# BEYOND THE INJECTIVE ASSUMPTION IN CAUSAL REPRESENTATION LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Causal representation learning aims to take some entangled observation, $x$, and recover the latent causal variables $z$ from which the observation was generated using a generative function $g(\cdot) : \mathcal{Z} \to \mathcal{X}$. While this problem is impossible in its full generality, there has been considerable recent progress in showing a variety of conditions in which the latents are identifiable. All of these approaches share the assumption that $g(\cdot)$ is injective: i.e. for any two observations $x_1$ and $x_2$, if $x_1 = x_2$ then the corresponding latent variables, $z_1$ and $z_2$ are equal. This assumption is restrictive but dropping it entirely would allow pathological examples that we could never hope to identify, so in order to make progress beyond injectivity, we need to make explicit the important classes of non-injective functions. In this paper we present a formal hierarchy over generative functions that includes injective functions and two non-trivial classes of non-injective functions—occluded observables and observable effects—that we argue are important for causal representation learning to consider. We demonstrate that the injective assumption is not necessary, by proving the first identifiability results in settings with occluded variables.

## 1 INTRODUCTION

Causal representation learning aims to take some entangled observations, $x$—such as images or videos—that were generated via some generative function $g(\cdot) : \mathcal{Z} \to \mathcal{X}$, and recover for each $x$ the latent causal variables $z$ used to generate it. As a running example, think of $z$ as the latent variables describing the properties of objects in a scene and $g(.)$ as a camera or rendering engine that projects these variables to an image; our task is to "invert" this rendering function to recover the original variables up to some reasonable transformation. This task is impossible without further assumptions (Hyvärinen & Pajunen, 1999; Locatello et al., 2019), but there has been significant recent progress in proving identification by leveraging a large variety of assumptions on the distribution of $z$ and / or constraints on $g$. For example, one can leverage independent $z_i$ and auxiliary variables (Hyvarinen & Morioka, 2016; 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020a;b), (sparse) temporal dependencies (Locatello et al., 2020; Lachapelle et al., 2022; Lippe et al., 2022; Ahuja et al., 2022b), mechanism knowledge (Ahuja et al., 2022a), data augmentations (Von Kügelgen et al., 2021) or multiple views (Gresele et al., 2020).

All of these approaches share the assumption that $g(\cdot)$ is injective: i.e. for any two observations $x_1$ and $x_2$, if $x_1 = x_2$ then the corresponding latent variables, $z_1$ and $z_2$ are equal.[1] This assumption is obviously restrictive—e.g. objects of interest may be occluded, or we may want to make inferences about properties like the force exerted on an object which is not directly observable—but dropping this assumption entirely would allow pathological examples that we could never hope to identify. So in order to relax this injective assumption, we first need to define the classes of problems that we might hope to solve.

In this paper, we take inspiration from Pearl's Causal Hierarchy (Pearl & Mackenzie, 2018) over causal queries, to define an analogous hierarchy over latent variables that an inference algorithm may need to address. We propose a hierarchy with four levels: the first is the familiar injective

[1] Some papers allow for observation noise such that $x = g(z) + \epsilon$, in which case these point-wise comparisons refer to the the output of $g$ before the noise is applied.

setting that we describe above. The second level considers variables which are *occluded* in the images. We define occlusion broadly to be any setting where the variable of interest is in principle observable under some view but remains ambiguous under other views. For example, a sphere that rolls behind a tree could have been observed from a different camera angle, but is not observed in the current view. Analogously, if a biologist is interested in both the shape of a cell—which can be viewed under a microscope—and the genes that the same cell is expressing—which can be observed with RNA sequencing techniques—then some the variables of interest are "occluded" if we can either use a microscope or sequence the cell.

The third level of our hierarchy are variables with *observable effects*: they cannot be inferred unambiguously with any available view, but their effect is apparent in the statistical behaviour of the observable variables. When learning from simulated data, a simple example of these variables is the parameters of the underlying physics engine used to model the environment's dynamics. For example one might want to infer the masses of the objects in a scene or the gravity constants. Explicit inference over these parameters is typically not necessary to predict the dynamics of a system, but if we want to be able to predict how the system would behave under interventions on these variables and explain their effects, modelling them explicitly becomes important. Indeed, many of the most interesting discoveries from the sciences—atoms and subatomic particles, heritable traits from genes, etc.—involved explaining observable features of the world as the result of variables that we cannot directly observe.

Finally, latent variables from the fourth level of our hierarchy are not constrained by any data that we could observe, and hence it is impossible to have any identifiability guarantees over the underlying variables. The most popular recent example of this problem is in "out-painting"[2] (Ramesh et al., 2022), where a model attempts to infer what was outside the field of view in a photograph or famous painting. In this setting there are no hard constraints on the variables that can be inferred far away from the original image, and hence we do not expect injectivity to hold.

In principle there may be many ways of partitioning non-injective problems so it is worth noting the two main benefits of this hierarchy. First, if a practitioner is in control of the data collection process, then by separating sources of non-injectivity into problems that result from the choice of view—Level 2 problems—and that which results from a latent state that can never be observed under any view but could be the subject of interventions—Level 3 problems—this hierarchy shows how the collected data needs to change if the practitioner wants to leverage known disentanglement results (i.e. Level 1 injective results). For example, adding more views in the training data will not change how much of the latent state we can observe in a Level 3 problem, but a richer set of available views may make a Level 2 problem solvable with current disentanglement techniques.

