Towards Causal Representation Learning with Observable Sources as Auxiliaries

Kwonho Kim[1] Heejeong Nam[2] Inwoo Hwang[3] Sanghack Lee[1]

[1]Graduate School of Data Science, Seoul National University, Seoul, South Korea
[2]Boeing Korea, Seoul, South Korea
[3]Causal Artificial Intelligence Lab, Columbia University, New York, USA

Abstract

Causal representation learning, particularly in the context of nonlinear independent component analysis, aims to uncover the underlying latent variables from observed data, providing critical insights into the true generative processes. However, achieving the identifiability of these latent variables has been an obstacle due to the possibility of infinite spurious solutions. Prior works often rely on auxiliary variable assumptions that enforce conditional independence among latents. However, they require that auxiliary variables not be involved in the mixing function—a constraint that significantly limits the applicability in real-world settings as it is often difficult to obtain suitable label data that can serve as external side information. In this work, we address this challenge by leveraging observable sources as auxiliary variables, a more practical scenario. We also propose a novel framework that selects proper auxiliary variables to improve the recoverability of the latents while ensuring that identifiability conditions are satisfied. To the best of our knowledge, this is the first work to demonstrate identifiability under this setting, offering a more practical solution for causal representation learning. By exploiting the graphical structure of the latent variables, we enhance both identifiability and recoverability, extending the boundaries of current approaches to causal representation learning.

1 INTRODUCTION

Understanding the underlying generative process of observations is crucial for scientific discovery. In this context, causal representation learning (CRL) [Schölkopf et al., 2021], including nonlinear independent component analysis (ICA) [Hyvärinen et al., 2009], aims to recover latent variables from observed data. This approach holds significant promise for applications in areas such as healthcare [Sanchez et al., 2022], climate science [Yao et al., 2024a], and recommendation [Wang et al., 2022, 2024, Yang et al., 2024], as understanding the causal mechanisms can lead to better interpretability and improved generalization to new settings.

However, without proper assumptions, infinitely many spurious solutions could exist, yielding independent mixtures of the true sources [Hyvärinen and Pajunen, 1999]. This has made unsupervised learning of disentangled representations challenging from a theoretical perspective [Locatello et al., 2019]. Accordingly, recent studies employed the assumption that sources are conditionally independent given observed auxiliary variables $\mathbf{u}$, as shown in Fig. 1a [Hyvärinen and Morioka, 2016, Khemakhem et al., 2020], thereby achieving identifiability. Although the conditional independence assumption enables identifiability, it is often violated in practice, as dependencies between sources are frequently encountered in real-world settings such as computational biology [Cardoso, 1998, Theis, 2006].

Several approaches have attempted to relax the assumption of conditional independence by addressing the dependence on the source.source. For example, Lu et al. [2022] achieved identifiability upto component-wise transformation under source dependence by assuming a structured exponential family form with a parametric decomposition of sufficient statistics into factorizable part and correlated part. Zheng and Zhang [2023] also achieved the identification of latent sources under source dependence, up to a subspace-wise invertible transformation and permutation, given auxiliary variables but without relying on parametric forms, thereby providing a non-parametric generalization.

While there has been progress in addressing source dependence, *the use of observed sources as auxiliary variables*—specifically, those that participate in the mixing function—remains largely underexplored. This can pose a problem in the identifiability proof, where the log-determinant term cannot be eliminated. Nevertheless, using observed

Accepted for the 1st Workshop on Causal Abstractions and Representations (CAR) at UAI 2025 (CAR 2025).

sources as auxiliary variables is especially important in practical scenarios, as it reflects more general and realistic data-generating processes. This approach also opens up the possibility of extracting the auxiliary information directly from the data, even when explicit auxiliary variables are not available.

One illustrative case is that of a robotic arm carrying out a manipulation task The underlying latent sources may correspond to various physical parameters, such as the joint angles, torques, or the forces exerted by the arm, while the observed variables could consist of camera images capturing the robot's movements. In this context, we can treat arm angle information directly extracted from image data as observable sources, which provide only partial information about the true latent variables governing the system.

Moreover, scientific systems like robotic arms that are governed by physical laws can often be represented using causal graphs [Mooij et al., 2013, Baumann et al., 2022]. In such cases, the conditional independence relations implied by the graph (via d-separation) can reveal which subsets of latent variables are identifiable. For example, fixing the angle of a specific joint in a robotic arm can render the movements of the joints before and after it conditionally independent.

When considering the causal graph as more general data-generating process, the conditional independence between the latents can vary depending on which variables are conditioned. Thus, selecting proper auxiliary variables can determine the degree of identifiability. However, this topic, i.e., *how to exploit/select auxiliary variables leveerating graphical information*, has remained unexplored in recent studies.

We summarize our contributions as follows:

- To the best of our knowledge, we are the first to propose generalized setting and achieve identifiability with observed sources as auxiliaries in the context of causal representation learning.
- We introduce a principled framework for selecting a subset of auxiliary variables—when multiple are available—that maximizes identifiability, leveraging the conditional independence structure of the latent causal graph.
- We empirically validate our approach across various experimental settings, demonstrating that the representation effectively disentangles according to the conditional independence structure of the latent graph.

2 PRELIMINARY

We formalize *observed sources as auxiliaries* in terms of nonlinear ICA and causal representation learning to construct our problem setting. We use upper case letters for random variables or vectors, and lower case for their assignments. Bold letters represent a set of random variables or random variables which is not a singleton. We use $[d]$ to denote a set $\{1, 2, \cdots, d\}$.

2.1 OBSERVED SOURCES AS AUXILIARIES

Let $\mathbf{x} \in \mathbb{R}^m$ be an observation (e.g., image) which are generated from latent sources $\mathbf{z} \in \mathbb{R}^n$ with a mixing function g as follow:

$$\mathbf{x} = g(\mathbf{z}). \tag{1}$$

where g is an arbitrary invertible and smooth nonlinear function in the sense that its second-order derivatives exist. By adopting a Bayesian network, we represent a data-generating process regarding latent sources as

$$z_i = f_i(\mathrm{Pa}^{\mathcal{G}}(z_i), \epsilon_i), \quad \epsilon_i \sim p_{\epsilon_i}, \tag{2}$$

for all $i \in [n]$ where $\mathrm{Pa}^{\mathcal{G}}(\cdot)$ represents parent nodes on a known causal graph $\mathcal{G}$ consisting of nodes V and edges E. Nonlinear ICA considers independent latent sources, i.e.,

$$p(\mathbf{z}) = \prod_{i=1}^{n} p(z_i). \tag{3}$$

The primary goal is to recover inverse function g^{-1} and the true independent components $\mathbf{z} = (z_1, \cdots, z_n)$ solely from observations. However, the model is known to be unidentifiable only with i.i.d samples [Hyvärinen and Pajunen, 1999]. This means that the mapping from observations to independent sources cannot be uniquely determined based solely on the assumption of mutually independent sources.

Accordingly, various conditions have been explored to address non-identifiability. The most widely used condition is independence given auxiliary observable variables $\mathbf{u}$, i.e.,

$$p(\mathbf{z} \mid \mathbf{u}) = \prod_{i=1}^{n} p(z_i \mid \mathbf{u}), \tag{4}$$

where $\mathbf{u}$ can be a class label, time index, or historical information [Hyvarinen et al., 2019]. Fig. 1a illustrates a data generating process where u is the observable and $\mathbf{z} = (z_1, z_2, z_3, z_4)$ is the latent sources that generates the observation. However, they rely on the critical assumption that the auxiliary variable does not have a direct influence to the observation $\mathbf{x}$ to establish identifiability.

In this work, we consider a more generalized setup where auxiliary variables may directly participate in the mixing function, rather than being restricted to external side information. Specifically, we treat auxiliary variables as observed latent sources $\mathbf{z}_o \subset \mathbf{z}$ that directly participate in the mixing function, where the generative process is governed by a DAG $\mathcal{G}$ capturing arbitrary dependencies. Note that CRL also aims to recover the latent sources with arbitrary dependencies, but it remains unclear how to leverage observed

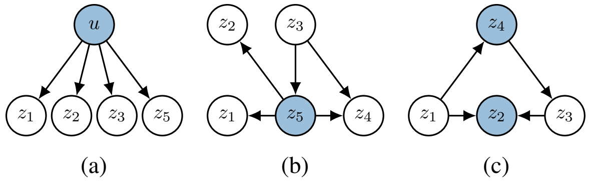


Figure 1: Examples of data generating process except mixing process, i.e. latent mechanism. Blue nodes represent the observable variables.

sources. We will discuss related works in nonlinear ICA and CRL in Sec. 5.

