x},\mathbf{y}^k}$. Consider two different tasks $t_k, t_{k'}$ with $k, k' \in [K]$, the corresponding data $\mathbf{x}, \mathbf{y}^k$ and $\mathbf{x}, \mathbf{y}^{k'}$ are *invariant* in the intersection of task-related features $\mathbf{z}_A$ with $A = S_k \cap S_{k'}$. Formally, let $r_k^{-1}(\{\mathbf{y}^k\})$ denotes the pre-image of $\mathbf{y}^k$, for which it holds

$$r_k^{-1}(\{\mathbf{y}^k\})_A = r_{k'}^{-1}(\{\mathbf{y}^{k'}\})_A, \tag{D.8}$$

showing alignment on the shared partition of the task-related latents. In the ideal case, each latent component $j \in [N]$ is *uniquely shared* by a subset of tasks, all factors of variation can be fully disentangled, which aligns with the theoretical claims by Lachapelle et al. [26], Fumero et al. [11].

Identification algorithms. We remark that the *sharing* mechanism in the context of multi-task learning fundamentally differs from that of multiview setup, thus resulting in different learning algorithms. Regarding learning, the shared partition of task-related latents is enforced to align up to the linear equivalence class (given a linear readout) instead of sample level L_2 alignment. Intuitively, this invariance principle can be interpreted as a soft version of the that in the multiview case. In practice, under the constraint of perfect classification, one employs (1) a sparsity constraint on the linear readout weights to enforce the encoder to allocate the correct task-specific latents and (2) an information-sharing term to encourage reusing latents across various tasks. Equilibrium can be obtained between these two terms only when the shared task-specific latent is element-wise identified (Defn. C.2). Thus, this soft invariance principle is jointly implemented by the sparsity constraint and information sharing regularization [11, Sec. 2.1].

D.5 Domain Generalization

High-level overview. Domain generalization aims at *out-of-distribution* performance. That is, learning an optimal encoder and predictor that performs well at some unseen test domain that preserves the same data symmetries as in the training data. At a high level, domain generalization representation learning [24, 69, 70, 30, 25] considers a similar framework as introduced for interventional CRL, with *independent* but *non-identically distributed* data, but additionally incorporated with external supervision and focusing more on model robustness perspective. While interventional CRL aims to identify the true latent factors of variations (up to some transformation), domain generalization learning focuses directly on *out-of-distribution* prediction, relying on some

invariance properties preserved under the distributional shifts. Due to the non-causal objective, new methodologies are motivated and tested on real-world benchmarks (e.g., VLCS [71], PACS [72], Office-Home [73], Terra Incognita [74], DomainNet [75]) and could inspire future real-world applicability of causal representation learning approaches.

Data generating process. The problem of domain generalizations is an *extension of supervised learning* where training data from multiple environments are available [76]. An environment is a dataset of i.i.d. observations from a joint distribution $P_{\mathbf{x}^k,\mathbf{y}^k}$ of the observables $\mathbf{x}^k \in \mathbb{R}^D$ and the label $\mathbf{y}^k \in \mathbb{R}$. The label $\mathbf{y}^k \in \mathbb{R}^m$ only depends on the invariant latents through a linear regression structural equation model [8, Assmp. 1], described as follows:

$$\begin{aligned} \mathbf{y}^k &= \mathbf{w}^* \mathbf{z}_A^k + \epsilon_k, \, \mathbf{z}_A^k \perp \epsilon_k \\ \mathbf{x}^k &= f(\mathbf{z}^k) \end{aligned} \tag{D.9}$$

where $\mathbf{w}^* \in \mathbb{R}^{D \times m}$ represents the ground truth relationship between the label $\mathbf{y}^k$ and the invariant latents $\mathbf{z}_A^k$. ϵ_k is some white noise with bounded variance and $f : \mathbb{R}^N \to \mathbb{R}^D$ denotes the shared mixing function for all $k \in [K]$ satisfying Asm. B.1. The set of environment distributions $\{P_{\mathbf{x}^k,\mathbf{y}^k}\}_{k\in[K]}$ generally differ from each other because of interventions or other distributional shifts such as covariates shift and concept shift. However, as the relationship between the invariant latents and the labels $\mathbf{w}^*$ and the mixing mechanism f are shared across different environments, the optimal risk remains invariant in the sense that

$$\mathcal{R}_k^*(\mathbf{w}^* \circ f^{-1}) = \mathcal{R}_{k'}^*(\mathbf{w}^* \circ f^{-1}), \tag{D.10}$$

where $\mathbf{w}^*$ denotes the ground truth relation between the invariant latents $\mathbf{z}_A^k$ an the labels $\mathbf{y}^k$ and f^{-1} is the inverse of the diffeomorphism mixing f(see eq. (D.9)). Note that this is a non-trivial ι property as the labels $\mathbf{y}^k$ only depend on the invariant latents $\mathbf{z}_A^k$, thus satisfying Defn. 2.1 (ii).

Identification algorithms. Different distributional invariance are enforced by interpolating and extrapolating across various environments. Among the countless contribution to the literature, *mixup* [69] linearly interpolates observations from different environments as a robust data augmentation procedure, Domain-Adversarial Neural Networks [70] support the main learning task discouraging learning domain-discriminant features, Distributionally Robust Optimization (DRO) [24] replaces the vanilla Empirical Risk objective minimizing only with respect to the worst modeled environment, Invariant Risk Minimization [30] combines the Empirical Risk objective with an invariance constraint on the gradient, and Variance Risk Extrapolation [25, V-REx], similar in spirit combines the empirical risk objective with an invariance constraint using the variance among environments. For a more comprehensive review of domain generalization algorithms, see Zhou et al. [77].

D.6 Further Explanations for Tab. 3

General clarification. Tab. 3 summarizes special cases of our invariance framework. For each work, we present their technical assumptions, the type of invariance, the implementation for the invariance and the sufficiency regularizers (to satisfy Constraints 3.1 and 3.2), and the type of identifiability they achieve. Note that this table is by no means exhaustive. Also, we omit some additional results and technical assumptions of individual papers for readability. A list of paragraphs is provided below for further clarification, as referenced in Tab. 3.

(a) Single-node intervention and parametric assumptions. Many existing CRL works that consider single node intervention per node require additional parametric assumptions, either on the mixing function [53, 18] or the latent causal model [17] or both [40], thus achieving (at least) element-wise identifiability (Defn. C.2). We conjecture these additional parametric assumptions serve two purposes: (1) to identify valid topological order of the interventional targets, as required by Asm. D.1 for Cor. D.1 (2) to get a more fine-grained identification level of affine transformation, as explained by Proposition C.2.

In the following, we restate the definition of linear latent SCM for reference:

Definition D.1 (Linear latent SCM [40, 17])**.** The latent variables $\mathbf{z}$ follows a linear SCM with Gaussian noise in the sense that

$$\mathbf{z} = A\mathbf{z} + \Gamma^{1/2}\epsilon, \tag{D.11}$$

where Γ is a diagonal matrix with positive entries, A encodes the underlying causal graph G and the ϵ is the standard Gaussian noise. For the sake of simplicity, we often define $B := \Gamma^{-1/2}(\mathrm{Id} - A)$

such that $\mathbf{z} = B^{-1}\epsilon$ to explicitly map from the exogenous noise ϵ to the latent variables $\mathbf{z}$. We use B_k to denote this matrix for the domain k.

(b) Multi-node intervention and linear mixing. Recently, Varici et al. [41] extends previous interventional CRL works to unknown multi-node interventions and achieves identifiability under the assumption of a linearly independent intervention signature matrix $M_{\text{int}} \in \{0,1\}^{N\times K}$ with each column k represents the intervened node in this environment k. The row-wise linear independence of M_{int} implies that each latent variable must have been intervened at least once. Let $M \in \{0,1\}^{N\times N}$ represent a submatrix of M_{int} with *linearly independent* columns. By multiplying M with its adjoint transpose $\text{adj}^\intercal(M)$, one obtains a matrix where each column has only one non-zero component. Applying the same transformation to the score change, this problem is reduced to a similar setting as a single node intervention per node, which can be intuitively explained using the same distributional invariance principle introduced earlier (App. D.2).

