# Towards Identifiable Latent Additive Noise Models

**Yuhang Liu**[1,2], **Zhen Zhang**[1,2], **Dong Gong**[3], **Erdun Gao**[1,2], **Biwei Huang**[5], **Mingming Gong**[4], **Anton van den Hengel**[1,2], **Kun Zhang**[6], **Javen Qinfeng Shi**[1,2]

[1] Responsible AI Research Centre, Australia
[2] Australian Institute for Machine Learning, Adelaide University, Australia
[3]School of Computer Science and Engineering, The University of New South Wales, Australia
[4]School of Mathematics and Statistics, The University of Melbourne, Australia
[5]Halicioğlu Data Science Institute, University of California San Diego, USA
[6]Department of Philosophy, Carnegie Mellon University, USA

yuhang.liu01@adelaide.edu.au

## Abstract

Causal representation learning (CRL) offers the promise of uncovering the underlying causal model by which observed data was generated, but the practical applicability of existing methods remains limited by the strong assumptions required for identifiability and by challenges in applying them to real-world settings. Most current approaches are applicable only to relatively restrictive model classes, such as linear or polynomial models, which limits their flexibility and robustness in practice. One promising approach to this problem seeks to address these issues by leveraging changes in causal influences among latent variables. In this vein we propose a more general and relaxed framework than typically applied, formulated by imposing constraints on the function classes applied. Within this framework, we establish partial identifiability results under certain conditions, including scenarios where only a subset of causal influences change. We then extend our analysis to a broader class of latent post-nonlinear models. Building on these theoretical insights, we develop a flexible method for learning latent causal representations. We demonstrate the effectiveness of our approach on synthetic and semi-synthetic datasets, and further showcase its applicability in a case study on human motion analysis, a complex real-world domain that also highlights the potential to broaden the practical reach of identifiable CRL models.

## 1 Introduction

Causal representation learning (CRL) aims to recover the latent variables and causal structures that give rise to high-dimensional observations, offering a principled perspective on modeling complex systems [56, 3]. By explicitly capturing underlying generative mechanisms, CRL enhances interpretability and supports robust generalization across environments, particularly under distribution shifts induced by interventions [51, 50]. Such capabilities make CRL particularly valuable in domains such as reinforcement learning and self-supervised learning, where uncovering latent causal factors can facilitate more general yet effective representations and enable more effective planning [47, 83]. While CRL holds clear advantages over correlation-based methods, yielding representations that are more robust and transferable, it remains difficult to realize these benefits in practice. This is largely due to the strong assumptions required for identifiability and the challenges associated with deploying existing models in realistic environments [6, 79, 20].

Leveraging changes in causal influences among latent variables has emerged as a promising strategy to enhance identifiability and improve estimation quality. Recent work in this direction has focused on developing theoretically grounded frameworks for identifiability, alongside practical methods

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

tailored for real-world applications [61, 45, 10, 43, 71, 9, 3, 66, 70]. Underlying these methods is the core intuition that changes in causal influences introduce asymmetries, e.g., pre- and post-change behaviors, into the system, which provides valuable signals that help achieve identifiability. Based on this idea, prior works have established various identifiability results for restricted function classes over latent variable models, including, but not limited to linear Gaussian models [45, 10], linear additive noise models [61, 15, 26], and polynomial models [43]. See Sec. B for more related work.

**Remaining Challenges.** Despite this progress, several limitations pose difficulties for broader applicability. Theoretically, many existing identifiability results rely on relatively strong assumptions, such as specific functional forms as mentioned above, which may not hold in complex or poorly understood real-world systems. Empirically, due to the relatively strong assumption required for changes in causal influences in latent space, real-world applications that conform to these assumptions remain limited. As a result, many identifiability results are primarily evaluated on synthetic datasets, typically Causal3DIdent [71]. Although there have been promising attempts to adapt these identifiability results into effective methods for real-world data, particularly in biological data [61, 84] and climate data [78], further efforts are needed to extend CRL to a wider range of real-world scenarios.

**Contributions.** In this work, we aim to advance the study of leveraging changes in causal influences by contributing to both theory and practical applications. *On the theoretical side*, we introduce a nonparametric condition that characterizes changes in causal influences between latent causal variables. Under this condition and standard assumptions from nonlinear ICA [24, 25, 29], we show that general latent additive noise models can be identified up to permutation and scaling. We further extend this result to a more realistic setting where the proposed condition holds for only a subset of the latent causal variables, resulting in partial identifiability. We also generalize these results to latent post-nonlinear models, which include additive noise models as a special case. *On the practical side*, we explore a novel real-world application: learning causal representations from human motion data. In this setting, the underlying latent causal system can be interpreted as dynamic motor control modules that govern human motion across different tasks [19, 17, 64]. This opens a new potential application for applying CRL to complex human-centered data.

## 2 Latent Additive Noise Models with Change in Causal Influences

We consider a general class of latent additive noise models, where the observed data $\mathbf{x}$ is generated from a set of latent causal variables $\mathbf{z} \in \mathbb{R}^\ell$. These latent causal variables are causally influenced by latent noise terms $\mathbf{n} \in \mathbb{R}^\ell$, and their causal relationships are represented by a directed acyclic graph (DAG). Importantly, we do not assume a fixed graph structure over the latent causal variables, allowing for flexible modeling of their dependencies and applicability across different settings. To account for the change of causal influences between latent causal variables $\mathbf{z}$, which may arise from environmental or contextual factors, we introduce a surrogate variable $\mathbf{u}$. This variable plays a central role in capturing how changes in external conditions are reflected in the observed data $\mathbf{x} \in \mathbb{R}^\ell$. The interpretation of $\mathbf{u}$ is application-dependent. In domain adaptation or generalization tasks, it may represent environmental factors that vary across domains. In time series forecasting [48], $\mathbf{u}$ can capture temporal indices reflecting evolving trends. In remote sensing [54], it may encode geographic attributes such as longitude and latitude that influence observations.

**Data Generation Process.** More specifically, we parameterize the latent causal generative models by assuming $\mathbf{n}$ follows an exponential family distribution given $\mathbf{u}$, and $\mathbf{z}$ and $\mathbf{x}$ are generated as follows:

$$p_{(\mathbf{T},\boldsymbol{\eta})}(\mathbf{n} \mid \mathbf{u}) := \prod_i \frac{\exp\Big(\sum_j T_{i,j}(n_i)\,\eta_{i,j}(\mathbf{u})\Big)}{Z_i(\mathbf{u})}, \quad (1a) \qquad z_i := \mathrm{g}_i^{\mathbf{u}}(\mathrm{pa}_i) + n_i, \quad (1b) \qquad \mathbf{x} := \mathbf{f}(\mathbf{z}). \quad (1c)$$

- In Eq. (1a): $Z_i(\mathbf{u})$ is the normalizing constant, and $T_{i,j}(n_i)$ is the sufficient statistic, with its natural parameter $\eta_{i,j}(\mathbf{u})$. We assume a two-parameter exponential family, following [60].
- In Eq. (1b): The term $\mathrm{g}_i^{\mathbf{u}}(\mathrm{pa}_i)$ shows how $\mathbf{u}$ influences the mapping $\mathrm{g}_i$ of parents $\mathrm{pa}_i$ to $z_i$, e.g., if $\mathrm{g}_i$ is modeled by a multilayer perceptron (MLP), $\mathbf{u}$ adjusts the weights of the MLP.
- In Eq. (1c): $\mathbf{f}$ represents a nonlinear mapping from $\mathbf{z}$ to $\mathbf{x}$.

The surrogate variable $\mathbf{u}$ captures distributional shifts in the latent noise variables, as reflected by changes in the natural parameters across $\mathbf{u}$ in Eq. (1a). This enables the adaptation of recent progress

in nonlinear ICA. More importantly, $\mathbf{u}$ also models changes in the underlying causal influences from parent variables to each latent causal variable. In particular, as shown in Eq. (1b), $\mathbf{u}$ modulates the functional form $\mathrm{g}_i^{\mathbf{u}}$, effectively characterizing how the parents $\mathrm{pa}_i$ influence $z_i$. By combining parent removal in $\mathrm{g}_i^{\mathbf{u}}$ with sufficiently diverse changes in $\eta_i(\mathbf{u})$, we can obtain a sufficient condition for identifying $\mathbf{z}$ (see Assumption (iii)).

## 3 Identifiability Results of Latent Additive Noise Models

In this section, we analyze identifiability in latent additive noise models by leveraging changes in causal influences across environments. We first present the complete identifiability result in Section 3.1, then extend to partial identifiability result in Section 3.2, addressing more general and realistic scenarios in which the proposed conditions hold for only a subset of latent variables. Finally, we generalize both complete and partial results to latent post-nonlinear models in Section 3.3.

### 3.1 Complete Identifiability Result

We explicitly introduce the surrogate variable $\mathbf{u}$ as described in the data generation process defined by Eqs. (1a)–(1c). This mechanism allows us to formulate key identifiability conditions in terms of $\mathbf{u}$. We now present the main identifiability result as follows:

**Theorem 3.1.** *Suppose latent causal variables $\mathbf{z}$ and the observed variable $\mathbf{x}$ follow the causal data generative models defined in Eqs. (1a)–(1c). Consider an estimated model from the same model class that matches $p(\mathbf{x} \mid \mathbf{u})$. The same labeled environment triples below are available to both models, while their corresponding latent target identities need not be known or the same. Assume the following holds:*

*(i) The function $\mathbf{f}$ is smooth and invertible.*

*(ii) For each $i$, let $\mathbf{u}_i$, $\mathbf{u}_i'$, and $\mathbf{u}_i''$ be the three environments. The matrix*

$$\mathbf{L}_i = \Big(\boldsymbol{\eta}_i(\mathbf{u}_i') - \boldsymbol{\eta}_i(\mathbf{u}_i),\ \boldsymbol{\eta}_i(\mathbf{u}_i'') - \boldsymbol{\eta}_i(\mathbf{u}_i)\Big) \tag{2}$$

*of size $2 \times 2$ is invertible, where $\boldsymbol{\eta}_i(\mathbf{u}) = (\eta_{i,1}(\mathbf{u}), \eta_{i,2}(\mathbf{u}))^\top$.*

*(iii) For each $i$, the three environments $\mathbf{u}_i$, $\mathbf{u}_i'$, and $\mathbf{u}_i''$ specified in Assumption (ii) satisfy, for every parent node $z_j \in \mathrm{pa}_i$ and all parent values,*

$$\frac{\partial \mathrm{g}_i^{\mathbf{u}}(\mathrm{pa}_i)}{\partial z_j} = 0, \qquad \mathbf{u} \in \{\mathbf{u}_i, \mathbf{u}_i', \mathbf{u}_i''\}. \tag{3}$$

*The parent domain is connected. For every $k \neq i$, the conditional distribution of $z_k$ given its parents remains unchanged across these three environments.*

*Then each true latent variable $z_i$ is affinely related to exactly one estimated latent variable $\hat{z}_{\pi(i)}$, as*

$$z_i = s_i \hat{z}_{\pi(i)} + c_i, \tag{4}$$

*for constants $s_i \neq 0$ and $c_i$, where $\pi$ is a permutation and all $\hat{z}_j$ are learned by matching the true conditional distribution $p(\mathbf{x} \mid \mathbf{u})$.*

*Proof.* A proof sketch and the full proof are provided in Appendix D. □

Assumptions (i)-(ii) are originally developed by nonlinear ICA [24, 25, 29, 60], and have also been adopted in several recent works on CRL (with different forms) [45, 43, 86, 49]. Basically, assumption (i) requires that information is preserved (i.e., no loss), while assumption (ii) requires that the variable $\mathbf{u}$ exerts sufficiently strong and diverse influences on the distributions of the independent noise variables. We here utilize these assumptions considering that these assumptions have been verified to be practicable in diverse real-world application scenarios [31, 76, 72].

Assumption (iii) provides the parent-removal anchor. Loosely speaking, it requires that, for each latent variable, there exist three environments in which the causal influences from its parents are

removed. This is achieved by requiring the gradient of $g_i^{\mathbf{u}}$ with respect to its parent variables to vanish in each of the three environments, so that $g_i^{\mathbf{u}_i}$ is constant in its parent arguments. Without such an anchor, an environment-invariant parent effect can create ambiguity, as illustrated in Remark 3.2.

Assumption (iii), for instance, could arise in the analysis of cell imaging data (i.e., $\mathbf{x}$), where various batches of cells are exposed to different small-molecule compounds (i.e., $\mathbf{u}$). each latent variable (i.e., $z_i$) represents the concentration level of a distinct group of proteins, with protein-protein interactions (e.g., causal influences among $z_i$) playing a significant role [14]. Research has revealed that the mechanisms of action of small molecules exhibit variations in selectivity [57], which can profoundly affect protein-protein interactions (i.e., $g_i^{\mathbf{u}}$). The assumption (iii) requires the existence of a specific $\mathbf{u} = \mathbf{u}_i$, such that the original causal influences can be disconnected. This parallels cases where small molecule compounds disrupt or inhibit protein-protein interactions (PPIs), effectively causing these interactions to cease [5]. Such molecules are commonly referred to as inhibitors of PPIs. Developing small molecule inhibitors for PPIs is a key focus in drug discovery [46, 8]. Additionally, gene editing technologies like CRISPR/Cas9 can effectively 'knock out' a protein or gene, leading to complete inhibition. Similarly, receptor antagonists can achieve full inhibition by completely blocking the activity of a receptor.

Assumption (iii) is one of the main structural conditions in our analysis. It specifies a sufficient type of change in causal influences that facilitates identifiability.

**Remark 3.2** (Types of Changes in Causal Influences That Facilitate Identifiability)**.** *Not all changes in causal influences lead to identifiability. Assumption (iii) specifies the types of changes in causal influences among latent causal variables contributing to identifiability.*

To clarify this point, consider the following example.

**Example 3.3.** Let $z_2 := \text{MLP}^{\mathbf{u}}(z_1) + n_2$, where $\text{MLP}^{\mathbf{u}}(z_1)$ can be decomposed as $\text{MLP}^{\mathbf{u}}(z_1) = \text{MLP}_1^{\mathbf{u}}(z_1) + \text{MLP}_2(z_1)$, with $\text{MLP}_1^{\mathbf{u}}(z_1)$ being the $\mathbf{u}$-dependent component and $\text{MLP}_2(z_1)$ being a $z_1$-dependent term invariant across $\mathbf{u}$. Both $\text{MLP}^{\mathbf{u}}$ and $\text{MLP}_1^{\mathbf{u}}$ belong to the same function class.

In this example, if assumption (iii) is violated, $z_2$ may become unidentifiable. While the causal influence from $z_1$ to $z_2$ changes across $\mathbf{u}$ due to the $\mathbf{u}$-dependent term $\text{MLP}_1^{\mathbf{u}}(z_1)$, the invariant term $\text{MLP}_2(z_1)$ induces a invariant causal link between $z_1$ and $z_2$ across $\mathbf{u}$, which leads to unidentifiable result. Specifically, the invariant $\text{MLP}_2(z_1)$ can be absorbed into the generative mapping $\mathbf{f}$, resulting in an alternative representation $z_2' := \text{MLP}_1^{\mathbf{u}}(z_1) + n_2$, which would generate the same observational data. That is, the original data generation process can be equivalently written as $\mathbf{x} = \mathbf{f}(z_1, z_2) = \mathbf{f} \circ \mathbf{f}_1(z_1, z_2')$, where $\mathbf{f}_1(z_1, z_2') = [z_1, z_2' + \text{MLP}_2(z_1)]$. Consequently, the model remains unidentifiable. The reason of this unidentifiability is the presence of the invariant $\text{MLP}_2(z_1)$, which maintains a constant causal influence of $z_1$ on $z_2$ across $\mathbf{u}$. Assumption (iii) mitigates this issue by eliminating such invariant component. It does so by constraining the function class satisfies: $\partial \text{MLP}^{\mathbf{u}=\mathbf{u}_2}(z_1)/\partial z_1 = 0$ and $\partial \text{MLP}_1^{\mathbf{u}=\mathbf{u}_2}(z_1)/\partial z_1 = 0$. As a result, $\text{MLP}_2(z_1)/\partial z_1 = 0$, which implies that $\text{MLP}_2(z_1)$ must be a constant, removing $z_1$-dependent term and thus ensuring identifiability. Assumption (iii) is closely related to the notion of perfect interventions studied in prior work [39, 38, 73, 9], since it requires environments in which the incoming causal influences on a latent variable are removed. Some existing methods assume known intervention targets, while others consider weakly supervised or unknown-target settings. In our setting, the grouping of environments into calibration blocks is observed, but the latent target of each block need not be known.