The second benefit is in defining settings for new theoretical results that show when disentanglement is possible without the injective assumption. We demonstrate this by showing how the mechanism-based disentanglement results of Ahuja et al. (2022a) can be extended to the occluded case. We show that one can partition the latent space into an occluded region and an observable region, and that mechanism knowledge can be used to disentangle the latter, while predictions on the former are constant. If we additionally leverage paths of mechanisms it is possible to infer the occluded latents.

**Related work.** This paper connects most closely to the recent literature on disentangled representation learning in nonlinear models. Several papers have shown that with inductive biases or auxiliary information, it is possible to identify representations (up to transformations such as scaling or permutations) in nonlinear independent component analysis (Gresele et al., 2020; Hyvarinen & Morioka, 2017; Hyvarinen et al., 2019; Lachapelle et al., 2022; Ahuja et al., 2022c; Von Kügelgen et al., 2021), autoencoders (Ahuja et al., 2022a;b), and deep generative models (Lippe et al., 2022; Khemakhem et al., 2020a;b; Locatello et al., 2020; Moran et al., 2022; Klindt et al., 2021; Brehmer et al., 2022; Yao et al., 2021; Xi & Bloem-Reddy, 2022). However, the starting point in all these papers is an injective mapping from latents variables to observations. In this paper, we study and characterize generative functions from latents to observations that do not satisfy injectivity. We establish a hierarchy of latent variable inference problems that violate injectivity but leave other implications in the observed data.

[2]There also exist out-painting examples from level 2 (occluded variables) in settings where it is in principle possible to view the out-painted part of the image. But this is clearly not the case for popular out-painting applications such as extending historical artworks.

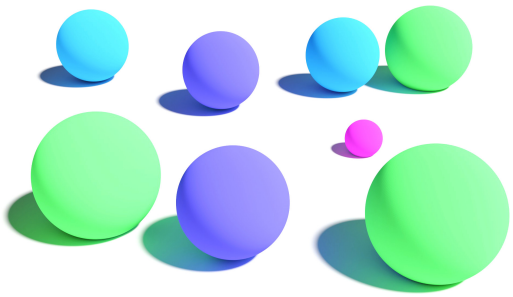

Figure 1: A image illustrating Example 1. If all the spheres and their properties are visible, then injectivity with respect to the set of latent variables holds.

We also draw connections to structural causal models (Pearl, 2009) to better interpret the variables in each level of the inference hierarchy, and is inspired by Pearl's Causal Hierarchy (Pearl & Mackenzie, 2018; Bareinboim et al., 2022). For example, the variables differ in their relationship to the observed variables as well as in the manipulations required to obtain observable implications. Moreover, level three of the hierarchy features inference problems about the causal mechanisms that govern latent variables. This connects to the work of Schölkopf et al. (2021), who discuss the open problem of learning causal models of latent variables.

The question of the relationship between observable and unobservable quantities has been studied extensively in the philosphy of science literature, particularly in the logical empiricist tradition (Boyd & Bogen, 2021). These distinctions have largely been abandoned in explaining the practice of human scientists, but they are more useful in causal representation learning where we ultimately aim to automate parts of scientific discovery. Our distinction between observable and unobservable, is related to the distinction between observational and theoretical language (Godfrey-Smith, 2021, page 28), but we do not place any restrictions on the apparatus used to collect observations.

## 2 Observation Hierarchy

We consider unsupervised representation learning, where we have access to observations $x$ in some data manifold $\mathcal{X}$. Each observation is generated by a set of $K$ latent variables $z = \{z_1, \cdots, z_K\} \in \mathcal{Z}$. We'll use uppercase $Z$ or $X$ to denote random variables and lowercase to refer to their realizations. We follow the nonlinear independent component analysis (ICA) literature but add the notion of view and propose the generative model,

$$z \sim P(Z); \quad x = g(z, V). \tag{1}$$

In this paper we make explicit the fact that $g(\cdot)$ is also parameterized by a view variable $V$ that controls our view of the data, i.e. how $\mathcal{Z}$ is mapped to $\mathcal{X}$. The view variable $V$ takes on values in a set, $\mathcal{V}$, that indexes the set of views that are possible in the data collection process. Our definition of views is deliberately general in order to capture both literal view movements—where $\mathcal{V}$ is the cross-product of all translations, rotations, pans, orbits, etc. that could be applied to a camera—and more abstract views, where $\mathcal{V}$ includes the set of different experimental apparatus (e.g. microscopes, telescopes, gene sequencing, particle detectors, etc.) that are feasibly available for a scientist to observe their system.

From a causal perspective, $g(Z, V)$ is just the structural equation that produces $X$ as a function of the latent variables, $Z$, and the view $V$. We can equivalently interpret $v$ as indexing a particular set of observation-specific interventions, each corresponding to a change in the mechanism, $g_v$, that produces $X$ from $Z$,

$$X = g_v(Z),\ v \in \mathcal{V}.$$

Note that in general, both the output space, $\mathcal{X}$, and the domain of $g_v$ and $\mathcal{Z}$, may change as a function of $v$, as different views may entail different observation types or data modalities.

