Interventional Causal Representation Learning

Kartik Ahuja[*] **Yixin Wang**[†] **Divyat Mahajan**[*] **Yoshua Bengio**[*‡]

Abstract

The theory of identifiable representation learning aims to build general-purpose methods that extract high-level latent (causal) factors from low-level sensory data. Most existing works focus on identifiable representation learning with observational data, relying on distributional assumptions on latent (causal) factors. However, in practice, we often also have access to interventional data for representation learning. How can we leverage interventional data to help identify high-level latents? To this end, we explore the role of interventional data for identifiable representation learning in this work. We study the identifiability of latent causal factors with and without interventional data, under minimal distributional assumptions on the latents. We prove that, if the true latent variables map to the observed high-dimensional data via a polynomial function, then representation learning via minimizing the standard reconstruction loss of autoencoders identifies the true latents up to affine transformation. If we further have access to interventional data generated by hard *do* interventions on some of the latents, then we can identify these intervened latents up to permutation, shift and scaling.

1 Introduction

Modern deep learning models like GPT-3 (Brown et al., 2020) and CLIP (Radford et al., 2021) are remarkable representation learners (Bengio et al., 2013). Despite the successes, these models continue to be far from the human ability to adapt to new situations (distribution shifts) or carry out new tasks (Geirhos et al., 2020; Bommasani et al., 2021). Humans encapsulate the causal knowledge of the world in a way that is highly reusable and recomposable (Goyal and Bengio, 2020), which helps them adapt to new tasks in an ever-distribution-shifting world. How to make modern deep learning models extract a similar causal understanding of the world? This question is central to the emerging field of causal representation learning (Schölkopf et al., 2021).

A core task in causal representation learning is *provable representation identification*, i.e. developing conditions under which representation learning algorithms can provably identify latent objects (or factors) and their causal relationships. Towards understanding this task, several notable works have shown that provable representation identification for arbitrary data generation process (DGP) is impossible if we only enforce the independence between the latent factors (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Yet, real data generation processes often have additional structures we can leverage to achieve provable representation identification. Such structures include the independence between the latent factors conditional on auxiliary information (Khemakhem et al., 2020a), the sparsity of the causal connections among the latents (Lachapelle et al., 2022), and the sparsity of the mechanisms that govern the variation of the latents (Locatello et al., 2020; Ahuja et al., 2022a; Klindt et al., 2020).

Despite these efforts toward provable representation identification, most existing works focus on representation learning with observational data, relying on distributional assumptions on latent (causal)

[*]Mila - Quebec AI Institute, Université de Montréal, Quebec, Canada
[†]University of Michigan, Ann Arbor, Michigan, USA
[‡]CIFAR Senior Fellow and CIFAR AI Chair

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

factors to achieve identification. However, in practice, we often also have access to interventional data for representation learning. For example, for representation learning in robotics, we often have access to interventional data from robotic manipulation experimentss (Collins et al., 2019); for genomics and neuroscience, we also often have access to interventional data from genetic perturbation experiments (Dixit et al., 2016) and from electrical stimulation experiments (Nejatbakhsh et al., 2021) respectively. In this work, we seek to understand how we can leverage such interventional data to identify high-level (causal) factors from low-level data. The key findings are summarized below.

- Under the assumption that the true latent factors map to the high dimensional observations via a finite degree multivariate polynomial, we first show that it is possible to achieve affine identification with respect to the true latents with minimal assumptions on the support of the true latents and no further distributional assumptions.
- If we also observe data where some latents undergo a hard do intervention (Pearl, 2009), then we can guarantee affine identification up to the block of these hard intervened latents. As a result, if only one latent variable undergoes a do intervention in some environments, then those latents are identified up to permutation, shift, and scaling.

2 Representation Identification with Observational and Interventional Data

We begin with setting up the representation learning problem. We then present a suite of identifiability results that explore the role of interventional data in achieving representation identification.

The data generating process. We consider a data generating process where the observations $x \in \mathcal{X} \subseteq \mathbb{R}^n$ are generated from some underlying latent variables $z \in \mathcal{Z} \subseteq \mathbb{R}^d$, with $n \geq d$. This data generating process follows

$$\begin{aligned} z &\sim \mathbb{P}_Z, \\ x &= g(z), \end{aligned} \tag{1}$$

where $\mathbb{P}_Z$ is the distribution from which the latent z is sampled, and x is the observed data point rendered from the underlying latent z using an injective decoder $g : \mathcal{Z} \to \mathbb{R}^n$. We denote $\mathcal{Z}$ as the support of $\mathbb{P}_Z$; as a consequence, the support of the observations x is $\mathcal{X} = g(\mathcal{Z})$.

The identifiable representation learning task. To perform representation learning, we aim to find an encoder—also known as the representation function—that can help us estimate the underlying true latent variables z. Specifically, the goal is to find an encoder $f : \mathcal{X} \to \mathbb{R}^d$ and a decoder $h : \mathbb{R}^d \to \mathbb{R}^n$ such that the encoder and decoder jointly satisfy the following reconstruction identity,[4]

$$h \circ f(x) = x \qquad \forall x \in \mathcal{X}. \tag{2}$$

Given the learned encoder f, the resulting representation is $\hat{z} = f(x)$, which holds the value of the latents that the encoder guesses. Note that Equation 2 is highly underspecified and cannot in general identify the latents: it can have many solutions such that the resulting representation $\hat{z}$ do not coincide with the true latents z. For example, if we take any solution f, h of the reconstruction identity, then $b \circ f, h \circ b^{-1}$ is another valid solution where b is an invertible map. However, in practice, exact identification of the latents is often neither necessary nor reasonable. For example, we may not care about the labels given to each latent, i.e., about the coordinate permutations of z. Thus in this work, we study conditions under which the true latents are identifiable up to certain transformations, e.g. affine transformations, coordinate permutations, etc.