**Remark 3.4** (Implicit Alignment of Calibration Blocks)**.** *The calibration blocks align latent variables across admissible representations without requiring their target identities. Within each block, the non-target conditional distributions remain unchanged, so the environment density ratios depend only on the parent-removed target variable.*

To clarify this point, consider the following example.

**Example 3.5.** Let $z_1 = n_1$ and $z_2 = g^{\mathbf{u}}(z_1) + n_2$. Suppose three observed environments form a calibration block for $z_2$. Within this block, the same parent-removing structural function is used for $z_2$, the conditional distribution of $z_1$ remains unchanged, and the noise parameters of $z_2$ change in two linearly independent directions. The two corresponding density ratios therefore depend only on $z_2$ and provide two independent contrasts of its sufficient statistics. In any alternative representation satisfying the same block structure, these ratios depend on a single estimated latent variable $\hat{z}_j$. Under the fixed two-parameter noise family, matching the two ratios identifies $z_2$ with $\hat{z}_j$ up to the

affine ambiguity in Theorem 3.1. Thus, the latent target identity need not be provided, although the environment block is observed.

### 3.2 Partial Identifiability Result

In practice, satisfying assumption (iii) for every causal influence from parent nodes to each child node can be challenging. When this assumption is violated for some nodes, full identifiability may not be achievable. Nevertheless, we can still derive partial identifiability results, as detailed below:

**Theorem 3.6.** *Suppose latent causal variables* $\mathbf{z}$ *and the observed variable* $\mathbf{x}$ *follow the causal generative models defined in Eqs.* (1a) - (1c)*, and the assumptions (i)-(ii) are satisfied, for each* $z_i$*,*

*(a) if condition (iii) is satisfied, then the true* $z_i$ *is related to the recovered one* $\hat{z}_j$*, obtained by matching the true marginal data distribution* $p(\mathbf{x}|\mathbf{u})$*, by the following relationship:* $z_i = s_j\hat{z}_j + c_j$*, where* $s_j$ *denotes scaling,* $c_j$ *denotes a constant,*

*(b) if condition (iii) is not satisfied, then* $z_i$ *is not identifiable without additional assumptions.*

*Proof.* A proof sketch and the full proof are provided in Appendix E. □

**Remark 3.7** (Identifiability does not require identifiable parents)**.** A latent variable $z_i$ covered by a calibration block can be identifiable even when some of its parents are not. Within this block, the parent effects on $z_i$ are removed and the non-target conditional distributions cancel from the environment density ratios. The proof therefore identifies $z_i$ directly and does not require its parents or the complete noise vector to be identified first.

**Remark 3.8** (Identifiable subset)**.** Theorem 3.6 separates the latent coordinates covered by calibration blocks from the remaining coordinates. Within this admissible comparison class, the remaining non-root coordinates admit the parent-dependent ambiguity constructed in Appendix E.

**Summary** The identifiability argument uses density ratios within each calibration block. The non-target conditional distributions cancel, while two independent changes in the target noise parameters identify the parent-removed target variable. When the calibration blocks cover all latent variables, Theorem 3.1 gives complete identifiability. When they cover only a subset, Theorem 3.6 identifies that subset and leaves the remaining variables unconstrained.

### 3.3 Extension to Latent Post-Nonlinear Models

While latent additive noise models, as defined in Eq. (1b), are general, their capacities are still limited, e.g., requiring additive noise. In this section, we generalize latent additive noise models to latent post-nonlinear models [85, 65, 28], which generally offer more powerful expressive capabilities than latent additive noise models. To this end, we replace Eq. (1b) by the following:

$$\bar{z}_i := \bar{\mathrm{g}}_i(z_i) = \bar{\mathrm{g}}_i(\mathrm{g}_i^{\mathbf{u}}(\mathrm{pa}_i) + n_i), \tag{5}$$

where $\bar{\mathrm{g}}_i$ denotes a invertible post-nonlinear mapping. It includes the latent additive noise models Eq. (1b) as a special case in which the nonlinear distortion $\bar{\mathrm{g}}_i$ does not exist. Based on this, we can identify $\bar{\mathbf{z}}$ up to component-wise invertible nonlinear transformation as follows:

**Corollary 3.9.** *Suppose latent causal variables* $\mathbf{z}$ *and the observed variable* $\mathbf{x}$ *follow the causal generative models defined in Eqs.* (1a)*,* (5) *and* (1c)*. Assume that conditions (i) - (iii) in Theorem 3.1 hold, then each true latent variable* $\bar{z}_i$ *is related to exactly one estimated latent variable* $\hat{\bar{z}}_j$*, which is learned by matching the true marginal data distribution* $p(\mathbf{x}|\mathbf{u})$*, by the following relationship:* $\bar{z}_i = M_j(\hat{\bar{z}}_j) + c_j$*, where* $M_j$ *and* $c_j$ *denote a invertible nonlinear mapping and a constant, respectively.*

*Proof.* A proof sketch and the full proof are provided in Appendix F. □

Similar to Theorem 3.6, we have the following partial identifiability result:

**Corollary 3.10.** *Suppose latent causal variables* $\mathbf{z}$ *and the observed variable* $\mathbf{x}$ *follow the causal generative models defined in Eqs.* (1a)*,* (5) *and* (1c)*. Under the condition that the assumptions (i)-(ii) are satisfied, for each* $\bar{z}_i$*, (a) if it is a root node or condition (iii) is satisfied, then the true* $\bar{z}_i$ *is related*

*to the recovered one $\hat{\bar{z}}_j$, obtained by matching the true marginal data distribution $p(\mathbf{x}|\mathbf{u})$, by the following relationship: $\bar{z}_i = M_j(\hat{\bar{z}}_j) + c_j$, where $M_j$ denotes a invertible mapping, $c_j$ denotes a constant, (b) if condition (iii) is not satisfied, then $\bar{z}_i$ is unidentifiable.*

*Proof.* See Appendix G. □

**Remark 3.11** (Sharing Properties)**.** Corollary 3.10 establishes that the properties outlined in Theorem 3.6, including remarks 3.7 to 3.8, remain applicable in latent post-nonlinear causal models.

## 4 Learning Latent Additive Noise Models

In this section, we translate our theoretical findings into a novel method for learning latent causal models. See Appendix J for the connection between theory and method. Our primary focus is on learning additive noise models, as extending the method to latent post-nonlinear models is straightforward, simply involving the utilization of invertible nonlinear mappings. Without loss of generality, due to permutation indeterminacy in latent space, we can naturally enforce a causal order $z_1 \succ z_2 \succ ..., \succ z_\ell$ [61, 45]. This does not imply that we require knowledge of the true causal order, refer to Appendix H for more details. As a result, we formulate a prior model as follows:

$$p(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^{\ell} p(z_i|\mathbf{z}_{<i}, \mathbf{m}_i(\mathbf{u}), \mathbf{u}) = \prod_{i=1}^{\ell} \mathcal{N}(\mu_{z_i}(\mathbf{z}_{<i} \odot \mathbf{m}_i(\mathbf{u}), \mathbf{u}), \delta^2_{z_i}(\mathbf{u})), \tag{6}$$

where we focus on latent Gaussian noise variables, to satisfy the exponential family assumption in Eq. (1a) and naturally allow the implementation of the reparameterization trick. Moreover, we introduce additional vectors $\mathbf{m}_i(\mathbf{u})$, by enforcing sparsity on $\mathbf{m}_i(\mathbf{u})$ and the component-wise product $\odot$, which facilitates learning the latent causal graph structure. In our implementation, we impose the L1 norm, though other methods may also be flexible, e.g., sparsity priors [13, 40]. We employ a variational posterior to approximate the true posterior $p(\mathbf{z}|\mathbf{x}, \mathbf{u})$:

$$q(\mathbf{z}|\mathbf{u}, \mathbf{x}) = \prod_{i=1}^{\ell} q(z_i|\mathbf{z}_{<i}, \mathbf{m}_i(\mathbf{u}), \mathbf{u}, \mathbf{x}) = \prod_{i=1}^{\ell} \mathcal{N}(\mu_{z_i}(\mathbf{z}_{<i} \odot \mathbf{m}_i(\mathbf{u}), \mathbf{u}, \mathbf{x}), \delta^2_{z_i}(\mathbf{z}_{<i} \odot \mathbf{m}_i(\mathbf{u}), \mathbf{u}, \mathbf{x})), \tag{7}$$

where the variational posterior shares the same parameter $\mathbf{m}_i$ to limit both the prior and the variational posterior, maintaining the same latent causal graph structure. In addition, we enforce a Gaussian posterior conditioned on the parent nodes, to align with the prior model and model assumption in Eqs. (1a) and (1b). Finally, we arrive at the objective:

$$\max \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}(\log p(\mathbf{x}|\mathbf{z})) - D_{\mathrm{KL}}(q(\mathbf{z}|\mathbf{x}, \mathbf{u})||p(\mathbf{z}|\mathbf{u})) - \gamma \sum\nolimits_i \|\mathbf{m}_i(\mathbf{u})\|_1^1, \tag{8}$$

where $D_{\mathrm{KL}}$ denotes the KL divergence, $\gamma$ denotes a hyperparameters to control the sparsity of latent causal structure. The objective is known as the evidence lower bound (ELBO), which serves as a lower bound of the log-likelihood. Under certain conditions, it can match the true data distribution, which is one of the requirements for our identifiability guarantee in Theorem 3.1. Moreover, the variational estimator is consistent in the sense that, as the number of data samples grows, the learned variational posterior converges to the true posterior [1] , ensuring reliable recovery of the underlying latent variables. Implementation details can be found in Appendix K.

## 5 Experiments

**Synthetic Data** We first conduct experiments on synthetic data, generated by the following process: we divide latent noise variables into $M$ segments, where each segment corresponds to one value of $\mathbf{u}$ as the segment label. Within each segment, the location and scale parameters are respectively sampled from uniform priors. After generating latent noise variables, we generate latent causal variables, and finally obtain the observed data samples by an invertible nonlinear mapping. See App. I for details.

[1] More strictly, this holds under the assumption that the variational posterior has sufficient capacity to represent the true posterior, and in the limit of the optimal ELBO solution [29]. Note that the sparsity regularization term with weight $\gamma$ may introduce bias, therefore, strict consistency holds when $\gamma$ is sufficiently small, such that the regularized solution remains close to the unregularized ELBO optimum.

We evaluate our proposed method, implemented with multilayer perceptrons (MLPs) and hence referred to as MLPs, to model the causal relations among latent causal variables, against established models: vanilla VAE [30], $\beta$-VAE [21], identifiable VAE (iVAE) [29], and latent polynomial models (Polynomials) [43]. Notably, the iVAE demonstrates the capability to identify true independent noise variables, subject to certain conditions, with permutation and scaling. Polynomials, while sharing similar assumptions with our proposed method, are prone to certain limitations. Specifically, they may suffer from numerical instability and face challenges due to the exponential growth in the number of terms. While the $\beta$-VAE is popular in disentanglement tasks due to its emphasis on independence among recovered variables, it lacks robust theoretical backing. Our evaluation focuses on two metrics: the Mean of the Pearson Correlation Coefficient (MPC) to assess performance, and the Structural Hamming Distance (SHD) to gauge the accuracy of the latent causal graphs. The result for iVAE is obtained by applying the method from [22] to the latent variables estimated by iVAE. Figure 1 illustrates the comparative performances of various methods, e.g., VAE and iVAE, across different models, e.g., models with different dimensions of latent variables. Based on MPC, the proposed method demonstrates satisfactory results, thereby supporting our identifiability claims. Additionally, Figure 2 presents how the proposed method performs when condition (iii) is not met. It is evident that condition (iii) is a sufficient condition characterizing the types of distribution shifts for identifiability in the context of latent additive noise models. These empirical findings align with our partial identifiability results. More results on high-dimensional synthetic image data can be found in Appendix L.

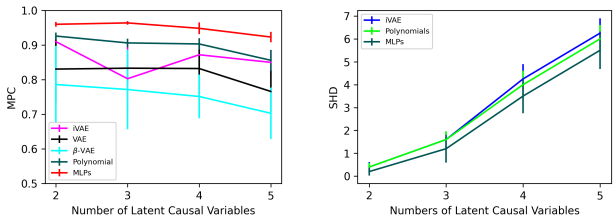



Figure 1: Performance comparison under latent additive Gaussian noise. Left: MPC scores for different methods, where the proposed MLPs method achieves the best performance, supporting our theoretical results. Right: SHD scores of the proposed method, Polynimals [43], and iVAE combined with the method from [22].

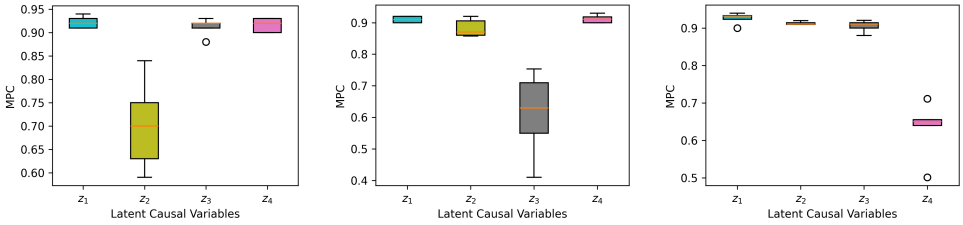
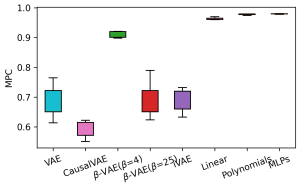


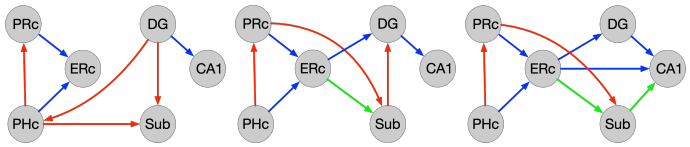

Figure 2: Performance of the proposed method under scenarios where condition (iii) is not satisfied regarding the causal influence of $z_1 \to z_2$ (consequently, $z_2 \to z_3$, and $z_3 \to z_4$). The results are in agreement with partial identifiability in Theorem 3.6, i.e., roughly speaking, latent variables that satisfy Condition (iii) are identifiable, while those that do not are not identifiable.

**Semi-Synthetic fMRI Data** Building on the works in [45, 43], we extended the application of the proposed method to the fMRI hippocampus dataset [34]. This dataset comprises signals from six distinct brain regions: perirhinal cortex (PRC), parahippocampal cortex (PHC), entorhinal cortex (ERC), subiculum (Sub), CA1, and CA3/Dentate Gyrus (DG). These signals, recorded during resting states, span 84 consecutive days from a single individual. Each day's data contributes to an 84-dimensional vector, e.g., $\mathbf{u}$. Our focus is on uncovering latent causal variables, therefore, we treat these six brain signals as such. Specifically, we assume that they undergo a random nonlinear mapping into the observable space, after which suitable methods can be applied to recover them.

Figure 3 presents the comparative results yielded by the proposed method alongside various other methods. Notably, the VAE, $\beta$-VAE, and iVAE models presume the independence of latent variables, rendering them incapable of discerning the underlying latent causal structure. Conversely, other methods, including latent linear models, latent polynomials, and latent MLPs, are able to accurately recover the latent causal structure with guarantees. We also include CausalVAE [77] as a baseline. Among these, the MLP models outperform the others in terms of MPC. In the study by [43], it is noted that linear relationships among the examined signals tend to be more prominent than nonlinear ones. This observation might lead to the presumption that linear models would be effective. However,




(a) MPC scores.