To deal with the dependence between sources, we can partition the latent sources into conditionally independent sets, $\mathbf{z}_{c_i}(i = 1, ..., d)$ where $\cup_{i=1}^{d} \mathbf{z}_{c_i} = \{z_1, \ldots, z_n\}$.Thus, Eq. (4) can be restated in the context of observed sources as auxiliaries:

$$p_{\mathbf{z}_{o^-}|\mathbf{z}_o}(\mathbf{z}_{o^-}|\mathbf{z}_o) = \prod_{j=1}^{d} p_{\mathbf{z}_{c_j}|\mathbf{z}_o}(\mathbf{z}_{c_j}|\mathbf{z}_o). \tag{5}$$

where $\mathbf{z}_o$ is observed sources and $\mathbf{z}_{o^-}$ is unobserved sources.

However, we do not consider the case where $\mathbf{z}_o$ is an empty set (i.e., an empty latent graph, without auxiliary variables), as this scenario requires additional assumptions such as structural sparsity for identifiability [Zheng and Zhang, 2023].

2.2 PROBLEM FORMULATION

Suppose a data generating process in Eq. (1) and Eq. (2). Our goal is to establish the identifiability of the independent latent sources (i.e., $\mathbf{z}_{c_i}$) up to certain subspace-wise invertible transformation and permutation, given the observations $\mathbf{x}$, observable sources $\mathbf{z}_o(\subseteq \mathbf{z})$, and the latent Bayesian network $\mathcal{G}$ which encodes the conditional independence relationships between the latent sources as shown in Fig. 1.

In contrast to previous works exploiting a specific factorization of the latent sources, e.g., Eq. (3) or Eq. (12), the knowledge of the latent Bayesian network $\mathcal{G}$ allows us to leverage diverse conditional independence relationships between the sources. Importantly, the partition of the latent sources into subspaces $\mathbf{z}_{o^-} = \cup_i \mathbf{z}_{c_i}$ determines the *degree* of the identifiability we could achieve (Thm. 4.3 of Zheng and Zhang [2023]). Therefore, it is crucial to capture the proper observable sources $\mathbf{z}_u \subseteq \mathbf{z}_o$ that entails fine-grained subspaces $\mathbf{z}_{c_j}$ mutually independent to each other conditioned on $\mathbf{z}_u$.

2.3 MOTIVATING EXAMPLES

We illustrate the concept of a *proper* auxiliary variable and *fine-grained* subspaces through a representative example in Fig. 1.

In Fig. 1b, it is straightforward to find the most fine-grained conditionally independent groups, i.e. $\{z_1\}$, $\{z_2\}$, $\{z_4, z_5\}$. However, as the number of sources or observed sources increases, finding auxiliary variables that make the latent grouping fine-grained can become computationally challenging.

In addition, there is an another problem in exploiting auxiliary variables. Considering the case where the multiple sources are observed as Fig. 1c, if all observed sources are conditioned, z_1 and z_3 cannot be disentangled. z_2 should not be conditioned to satisfy conditional independence. Although

It is true that dillema of which node to condition on typically arises in collider structures, but as the graph grows larger and more complex, the same node can act as a confounder for one group of nodes and as a collider for another. In such case, it becomes necessary to carefully choose which nodes to condition on in order to achieve optimal results.

3 METHOD

In this section, we establish identifiability in the presence of observable sources (Sec. 3.1). Based on conditions for identifiability, we introduce a framework with a graphical criterion to effectively leverage auxiliary variables that makes the conditionally independent latents more fine-grained (Sec. 3.2) and method to recover unobserved latents (Sec. 3.3).

3.1 IDENTIFIABILITY

We can consider existing approaches with auxiliary variables as the case that observed sources $\mathbf{z}_o$ do not have edges into the observations $\mathbf{x}$. (See Prop. 2 in Appendix B.) To deal with problems that the observed sources $\mathbf{z}_o$ are included in the mixing function, we assume that the mixing function is constrained to a specific form as Yang et al. [2022]. We adopt the proof of Zheng and Zhang [2023] for identifiability with dependent sources.

Proposition 1. *Suppose the following assumptions hold:*

1. *The observed data and sources are generated from Eq.* (1) *and Eq.* (5)
2. *The mixing function g is volume-preserving, i.e.,* $|det(\mathbf{J}_g(\mathbf{z}))| = 1$
3. *For every value of $\mathbf{z}_{o^-}$, there exists $2d$ values of $\mathbf{z}_o$, such that the $2d$ vectors $\mathbf{w}(\mathbf{z}_{o^-}, \mathbf{z}_{o_i})$ are linearly independent, where vector $\mathbf{w}(\mathbf{z}_{o^-}, \mathbf{z}_{o_i})$ is defined as fol-*

lows:

$$\mathbf{w}(\mathbf{z}_{o^-}, \mathbf{z}_{o_i}) = \big(\mathbf{v}(\mathbf{z}_{c_1}, \mathbf{z}_{o_i}), \ldots, \mathbf{v}(\mathbf{z}_{c_d}, \mathbf{z}_{o_i}), \\ \mathbf{v}'(\mathbf{z}_{c_1}, \mathbf{z}_{o_i}), \ldots, \mathbf{v}'(\mathbf{z}_{c_d}, \mathbf{z}_{o_i})\big)$$

where

$$\mathbf{v}(\mathbf{z}_{c_j}, \mathbf{z}_{o_i}) = \left(\frac{\partial \log p(\mathbf{z}_{c_j}|\mathbf{z}_{o_i})}{\partial z_{c_j}^{(l)}}, \ldots, \frac{\partial \log p(\mathbf{z}_{c_j}|\mathbf{z}_{o_i})}{\partial z_{c_j}^{(h)}}\right),$$

$$\mathbf{v}'(\mathbf{z}_{c_j}, \mathbf{z}_{o_i}) = \left(\frac{\partial^2 \log p(\mathbf{z}_{c_j}|\mathbf{z}_{o_i})}{\partial (z_{c_j}^{(l)})^2}, \ldots, \frac{\partial^2 \log p(\mathbf{z}_{c_j}|\mathbf{z}_{o_i})}{\partial (z_{c_j}^{(h)})^2}\right)$$

and $\mathbf{z}_{c_j} = (z_{c_j^{(l)}}, \ldots, z_{c_j^{(h)}})$.

Then all the components of $\mathbf{z}_{o^-}$ *(i.e.,* $\mathbf{z}_{c_i}$ *where* $c_i \in \{c_1, \ldots, c_d\}$*) is identifiable up to a subspace-wise invertible transformation and a subspace-wise permutation.*

Most prior works on CRL achieve identifiability by assuming the mixing function is fixed across environments. Under a common invertible mixing g, one can write the log-likelihood of the data under two domains and subtract them, causing the Jacobian log-determinant terms to cancel (since the same g applies in both cases). In such settings the latent distributions change across domains while g remains invariant, so the log-determinant terms disappear when differencing log-likelihoods.

In our setting, by contrast, the observed source (domain label) is used to index the mixing function, so g varies with the source. As a result the usual cancellation does not occur and the standard identifiability proof breaks down. To address the problem, we constrain the mixing to be volume-preserving (i.e., $|\det(\mathbf{J}_g(\mathbf{z}))| = 0$ everywhere). With the volume-preserving assumption on the mixing function, the Jacobian determinant remains constant at 1, making the log-determinant term equal to zero. Details of the proof are in Appendix B.

3.2 SELECTION ON OBSERVABLES

According to the Prop. 1, the conditional independence determines the number of recoverable sources in the identifiability of latent variables and our goal is first to capture mutually independent groups of nodes given observable sources and the known causal graph. However, a naive approach of leveraging all observed sources might not capture conditional independence relationships, i.e., $z_1 \not\perp\!\!\!\perp z_3 \mid z_2, z_4$ in Fig. 1c. It is necessary to capture a proper subset of observed sources that entails the most fine-grained groups of mutually independent sources, and ultimately, leads to the most granular identifiability.