Overview of results. Below we present a suite of identifiability results that explore the role of interventional data for representation identification. We first show that, when the true decoder g is a polynomial function of z with a known degree and the learned decoder h is also a polynomial with the same degree, then the latent z can be identified from observational data up to affine transformations using the encoder learned from the reconstruction identity (Equation 2). We then extend this result to settings where we do not know the exact degree of g but only its upper bound. We also

[4]The identity requires the reconstruction at all points the support of $\mathcal{X}$. We can also extend our results for settings (e.g., $\mathcal{X}$ is a continuous random vector) where the identity holds almost everywhere in $\mathcal{X}$.

provide approximate affine identification guarantees for the setting when the true decoder g can be ϵ-approximated by a polynomial function. Next, we study the identifiability of the latent z when additional interventional data is available. We prove that, if we observe data where some latents undergo a hard do intervention (Pearl, 2009), then we can achieve affine identification up to the block of the latents that underwent the hard do interventions. As a result, if only one latent variable underwent a hard do intervention in some environments, then that latent variable is identifiable up to shift and scaling.

2.1 Affine representation identification with observational data

We first establish an affine identification result for representation learning from observational data, relying on the decoders being multivariate polynomial functions. We begin with the few assumptions required.

Assumption 1. *The interior of the support of Z (denoted as $\mathcal{Z}$) is a non-empty subset of $\mathbb{R}^d$.*[5]

Assumption 2. *The decoder g is a polynomial of degree p whose corresponding coefficient matrix G (a.k.a. the weight matrix) has full column rank. Specifically, the decoder g is determined by the coefficient matrix G as follows,*

$$g(z) = G[1, z, z \quad z, \quad , \underbrace{z \qquad z}_{p \text{ times}}]^\top \qquad \forall z \in \mathbb{R}^d, \tag{3}$$

where $\quad$ represents the Kronecker product with all distinct entries; for example, if $z = [z_1, z_2]$, then $z \quad z = [z_1^2, z_1 z_2, z_2^2]$.

The assumption that the matrix $G \in \mathbb{R}^{n \times q}$ has a full column rank of q implies that the decoder g is guaranteed to be injective; see the appendix for a proof of this claim. This injectivity condition on g is common in identifiable representation learning since otherwise the problem of identification will become ill-defined: multiple latents can give rise to the same observation x.

We note that the full-column-rank condition for G in Assumption 2 imposes an implicit constraint on the dimensionality n of the data; it requires that the dimensionality n is greater than the number of terms in the polynomial of degree p, namely $n = O(d^p)$, where d is the dimensionality of z.

Assumption 3. *The learned decoder h is a polynomial of degree p whose corresponding coefficient matrix H has full column rank. Similar to the decoder g, the encoder h is determined by the coefficient matrix H as follows,*

$$h(z) = H[1, z, z \quad z, \quad , \underbrace{z \qquad z}_{p \text{ times}}]^\top \qquad \forall z \in \mathbb{R}^d, \tag{4}$$

where $\quad$ represents the Kronecker product with all distinct entries.

Under Assumptions 1 to 3, we show that the representation $\hat{z}$ resulting from solving the reconstruction identity must identify the true latents z up to affine transformations. Specifically, by leveraging the relationship $x = g(z)$ in Equation 1, we write the representation $\hat{z} = f(x)$ as $\hat{z} = f \quad g(z) = a(z)$ with $a \ , \ f \quad g$. We then show that the a function must be an affine transformation.

Theorem 1. *If the data generation follows Equation 1 and Assumptions 1 to 3 hold, then any encoder f and decoder h that solve the reconstruction identity (Equation 2) must achieve affine identification: $\forall z \in \mathcal{Z}$, we must have $\hat{z} = Az + c$, where $\hat{z} = f(x)$ is the output of the encoder and z is the true latent. Moreover, A is an invertible $d \quad d$ matrix and $c \in \mathbb{R}^d$.*

The proof is in the Appendix. To understand its intuition, we consider one-dimensional latent z, three-dimensional observation x, and the true decoder g and the learned decoder h each being a degree-two polynomial. We first solve the reconstruction identity on all x, which gives $h(\hat{z}) = g(z)$, and equivalently $H[1, \hat{z}, \hat{z}^2]^\top = G[1, z, z^2]^\top$. It implies that $\hat{z}$ is at most a degree-two polynomial of z because both h and g are degree-two polynomials with full-column-rank coefficient matrices. It also implies that $\hat{z}^2$ is also a polynomial of at most degree two in z. We next argue that $\hat{z}$ must be a degree-one polynomial of z by contradiction. If $\hat{z}$ is a degree-two polynomial of z, then $\hat{z}^2$ is degree four, which contradicts the fact that $\hat{z}^2$ is at most degree two in z. Therefore, $\hat{z}$ is a degree-one polynomial in z.

[5] Here we work with $(\mathbb{R}^d; \|\cdot\|_2)$ as the metric space. A point is said to be in the interior if there exists an $\quad$ ball containing that point that is strictly in the set. The set of all the interior points defines the interior of the set.