(b1) Linear (b2) Polynomials (b3) MLPs

Figure 3: (a) MPC scores achieved by different methods. Notably, the proposed MLPs achieve an outstanding average MPC of 0.981, outperforming polynomials (0.977) and linear models (0.965). (b) Recovered latent causal structures using (b1) latent linear models, (b2) latent polynomials, and (b3) latent MLPs. Results for linear models and polynomials are sourced from [43]. Blue edges align with known anatomical connectivity, red edges violate anatomical constraints, and green edges are reversed directions.

this is not necessarily the case, as these models can still yield suboptimal outcomes. In contrast, MLPs demonstrate superior performance in term of MPC, particularly when compared to polynomial models, which are prone to instability and exponential growth issues. The effectiveness of MLPs is further underscored by their impressive average MPC score of 0.981. It is important to emphasize that while the improvement in MPC over the proposed method (also achieving 0.981) may appear marginal, compared to prior methods such as linear models (MPC 0.965) and polynomial models (MPC 0.977), this seemingly "slight" gain in MPC corresponds to a substantial difference in the recovered graph structures, which is visually illustrated in Figure 3 (b3). Moreover, we found that CausalVAE consistently produces fully connected graphs across all different random seeds, resulting in an SHD of 9.0 ± 0.0. MLP-based model achieves an SHD of 4.75 ± 0.22, outperforming the polynomial model (5.5 ± 0.25) and the linear model (5.0 ± 0.28). Here, we note that although the polynomial model may underperform the linear model on average, in this particular example its ability to capture non-linear relationships allows it to achieve a lower SHD.

To further assess the statistical significance of these results, we additionally perform a comparison based on the negative-control (NC) procedure in Sec. 6 of [52]. For MLP-based model, we obtain SHD = 4.75 (95% Confidence Interval (CI): [4, 5]), compared to the NC baseline with SHD = 9.57 (CI: [6, 13], $p = 0.0085$). For polynomial model, SHD = 5.5 (CI: [5, 6]) versus NC: SHD = 9.49 (CI: [6, 13], $p = 0.024$). For the linear model, SHD = 5 (CI: [4, 6]) versus NC: SHD = 9.47 (CI: [6, 13], $p = 0.02$). These results provide additional evidence for the advantage of our MLP-based model.

**Real Human Motion Dataset** Lastly, we apply the proposed method to a potential real-world application: human motion analysis. The human nervous system adapts flexibly to different motor tasks (e.g., walking, running, or fine hand manipulation). In this setting, the observed data $\mathbf{x}$, i.e., human skeleton poses captured via motion capture, is influenced by different motor tasks $\mathbf{u}$ [12, 82]. It is natural to consider that human motion is governed by a set of underlying latent variables (i.e., $\mathbf{z}$) [55, 63, 64, 19]. Each $z_i$ capture patterns analogous to motor neuron activation dynamics [19]. The interactions among these latent variables capture the coordination dynamics between control modules involved in executing the task. Crucially, previous studies demonstrate that these interactions are task-dependent and reconfigurable (i.e., $g_i^{\mathbf{u}}$) [17, 7, 16, 53]. For example, compared to gait velocity in the single-task condition, the networks associated with gait velocity in the dual-task condition were associated with greater functional connectivity in supplementary motor and prefrontal regions [82]. Dynamic causal modeling was applied for neuronal states of the regions of interest the motor task time series to estimate endogenous and context-dependent effective connectivity [53]. Therefore, modeling human motion with latent variables whose interaction strengths vary across motor tasks offers a biologically plausible framework for capturing the flexibility and task-specific nature of human motor control.



Figure 4: Some sample examples of the data used.

We use the Human3.6M dataset, which provides a diverse set of 17 motion tasks, e.g., discussion, smoking, taking photos, and talking on the phone. Following pre-processing steps from prior works (see Appendix M.1), we construct a filtered subset comprising 7 subjects and 15 motion tasks,

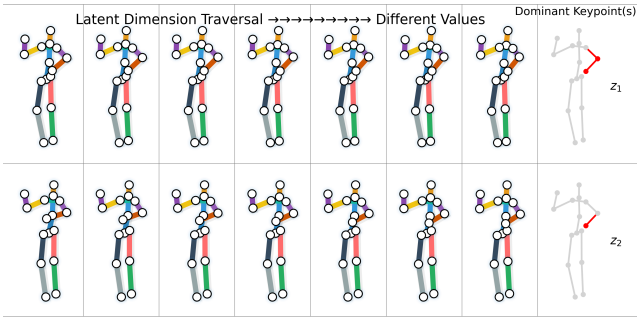



Figure 6: Intervention results on $z_1$ and $z_2$ show that intervening on $z_1$ induces changes from the elbow to the wrist, whereas intervening on $z_2$ affects only the wrist. Together with the corresponding entry in the estimated adjacency matrix in Figure 5, this suggests that the proposed method has learned a causal relationship from elbow to wrist in the right hand.

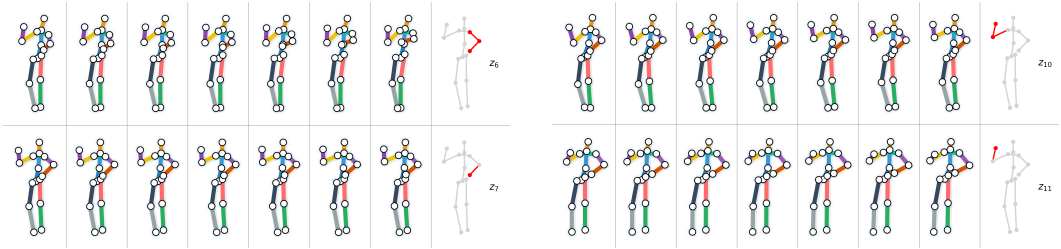



Figure 7: Visualization of interventions on the learned latent variables by the proposed method. The left show that intervention on $z_6$ leads to changes from the shoulder joint to the wrist joint, whereas intervention on $z_7$ affects only the wrist joint, indicating the causal relationship from the shoulder to the wrist. The right reveals that intervention on $z_{10}$ leads to changes from the elbow to the wrist, whereas intervention on $z_{11}$ affects only the wrist, supporting the causal relationship from the elbow to the wrist.

resulting in a total of 105 distinct values of the condition variable $\mathbf{u}$. For each condition, we obtain 1,040 samples. Each sample is represented as a $2 \times 16$ matrix, where 2 denotes the spatial coordinates, and 16 corresponds to skeletal keypoints, such as shoulders, elbows, and knees. Figure 4 shows some samples used in experiments after preprocessing.

Due to the absence of groundtruth, we evaluate the proposed method using an intervention based protocol. Specifically, if the model learns a causal relation $z_i \to z_j$, we expect the following properties to hold: (1) intervention on $z_i$ induces structured changes in the observation space associated with both $z_i$ and its descendants (e.g., $z_j$). We focus on dominant changes in observation space as precise disentanglement remains challenging even in synthetic settings.; (2) intervening on $z_j$ does not induce corresponding changes in the observation space associated with $z_i$; (3) such intervention results are aligned with the presence of an edge between $z_i$ and $z_j$ in the learned causal graph. As shown in Figure 6, intervening on $z_1$ induces changes in both the elbow and wrist regions, while intervening on $z_2$ primarily affects the wrist. Combined with the learned adjacency matrix in Figure 5, which indicates an edge between $z_1$ and $z_2$, this asymmetric intervention pattern suggests a causal relation from elbow to wrist. This finding aligns with established biomechanical principles of intersegmental limb dynamics. In particular, it has been demonstrated that the nervous system predicts and compensates for interaction torques arising from shoulder and elbow movements to adjust wrist muscle activity reflexively, reflecting an internal model of limb dynamics [32]. This agreement provides qualitative evidence that the learned representations capture meaningful causal structure. Similar results are observed for the shoulder-to-wrist causal relation in Figure 7, further supporting the proposed method. See Appendix M for more results.

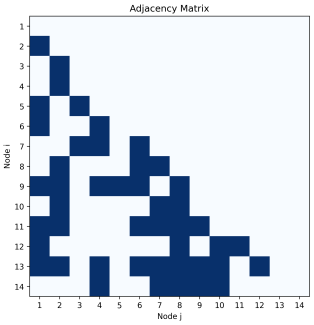



Figure 5: The estimated adjacency matrix by our method.

## 6 Conclusion

In this work, we study the identifiability of latent additive noise models under distribution shifts. We introduce a condition that characterizes the types of changes in causal influences that enable identifiability, and further extend this analysis to a partial identifiability setting, where only a subset of such conditions is satisfied. We also generalize the results to latent post-nonlinear causal models, expanding the applicability of the theoretical framework. Building on these insights, we develop a method for learning latent causal representations, and evaluate it on both synthetic and real-world datasets. In particular, we present an application to human motion data, where the learned representations exhibit structured and interpretable behavior under intervention. We hope that this work provides useful insights toward developing practical approaches for causal representation learning.

## References


[1] Jeffrey Adams, Niels Hansen, and Kun Zhang. Identification of partially observed linear causal models: Graphical conditions for the non-gaussian and heterogeneous cases. In *NeurIPS*, 2021.

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

[3] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *ICML*, pages 372–407. PMLR, 2023.

[4] Animashree Anandkumar, Daniel Hsu, Adel Javanmard, and Sham Kakade. Learning linear bayesian networks with latent variables. In *ICML*, pages 249–257, 2013.

[5] Michelle R Arkin and James A Wells. Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nature reviews Drug discovery*, 3(4):301–317, 2004.

[6] Simon Bing, Tom Hochsprung, Jonas Wahl, Urmi Ninad, and Jakob Runge. Invariance & causal representation learning: Prospects and limitations. *Transactions on Machine Learning Research*, 2024.

[7] Emilio Bizzi, Vincent CK Cheung, Andrea d'Avella, Philippe Saltiel, and Matthew Tresch. Combining modules for movement. *Brain research reviews*, 57(1):125–133, 2008.

[8] Damir Bojadzic, Oscar Alcazar, Jinshui Chen, Sung-Ting Chuang, Jose M Condor Capcha, Lina A Shehadeh, and Peter Buchwald. Small-molecule inhibitors of the coronavirus spike: Ace2 protein–protein interaction as blockers of viral attachment and entry for sars-cov-2. *ACS infectious diseases*, 7(6):1519–1534, 2021.

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

[10] Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. *arXiv preprint arXiv:2306.02235*, 2023.

[11] Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. In *NeurIPS*, 2019.

[12] Germana Cappellini, Yuri P Ivanenko, Richard E Poppele, and Francesco Lacquaniti. Motor patterns in human walking and running. *Journal of neurophysiology*, 95(6):3426–3437, 2006.

[13] Carlos M Carvalho, Nicholas G Polson, and James G Scott. Handling sparsity via the horseshoe. In *Artificial intelligence and statistics*, pages 73–80. PMLR, 2009.

[14] Srinivas Niranj Chandrasekaran, Hugo Ceulemans, Justin D Boyd, and Anne E Carpenter. Image-based profiling for drug discovery: due for a machine-learning upgrade? *Nature Reviews Drug Discovery*, 20(2):145–159, 2021.

[15] Tianyu Chen, Kevin Bello, Francesco Locatello, Bryon Aragam, and Pradeep Ravikumar. Identifying general mechanism shifts in linear causal representations. *Advances in Neural Information Processing Systems*, 37:42405–42429, 2024.

[16] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. *Nature neuroscience*, 6(3):300–308, 2003.

[17] Julien Doyon and Habib Benali. Reorganization and plasticity in the adult brain during learning of motor skills. *Current opinion in neurobiology*, 15(2):161–167, 2005.

[18] Benjamin Frot, Preetam Nandy, and Marloes H Maathuis. Robust causal structure learning with some hidden variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):459–487, 2019.

[19] Juan A Gallego, Matthew G Perich, Lee E Miller, and Sara A Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978–984, 2017.

[20] Juan L Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In *International Conference on Machine Learning*, pages 18143–18169. PMLR, 2025.

[21] I. Higgins, Loïc Matthey, A. Pal, Christopher P. Burgess, Xavier Glorot, M. Botvinick, S. Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[22] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, and B. Schölkopf. Causal discovery from heterogeneous/nonstationary data. *JMLR*, 21(89), 2020.

[23] Biwei Huang, Charles Jia Han Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. *NeurIPS*, 35:5549–5561, 2022.

[24] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *NeurIPS*, 29, 2016.

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

[26] Jikai Jin and Vasilis Syrgkanis. Learning linear causal representations from general environments: Identifiability and intrinsic ambiguity. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[27] Nan Rosemary Ke, Aniket Didolkar, Sarthak Mittal, Anirudh Goyal, Guillaume Lajoie, Stefan Bauer, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Christopher Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021.

[28] Grigor Keropyan, David Strieder, and Mathias Drton. Rank-based causal discovery for post-nonlinear models. In *International Conference on Artificial Intelligence and Statistics*, pages 7849–7870. PMLR, 2023.

[29] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *AISTAS*, pages 2207–2217. PMLR, 2020.

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

[31] Lingjing Kong, Shaoan Xie, Weiran Yao, Yujia Zheng, Guangyi Chen, Petar Stojanov, Victor Akinwande, and Kun Zhang. Partial disentanglement for domain adaptation. In *ICML*, pages 11455–11472. PMLR, 2022.

[32] Isaac L Kurtzer, J Andrew Pruszynski, and Stephen H Scott. Long-latency reflexes of the human arm reflect an internal model of limb dynamics. *Current biology*, 18(6):449–453, 2008.

[33] Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. *arXiv preprint arXiv:2107.10098*, 2021.

[34] Timothy O. Laumann and Russell A. Poldrack, 2015.

[35] Xiu-Chuan Li, James Kwok, Jiaxian Guo, and Tongliang Liu. Causal effect identifiability in the presence of latent confounders without auxiliary variables. In *Forty-third International Conference on Machine Learning*, 2026.

[36] Xiuchuan Li, Jun Wang, and Tongliang Liu. Recovery of causal graph involving latent variables via homologous surrogates. In *International Conference on Learning Representations*, 2025.

[37] Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[39] Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *ICML*, pages 13557–13603. PMLR, 2022.

[40] Yuhang Liu, Wenyong Dong, Lei Zhang, Dong Gong, and Qinfeng Shi. Variational bayesian dropout with a hierarchical prior. In *CVPR*, 2019.

[41] Yuhang Liu, Dong Gong, Yichao Cai, Erdun Gao, Zhen Zhang, Biwei Huang, Mingming Gong, Anton Hengel, and Javen Qinfeng Shi. I predict therefore i am: Is next token prediction enough to learn human-interpretable concepts from data? In *International Conference on Learning Representations*, 2026.

[42] Yuhang Liu, Zhen Zhang, Dong Gong, Erdun Gao, Biwei Huang, Mingming Gong, Anton Hengel, Kun Zhang, and Javen Qinfeng Shi. Beyond dags: A latent partial causal model for multimodal learning. In *International Conference on Learning Representations*, 2026.

[43] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. In *The Twelfth International Conference on Learning Representations*, 2024.

[44] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Latent covariate shift: Unlocking partial identifiability for multi-source domain adaptation. *Transactions on Machine Learning Research*, 2025.

[45] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models, 2026.

[46] Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, and Jianyou Shi. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal transduction and targeted therapy*, 5(1):213, 2020.

[47] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021.

[48] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

[49] Ignavier Ng, Shaoan Xie, Xinshuai Dong, Peter Spirtes, and Kun Zhang. Causal representation learning from general environments under nonparametric mixing. In *International Conference on Artificial Intelligence and Statistics*, 2025.

[50] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

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

[52] Anne Helby Petersen. Are you doing better than random guessing? a call for using negative controls when evaluating causal discovery algorithms. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, pages 3465–3479, 2025.

[53] Anne K Rehme, Simon B Eickhoff, and Christian Grefkes. State-dependent differences between functional and effective connectivity of the human cortical motor system. *Neuroimage*, 67:237–246, 2013.

[54] Marc Rußwurm, Sherrie Wang, Marco Korner, and David Lobell. Meta-learning for few-shot land cover classification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition workshops*, pages 200–201, 2020.

[55] Richard A Schmidt, Timothy D Lee, Carolee Winstein, Gabriele Wulf, and Howard N Zelaznik. *Motor control and learning: A behavioral emphasis*. Human kinetics, 2018.

[56] 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.

[57] Duncan E Scott, Andrew R Bayly, Chris Abell, and John Skidmore. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nature Reviews Drug Discovery*, 15(8):533–550, 2016.

[58] Shohei Shimizu, Patrik O Hoyer, and Aapo Hyvärinen. Estimation of linear non-gaussian acyclic models for latent factors. *Neurocomputing*, 72(7-9):2024–2027, 2009.