Formally, we aim to discover a conditional independence structure that partition $\mathbf{z}_{o^-}$ into the most fine-grained subgroups such that:

Algorithm 1 Selection on Observables

```
Input: graph $G$, observed set $O$
Output: conditioning set $C$
$C \leftarrow$ {nodes acting only as confounders on $G$ }
$O \leftarrow O \setminus$ {nodes acting only as colliders on $G$ }
$max \leftarrow 0$
for each subset $T \subseteq O$ do
    $S \leftarrow Partition(G, T, O)$
    if $|S| > max$ or ($|S| = max$ and $|T| < |C|$) then
        $max \leftarrow |S|$
        $C \leftarrow T$
    end if
end for
return $C$
```

$$\mathbf{z}_{c_i} \perp\!\!\!\perp \mathbf{z}_{c_j} \mid \mathbf{z}_u, \quad \text{for all } i \neq j, \tag{6}$$

where $\mathbf{z}_u \subseteq \mathbf{z}_o$, $\cup_i \mathbf{z}_{c_i} \subseteq \mathbf{z} \setminus \mathbf{z}_o$, and $\mathbf{z}_{c_i} \cap \mathbf{z}_{c_j} = \emptyset$ for all $i \neq j$. Importantly, satisfying a fine-grained conditional independence condition enables the identification of a greater number of latent variables. This ensures a more precise disentanglement of the underlying causal structure, leading to improved recoverability and manipulability of the true latent factors.

We propose a strategy that selects the most fine-grained conditionally independent groups of the latents with the minimum set of observed sources in Alg. 1. The algorithm initializes a candidate set by including only nodes that act as confounders and excluding those that act solely as colliders, in order to account for nodes that may serve as both. The *Partition* algorithm counts the number of groups that satisfy conditional independence by running *Bayes-ball* [Shachter, 1998] algorithm repeatedly. Finally, the algorithm outputs the conditioning set that results in the largest number of groups, i.e., the most fine-grained partitioning.

Example Consider the latent graph in the Fig. 1c. Observed set $O = \{z_2, z_4\}$. We will iterate all the subsets of O, i.e., $\{z_2\}, \{z_4\}, \{z_2, z_4\}$.[1] Firstly, with conditioning set $\{z_2\}$, the partition process is as follow:

1. Started from z_1, the *result* contains z_1.
2. *Bayes-ball* algorithm get input as G, $\{z_2\}$ and *result*.
3. In the *Bayes-ball*, path from z_1 to z_3 through z_2 cannot be d-separated because z_2 works as collider.
4. The path from z_1 to z_3 also cannot be d-separated.
5. The *result* is $\{\{z_1, z_3\}\}$ except for observed source z_2.

With conditioning set $\{z_4\}$, by following same process, the *result* will be $\{\{z_1\}, \{z_3\}\}$. The conditioning set $\{z_2, z_4\}$

[1] $\emptyset$ cannot be considered due to the condition for the identifiability.

makes the result to be $\{\{z_1, z_3\}\}$. Hence, the selection result will be $\{z_4\}$ for the most fine-grained conditionally independent latents.

3.3 LEARNING TO RECOVER

To construct a representation that satisfies the identifiability conditions in Prop. 1, we enforce volume preservation in the encoder by adopting General Incompressible-flow Network (GIN) [Sorrenson et al., 2020] as our encoder. In addition to volume preservation, we also impose a graphical constraint via a structural neural network to preserve dependencies among latent variables that are not assumed to be independent, reflecting the known latent causal structure to strengthen disentanglement.

Volume-preservation While GIN originally optimizes only the log-likelihood of the conditional distribution given the auxiliary variables, we factorize the log-likelihood of the distribution as follows:

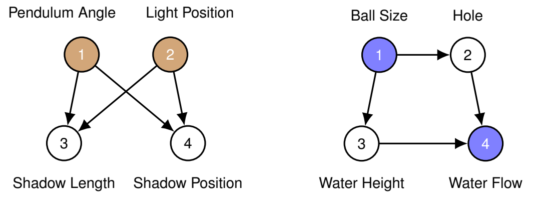
$$\log p_{\hat{g}^{-1}}(\mathbf{x}) = \log p(\hat{\mathbf{z}}) = \log p(\mathbf{z}_u) + \sum_i \log p(\hat{\mathbf{z}}_{u_i^-} \mid \mathbf{z}_u),$$

where $\hat{\mathbf{z}}_{u_i^-} = \hat{\mathbf{z}} \setminus \hat{\mathbf{z}}_{u_i}$. By factorizing the log-likelihood of the distribution, we can naturally address the issue that the information from the auxiliary variable is directly entangled with the observations. The preceding term will serve to absorb information about $\mathbf{z}_u$ from $\mathbf{x}$ while the latter term enforces the components of $\mathbf{z}_{u^-}$ to be independent given $\mathbf{z}_u$ by modeling them as a multivariate normal distribution with zero off-diagonal elements.

Graphical constraint Besides, $\hat{\mathbf{z}}_{u^-}$ contain the information of sources that are observed but not selected (expressed as $\mathbf{z}_n$), i.e., $\hat{\mathbf{z}}_{u^-} = \{\hat{\mathbf{z}}_{o^-}, \hat{\mathbf{z}}_n\}$. We need to keep the relationship between $\hat{\mathbf{z}}_n$ and $\hat{\mathbf{z}}_{o^-}$ which is not independent (relationship between z_2 and z_1, z_3 in Fig. 1c).

To deal with this problem, we leverage the structural neural net to enforce the relationship between $\hat{\mathbf{z}}_{o^-}$ and $\hat{\mathbf{z}}_n$. A structural neural network is designed based on the latent graph $\mathcal{G}$ and not selected label $\mathbf{z}_n$. Specifically, $\mathbf{z}_n$ is predicted by arbitrary dimensions of $\hat{\mathbf{z}}_{u^-}$ working as parents of $\mathbf{z}_n$. Since we do not know exactly which dimension of the representation corresponds to which true latent variable, we rely only on the number of parents of $\mathbf{z}_n$. For example, in Fig. 1c, true z_2 is predicted by the certain dimension of the estimated representation given the other dimensions $(\hat{z}_1, \hat{z}_3)$, naturally reflecting the causal structure. The full objective function is:

$$\mathcal{L}(\theta) = \mathbb{E}_{(\mathbf{x}, \mathbf{z}_u, \mathbf{z}_n) \in \mathcal{D}} \Bigg[\log p(\mathbf{z}_u) + \sum_i \log p(\hat{\mathbf{z}}_{u_i^-} \mid \mathbf{z}_u) + \log p(\mathbf{z}_n \mid \mathrm{Pa}^{\mathcal{G}}(\mathbf{z}_n)) \Bigg]. \quad (7)$$



(a) Pendulum System (b) Flow System

Figure 2: Causal graphs for two systems. Colored nodes are observed sources: (a) Pendulum and (b) Flow.

4 EXPERIMENT

We conduct experiments to empirically validate both the effectiveness of the selection procedure and the capability of our proposed architecture in leveraging observable sources.

4.1 EXPERIMENTAL SETUP

Data Reflecting the setup of observable sources, we consider synthetic datasets generated from the three graphs in Fig. 1: $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{z}_o^{(i)})\}_{i=1}^N$, where N is the sample size and $\mathbf{z}_o^{(i)}$ is the observed sources corresponding to the data point $\mathbf{x}^{(i)}$. When we run our selection procedure given a graph to choose the best combination of the auxiliary variables, $\mathbf{z}_o$ will be partitioned into $\mathbf{z}_u$ and $\mathbf{z}_n$.

The data was generated using a linear Structural Causal Model (SCM) where each variable is determined by a linear combination of its parents and an additive noise term:

$$X_i = \sum_{j \in \mathrm{pa}(X_i)} \beta_{ij} X_j + \varepsilon_i, \quad (8)$$

where β_{ij} are sampled uniformly from $[0.5, 1.0]$, and ε_i is the additive noise term with coefficient fixed to 1.0.

To further demonstrate the effectiveness of our method on high-dimensional data, we used the Pendulum and modified Flow datasets Yang et al. [2021], which consist of structured, systematically sampled image data. Corresponding latent causal graphs are shown in Fig. 2. The implementation details is in Appendix D.

Metrics After training the proposed method, we measure Disentanglement, Completeness, Informativeness (DCI) metric [Eastwood and Williams, 2018] based on Mean Correlation Coefficient (MCC) matrix which is a widely accepted metric in the literature for measuring the degree of identifiability [Hyvärinen and Morioka, 2016]. We assess how well the learned representation aligns with the independence structure of the underlying graph.

Specifically, the MCC matrix is defined as:

$$\mathbf{MCC}_{ij} = \mathrm{corr}(z_i, \hat{z}_j), \quad (9)$$

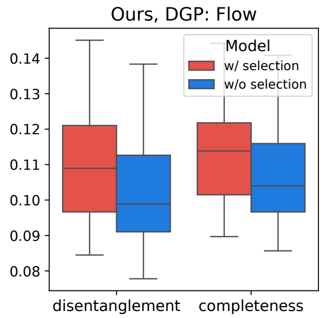


Figure 3: Ablation study on the selection procedure for Ours.

where each entry $\mathbf{MCC}_{ij}$ represents the Pearson correlation coefficient between the true latent variable z_i and the estimated latent variable $\hat{z}_j$. The optimal permutation σ^* is selected to maximize the total correlation, ensuring that each estimated latent variable is matched to the most similar true latent variable.