(c) Paired single-node intervention per node under nonparametric assumptions. In the nonparametric settings, several works [14, 19] have shown element-wise latent variable identification under sufficiently different paired perfect intervention per node. By having two sufficiently different interventions per node, one introduces invariance on the interventional target across these paired interventional environments. This invariance property can be enforced using the score differences [19] or algorithmically by performing model selection [14], as elaborated in App. D.2.

(d) Variant latents identification under independence. While some papers states main identification results on the variant partition, it can be explained by Thm. 3.1 and Proposition 3.3 stating that the variant block can be identified under independence and invertible encoder. For example, Wendong et al. [43, Thm. 4.5] shows block-identifiability on the intervened (variant) latents under [43, Assumption 4.4] of block-wise independence between the invariant and variant blocks.

(e) Invariance regularizers in multitask CRL Under the assumption of knowing the number of latent variables, Lachapelle et al. [26] solves a bi-level optimization problem, enforcing $L_{2,1}$ sparsity on individual task readouts in the inner problem. Coupled with a backbone shared across all tasks, this implicitly encourages discovering the ground truth overlapping partition of task support. Fumero et al. [11] lifted the constraint of assuming the known number of latents by incorporating an additional information-sharing regularizer, as explained in [11, Sec. 2.1].

(f) Invariance regularizers in domain generalization. While Sagawa et al. [24] directly optimize for the worst-case risk, a link can be drawn between this objective and the risk invariance: Given a pair of linear head $\mathbf{w}$ and encoder g shared across $[K]$ domains, let the order of risks be $\mathcal{R}^{\pi_1} \geq \mathcal{R}^{\pi_2} \dots \mathcal{R}^{\pi_K}$. Since $\mathcal{R}^{\pi_1}$ is lower bounded by $\mathcal{R}^{\pi_2}$ the minimum of the training objective in Sagawa et al. [24] ($\max_{k\in[K]} \mathcal{R}^k(w,g)$) is obtained when $\mathcal{R}^{\pi_1} = \mathcal{R}^{\pi_2}$. Then we have $\mathcal{R}^{\pi_1} = \mathcal{R}^{\pi_2} \geq \dots \geq \mathcal{R}^{\pi_K}$, and the next minimum will be obtained when $\mathcal{R}^{\pi_1} = \mathcal{R}^{\pi_2} = \mathcal{R}^{\pi_3}$, and so on so forth. The optimization procedure stops when the risks are equally minimized across all domains.

[25] minimizes variance between domain risks to enforce the risk invariance. We formally show these two are equivalent in the following. Note that the invariance principle for risk alignment can be formulated as

$$(\mathcal{R}_k - \mathcal{R}_{k'})^2 \tag{D.12}$$

According to Zhang et al. [78], variance can be equivalently expressed as pair-wise distances between the samples. Hence, we can reformulate the risk variance term in [24] as follows:

$$\text{Var}\left[\mathcal{R}\right] = \frac{1}{K^2} \sum_{k,k'\in[K]} \frac{1}{2} \left(\mathcal{R}_k - \mathcal{R}_{k'}\right)^2,$$

showing that the variance regularization in [25] enforces risk invariance.

D.7 Notable Cases Not Directly Covered by the Theory

Some works not listed in Tab. 3 cannot yet be directly explained by our invariance frameworks but are rather loosely connected. One representative line of work [20, 79, 80, 56] relies on the sparsity assumption in the latent dependency to achieve latent variable and graph identification. This assumption is closely related to the *sparse mechanism shift* hypothesis in causal representation learning [1], stating small distributional changes should not affect all causal variables but only a small subset of these. Note that the sparsity constraint is often formulated as the estimator (either for the graph [26, 56] or of the latents [80]) should be at least sparse as the ground truth one, maximizing the

cardinality of the unaffected (invariant) part. Some theoretical results do not rely on multiple data pockets that share certain invariance properties but directly employ specific properties within the observational data, such as independent support [15], or shared cluster membership [62, 81]. Some works [82] follow an orthogonal proof technique originating from the *nonlinear ICA with auxiliary variable* line of work [83]. Their proofs often rely on linear independence derived from the statistical diversity of various underlying data distributions instead of shared invariance properties. Our framework thus does not trivially include them.

E Proofs

This section includes formal proofs for the theoretical statements of the paper.

E.1 Assumption Justification

We justify the Defn. 2.1 (ii) by showing negative results under violation of this assumption, i.e., trivially invariant latent variables are not identifiable.

Proposition E.1 (General non-identifiability of trivially invariant latent variables)**.** *Consider the setup in Thm. 3.1, w.l.o.g we assume $\mathfrak{I} = \{\iota\}$ and ι is trivial in the sense that assumption (ii) in Defn. 2.1 is violated. Then, the corresponding invariant partition $\mathbf{z}_A^k$ is not identifiable for any $k \in [K]$.*

Proof. We provide a counter example as follows: Define a trivial ι-property as "if the first component is greater than zero on $A = \{1\}$ of some two dimensional latents $\mathbf{z}$". Formally,

$$\iota(\mathbf{z}_1) = \mathbf{1}[\mathbf{z}_1 > 0].$$

Consider a mixing function $f = id$ and an invertible encoder $g(\mathbf{x}) = g(f(\mathbf{z})) = [\mathbf{z}_1 + \mathbf{z}_2, \mathbf{z}_2]$ satisfying the sufficiency constraint (Constraint 3.2). Define $h_1 = h_2 = [g \circ f]_A$. Then for some realizations $z, \tilde{z}$ with $z_1 + z_2 > 0$ and $\tilde{z}_1 + \tilde{z}_2 > 0$ we have $\iota(h(\mathbf{z})) = \iota(h(\tilde{\mathbf{z}}))$. However, h_1, h_2 can not disentangle $\mathbf{z}_1$, showing non-identifiability for the invariant partition $\mathbf{z}_A$. □

Link between Defn. 2.1 (ii) and interventional discrepancy. In the following, we elaborate how Defn. 2.1 (ii) resembles the most common assumption in interventional causal representation learning, the interventional discrepancy [43, 19]. Note that this assumption may termed differently as *sufficient variability* [14, 22], *interventional regularity* [53, 41], but the mathematical formulation remain the same. We begin with restating this assumption:

Assumption E.1 (Interventional discrepancy [43])**.** Given $k \in [K]$, let p_{t_k} denote the causal mechanism of the intervened variable $\mathbf{z}_{t_k}$ with $t_k \in [N]$. We say a stochastic intervention $\tilde{p}_k$ satisfies interventional discrepancy if

$$\frac{\partial \log p_{t_k}}{\partial \mathbf{z}_{t_k}}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\mathrm{pa}(t_k)}) \neq \frac{\partial \log \tilde{p}_{t_k}}{\partial \mathbf{z}_{t_k}}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\mathrm{pa}(t_k)}) \qquad \text{almost everywhere } (a.e.).$$

Proof. We show that any cases violating the interventional discrepancy assumption also violates Defn. 2.1 (ii) and vice versa. Suppose for a contradiction that there exists $t_k \in [N]$ that is intervened in environment $k \in [K]$, and there is a non-empty interior $U \subset \mathbb{R}$ with non-zero measure where the interventional discrepancy is violated, i.e., for all $z_{t_k} \in U$, it holds

$$\frac{\partial \log p_{t_k}}{\partial z_{t_k}}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\mathrm{pa}(t_k)}) = \frac{\partial \log \tilde{p}_{t_k}}{\partial z_{t_k}}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\mathrm{pa}(t_k)}) \tag{E.1}$$