[59] Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *JMLR*, 7(2), 2006.

[60] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

[61] Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *International conference on machine learning*, pages 32540–32560. PMLR, 2023.

[62] Stefan G Stark, Joanna Ficek, Francesco Locatello, Ximena Bonilla, Stéphane Chevrier, Franziska Singer, Tumor Profiler Consortium, Gunnar Rätsch, and Kjong-Van Lehmann. Scim: universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36, 12 2020.

[63] Karel Svoboda and Nuo Li. Neural mechanisms of movement planning: motor cortex and beyond. *Current opinion in neurobiology*, 49:33–41, 2018.

[64] Graham W Taylor, Geoffrey E Hinton, and Sam Roweis. Modeling human motion using binary latent variables. *Advances in neural information processing systems*, 19, 2006.

[65] Kento Uemura, Takuya Takagi, Kambayashi Takayuki, Hiroyuki Yoshida, and Shohei Shimizu. A multivariate causal discovery based on post-nonlinear model. In *Conference on Causal Learning and Reasoning*, pages 826–839. PMLR, 2022.

[66] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023.

[67] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning: Linear and general transformations. *Journal of Machine Learning Research*, 26(112):1–90, 2025.

[68] Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. General identifiability and achievability for causal representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 2314–2322. PMLR, 2024.

[69] Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Linear causal representation learning from unknown multi-node interventions. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[70] Julius von Kügelgen, Michel Besserve, Liang Wendong, Luigi Gresele, Armin Kekić, Elias Bareinboim, David Blei, and Bernhard Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. *Advances in Neural Information Processing Systems*, 36, 2023.

[71] 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. In *NeurIPS*, 2021.

[72] Xinyi Wang, Michael Saxon, Jiachen Li, Hongyang Zhang, Kun Zhang, and William Yang Wang. Causal balancing for domain generalization. *arXiv preprint arXiv:2206.05263*, 2022.

[73] Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. *Advances in Neural Information Processing Systems*, 36:32481–32520, 2023.

[74] Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *NeurIPS*, 2020.

[75] Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *ICML*, pages 24370–24387. PMLR, 2022.

[76] Shaoan Xie, Lingjing Kong, Mingming Gong, and Kun Zhang. Multi-domain image generation and translation with identifiability guarantees. In *The Eleventh International Conference on Learning Representations*, 2022.

[77] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. In *CVPR*, 2021.

[78] Dingling Yao, Caroline Muller, and Francesco Locatello. Marrying causal representation learning with dynamical systems for science. *arXiv preprint arXiv:2405.13888*, 2024.

[79] Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *The Thirteenth International Conference on Learning Representations*, 2025.

[80] Weiran Yao, Guangyi Chen, and Kun Zhang. Learning latent causal dynamics. *arXiv preprint arXiv:2202.04828*, 2022.

[81] 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.

[82] Jennifer Yuan, Helena M Blumen, Joe Verghese, and Roee Holtzer. Functional connectivity associated with gait velocity during walking and walking-while-talking in aging: A resting-state fmri study. *Human brain mapping*, 36(4):1484–1493, 2015.

[83] Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[84] Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. *arXiv preprint arXiv:2307.06250*, 2023.

[85] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009.

[86] Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *arXiv preprint arXiv:2402.05052*, 2024.

[87] Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *NeurIPS*, 2018.

# Part I
# Appendix

## Table of Contents


**A Limitations and Discussions** **16**

**B Related Work** **16**

**C Definition and Lemmas** **17**

**D The Proof of Theorem 3.1** **21**

**E The Proof of Theorem 3.6** **23**

**F The Proof of Corollary 3.9** **24**

**G The Proof of Corollary 3.10** **25**

**H Understanding Enforcing Causal Order** **25**

**I Data Details** **26**

**J The Connection Between Theory and Method** **26**

**K Implementation Framework** **27**

**L Experiments on High-Dimensional Synthetic Image Data** **27**

**M Details and More Results of Experiments on Human Motion Data** **28**

M.1 Preprocessing . . . . . 28

M.2 Evaluation Details . . . . . 29

M.3 More Results . . . . . 30

## A Limitations and Discussions

It should be noted that our framework relies on the Assumptions (i)-(iii), as well as the generative model defined in Eqs. (1a)–(1c), which are inherently untestable in practice. This limitation is not unique to our work but is a common challenge across the causal representation learning community. Precisely because of this, there is strong motivation to continue relaxing such assumptions and extending the scope of current theoretical results, which is what this work aims to do. In addition, empirical evaluations on real data provide evidence that our approach is effective, demonstrating that the insights derived from our theoretical analysis can meaningfully guide practical modeling, despite the untestable nature of the underlying assumptions.

## B Related Work

Given the challenges associated with identifiability in causal representation learning, numerous existing works tackle this issue by introducing specific assumptions. We categorize these related works into three primary parts based on the nature of these assumptions.

**Special graph structure** Some progress in achieving identifiability centers around the imposition of specific graphical structure constraints [59, 58, 4, 18, 11, 74, 75, 33, 36, 35]. Essentially, these graph structure assumptions reduce the space of possible latent causal representations or structures, by imposing specific rules for how variables are connected in the graph. One popular special graph structure assumption is the presence of two pure children nodes for each causal variable [74, 75, 23]. Very recently, the work in [1] provides a viewpoint of sparsity to understand previous various graph structure constraints.However, any complex causal graph structures may appear in real-world scenarios, beyond the pure sparsity assumption. In contrast, our approach adopts a model-based representation for latent variables, allowing arbitrary underlying graph structures.

**Temporal Information** The temporal constraint that the effect cannot precede the cause has been applied in causal representation learning [81, 39, 80, 38, 37, 42]. The success of utilizing temporal information to identify causal representations can be attributed to its innate ability to establish causal direction through time delay. By tracking the sequence of events over time, we gain the capacity to infer latent causal variables. In contrast to these approaches, our focus lies on discovering instantaneous causal relations among latent variables.

**Changes in Causal Influences** Recent advances have significantly developed the use of changes in causal influences within latent space as a means for identifying causal representations [71, 44, 45, 43, 9, 3, 41, 61, 10, 66, 70, 2, 69, 68]. Several of these works leverage such changes in conjunction with model constraints on the mapping from latent to observed space, such as assuming linear or polynomial [61, 66, 3, 84, 69, 26]. In contrast, our approach allows for flexibility by employing MLPs for this mapping. In addition, some works focus on imposing model assumptions on the latent causal variables, such as linear Gaussian models [10, 45], linear additive noise models [61, 15], and polynomial additive noise models [43]. In contrast, our work considers both additive noise models and the more general post-nonlinear models, thereby broadening the scope of identifiable causal structures. Furthermore, prior work often relies on paired data before and after random, unknown interventions [2, 9], a requirement that has been relaxed in recent work [67] to using data from two uncoupled intervention environments, whereas our method operates on fully unpaired data, a more realistic setting for applications such as biology [61, 62]. Additionally, some approaches require single-node interventions [70, 10], whereas our framework allows interventions on one node while other nodes may also be simultaneously affected. In contrast to [49], our framework imposes exponential-family assumptions on the latent noise variables and achieves identifiability with fewer environments in certain settings, whereas theirs relies on nonparametric assumptions and typically requires a larger number of diverse environments.

## C Definition and Lemmas

For ease of exposition in the following sections, we first introduce the following definition and lemmas.

**Definition C.1.** [Mapping $\mathbf{h}^{\mathbf{u}}$ From $\mathbf{n}$ to $\mathbf{z}$] Without loss of generality, since any directed acyclic causal graph admits a topological ordering, we can fix an indexing of the latent causal variables [61], such that $z_1 \prec z_2 \prec \cdots \prec z_\ell$. Together with additive noise model assumption in Eq. (1b), we have:

$$z_j = g_j^{\mathbf{u}}(z_{1:j-1}) + n_j, \qquad j = 1, \ldots, \ell. \tag{9}$$

Then the mapping $\mathbf{h}^{\mathbf{u}} : \mathbf{n} \mapsto \mathbf{z}$ induced by these equations is defined recursively by

$$h_1^{\mathbf{u}}(n_1) := n_1, \tag{10}$$

and, for each $j \geq 2$,

$$h_j^{\mathbf{u}}(n_{1:j}) := g_j^{\mathbf{u}}\big(h_1^{\mathbf{u}}(n_1),\, h_2^{\mathbf{u}}(n_{1:2}),\, \ldots,\, h_{j-1}^{\mathbf{u}}(n_{1:j-1})\big) + n_j. \tag{11}$$

The overall mapping is

$$\mathbf{h}^{\mathbf{u}}(\mathbf{n}) := \big[h_1^{\mathbf{u}}(n_1),\, h_2^{\mathbf{u}}(n_{1:2}),\, \ldots,\, h_\ell^{\mathbf{u}}(n_{1:\ell})\big]. \tag{12}$$

**Lemma C.2.** *Let $\mathbf{h}^{\mathbf{u}}$ denote the mapping from $\mathbf{n}$ to $\mathbf{z}$ as defined in Definition C.1. Then, $\mathbf{h}^{\mathbf{u}}$ is invertible, and its Jacobian matrix is lower triangular with ones on the diagonal. Consequently, $|\det \mathbf{J}_{\mathbf{h}^{\mathbf{u}}}| == 1$, and hence $|\det \mathbf{J}_{(\mathbf{h}^{\mathbf{u}})^{-1}}| = 1$ .*

*Proof.* The result follows directly from the structural form of the generative process. According to Eq. (1b), each variable $z_i$ depends only on its parents and the corresponding noise variable $n_i$. This allows us to recursively express each $z_i$ in terms of its ancestral noise variables and $n_i$.

According to the definition of $\mathbf{h}^{\mathbf{u}}$ in Definition C.1, we have:

$$z_1 = \underbrace{n_1}_{h_1^{\mathbf{u}}(n_1)}, \tag{13}$$

$$z_2 = g_2^{\mathbf{u}}(z_1) + n_2 = \underbrace{g_2^{\mathbf{u}}(n_1) + n_2}_{h_2^{\mathbf{u}}(n_1, n_2)}, \tag{14}$$

$$z_3 = \underbrace{g_3^{\mathbf{u}}(n_1, g_2^{\mathbf{u}}(n_1) + n_2) + n_3}_{h_3^{\mathbf{u}}(n_1, n_2, n_3)}, \tag{15}$$

$$\vdots$$

Due to the structure of the additive noise model and the acyclicity of the underlying causal graph (DAG), the mapping $\mathbf{h}^{\mathbf{u}}$ is invertible. Moreover, its Jacobian matrix is lower triangular with ones on the diagonal, which directly implies that $|\det \mathbf{J}_{\mathbf{h}^{\mathbf{u}}}| = 1$, thus $|\det \mathbf{J}_{(\mathbf{h}^{\mathbf{u}})^{-1}}| = 1/|\det \mathbf{J}_{\mathbf{h}^{\mathbf{u}}}| = 1$.. □

**Lemma C.3.** *The mapping between the latent causal variables $\mathbf{z}$ and the estimated latent variables $\hat{\mathbf{z}}$, obtained by matching the marginal data distribution $p(\mathbf{x} \mid \mathbf{u})$, does not depend on the environment variable $\mathbf{u}$.*

*Proof.* From Eq. (1c), the observed variable is generated as

$$\mathbf{x} = \mathbf{f}(\mathbf{z}), \tag{16}$$

where $\mathbf{f}$ is smooth and invertible by assumption (i). Suppose there exists an alternative solution such that

$$\mathbf{x} = \hat{\mathbf{f}}(\hat{\mathbf{z}}), \tag{17}$$

where $\hat{\mathbf{f}}$ is also invertible and $\hat{\mathbf{z}}$ is learned by matching the marginal distributions $p(\mathbf{x} \mid \mathbf{u})$.

Since both $\mathbf{f}$ and $\hat{\mathbf{f}}$ are invertible, we can write

$$\hat{\mathbf{z}} = \hat{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z}). \tag{18}$$

Because neither $\mathbf{f}$ nor $\hat{\mathbf{f}}$ depends on $\mathbf{u}$, the resulting mapping between $\mathbf{z}$ and $\hat{\mathbf{z}}$ is independent of $\mathbf{u}$. □

**Illustrative Example.** Consider the following structural equations:

$$z_1 = n_1, \tag{19}$$
$$z_2 = \lambda(\mathbf{u}) \cdot z_1 + n_2, \tag{20}$$
$$x_1 = z_1, \tag{21}$$
$$x_2 = z_2^3 + z_1, \tag{22}$$

In this example, although the latent variable $z_2$ depends on the environment $\mathbf{u}$ through $\lambda(\mathbf{u})$, the observed variables $\mathbf{x} = (x_1, x_2)$ are functions of $\mathbf{z} = (z_1, z_2)$, and this generative mapping does not involve $\mathbf{u}$.

Now suppose the estimated latent variables $\hat{\mathbf{z}} = (\hat{z}_1, \hat{z}_2)$ from $\mathbf{x}$. Since $x_1 = z_1$, we have $\hat{z}_1 = x_1$. Moreover, from $x_2 = z_2^3 + z_1 = z_2^3 + x_1$, we can uniquely solve for $z_2$ as

$$\hat{z}_2 = (x_2 - x_1)^{1/3}. \tag{23}$$

Therefore, the inverse mapping

$$\hat{\mathbf{z}} = \big(x_1,\ (x_2 - x_1)^{1/3}\big), \tag{24}$$

is independent of $\lambda(\mathbf{u})$, even though $\lambda(\mathbf{u})$ affects the distribution of the latent variables. Since the forward mapping from $\mathbf{z}$ to $\mathbf{x}$ is also independent of $\mathbf{u}$, it follows that the induced mapping between $\mathbf{z}$ and $\hat{\mathbf{z}}$ does not depend on $\mathbf{u}$.

**Canonical densities.** We use the following canonical statistics and full noise supports:

| Family | $\mathcal{S}$ | $\mathbf{T}(t)$ | $q(t)$ |
|---|---|---|---|
| Gaussian | $\mathbb{R}$ | $(t, t^2)^\top$ | 1 |
| Gamma | $(0, \infty)$ | $(\log t, t)^\top$ | 1 |
| Inverse Gamma | $(0, \infty)$ | $(\log t, t^{-1})^\top$ | 1 |
| Inverse Gaussian | $(0, \infty)$ | $(t, t^{-1})^\top$ | $t^{-3/2}$ |
| Beta | $(0, 1)$ | $(\log t, \log(1-t))^\top$ | 1 |

All parameters belong to the usual proper, nondegenerate parameter domains. A fixed invertible change of statistic basis does not change any rank condition below.