Based on the computed MCC matrix, we evaluate models with Disentanglement and Completeness among DCI metrics:

$$D = 1 - H(P_{i,\cdot}), \tag{10}$$

$$C = 1 - H(P_{\cdot,j}), \tag{11}$$

where $P_{i,j}$ is the value from the MCC matrix, representing the contribution of the estimated latent variable $\hat{z}_j$ to the true latent factor z_i. The entropy function $H(\cdot)$ measures the dispersion of importance values across dimensions, ensuring that a lower entropy corresponds to a more structured and disentangled representation. Disentanglement (D) quantifies whether each estimated latent variable captures at most one true latent factor, computed by applying row-wise entropy over $P_{i,\cdot}$. Completeness (C) assesses whether each true latent factor is captured by a single estimated latent variable, computed via column-wise entropy over $P_{\cdot,j}$. Since both scores range from 0 to 1, higher values indicate better structured representations with minimal mixing between factors. Further discussions on why MCC alone is insufficient for evaluation are provided in the Appendix A. All the metrics are measured over 20 repetitions.

4.2 EMPIRICAL RESULTS

Effectiveness of selection We conducted an ablation study on the selection procedure for our architecture. The experiments are based on the data-generating process illustrated in Fig. 2b, where the differences in results arise depending on the selection procedure. Fig. 3 shows the change in DCI metric for our model before and after selection. The selection procedure improves disentanglement in the representation as shown in Fig. 3. For the graph in Fig. 2b, using all observed sources as auxiliary variables without a selection procedure breaks the conditional independence between **Water Height** and **Hole**, leading to entangled representations.

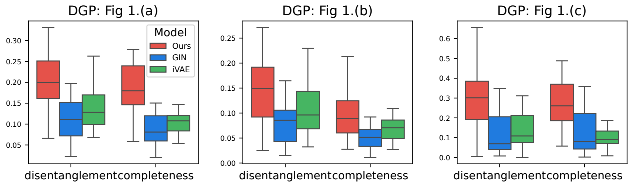


Figure 4: Comparison plot for DCI metric between Ours, GIN, and iVAE.

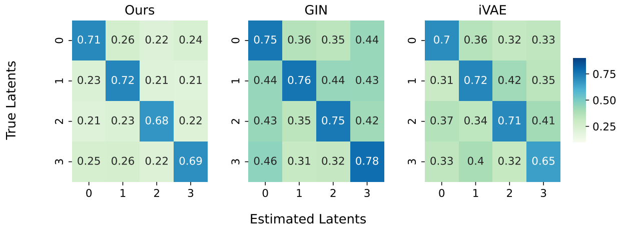


Figure 5: Mean correlation matrices of Ours, GIN applying selection, and iVAE applying selection matched with the best permutation on the setting of Fig. 1a.

Effectiveness of architecture To verify the effectiveness of our proposed architecture, we choose GIN [Sorrenson et al., 2020] and iVAE [Khemakhem et al., 2020] as baseline models. GIN is used as the encoder in our architecture, ensuring the volume-preserving property but not designed to handle observed sources. iVAE is also not designed to handle partially observed sources. Furthermore, it does not impose any constraints on the mixing function and solely relies on a multivariate normal distribution as the prior, ensuring that each latent variable is conditionally factorizable. For a fair comparison, all experiments are conducted using the same auxiliary variables filtered through the selection procedure.

Fig. 4 demonstrates that our proposed method outperforms other approaches in terms of the DCI metric. Our proposed method maximizes the likelihood of a conditionally factorizable distribution for the remaining components while simultaneously excluding the information of auxiliary variables mixed with the observation $\mathbf{x}$. This prevents spurious correlations in the representation by ensuring that the information of $\mathbf{z}_u$, which is related to unobserved latents, does not mix into the representation.

We further analyzed the results with the MCC matrix for a more detailed examination. The proposed architecture shows a comparable MCC score (mean of diagonal terms) as GIN and iVAE, as illustrated in Fig. 5 for the DGP of Fig. 1a. However, looking at the MCC matrix, we can observe that both GIN and iVAE show high correlations with the other latents besides the true latent, even when matched

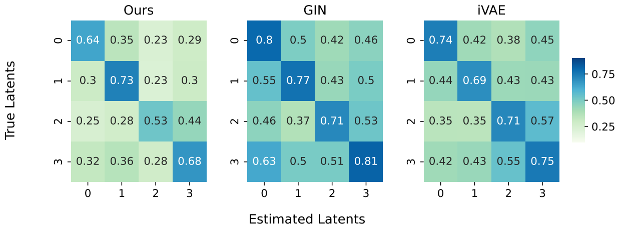


Figure 6: Mean correlation matrices of Ours, GIN applying selection, and iVAE applying selection matched with the best permutation on the setting of Fig. 1b.

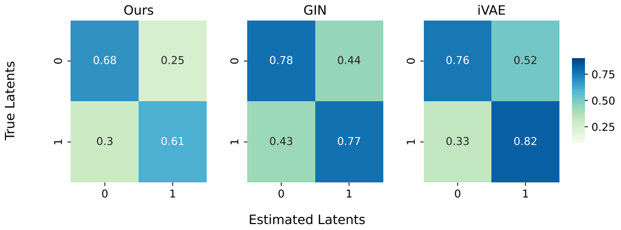


Figure 7: Mean correlation matrices of Ours, GIN applying selection and iVAE applying selection matched with the best permutation on the setting of Fig. 1c.

with the best permutation. This suggests that manipulating a specific dimension of the representation simultaneously affects other latents, indicating that the representation is not well disentangled.

Considering the DGP in Fig. 1b, the ideal disentanglement is that $\hat{z}_1$, $\hat{z}_2$, ($\hat{z}_3$, and $\hat{z}_4$) are conditionally independent. The result of our architecture for MCC matrix in Fig. 6 represents the almost ideal disentanglement, while the other methods still show entangled results. As the conditional independence in DGP in Fig. 1b does not ensure each latent to be identified, but block-identified, the MCC score might be lower. Even in this case, GIN and iVAE, which do not consider observed sources, show a high MCC score because of spurious correlation. Likewise, on the DGP (Fig. 1c), our architecture yields a disentangled MCC matrix as expected.

High-Dimensional data We also conduct the experiments on the Pendulum and Flow datasets from Yang et al. [2021]. The images are generated by a latent mechanism shown in Fig. 2. The images have a size of 4 × 96 × 96. For the Flow dataset, the auxiliary variable **Ball Size** is determined through the selection process. In the case of the Pendulum dataset, all observed latents should be selected as auxiliary variables to ensure the conditional independence of the unobserved latent variables.

As illustrated in Fig. 8, our proposed method demonstrated performance that is comparable to or superior to other models. Unlike the results on synthetic data, the GIN model exhibited strong performance because its normalizing flow-based architecture is more suitable for handling image data in terms of model capacity. The MCC matrices (Figs. 9 and 10) also show that our method learns disentangled representations for unobserved latent variables. It suggests that the learned representations align well with the conditional independence structure of the underlying latent graph in Figs. 2a and 2b.

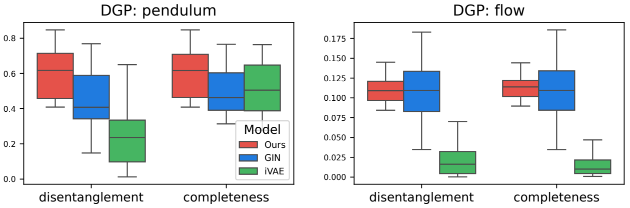


Figure 8: Comparison plot for DCI metric between Ours, GIN , and iVAE on high-dimensional data.

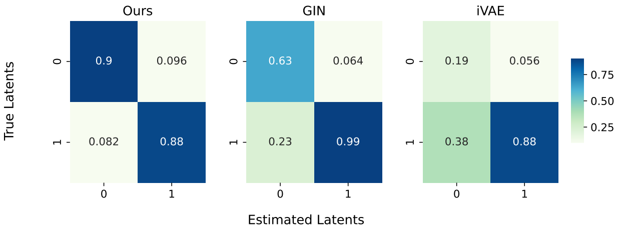


Figure 9: Mean correlation matrices of Ours, GIN applying selection and iVAE applying selection matched with the best permutation on the Pendulum dataset.