Under a single node imperfect intervention, the complementary set of the transitive closure of t_k, i.e., $A := [N] \setminus \mathrm{TC}(t_k)$ remain marginally invariant:

$$\iota(\mathbf{z}_A) = p_{\mathbf{z}_A} = \tilde{p}_{\mathbf{z}_A}.$$

W.l.o.g, we assume $A = \{1, \dots, t_k - 1\}$, define a function $h : \mathbb{R}^N \to \mathbb{R}^{|A|}$ with

$$h(\mathbf{z}) = [\mathbf{z}_1, \dots, \mathbf{z}_{t_k - 2}, \mathbf{z}_{t_k}]$$

that omits the $t_k - 1$-th component of $\mathbf{z}$ but includes the variant component t_k. Note that the marginal of $\mathbf{z}_{t_k}$ after intervention remains invariant within U because

$$
\begin{aligned}
p(\mathbf{z}_{t_k}) &= \int p_{t_k}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\text{pa}(t_k)}) p(\mathbf{z}_{\text{pa}(t_k)}) d\mathbf{z}_{\text{pa}(t_k)} && \text{pa}(t_k) \in A \\
&= \int p_{t_k}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\text{pa}(t_k)}) \tilde{p}(\mathbf{z}_{\text{pa}(t_k)}) d\mathbf{z}_{\text{pa}(t_k)} && eq.\ (\text{E.1}) \text{ and both } p_k, \tilde{p}_k \text{ pdfs} \\
&= \int \tilde{p}_{t_k}(\mathbf{z}_{t_k} \mid \mathbf{z}_{\text{pa}(t_k)}) \tilde{p}(\mathbf{z}_{\text{pa}(t_k)}) d\mathbf{z}_{\text{pa}(t_k)} \\
&= \tilde{p}(\mathbf{z}_{t_k}).
\end{aligned}
$$

Therefore, we have $\iota(h(\mathbf{z})) = \iota(h(\tilde{\mathbf{z}}))$ (with $\tilde{\mathbf{z}}$ noting the latent vectors under intervention) contradicting Defn. 2.1 (ii). The other direction (violating Defn. 2.1 (ii) implies violating Asm. E.1) can be proved using the same example.

□

E.2 Proof for Thm. 3.1

Our proof consists of the following steps:

1. We construct the optimal encoders G^* (Defn. 3.2) and selectors Φ^* (Defn. 3.4) that solves the constrained optimization problem in Thm. 3.1.
2. We show that, for any invariance property $\iota_i \in \mathfrak{I}$ and any observation $\mathbf{x}^k$ in the corresponding ι_i-equivalent subset $\mathbf{x}_{V_i}$, the selected representation $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)$ cannot contain any other information than the invariant partition $\mathbf{z}^k_{A_i}$.
3. Lastly, we prove that selected representation $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)$ relates to the ground truth invariant partition $\mathbf{z}^k_{A_i}$ through a diffeomorphism $h_k : \mathbb{R}^{|A_i|} \to \mathbb{R}^{|A_i|}$ for all invariance property $\iota_i \in \mathfrak{I}$ and for any observable $\mathbf{x}^k$ from the ι_i-equivalent subset $\mathbf{x}_{V_i}$; in other words, $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)$ block-identifies $\mathbf{z}^k_{A_i}$ in the sense of Defn. 3.1.

Lemma E.1 (Existence of optimal encoders and selectors)**.** *Consider a set of observables $\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^K\} \in \mathcal{X}$ generated from § 2 satisfying Asm. 2.1, then there exists optimal encoders G^* (Defn. 3.2) and selectors Φ^* (Defn. 3.4) which satisfy both Constraints 3.1 and 3.2.*

Proof. The optimal encoders can be constructed as the set of the inverse of the ground truth mixing functions:

$$G^* = \{f_k^{-1}\}_{k \in [K]}, \tag{E.2}$$

f_k^{-1} is smooth and invertible following Asm. B.1. By definition, for each $k \in [K]$, we have:

$$f_k^{-1}(\mathbf{x}^k) = \mathbf{z}^k \in \mathcal{Z}^k. \tag{E.3}$$

Next, we define the optimal selector $\Phi^* = \{\phi^{(i,k)}\}_{i \in [n_{\mathfrak{I}}], k \in [K]}$ such that for all $i \in n_{\mathfrak{I}}, k \in [K]$, it holds

$$\phi^{(i,k)} \oslash \mathbf{z}^k = \mathbf{z}^k_{A_i}. \tag{E.4}$$

Thus, the invariance constraint (Constraint 3.1) is trivially satisfied as given by § 2. The optimal encoder f_k^{-1} is smooth and invertible following Asm. B.1 so the sufficiency constraint (Constraint 3.2) is also satisfied. Hence, we have shown the optimum of the constrained optimization problem in Thm. 3.1 exists. □

Lemma E.2 (Invariant component isolation)**.** *Consider the same set of observables $\mathcal{S}_{\mathbf{x}}$ as introduced in Lemma E.1, then for any set of smooth encoders G (Defn. 3.2), Φ (Defn. 3.4) that satisfy the invariance condition (Constraint 3.1), the learned representation $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)$ can only be dependent on the invariant latent variables $\mathbf{z}^k_{A_i} := \{\mathbf{z}^k_j : j \in A_i\}$, not any non-invariant variables $\mathbf{z}^k_q$ with $q \in A_i^c := [N] \setminus A_i$.*

Proof. This proof directly follows Defn. 2.1 (ii). Define

$$h_k^i := \phi^{(i,k)} \oslash g_k \circ f_k \quad k \in [K]. \tag{E.5}$$

By Constraint 3.1, for all $\iota_i \in \mathfrak{I}$, we have

$$\iota_i(h_k^i(\mathbf{z}^k)) = \iota_i(h_{k'}^i(\mathbf{z}^{k'})) \quad a.s. \qquad \forall k \neq k' \in [K]. \tag{E.6}$$

According to Defn. 2.1 (ii), for all $i \in [n_{\mathfrak{I}}], k \in V_i, h_k^i$ cannot directly depends on any other latent component $\mathbf{z}_q$ with $q \notin A_i$. Therefore, we have shown that h_k^i is a function of $\mathbf{z}_{A_i}^k$, for all $i \in [n_{\mathfrak{I}}], k \in V_i$. □

Theorem 3.1 (Identifiability of multiple invariant blocks)**.** *Consider a set of observables $\mathcal{S}_{\mathbf{x}} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^K\} \in \mathcal{X}$ generated from § 2 satisfying Asm. 2.1. Let G, Φ be the set of smooth encoders (Defn. 3.2) and selectors (Defn. 3.4) that satisfy Constraints 3.1 and 3.2, then the invariant component $\mathbf{z}_{A_i}^k$ is block-identified (Defn. 3.1) by $\phi^{(i,k)} \oslash g_k$ for all $\iota_i \in \mathfrak{I}, k \in [K]$.*

Proof. Lem. E.1 verifies that there exists such optimum which satisfies both invariance and sufficiency conditions (Constraints 3.1 and 3.2). Following Lem. E.2, the composition $\phi^{(i,k)} \oslash g_k$ can only encode information related to the invariant latent subset A_i specified by the invariance property $\iota_i \in \mathfrak{I}$ for all $k \in V_i$. As given by Constraint 3.2, $\phi^{(i,k)} \oslash g_k$ contain all information the ground truth invariant latents $\mathbf{z}_{A_i}$ for i with $k \in V_i$. Therefore, the selected representation $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)$ relates to the ground truth invariant partition $\mathbf{z}_{A_i}$ through some diffeomorphism, i.e., $\mathbf{z}_{A_i}$ is blocked-identified by $\phi^{(i,k)} \oslash g_k(\mathbf{x}^k)$ for all invariance property $\iota_i \in \mathfrak{I}$ and observable $k \in V_i$, . □