**Lemma C.4** (Nonconstant target contrasts with varying locations)**.** *Fix one of the five families. Let $p_r(t) = p_{\boldsymbol{\eta}_r}(t - d_r)$ for $r = 0, 1, 2$, and assume $J = \bigcap_{r=0}^{2}(d_r + \mathcal{S})$ is nonempty. If the two columns of $[\boldsymbol{\eta}_1 - \boldsymbol{\eta}_0, \boldsymbol{\eta}_2 - \boldsymbol{\eta}_0]$ are linearly independent, then at least one of $F_r(t) = \log[p_r(t)/p_0(t)]$, $r = 1, 2$, is nonconstant on $J$.*

*Proof.* Write $x = t - d_r$ in the following table. The logarithmic derivatives of the shifted densities, and their canonical parameters, are

| Family | $(\log p_r)'(t)$ | $\boldsymbol{\eta}_r$ |
|---|---|---|
| Gaussian | $-\tau_r(x - \mu_r)$ | $(\tau_r\mu_r, -\tau_r/2)^\top$ |
| Gamma | $(\alpha_r - 1)/x - \beta_r$ | $(\alpha_r - 1, -\beta_r)^\top$ |
| Inverse Gamma | $-(\alpha_r + 1)/x + \beta_r/x^2$ | $(-\alpha_r - 1, -\beta_r)^\top$ |
| Inverse Gaussian | $-a_r - 3/(2x) + b_r/x^2$ | $(-a_r, -b_r)^\top$ |
| Beta | $(\alpha_r - 1)/x + (\beta_r - 1)/(x - 1)$ | $(\alpha_r - 1, \beta_r - 1)^\top$ |

Here all shape, rate, and precision parameters are positive, and $a_r, b_r > 0$ for inverse Gaussian. Suppose both contrasts were constant on $J$. Then the three rational functions in the middle column would agree on an interval and hence as rational functions. Comparing the leading terms as $t \to \infty$ would give, respectively,

$$\tau_0 = \tau_1 = \tau_2, \quad \beta_0 = \beta_1 = \beta_2, \quad \alpha_0 = \alpha_1 = \alpha_2, \quad a_0 = a_1 = a_2, \quad \alpha_0 + \beta_0 = \alpha_1 + \beta_1 = \alpha_2 + \beta_2.$$

For Beta this comparison is between rational functions after clearing denominators; it is not a claim that the density is positive at infinity. In each case a nonzero linear functional of $\boldsymbol{\eta}_r$ is constant in $r$. Both parameter-contrast columns would lie in its one-dimensional kernel, contradicting rank two. □

**Lemma C.5** (An algebraic exponential)**.** *Let $H$ be a real-analytic function on a connected open interval. If both $H(t)$ and $\exp(H(t))$ are algebraic functions of $t$, then $H$ is constant. Here algebraic means satisfying a nonzero polynomial relation with $t$.*

*Proof.* Suppose $H$ is nonconstant. Restrict to an interval where $H' \neq 0$ and let $K$ be its local inverse. Interchanging the variables in a polynomial relation for $H$ shows that $K$ is algebraic. Composition of algebraic functions is algebraic: a polynomial relation for the outer function has algebraic coefficients after substitution, and eliminating those coefficients gives a polynomial over the original variable. Equivalently, the two successive finite algebraic extensions are finite, which follows by multiplying their finite bases. Thus $\exp(s) = \exp(H(K(s)))$ would be algebraic on an open interval. There would be polynomials $p_0, \dots, p_m$, with $p_m \neq 0$, such that

$$\sum_{k=0}^{m} p_k(s) e^{ks} = 0.$$

The left-hand side is real analytic on $\mathbb{R}$, so this identity holds for every real $s$. If $m = 0$ it is already impossible. Otherwise, divide by $e^{ms}$ and let $s \to \infty$. All lower-exponential terms tend to zero, forcing the nonzero polynomial $p_m(s)$ to tend to zero, a contradiction. □

**Lemma C.6** (Scalar identification with varying removal constants)**.** *Fix one named family from the table above. For $r = 0, 1, 2$, let $N_r$ and $\hat{N}_r$ have full-support distributions in that family, with parameters $\boldsymbol{\eta}_r$ and $\hat{\boldsymbol{\eta}}_r$. Let $d_r, e_r \in \mathbb{R}$. Suppose $w : \mathbb{R} \to \mathbb{R}$ is a global $C^2$ diffeomorphism such that*

$$w(e_r + \hat{N}_r) \overset{d}{=} d_r + N_r, \qquad r = 0, 1, 2. \tag{25}$$

*For the four non-Gaussian families, suppose also that $w$ is real analytic on $\mathbb{R}$. Assume both target-noise contrast matrices have rank two. Then $w(t) = st + c$ for all $t \in \mathbb{R}$, where $s \neq 0$. For the three positive-half-line families $s > 0$, and for Beta $s \in \{-1, 1\}$. No equality among $d_0, d_1, d_2$, or among $e_0, e_1, e_2$, is required.*

*Proof.* A global scalar diffeomorphism is either increasing or decreasing. For a positive-half-line family, it must be increasing: a decreasing global diffeomorphism sends a right half-line onto a left half-line, not onto the support of the true marginal. Full support gives $w(e_r) = d_r$. For Beta, the endpoints are mapped either in their original order or in reverse order.

**Gaussian.** In the baseline environment, $e_0 + \hat{N}_0$ and $d_0 + N_0$ are nondegenerate Gaussian variables. The unique increasing transport between their strictly increasing distribution functions is affine; the decreasing transport is obtained by reflection and is also affine. Explicitly, if their means and standard deviations are $\hat{m}, \hat{\sigma}$ and $m, \sigma$, then $w(t) = m + \varepsilon\sigma(t - \hat{m})/\hat{\sigma}$ with $\varepsilon \in \{-1, 1\}$.

**Gamma.** Put $\rho(t) = w(e_0 + t) - d_0$ for $t > 0$. The global $C^2$ condition at the endpoint gives

$$\rho(t) = kt + O(t^2), \qquad \rho'(t) = k + O(t), \qquad k > 0.$$

Let the true and estimated baseline noises have shape–rate parameters $(\alpha, \beta)$ and $(\hat{\alpha}, \hat{\beta})$. The density transport identity is $\hat{p}(t) = p(\rho(t))\rho'(t)$. Comparing its powers at $t \downarrow 0$ yields $\hat{\alpha} = \alpha$. Equal-shape Gamma laws differ only by scale, so uniqueness of increasing quantile transport gives $\rho(t) = (\hat{\beta}/\beta)t$ on $(0, \infty)$. Thus $w$ is affine on that half-line. Real analyticity extends the identity to all of $\mathbb{R}$.

**Inverse Gamma.** Use the same $\rho$ and $k$. The log-density transport near zero gives

$$-\hat{\beta}/t - (\hat{\alpha} + 1)\log t + O(1) = -\beta/(kt) - (\alpha + 1)\log t + O(1).$$

Multiplying by $t$ gives $\hat{\beta} = \beta/k$. After subtracting the equal $1/t$ terms, comparison of the logarithmic terms gives $\hat{\alpha} = \alpha$. Equal-shape inverse-Gamma laws differ only by scale; hence $\rho(t) = (\beta/\hat{\beta})t$. Again, analyticity extends the affine identity to $\mathbb{R}$.

**Beta.** If $w$ is increasing, then $\rho(t) = w(e_0 + t) - d_0$ maps $(0, 1)$ onto itself and has finite nonzero derivatives at both endpoints. Comparison of the density powers at zero and one gives $\hat{\alpha} = \alpha$ and $\hat{\beta} = \beta$. Uniqueness of increasing quantile transport implies $\rho(t) = t$. If $w$ is decreasing, the two shape parameters are exchanged and $\rho(t) = 1 - t$. Analyticity extends the resulting affine identity from the interval to $\mathbb{R}$.

**Inverse Gaussian.** Write the full shifted densities as

$$p_r(s) = C_r(s-d_r)^{-3/2}\exp\{-a_r(s-d_r) - b_r/(s-d_r)\}, \qquad s > d_r,$$

where $a_r, b_r > 0$ and $\boldsymbol{\eta}_r = (-a_r, -b_r)^\top$. Their logarithmic derivatives are

$$S_r(s) = -a_r - \frac{3}{2(s-d_r)} + \frac{b_r}{(s-d_r)^2}.$$

Let $r_1 = S_1 - S_0$ and $r_2 = S_2 - S_0$. These two rational functions are linearly independent. To verify this, consider the possible equalities between their denominator locations. If all $d_r$ agree, independence is exactly rank two of the contrasts in $(a_r, b_r)$. If the three locations are distinct, each contrast has a nonzero double-pole term at a location absent from the other. If $d_0 = d_1 \neq d_2$, the double-pole term at $d_2$ forces the coefficient of $r_2$ in a linear relation to vanish, and $r_1 \neq 0$ by rank two; the case $d_0 = d_2 \neq d_1$ is identical. Finally, if $d_1 = d_2 \neq d_0$, the nonzero simple-pole coefficients at these locations force any proportionality factor to be one. Then $r_1 = r_2$ would imply $(a_1, b_1) = (a_2, b_2)$, again violating rank two. These comparisons can equivalently be made by clearing denominators and comparing polynomial coefficients.

Thus $R = r_1/r_2$ is a nonconstant rational function. By Eq. (25), on a sufficiently large tail where all densities are positive,

$$p_r(w(t))w'(t) = \hat{p}_r(t).$$

Taking ratios with $r = 0$ and differentiating gives $r_a(w(t))w'(t) = \hat{r}_a(t)$ for $a = 1, 2$. Consequently

$$R(w(t)) = \hat{R}(t),$$

where $\hat{R}$ is also rational. Clearing denominators yields a nonzero polynomial $P$ with $P(t, w(t)) = 0$. Thus $w$ is algebraic on the tail. Its derivative is also algebraic, by implicit differentiation of a minimal polynomial on an interval where its derivative in the second variable is nonzero.

Set $U(t) = w(t) - d_0$ and $V(t) = t - e_0$. Baseline density transport gives

$$\exp\{H(t)\} = \frac{\hat{C}_0}{C_0}\frac{(U(t)/V(t))^{3/2}}{w'(t)},$$

$$H(t) = -a_0 U(t) - \frac{b_0}{U(t)} + \hat{a}_0 V(t) + \frac{\hat{b}_0}{V(t)}. \tag{26}$$

Both $H$ and the right-hand side are algebraic on an open interval. Lemma C.5 makes $H$ constant there; analyticity extends this to the positive baseline half-line. For a constant $K > 0$ it follows that

$$w'(t)\left(\frac{V(t)}{U(t)}\right)^{3/2} = K.$$

Integrating gives

$$U(t)^{-1/2} = KV(t)^{-1/2} + C.$$

Since $w$ is an increasing global diffeomorphism, both $U(t)$ and $V(t)$ tend to infinity as $t \to \infty$, so $C = 0$. Therefore $w(t) = d_0 + K^{-2}(t - e_0)$ on the half-line, and hence on $\mathbb{R}$ by analyticity. This argument uses the complete inverse-Gaussian density, not only its two sufficient statistics. □

## D The Proof of Theorem 3.1

**Proof sketch** First, we compare the conditional data distributions within each parent-removal block. The non-target conditional factors cancel, leaving density ratios that depend only on the corresponding target noise. Second, the true latent causal variables $\mathbf{z}$ are related to the estimated ones through an invertible map independent of $\mathbf{u}$ (Lemma C.3). The density-ratio identities make this map component-wise, up to permutation. Finally, the specified noise family and parent removal restrict each component to an affine transformation. The removal constants need not coincide across environments.

*Proof.* The proof of Theorem 3.1 unfolds in three steps. Before proceeding, we introduce the following notation.

**Notation** We consider two sets of model parameters:

$$\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \mathbf{h}^{\mathbf{u}}, \boldsymbol{\eta}), \qquad \hat{\boldsymbol{\theta}} = (\hat{\mathbf{f}}, \hat{\mathbf{T}}, \hat{\mathbf{h}}^{\mathbf{u}}, \hat{\boldsymbol{\eta}}), \tag{27}$$

where $\boldsymbol{\theta}$ corresponds to the true generative model and $\hat{\boldsymbol{\theta}}$ to an admissible estimated model. Recall that $\mathbf{T}$ denotes the sufficient statistics of the noise variables, and $\boldsymbol{\eta}$ their natural parameters. The function $\mathbf{f}$ maps latent causal variables to observations. The mapping $\mathbf{h}^{\mathbf{u}}$ (Definition C.1) maps noise variables to causal variables, $\mathbf{z} = \mathbf{h}^{\mathbf{u}}(\mathbf{n})$.

Both parameter sets induce the same conditional data distributions:

$$p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{u}) = p_{\hat{\boldsymbol{\theta}}}(\mathbf{x} \mid \mathbf{u}) \tag{28}$$

for all calibration environments. At a fixed observation, write $\mathbf{z} = \mathbf{f}^{-1}(\mathbf{x})$ and $\hat{\mathbf{z}} = \hat{\mathbf{f}}^{-1}(\mathbf{x})$. The associated noise arguments are $\mathbf{n}^{\mathbf{u}} = (\mathbf{h}^{\mathbf{u}})^{-1}(\mathbf{z})$ and $\hat{\mathbf{n}}^{\mathbf{u}} = (\hat{\mathbf{h}}^{\mathbf{u}})^{-1}(\hat{\mathbf{z}})$; they generally depend on $\mathbf{u}$. We use continuous density versions on their positive-density interiors, so the density equalities hold pointwise there.

**Step I: Comparison of the target noise distributions.** Expanding Eq. (28) by the change of variables formula and taking logarithms yields

$$\begin{aligned} &\log|\det \mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \log|\det \mathbf{J}_{(\mathbf{h}^{\mathbf{u}})^{-1}}(\mathbf{z})| + \log p_{(\mathbf{T},\boldsymbol{\eta})}(\mathbf{n}^{\mathbf{u}} \mid \mathbf{u}) \\ &\quad = \log|\det \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \log|\det \mathbf{J}_{(\hat{\mathbf{h}}^{\mathbf{u}})^{-1}}(\hat{\mathbf{z}})| + \log p_{(\hat{\mathbf{T}},\hat{\boldsymbol{\eta}})}(\hat{\mathbf{n}}^{\mathbf{u}} \mid \mathbf{u}). \end{aligned} \tag{29}$$

The complete exponential-family noise density satisfies

$$\log p_{(\mathbf{T},\boldsymbol{\eta})}(\mathbf{n} \mid \mathbf{u}) = \sum_k \log q_k(n_k) + \mathbf{T}(\mathbf{n})^\top \boldsymbol{\eta}(\mathbf{u}) - \sum_k \log Z_k(\mathbf{u}), \tag{30}$$

where $q_k$ is the fixed carrier. By Lemma C.2, both additive-noise maps have unit absolute Jacobian determinant. Consequently, Eq. (29) reduces to

$$\begin{aligned} &\log|\det \mathbf{J}_{\mathbf{f}^{-1}}(\mathbf{x})| + \sum_k \log p_{\boldsymbol{\eta}_k(\mathbf{u})}\big(z_k - \mathrm{g}_k^{\mathbf{u}}(\mathrm{pa}_k)\big) \\ &\quad = \log|\det \mathbf{J}_{\hat{\mathbf{f}}^{-1}}(\mathbf{x})| + \sum_k \log p_{\hat{\boldsymbol{\eta}}_k(\mathbf{u})}\big(\hat{z}_k - \hat{\mathrm{g}}_k^{\mathbf{u}}(\widehat{\mathrm{pa}}_k)\big). \end{aligned} \tag{31}$$

Each density here includes its carrier and support.

Fix $i$ and write its calibration block as $\mathbf{u}_i^0 = \mathbf{u}_i$, $\mathbf{u}_i^1 = \mathbf{u}_i'$, and $\mathbf{u}_i^2 = \mathbf{u}_i''$. Let $j_i$ be the unknown target of the same labeled block in the estimated model, as required by the admissible model class. Parent removal and connectedness of the parent domains give

$$\mathrm{g}_i^{\mathbf{u}_i^r} \equiv d_{ir}, \qquad \hat{\mathrm{g}}_{j_i}^{\mathbf{u}_i^r} \equiv \hat{d}_{ir}, \qquad r = 0, 1, 2. \tag{32}$$

The constants may depend on $r$.

Define the resulting target marginal densities and their log-density ratios by

$$p_{ir}(t) := p_{\boldsymbol{\eta}_i(\mathbf{u}_i^r)}(t - d_{ir}), \qquad F_{ir}(t) := \log \frac{p_{ir}(t)}{p_{i0}(t)}, \qquad r = 1, 2, \tag{33}$$

with $p_{i0}$ defined in the same way. Define $\hat{p}_{ir}$ and $\hat{F}_{ir}$ analogously for target $j_i$.

Subtract Eq. (31) at $\mathbf{u}_i^0$ from its value at $\mathbf{u}_i^r$. The decoder Jacobians and all non-target conditional factors cancel. Hence

$$F_{ir}(z_i) = \hat{F}_{ir}(\hat{z}_{j_i}), \qquad r = 1, 2, \tag{34}$$

on the common positive-density interior of this block. By Assumption (ii) and Lemma C.4, at least one of $F_{i1}, F_{i2}$ is nonconstant.

**Step II: The mapping between the latent causal variables.** Since $\mathbf{x} = \mathbf{f}(\mathbf{z}) = \hat{\mathbf{f}}(\hat{\mathbf{z}})$, we have

$$\mathbf{z} = \mathbf{\Phi}(\hat{\mathbf{z}}), \qquad \mathbf{\Phi} := \mathbf{f}^{-1} \circ \hat{\mathbf{f}}. \tag{35}$$

This is an environment-independent diffeomorphism (Lemma C.3).