We also observed that the representations in Flow were more entangled compared to Pendulum. There exists an observed but unselected variable z_n (**Water Flow**), which introduces additional graph constraints. The graph constraints may conflict with the term enforcing conditional independence, making the learning process more challenging. Addressing this challenge remains an avenue for future work.

4.3 LATENT TRAVERSE

For better comprehensibility, we further extend our model to the image reconstruction task and perform latent traversal to assess whether the factors have been disentangled effectively. We conducted experiments on the pendulum dataset as shown in Fig. 2a, choosing the pendulum angle and light position as selected variables. To efficiently extract relevant features from high-dimensional image data and visualize disentangled factors, an extra encoder-decoder architecture with an additional MSE (Mean Squared Error) loss was adopted to ensure successful compression and reconstruction of the images.

The encoder initially compresses the image into exogenous latent variables corresponding to the number of nodes in the causal graph. This set of variables is then passed through

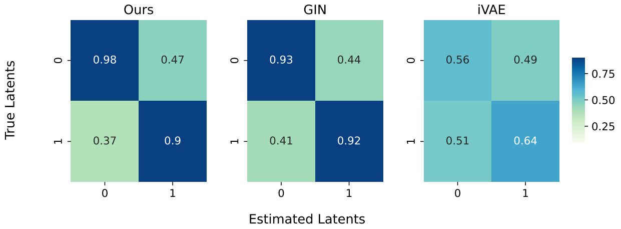


Figure 10: Mean correlation matrices of Ours, GIN applying selection and iVAE applying selection matched with the best permutation on the Flow dataset.

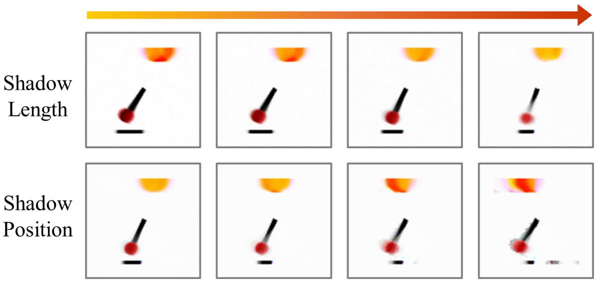


Figure 11: Latent traversal results for unobserved variables. The upper and lower rows show reconstructed images by traversing the variables for shadow length and shadow position, respectively.

our model, generating endogenous latent variables of the same dimensionality. The decoder follows a scene-mixture approach, where each scalar from the endogenous latent variables passes through multiple fully connected layers to generate a full-size image. The final output is then reconstructed by averaging these images.

Fig. 11 presents the results of generating counterfactual images by traversing unobserved latent variables after training our model with the reconstruction objective. As shown in the upper row, traversing the variable associated with shadow length gradually decreases its extent in the reconstructed images. Similarly, modifying the latent variable corresponding to shadow position causes the shadow to shift progressively to the right while mostly preserving the other factors. The successful disentanglement of unobserved latent variables further demonstrates the model's effectiveness in its transferability.

5 RELATED WORK

One of the key obstacles in CRL is the dependence among latent sources induced by underlying causal mechanisms. It directly violates the assumption of conditionally independent sources, which underlies the identifiability of many nonlinear ICA approaches that rely on conditionally factorized priors [Khemakhem et al., 2020]. To address this issue, several works explicitly incorporate a known or assumed causal graph over the latent variables to model source dependencies. For example, Yang et al. [2021] (CausalVAE) propose a structured variational autoencoder where the latent variables follow a predefined causal DAG, enabling do-interventions in the latent space. Similarly, Pan and Bareinboim [2024] (ANCM) handle non-Markovian generative processes by modeling image generation with an augmented causal graph that captures temporally entangled latent factors. While these methods provide a framework for incorporating causal structure into representation learning, they operate under a fully supervised setting, assuming access to structured semantic labels or ground-truth causal factors. Moreover, they are primarily focused on image generation and counterfactual editing tasks, rather than the general identifiability or recovery of latent sources from more weakly supervised or observational data.

To achieve identifiability under such dependencies, many methods rely on interventional data which can be impractical in real-world settings [Lippe et al., 2023, Liang et al., 2023, Li et al., 2024]. In particular, the Liang et al. [2023] (CauCa) assumes a Markovian graph and leverages interventions for identifiability, while Li et al. [2024] (CRID) handles more general non-Markovian settings by explicitly modeling unobserved confounders. Both of CauCA and CRID share with our approach the use of causal graph to guide recovery, suggesting that our method could be extended to non-Markovian settings in future work.

As an alternative, recent efforts have aimed to prove identifiability from observational data alone. For example, Yao et al. [2024b] introduce a method based on block-identifiability [Kügelgen et al., 2021], which extracts shared latent variables from multiple views using contrastive learning and entropy regularization. Zhang et al. [2024] show that assuming structural sparsity among the sources enables identifiability without any explicit causal graph. While these works relax assumptions on data collection, they rely on indirect structural constraints. In contrast, we investigate how to select or exploit observed sources as auxiliary variables under a known causal structure to recover latent sources. This approach retains the strengths of causal modeling while improving recoverability in settings where full interventions or disentangled views are unavailable.

6 CONCLUSION

CRL aims to uncover latent variables in real-world systems. Our work is the first to achieve identifiability with observed sources by incorporating auxiliary variables into the mixing function. We also introduce a framework for selecting auxiliary variables to improve recoverability by leveraging the causal structure. Empirical results show that our method outperforms others in identifying true latent variables, effectively mitigating spurious correlations arising from observable sources.

References

Dominik Baumann, Friedrich Solowjow, Karl Henrik Johansson, and Sebastian Trimpe. Identifying causal structure in dynamical systems. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL `https://openreview.net/forum?id=X2BodlyLvT`.

J.-F. Cardoso. Multidimensional independent component analysis. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 4, pages 1941–1944 vol.4, 1998. doi: 10.1109/ICASSP.1998.681443.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=By-7dz-AZ`.

Dan Geiger, Thomas Verma, and Judea Pearl. d-separation: From theorems to algorithms. In Max HENRION, Ross D. SHACHTER, Laveen N. KANAL, and John F. LEMMER, editors, *Uncertainty in Artificial Intelligence*, volume 10 of *Machine Intelligence and Pattern Recognition*, pages 139–148. North-Holland, 1990.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In R. Garnett, D.D. Lee, U. von Luxburg, I. Guyon, and M. Sugiyama, editors, *Advances in Neural Information Processing Systems*, number NIPS 2016 in Advances in neural information processing systems, pages 3772–3780, United States, 2016. Neural Information Processing Systems Foundation. Annual Conference on Neural Information Processing Systems, NIPS ; Conference date: 05-12-2016 Through 10-12-2016.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: existence and uniqueness results. *Neural Netw.*, 12(3), 1999.

Aapo Hyvärinen, Jarmo Hurri, Patrik O Hoyer, Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. *Independent component analysis*. Springer, 2009.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 16–18 Apr 2019.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 26–28 Aug 2020.

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 A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL `https://openreview.net/forum?id=4pf_pOo0Dt`.

A. Li, E. Pan, and E. Bareinboim. Disentangled representation learning in non-markovian causal systems. Technical Report R-110, Causal Artificial Intelligence Lab, Columbia University, May 2024.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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*, 2023.

Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. Best Paper Award.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=-e4EXDWXnSn`.

Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, page 440–448, Arlington, Virginia, USA, 2013. AUAI Press.

Yushu Pan and Elias Bareinboim. Counterfactual image editing. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL `https://openreview.net/forum?id=OXzkw7vFIO`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *Royal Society Open Science*, 9(8):220638, 2022.

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. doi: 10.1109/JPROC.2021.3058954.

Ross D. Shachter. Bayes-ball: Rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams). In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, page 480–487, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin), 2020. URL `https://arxiv.org/abs/2001.04872`.

Fabian Theis. Towards a general independent subspace analysis. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006. URL `https://proceedings.neurips.cc/paper_files/paper/2006/file/20479c788fb27378c2c99eadcf207e7f-Paper.pdf`.

Siyu Wang, Xiaocong Chen, and Lina Yao. On causally disentangled state representation learning for reinforcement learning based recommender systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2390–2399, 2024.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. Causal representation learning for out-of-distribution recommendation. In *Proceedings of the ACM Web Conference 2022*, pages 3562–3571, 2022.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9593–9602, 2021.

Xiaojiang Yang, Yi Wang, Jiacheng Sun, Xing Zhang, Shifeng Zhang, Zhenguo Li, and Junchi Yan. Nonlinear ICA using volume-preserving transformations. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=AMpki9kp8Cn`.