**Example 1.** *Each observation $X$ depicts a 3d scene of spheres (see Figure 1). The latent variable $Z$ encodes the position of each sphere. The set of views $\mathcal{V}$ might include different camera angles that could be used to capture the scene.*

With the setup in place, we will introduce the hierarchy of representation learning settings. First, we connect the proposed generative model to the standard setting in which disentangled representation

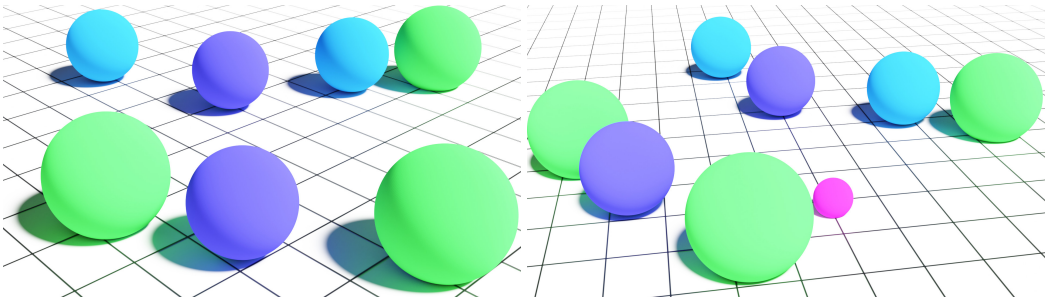

Figure 2: Example of occlusion. In the view on the left, the position of the small pink sphere is not injective, but there exists a counterfactual view in which injectivity holds.

learning is possible (with additional assumptions). Then, we use the generative model parameterized by views to define three new settings that are not addressed by existing disentanglement results.

### 2.1 Level 1: Always Observable

The base case corresponds to the standard setting where the injective assumption holds for all latents. Let $\mathcal{V}_{\text{obs}} \subseteq \mathcal{V}$, denote the subset of views available to collect the observations, then,

**Definition 1.** *The latent variables are* **always observable** *if the generative function, $g(\cdot, V)$ is injective for all $v \in \mathcal{V}_{obs}$, such that $g(z_i, v) = g(z_j, v)$ implies $z_i = z_j$.*

In most existing work, $\mathcal{V}_{\text{obs}}$, is just a singleton. Implicitly, $\mathcal{V}_{\text{obs}}$ being a singleton means that one does not have access to any additional views of the system. In this case, we can ignore the parameter $V$ and write $x = g(z)$ which recovers the standard nonlinear ICA model.

It is important emphasize that we use the term, "observable" to refer *latent* variables that are injective with respect to our observations, $x$, but we still need an inference step to infer the realization of these variables from $x$ (i.e. observ*able* latent variables are not directly observ*ed*). The injective assumption, put simply, says that any change in the latent space must be observable in the observation space. In Example 1, this implies that any change to the positions of the spheres, $Z$, is reflected in a pixel-space change in how the spheres are rendered. This seems simple enough, but it is worth noting that in scenes with multiple objects (such as those shown in Example 1), even the question of whether or not the injective assumption holds is subtle, because it depends on the space, $\mathcal{Z}$, in which $z$ is represented. If we use the common assumption that $z \in \mathbb{R}^k$ where $k$ is the product of the three dimensions required to represent the $(x, y, z)$ locations for each sphere and the number of spheres, then we can apply a permutation to $z$ that swaps the positions of the spheres without changing how the image is rendered, and hence $g(\cdot)$ is not injective; this is no longer a problem if $\mathcal{Z}$ is represented as a set instead of a vector, but that requires different inference techniques (Anonymous, 2022).

Now, consider the spheres from Example 1. Suppose there are two spheres and one is bigger than the other. Then, the smaller might move behind the larger sphere from the perspective of some camera angle, $v$, such that small changes to the occluded sphere's position will not be observable. These images now violate the injective assumption, and as such, existing disentanglement results cannot be leveraged in this setting.

### 2.2 Level 2: Occluded Observables

The most natural extension from injectivity is to occluded latents that could, in principle, be observed. Consider Figure 2 (left) where the small pink sphere is occluded by the larger green sphere and as such, it could move a small amount without changing how the scene is rendered. This image is not injective because there exist $v$ and $z_1 \neq z_2$ such that,

$$g(z_1, v) = g(z_2, v).$$

That is, the dimension of $z_1$ that encodes the position of the pink sphere can change ($z_1 \neq z_2$) but from a particular camera angle, encoded by $v$, the image is exactly the same. Clearly, if we had looked behind the green sphere (Figure 2, right), then the pink sphere would no longer be

occluded. This example illustrates how we define occluded observables: a latent variable is an occluded observable if injectivity fails under the view that rendered $x$, but there exists an alternative view under which injectivity holds. Or put differently, $g(\cdot, v)$ is not injective in the *factual* view, $v$, but there exists a *counterfactual* view, $v'$, for which $g(\cdot, v')$ is injective.