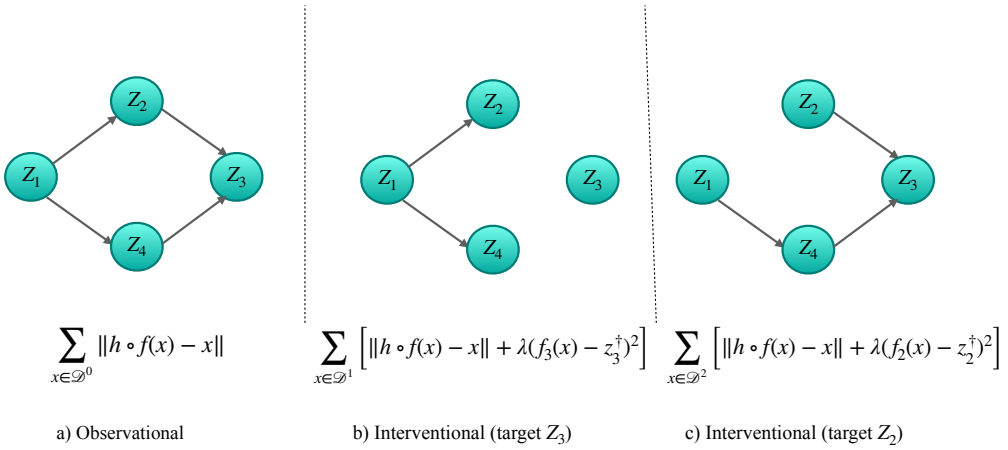


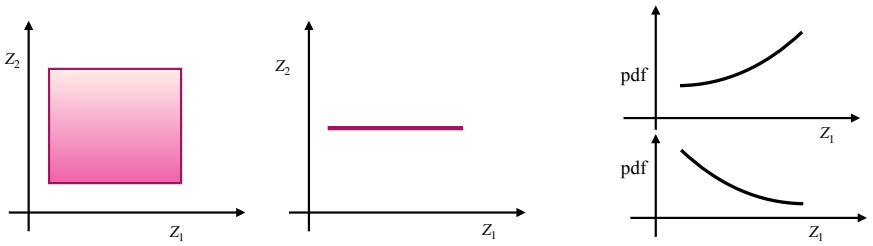
Figure 1: Illustrating the data generation process using a simple SCM. Figure 1a shows the causal DAG and the associated reconstruction loss used on observational data. Figure 1b and c show the intervened causal DAG and associated reconstruction loss, along with a penalty due to the do intervention constraints on the decoder, with z_3^* (resp. z_2^*) as the value of the intervention in Figure 1b (resp. in Figure 1c).

Extensions of Theorem 1 to polynomials with an unknown degree p. Theorem 1 requires that the degree p is known. We extend it to settings where p is unknown but an upper bound on the value p is known, which we denote as s. (It implicitly requires that the dimensionality n of the data must be sufficiently large, i.e., $n = O(d^s)$, since the coefficient matrix of H must have full column rank.) To achieve affine identification in this setting, we perform an iterative procedure. The learner first tries to solve the reconstruction identity with a polynomial $h(\,)$ of degree equal to the upper bound s. If $s > p$, then we can show that there exists no solution to the reconstruction identity (Equation 2); see Appendix for the detailed justification. The learner then decreases the degree and searches over all full-rank polynomials $h(\,)$ of degree $s \quad 1$. She repeats this procedure until the degree is p, which is when a solution to the reconstruction identity exists.

Extensions of Theorem 1 beyond polynomial decoders. While Theorem 1 assumes polynomial decoders, we know that, from Stone-Weiresstrass theorem (Rudin et al., 1976), a continuous map on a closed and bounded set can always be approximated with a high dimensional polynomial. Inspired by this result, we extend our results to a class of maps $g(\,)$ that are ϵ-approximable by a polynomial of sufficiently high degree in the Appendix. We show that the map $a(\,)$ that connects $\hat{z}$ to true z must be an approximately linear map, namely the polynomial expansion of the map $a(\,)$ must have sufficiently small weights (in terms of their norm) on the higher-order (degree greater than equal to two) terms.

2.2 Representation identification with interventional data

In the previous section, we considered representation identification from some observational data where the data is generated from Equation 1. The latent variable Z is drawn from an arbitrary distribution P_Z, which does not necessarily come from any structural causal model. We next study how interventional data could enhance representation identification. We consider the general case where at least one component of Z, say the i^{th} component, is set to a fixed value, and the remaining components are sampled from a distribution Q_{Z_i}. Note Q_{Z_i} does not have to be equal to the distribution $P_{Z_i|z_i=z}$ over samples Z_{-i} generated from Equation 1 when $z_i = z^*$. The



a) Support of Z for observational data

b) Support of Z for interventional data (Z_2 undergoes a hard do intervention)

c) Distribution of Z_1 can take arbitrary form provided the support has a non-empty interior.

Figure 2: Illustration of the assumptions on the support. In Figure 2a, we show that the support of both Z_1, Z_2 has a non-empty interior. In Figure 2b, hard do intervention occurs on Z_2, the support of Z_1 has a non-empty interior. In Figure 2c, we show that for the setting corresponding to Figure 2b, the distribution Z_1 can take arbitrary form as long as the assumption on the support is met.

data-generating process is written as follows

$$\begin{aligned} z_i &\quad z^*, \\ z_{-i} &\quad \mathsf{Q}_{Z\ i}, \\ x &\quad g(z), \end{aligned} \tag{5}$$

where the variable z_i is fixed to be equal to z^*; the remaining $d \quad 1$ variables in z (denoted as z_{-i}) are sampled from $\mathsf{Q}_{Z\ i}$; and the function g is the true decoder that generates observed x from z. The results we present below only require a restriction on the support of $\mathsf{Q}_{Z\ i}$ and this flexibility allows $\mathsf{Q}_{Z\ i}$ to model standard do interventions (Pearl, 2009), i.e., $Q_{Z\ i} = \mathsf{P}_{Z\ i|do(z_i = z\ )}$ as we illustrate below. Our DGP also allows the possibility that $\mathsf{Q}_{Z\ i} = \mathsf{P}_{Z\ i|z_i = z\ }$.