Xinxin Yang, Xinwei Li, Zhen Liu, Yannan Wang, Sibo Lu, and Feng Liu. Disentangled causal representation learning for debiasing recommendation with uniform data. *Applied Intelligence*, pages 1–16, 2024.

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

Dingling Yao, Danru Xu, Sebastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. In *The Twelfth International Conference on Learning Representations*, 2024b.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. In *Forty-first International Conference on Machine Learning*, 2024.

Yujia Zheng and Kun Zhang. Generalizing nonlinear ICA beyond structural sparsity. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Towards Causal Representation Learning with Observable Sources as Auxiliaries (Supplementary Material)

Kwonho Kim[1] **Heejeong Nam**[2] **Inwoo Hwang**[3] **Sanghack Lee**[1]

[1]Graduate School of Data Science, Seoul National University, Seoul, South Korea
[2]Boeing Korea, Seoul, South Korea
[3]Causal Artificial Intelligence Lab, Columbia University, New York, USA

A DISCUSSION

Related Work To deal with the case of Fig. 1b, we can partition the latent sources into conditionally independent sets, $\mathbf{z}_{c_i}(i = 1, ..., d)$ where $\cup_{i=1}^{d}\mathbf{z}_{c_i} = \{z_1, \ldots, z_n\}$. It enables a more general formulation of Eq. (3) as Zheng and Zhang [2023]:

$$p_{\mathbf{z}|\mathbf{u}}(\mathbf{z}|\mathbf{u}) = \prod_{i=1}^{n_i} p_{z_i}(z_i) \prod_{j=1}^{d} p_{\mathbf{z}_{c_j}|\mathbf{u}}(\mathbf{z}_{c_j}|\mathbf{u}). \tag{12}$$

where n_i is the number of mutually independent sources. Zheng and Zhang [2023] partition all the sources into a set of mutually independent sources $\mathbf{z}_I$ and a set of variables in which do not need to be independent $\mathbf{z}_{o^-} = \cup_{i=1}^{d}\mathbf{z}_{c_i}$. In **??**, we further generalize Eq. (12) into the setting with observed sources, which includes Eq. (12) as a special case in that $\mathbf{u}$ is independent from DGP.

$$p_{\mathbf{z}_{o^-}|\mathbf{z}_o}(\mathbf{z}_{o^-}|\mathbf{z}_o) = \prod_{i=1}^{n_i} p_{z_i}(z_i) \prod_{j=1}^{d} p_{\mathbf{z}_{c_j}|\mathbf{z}_o}(\mathbf{z}_{c_j}|\mathbf{z}_o). \tag{13}$$

where $\mathbf{z}_o$ is observed sources and $\mathbf{z}_{o^-}$ is unobserved sources. The former term corresponds to the case without auxiliary variables, which is beyond the scope of our study and thus not considered further.

Metrics After training the proposed method, we measure Disentanglement, Completeness, Informativeness (DCI) metric [Eastwood and Williams, 2018] to measure the degree of identifiability based on Mean Correlation Coefficient (MCC) matrix which is a widely accepted metric in the literature for measuring the degree of identifiability [Hyvärinen and Morioka, 2016]. Specifically, MCC metric is expressed as follows:

$$\text{MCC}(\mathbf{z}, \hat{\mathbf{z}}) = \frac{1}{n} \max_{\sigma \in S_n} \sum_{i=1}^{n} \text{corr}(z_i, \hat{z}_{\sigma(i)}),$$

where $\sigma \in S_n$ is a permutation of the set of indices. If the model successfully recovers the latent variables, MCC will match estimation with the most similar distributions to true latent (i.e., the highest correlation). However, the MCC metric alone is insufficient for measuring the degree of identifiability in scenarios involving partially observable sources since spurious correlation can arise without disentangling the information of $\mathbf{z}_o$ due to the information from the auxiliary variable $\mathbf{z}_o$ being entangled with the observation $\mathbf{x}$.

The Fig. 12 demonstrates the insufficiency of MCC score in evaluating the degree of identifiability. The MCC scores of the GIN and iVAE models are around 0.7, suggesting that they recover the true latents reasonably well. However, examining the correlation matrix reveals that the estimated latents also show high correlations with dimensions other than the one with the highest correlation. This is because existing methods do not account for cases where the mixing function includes auxiliary variables, leading to information from the auxiliary variables being entangled in the estimated latents.

Accordingly, we leverage the DCI metric [Eastwood and Williams, 2018] to evaluate whether the learned representation correctly models the conditional independence structure of the graph without spurious correlation. The DCI metric evaluates

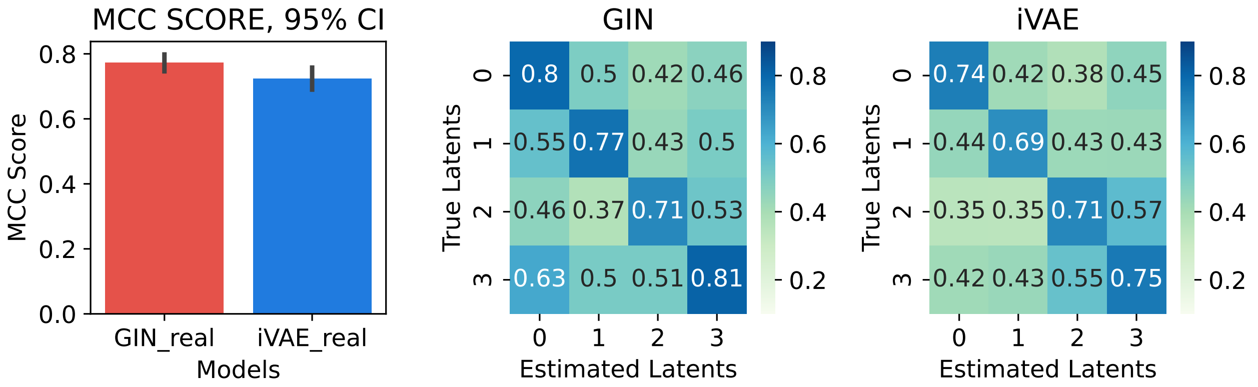


Figure 12: MCC score and mean correlation matrices of GIN and iVAE matched with the best permutation on the setting of Fig. 1b.

the performance of disentanglement, completeness, and informativeness of representation by measuring the entropy of the importance matrix (in our case, a MCC matrix) If the true sources are well identified without spurious correlation, the representation will be highly disentangled with complete information.

B THEORETICAL ANLAYSIS

Firstly, we begin with the definition of identifiability, which is the goal of nonlinear ICA and causal representation learning. By adopting a Structural Causal Model (SCM, Pearl [2009]), we represent a data-generating process regarding latent sources as

$$z_i = f_i(\text{Pa}^{\mathcal{G}}(z_i), \epsilon_i), \quad \epsilon_i \sim p_{\epsilon_i}, \tag{14}$$

for all $i \in [n]$ where $\text{Pa}^{\mathcal{G}}(\cdot)$ represents parent nodes on a latent causal graph $\mathcal{G}$ consisting of nodes V and edges E.

Definition 1. *(Identifiability). Suppose the observations* $\mathbf{x}$ *are generated by true latent mechanism specified by* $\Theta = (\mathbf{f}, p(\boldsymbol{\epsilon}), \mathbf{g})$ *given in Eqs.* (1) *and* (2)*. The learned generative model parameterized by* $\hat{\Theta} = (\hat{\mathbf{f}}, \hat{p}(\boldsymbol{\epsilon}), \hat{\mathbf{g}})$ *is observationally equivalent to the true model if the model distribution* $p_{\hat{\Theta}}(\mathbf{x})$ *matches the data distribution* $p_{\Theta}(\mathbf{x})$ *for any value of* $\mathbf{x}$*. Let* A *be an arbitrary invertible transformation. We say that the model is identifiable up to* A *if*

$$p_{\hat{\Theta}}(\mathbf{x}) = p_{\Theta}(\mathbf{x}) \implies \hat{\mathbf{g}} = \mathbf{g} \circ A. \tag{15}$$

Once the mixing function g is identified, the latent variables can be identified up to A:

$$\begin{aligned}\hat{\mathbf{z}} = \hat{\mathbf{g}}^{-1}(\mathbf{x}) &= (A^{-1} \circ \mathbf{g}^{-1})(\mathbf{x}) \\ &= A^{-1}(\mathbf{g}^{-1}(\mathbf{x})) \\ &= A^{-1}(\mathbf{z}).\end{aligned}$$