E.3 Proofs for Generalization of Variant Latents

Proposition 3.2 (General non-identifiability of variant latent variables)**.** *Consider the setup in Thm. 3.1, let $A := \bigcup_{i \in [n_{\mathfrak{I}}]} A_i$ denote the union of block-identified latent indices and $A^{\mathrm{c}} := [N] \setminus A$ the complementary set where no ι-invariance $\iota \in \mathfrak{I}$ applies, then the variant latents $\mathbf{z}_{A^{\mathrm{c}}}$ cannot be identified.*

Proof. We provide a simple counter example with two latent variables $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]$, with the mixing function f being the identity map id. W.l.o.g. we assume the invariant partition to be $A = \{1\}$. According to Thm. 3.1, the invariant latent variable can be identified up to a certain bijection $h : \mathbb{R} \to \mathbb{R}$. Let $\hat{\mathbf{z}}$ be the estimated representation:

$$\hat{\mathbf{z}} = [h(\mathbf{z}_1), \mathbf{z}_2 - \mathbf{z}_1] \tag{E.7}$$

with the estimated mixing function $\hat{f} : \mathbb{R}^2 \to \mathbb{R}^2$:

$$\hat{f}(\hat{\mathbf{z}}) = [h^{-1}(\hat{\mathbf{z}}_1), \hat{\mathbf{z}}_2 + h^{-1}(\hat{\mathbf{z}}_1)], \tag{E.8}$$

then we obtain the same observations $\hat{f}(\hat{\mathbf{z}}) = f(\mathbf{z})$ whereas $\hat{\mathbf{z}}_2$ consists of a mixing of $\mathbf{z}_1$ and $\mathbf{z}_2$, showing the variant latent variable $\mathbf{z}_2$ can not be identified. □

Proposition 3.3 (Identifiability of variant latent under independence)**.** *Consider an optimal encoder $g \in G^*$ and optimal selector $\phi \in \Phi^*$ from Thm. 3.1 that jointly identify an invariant block $\mathbf{z}_A$ (we omit subscriptions k, i for simplicity), then $\mathbf{z}_{A^{\mathrm{c}}}(A^{\mathrm{c}} := [N] \setminus A)$ can be identified by the complementary encoding partition $(1 - \phi) \oslash g$ only if*

(i) g is invertible in the sense that $I(\mathbf{x}, g(\mathbf{x})) = H(\mathbf{x})$;

(ii) $\mathbf{z}_{A^{\mathrm{c}}}$ is independent on $\mathbf{z}_A$.

Proof. The mutual information between the observation $\mathbf{x} \in \mathcal{S}_{\mathbf{x}}$ and the optimal encoder $g \in G^*$ from Thm. 3.1 writes:

$$I(\mathbf{x}, g(\mathbf{x})) = H(\mathbf{x}) - H(\mathbf{x} \mid g(\mathbf{x})),$$

following condition (i) in Proposition 3.3, the second term (conditional entropy) must equal zero: $H(\mathbf{x} \mid g(\mathbf{x})) = 0$.

Writing the $\mathbf{x} = f(\mathbf{z}_A, \mathbf{z}_{A^c})$, we have

$$H(\mathbf{x} \mid g(\mathbf{x})) = H(f(\mathbf{z}_A, \mathbf{z}_{A^c}) \mid g(\mathbf{x})) = H(\mathbf{z}_A, \mathbf{z}_{A^c} \mid g(\mathbf{x})),$$

because the mixing function f is deterministic as given by Defn. B.2.

Note that $g(\mathbf{x})$ can be decomposed into two separate partitions: $\phi \oslash g(\mathbf{x}), (1-\phi) \oslash g(\mathbf{x})$; thus we can write the conditional entropy as

$$\begin{aligned} H(\mathbf{x} \mid g(\mathbf{x})) &= H(\mathbf{z}_A, \mathbf{z}_{A^c} \mid \phi \oslash g(\mathbf{x}), (1-\phi) \oslash g(\mathbf{x})) \\ &= H(\mathbf{z}_{A^c} \mid \mathbf{z}_A, \phi \oslash g(\mathbf{x}), (1-\phi) \oslash g(\mathbf{x})) + H(\mathbf{z}_A \mid \phi \oslash g(\mathbf{x}), (1-\phi) \oslash g(\mathbf{x})) \end{aligned}$$

Given that $\phi \oslash g(\mathbf{x})$ block identifies $\mathbf{z}_A$, $(1-\phi) \oslash g(\mathbf{x}))$ cannot contain any information about $\mathbf{z}_A$, hence we can simplify the second term as

$$H(\mathbf{z}_A \mid \phi \oslash g(\mathbf{x}))$$

Using the additional mutual independence assumption between $\mathbf{z}_A$ and $\mathbf{z}_{A^c}$ (Proposition 3.3 (ii)), we can rewrite the first term as

$$H(\mathbf{z}_{A^c} \mid (1-\phi) \oslash g(\mathbf{x})).$$

As a result, the condition entropy $H(\mathbf{x} \mid g(\mathbf{x}))$ can be decomposed as

$$H(\mathbf{x} \mid g(\mathbf{x})) = H(\mathbf{z}_A \mid \phi \oslash g(\mathbf{x})) + H(\mathbf{z}_{A^c} \mid (1-\phi) \oslash g(\mathbf{x})) = 0.$$

Since $H(\mathbf{z}_A \mid \phi \oslash g(\mathbf{x})) = 0$ following Constraint 3.2, the second term also must be zero, i.e., $H(\mathbf{z}_{A^c} \mid (1-\phi) \oslash g(\mathbf{x})) = 0$, which is satisfied only if $(1-\phi) \oslash g(\mathbf{x})$ is a invertible function of $\mathbf{z}_{A^c}$. That is, $(1-\phi) \oslash g(\mathbf{x})$ block-identifies $\mathbf{z}_{A^c}$. □

E.4 Proofs for Granularity of Latent Variable Identification

Proposition C.1 (Granularity of identification)**.** *Affine-identifiability (Defn. C.3) implies element-identifiability (Defn. C.2) and block affine-identifiability (Defn. C.1) while element-identifiability and block affine-identifiability implies block-identifiability (Defn. 3.1).*

Proof. The diagonal matrix Λ in eq. (C.3) is invertible and thus also a diffeomorphism h (eq. (C.2)). Hence, affine-identifiability implies element-identifiability. Affine-identifiability provides identification results with block-size one thus implies block affine-identifiability. On the other hand, block affine-identifiability is block-identifiability with affine bijection h and element-identifiability defines a special case of block-identifiability where each latent component $\mathbf{z}_i$ is an individual block. □

Proposition C.2 (Transition between identification levels)**.** *The transition between different levels of latent variable identification (Fig. 2) can be summarized as follows:*

(i) Element-level identifiability (Defns. C.2 and C.3) can be obtained from block-wise identifiability (Defns. 3.1 and C.1) when each individual latent constitutes an invariant block;

(ii) Identifiability up to an affine transformation (Defns. C.1 and C.3) can be obtained from general identifiability on arbitrary diffeomorphism (Defns. 3.1 and C.2) by additionally assuming that both the ground truth mixing function and decoder are finite degree polynomials of the same degree.

Proof. The proof for (i) is trivial in the sense that identification of block with size one boils down to the identification on the element level. (ii) directly follows Ahuja et al. [15, Thm. 4.4] and Zhang et al. [18, Lem. 1], stating that when both ground truth mixing function and decoder are finite degree polynomials of the same degree, the *invertible* encoder learns a representation that is affine linear to the ground truth latents, i.e., $\hat{\mathbf{z}} = \mathbf{L} \cdot \mathbf{z} + \mathbf{b}$ with $\mathbf{L} \in \mathbb{R}^{N \times N}$.