Notation-wise, we use $\mathcal{D}^0$ to denote the observational data generated from Equation 1, and use $\mathcal{D}^i$ to denote the data generated from Equation 5 when i^{th} latent variable is fixed to z^*. We denote the support of Z for DGP in Equation 5 as $\mathcal{Z}^i$. We denote the support of the latents other than z_i, i.e., the support of $\mathsf{Q}_{Z\ i}$, as $\mathcal{Z}^i$. We also denote the corresponding support of x as $\mathcal{X}^i$, where $\mathcal{X}^i = g(\mathcal{Z}^i)$.

Given the observational data ($\mathcal{D}^0$) and the data from equation (5) ($\mathcal{D}^i$), we perform representation learning via the reconstruction identity as follows,

$$h \quad f(x) = x, \qquad \forall x \in \mathcal{X} \ [\ \mathcal{X}^i. \tag{6}$$

We further need to enforce the constraint on the encoder such that, for all the data points $x \in \mathcal{X}^i$, the intervened component (say the kth component) must take some fixed value $z^\dagger$ due to the intervention:

$$f_k(x) = z^\dagger \qquad x \in \mathcal{X}^i, \tag{7}$$

where $f_k(x)$ denotes the k^{th} component of $f(x)$ and is required to take some fixed value $z^\dagger$ for all $x \in \mathcal{X}^i$.

Illustrating interventions using a structural causal model for Z. For this example only, we make an additional assumption that Z is drawn from a structural causal model, and consider the setting where we have access to both observational and interventional data. In the interventional data, we assume exactly one latent undergoes a hard do intervention as in Equation 5. In performing representation learning with the reconstruction identity, we further enforce the constraint (Equation 7) that exactly one component of the encoder also takes a fixed value. We illustrate this setting with observational and interventional data using a simple example. Suppose $Z = [Z_1, Z_2, Z_3, Z_4]$ is drawn from a structural causal model with the underlying directed acyclic graph (DAG) displayed in Figure 1a. Figure 1b (resp. Figure 1c) shows the DAG when Z_3 (resp. Z_2) undergoes a hard intervention and is set to z_3^* (resp. z_2^*). Under Figure 1a, we write down the reconstruction loss based on reconstruction identity. Under Figure 1b (resp. Figure 1c), we write down the reconstruction loss

based on the reconstruction identity (Equation 6), along with the penalty that enforces the constraint as in Equation 7.

Below we make an assumption on the support of the latents and state the identifiability result given both observational and interventional data.

Assumption 4. *The interior of support of latents other than i, $\mathcal{Z}^i$, is a non-empty subset of* $\mathbb{R}^{d-1}$.

Theorem 2. *Suppose the observational data is generated from Equation 1 and the interventional data is generated from Equation 5. If Assumptions 1 to 4 are satisfied, then the intervened latent z_i is identified up to shift and scaling, and the other latents z_{-i} are identified up to affine transformations: the solution to Equations 6 and 7 must satisfy $\hat{z}_k = az_i + b$ and $\hat{z}_{-k} = Ez_{-i} + f$, where $\hat{z}_{-k}$ denotes the estimate of the latents other than $\hat{z}_k$, and z_{-i} denotes the vector of true latents other than z_i. Moreover, $a \in \mathbb{R}, b \in \mathbb{R}$, $E \in \mathbb{R}^{d-1 \times d-1}$, and $f \in \mathbb{R}^{d-1}$.*

The proof of Theorem 2 is in the Appendix. We provide some intuition here. First, given Assumptions 1 to 3, we can already achieve affine identification due to Theorem 1. As a consequence, we have $\hat{z}_k = a_{-i}^\top z_{-i} + az_i + b$, where z_{-i} includes all entries of z other than z_i, and a_{-i} is a vector of the corresponding coefficients. Next, because both $\hat{z}_k$ and z_i are set to a fixed value, we have that $a_{-i}^\top z_{-i}$ must also take a fixed value for all values of $z_{-i} \in \mathcal{Z}^i$. Finally, we argue $a_{-i} = 0$ by contradiction. If $a_{-i} \neq 0$, then any changes to z_{-i} in the direction of a_{-i} will also reflect as a change $\hat{z}_k$, which contradicts the fact that $\hat{z}_k$ takes a fixed value. We thus conclude that $a_{-i} = 0$.

We note that Theorem 2 does not rely on any distributional assumptions (e.g., parametric assumptions) on Z; not does it rely on the nature of graphical model for Z (e.g., Z factorizes according to a certain DAG or a Markov random field). The only assumption we require is on the support of Z in observational data and interventional data; we require that the support of the variables must have a non-empty interior if they do not undergo any hard do interventions. We illustrate this support assumption in Figure 2.

More generally, Theorem 2 can be extended to setting with data from multiple environments. One such setting is where each environment corresponds to a hard do intervention on a distinct latent variable. Under the same assumptions of Theorem 2, we can identify each of the intervened latents up to permutation (since we do not know the index of the intervened latents), shift, and scaling. This setting requires that at most one of the latent variables undergoes a hard do intervention in each environment. A further extension of this setting is where multiple latent variables can undergo hard do interventions in an environment. In this setting, we can follow the exact proof recipe of Theorem 2 and achieve affine identification with respect to the block of hard intervened latents and block of remaining latents separately.