One subtlety is that, in the worst case, each latent variable $Z_k \in Z$ may have needed to be viewed from a different counterfactual view, $v'_k$, to render it injective. To accommodate this, Level 2 is formally characterized in terms of each $z_k$ as the existence of a view under which different latents imply different images (the contrapositive of the standard injective assumption),

**Definition 2.** *An observation pair $g(z_1, v) = g(z_2, v)$ with $z_1 \neq z_2$ has* **occluded observables** *if,*

$$\exists v'_k \in \mathcal{V} \text{ such that } z_{1k} \neq z_{2k} \implies g(z_{1k}, z_{1\bar{k}}, v'_k) \neq g(z_{2k}, z_{2\bar{k}}, v'_k) \quad \text{for all } k = 1, \ldots, K,$$

*where $z_{i\bar{k}}$ are the values of all but latent variable $k$ in the realized vector $z_i$.*

Put more plainly, in Level 2, there may exist latent variables that are not observable in the views we have available, but every latent variable could be made observable from a view that the modeler is willing to obtain.

Level 2 has a causal interpretation. Consider the causal graph for a view $v$ that defines the mechanism $X = g_v(Z_{\text{obs}})$, where only a strict subset $Z_{\text{obs}} \subsetneq Z$ of the latents are observable causal parents of $X$, then any variables in $Z \setminus Z_{\text{obs}}$ that are causal parents of $X$ in some other view $v'$ can be regarded as occluded observables. This is because, by definition of missing causal arrows, we can perturb any latent in $Z \setminus Z_{\text{obs}}$ without changing our observations of $X$ – this corresponds to non-injectivity. A typical example where this would occur is in a sequential setting where we receive infrequent observations, $\{X_1, X_T\}$. The latents from the intervening time steps, $\{Z_2, \ldots, Z_{T-1}\}$, are not observable in the current view, but may be observable if it is possible change the view by collecting observations at the required times.

This graphical criterion is not sufficient, however, to define occluded observables. To see this, consider a sequence of images from Example 1. The position of each sphere is clearly a causal parent of the sequence of images because interventions on the distribution positions would change the distribution of observations. However, for most views, there exist particular realizations of these positions where where some spheres are occluded by others, thereby violating injectivity. This example illustrates that injectivity is a stronger requirement than requiring causal arrows from $Z$ to $X$ because it is a condition on point-wise dependence between $z$ and $x$, rather than a distributional dependence.

### 2.3 Level 3: Observable effects

In the figure, although the pink sphere is occluded by the larger green sphere, there is a camera angle that makes the pink sphere visible. That is, there is a view (specified by a particular camera angle) where changes to the pink sphere's position is observable. Unfortunately, not all latent variables leave pointwise observable implications, even in a counterfactual view.

Instead of the positions of the spheres in the image, consider the downward force that each sphere exerts. The relationship between force—a latent variable—and images is non-injective: manipulating a sphere's downward force (e.g., by changing its mass) does not change the image. Moreover, it's uncommon for a modeler to have access to a view where the implications of changing a sphere's force is observed. There are, of course, exceptions: if a weigh-scale is introduced into the scene, or the spheres are viewed immediately before and after collisions with each other. If such views exist, we have an inference problem characterized by Level 2.

To differentiate Level 3 from Level 2, we first formally define these latent variables,

**Definition 3.** *A latent variable $U$ is* **unobservable** *if the generative function, $g(\cdot, V)$ is non-injective for all $v \in \mathcal{V}$, such that there exists $z_i \neq z_j$ for which $g(z_i, v) = g(z_j, v)$.*

On its own, this definition is far too broad to be useful: virtually everything in the universe is an unobservable latent variable, and we have no hope of inferring their unobservable values. The unobservables that are important, are those variables that we can use to explain the dynamics of our local environment as a function of their effects. We typically achieve this by experimentally perturbing these latents (e.g. applying forces) and inferring their effects indirectly via their effect on

observable latents, such as the positions, velocities and accelerations of the object in our view. To formalize this intuition, consider the distribution over observations, $P(X)$.

**Definition 4.** *An unobservable latent variable $U$ has an* **observable effect***, if there exists an intervention do$(U = u)$ such that,*

$$P(X) \neq P(X; do(U = u)). \tag{2}$$

That is, when we intervene and fix value of $U$ to $u$ and sample observations $x'$ from the corresponding intervened causal model, the distribution over the original observations, $P(X)$ and the distribution $P(X')$, are not equal. For example, imagine intervening on downward force by changing the gravity constant i.e., imagine spheres on the moon. The distribution of original images and lunar images would be distinct; on the moon, the spheres would bounce higher and be slowed less by friction. One subtlety is that to observe the causal effects of manipulating $U$, we may also need to render the observations $x$ with a different view $v \notin \mathcal{V}_{\text{obs}}$, so there may exist scenarios where we also have to account for occlusions.

As with occluded observables, latents that are unobservable have a causal interpretation. Consider the expanded structural causal model (SCM) that describes full system of variables,

$$u \leftarrow f_u(\epsilon_u); \quad z \leftarrow f_z(u, \epsilon_z), \quad x \leftarrow g_v(z), \tag{3}$$

where the noise variables $\epsilon_u$ and $\epsilon_z$ are jointly independent. The latent variables $U$ govern the distribution of $Z$, which capture factors like the positions of spheres in images. For many standard models of $f_z(\cdot)$, e.g., an additive noise model, $f_z = u + \epsilon_z$, the resulting mapping between $U$ and $X$ is not injective, because $g(u_1 + \epsilon_{z,1}) = g(u_2 + \epsilon_{z,2})$ could occur either because $u_1 = u_2$ or $u_1 \neq u_2$ and realizations of $\epsilon_z$ offset their differences. In these cases, the latents $U$ are unobservable. However, the SCM tells us that intervening on $U$ changes the distribution of $X$. From the SCM perspective, the distinction between Level 2 and Level 3 is that while Level 2 corresponds to direct interventions on $X$ by changing the view mechanism $g_v(Z)$, Level 3 involves manipulating the latent distribution $P(Z)$ by intervening on $U$.