□

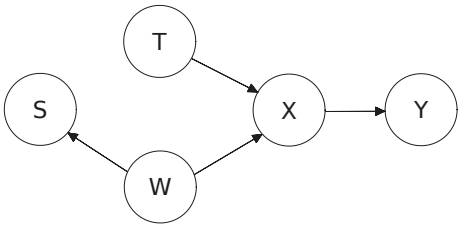
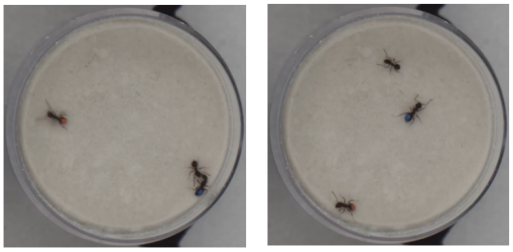


Figure 3: Causal Model for generic partially annotated scientific experiment: T treatment, $\boldsymbol{W}$ experimental settings, $\boldsymbol{X}$ high-dimensional observation, Y outcome, S annotation flag. Figure and caption adapted from [27, Fig. 1]



(a) Grooming (blue to focal) (b) No Action

Figure 4: Examples of high-dimensional observations $\boldsymbol{X}$ with corresponding annotated social behaviour Y (grooming). Figure and caption adapted from [27, Fig. 2]

E.5 Proof for Cor. D.1

Corollary D.1 (Identifiability from single node interventions per node [38])**.** *Given N environments $\{k_1, \dots, k_N\} \subseteq [K]$ satisfying Asm. D.1, the ground truth latent variables $\mathbf{z}$ can be identified up to element-wise diffeomorphism (Defn. C.2) by combining both marginal and score invariances (eqs.* (D.3) *and* (D.4)*) under our framework (Thm. 3.1).*

Proof. We consider a coarse-grained version of the underlying causal graph consisting of a block-node $\mathbf{z}_{[N-1]} := \{\mathbf{z}_1, \dots, \mathbf{z}_{N-1}\}$ and the leaf node $\mathbf{z}_N$ with $\mathbf{z}_{[N-1]}$ causing $\mathbf{z}_N$ (i.e., $\mathbf{z}_{[N-1]} \to \mathbf{z}_N$). We first select a pair of environments $V = \{0, k_N\}$ consisting of the observational environment and the environment where the leaf node $\mathbf{z}_N$ is intervened upon. According to eq. (D.3), the *marginal invariance* holds for the partition $A = [N-1]$, implying identification on $\mathbf{z}_{[N-1]}$ from Thm. 3.1. At the same time, when considering the set of environments $V' = \{0, k_1, \dots, k_{N-1}\}$, the leaf node N is the only component that satisfy *score* invariance across all environments V', because N is not the parent of any intervened node (also see [53, Lemma 4]). So here we have another invariant partition $A' = \{N\}$, implying identification on $\mathbf{z}_N$ (Thm. 3.1). By jointly enforcing the marginal and score invariance on A and A' under a sufficient encoder (Constraint 3.2), we identify both $\mathbf{z}_{[N-1]}$ as a block and $\mathbf{z}_N$ as a single element. Formally, for the parental block $\mathbf{z}_{[N-1]}$, we have:

$$\hat{\mathbf{z}}^k_{[N-1]} = g_{:N-1}(\mathbf{x}^k) \qquad \forall k \in \{0, k_1, \dots, k_N\} \tag{E.9}$$

where $g_{:N-1}(\mathbf{x}^k) := [g(\mathbf{x}^k)]_{:N-1}$ relates to the ground truth $\mathbf{z}_{[N-1]}$ through some diffeomorphism $h_{[N-1]} : \mathbb{R}^{N-1} \to \mathbb{R}^{N-1}$ (Defn. 3.1). Now, we can remove the leaf node N as follows: For each environment $k \in \{0, k_1, \dots, k_{N-1}\}$, we compute the pushforward of $P_{\mathbf{x}^k}$ using the learned encoder $g_{:N-1} : \mathcal{X}^k \to \mathbb{R}^{N-1}$:

$$P_{\hat{\mathbf{z}}^k_{[N-1]}} = g_{\#}(P_{\mathbf{x}^k})$$

Note that the estimated representations $P_{\hat{\mathbf{z}}^k_{[N-1]}}$ can be seen as a new observed data distribution for each environment k that is generated from the subgraph $\mathcal{G}_{-N}$ without the leaf node N. Using an iterative argument, we can identify all latent variables element-wise (Defn. C.2). □

F Implementation Details

This section provides further details about the experiment settings of § 4, including a formal introduction to the ISTAnt dataset, highlighted open challenges (App. F.1), and additional training settings for reproducibility (App. F.2).

F.1 Case Study: ISTAnt

Problem. Despite the majority of causal representation learning algorithms being designed to enforce the identifiability of some latent factors and tested on controlled synthetic benchmarks, there are a plethora of real-world applications across scientific disciplines requiring representation learning to answer causal questions [84–87]. Recently, Cadei et al. [27] introduced ISTAnt, the first real-world representation learning benchmark with a real causal downstream task (treatment effect

Model/Hyper-parameters	Value(s)
Encoder	DINOv2 [88]
Encoder (token)	class
MLP (head): hidden layers	1
MLP (head): hidden nodes	256
MLP (head): activation function	ReLU + Sigmoid output
Tass	or
Dropout	No
Regularization	No
Loss	BCELoss (with positive weighting)
Loss: Positive Weight	$\frac{\sum_{i=1}^{n_s} 1-Y_i}{\sum_{i=1}^{n_s} Y_i}$
Learning Rate	0.0005
Optimizer	Adam ($\beta_1 = 0.9, \beta_2 = 0.9, \epsilon = 10^{-8}$)
Batch Size	128
Epochs	15
Seeds	`range(20)`

Table 1: Model and training details for the case study on ISTAnt (§ 4.1). Table adapted from [27, Tab. 4]

estimation). This benchmark highlights different challenges (sources of biases) that could arise from machine learning pipelines even in the simplest possible setting of a randomized controlled trial. Videos of ants triplets are recorded, and a per-frame representation has to be extracted for supervised behavior classification to estimate the Average Treatment Effect of an intervention (exposure to a chemical substance). Beyond desirable identification result on the latent factors (implying that the causal variables are recovered without bias), no clear algorithm has been proposed yet on minimizing the Treatment Effect Bias (TEB) [27]. One of the challenges highlighted by Cadei et al. [27] is that in practice, there is both covariate and concept shifts due to the effect modification from training on a non-random subset of the RCT because, for example, ecologists do not label individual frames but whole video recordings. Figs. 3 and 4 shows the underlying causal graph and example input.

Solution. Relying on our framework, we can explicitly aim for low TEB by leveraging *known data symmetries* from the experimental protocol. In fact, the causal mechanism ($P(Y^e|do(\boldsymbol{X}^e = \boldsymbol{x}))$ stays invariant among the different experiment settings (i.e., individual videos or position of the petri dish). This condition can be easily enforced by existing domain generalization algorithms. For exemplary purposes, we choose Variance Risk Extrapolation [25, V-REx], which directly enforces both the invariance sufficiency constraints (Constraints 3.1 and 3.2) by minimizing the Empirical Risk together with the risk variance inter-environments.

Implementation details All training settings follow the best-performing settings from [27], which we restate in Tab. 1 for reference.

Discussion. Interestingly, Gulrajani and Lopez-Paz [89] empirically demonstrated that no domain generalization algorithm consistently outperforms Empirical Risk Minimization in *out-of-distribution* prediction. However, in this application, our goal is not to achieve high out-of-distribution accuracy but rather to identify a representation that is invariant to the effect modifiers introduced by the data labeling process. This experiment serves as a clear example of the paradigm shift of CRL via the invariance principle. While existing CRL approaches design algorithms based on specific assumptions that are often challenging to align with real-world applications, our approach begins from the application perspective. It allows for the specification of known data symmetries and desired properties of the learned representation, followed by selecting an appropriate implementation for the distance function (potentially from existing methods). Ultimately, identifiability hinges on the guarantee of asymptotic consistency in the estimates.