3 Conclusion

We studied the role of interventional data (level-two data) in causal representation learning. We show that, under minimal distributional conditions, latent factors are identifiable up to affine transformations from observational data under a polynomial decoder assumption. With additional interventional data, the latent variables that undergo do interventions can further be identified up to permutation, shift, and scaling. Extending these identifiability results beyond polynomials to more general decoder functions g is an interesting venue for future work. One may also consider introducing appropriate notions of approximate identification to further extend Theorems 1 and 2. Finally, we focus on interventional data from hard do interventions in this work and crucially used the nature of hard interventions in Theorem 2. Extending the results to soft interventions for flexible families of DAGs can be another fruitful direction.

A Proofs and Technical Details

A.1 Proof of Theorem 1

We restate the theorems from the main body of the paper for convenience.

Lemma 1. *If the matrix G that defines the polynomial g is full rank, then g is injective.*

Proof Suppose this is not the case and $g(z_1) = g(z_2)$ for some $z_1 \neq z_2$. Thus

$$
\begin{aligned}
G \begin{bmatrix} 1 \\ z_1 \\ z_1 \quad z_1 \\ \vdots \\ \underbrace{z_1 \qquad\qquad z_1}_{p \text{ times}} \end{bmatrix} &= G \begin{bmatrix} 1 \\ z_2 \\ z_2 \quad z_2 \\ \vdots \\ \underbrace{z_2 \qquad\qquad z_2}_{p \text{ times}} \end{bmatrix} \\
\implies G \begin{bmatrix} 0 \\ (z_1 \quad z_2) \\ z_1 \quad z_1 \quad z_2 \quad z_2 \\ \vdots \\ \underbrace{z_1 \qquad z_1}_{p \text{ times}} \quad \underbrace{z_2 \qquad z_2}_{p \text{ times}} \end{bmatrix} &= 0
\end{aligned}
\tag{8}
$$

Since $z_1 \neq z_2$ we find a non-zero vector in the null space of G which contradicts the fact that G has full column rank. Therefore, it cannot be the case that $g(z_1) = g(z_2)$ for some $z_1 \neq z_2$. Thus g has to be injective.

Theorem 3 (Restatement of Theorem 1)**.** *If the data generation follows Equation 1 and Assumptions 1 to 3 hold, then any encoder f and decoder h that solve the reconstruction identity (Equation 2) must achieve affine identification: $\forall z \in \mathcal{Z}$, we must have $\hat{z} = Az + c$, where $\hat{z} = f(x)$ is the output of the encoder and z is the true latent. Moreover, A is an invertible $d \times d$ matrix and $c \in \mathbb{R}^d$.*

Proof. We start by restating the reconstruction identity. For all $x \in \mathcal{X}$

$$
\begin{aligned}
& h \circ f(x) = x \\
& h(\hat{z}) = g(z) \\
& H \begin{bmatrix} 1 \\ \hat{z} \\ \hat{z} \quad \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad\qquad \hat{z}}_{p \text{ times}} \end{bmatrix} = G \begin{bmatrix} 1 \\ z \\ z \quad z \\ \vdots \\ \underbrace{z \qquad\qquad z}_{p \text{ times}} \end{bmatrix}
\end{aligned}
\tag{9}
$$

Following the assumptions, h is restricted to be polynomial but f bears no restriction. If $H = G$ and $f = g^{-1}$, we get the ideal solution $\hat{z} = z$, thus a solution to the above identity exists. Since H has full column rank, we can select q rows of H such that $\tilde{H} \in \mathbb{R}^{q \times q}$ and $\operatorname{rank}(\tilde{H}) = q$. Denote the corresponding matrix G that selects the same rows as $\tilde{G}$. We restate the identity in Equation 9 in terms of $\tilde{H}$ and $\tilde{G}$ as follows.

$$
\begin{aligned}
H \begin{bmatrix} 1 \\ \hat{z} \\ \hat{z}\ \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{p \text{ times}} \end{bmatrix} &= G \begin{bmatrix} 1 \\ z \\ z\ z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \\
\begin{bmatrix} 1 \\ \hat{z} \\ \hat{z}\ \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{p \text{ times}} \end{bmatrix} &= H^{-1} G \begin{bmatrix} 1 \\ z \\ z\ z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \\
\hat{z} = A \begin{bmatrix} 1 \\ z \\ z\ z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} & \\
\hat{z} = A_1 z + A_2\, z\ \ z + \quad A_p\ \underbrace{z \qquad\qquad z}_{p \text{ times}} + c &
\end{aligned}
\tag{10}
$$

Suppose at least one of $A_2, \quad , A_p$ is non-zero. Among the matrices $A_2, \quad , A_p$ which are non-zero, pick the matrix A_k with largest index k. Suppose row i of A_k has some non-zero element. Now consider the element in the row in the LHS of (10) corresponding to $\hat{z}_i^p$. Observe that $\hat{z}_i^p$ is a polynomial of z of degree kp, where $k \quad 2$. In the RHS, we have a polynomial of degree at most p. The equality between LHS and RHS is true for all $z \ 2 \ Z$ (and correspondingly all $x \ 2 \ X$). The difference of LHS and RHS is an analytic function. Note that Z has a non-empty interior, which implies Z has a positive Lebesgue measure in R^d. Therefore, from Mityagin (2015) it follows that the LHS is equal to RHS on entire R^d.

If two polynomials are equal everywhere, then their respective coefficients have to be the same. Based on supposition, LHS has non zero coefficient for terms with degree kp while RHS has zero coefficient for terms higher than degree p. This leads to a contradiction. As a result, none of $A_2, \quad , A_p$ can be non-zero. Thus $\hat{z} = A_1 z + c$.

Note that A_1 is also invertible. Suppose A_1 was not invertible, then we take a latent z and perturb it in the direction in the null space of A_1. Note that under this perturbation $\hat{z}$ does not change but z changes. Since z changes x has to change. However, $\hat{z}$ is same so the reconstructed $\hat{x}$ has to be the same. This leads to a violation of the reconstruction identity, which is a contradiction. Therefore, A_1 is invertible.