**Unobservable latents and science.** At this point, a reader might be asking a good question: when do we care about inferring unobservable latents, given that they are once removed from the observations themselves? These variables, like force, capture dynamics in the world but can only be viewed via experiments (i.e., interventions). However, many of the most important scientific discoveries, from the discovery of atoms to planetary motion, involves inferences about unobservables latents. Indeed, a key role of the physical sciences is explaining observable phenomena in terms of unobservable variables. This connection is intentional: our definition of the distinction between observables and unobservables is inspired by the distinction between scientific theories and observations in philosophers of science (Boyd & Bogen, 2021), where theories correspond to unobservables that explain observable data.

Although this level is defined by the observable effects of interventions, Level 3 inferences do not strictly require intervening on $U$. For example, we cannot manipulate latents like the gravity constant or the mass of a planet. Instead we explain observations in the context of a specific theory (mechanism) that describes how the dynamics of a system should evolve. We believe an important open problem for causal representation learning is understanding how to formalize this procedure to posit mechanisms that explain high dimensional observations, and use them to infer Level 3 latents.

**Level 4: Unconstrained latents** Consider extrapolating beyond the boundaries of a scene, or a painting (i.e., out-painting). These extrapolations require inferences about latent variables that leave no observable implications on the scene or painting. What characterizes this setting is that we want to infer latent variables $\tilde{U}$ such that,

$$P(X) = P(X; \text{do}(\tilde{U} = u)).$$

That is, unlike in Level 3, intervening on $\tilde{U}$ does not change the distribution of $X$.

Of course, if we never have anything observable or testable, then there is no hope of identifiability. It is worth noting, however, that this "impossible" case is not completely without structure: any extrapolation still needs to be consistent with what we observe. For example, if we ask a modern generative model to guess parts of an image that are outside the field of view, we see parts of the image that we can never observe (so not identifiable) but the results are still image-like.

# 3 IDENTIFICATION RESULTS

The goal in this section is to characterize the identification of representations learned from Level 2 observations. The feature of Level 2 observations is that there there exists a counterfactual view with which the generative function would be injective. If the modeler could gather observations using that view, they could appeal to existing results on disentangled representation learning. We instead consider the case where the modeler cannot access observations from the counterfactual, or wants to use the existing observations, which contain occluded latents. To this end, we develop identification results using temporal observations, following Ahuja et al. (2022a).

## 3.1 PRELIMINARIES

For simplicity, we assume that the modeler only has access to a single view $v$ and the generative function $g_v(Z)$ is non-injective. Because of the absence of other views, we drop the explicit dependence on $V$ from the generative function and write $g(Z)$. The data generating process (DGP) follows from Ahuja et al. (2022a) which studies equation 1 in a temporal setting, where the latents that generate observations evolve according to a mechanism $m$ such that,

$$x_t \leftarrow g(z_t); \quad z_{t+1} \leftarrow m(z_t). \tag{4}$$

In Ahuja et al. (2022a), the true generative function $g$ is bijective with respect to the data manifold, and as a result, their generative process and theoretical results correspond to Level 1 in the hierarchy. As a first step to generalizing their results to Level 2, we first need to relax the bijective requirement and re-derive their main result under a weaker set of assumptions.

Our hypothesis class is the set of all encoder / decoder pairs, $(\tilde{f}, \tilde{g})$ over which we can search for solutions to the representation learning problem. We want to know whether the true encoder, $f$, is identifiable—i.e. we ask if $f$ is the unique solution to our learning problem—so we can restrict this hypothesis class to only those encoder / decoder pairs that can both perfectly reconstruct the observations and reconstruct temporal pairs after applying a known mechanism, $m$, such that,

$$x_{t+1} = \tilde{g} \circ m \circ \tilde{f} \circ x_t; \quad x = \tilde{g} \circ \tilde{f} \circ x. \tag{5}$$

Instead of considering the set of bijective functions from $\mathcal{Z}$ to $\mathcal{X}$ as our hypothesis class, we relax the bijectivity requirement and consider functions $\tilde{g}$ ($\tilde{f}$) that have a right inverse (left inverse). The significance of this change is that now two different realizations of latents could be decoded to the same image, which is a defining property of non-injective observations.

Before moving to non-injective generative functions, we first need to show that right (left) inverses suffice to re-derive Theorem 1 of Ahuja et al.. Although we consider encoders and decoders that permit different latent values to decode to the same observation, if the true generative function is in fact bijective, the constraints on the hypothesis class from equation 5 suffice to recover representations up to "equivariances" of the mechanism, which are any invertible functions, $a$, that commute with the mechanism such that $m \circ a(z) = a \circ m(z)$. The caveat is this result now only holds on a restricted domain defined by the encoder. Consider the set $\mathcal{Z}' = \tilde{f}(\mathcal{X})$ that contains all the points in the image of the encoder. We can define a new mechanism $m' : \mathcal{Z}' \to \mathcal{Z}$ such that $m'(z) = m(z)$ for each $z \in \mathcal{Z}'$. That is, the mechanism $m'$ is same as $m$ but has its domain restricted to $\mathcal{Z}'$.