Table 2: Training setup for synthetic ablations in § 4.2.

Parameter	Value
Mixing function	3-layer MLP
Encoder	3-layer MLP
Decoder	3-layer MLP
Hidden dim	128
Activation	Leaky-ReLU
Optimizer	Adam
Adam: learning rate	1e-4
Adam: beta1	0.9
Adam: beta2	0.999
Adam: epsilon	1e-8
Batch size	4000
Sample size	200,000
# Epochs	500

F.2 Synthetic Ablation with "Ninterventions"

The numerical data is generated using a linear Gaussian additive noise model as follows:

$$\begin{aligned} p(\mathbf{z}_1) &= \mathcal{N}(\mu_1, \sigma_1^2) \\ p(\mathbf{z}_2 \mid \mathbf{z}_1) &= \mathcal{N}(\alpha_1 \cdot \mathbf{z}_1 + \beta_1, \sigma_2^2) \\ p(\mathbf{z}_3 \mid \mathbf{z}_2) &= \mathcal{N}(\alpha_2 \cdot \mathbf{z}_2 + \beta_2, \sigma_3^2) \\ \tilde{p}(\mathbf{z}_2) &= \mathcal{N}(\tilde{\mu}_2, \tilde{\sigma}_2^2) \end{aligned} \tag{F.1}$$

We choose $\mu_1 = 10.5, \sigma_1 = 0.8, \alpha_1 = 0.02, \beta_1 = 0, \sigma_2 = 0.5, \alpha_2 = 1, \beta_2 = 3, \sigma_3 = 1, \tilde{\sigma}_2 = 0.02$. We sample three independent $\tilde{\mu}_2$ according to a uniform distribution $\mathrm{Unif}[2, 5]$ to validate the consistency of the identification results.

For the training, we employ a simple auto-encoder architecture implementing both encoder and decoder as 3-Layer MLP. We enforce the marginal invariance using the Max Mean Discrepancy loss (MMD) on the first and last component $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_3$. Formally, the objective function writes

$$\mathcal{L}(g, \hat{f}) = \mathbb{E}_{\mathbf{x}, \tilde{\mathbf{x}}} \left[\left\| \hat{f}(g(\mathbf{x})) - \mathbf{x} \right\|_2^2 + \left\| \hat{f}(g(\tilde{\mathbf{x}})) - \mathbf{x} \right\|_2^2 \right] + \mathrm{MMD}(g(\mathbf{x})_{[1,3]}, g(\tilde{\mathbf{x}})_{[1,3]}),$$

where $\mathbf{x}, \tilde{\mathbf{x}}$ denote the observational and ninterventional data, respectively.

Further training details are summarized in Tab. 2

G Further Discussions and Connections to Other Fields

In this paper, we take a closer look at the wide range of causal representation learning methods. Interestingly, we find that the differences between them may often be more related to "semantics" than to fundamental methodological distinctions. We identified two components involved in identifiability results: preserving information of the data and a set of known invariances. Our results have two immediate implications. First, they provide new insights into the "causal representation learning problem," particularly clarifying the role of causal assumptions. We have shown that while learning the graph requires traditional causal assumptions such as additive noise models or access to interventions, identifying the causal variables may not. This is an important result, as access to causal variables is standalone useful for downstream tasks, e.g., for training robust downstream predictors or even extracting pre-treatment covariates for treatment effect estimation [90], even without knowledge of the full causal graph. Second, we have exemplified how causal representation can lead to successful applications in practice. We moved the goal post from a characterization of specific assumptions that lead to identifiability, which often do not align with real-world data, to a general recipe that allow practitioners to specify known invariances in their problem and learn representations that align with them. In the domain generalization literature, it has been widely observed that invariant training methods often do not consistently outperform empirical risk minimization (ERM). In our experiments, instead, we have demonstrated that the specific invariance enforced by

V-REx [25] entails good performance in our causal downstream task (§ 4.1). Our paper leaves out certain settings concerning identifiability that may be interesting for future work, such as discrete variables and finite samples guarantees.

One question the reader may ask, then, is "*so what is exactly causal in causal representation learning?*". We have shown that the identifiability results in typical causal representation learning are primarily based on invariance assumptions, which do not necessarily pertain to causality. We hope this insight will broaden the applicability of these methods. At the same time, we used causality as a language describing the "parameterization" of the system in terms of latent causal variables with associated known symmetries. Defining the symmetries at the level of these causal variables gives the identified representation a causal meaning, important when incorporating a graph discovery step or some other causal downstream task like treatment effect estimation. Ultimately, our representations and latent causal models can be "true" in the sense of [28] when they allow us to predict "causal effects that one observes in practice". Overall, our view also aligns with "phenomenological" accounts of causality [91], that define causal variables from a set of elementary interventions. In our setting too, the identified latent variables or blocks thereof are directly defined by the invariances at hand. From the methodological perspective, all is needed to learn causal variables is for the symmetries defined over the causal latent variables to entail some statistical footprint across pockets of data. If variables are available, learning the graph has a rich literature [92], with assumptions that are often compatible with learning the variables themselves. Our general characterization of the variable learning problem opens new frontiers for research in representation learning:

G.1 Representational Alignment and Platonic Representation

Several works ([93–96]) have highlighted the emergence of similar representations in neural models trained independently. In [96] is hypothesized that neural networks, trained with different objectives on various data and modalities, are converging toward a *shared* statistical model of reality within their representation spaces. To support this hypothesis, they measure the alignment of representations proposing to use a mutual nearest-neighbor metric, which measures the mean intersection of the k-nearest neighbor sets induced by two kernels defined on the two spaces, normalized by k. This metric can be an instance to the distance function in our formulation in Thm. 3.1. Despite not being optimized directly, several models in multiple settings (different objectives, data and modalities) seem to be aligned, hinting at the fact that their individual training objectives may be respecting some unknwon symmetries. A precise formalization of the latent causal model and identifiability in the context of foundational models remains open and will be objective for future research.

G.2 Environment Discovery

Domain generalization methods generalize to distributions potentially far away from the training, distribution, via learning representations invariant across distinct environments. However this can be costly as it requires to have label information informing on the partition of the data into environments. Automatic environment discovery ([97–99]) attempts to solve this problem by learning to recover the environment partition. This is an interesting new frontier for causal representation learning, discovering data symmetries as opposed to only enforcing them. For example, this would correspond to having access to multiple interventional distributions but without knowing which samples belong to the same interventional or observational distribution. Discovering that a data set is a mixture of distributions, each being a different intervention on the same causal model, could help increase applicability of causal representations to large obeservational data sets. We expect this to be particularly relevant to downstream tasks were biases to certain experimental settings are undesirable, as in our case study on treatment effect estimation from high-dimensional recordings of a randomized controlled trial.

G.3 Geometric Deep Learning

Geometric deep learning (GDL) ([32, 33]) is a well estabilished learning paradigm which involves encoding a geometric understanding of data as an inductive bias in deep learning models, in order to obtain more robust models and improve performance. One fundamental direction for these priors is to encode symmetries and invariances to different types of transformations of the input data, e.g. rotations or group actions ([31, 100]), in representational space. Our work can be fundamentally related with this direction, with the difference that we don't aim to model *explicitly* the transformations of the input space, but the invariances defined at the latent level. While an initial connection has been developed for disentanglement [101, 102], a precise connection between GDL and causal

representation learning remains a open direction. We expect this to benefit the two communities in both directions: (i) by injecting geometric priors in order to craft better CRL algorithms and (ii) by incorporating causality into successful GDL frameworks, which have been fundamentally advancing challenging real-world problems, such as protein folding ([103]).