Choose a nonconstant $F_{ir}$ and substitute Eq. (35) into Eq. (34). For any $k \neq j_i$, differentiation gives

$$F'_{ir}\big(\Phi_i(\hat{\mathbf{z}})\big)\frac{\partial \Phi_i}{\partial \hat{z}_k} = 0. \tag{36}$$

If all noise coordinates are Gaussian, the identity holds on $\mathbb{R}^\ell$ and $F_{ir}$ is a nonconstant polynomial of degree at most two. Its derivative has at most one zero. Because $\nabla \Phi_i$ is nowhere zero, the preimage of this zero has empty interior. Thus $\partial \Phi_i / \partial \hat{z}_k = 0$ everywhere by continuity.

Otherwise, the observation maps and their inverses are globally real analytic by Assumption (i). For every named noise family, including a Gaussian target in a mixed-family model, $F_{ir}$ is nonconstant and real analytic on its common target-support interval. The image under $\Phi_i$ of the nonempty block overlap is open, so there is a nonempty open subset on which $F'_{ir}(\Phi_i) \neq 0$. Equation (36) makes $\partial \Phi_i / \partial \hat{z}_k$ vanish there. Global real analyticity extends this equality to $\mathbb{R}^\ell$.

Consequently,

$$\Phi_i(\hat{\mathbf{z}}) = w_i(\hat{z}_{j_i}). \tag{37}$$

The Jacobian of $\mathbf{\Phi}$ is nonsingular, so $w_i'$ is nowhere zero. Surjectivity of $\mathbf{\Phi}$ makes $w_i$ onto $\mathbb{R}$. It is therefore a global scalar diffeomorphism, real analytic whenever the model contains a non-Gaussian noise coordinate.

**Step III: Affine identification and permutation.** In every environment $\mathbf{u}_i^r$ of the block, the target is parent-free in both models. Matching the marginal distributions in Eq. (37) therefore gives

$$w_i\big(\hat{d}_{ir} + \hat{N}_{j_i}^{\mathbf{u}_i^r}\big) \overset{d}{=} d_{ir} + N_i^{\mathbf{u}_i^r}, \qquad r = 0, 1, 2. \tag{38}$$

Both noises belong to the same specified family and have its full support. Lemma C.6, with the local rank condition in Assumption (ii), gives $w_i(t) = s_i t + c_i$, where $s_i \neq 0$. For Gaussian noise, this is simply the unique monotone transport between the two Gaussian target marginals in any baseline environment. The other four cases are proved in the cited auxiliary lemma.

In particular, the baseline target noises $n_i^{(i)} := z_i - d_{i0}$ and $\hat{n}_{j_i}^{(i)} := \hat{z}_{j_i} - \hat{d}_{i0}$ satisfy

$$n_i^{(i)} = s_i \hat{n}_{j_i}^{(i)} + b_i, \qquad b_i = s_i \hat{d}_{i0} + c_i - d_{i0}. \tag{39}$$

Equivalently, parent removal gives $z_i = s_i \hat{z}_{j_i} + c_i$.

If two distinct true coordinates had the same $j_i$, two rows of $\mathbf{J}_{\mathbf{\Phi}}$ would be proportional, contradicting its nonsingularity. Thus $j_i = \pi(i)$ for a permutation $\pi$. Let $P_{i,\pi(i)} = 1$ and $\mathbf{D} = \operatorname{diag}(s_1, \ldots, s_\ell)$. We conclude that

$$\mathbf{z} = \mathbf{D}\mathbf{P}\hat{\mathbf{z}} + \mathbf{c}. \tag{40}$$

This completes the proof. □

## E The Proof of Theorem 3.6

**Proof sketch** As outlined in the proof of Theorem 3.1, the identification argument can be applied separately to each parent-removal block. Under Assumptions (i)–(ii), when Assumption (iii) holds for a subset of latent variables, their corresponding reference noise variables can still be identified up to scaling and translation, without requiring parent removal for the remaining variables. Moreover, the invertible mapping between the true latent causal variables $\mathbf{z}$ and the estimated ones is independent of $\mathbf{u}$ (Lemma C.3). Combining these results with parent removal gives the positive partial identifiability result. For a non-root coordinate without an imposed parent-removal block, an environment-independent triangular shear gives an observationally equivalent admissible model and establishes the non-identifiability statement.

*Proof.* Let $S \subseteq \{1, \dots, \ell\}$ denote the set of latent variables whose designated blocks satisfy Assumption (iii). Assumptions (i)–(ii) hold for all coordinates. We use the same admissible comparison class and labeled blocks as in Theorem 3.1, imposing parent removal on the same selected block labels in both models, although their target identities need not be known or the same. We show that every $z_i$ with $i \in S$ is identifiable up to a component-wise affine transformation.

For any $i \in S$, consider its three environments $\mathbf{u}_i^0 = \mathbf{u}_i$, $\mathbf{u}_i^1 = \mathbf{u}_i'$, and $\mathbf{u}_i^2 = \mathbf{u}_i''$. Let $j_i$ denote the unknown target of the same block in the estimated model. As in Step I of Appendix D, parent removal and the invariance of all non-target conditional factors give

$$F_{ir}(z_i) = \hat{F}_{ir}(\hat{z}_{j_i}), \qquad r = 1, 2, \tag{41}$$

on the common positive-density interior of the block, where $F_{ir}$ and $\hat{F}_{ir}$ are the complete target log-density ratios. Applying Steps II and III of Appendix D to this block, under the same regularity and noise-family assumptions, we obtain

$$n_i^{(i)} = s_i \hat{n}_{j_i}^{(i)} + b_i, \qquad s_i \neq 0, \tag{42}$$

where

$$n_i^{(i)} = z_i - d_{i0}, \qquad \hat{n}_{j_i}^{(i)} = \hat{z}_{j_i} - \hat{d}_{i0}. \tag{43}$$

Here $d_{i0}$ and $\hat{d}_{i0}$ are the respective constant parent contributions in the baseline environment $\mathbf{u}_i^0$. The positive identification argument only uses the calibration block for $i$ and does not require Assumption (iii) to hold for any $k \notin S$.

By Lemma C.3, the true and estimated latent causal variables are related through the environment-independent invertible mapping

$$\mathbf{z} = \mathbf{\Phi}(\hat{\mathbf{z}}), \qquad \mathbf{\Phi} := \mathbf{f}^{-1} \circ \hat{\mathbf{f}}. \tag{44}$$

Substituting the expressions of $n_i^{(i)}$ and $\hat{n}_{j_i}^{(i)}$ into Eq. (42) gives

$$z_i = s_i \hat{z}_{j_i} + c_i, \qquad c_i = d_{i0} - s_i \hat{d}_{i0} + b_i. \tag{45}$$

Finally, suppose two distinct indices $i, k \in S$ had $j_i = j_k$. Then Eq. (45) would make the $i$-th and $k$-th rows of $\mathbf{J}_{\mathbf{\Phi}}$ proportional, contradicting its invertibility. Thus the correspondence $\pi(i) = j_i$ is injective on $S$. Therefore every $z_i$, $i \in S$, is affinely identifiable up to an injective matching with the estimated coordinates.

For part (b), let $i \notin S$. A root satisfies the parent-removal condition vacuously, so $z_i$ has a parent $z_p$. Choose a topological ordering of the true acyclic supergraph, so that $p < i$, and define the triangular map $\hat{\mathbf{z}} = \mathbf{\Psi}_\alpha(\mathbf{z})$ by

$$\hat{z}_i = z_i - \alpha z_p, \qquad \hat{z}_k = z_k \quad (k \neq i), \qquad \alpha \neq 0. \tag{46}$$

Choose $\alpha$ so that $\mathrm{g}_i^{\mathbf{u}}(\mathrm{pa}_i) - \alpha z_p$ is not constant in its parent arguments throughout the designated block for $i$. At most one value of $\alpha$ can violate this requirement: two such values would imply that a nonzero multiple of $z_p$ is constant on an open parent domain.

The inverse map has $z_i = \hat{z}_i + \alpha \hat{z}_p$ and $z_k = \hat{z}_k$ for $k \neq i$. Thus $\mathbf{\Psi}_\alpha$ is an environment-independent global analytic diffeomorphism with unit Jacobian determinant. Define

$$\hat{\mathbf{f}} = \mathbf{f} \circ \mathbf{\Psi}_\alpha^{-1}. \tag{47}$$

Rewrite every structural equation by substituting $\mathbf{z} = \boldsymbol{\Psi}_\alpha^{-1}(\hat{\mathbf{z}})$. In particular,

$$\hat{z}_i = \mathrm{g}_i^{\mathbf{u}}(\widehat{\mathrm{pa}}_i) - \alpha \hat{z}_p + n_i. \tag{48}$$

The other transformed mechanisms are obtained by the same substitution and still depend only on predecessors in this topological ordering. They may introduce additional parent edges, which are permitted by the unrestricted DAG and mechanism class. The transformed model retains exactly the same independent noises, their families, and their natural parameters, so all noise-rank conditions remain valid.

To verify the block conditions, write $\rho_k^{\mathbf{u}}(z_k \mid \mathbf{z}_{<k})$ for the original conditional noise factor, conditioning on all predecessors rather than only its parents. Triangularity and the unit diagonal give

$$\hat{\rho}_k^{\mathbf{u}}(\hat{z}_k \mid \hat{\mathbf{z}}_{<k}) = \rho_k^{\mathbf{u}}\big((\boldsymbol{\Psi}_\alpha^{-1}(\hat{\mathbf{z}}))_k \mid (\boldsymbol{\Psi}_\alpha^{-1}(\hat{\mathbf{z}}))_{<k}\big)\,. \tag{49}$$

The change of arguments is independent of $\mathbf{u}$, so every non-target conditional invariance in Assumption (ii) is preserved. For each protected target $k \in S$, we have $k \neq i$, and a structural function constant in its parent arguments remains constant after the substitution. Thus all required parent-removal conditions for $S$ are preserved as well. The constructed alternative may retain the original target identities of all labeled blocks. Decoder regularity is preserved by composition with the linear shear, and the common block overlap is preserved because the observational distributions are unchanged.

Indeed,

$$\hat{\mathbf{f}}(\hat{\mathbf{z}}) = \mathbf{f}(\mathbf{z}) \tag{50}$$

samplewise in every environment. Nevertheless,

$$z_i = \hat{z}_i + \alpha \hat{z}_p \tag{51}$$

is not a function of any single estimated coordinate on a nonempty joint positive-density interior, since $p \neq i$ and $\alpha \neq 0$. Therefore $z_i$ is not identifiable even up to an arbitrary component-wise invertible transformation within this admissible comparison class. This establishes part (b).

□

## F The Proof of Corollary 3.9

**Proof sketch** The proof proceeds intuitively as follows. Since each function $\bar{g}_i$ in Eq. (5) is invertible, we can compose $\mathbf{f}$ with $\bar{\mathbf{g}}$ to obtain a new observation mapping $\widetilde{\mathbf{f}}$. Under the stated regularity conditions, this composition satisfies the observation-map assumption of Theorem 3.1. We can therefore apply Theorem 3.1 to the underlying latent additive noise model and then use the component-wise invertibility of $\bar{\mathbf{g}}$ to obtain the stated result.

*Proof.* Since $\bar{\mathbf{z}} = \bar{\mathbf{g}}(\mathbf{z})$ in Eq. (5) and $\mathbf{x} = \mathbf{f}(\bar{\mathbf{z}})$, we can define $\widetilde{\mathbf{f}} := \mathbf{f} \circ \bar{\mathbf{g}}$, so that $\mathbf{x} = \widetilde{\mathbf{f}}(\mathbf{z})$. The estimated model similarly satisfies $\mathbf{x} = \hat{\widetilde{\mathbf{f}}}(\hat{\mathbf{z}})$, where $\hat{\widetilde{\mathbf{f}}} := \hat{\mathbf{f}} \circ \hat{\bar{\mathbf{g}}}$. By the stated regularity assumptions, both compositions are invertible and satisfy Assumption (i) of Theorem 3.1, including the additional regularity required for the non-Gaussian families. The underlying additive noise models retain the same calibrated noise-variation and parent-removal conditions. Thus the proof in Appendix D applies, and Eq. (40) gives $z_i = s_i \hat{z}_{\pi(i)} + c_i$ with $s_i \neq 0$ for a permutation $\pi$.

Finally, given the component-wise invertible nonlinear mappings between $\bar{\mathbf{z}}$ and $\mathbf{z}$,

$$\bar{\mathbf{z}} = \bar{\mathbf{g}}(\mathbf{z}), \tag{52}$$

and between the estimated variables,

$$\hat{\bar{\mathbf{z}}} = \hat{\bar{\mathbf{g}}}(\hat{\mathbf{z}}), \tag{53}$$

we substitute their inverses into Eq. (40) to obtain

$$\bar{g}_i^{-1}(\bar{z}_i) = s_i \hat{\bar{g}}_{\pi(i)}^{-1}(\hat{\bar{z}}_{\pi(i)}) + c_i. \tag{54}$$

Consequently,

$$\bar{z}_i = \bar{g}_i\Big(s_i \hat{\bar{g}}_{\pi(i)}^{-1}(\hat{\bar{z}}_{\pi(i)}) + c_i\Big) =: M_i(\hat{\bar{z}}_{\pi(i)}). \tag{55}$$

Since $s_i \neq 0$ and both $\bar{g}_i$ and $\hat{\bar{g}}_{\pi(i)}$ are invertible, $M_i$ is an invertible component-wise transformation. This has the stated form with the additional outer translation constant set to zero. This completes the proof. □

## G The Proof of Corollary 3.10

*Proof.* As in Appendix F, define the effective observation maps $\widetilde{\mathbf{f}} = \mathbf{f} \circ \bar{\mathbf{g}}$ and $\hat{\widetilde{\mathbf{f}}} = \hat{\mathbf{f}} \circ \hat{\bar{\mathbf{g}}}$. By the stated model-class assumptions, both maps satisfy Assumption (i), including its global regularity requirements. The underlying additive models have the same labeled blocks, noise families, and conditions of Assumption (ii); parent removal is imposed on the same selected block labels. The partial-identification argument in Appendix E therefore applies. We consider the following two cases.

- If $z_i$ is a root node or Assumption (iii) holds for $z_i$, the proof in Appendix E gives

$$z_i = s_i \hat{z}_j + c_i, \qquad s_i \neq 0. \tag{56}$$

  Since $\bar{z}_i = \bar{g}_i(z_i)$ and $\hat{\bar{z}}_j = \hat{\bar{g}}_j(\hat{z}_j)$, we obtain

$$\bar{z}_i = \bar{g}_i \left( s_i \hat{\bar{g}}_j^{-1}(\hat{\bar{z}}_j) + c_i \right) =: M_j(\hat{\bar{z}}_j). \tag{57}$$

  where $M_j$ is invertible because it is a composition of invertible mappings. This has the stated form with the additional outer translation constant set to zero. For a root, parent removal is automatic, but the block conditions in Assumption (ii) are still required.

- If Assumption (iii) does not hold for a non-root variable $z_i$, apply the shear construction in Appendix E to its underlying additive model and effective decoder $\widetilde{\mathbf{f}}$. It gives an observationally equivalent admissible model with $z_i = \hat{z}_i + \alpha \hat{z}_p$, where $p \neq i$ and $\alpha \neq 0$. To express it as a post-nonlinear model, take $\hat{\bar{\mathbf{g}}} = \mathrm{id}$ and define

$$\hat{\mathbf{f}} = \widetilde{\mathbf{f}} \circ \mathbf{\Psi}_\alpha^{-1}. \tag{58}$$

  Identity post-nonlinear maps are allowed because the post-nonlinear class includes additive-noise models. Its effective decoder is $\hat{\widetilde{\mathbf{f}}} = \widetilde{\mathbf{f}} \circ \mathbf{\Psi}_\alpha^{-1}$, which retains the required regularity. Thus it is an admissible observationally equivalent post-nonlinear model.

  Suppose, for contradiction, that $\bar{z}_i = M_j(\hat{\bar{z}}_j) + c_j$ for some invertible mapping $M_j$. Then

$$z_i = \bar{g}_i^{-1} \left( M_j(\hat{\bar{g}}_j(\hat{z}_j)) + c_j \right), \tag{59}$$

  implying that $z_i$ is a function of $\hat{z}_j$ alone, contradicting the shear relation. Therefore, the non-identifiability also carries over to the post-nonlinear model.