**Theorem 1.** *Define $\mathcal{E}$ as the set of equivariances of $m'$. If the data generating process in equation 4 holds and the true generative function $g$ is bijective, then the encoders that satisfy equation 5 identify the true latent up to the equivariances of $m'$.*

**Remarks.** The proof is in the appendix. The significance of this result is, if parts of the observation space are in fact injective, then this hypothesis class of solutions can match the known identifiability results. In the next part, we establish that indeed, some observations satisfy injectivity, and use this theorem and mechanisms to constrain the values of latents during occlusion.

## 3.2 LEVEL 2 IDENTIFICATION

What makes identifying representations challenging from non-injective observations is that at some time points, latents can become occluded and from the observations at these points the image appears

static as the occluded latents change. To develop identification results, first we establish that just before and after the non-injective time points, the observations are injective, and the corresponding latent variables are identified up to some transformation at these points. Then, we use mechanisms to constrain the set of values that the latents can take on during the occlusion.

**Setup.** We modify the DGP in Eqn. 4 so that now, the true generative function $g$ is surjective. A tuple $(x, x') \in \mathcal{X} \times \mathcal{X}$ is realizable under this DGP if $\exists z \in \mathcal{Z}$ for which $x = g(z)$ and $x' = g \circ m(z)$. The set of all realizable tuples is $\mathcal{X}^{\mathsf{r}}$.

A key observation is that we can divide the image of $g$, i.e., $\mathcal{X}$, into two parts. $\mathcal{X}_{\mathsf{in}}$ consists of all the points in the image for which there exists exactly one $z$ such that $x = g(z)$ and $\mathcal{X}_{\mathsf{nin}} = \mathcal{X} \setminus \mathcal{X}_{\mathsf{in}}$. We can further define $\mathcal{Z}_{\mathsf{in}}$ as the preimage of $\mathcal{X}_{\mathsf{in}}$ under $g$ and $\mathcal{Z}_{\mathsf{nin}} = \mathcal{Z} \setminus \mathcal{Z}_{\mathsf{in}}$ as the preimage of $\mathcal{X}_{\mathsf{nin}}$ under $\tilde{g}$.

Consider the points in the set $\mathcal{X}_{\mathsf{in}}$. If we restrict $\tilde{g}$'s domain to $\mathcal{Z}_{\mathsf{in}}$, then $\tilde{g}$ is invertible. For each point $x \in \mathcal{X}_{\mathsf{in}}$ the subsequent point that is generated is $x' = g \circ m \circ g^{-1}(x)$. We denote the set of the tuples generated from $\mathcal{X}_{\mathsf{in}}$ as $\mathcal{X}^{\mathsf{r}}_{\mathsf{in}} \subseteq \mathcal{X}^{\mathsf{r}}$. Following the previous section, we define $\mathcal{Z}'_{\mathsf{in}} = \tilde{f}(\tilde{X}_{\mathsf{in}})$ to be the set of all the points in the image of the encoder. Define a new mechanism $m_{\mathsf{in}} : \mathcal{Z}'_{\mathsf{in}} \to \mathcal{Z}$ as follows for each $z \in \mathcal{Z}'_{\mathsf{in}}, m_{\mathsf{in}}(z) = m(z)$. Thus $m_{\mathsf{in}}$ is same as $m$ but has domain restricted to $\mathcal{Z}'_{\mathsf{in}}$.

**Theorem 2.** *If the data generating process in equation 4 holds and the true generative function $g$ is surjective, then the encoders that satisfy the equation 5 identify the true latents associated with data in $\mathcal{X}_{\mathsf{in}}$ up to the equivariances of $m_{\mathsf{in}}$.*

The proof of the above theorem is exactly the same as the previous theorem. Denote the set of equivariances of $m_{\mathsf{in}}$ as $\mathcal{E}_{\mathsf{in}}$. The identity in equation 5 enforces constraint on data on entire $\mathcal{X}$ while the above theorem only describes the identification guarantees for the latents associated with the points in $\mathcal{X}_{\mathsf{in}}$.

The discussion above exploits points only in $\mathcal{X}_{\mathsf{in}}$ and thus only speaks to the identification of the corresponding latents. We now consider a particular case where the mechanism consists of offsets to the latent variables which lets us derive identification results up to permutations and scaling on the injective region and describe the output of the encoder on the non-injective region.