Table 3: **A non-exhaustive summary of existing identifiability results for Causal Representation Learning.** All of the listed works assume injectivity of the mixing function and causal sufficiency (Markovianity) for the causal latent variables. Many listed papers depend on further technical assumptions and could yield additional results. For clarity, these are omitted; see references for details. In the table, "not assigned" means that the practical method did not directly enforce the invariance principle but considered other algorithmic designs that still implicitly preserve the data symmetries..

Work	Causal Model	Mixing Function	Invariance	Source of invariance, Inv. subset A	Invariance reg.	Sufficiency reg.	Identifiability	Expl.
Squires et al. [40, Thms. 1 & 2]	linear	linear	distributional	perfect intervention per node	$\mathrm{rank}(H^\intercal \Delta_k H) \overset{!}{=} 1$ for source nodes; linear encoder $g(\mathbf{x}) = H\mathbf{x}$, where $\Delta_k := B_k^\intercal B_k - B_0^\intercal B_0, \mathbf{z} = B_k^{-1}\epsilon$	g invertible by assumption	affine-id. and partial order preserving graph-id.	(a)
Ahuja et al. [44, Thm. 2]	nonparam.	finite-deg. poly.	marginal	single-node imperfect interventions on variant latents	$\sum_{k,k'} \sum_{j \in A} \mathrm{MMD}(p^k_{[g(\mathbf{x})]_j}, p^{k'}_{[g(\mathbf{x})]_j})$	$\sum_k \mathbb{E}_{\mathbf{x}^k} \left\lVert \hat{f}(g(\mathbf{x}^k)) - \mathbf{x}^k \right\rVert_2^2$	block affine-id.	-
Ahuja et al. [44, Thm. 3]	nonparam.	finite-deg. poly.	marginal	multi-node imperfect interventions on variant latents	$\sum_{k,k'} \sum_{j \in A} \mathrm{MMD}(p^k_{[g(\mathbf{x})]_j}, p^{k'}_{[g(\mathbf{x})]_j})$	$\sum_k \mathbb{E}_{\mathbf{x}^k} \left\lVert \hat{f}(g(\mathbf{x}^k)) - \mathbf{x}^k \right\rVert_2^2$	block affine-id.	-
Ahuja et al. [44, Thm. 4]	nonparam.	finite-deg. poly.	marginal support	imperfect interventions on variant latents	$\sum_{k,k'} \sum_{j \in A} \left\lVert \mathrm{bnd}(\hat{\mathcal{Z}}^k_j) - \mathrm{bnd}(\hat{\mathcal{Z}}^{k'}_j) \right\rVert_2^2$	$\sum_k \mathbb{E}_{\mathbf{x}^k} \left\lVert \hat{f}(g(\mathbf{x}^k)) - \mathbf{x}^k \right\rVert_2^2$	block affine-id.	-
Buchholz et al. [17]	linear Gaussian	nonparam.	marginal	perfect intervention per node	$-\mathbb{E}_{l \sim \mathcal{U}(\{0,k\})} \mathbb{E}_{\mathbf{x}^l} \ln\left(e^{\mathbf{1}_{l=k} g_k(\mathbf{x}^l)}\right)$	$\mathbb{E}_{l \sim \mathcal{U}(\{0,k\})} \mathbb{E}_{\mathbf{x}^l} \ln\left(e^{g_k(\mathbf{x}^l)} + 1\right)$	affine id. + graph id.	(a)

Work	Causal Model	Mixing Function	Invariance	Source of invariance, Inv. subset A	Invariance reg.	Sufficiency reg.	Identifiability	Expl.
Varici et al. [53, Thm. 16]	nonparam.	linear	distributional	perfect intervention per node	$\left\|\Delta^s_{\mathbf{x}}(U^\intercal)\right\|_0$. For all $j,k\in[N]$, its element $[\Delta^s_{\mathbf{x}}(U^\intercal)]_{j,k}=$ $\mathbf{1}([U^\intercal S(\mathbf{x}^0)]_j \overset{P_{\mathbf{x}^{0,k}}}{\neq} [U^\intercal S(\mathbf{x}^k)]_j)$, $g(\mathbf{x}):=U^+\mathbf{x}$	g invertible by assumption	affine-id. + graph-id.	(a)
Varici et al. [53, Thm. 13]	nonparam.	linear	distributional	imperfect intervention per node	$\left\|\Delta^s_{\mathbf{x}}(U^\intercal)\right\|_0$. For all $j,k\in[N]$, its element $[\Delta^s_{\mathbf{x}}(U^\intercal)]_{j,k}=$ $\mathbf{1}([U^\intercal S(\mathbf{x}^0)]_j \overset{P_{\mathbf{x}^{0,k}}}{\neq} [U^\intercal S(\mathbf{x}^k)]_j)$, $g(\mathbf{x}):=U^+\mathbf{x}$	g invertible by assumption	block affine-id. + graph-id.	(a)
Varici et al. [19, Thm. 3]	nonparam.	nonparam.	interventional target	paired perfect intervention per node	$\min\|\Delta^s(g)\|_0$ s.t. it is diagonal. $\Delta^s(g)_{j,k}=$ $\mathbb{E}\left[\left\|[S(g(\mathbf{x}^k))-S(g(\mathbf{x}^{k'}))]_j\right\|\right]$	g invertible by assumption	element-id. + graph-id.	(c)
Varici et al. [41, Thm. 1]	nonparam.	linear	distributional	linearly independent multi-node perfect intervention	Linear encoder $g(\mathbf{x})=H\mathbf{x}$, $H^*_i\in\mathrm{im}(\Delta s_{\mathbf{x}}\mathbf{w}_i)\backslash\mathrm{span}(H^*_{[i-1]})$ such that the $\dim$ of $\mathrm{proj}_{\mathrm{null}\left(H^*_{[i-1]}\right)}\mathrm{im}(\Delta S_{\mathbf{x}}\mathbf{w}_i)$ equals one.	g invertible by assumption	affine id. + graph id.	(b)

Work	Causal Model	Mixing Function	Invariance	Source of invariance, Inv. subset A	Invariance reg.	Sufficiency reg.	Identifiability	Expl.
Varici et al. [41, Thm. 2]	nonparam.	linear	distributional	linearly independent multinode imperfect intervention	Linear encoder $g(\mathbf{x})=H\mathbf{x}$, $H_i^* \in \mathrm{im}(\Delta s_{\mathbf{x}} \mathbf{w}_i) \setminus \mathrm{span}(H_{[i-1]}^*)$ such that the $\dim$ of $\mathrm{proj}_{\mathrm{null}\left(H_{[i-1]}^*\right)} \mathrm{im}(\Delta S_{\mathbf{x}} \mathbf{w}_i)$ equals one.	g invertible by assumption	block affine-id. + graph id.	(b)
Zhang et al. [18]	nonparam.	finite-deg. poly.	distributional	imperfect intervention per node	$-\sum_k \mathrm{MMD}(q_{\tilde{\mathbf{x}}^k}, p_{\mathbf{x}^k})$ where $\tilde{\mathbf{x}}^k$ the generated "counterfactual" pair through VAE	$-\sum_k \mathbb{E}_{\mathbf{x}^k} \log p(\mathbf{x}^k \mid g(\mathbf{x}^k))$	affine-id. + graph id.	(a)
Wendong et al. [43, Thm. 4.5]	nonparam.	nonparam.	marginal	marginal invariance from multiple fat-hand interventions on the same set of interventional targets I, invariant partition $A := [N] \setminus I$	model selection	$-\sum_k \log p_{\boldsymbol{\theta}}^k(\mathbf{x}^k)$	block-id. (known graph)	(d)
von Kügelgen et al. [14, Thm. 4.1]	nonparam.	nonparam.	interventional target	paired perfect intervention per node	model selection	$-\sum_k \log p_{\boldsymbol{\theta}}^k(\mathbf{x}^k))$	element-id. + graph-id	(c)
von Kügelgen et al. [38]	nonparam.	nonparam.	sample level on all realizations of z_A^k	one imperfect fat-hand intervention	$\left\lVert g(\mathbf{x}^1)_{\hat{A}} - g(\mathbf{x}^2)_{\hat{A}} \right\rVert_2$	$-\sum_k H(g(\mathbf{x}^k)_{\hat{A}}),\ k \in \{1,2\}$	block-id.	-