□

## H Understanding Enforcing Causal Order

In the inference model, we naturally enforce a causal order $z_1 \succ z_2 \succ \cdots \succ z_\ell$ without requiring specific semantic information. This does not imply that we need to know the true causal order a prior. Instead, we leverage the permutation indeterminacy in latent space, as demonstrated in [45].

For instance, suppose the underlying latent causal variables correspond to properties such as the size ($z_1$) and color ($z_2$) of an object. Permutation indeterminacy implies that we cannot guarantee whether the recovered latent variable $\hat{z}_1$ represents the size or the color. This ambiguity in the latent space, however, offers an advantage: by predefining a causal order, we enforce that $\hat{z}_1$ causes $\hat{z}_2$, without explicitly specifying the semantic meaning of these variables.

Due to the identifiability guarantee, $\hat{z}_1$, as the first node in the predefined causal order, will learn the semantic information of the first node in the true underlying causal order, e.g., the size. Similarly, $\hat{z}_2$, as the second node in the predefined causal order, will be assigned to learn the semantic feature of the second node (e.g., color). As a result, we can naturally establish a causal fully-connected graph by pre-defining causal order, ensuring the estimation of a directed acyclic graph (DAG) in inference model and avoiding DAG constraints, such as those proposed by [87].

As an exception to the coefficient sampling rule above, for each target $z_i$, we designate three calibration environments in which $\lambda_{k,i}(\mathbf{u}) = 0$ for every parent $z_k \in \mathrm{pa}_i$. Within each triple, all non-target structural mechanisms and noise parameters are held fixed, while the target Gaussian parameters vary so that $\mathbf{L}_i$ in Assumption (ii) is invertible.

# I Data Details

**Synthetic Data** We follow nonlinear ICA by generating nonstationary Gaussian noise. The data is divided into M segments, each containing L samples. The environment variable u corresponds to the segment index and is uniformly distributed over [1,M]. Within each segment, latent noise variables are sampled from Gaussian distributions with segment-dependent mean and variance. More details, refer to https://github.com/ilkhem/icebeem/blob/master/data/imca.py. Based on the sampled noise, we construct latent causal variables according to Eq. (1b). This step distinguishes our setting from nonlinear ICA. The sampled causal variables are then mapped to observation data via a nonlinear mixing function, implemented as an MLP, the same as in nonlinear ICA.

In our experimental results using synthetic data, we utilize 50 segments, with each segment containing a sample size of 1000. Furthermore, we explore latent causal or noise variables with dimensions of 2, 3, 4, and 5, respectively. Specifically, our analysis centers around the following structural causal model:

$$n_i :\sim \mathcal{N}(\alpha, \beta), \tag{60}$$
$$z_1 := n_1, \tag{61}$$
$$z_2 := \lambda_{1,2}(\mathbf{u}) \sin(z_1) + n_2, \tag{62}$$
$$z_3 := \lambda_{2,3}(\mathbf{u}) \cos(z_2) + n_3, \tag{63}$$
$$z_4 := \lambda_{3,4}(\mathbf{u}) \log(z_3^2) + n_4, \tag{64}$$
$$z_5 := \lambda_{3,5}(\mathbf{u}) \exp(\sin(z_3^2)) + n_5. \tag{65}$$

In this context, both $\alpha$ and $\beta$ for Gaussian noise are drawn from uniform distributions within the ranges of $[-2.0, 2.0]$ and $[0.1, 3.0]$, respectively. The values of $\lambda_{i,j}(\mathbf{u})$ are sampled from a uniform distribution spanning $[-2.0, -0.1] \cup [0.1, 2.0]$. As an exception to the coefficient sampling rule above, for each target $z_i$, we designate three calibration environments in which $\lambda_{k,i}(\mathbf{u}) = 0$ for every parent $z_k \in \mathrm{pa}_i$. Within each triple, all non-target structural mechanisms and noise parameters are held fixed. After sampling the latent variables, we use a random three-layer feedforward neural network as the mixing function, as described in [24, 25, 29].

**Synthetic Data for Partial Identifiability** In our experimental results, which utilized synthetic data to explore partial identifiability, we modified the Eqs (60)-(64) by

$$\dot{z}_i := z_i + z_{i-1}. \tag{66}$$

In this formulation, for each $i$, there exists a $z_{i-1}$ that remains unaffected by $\mathbf{u}$, thereby violating condition (iii). The modified parent-dependent term has an environment-invariant coefficient and remains present in the calibration environments, thereby violating the parent-removal requirement in Assumption (iii) for the modified coordinates.

**Image Data** In our experimental results using image data, we consider the following latent structural causal model:

$$n_i :\sim \mathcal{N}(\alpha, \beta), \tag{67}$$
$$z_1 := n_1 \tag{68}$$
$$z_2 := \lambda_{1,2}(\mathbf{u})(\sin(z_1) + z_1) + n_2, \tag{69}$$
$$z_3 := \lambda_{2,3}(\mathbf{u})(\cos(z_2) + z_2) + n_3, \tag{70}$$

where both $\alpha$ and $\beta$ for Gaussian noise are drawn from uniform distributions within the ranges of $[-2.0, 2.0]$ and $[0.1, 3.0]$, respectively. The values of $\lambda_{i,j}(\mathbf{u})$ are sampled from a uniform distribution spanning $[-2.0, -0.1] \cup [0.1, 2.0]$. The same calibration-block construction is applied to the image data.

# J The Connection Between Theory and Method

In this section, we clarify the connection between the theoretical results in Section 3 and the practical learning framework in Section 4. In particular, we explain how the proposed method can be viewed as a *constructive realization* of the likelihood-based objective underlying the identifiability guarantees.

**From Identifiability to Likelihood-Based Learning** The identifiability result in Theorem 3.1 establishes recovery of latent variables under the condition that the learned model matches the true conditional data distribution $p(\mathbf{x} \mid \mathbf{u})$. Therefore, the central requirement is *likelihood matching*, i.e.,

$$p_\theta(\mathbf{x} \mid \mathbf{u}) = p_{\text{data}}(\mathbf{x} \mid \mathbf{u}).$$

To realize this in practice, we adopt a variational inference framework, where the log-likelihood is approximated via the evidence lower bound (ELBO). This naturally leads to a VAE-based formulation as a flexible and scalable approach for density estimation.

**Instantiating the Generative Model** The theoretical data-generating process (Eqs. (1a)-(1c)) assumes: (i) latent noise follows an exponential family conditioned on $\mathbf{u}$, (ii) latent variables follow an additive noise model, and (iii) observations are generated via an invertible mapping.

To translate this into a learnable model, we make the following instantiations:

- We instantiate the latent noise distribution as Gaussian. This choice satisfies the exponential family assumption in Eq. (1a) and enables efficient optimization via the reparameterization trick.
- Under the additive structure in Eq. (1b), we obtain a conditional Gaussian prior, which corresponds to Eq. (6) in the method.

A key aspect of the theory is that the environment variable $\mathbf{u}$ modulates both: (i) the distribution of latent noise, and (ii) the causal mechanisms $\mathrm{g}_i^{\mathbf{u}}$. In the Gaussian instantiation, this is implemented by allowing $\mathbf{u}$ to control the mean and variance as in Eq. (6), thereby capturing environment-dependent causal mechanisms.

**Connection to Assumption (iii) and Structure Learning** A central requirement for identifiability is Assumption (iii), which characterizes changes in causal influences across environments. In our method, this is implemented through the gating variables $\mathbf{m}_i(\mathbf{u})$ in Eqs. (6) and (7), which modulate the contribution of parent variables. By enforcing sparsity on $\mathbf{m}_i(\mathbf{u})$, the model is encouraged to select environment-dependent causal relationships, thereby aligning with the mechanism change condition required by Assumption (iii).

**Variational Inference and Objective** Since the true posterior $p(\mathbf{z} \mid \mathbf{x}, \mathbf{u})$ is intractable, we introduce a variational approximation, as defined in Eq. (7). The posterior mirrors the structure of the prior, and the same gating parameters $\mathbf{m}_i(\mathbf{u})$ are shared. This ensures structural consistency and reduces mismatch between inference and generation.

Combining the above components, we optimize the ELBO, as defined in Eq. (6). Under sufficient model capacity and data, this objective approximates likelihood maximization, thereby aligning with the requirement of Theorem 3.1.

## K Implementation Framework

We perform all experiments using the GPU RTX 4090, equipped with 32 GB of memory. Figure 8 illustrates our proposed method for learning latent nonlinear models with additive Gaussian noise. In our experiments with synthetic and fMRI data, we implemented the encoder, decoder, and MLPs using three-layer fully connected networks, complemented by Leaky-ReLU activation functions. For optimization, the Adam optimizer was employed with a learning rate of 0.001. In the case of image data experiments, the prior model also utilized a three-layer fully connected network with Leaky-ReLU activation functions. The encoder and decoder designs are detailed in Table 1 and Table 2, respectively.

## L Experiments on High-Dimensional Synthetic Image Data

We further validate our proposed identifiability results and methodology using images from the chemistry dataset introduced by [27]. This dataset is representative of chemical reactions where the state of one element can influence the state of another. The images feature multiple objects with

| Layer | Output / Activation |
|---|---|
| Conv2d(3, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| Conv2d(32, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| Conv2d(32, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| Conv2d(32, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| Linear(32×32×4 + size($\mathbf{u}$), 30) | Leaky-ReLU |
| Linear(30, 30) | Leaky-ReLU |
| Linear(30, 3×2) | - |

Table 1: Encoder for the image data.

| Layer | Output / Activation |
|---|---|
| Linear(3, 30) | Leaky-ReLU |
| Linear(30, 30) | Leaky-ReLU |
| Linear(30, 32×32×4) | Leaky-ReLU |
| ConvTranspose2d(32, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| ConvTranspose2d(32, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| ConvTranspose2d(32, 32, 4, stride=2, padding=1) | Leaky-ReLU |
| ConvTranspose2d(32, 3, 4, stride=2, padding=1) | - |

Table 2: Decoder for the image data.

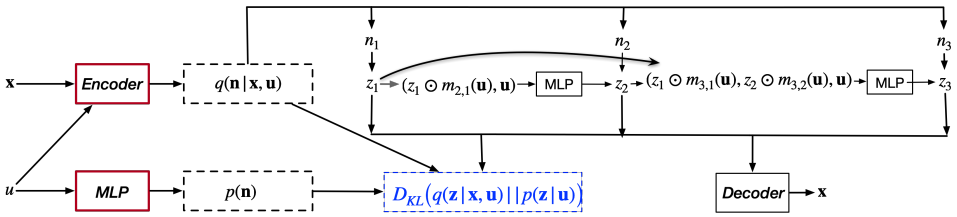



Figure 8: Implementation Framework to learn latent nonlinear models (i.e., MLP) with Gaussian noise. Here we demonstrate the method using 3 latent variables, however, our approach is versatile and can be effectively generalized to accommodate much larger graphs.

fixed positions, but their colors, representing different states, change according to a predefined causal graph. To align with our theoretical framework, we employ a nonlinear model with additive Gaussian noise for generating latent variables that correspond to the colors of these objects. The established latent causal graph within this context indicates that the 'diamond' object (denoted as $z_1$) influences the 'triangle' ($z_2$), which in turn affects the 'square' ($z_3$). Figure 9 provides a visual representation of these observational images, illustrating the causal relationships in a tangible format.

Figure 11 presents MPC outcomes as derived from various methods. Among these, the proposed method demonstrates superior performance. In addition, both the proposed method (MLPs) and Polynomials can accurately learn the causal graph with guarantee. However, Polynomial encounters issues such as numerical instability and exponential growth in terms, which compromises its performance in MPC, as seen in Figure 11. This superiority of MLPs is further evidenced in the intervention results, as depicted in Figure 10, compared with results of Polynomial shown in Figure 12. Additional traversal results concerning the learned latent variables from other methodologies are detailed in Figure 13 (VAE), Figure 14 ($\beta$-VAE) and Figure 15 (iVAE). For these methods without identifiability, traversing any learned variable results in a change in color across all objects.

## M Details and More Results of Experiments on Human Motion Data

### M.1 Preprocessing

We adopt a two-stage preprocessing pipeline to construct 2D pose sequences from the Human3.6M dataset. First, we extract ground-truth 3D joint positions provided in the dataset. Each 3D pose is transformed from world coordinates into the camera coordinate frame using the associated extrinsic

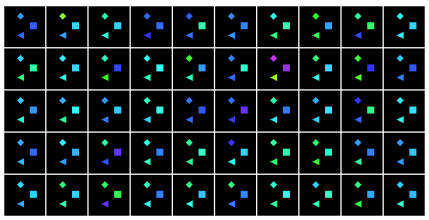


Figure 9: Samples generated by using a modified version of the chemistry dataset originally presented in [27]. In this adaptation, the objects' colors (representing different states) change in accordance with a specified causal graph, e.g., 'diamond' causes 'triangle', and 'triangle' causes 'square'.

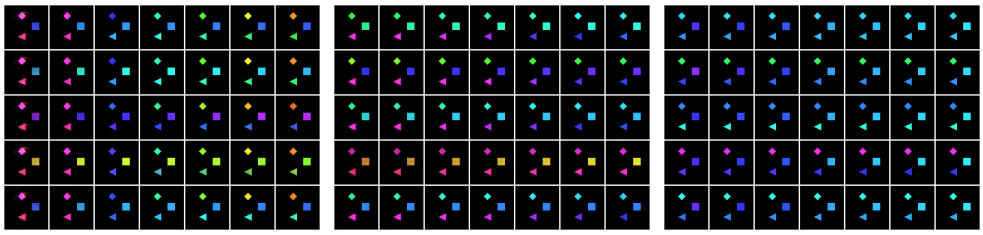


Figure 10: From left to right, the interventions are applied to the causal representations $z_1$, $z_2$, and $z_3$ learned by the proposed method (MLPs), respectively. The vertical axis represents different samples, while the horizontal axis represents the enforcement of various values on the learned causal representation.

parameters (rotation and translation). Subsequently, we apply a perspective projection using the intrinsic parameters (focal length and principal point), and the resulting 2D coordinates are converted into image-space pixel positions based on the camera resolution. This process follows the implementation provided by https://github.com/facebookresearch/VideoPose3D/blob/main/data/prepare_data_h36m.py. Corrupted sequences (e.g., `Directions` for subject S11) are excluded, and only valid data is retained. The final output is stored as structured 2D keypoint arrays indexed by subject, action, and camera.

In the second stage, for each unique (subject, action) pair, we keep only the sequence from the first camera view. A unique one-hot vector is assigned to each pair, which is then concatenated to the 2D joint coordinates of every frame across all joints. To ensure balanced representation among all subject-action categories, we uniformly sample the same number of frames from each sequence based on the shortest available sequence length. This balanced and encoded dataset is then prepared for subsequent training tasks. As a result, we obtain a final dataset comprising 105 contexts $i.e.$, $\mathbf{u}$, each containing 1040 frames, with each frame represented by 2D coordinates of 16 joints.

### M.2 Evaluation Details

After training, the proposed method learns (i) latent representations $\mathbf{z}$, and (ii) the causal relations among them, represented by a directed graph. To evaluate the learned representations, we perform *interventions* on each learned latent variable. In particularly, given a target variable $z_i$, we do an intervention by modifying its value while respecting the learned causal structure:

- Variables preceding $z_i$ in the causal order are kept unchanged;
- $z_i$ is replaced by an intervened value;
- Descendant variables of $z_i$ are updated according to the learned causal mechanisms.

The resulting intervened representations are passed through the decoder to obtain outputs in the observation space. By comparing these with the original input, we analyze how the intervention on $z_i$ affects the observation space.