### 3.2.1 Analyzing known offsets case

Define the latent space as a $d-$dimensional hypercube $\mathcal{Z} = [0, 1]^d$ and for simplicity, suppose that corner of the cube is occluded such that, $\mathcal{Z}_{\mathsf{nin}} = [v, 1]^d$, where $v > 0$ is parameterizes the view that defines which subset of $\mathcal{Z}$ is occluded. The remainder of the space is observable, such that $\mathcal{Z}_{\mathsf{in}} = \mathcal{Z} \setminus \mathcal{Z}_{\mathsf{nin}}$. As before, let $x \leftarrow g(z)$, but restrict $g : \mathcal{Z} \to \mathcal{X}$ to be a piece-wise analytic function defined as follows: $g_1 : \mathcal{Z}_{\mathsf{in}} \to \mathcal{X}_{\mathsf{in}}$, is an injective, analytic rendering function defined on the visible subset of $\mathcal{Z}$; we can think of the output of $g_1$ as rendering non-occluded images of the full scene—indeed with $v = 1$, this is just the standard injective setting. Suppose $g_2 : \mathcal{Z}_{\mathsf{nin}}$ is a constant function, since the image $x$ is unchanged while the latents are occluded. We then define $g$ as,

$$g_v(z) = \begin{cases} g_1(z) \text{ if } z \in \mathcal{Z}_{\mathsf{in}}(v) \\ g_2(z) \text{ if } z \in \mathcal{Z}_{\mathsf{nin}}(v) \end{cases}$$

Under this data generating process, we show that by assuming access to the same mechanisms as those used in Ahuja et al. (2022a), we can derive analogous results without requiring injectivity. The "mechanism" that acts on $z$ is $m(z) = z + \delta_i$ for a set of $d$ offsets, $\Delta = \{\delta_i\}_{i=1}^d$ such that,

$$\tilde{g} \circ m \circ \tilde{f}(x) = x' \Rightarrow \tilde{g}(\tilde{f}(x) + \delta') = x' \tag{6}$$

We need these pertubations to be diverse enough that they span the latent space.

**Assumption 1.** *The dimension of the span of the true and estimated perturbations $\Delta$ and $\Delta'$ is $d$, i.e., dim (span($\Delta$)) $= d$.*

Like Ahuja et al., we need a regularity condition on $a(\cdot)$, but we differ slightly in that $a(\cdot)$ is constrained to the restricted space $\mathcal{Z}'$,

**Assumption 2.** *$a : \mathcal{Z}' \to \mathcal{Z}'$ is an analytic function. For each component $i \in \{1, \cdots, d\}$ of $a(z)$ and each component $j \in \{1, \cdots, d\}$ of $z$, define the set $\mathcal{S}^{ij} = \{\theta \mid \nabla_j a_i(z + b) = \nabla_j a_i(z) + \nabla_j^2 a_i(\theta) b, z \in \mathbb{R}^d\}$, where $b$ is a fixed vector in $\mathbb{R}^d$. Each set $\mathcal{S}^{ij}$ has a non-zero Lebesgue measure in $\mathbb{R}^d$.*

With these assumptions, we can show an analogous result to Theorem 2 of Ahuja et al. (2022a).

**Proposition 1.** *If Assumptions 1, and 2 hold, and the data is generated according to equation 4 then the encoder that solves equation 6 identifies true latents on the restricted space up to offsets, i.e. $\hat{z} = z + c$ for all $z \in \mathcal{Z}'_{\text{in}}$, where $c \in \mathbb{R}^d$ is an offset. For all $z \in \mathcal{Z}_{\text{nin}}$, the encoder outputs a constant.*

*Proof.* From Theorem 2, we know that we can achieve identification up to equivariance of $m$ with a restricted domain. For $z \in \mathcal{Z}'_{\text{in}}$ we have,

$$m \circ a = a \circ m \Rightarrow a(z) + \delta = a(z') \tag{7}$$

We now can use the same steps that are used in Ahuja et al. (2022b) and show that $a$ is offset function. The key difference between the setup in Ahuja et al. (2022b) and this is that in that work the identity holds on the entire support of the latent $\mathcal{Z}$ and not a subset $\mathcal{Z}'_{\text{in}}$ defined by the encoder.

Now, for all $z \in \mathcal{Z}_{\text{nin}}$, since $x$ is constant, $\hat{z} = \tilde{f}(x)$ is also a constant. Hence, we get $a$ is offset function of $\mathcal{Z}_{\text{in}}$ and a constant function on $\mathcal{Z}_{\text{nin}}$. □

**Beyond known mechanisms** We can relax the assumption of full knowledge of the offsets, to instead assuming that the learner knows that either some set of (unknown) dense offsets or sparse offsets were applied. Under the same diversity assumptions on the span of the set of offsets, we would have have recovered identification up to either linear or a diagonal / scaling factor respectively, analogous to those in Ahuja et al. (2022b) on the restricted support.

From an algorithmic perspective is, we can understand these results as showing that next frame prediction through mechanism knowledge ($\tilde{g} \circ m \circ f(x) = x'$) and reconstruction $\tilde{g} \circ \tilde{f}$ is sufficient to achieve standard identification results in the injective regions of the function and constant value in other regions. In order to make non-trivial inferences over the latents in the non-injective region, we can leverage constraints that mechanism-induced paths through the occluded region provide.

### 3.2.2 Path based constraints

Finally, we observe that knowledge of the mechanisms makes it possible to go beyond a constant prediction and bound the latents for $z \in \mathcal{Z}_{\text{nin}}$. First note that in Propsition 1, we show that for $\mathcal{Z}_{\text{in}}$, the latents are identified up to a constant, such that $\hat{z} = z + c$. On $\mathcal{Z}_{\text{nin}}$, our predictions are just a constant, $\hat{z} = b$, because the images are constant for all occluded latents $z \in \mathcal{Z}_{\text{nin}}$. But, for any path that includes a point in $\mathcal{Z}_{\text{in}}$ before arriving in the non-injective region, $\mathcal{Z}_{\text{nin}}$, we can estimate the occluded latent using the fact that the path consists of repeated applications of some mechanism from an observable point.