Proposition 2. *Suppose the following assumptions hold:*

1. *The observed data and sources are generated from Eq.* (1) *and Eq.* (5)
2. *The observable sources do not have direct edge into the observation* $\mathbf{x}$*, i.e.,* $\frac{\partial \mathbf{x}}{\partial \mathbf{z}_o} = 0$
3. *For every value of* $\mathbf{z}_D$*, there exists* $2d+1$ *values of* $\mathbf{z}_o$*, such that the* $2d$ *vectors* $\mathbf{w}(\mathbf{z}_D, \mathbf{z}_{o_i}) - \mathbf{w}(\mathbf{z}_D, \mathbf{z}_{o_0})$ *are linearly independent, where vector* $\mathbf{w}(Z_D, \mathbf{z}_{o_i})$ *is defined as follows:*

$$\begin{aligned}\mathbf{w}(\mathbf{z}_D, \mathbf{z}_{o_i}) = \big(&\mathbf{v}(\mathbf{z}_{c_1}, \mathbf{z}_{o_i}), \ldots, \mathbf{v}(\mathbf{z}_{c_d}, \mathbf{z}_{o_i}), \\ &\mathbf{v}'(\mathbf{z}_{c_1}, \mathbf{z}_{o_i}), \ldots, \mathbf{v}'(\mathbf{z}_{c_d}, \mathbf{z}_{o_i})\big)\end{aligned}$$

where

$$\mathbf{v}(\mathbf{z}_{c_j}, \mathbf{z}_{o_i}) = \left(\frac{\partial \log p(\mathbf{z}_{c_j} \mid \mathbf{z}_{o_i})}{\partial z_{c_j}^{(l)}}, \ldots, \frac{\partial \log p(\mathbf{z}_{c_j} \mid \mathbf{z}_{o_i})}{\partial z_{c_j}^{(h)}} \right),$$
$$\mathbf{v}'(\mathbf{z}_{c_j}, \mathbf{z}_{o_i}) = \left(\frac{\partial^2 \log p(\mathbf{z}_{c_j} \mid \mathbf{z}_{o_i})}{\partial (z_{c_j}^{(l)})^2}, \ldots, \frac{\partial^2 \log p(\mathbf{z}_{c_j} | \mathbf{z}_{o_i})}{\partial (z_{c_j}^{(h)})^2} \right)$$

and $\mathbf{z}_{c_j} = (z_{c_j^{(l)}}, \ldots, z_{c_j^{(h)}})$.

Then all components of $\mathbf{z}_D$ *(i.e.,* $\mathbf{z}_{c_i}$ *where* $c_i \in \{c_1, \ldots, c_d\}$*) is identifiable up to a subspace-wise invertible transformation and a subspace-wise permutation.*

Proof. Let $h : \mathbf{z}_{o^-} \to \hat{\mathbf{z}}_{o^-}$ denote the transformation from true sources to estimated sources. Thus, we can derive $\hat{g} = g \circ h^{-1}(\mathbf{z}_{o^-})$ equivalently as

$$\mathbf{J}_g(\mathbf{z}_{o^-}) = \mathbf{J}_{\hat{g} \circ h}(\mathbf{z}_{o^-}) = \mathbf{J}_{\hat{g}}(\hat{\mathbf{z}}_{o^-}) \mathbf{J}_h(\mathbf{z}_{o^-})$$

by using chain rule repeatedly. $\mathbf{J}_h(\mathbf{z}_{o^-})$ must be invertible and have a non-zero determinant because $\mathbf{J}_{\hat{g}}(\hat{\mathbf{z}}_{o^-})$ and $\mathbf{J}_g(\mathbf{z}_{o^-})$ have full column rank. The change of variable rule and Assumption 2 make the following equations hold:

$$p(\mathbf{z}_{o^-} \mid \mathbf{z}_o) \cdot |\det(\mathbf{J}_{h^{-1}}(\hat{\mathbf{z}}_{o^-}))| = p(\hat{\mathbf{z}}_{o^-} \mid \mathbf{z}_o).$$

By taking logarithm on both sides, we can obtain

$$\log p(\mathbf{z}_{o^-} \mid \mathbf{z}_o) + \log |\det(\mathbf{J}_{h^{-1}}(\hat{\mathbf{z}}_{o^-}))| = \log p(\hat{\mathbf{z}}_{o^-} \mid \mathbf{z}_o).$$

According to the Assumption 1[1] and $\cup_i \mathbf{z}_{c_i} = \mathbf{z} \setminus \mathbf{z}_o$, the joint log densities can be factorized as

$$\sum_{j=c_1}^{c_d} \log p(\mathbf{z}_j \mid \mathbf{z}_o) + \log |\det(\mathbf{J}_{h^{-1}}(\hat{\mathbf{z}}_{o^-}))| = \sum_{j=c_1}^{c_d} \log p(\hat{\mathbf{z}}_j \mid \hat{\mathbf{z}}_o).$$

Thus, for $\mathbf{z}_o = \mathbf{z}_{o_0}, \ldots \mathbf{z}_{o_{2d}}$, we have $2d+1$ equations. Subtracting each equation corresponding to $\mathbf{z}_{o_1}, \ldots, \mathbf{z}_{o_2 d}$ with the equation corresponding to $\mathbf{z}_{o_0}$ results in $2d$ equations:

$$\sum_{i=c_1}^{c_d} (\log p(\mathbf{z}_i \mid \mathbf{z}_{o_j}) - \log p(\mathbf{z}_i \mid \mathbf{z}_{o_0})) = \sum_{i=c_1}^{c_d} (\log p(\hat{\mathbf{z}}_i \mid \mathbf{z}_{o_j}) - \log p(\hat{\mathbf{z}}_i \mid \mathbf{z}_{o_0})) \tag{16}$$

Take the derivatives of both sides of Eq. (16) with respect to $\hat{z}_k$ and $\hat{z}_v$ where $k, v \in \{1, \ldots, n\}$ and they are not indices of the same subspace. It is clear that the RHS of Eq. (16) equals to zero because k and v are not indices of the same subspace. For the i-th term of the summation on the LHS, we can get following equations:

$$\sum_{l=i^{(l)}}^{i^{(h)}} \left(\left(\frac{\partial^2 \log p(\mathbf{z}_i \mid \mathbf{z}_{o_j})}{(\partial z_l)^2} - \frac{\partial^2 \log p(\mathbf{z}_i \mid \mathbf{z}_{o_0})}{(\partial z_l)^2} \right) \cdot \frac{\partial z_l}{\partial \hat{z}_k} \frac{\partial z_l}{\partial \hat{z}_v} \right. \tag{17}$$
$$\left. + \left(\frac{\partial \log p(\mathbf{z}_i \mid \mathbf{z}_{o_j})}{\partial z_l} - \frac{\partial \log p(\mathbf{z}_i \mid \mathbf{z}_{o_0})}{\partial z_l} \right) \cdot \frac{\partial^2 z_l}{\partial \hat{z}_k \partial \hat{z}_v} \right) = 0,$$

where i_l and i_h are the minimum and maximum indices of elements in $\mathbf{z}_i = (z_{il}, \ldots, z_{ih})$. By iterating i from c_1 to c_d, we can also iterate l from 0 to n. Thus, there exists a linear system with a $2d \times 2d$ coefficient matrix.

Considering Assumption 3, the coefficient matrix of the linear system has full rank. The only solution of Eq. (17) is $\frac{\partial z_l}{\partial \hat{z}_k} \frac{\partial z_l}{\partial \hat{z}_v} = 0$ and $\frac{\partial^2 z_l}{\partial \hat{z}_k \partial \hat{z}_v} = 0$. Note that $\frac{\partial z_l}{\partial \hat{z}_k}$ and $\frac{\partial z_l}{\partial \hat{z}_v}$ cannot be both zero because of invertibility of h. Therefore, k can only be the index of an estimated source from one independent subspace, which, together with the invertibility, leads to the conclusion that $\mathbf{z}_{o^-}$ is a composition of an invertible subspace-wise transformation and a subspace-wise permutation of $\hat{Z}_D$.

So it is the mapping from $\hat{\mathbf{z}}_{o^-}$ to $\mathbf{z}_{o^-}$ since the subspace-wise transformation is invertible and the inverse of a block-wise permutation matrix is still a block-wise invertible matrix. □

We now establish identifiability in the presence of partially observable sources, where an auxiliary variable directly influences the observation $\mathbf{x}$ through the mixing function. This constitutes the proof of Prop. 1.