Work	Causal Model	Mixing Function	Invariance	Source of invariance, Inv. subset A	Invariance reg.	Sufficiency reg.	Identifiability	Expl.
Daunhawer et al. [39]	nonparam.	nonparam.	sample level on all realizations of z_A^k	one imperfect fat-hand intervention,	$\left\|g_1(\mathbf{x}^1)_{\hat{A}} - g_2(\mathbf{x}^2)_{\hat{A}}\right\|_2$	$-\sum_k H(g_k(\mathbf{x}^k)_{\hat{A}})$, $k \in \{1,2\}$	block-id.	-
Ahuja et al. [36]	nonparam.	nonparam.	sample level on all realizations of z_A^k	one imperfect fat-hand intervention	$\left\|g(\mathbf{x}^1)_{\hat{A}} - g(\mathbf{x}^2)_{\hat{A}} + \delta\right\|_2$	$-\sum_k \mathbb{E}_{\mathbf{x}^k} \log p(\mathbf{x}^k \mid g(\mathbf{x}^k))$, $k \in \{1,2\}$	block-id.	-
Locatello et al. [35]	nonparam.	nonparam.	sample level	one imperfect fat-hand intervention	avg. encoding	$-\sum_k \mathbb{E}_{\mathbf{x}^k} \log p(\mathbf{x}^k \mid g(\mathbf{x}^k))$, $k \in \{1,2\}$	block-id.	-
Yao et al. [34, Thm. 3.2]	nonparam.	nonparam.	sample level on all realizations of z_A^k	partial observability	$\sum_{k,k' \in [K]} \left\|g_k(\mathbf{x})_{\hat{A}} - g_{k'}(\tilde{\mathbf{x}})_{\hat{A}}\right\|_2$	$-\sum_{k \in [K]} H(g_k(\mathbf{x})_{\hat{A}})$	block-id.	-
Yao et al. [34, Thm. 3.8]	nonparam.	nonparam.	sample level on all realizations of $z_{A_i}^k$	partial observability, $k \in V_i$	$\sum_{k,k' \in V_i} \left\|g_k(\mathbf{x})_{\hat{A}(i,k)} - g_{k'}(\tilde{\mathbf{x}})_{\hat{A}(i,k')}\right\|_2$	$-\sum_{k \in [K]} H(t_k \circ g_k(\mathbf{x}))$	block-id	-
Brehmer et al. [16]	nonparam.	nonparam.	sample level	perfect intervention per node	$D_{\mathrm{KL}}\big(q(\mathcal{I}, \hat{\mathbf{z}}^{1,2} \mid \mathbf{x}^{1,2}) \| p(\mathcal{I}, \hat{\mathbf{z}}^{1,2})\big)$ where $\hat{\mathbf{z}}^k := g(\mathbf{x}^k), k \in \{1,2\}$	$-\sum_k \mathbb{E}_{\mathbf{x}^k} \log p(\mathbf{x}^k \mid g(\mathbf{x}^k))$, $k \in \{1,2\}$	element-id.	-
Lippe et al. [22]	nonparam.	nonparam.	transitional invariance on a distributional level	known-target interventions $\mathcal{I}_t$, invariant partition $A := [N] \setminus \mathcal{I}_t$	$-H(\hat{\mathbf{z}}^t_{\hat{A}^t} \mid \hat{\mathbf{z}}^{t-1})$ where $\hat{\mathbf{z}}^t := g(\mathbf{x}^t)$	$-p(\mathbf{x}^t \mid \mathbf{x}^{t-1}, \mathcal{I}_t)$	block-id.	-

Work	Causal Model	Mixing Function	Invariance	Source of invariance, Inv. subset A	Invariance reg.	Sufficiency reg.	Identifiability	Expl.
Lippe et al. [21]	nonparam.	nonparam.	transitional invariance on a distributional level	known-target, partially perfect interventions $\mathcal{I}_t$, invariant partition $A := [N] \setminus \mathcal{I}_t$	$-H(\hat{\mathbf{z}}^t_{\hat{A}^t} \mid \hat{\mathbf{z}}^{t-1})$ where $\hat{\mathbf{z}}^t := g(\mathbf{x}^t)$	$-p(\mathbf{x}^t \mid \mathbf{x}^{t-1}, \mathcal{I}_t)$	block-id.	-
Lippe et al. [23]	nonparam.	nonparam.	transitional invariance on a distributional level	binary interventions (interventional target unknown)	$D_{\mathrm{KL}}(q(\hat{\mathbf{z}}^t \mid \mathbf{x}^t) \parallel p(\hat{\mathbf{z}}^t \mid \hat{\mathbf{z}}^{t-1}, \mathbf{r}^t))$, $\mathbf{r}^t$ observed regime variable	$-\log p(\mathbf{x}^t \mid \hat{\mathbf{z}}^t)$	block-id.	-
Lachapelle et al. [26]	nonparam.	nonparam.	task support	task distribution, overlapping task supports, number of causal variables known	$\sum_t \left\lVert \hat{\mathbf{w}}^{(t)} \right\rVert_{2,1}$	$\sum_t \mathcal{R}(\hat{\mathbf{w}}^{(t)} \circ g)$	affine-id.	(e)
Fumero et al. [11]	nonparam.	nonparam.	task support	task distribution, overlapping task supports	$H(\tilde{\mathbf{w}}) + \sum_t \left\lVert \hat{\mathbf{w}}^{(t)} \right\rVert_1$	$\sum_t \mathcal{R}(\hat{\mathbf{w}}^{(t)} \circ g)$	element-id.	(e)
Sagawa et al. [24]	nonparam.	nonparam.	risk	invariant relationship between label and invariant features, preserved under covariate shift	$\max_{k \in [K]} \mathcal{R}^k(\mathbf{w} \circ g)$	$\max_{k \in [K]} \mathcal{R}^k(\mathbf{w} \circ g)$	NA	(f)

Work	Causal Model	Mixing Function	Invariance	Source of invariance, Inv. subset A	Invariance reg.	Sufficiency reg.	Identifiability	Expl.
Arjovsky et al. [30]	nonparam.	nonparam.	risk	invariant relationship between label and invariant features, preserved under covariate shift	$\|\nabla_{\mathbf{w},\mathbf{w}=1}\mathcal{R}^k(\mathbf{w}\circ g)\|^2$	$\sum_{k\in[K]}\mathcal{R}^k(\mathbf{w}\circ g)$	NA	-
Krueger et al. [25]	nonparam.	nonparam.	risk	invariant relationship between label and invariant features, preserved under covariate shift	$\mathrm{Var}(\{\mathcal{R}^k(\mathbf{w}\circ g)\}_{k\in[K]})$	$\sum_{k\in[K]}\mathcal{R}^k(\mathbf{w}\circ g)$	NA	(f)
Ahuja et al. [8]	nonparam.	nonparam.	risk	invariant relationship between label and invariant features, preserved under covariate shift	$\|\nabla_{\mathbf{w},\mathbf{w}=1}\mathcal{R}^k(\mathbf{w}\circ g)\|^2$	$\sum_{k\in[K]}\mathcal{R}^k(\mathbf{w}\circ g)+\mathrm{Var}(\mathcal{R})$	NA	-