□

A.2 Extensions to polynomial $g(\ )$ with unknown degree

We provide further explanation for the case when we do not know the degree. The learner starts with solving the reconstruction identity by setting the degree of $h(\ )$ to be s; here we assume H has full rank (this implicitly requires that n is greater than the number of terms in the polynomial of degree s).

$$H \begin{bmatrix} 1 \\ \hat{z} \\ \hat{z}\ \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{s \text{ times}} \end{bmatrix} = G \begin{bmatrix} 1 \\ z \\ z\ z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \tag{11}$$

We can restrict H to rows such that it is a square invertible matrix $\bar{H}$. Denote the corresponding restriction of G as $\bar{G}$. The equality is stated as follows.

$$\begin{bmatrix} 1 \\ \hat{z} \\ \hat{z}\ \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{s \text{ times}} \end{bmatrix} = \bar{H}^{-1}\bar{G} \begin{bmatrix} 1 \\ z \\ z\ z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \tag{12}$$

If $s > p$, then $\underbrace{\hat{z} \qquad \hat{z}}_{s \text{ times}}$ is a polynomial of degree at least $p + 1$. Since the RHS contains a polynomial of degree at most p the two sides cannot be equal over a set of values of z with positive Lebesgue measure in R^d. Thus the reconstruction identity will only be satisfied when $s = p$. Thus we can start with the upper bound and reduce the degree of the polynomial on LHS till the identity is satisfied.

A.3 Extensions from polynomials to ϵ-approximate polynomials

We now discuss how to relax the polynomial assumption we discussed above. Suppose g is a continuous function that can be ϵ-approximated by a polynomial of degree p on entire R^d. If we continue to use h as a polynomial, then satisfying the exact reconstruction is not possible. Instead, we enforce approximate reconstruction as follows. For all $x \in X$, we want

$$\|h(f(x)) - x\| \leq \epsilon, \tag{13}$$

where ϵ is the tolerance on reconstruction error. We assume that $h(\cdot)$ is expressive enough that the above identity is satisfied up to ϵ tolerance. Recall $\hat{z} = f(x)$. We further simplify it as $\hat{z} = f \circ g(z) = a(z)$. We also assume that a can be η-approximated on entire R^d with a polynomial of sufficiently high degree say q. We write this as follows. For all $z \in \mathsf{R}^d$,

$$\begin{aligned} \left\| \hat{z} - \begin{bmatrix} z \\ z\ z \\ \vdots \\ \underbrace{z \qquad z}_{q \text{ times}} \end{bmatrix} \right\| &\leq \eta, \\ \left\| \hat{z} - \left({}_1 z + {}_2 z\ z + \cdots + {}_p \underbrace{z \qquad z}_{q \text{ times}} \right) \right\| &\leq \eta. \end{aligned} \tag{14}$$

We want to show that the norm of $_k$ for all $k \geq 2$ is sufficiently small. We state some assumptions needed in theorem below.

Assumption 5. *Encoder f does not take values near zero, i.e., $f_i(x) \geq \gamma\eta$ for all $x \in X$ and for all $i \in \{1, \ldots, d\}$, where $\gamma > 1$. The absolute value of each element in $\bar{H}^{-1}\bar{G}$ is bounded by a fixed constant. Consider the absolute value of the singular values of $\bar{H}$; we assume that the smallest absolute value is strictly positive and bounded below by ζ.*

Theorem 4. *Suppose $f : \mathsf{R}^n \to \mathsf{R}^d$ and $g : \mathsf{R}^d \to \mathsf{R}^n$ are functions such that g and $a = f \circ g$ can be approximated by polynomials on entire R^d with $\frac{\epsilon}{2}$ and η tolerance respectively. If $Z = [-z_{\max}, z_{\max}]^d$, where $z_{\max}$ is sufficiently large, and Assumptions 1, 3 and 5 hold, then the polynomial approximation of a (recall $\hat{z} = a(z)$) corresponding to solutions of approximate reconstruction identity in Equation 13 is approximately linear, i.e., the norms of the weights on higher order terms is sufficiently small.*

Proof sketch. We start by restating the approximate reconstruction identity. We use the fact that g can be approximated with a polynomial of say degree p to simplify the identity below.

$$\left\| H \begin{bmatrix} \hat{z} \\ \hat{z} \;\; \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{p \text{ times}} \end{bmatrix} \quad G \begin{bmatrix} z \\ z \;\; z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \right\| \quad \left\| G \begin{bmatrix} z \\ z \;\; z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \quad g(z) \right\| \quad \epsilon \tag{15}$$

Since H is full rank, we select rows of H such that $\bar{H}$ is square and invertible. The corresponding selection for G is denoted as $\bar{G}$. We write the identity in terms of these matrices as follows.

$$\begin{aligned} &\left\| \bar{H} \begin{bmatrix} \hat{z} \\ \hat{z} \;\; \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{p \text{ times}} \end{bmatrix} \quad \bar{G} \begin{bmatrix} z \\ z \;\; z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \right\| \quad \frac{3\epsilon}{2} \\ &\left\| \begin{bmatrix} \hat{z} \\ \hat{z} \;\; \hat{z} \\ \vdots \\ \underbrace{\hat{z} \qquad \hat{z}}_{p \text{ times}} \end{bmatrix} \quad \bar{H}^{-1}\bar{G} \begin{bmatrix} z \\ z \;\; z \\ \vdots \\ \underbrace{z \qquad z}_{p \text{ times}} \end{bmatrix} \right\| \quad \frac{3\epsilon}{2|\sigma_{\min}(\bar{H})|} \end{aligned} \tag{16}$$

where $|\sigma_{\min}(\bar{H})|$ is the singular value with smallest absolute value corresponding to the matrix $\bar{H}$. Now we write that the polynomial that approximates $\hat{z}_i = a_i(z)$ as follows.