Proof. Assume observational equivalence between estimated and true model, i.e. $p_g(\mathbf{x}) = p_{\hat{g}}(\mathbf{x})$. The change of varialbe rule makes following equations to hold:

$$p(\mathbf{x}) = p(\mathbf{z}) \cdot |\det(\mathbf{J}_{g^{-1}})(\mathbf{x})| = p(\hat{\mathbf{z}}) \cdot |\det(\mathbf{J}_{\hat{g}^{-1}})(\mathbf{x})|$$

Since $p(\mathbf{z}) = p(\mathbf{z}_{o^-} \mid \mathbf{z}_o) \cdot p(\mathbf{z}_o)$,

$$p(\mathbf{z}_{o^-} \mid \mathbf{z}_o) \cdot p(\mathbf{z}_o) \cdot |\det(\mathbf{J}_{g^{-1}})(\mathbf{x})| = p(\hat{\mathbf{z}}_{o^-} \mid \hat{\mathbf{z}}_o) \cdot p(\hat{\mathbf{z}}_o) \cdot |\det(\mathbf{J}_{\hat{g}^{-1}})(\mathbf{x})|$$

also can hold. Note that $p(\hat{\mathbf{z}}_o)$ can be replaced by $p(\mathbf{z}_o)$ because $\mathbf{z}_o$ is already observed.

$$p(\mathbf{z}_{o^-} \mid \mathbf{z}_o) \cdot |\det(\mathbf{J}_{g^{-1}})(\mathbf{x})| = p(\hat{\mathbf{z}}_{o^-} \mid \mathbf{z}_o) \cdot |\det(\mathbf{J}_{\hat{g}^{-1}})(\mathbf{x})|$$

By taking logarithm on both sides, we can obtain

$$\log p(\mathbf{z}_{o^-} \mid \mathbf{z}_o) + \log |\det(\mathbf{J}_{g^{-1}})(\mathbf{x})| = \log p(\hat{\mathbf{z}}_{o^-} \mid \mathbf{z}_o) + \log |\det(\mathbf{J}_{\hat{g}^{-1}})(\mathbf{x})|.$$

According to the Assumption 1, Assumption 2 and $\cup_i \mathbf{z}_{c_i} = \mathbf{z} \setminus \mathbf{z}_o$, the joint log densities can be factorized as

$$\sum_{j=c_1}^{c_d} \log p(\mathbf{z}_j \mid \mathbf{z}_o) = \sum_{j=c_1}^{c_d} \log p(\hat{\mathbf{z}}_j \mid \mathbf{z}_o).$$

Thus, for $\mathbf{z}_o = \mathbf{z}_{o_0}, \ldots \mathbf{z}_{o_{2d-1}}$, we have $2d$ equations. Take the derivatives of both sides of above equation with respect to $\hat{z}_k$ and $\hat{z}_v$ where $k, v \in \{1, \ldots, n\}$ and they are not indices of the same subspace. It is clear that the RHS of Eq. (16) equals to zero because k and v are not indices of the same subspace. For the i-th term of the summation on the LHS, we can get following equations:

$$\sum_{l=i^{(l)}}^{i^{(h)}} \left(\left(\frac{\partial^2 \log p(\mathbf{z}_i \mid \mathbf{z}_o)}{(\partial z_l)^2} \right) \cdot \frac{\partial z_l}{\partial \hat{z}_k} \frac{\partial z_l}{\partial \hat{z}_v} + \left(\frac{\partial \log p(\mathbf{z}_i \mid \mathbf{z}_o)}{\partial z_l} \right) \cdot \frac{\partial^2 z_l}{\partial \hat{z}_k \partial \hat{z}_v} \right) = 0, \tag{18}$$

where i_l and i_h are the minimum and maximum indices of elements in $\mathbf{z}_i = (z_{il}, \ldots, z_{ih})$. By iterating i from c_1 to c_d, we can also iterate l from 0 to n.

Considering Assumption 3, the coefficient matrix of the linear system has full rank. The only solution of Eq. (18) is $\frac{\partial z_l}{\partial \hat{z}_k} \frac{\partial z_l}{\partial \hat{z}_v} = 0$ and $\frac{\partial^2 z_l}{\partial \hat{z}_k \partial \hat{z}_v} = 0$. Note that $\frac{\partial z_l}{\partial \hat{z}_k}$ and $\frac{\partial z_l}{\partial \hat{z}_v}$ cannot be both zero because of invertibility of h. Therefore, k can only be the index of an estimated source from one independent subspace, which, together with the invertibility, leads to the conclusion that $\mathbf{z}_{o^-}$ is a composition of an invertible subspace-wise transformation and a subspace-wise permutation of $\hat{Z}_D$. So it is the mapping from $\hat{Z}_D$ to $\mathbf{z}_{o^-}$ since the subspace-wise transformation is invertible and the inverse of a block-wise permutation matrix is still a block-wise invertible matrix. □

C BAYES-BALL ALGORITHM

The best known criterion for conditional independence is *d-separation* [Geiger et al., 1990]. We want to find clusters with inter-cluster d-connectedness and intra-cluster d-separation.

We exploit *Bayes-ball* algorithm to examine the conditional independence of two node sets on the given graph $\mathcal{G}$. The Bayes-ball algorithm can be extended to partition graph. It returns a set of nodes dependent to an input node set.

D EXPERIMENTAL DETAILS

The implementation of the experiments is based on Liang et al. [2023]. Following tables are hyperparameters for learning Ours, GIN and iVAE.

Algorithm 2 Graph Partition by Conditional Independence

```
**Input**: graph $G = (\mathbf{V}, \mathbf{E})$, condition $C$, observed set $O$
**Output**: a set of d-connected node clusters $R$
$R \leftarrow \emptyset$
**for** each node $n$ in $\mathbf{V} \setminus O$ **do**
    **if** $\exists_{\mathbf{C} \in R}\, n \in \mathbf{C}$ **then**
        continue
    **end if**
    $result \leftarrow \{n\}$;
    **while** $result$ updated **do**
        $result \leftarrow \textsc{BayesBall}(G, C, O, result)$
    **end while**
    Add $result$ to $R$
**end for**
**return** $R$
```

Table 1: Hyperparameters for different models.

Ours	
LR scheduler	Cosine
Learning rate	0.01
Number of flows	8
Optimizer	Adam
Batch size	1024
Training epochs	20

(a) Synthetic data

GIN	
LR scheduler	-
Learning rate	0.01
Number of flows	8
Optimizer	Adam
Batch size	1024
Training epochs	20

(b) Synthetic data

iVAE	
Number of layers	3
Learning rate	0.0001
Hidden dim	4096
Optimizer	Adam
Batch size	32
Training epochs	20

(c) Synthetic data

Ours	
LR scheduler	Cosine
Learning rate	0.001
Number of flows	8
Optimizer	Adam
Batch size	1024
Training epochs	50

(d) High-dimensional data

GIN	
LR scheduler	-
Learning rate	0.001
Number of flows	8
Optimizer	Adam
Batch size	1024
Training epochs	40

(e) High-dimensional data

iVAE	
Number of layers	3
Learning rate	0.0001
Hidden dim	4096
Optimizer	Adam
Batch size	1024
Training epochs	80

(f) High-dimensional data

Algorithm 3 Bayes Ball Algorithm for d-connected nodes

```
Input: Graph $G$, Conditioning Set $C$, Observed Set $O$, Set of nodes $R$
Output: Updated set of d-connected nodes $R$
Initialize an empty set $V$ FOR visited nodes
Initialize an empty queue $Q$
for each node $n$ in $R$ do
    Add $(n, \text{up})$ to $Q$
end for
while $Q$ is not empty do
    $(node, direction) \leftarrow Q.pop()$
    if $node \in V$ then
        continue
    end if
    Add $node$ to $V$
    if $node \in C$ and $direction \neq$ down then
        continue
    end if
    if $direction =$ up then
        for each $parent$ of $node$ in $G$ do
            Add $(parent, \text{up})$ to $Q$
        end for
        for each $child$ of $node$ in $G$ do
            Add $(child, \text{down})$ to $Q$
        end for
    else if $direction =$ down then
        Initialize $check \leftarrow$ false
        for each descendant $d$ of $node$ in $G$ do
            if $d \in C$ then
                $check \leftarrow$ true
                break
            end if
        end for
        if $node \in C$ or $check =$ true then
            for each $parent$ of $node$ in $G$ do
                Add $(parent, \text{up})$ to $Q$
            end for
        else
            for each $child$ of $node$ in $G$ do
                Add $(child, \text{down})$ to $Q$
            end for
        end if
    end if
    if $node \notin C$ and $node \notin O$ then
        Add $node$ to $R$
    end if
end while
return $R$
```