Based on these intervention effects, we first examine whether changing $z_i$ induces changes in components associated with another latent variable $z_j$, which suggests a potential dependency. We further assess directional asymmetry: if intervening on $z_i$ affects $z_j$, but intervening on $z_j$ does not affect $z_i$, this provides evidence of a directional relationship from $z_i$ to $z_j$.

| | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|
| $\hat{z}_1$ | 0.089 | 0.094 | 0.857 |
| $\hat{z}_2$ | 0.606 | 0.620 | 0.070 |
| $\hat{z}_3$ | 0.811 | 0.681 | 0.042 |

| | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|
| $\hat{z}_1$ | 0.067 | 0.582 | 0.628 |
| $\hat{z}_2$ | 0.958 | 0.065 | 0.046 |
| $\hat{z}_3$ | 0.117 | 0.429 | 0.765 |

| | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|
| $\hat{z}_1$ | 0.095 | 0.631 | 0.683 |
| $\hat{z}_2$ | 0.156 | 0.758 | 0.705 |
| $\hat{z}_3$ | 0.980 | 0.126 | 0.028 |

| | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|
| $\hat{z}_1$ | 0.862 | 0.281 | 0.003 |
| $\hat{z}_2$ | 0.553 | 0.868 | 0.123 |
| $\hat{z}_3$ | 0.225 | 0.312 | 0.918 |

| | $z_1$ | $z_2$ | $z_3$ |
|---|---|---|---|
| $\hat{z}_1$ | 0.912 | 0.501 | 0.024 |
| $\hat{z}_2$ | 0.162 | 0.893 | 0.101 |
| $\hat{z}_3$ | 0.089 | 0.139 | 0.948 |

Figure 11: MPC obtained by different methods on the image dataset. From top to bottom and left to right: VAE, $\beta$-VAE, iVAE, Polynomials, and the proposed method (MLPs). The proposed method performs better than others, which is not only in line with our identifiability claims but also highlights the flexibility of MLPs.

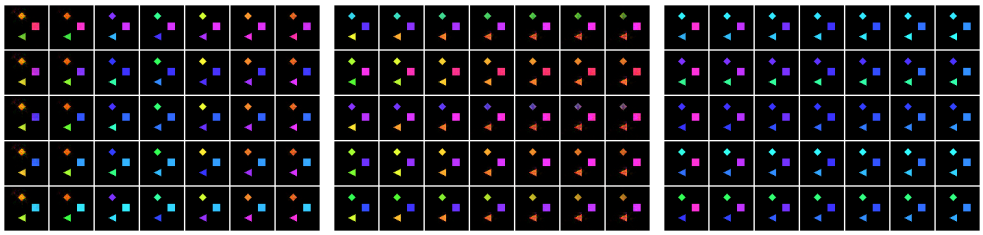


Figure 12: From left to right, the interventions are applied to the causal representations $z_1$, $z_2$, and $z_3$ learned by Polynomials, respectively. The vertical axis represents different samples, while the horizontal axis represents the enforcement of various values on the learned causal representation.

We then jointly consider these observations together with the learned causal graph. If the graph contains an edge $z_i \to z_j$, and the intervention experiments exhibit consistent directional effects, we interpret this as evidence that the proposed method has learned a plausible causal relationship from $z_i$ to $z_j$.

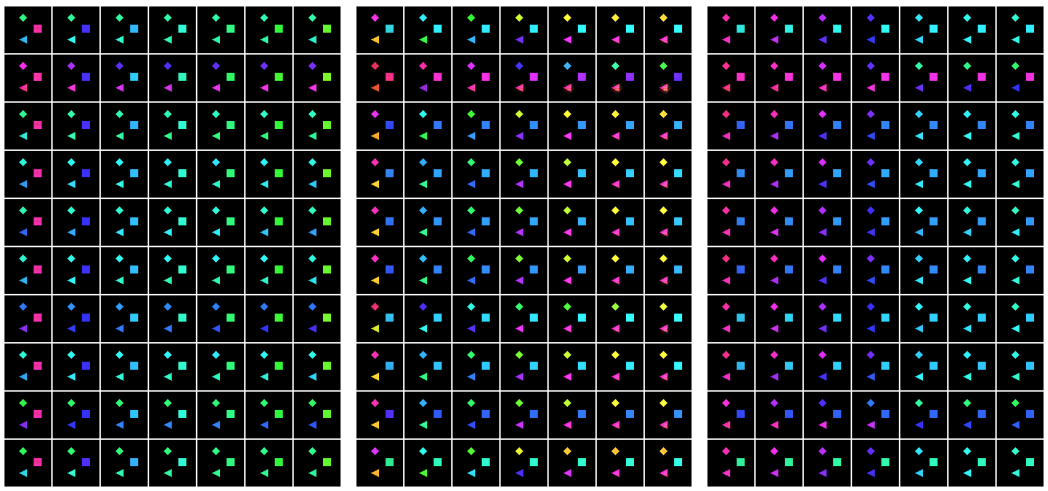
In the human motion setting, these relationships can be further supported by biomechanical principles. For example, proximal joints (e.g., shoulder and elbow) are known to influence distal joints (e.g., wrist) through intersegmental dynamics. Therefore, agreement between the learned relationships and such known dependencies provides additional qualitative support for the learned causal structure.

### M.3 More Results

In our implementation, we empirically set the number of latent variables to 14. The model is trained using the Adam optimizer with a learning rate of 1e-3 for 7000 epochs. We use the encoder designed to effectively encode 2D keypoint sequences. We employ an encoder to effectively encode 2D keypoints, where each input frame consists of multiple keypoint coordinates augmented with a subject-action condition vector $\mathbf{u}$. The input is first projected via a linear layer into a higher-dimensional feature space, enhancing its representational capacity. This is followed by a stack of Mixer layers, which alternate between mixing information across spatial (e.g., keypoint) and feature dimensions, thereby capturing complex dependencies both spatially and channel-wise. After all Mixer blocks, a layer normalization is applied to stabilize training. The used decoder applies multiple Mixer layers to iteratively mix spatial and channel information, followed by layer normalization for stable training. Finally, a linear layer projects the hidden features back to the keypoint coordinate dimension, producing an output that matches the original input shape of $2 \times 14$ keypoint coordinates, representing the reconstructed 2D coordinates. This decoder architecture symmetrically complements the encoder by reversing the compositional token embedding process, enabling effective recovery of keypoint positions from latent representations.

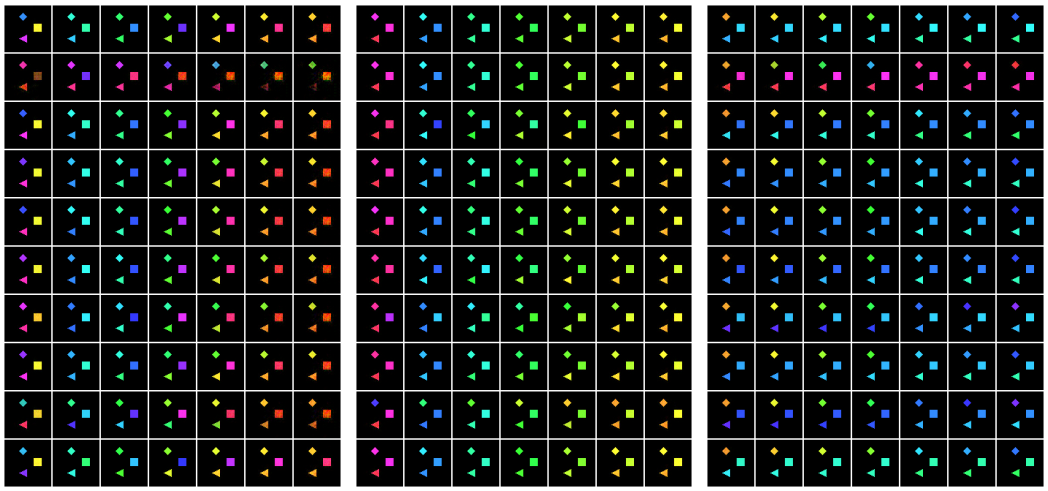
Figures 16–18 illustrate the results of intervention on each learned latent variables by our method. As discussed in the main manuscript, and supported by the estimated adjacency matrix shown in Figure 5, we observe that certain latent variables—specifically $z_1$ and $z_2$, $z_6$ and $z_7$, as well as $z_{10}$ and $z_{11}$—exhibit potential causal relationships. These include plausible dependencies such as from the shoulder to the wrist joint, and from the elbow to the wrist. Such findings are consistent with biomechanical principles of intersegmental limb dynamics.

For comparison, we also implemented the latent polynomial models proposed by [43]. As shown in Figures 19–21, the learned latent representations in this baseline tend to be more entangled, lacking the interpretable structure.



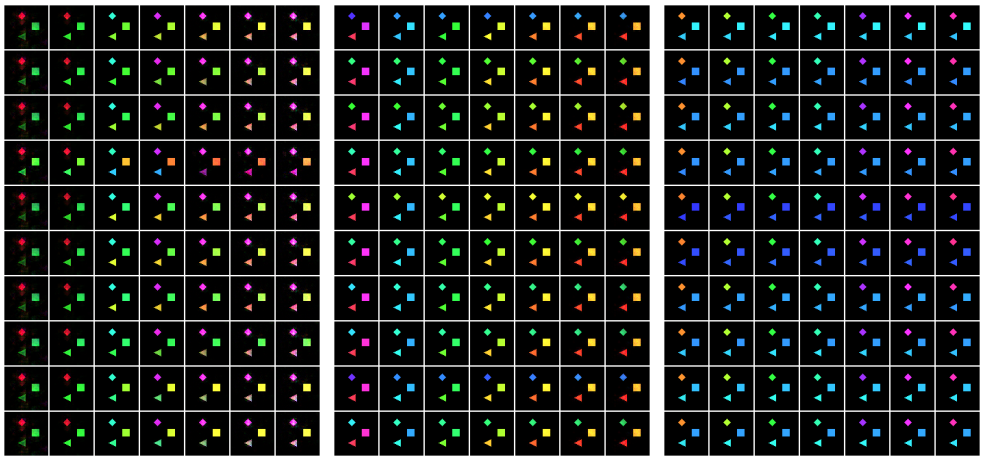
Traversals on $z_1$ Traversals on $z_2$ Traversals on $z_3$

Figure 13: The traversal results achieved using VAE on image datasets are depicted. On this representation, the vertical axis corresponds to different data samples, while the horizontal axis illustrates the impact of varying values on the identified causal representation. According to the latent causal graph's ground truth, the 'diamond' variable (denoted as $z_1$) influences the 'triangle' variable ($z_2$), which in turn affects the 'square' variable ($z_3$). Notably, modifications in each of the learned variables lead to observable changes in the color of all depicted objects.



Traversals on $z_1$ Traversals on $z_2$ Traversals on $z_3$

Figure 14: The traversal results achieved using $\beta$-VAE on image datasets are depicted. On this representation, the vertical axis corresponds to different data samples, while the horizontal axis illustrates the impact of varying values on the identified causal representation. According to the latent causal graph's ground truth, the 'diamond' variable (denoted as $z_1$) influences the 'triangle' variable ($z_2$), which in turn affects the 'square' variable ($z_3$). Notably, modifications in each of the learned variables lead to observable changes in the color of all depicted objects.



Traversals on $z_1$ Traversals on $z_2$ Traversals on $z_3$

Figure 15: The traversal results achieved using iVAE on image datasets are depicted. On this representation, the vertical axis corresponds to different data samples, while the horizontal axis illustrates the impact of varying values on the identified causal representation. According to the latent causal graph's ground truth, the 'diamond' variable (denoted as $z_1$) influences the 'triangle' variable ($z_2$), which in turn affects the 'square' variable ($z_3$). Notably, modifications in each of the learned variables lead to observable changes in the color of all depicted objects.




Figure 16: Complete Results of Intervention on the Estimated Latent Variables $z_1$ to $z_5$ by the Proposed Method.




Figure 17: Complete Results of Intervention on the Estimated Latent Variables $z_6$ to $z_{10}$ by the Proposed Method.

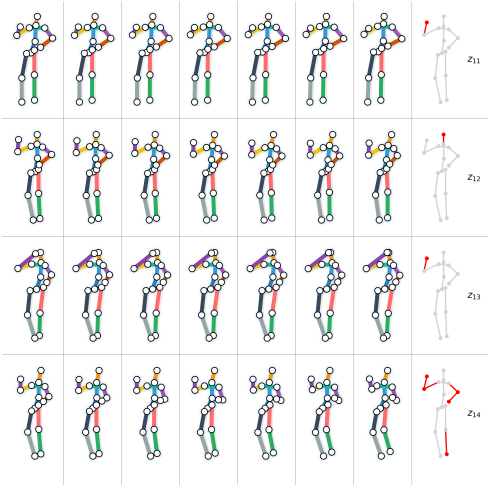



Figure 18: Complete Results of Intervention on the Estimated Latent Variables $z_{11}$ to $z_{14}$ by the Proposed Method.

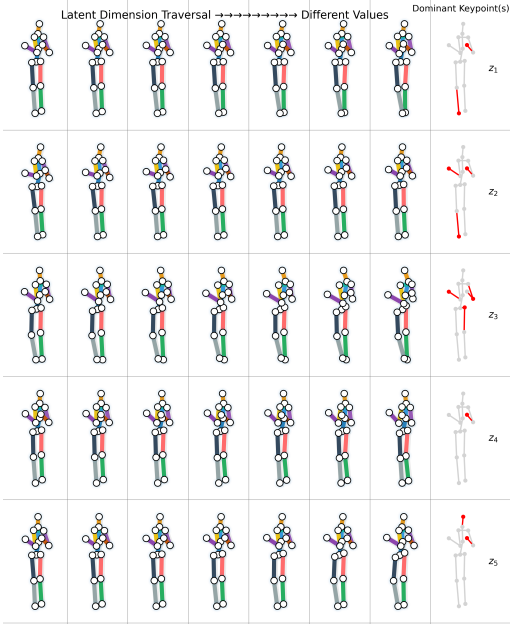



Figure 19: Complete Results of Intervention on the Estimated Latent Variables $z_1$ to $z_5$ by Latent Polynomial Model.

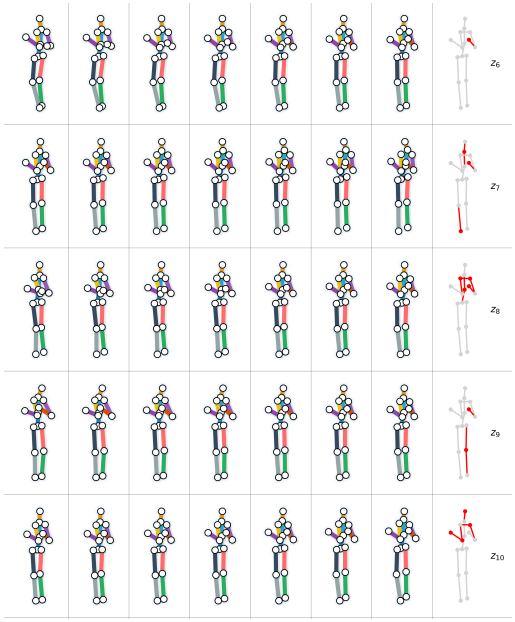



Figure 20: Complete Results of Intervention on the Estimated Latent Variables $z_6$ to $z_{10}$ by Latent Polynomial Model.




Figure 21: Complete Results of Intervention on the Estimated Latent Variables $z_{11}$ to $z_{14}$ by Latent Polynomial Model.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately summarize both the theoretical contributions (identifiability conditions and extensions) and the proposed method with its empirical validation, without overstating the scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations such as reliance on theoretical assumptions that are not directly verifiable in practice, similar to most of work in the community.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical results are stated with explicit assumptions, and complete proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of datasets, preprocessing steps, model design, training objectives, and evaluation protocols, with additional implementation details included in the appendix, enabling reproducibility of the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code and implementation details are not publicly released at the time of submission, but will be released upon acceptance.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies dataset splits, preprocessing, model architecture, training objectives, hyperparameters, and optimization details, with additional implementation specifics provided in the appendix to ensure clarity and understanding of the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments are conducted with random seeds to ensure reproducibility, and multiple runs are used where applicable to verify the stability of the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24GB memory. We do not report detailed computational cost, as the experiments are relatively lightweight and complete within a short time.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer:[Yes]

   Justification: The research complies with the NeurIPS Code of Ethics, as it does not involve human subjects, sensitive data, or any unethical practices.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper does not explicitly discuss broader societal impacts, as the work is primarily theoretical and methodological, with no direct or immediate societal deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not involve the release of high-risk data or models, and therefore safeguards for responsible release are not applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer:[Yes]

    Justification: All datasets and prior works used in the paper are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, models, or other assets requiring documentation.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or experiments with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects or participant-based studies, and thus no IRB approval or risk disclosure is required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The paper does not use LLMs as part of the core methodology or experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.