In particular, when the mechanisms are known offsets we can get offset identification for any point in $\mathcal{Z}_{\text{nin}}$ up to a constant offset by leveraging these paths. With know offsets, a path amounts to a sequence of offsets of length $k$, such that $\delta^{(k)} = \sum_{t=1}^{k} \delta_{t,i(t)}$ where each $\delta_{t,i(t)} \in \Delta$ is the offset applied after $t$ steps. But then for any $z_k \in \mathcal{Z}_{\text{nin}}$, given a path $\delta^{(k)}$ from $z_0 \in \mathcal{Z}_{\text{in}}$ to $z_k$, we can recover $z_k$ up to an offset, by letting $\hat{z}_k = \hat{z}_0 + \delta^{(k)} = z_0 + c + \delta^{(k)} = z_k + c$.

## 4 Discussion

This paper presented a hierarchy of problems that relax the injective assumption, which are characterized by the effect of views and the observable effects of experiments. We demonstrated that it is possible to make non-trivial identification claims by extending Ahuja et al. (2022a)'s approach to the occluded case. An important future direction is characterizing Level 3 of the hierarchy and expand the occluded results.

## REFERENCES


Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=g5ynW-jMq4M.

Kartik Ahuja, Jason Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In *Neural Information Processing Systems*, 2022b.

Kartik Ahuja, Divyat Mahajan, Vasilis Syrgkanis, and Ioannis Mitliagkas. Towards efficient representation identification in supervised learning. In *First Conference on Causal Learning and Reasoning*, 2022c. URL https://openreview.net/forum?id=7UwoSnMDXWE.

Anonymous. *Concurrent Submission*. 2022.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. 2022.

Nora Mills Boyd and James Bogen. Theory and Observation in Science. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*, 2022.

Peter Godfrey-Smith. Theory and reality. In *Theory and Reality*. University of Chicago Press, 2021.

Luigi Gresele, Paul K Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Uncertainty in Artificial Intelligence*, pp. 217–227. PMLR, 2020.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

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

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

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217. PMLR, 2020a.

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

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. URL https://openreview.net/forum?id=EbIDjBynYJ8.

Sebastien Lachapelle, Pau Rodriguez, 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. URL https://openreview.net/forum?id=dHsFFekd_-o.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Citris: Causal identifiability from temporal intervened sequences. *arXiv preprint arXiv:2202.03169*, 2022.

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

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Gemma E Moran, Dhanya Sridhar, Yixin Wang, and David M Blei. Identifiable variational autoencoders via sparse decoding. *Transactions of Machine Learning Research*, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

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.

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, 2021.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in latent variable models: Characterization and strong identifiability, 2022. URL https://arxiv.org/abs/2206.00801.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*, 2021.

## A Proof of Theorem 1

*Proof.* Define the set of all encoders that solve the identity in equation 5 as $\mathcal{G}_{\mathsf{id}}$.

We first make an observation about the right inverses below. For all $x \in \mathcal{X}$ we have

$$\tilde{g} \circ \tilde{f}(x) = x \tag{8}$$

Suppose we take a $z \in \mathcal{Z}'$. By definition of $\mathcal{Z}'$, we can find an $x \in \mathcal{X}$ such that $z = \tilde{f}(x)$. We exploit this in the simplifcation below.

$$\tilde{f} \circ \tilde{g}(z) = \tilde{f} \circ \tilde{g} \circ \tilde{f}(x) = \tilde{f}(x) = z \tag{9}$$

Therefore, $\tilde{f} \circ \tilde{g} = \mathsf{I}$, where $\mathsf{I} : \mathcal{Z}' \to \mathcal{Z}'$ is the identity map on the image of the encoder.

From equation 8, we know that for each $x \in \mathcal{X}$, there exists a $z = \tilde{f}(x) \in \mathcal{Z}'$ such that $\tilde{g}(z) = x$.

For all $x \in \mathcal{X}$ because $\tilde{f}$ and $\tilde{g}$ satisfy equation 5, we know,

$$\begin{aligned} \tilde{g} \circ m \circ \tilde{f}(x) &= g \circ m \circ g^{-1}(x) \\ g^{-1} \circ \tilde{g} \circ m \circ \tilde{f}(x) &= m \circ g^{-1}(x) \\ a \circ m \circ \tilde{f}(x) &= m \circ g^{-1}(x) \end{aligned} \tag{10}$$

In the above, we use the fact that for each $x \in \mathcal{X}$, there exists exactly one $z \in \mathcal{Z}'$ such that $\tilde{g}(z) = x$ and get

$$a \circ m \circ \tilde{f} \circ \tilde{g}(z) = m \circ g^{-1} \circ \tilde{g}(z) \tag{11}$$

We showed above that for each $z \in \mathcal{Z}'$, $\tilde{f} \circ \tilde{g}(z) = z$. Therefore, we finally get for all $z \in \mathcal{Z}'$

$$a \circ m(z) = m \circ a(z) \tag{12}$$

Therefore, the equivariance identity has to hold but over a subset of the domain. In other words, we can identify the true latents up to equivariances of $m'$ defined only on this restricted domain. $\square$