$$|\hat{z}_i - \theta_1^\top z - \theta_2^\top z \cdot z - \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}}| \leq \eta \tag{17}$$

$$\begin{aligned} \hat{z}_i &\leq \theta_1^\top z + \theta_2^\top z \cdot z + \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}} + \eta \\ \hat{z}_i &\geq \theta_1^\top z + \theta_2^\top z \cdot z + \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}} - \eta \end{aligned} \tag{18}$$

From Assumption 5 we know that $\hat{z}_i \geq \gamma\eta$, where $\gamma > 2$. It follows from the above equation that

$$\theta_1^\top z + \theta_2^\top z \cdot z + \cdots + \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}} \geq (\gamma - 1)\eta \geq 0 \tag{19}$$

For $\hat{z}_i \geq \gamma\eta$, we track how $\hat{z}_i^p$ grows below.

$$\begin{aligned} \hat{z}_i &\geq \theta_1^\top z + \theta_2^\top z \cdot z + \cdots + \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}} \geq \eta \geq 0 \\ \hat{z}_i^p &\geq (\theta_1^\top z + \theta_2^\top z \cdot z + \cdots + \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}} \geq \eta)^p \\ \hat{z}_i^p &\geq (\theta_1^\top z + \theta_2^\top z \cdot z + \cdots + \theta_q^\top \underbrace{z \cdots z}_{q \text{ times}})^p (1 - \frac{1}{\gamma - 1})^p \end{aligned} \tag{20}$$

We consider $z = [z_{\max}, \quad , z_{\max}]$. Consider of the terms $\theta_{ij} z_{\max}^k$ inside the polynomial in the RHS above. We assume all components of θ are positive. Suppose $\theta_{ij} \quad \frac{1}{z_{\max}^{k \quad \kappa \quad 1}}$, where $\kappa \in (0,1)$, then the RHS in Equation 20 grows at least $z_{\max}^{(1+\kappa)p} \left(\frac{\gamma-2}{\gamma-1}\right)^p$. From Equation 16, $\hat{z}_i^p$ is very close to degree p polynomial in z. Under the assumption that the terms in $H^{-1}G$ are bounded by a constant the polynomial of degree p grows at at most $z_{\max}^p$. The difference in growth rates the Equation 16 is an increasing function of $z_{\max}$ for ranges where $z_{\max}$ is sufficiently large. Therefore, the reconstruction identity in Equation 16 cannot be satisfied for points in the neighborhood of $z = [z_{\max}, \quad , z_{\max}]$. Therefore, $\theta_{ij} < \frac{1}{z_{\max}^{k \quad \kappa \quad 1}}$. We can consider other vertices of the hypercube Z and conclude that $|\theta_{ij}| < \frac{1}{z_{\max}^{k \quad \kappa \quad 1}}$.

A.4 Proof of Theorem 2

Theorem 5 (Restatement of Theorem 2)**.** *Suppose the observational data is generated from Equation 1 and the interventional data is generated from Equation 5. If Assumptions 1 to 4 are satisfied, then the intervened latent z_i is identified up to shift and scaling, and the other latents z_{-i} are identified up to affine transformations: the solution to Equations 6 and 7 must satisfy $\hat{z}_k = az_i + b$ and $\hat{z}_{-k} = Ez_{-i} + f$, where $\hat{z}_{-k}$ denotes the estimate of the latents other than $\hat{z}_k$, and z_{-i} denotes the vector of true latents other than z_i. Moreover, $a \in \mathbb{R}, b \in \mathbb{R}$, $E \in \mathbb{R}^{d-1 \times d-1}$, and $f \in \mathbb{R}^{d-1}$.*

Proof. First note that since Assumptions 1-3 hold, we can continue to use the result from Theorem 1. From Theorem 1, it follows that the estimated latents $\hat{z}$ are an affine function of the true z. $\hat{z}_i = a^\top z + b, \forall z \in Z \setminus Z^i$, where $a \in \mathbb{R}^d, b \in \mathbb{R}$.

We write $z \in Z^i$ as $[z^*, z_{-i}]$. We consider a $z \in Z^i$ such that z_{-i} is in the interior of Z^i. We can write $\hat{z}_i = a_i z^* + a_{-i}^\top z_{-i} + b$, where a_{-i} is the vector of the values of coefficients in a other than the coefficient of i^{th} dimension, a_i is i^{th} component of a, z_{-i} is the vector of values in z other than z_i. From the constraint in Equation 7 it follows that for all $z \in Z^i$, $\hat{z}_i = z^\dagger$. We use these expressions to carry out the following simplification.

$$a_{-i}^\top z_{-i} = z^\dagger \quad a_i z^* \quad b \tag{21}$$

Consider another data point $z^{'} \in Z^i$ from the same interventional distribution such that $z_{-i}^{'} = z_{-i} + \theta e_j$ is in the interior of Z^i, where e_j is vector with one in j^{th} coordinate and zero everywhere else. From Assumption 4, we know that there exists a small enough θ such that $z_{-i}^{'}$ is in Z^i . Since the point is from the same interventional distribution $z_i^{'} = z^*$. For $z_{-i}^{'}$ we have

$$a_{-i}^\top z'_{-i} = z^\dagger \quad az^* \quad b \tag{22}$$

We take a difference of the two equations (21) and (22) to get

$$a_{-i}^\top (z_{-i} \quad z_{-i}^{'}) = \theta a_{-i}^\top e_j = 0. \tag{23}$$

From the above, we get that the j^{th} component of a_{-i} is zero. We can repeat the above argument for all j and get that $a_{-i} = 0$. □

B Related Work

This work is related to multiple threads of work in identifiable representation learning. We discuss them in groups based on the type of information they leverage for representation identification.

Time-series datasets. Several works have leveraged the structure of latent variables in time-series data to achieve identification. The canonical data generating process in these works follows $x_t \quad g(z_t)$ and the latents z_t evolve under a structured time-series. Early works in this area consider non-stationary evolution of latents (i.e. assuming no dependence between time frames) (Hyvarinen and Morioka, 2016) and then came the models that considered stationary Markovian evolution (Hyvarinen and Morioka, 2017). In recent years, these models have been generalized significantly in works like Lachapelle et al. (2022); Ahuja et al. (2021); Lippe et al. (2022).

Contrastive observation-based datasets. In another family of works including Zimmermann et al. (2021); Von Kügelgen et al. (2021); Brehmer et al. (2022); Locatello et al. (2020); Ahuja et al. (2022a), one assumes access to contrastive observation pairs $(x, \tilde{x})$. For instance, an image and its rotated version can serve as contrastive observation pairs $(x, \tilde{x})$ (Zimmermann et al., 2021). In works such as Brehmer et al. (2022), the pair $(x, \tilde{x})$ corresponds to a data point pre and post-intervention on the latents. The data generation process is similar to time-series datasets in several aspects but there are a few key differences, including (a) the points $(x, \tilde{x})$ are not necessarily ordered by time, and (b) there may not exist any causal connections between the latents associated with x and $\tilde{x}$, unlike those in time-series datasets (e.g., Lachapelle et al. (2022); Lippe et al. (2022)).

Auxiliary information datasets. In the third line of work (Khemakhem et al., 2020a,b; Ahuja et al., 2022b), one assumes access to the high-dimensional observation x (e.g., an image) and some auxiliary information u. If u is the label of the image, we obtain standard supervised learning datasets. If $u = \tilde{x}$ is the positive pair (e.g., rotation of the image), we obtain contrastive observation-based datasets. If u is the time stamp and previous time information x_{t-1}, then we obtain the time-series datasets. This line of work (Khemakhem et al., 2020a) often relies on strong assumptions on the interaction between latents and auxiliary information (e.g., latents are independent conditioned on auxiliary information) to guarantee provable identification.

On the role of interventional data in causal representation learning. Finally, we contrast our work with the existing works discussed above in terms of the type of information we leverage for representation identification: we leverage interventional knowledge (level-two knowledge in Pearl's ladder of causation (Bareinboim et al., 2022)) while most existing works leverage either observational data (level-one knowledge) (e.g. Khemakhem et al., 2020a) or counterfactual information (level-three knowledge) (e.g. Brehmer et al., 2022). Specifically in this work, we focus on studying "to what extent can we identify the latent causal variables if the data comprises different interventional distributions?" This question about causal representation learning shares the same spirit with causal discovery using interventional data, where we seek to understand how different interventional distributions help identify the underlying causal graph (Yang et al., 2018); both tasks rely on interventional data but they target different causal inference goals.

In contrast to our work, most existing works have leveraged other types of information for representation identification. For example, Brehmer et al. (2022) assume that the representation learner has access to a pair of pre- and post-intervention observations, and the data generation process therein requires their noise to be set to the same realization across the pair of points at all the nodes except the intervened nodes. Therefore, they leverage counterfactual information (level-three knowledge) for representation identification. As another set of examples, Lippe et al. (2022); Lachapelle et al. (2022) leverage pre- and post-intervention observations in adjacent time frames to study the causal relationships between the latents. Other works (e.g. Khemakhem et al. (2020a,b); Ahuja et al. (2022b)) directly work with observational data, i.e., level-one knowledge, and do not require or take advantage of interventional data. Meanwhile, these works often achieve identification guarantees by making strong assumptions on the structure of the underlying causal connections between the latents, relying on observations of auxiliary information such as the label, or capitalizing on parametric assumptions on the distribution of the latent.

References

Ahuja, K., Hartford, J., and Bengio, Y. (2021). Properties from mechanisms: an equivariance perspective on identifiable representation learning. *arXiv preprint arXiv:2110.15796*.

Ahuja, K., Hartford, J., and Bengio, Y. (2022a). Weakly supervised representation learning with sparse perturbations. *arXiv preprint arXiv:2206.01101*.

Ahuja, K., Mahajan, D., Syrgkanis, V., and Mitliagkas, I. (2022b). Towards efficient representation identification in supervised learning. *arXiv preprint arXiv:2204.04606*.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. (2022). Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Collins, J., Howard, D., and Leitner, J. (2019). Quantifying the reality gap in robotic manipulation tasks. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 6706–6712. IEEE.

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., et al. (2016). Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Goyal, A. and Bengio, Y. (2020). Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*.

Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in neural information processing systems*, 29.

Hyvarinen, A. and Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR.

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

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020a). Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR.

Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. (2020b). Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. *Advances in Neural Information Processing Systems*, 33:12768–12778.

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. (2020). Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*.

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

Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, S. (2022). Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR.

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

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR.

Mityagin, B. (2015). The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*.

Nejatbakhsh, A., Fumarola, F., Esteki, S., Toyoizumi, T., Kiani, R., and Mazzucato, L. (2021). Predicting perturbation effects from resting activity using functional causal flow. *bioRxiv*, pages 2020–11.

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rudin, W. et al. (1976). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning 2021. *arXiv preprint arXiv:2102.11107*.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467.

Yang, K., Katcoff, A., and Uhler, C. (2018). Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR.