On Causal Representation Learning with Internal 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]Brown University, Providence, Rhode Island, USA
[3]Causal Artificial Intelligence Lab, Columbia University, New York, USA

Abstract

Causal representation learning typically achieves identifiability using auxiliary variables external to the mixing process. This fails when observable sources act as *internal* mixing inputs whose entanglement invalidates standard volume-change proofs. We introduce a framework that treats observable sources as internal auxiliaries. We establish identifiability when all observed internal sources are used as conditioning variables. Under a volume-preserving mixing assumption, suitable variability conditions, and additional alignment conditions on the learned model, the unobserved source subspaces are identifiable up to an Independent Subspace Analysis (ISA)-style equivalence class—a permutation of conditionally independent subspaces with within-subspace invertible transformations. In practice, we approximate the volume-preserving restriction with encoders that stabilize the induced distortion. Leveraging the known causal graph, we further propose a scheme that selects conditioning sources so as to preserve fine-grained factorizations of the unobserved sources. Experiments show improved recovery over representative auxiliary-based baselines in settings where treating internal sources as external conditioners can induce cross-factor leakage.

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., 2001], aims to recover latent variables from observed data. This approach holds great promise for applications in areas such as healthcare [Sanchez et al., 2022], climate science [Yao et al., 2024a], and recommendation systems [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, causal representation learning faces challenges in the absence of appropriate supervision or inductive biases, remaining vulnerable to infinitely many spurious solutions [Hyvärinen and Pajunen, 1999, Locatello et al., 2019]. To ensure identifiability of the generative process, an established line of research introduces the assumption that the data-governing latent sources are conditionally independent given the auxiliary variables $\mathbf{u}$ [Hyvärinen and Morioka, 2016, Khemakhem et al., 2020a,b] as illustrated in Fig. 1a. Moreover, the auxiliary must be sufficiently informative to induce (subspace-)conditional independence among the sources, which can be a strong requirement in practice.

Although this assumption enables identifiability, it relies on the auxiliary variables being *external* to the mixing process—modulating the source distribution without participating in the generation mapping itself. Standard identifiability proofs, therefore, hinge on the invariance of the mixing function's volume change across different auxiliary contexts. However, this formulation faces intrinsic limitations when the auxiliary information serves as a direct input to the mixing mechanism. In such cases, the observed sources do not merely condition the unobserved sources but alter the mixing process, rendering the invariance assumptions of prior works invalid. Although recent advances [Lu et al., 2022, Zheng and Zhang, 2023] relax conditional independence requirements, they maintain the premise that auxiliaries remain external to the mixing mechanism. Our work targets the scenario where this premise is violated.

Despite these theoretical challenges, treating a subset of source variables as observable is highly motivated by real-world physical systems where the factors driving the system are inseparable from the generative process. An illustrative example is a robotic arm manipulation task: the underlying latent sources correspond to physical parameters like joint

[*]Corresponding author: sanghack@snu.ac.kr.

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3061–3082.

angles or torques, while observations consist of camera images. Here, the arm angle is not an external label but an intrinsic component of the system state (an internal source) that directly determines the visual observation via the mixing function. While leveraging internal sources facilitates the identification of unobserved sources, it necessitates addressing the theoretical hurdles of mixing entanglement.

Moreover, physical systems like robotic arms 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 conditioning choices yield finer source factorizations. For example, conditioning on the measured 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 a more general data-generating process, the conditional independence between sources can vary depending on which variables are conditioned on. Thus, selecting appropriate auxiliary variables can determine the granularity of the conditional factorization of the unobserved sources used for recovery. However, this topic, i.e., *how to exploit or select auxiliary variables using graphical information*, has remained unexplored.

To this end, we propose a framework treating observable sources as *internal* auxiliaries. A key challenge is that these internal inputs induce observation-dependent geometric distortion, preventing the standard cancellation of Jacobian determinants. To address this, our contributions are threefold: (i) we formalize observable-sources CRL and show why standard proofs fail under mixing entanglement; (ii) we prove that, when all observed internal sources are conditioned on ($O = C$), the unobserved source subspaces are identifiable under a volume-preserving (VP) mixing assumption, suitable variability conditions, and learned-model alignment assumptions, up to an ISA-style equivalence class—subspace permutation and within-subspace invertible transformations; and (iii) we develop a graph-aware selection scheme for conditioning sources that preserve fine-grained conditional factorizations, and validate it using a volume-preserving flow-based model that yields disentanglement consistent with graph-implied independencies.

2 RELATED WORK

Latent Dependencies and Causal Structures Standard nonlinear ICA frameworks assume unconditionally independent sources or rely on conditionally factorized priors [Hyvärinen et al., 2001, Khemakhem et al., 2020a]. To relax these independence assumptions, several works have incorporated explicit causal graphs. For instance, CausalVAE [Yang et al., 2021] and ANCM [Pan and Bareinboim, 2024] utilize structured variational autoencoders where latent variables encode explicit causal structure. While these methods integrate causal structures for generation and counterfactual editing, they generally operate under fully supervised settings or lack rigorous identifiability guarantees for recovering true sources from purely observational data.

Interventional Approaches A substantial body of work establishes identifiability by leveraging interventional data to break the symmetry of observational distributions. Methods like CauCa [Liang et al., 2023] and CRID [Li et al., 2024] utilize perfect interventions or model unknown confounders to identify causal variables in Markovian and non-Markovian settings, respectively. Although these approaches share our use of causal graphs to guide recovery, they depend on interventions, which distinguishes them from our framework that targets the observational regime.

Identifiability via Auxiliaries and Constraints Closest to our work are methods seeking observational identifiability via auxiliary variables or structural constraints, such as external auxiliary modulation [Khemakhem et al., 2020a,b], multi-view consistency [Yao et al., 2024b], or no-auxiliary identifiability constraints [Kivva et al., 2022, Willetts and Paige, 2021]. Another line uses two distinct sparsity notions: *mixing-Jacobian sparsity*, which constrains the support of the mixing function's Jacobian [Zheng et al., 2022], and *latent-graph sparsity*, which instead constrains the Markov network recovered over the latent sources [Zhang et al., 2024]. Zheng and Zhang [2023] combine auxiliary-variable modulation with beyond-sparsity conditions. However, a key limitation of these frameworks—and general auxiliary-based identifiability theories [Zheng and Zhang, 2023]—is the premise that auxiliary variables remain *external* to the mixing process. This separation lets Jacobian determinant terms cancel across auxiliary values or environments in common identifiability arguments. In contrast, our work addresses the entangled regime where observable sources act as *internal* inputs to the mixing function. In this setting, the standard cancellation technique fails because the geometric distortion induced by the mixing becomes dependent on the observed values. Consequently, we adopt VP mixing as a theoretically grounded *sufficient* condition to control observation-dependent distortion and obtain identifiability in the internal-auxiliary regime. Unlike prior works that treat auxiliaries as fixed given variables, we actively leverage the known causal graph to *select* internal conditioning sources that preserve fine-grained conditional factorizations.

3 PRELIMINARIES AND PROBLEM FORMULATION

To frame our problem setting, we use lowercase letters for scalar source variables (e.g., s_i and z_i) and bold lowercase letters for vector-valued quantities (e.g., $\mathbf{s}$, $\mathbf{o}$, $\mathbf{z}$, and $\mathbf{x}$). Uppercase letters such as O and C denote index sets or graph node sets. We write $[d]$ to denote the set $\{1, 2, \cdots, d\}$.

Table 1: Notation for internal and external auxiliary settings.

Symbol	Meaning
$\mathbf{s}$	full pre-mixing source vector
$\mathbf{x}$	raw post-mixing observation, $\mathbf{x} = g(\mathbf{s})$
O	indices of observed internal sources
$C \subseteq O$	selected conditioning indices
$O^- = [n] \setminus O$	indices of unobserved target sources
$C^- = [n] \setminus C$	indices not selected for conditioning
$\mathbf{o} = \mathbf{s}_O$	observed internal source vector
$\mathbf{o}_C$	selected observed source vector used for conditioning
$\mathbf{o}_n = \mathbf{s}_{O \setminus C}$	observed but unselected source vector
$\mathbf{z} = \mathbf{s}_{O^-}$	unobserved target source vector
$\mathbf{u}$	external auxiliary that does not enter g

Data Generating Process We define $\mathbf{s} \in \mathbb{R}^n$ as the full vector of pre-mixing sources and $\mathbf{x} \in \mathbb{R}^m$ as an observation (e.g., image). Let $O \subseteq [n]$ denote the indices of observed internal sources and $O^- = [n] \setminus O$ its complement; we write $\mathbf{o} = \mathbf{s}_O$ for the observed source vector and $\mathbf{z} = \mathbf{s}_{O^-}$ for the unobserved source vector. The observation is generated as

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

For the theoretical identifiability results, we assume $m = n$ so that g is an invertible, twice-differentiable mixing function. In the high-dimensional image experiments, the main disentanglement runs apply the VP flow at the observation dimension, while the learned compression stage is used only for the reconstruction/traversal experiments in Sec. 5.2.

Definition 1 (Internal auxiliary)**.** *An internal auxiliary is an observed pre-mixing source coordinate, i.e., a component of* $\mathbf{s}$ *indexed by* O*, that is available at inference time and participates directly in the mixing map* g *together with the unobserved sources. It is therefore a causal handle inside the latent system, not the raw observation* $\mathbf{x}$ *and not an external auxiliary* $\mathbf{u}$*. Representation learning remains necessary because the target coordinates* $\mathbf{z} = \mathbf{s}_{O^-}$ *must still be recovered from* $\mathbf{x} = g(\mathbf{o}, \mathbf{z})$*.*

Using a Bayesian network, we represent the data-generating process (DGP) over latent sources as

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

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. When we read conditional independences from $\mathcal{G}$ using d-separation, we assume the source distribution is Markov and faithful with respect to this latent Bayesian network.

3.1 LEVERAGING CONDITIONAL INDEPENDENCE

The goal of CRL is to recover the inverse mapping g^{-1} and the true source vector $\mathbf{s} = (s_1, \cdots, s_n)$ from observations. The simplest setup might consider fully independent sources as in nonlinear ICA with the source factorization of Eq. 3.

$$p(\mathbf{s}) = \textstyle\prod_{i=1}^n p(s_i). \tag{3}$$

However, with only i.i.d. samples, this problem is provably unidentifiable: one cannot uniquely map observations to mutually independent sources under that assumption alone [Hyvärinen and Pajunen, 1999]. To ensure identifiability, many researchers have relied on the conditional independence assumption given observable variables.

Fig. 1a depicts the setup of CRL with auxiliary variables: the source vector $\mathbf{s} = (s_1, s_2, s_3, s_4)$ is mutually independent when conditioning on an observable auxiliary variable u,

$$p(\mathbf{s} \mid \mathbf{u}) = \textstyle\prod_{i=1}^n p(s_i \mid \mathbf{u}),$$

where $\mathbf{u}$ can be a class label, time index, or historical information [Hyvarinen et al., 2019].

Fig. 1b similarly assumes that the observable variable is given as an auxiliary input outside the data-generating process. This corresponds to Independent Subspace Analysis (ISA), which recovers sources only up to conditionally independent subspaces. This can be written as

$$p(\mathbf{s} \mid \mathbf{u}) = \textstyle\prod_{i=1}^d p(\mathbf{s}_{c_i} \mid \mathbf{u}), \tag{4}$$

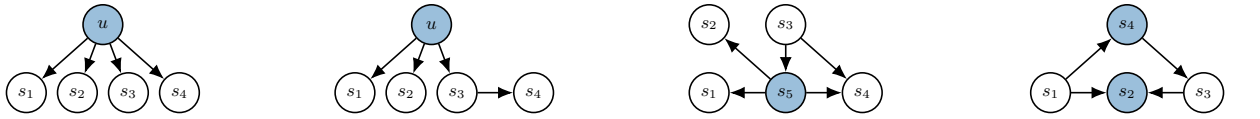
where $\{\mathbf{s}_{c_i}\}_{i=1}^d$ is a partition of the latent sources such that $\cup_{i=1}^d \mathbf{s}_{c_i} = \{s_1, \ldots, s_n\}$, and each $\mathbf{s}_{c_i}$ denotes a conditionally dependent subspace.

Figs. 1c and 1d depict our setting. Unlike Figs. 1a and 1b, the observable variables directly participate in the data-generating process. As in ISA, we do not require full conditional independence among latent sources. Let $\mathbf{o} = \mathbf{s}_O$ denote the subvector of internal sources that are directly observed. Concretely, the observed sources $\mathbf{o}$ directly enter the mixing; to obtain the selected conditional factorization used in Prop. 1, we condition on a *selected* sub-vector $\mathbf{o}_C$ of $\mathbf{o}$ (index set $C \subseteq O$; the selection procedure is given in Sec. 4.2). The generative process is then captured by DAG $\mathcal{G}$ over all sources, which encodes their arbitrary dependencies.

Accordingly, Eq. 4 is reformulated in our setting as:

$$p(\mathbf{z} \mid \mathbf{o}_C) = \textstyle\prod_{j=1}^d p(\mathbf{z}_{c_j} \mid \mathbf{o}_C), \tag{5}$$

where $\mathbf{z} \equiv \mathbf{s}_{O^-}$ denotes the subvector of unobserved source coordinates, i.e., the coordinates of $\mathbf{s}$ not indexed by O. We also write $\mathbf{o}_n = \mathbf{s}_{O \setminus C}$ for the observed but unselected sources. Note that we exclude the degenerate case $O = \emptyset$ (i.e., no auxiliary variables), since without any conditioning the problem falls into a different regime that demands extra assumptions for identifiability, such as structural sparsity [Zheng and Zhang, 2023]. We defer the discussion of how to select among multiple observed sources to Sec. 4.2.



(a) Single external auxiliary for ICA

(b) ISA subspaces with external auxiliary

(c) *Ours.* Single internal observed source

(d) *Ours.* Multiple internal observed sources

Figure 1: Latent-only graphs in different setups. (a) u is a single *external* auxiliary for ICA; (b) latent sources form ISA-style independent subspaces; (c,d) blue nodes are *internal* observed sources entering the mixing $\mathbf{x} = g(\mathbf{s})$.

3.2 PROBLEM FORMULATION

Suppose a DGP follows Eq. 1 and Eq. 2. Our goal is to recover conditionally independent unobserved source subspaces (i.e., $\mathbf{z}_{c_i}$) up to subspace-wise invertible transformations and permutations, given observations $\mathbf{x}$, observable internal sources $\mathbf{o} = \mathbf{s}_O$, and the latent Bayesian network $\mathcal{G}$ over the full source vector $\mathbf{s}$, as shown in Fig. 1. We assume that $\mathcal{G}$ is known, allowing us to leverage diverse conditional independence relationships between the sources. Importantly, the partition $\{\mathbf{z}_{c_i}\}_i$ of the unobserved source vector $\mathbf{z}$ determines how finely the unobserved source subspaces can be recovered, following the usual ISA-style equivalence class for grouped sources [Theis, 2006, Zheng and Zhang, 2023]. Therefore, it is crucial to select the appropriate observable source vector $\mathbf{o}_C$ (with $C \subseteq O$) that yields fine-grained subspaces $\mathbf{z}_{c_j}$ that are mutually independent given $\mathbf{o}_C$. The identifiability result in Sec. 4.1 is stated for the case in which the observed internal sources entering the proof are exactly the selected conditioning sources ($O = C$). When observed-but-unselected sources remain, one conservative formal route is to include them among the unconditioned source coordinates, defining $\mathbf{y} = (\mathbf{z}, \mathbf{o}_n)$ and seeking block identifiability under additional factorization and variability assumptions on $p(\mathbf{y} \mid \mathbf{o}_C)$. However, the resulting blocks may be coarser than the desired fine-grained partition of $\mathbf{z}$, especially when $\mathbf{o}_n$ couples multiple unobserved source blocks. In the present framework, we therefore handle such sources through the graph-structured objective.

Motivating Examples In simple cases such as Fig. 1c, the finest conditionally independent groups are easy to read off, i.e., $\{s_1\}$, $\{s_2\}$, $\{s_3, s_4\}$. In Fig. 1d, however, conditioning on all observed sources prevents disentangling s_1 and s_3; s_2 should be left unconditioned to obtain the finer partition. As the number of latent or observed sources grows, finding conditioning variables for fine-grained grouping becomes computationally demanding, because a single node can act as a confounder for one group and a collider for another.

4 PROPOSED FRAMEWORK

We first state an identifiability result for the selected-source case covered by the proof, where the observed internal sources entering the argument are exactly the conditioning sources ($O = C$; Sec. 4.1). Based on this conditional factorization, we introduce a framework with a graphical criterion to leverage auxiliary variables that yield a finer-grained partition of conditionally independent source groups (Sec. 4.2) and a method to recover unobserved sources (Sec. 4.3).

4.1 IDENTIFIABILITY FOR INTERNAL SOURCES

To deal with the problem that the observed sources $\mathbf{o}$ are included in the mixing function, we assume that the mixing function is constrained to a volume-preserving form, following Yang et al. [2022]. We build on the auxiliary-variable identifiability strategy for dependent source groups [Hyvarinen et al., 2019, Zheng and Zhang, 2023], using a full block-Hessian variability condition for multidimensional source subspaces.

Proposition 1 (Identifiability with internal observed sources)**.** *Let $q = \sum_{j=1}^{d} \dim(\mathbf{z}_{c_j}) = \dim(\mathbf{z})$ denote the dimension of the partitioned unobserved coordinates. For each block, let $m_j = \dim(\mathbf{z}_{c_j})$, $\phi_j(\mathbf{z}_{c_j}, \mathbf{o}_C) = \log p(\mathbf{z}_{c_j} \mid \mathbf{o}_C)$, and define $M = q + \sum_{j=1}^{d} m_j(m_j+1)/2$. Suppose the following assumptions hold:*

1. *We consider the selected-source case $O = C$, so $\mathbf{o}_n = \emptyset$ and the source vector in the change-of-variables argument is $\mathbf{s} = (\mathbf{o}_C, \mathbf{z})$. The learned model is restricted to the observed-coordinate preserving class: for all relevant $(\mathbf{o}_C, \mathbf{z})$, $\hat{g}^{-1}(g(\mathbf{o}_C, \mathbf{z})) = (\mathbf{o}_C, \hat{\mathbf{z}})$, and the marginal law of $\mathbf{o}_C$ is shared by the true and learned models. The observed data and sources are generated from Eq. 1, and both true and learned target-source distributions satisfy the conditional factorization in Eq. 5, i.e., given the selected conditioning set $\mathbf{o}_C$.*
2. *The true and learned mixing functions are volume-preserving over these source coordinates, i.e., $|\det(\mathbf{J}_g(\mathbf{s}))| = |\det(\mathbf{J}_{\hat{g}}(\hat{\mathbf{s}}))| = 1$, where $\mathbf{s} = (\mathbf{o}_C, \mathbf{z})$ and $\hat{\mathbf{s}} = (\mathbf{o}_C, \hat{\mathbf{z}})$.*
3. *For every value of $\mathbf{z}$, there exist M values of $\mathbf{o}_C$ such that the M full score-curvature vectors $\mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{o}_C^{(i)})$ are linearly independent, where*

$$\mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{o}_C^{(i)}) = \big(\{\nabla_{\mathbf{z}_{c_j}} \phi_j(\mathbf{z}_{c_j}, \mathbf{o}_C^{(i)})\}_{j=1}^{d}, \{\text{vech}(\nabla^2_{\mathbf{z}_{c_j}} \phi_j(\mathbf{z}_{c_j}, \mathbf{o}_C^{(i)}))\}_{j=1}^{d}\big).$$

Then, under the standing assumptions in App. B—in particular the conditioning-independent reparameterization in App. Assumption 3 and the block-matching condition in App. Assumption 6—the differentiated likelihood equations imply block-support exclusion: two learned blocks cannot both affect the same true block, and all the unobserved source subspaces $\{\mathbf{z}_{c_i}\}_{i=1}^{d}$ are identifiable up to subspace-wise invertible transformations and permutations.

The technical regularity conditions, together with two substantive standing assumptions—a conditioning-independent source-coordinate reparameterization ($\mathbf{z} = r(\hat{\mathbf{z}})$) and block-matching between the true and learned partitions—used to complete the subspace-wise proof are stated in App. B. The full block-Hessian condition above extends the score-curvature variability assumptions used in auxiliary-variable nonlinear ICA [Hyvarinen et al., 2019] to multidimensional source subspaces, as in nonlinear ISA [Setlur et al., 2020]. It includes within-block mixed second derivatives that are absent in singleton-block settings.

Prop. 1 should therefore be read as an identifiability result for the case in which the observed internal sources entering the proof are exactly the selected conditioning sources. When $\mathbf{o}_n = \mathbf{s}_{O \setminus C} \neq \emptyset$, applying the same proof to $\mathbf{z}$ alone would omit the density contribution of the observed-but-unselected sources, e.g., terms of the form $p(\mathbf{o}_n \mid \mathbf{z}, \mathbf{o}_C)$. In that case, the present framework handles such sources through the graph-structured objective (Sec. 4.3); incorporating them directly into the theorem would require an enlarged source vector and additional assumptions. This enlarged-source route can create multidimensional blocks, which is the setting addressed by the full block-Hessian variability condition in Prop. 1, Assumption 3. The conclusion is also stated within the model class fixed by App. Assumptions 3 and 6. The block-support exclusion step alone shows only that each true block is affected by at most one learned block, and these standing assumptions complete the final block alignment.

Most prior works on CRL achieve identifiability by assuming the mixing function is fixed across environments. Under a common invertible mixing g, auxiliary-variable proofs compare log-likelihoods across domains: the latent density terms change, but the Jacobian log-determinant is invariant and cancels when subtracting them.

In our setting, by contrast, observed sources (internal auxiliaries) enter the mixing mechanism: the mixing $g(\mathbf{o}, \mathbf{z})$ is a single, globally fixed map, but the local geometry of the slice $g(\mathbf{o}, \cdot)$ changes with the observed values $\mathbf{o}$. As a result, the usual cancellation does not occur, and the standard identifiability proof breaks down. To address the problem, we restrict both the true and learned mixing maps to the volume-preserving class (i.e., $\log|\det(\mathbf{J}_g(\mathbf{s}))| = 0$ and similarly for $\hat{g}$ over $\hat{\mathbf{s}}$). With this assumption, the Jacobian log-determinant terms remain zero (App. B).

Remark 1. *(Connection between auxiliary variables and observable sources). Existing approaches with external auxiliary variables (e.g., Fig. 1a) can be viewed as the special case in which the auxiliary does not enter the mixing, i.e., it has no edge into the observation* $\mathbf{x}$*. Prop. 2 gives the corresponding external-auxiliary statement in App. B. The cancellation mechanism differs across the two settings. For an external auxiliary, the log-determinant term is removed by differencing across auxiliary values* $\mathbf{u}$ *in Prop. 2. For internal sources, it is set to zero by the volume-preserving assumption. Thus Prop. 2 should be read as a parallel result rather than a literal corollary.*

4.2 GRAPH-GUIDED VARIABLE SELECTION

Prop. 1, stated for $O = C$, shows that the selected conditional factorization determines the granularity at which the unobserved sources are partitioned for recovery. Motivated by this, our goal is first to identify mutually independent groups of nodes given observable sources and the known causal graph. However, naively conditioning on all observed sources can miss conditional independences, i.e., $s_1 \not\perp\!\!\!\perp s_3 \mid s_2, s_4$ in Fig. 1d. We therefore seek an observed-source subset that yields the finest mutually independent groups of unobserved sources.

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

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

where $C \subseteq O$ indexes the selected observed sources, $\{\mathbf{z}_{c_i}\}_i$ partitions $\mathbf{z}$, and $\mathbf{z}_{c_i} \cap \mathbf{z}_{c_j} = \emptyset$ for all $i \neq j$. Importantly, satisfying a fine-grained conditional independence condition provides a conditional factorization with a greater number of independent source subspaces (i.e., finer groups). This gives the model a more precise structure-aware grouping of the unobserved sources, improving recoverability in practice.

We propose a strategy that selects a minimum-size conditioning set yielding the largest number of conditionally independent source groups in Alg. 1. To avoid relying on a global collider classification, the algorithm enumerates all nonempty subsets $T \subseteq O$ and lets d-separation determine how collider opening and blocking affect each candidate. The *Partition* algorithm counts conditionally independent groups using the *Bayes-ball* [Shachter, 1998] algorithm repeatedly: it forms connected components of the pairwise d-connection relation among the unobserved nodes $O^- = [n] \setminus O$ given T, iterating Bayes-ball until no component grows. Finally, the algorithm outputs the minimum-size conditioning set that results in the largest number of groups, i.e., the most fine-grained partitioning.

Example In Fig. 1d, the observed set is $O = \{s_2, s_4\}$. The node s_2 has no children and acts as a collider on the path

Algorithm 1: Selecting Observables for Conditioning

Input: graph G, observed set O
Output: conditioning set C
$C \leftarrow \emptyset$; $max \leftarrow -1$
for each nonempty subset $T \subseteq O$ **do**
  $S \leftarrow Partition(G, T, O)$ ▷ find independent groups
  if $|S| > max$ **or** ($|S| = max$ **and** $|T| < |C|$) **then**
    $max \leftarrow |S|$; $C \leftarrow T$
return C

$s_1 \to s_2 \leftarrow s_3$, so conditioning on it opens that path rather than separating the unobserved sources. Rather than pruning such nodes a priori, Alg. 1 evaluates all nonempty subsets of O, i.e., $\{s_2\}, \{s_4\}, \{s_2, s_4\}$.[1] First, with the conditioning set $\{s_2\}$, the partition process is as follows: Starting from s_1, the *result* contains s_1, and the *Bayes-ball* algorithm takes G, $\{s_2\}$, and this *result* as input. The path from s_1 to s_3 through s_2 cannot be d-separated because conditioning on the collider s_2 opens that path, and the remaining path from s_1 to s_3 through s_4 is also open because s_4 is not conditioned on. Hence the *result* is $\{\{s_1, s_3\}\}$ excluding the observed source s_2. With the conditioning set $\{s_4\}$, by following the same process, the *result* will be $\{\{s_1\}, \{s_3\}\}$. The conditioning set $\{s_2, s_4\}$ causes the result to be $\{\{s_1, s_3\}\}$. Hence, the selection result will be $\{s_4\}$ for the most fine-grained conditionally independent source groups.

4.3 STRUCTURE-AWARE FLOW ARCHITECTURE

To construct a representation aligned with Prop. 1, which covers the selected-source case, and the broader graph-aware recovery objective, we enforce volume preservation by adopting the 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 encourage dependencies among source variables that are not assumed to be independent, reflecting the known latent causal structure to strengthen disentanglement.

Volume-preservation While GIN optimizes the log-likelihood of the conditional distribution given auxiliary variables, we use the following conditionally factorized objective term for the selected conditioning sources and target subspaces: $\mathcal{L}_{\text{factor}} = \log p_{\text{align}}(\hat{\mathbf{o}}_C \mid \mathbf{o}_C) + \sum_{j=1}^{d} \log p(\hat{\mathbf{z}}_{c_j} \mid \mathbf{o}_C)$, where $\{\mathbf{z}_{c_j}\}_{j=1}^{d}$ partitions the target coordinates in $\mathbf{z} = \mathbf{s}_{O^-}$; observed but unselected sources $\mathbf{o}_n$ are outside this selected factorization and are handled by the graphical constraint below. When $C \subset O$, Eq. 5 is a marginal factorization target for the unobserved coordinates only; it is not by itself the likelihood factorization required by Prop. 1 because $\mathbf{o}_n$ still enters the mixing. This factorized objective term addresses the issue where information from the selected auxiliary variables is directly entangled with the observations. The first term is a supervised Gaussian alignment likelihood that matches the learned selected coordinates $\hat{\mathbf{o}}_C$ to the observed selected sources $\mathbf{o}_C$, while the second term encourages the target subspaces in $\mathbf{z}$ to be conditionally independent given $\mathbf{o}_C$ by modeling them with a conditionally factorized diagonal Gaussian structure.

Graphical constraint Furthermore, the remaining representation $\hat{\mathbf{s}}_{C^-}$, where $C^- = [n] \setminus C$, contains coordinates used to model the unobserved sources $\mathbf{z}$ and to predict the observed but unselected sources $\mathbf{o}_n$. We encourage the representation to retain predictive dependencies between $\mathbf{o}_n$ and $\mathbf{z}$ (the relationship between s_2 and $\{s_1, s_3\}$ in Fig. 1d). To model this dependency, we design a structural neural network, based on the latent graph $\mathcal{G}$, that encourages the non-selected learned representation $\hat{\mathbf{s}}_{C^-}$ to remain predictive of the observed-but-unselected source $\mathbf{o}_n$. Specifically, for each observed-but-unselected source, the predictor reads the representation coordinates at its graph-predecessor indices together with its own coordinate slot, following a fixed graph-indexed slot convention. For example, in Fig. 1d, s_2 has two parents, so the predictor for s_2 receives two predecessor-index coordinates from $\hat{\mathbf{s}}_{C^-}$ together with the s_2 target-coordinate slot; its input dimension is the parent count plus one. The full objective function is:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathcal{D}}[\log p_{\text{align}}(\hat{\mathbf{o}}_C \mid \mathbf{o}_C) + \textstyle\sum_j \log p(\hat{\mathbf{z}}_{c_j} \mid \mathbf{o}_C) + \sum_{i \in O \setminus C} \log p\big(o_i \mid \hat{f}_i(\hat{\mathbf{s}}_i^{\text{slot}})\big)]. \tag{7}$$

Here, $\hat{\mathbf{s}}_i^{\text{slot}}$ denotes the fixed graph-predecessor coordinate slots together with the target-coordinate slot. The graphical term is therefore a parent-count-plus-one structural prediction regularizer under this fixed slot convention.[2]

Implementation note The volume-preserving encoder restricts architectural choices, as invertible flow backbones such as GIN require matched input–output dimensionality between $\mathbf{x}$ and the learned representation $\hat{\mathbf{s}}$. In our theoretical analysis, this restriction serves as a sufficient condition for controlling observation-dependent distortion in the internal-auxiliary regime and establishing the identifiability result in Prop. 1. Empirically, prior work suggests that flow-based models can still be applied to high-dimensional image data under such constraints by effectively allocating redundancy to low-variance directions [Yang et al., 2022], making GIN a practical backbone for our framework. When image experiments use a learned encoder or compression stage, the VP restriction applies to the latent-level flow after that stage, so these experiments should be read as empiri-

[1] $\emptyset$ is excluded because the auxiliary-based factorization considered here requires a nonempty conditioning signal.

[2] The released code applies the diagonal Gaussian term to all non-selected coordinates and excludes observed-but-unselected coordinates only at evaluation.

cal approximations rather than direct instances of Prop. 1, which assumes a square invertible mixing map ($m = n$).

5 EMPIRICAL EVALUATION

We conducted experiments to validate both the effectiveness of the selection procedure and the ability of our proposed architecture to leverage observable sources.

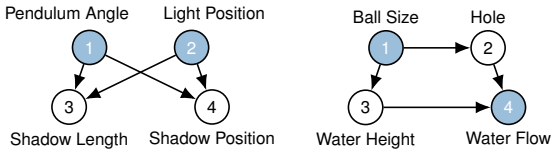
Data We use synthetic datasets from Figs. 1a, 1c, and 1d, with Fig. 1a serving as an external-auxiliary control and the latter two as internal-source settings: $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{o}^{(i)})\}_{i=1}^{N}$, where N is the sample size and $\mathbf{o}^{(i)}$ denotes the observed internal sources corresponding to the data point $\mathbf{x}^{(i)}$ (with $\mathbf{u}$ playing the analogous external-auxiliary role for Fig. 1a). When we run the graph-guided selection procedure, $\mathbf{o}$ is partitioned into the selected conditioning sources $\mathbf{o}_C$ and the observed-but-unselected sources $\mathbf{o}_n$.

The synthetic source vectors were generated using a linear Structural Causal Model (SCM): $s_i = \sum_{j \in \mathrm{pa}(s_i)} \beta_{ij} s_j + \varepsilon_i$. Here, β_{ij} are sampled uniformly from $[0.5, 1.0]$, and the noise coefficient is fixed at 1.0. The synthetic observation $\mathbf{x}$ is generated by a nonlinear invertible coupling-flow mixing map g with eight FrEIA coupling blocks. The VP setting uses GIN coupling blocks, each followed by a random permutation, with a shared fully connected subnet architecture and clamp value 2.0; the controlled non-VP ablation in App. E replaces only the GIN coupling block with a Glow-style affine coupling block while keeping the depth, subnet, permutations, optimizer, and SCM fixed. For high-dimensional experiments, we use the Pendulum and modified Flow datasets [Yang et al., 2021], which consist of structured, systematically sampled image data (Fig. 2). We provide implementation details in App. D.

Metrics We report DCI-style (Disentanglement, Completeness, and Informativeness) scores [Eastwood and Williams, 2018] computed from the Mean Correlation Coefficient (MCC) matrix rather than predictor-based feature importances. Although MCC is standard in nonlinear ICA and CRL, it can be misleading when the target equivalence class is broader than element-wise permutation and scaling [Hyvärinen and Morioka, 2016, Joshi et al., 2026]; we therefore also inspect graph-implied independence patterns.

Specifically, for an evaluation target of dimension q, the MCC matrix is defined as $\mathbf{MCC}_{ij} = |\mathrm{corr}(z_i, \hat{z}_j)|$, where each entry $\mathbf{MCC}_{ij}$ represents the absolute Pearson correlation coefficient between the true unobserved source z_i and the estimated $\hat{z}_j$. The optimal permutation σ^* is selected to maximize the total absolute correlation, matching each estimated target coordinate $\hat{z}_j$ to the most similar true unobserved source coordinate z_i.

Based on the computed MCC matrix, we evaluate models with Disentanglement $D = \frac{1}{q}\sum_{j=1}^{q}(1 - H(P^D_{\cdot|j}))$ and Completeness $C_{\mathrm{comp}} = \frac{1}{q}\sum_{i=1}^{q}(1 - H(P^C_{\cdot|i}))$. Here, $P^D_{i|j} = \mathbf{MCC}_{ij} / \sum_{i'} \mathbf{MCC}_{i'j}$ is the column-normalized contribution of true unobserved source z_i to estimated target coordinate $\hat{z}_j$, and $P^C_{j|i} = \mathbf{MCC}_{ij} / \sum_{j'} \mathbf{MCC}_{ij'}$ is the row-normalized contribution of estimated target coordinate $\hat{z}_j$ to true unobserved source z_i. The entropy function $H(\cdot)$, normalized by the logarithm of the dimension, 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 target coordinate captures at most one true unobserved source, via column-wise entropy over $P^D_{\cdot|j}$. Completeness (C_{comp}) assesses whether each true unobserved source is captured by a single estimated target coordinate, via row-wise entropy over $P^C_{\cdot|i}$. Since both scores range from 0 to 1, higher values indicate better-structured representations with minimal mixing between factors. Further discussion of why MCC alone is insufficient for evaluation is provided in App. A. The main experiments are evaluated over 20 repetitions; the additional ablations in App. E use 25 seeds as stated there.



(a) Pendulum System (b) Flow System

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

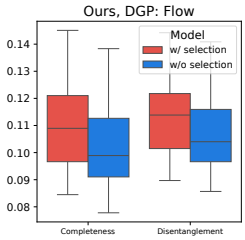


Figure 3: Ablation study on the selection procedure.

5.1 SOURCE RECOVERABILITY

Effectiveness of selection We conducted an ablation study on the selection procedure for our architecture using the DGP illustrated in Fig. 2b. Fig. 3 compares MCC-based DCI-style scores before and after selection. The selection procedure improves disentanglement in the representation. Under the Markov and faithfulness assumptions, conditioning on all observed sources in Fig. 2b opens the relevant collider path between **Water Height** and **Hole**, which is associated with more entanglement in finite-sample runs.

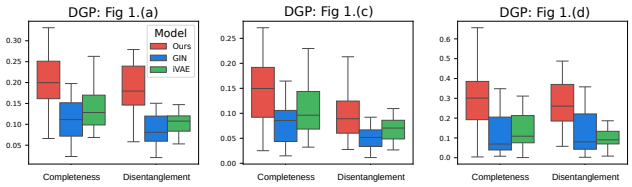


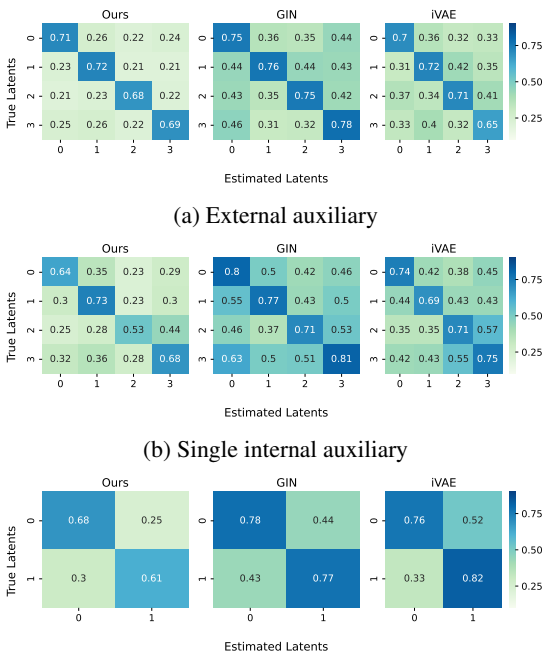
Figure 4: MCC-based DCI-style disentanglement/completeness scores for our method, GIN, and iVAE.

Effectiveness of architecture To evaluate the effectiveness of our architecture, we selected GIN [Sorrenson et al., 2020] and iVAE [Khemakhem et al., 2020a] as representative baselines. We explicitly exclude recent methods addressing non-invertible or external-auxiliary settings (e.g., Lu et al. [2022], Chen et al. [2024]) due to the fundamental structural differences from our internal-auxiliary framework. While we employ GIN as the backbone encoder to ensure the volume-preserving property, the standard GIN baseline can receive selected variables as external conditioners but does not model them as internal inputs to the mixing mechanism. Similarly, iVAE is not designed to handle partially observed sources, as it treats auxiliaries as external to the mixing. Furthermore, iVAE does not impose any constraints on the mixing function and relies on a factorized Gaussian prior over latent coordinates. The comparisons are intended to isolate the benefit of modeling internal auxiliaries and using the known graph; they are not information-identical comparisons, because the graph-structured prediction term is available only to our method.

Fig. 4 shows that our proposed method outperforms the baselines under the MCC-based DCI-style metric. Our method maximizes a conditionally factorized likelihood for the target subspaces $\hat{\mathbf{z}}_{c_j}$ while encouraging the selected-source information to appear in the aligned coordinates. This discourages spurious correlations in the representation by reducing leakage from $\mathbf{o}_C$ into $\hat{\mathbf{z}}$.

We further inspect the MCC matrices. The proposed architecture shows a comparable MCC score (mean of diagonal terms) to GIN and iVAE (Fig. 5a for the DGP of Fig. 1a). However, even after best-permutation matching, GIN and iVAE retain high off-target correlations. This suggests that changing one learned dimension is associated with changes in other target sources, indicating poor disentanglement.

Next, for the DGP in Fig. 1c, the target is subspace-wise recovery: $\{s_1\}$, $\{s_2\}$, and the dependent block $\{s_3, s_4\}$ should be separated. Our architecture's MCC matrix (Fig. 5b) shows a cleaner block structure, while the other methods still show cross-block entanglement. Because this conditional independence structure targets block-level recovery rather than variable-wise recovery, a lower diagonal MCC is not necessarily problematic. Even in this case, GIN and iVAE, which do not model observed sources as internal auxiliaries, may show high MCC scores due to spurious correlation. For



(a) External auxiliary

(b) Single internal auxiliary

(c) Multiple internal auxiliaries

Figure 5: Mean correlation matrices for our model, GIN with selection, and iVAE with selection, matched with the best permutation for the settings of Figs. 1a, 1c, and 1d.

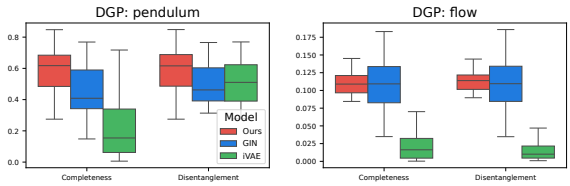


Figure 6: Comparison of MCC-based DCI-style disentanglement and completeness scores among our method, GIN, and iVAE on high-dimensional data.

the DGP in Fig. 1d, our architecture also yields the expected disentangled pattern (Fig. 5c). This case lies outside the direct scope of Prop. 1 because $\mathbf{o}_n = \{s_2\} \neq \emptyset$. The recovery is driven by the graph-structured objective rather than by the theorem alone, consistent with the scope remark in Sec. 3.2.

High-Dimensional Data We also conducted experiments on the Pendulum and Flow datasets [Yang et al., 2021], whose latent mechanisms (Fig. 2) generate $4\times96\times96$ images. For Flow, the selection process chooses **Ball Size** as the conditioning source, while **Water Flow** remains observed but unselected. For Pendulum, the observed pendulum angle and light position are both selected to render unobserved sources conditionally independent.

As illustrated in Fig. 6, our proposed method performs comparably to or better than the baselines. GIN is stronger here than on the synthetic data, likely because its flow-based

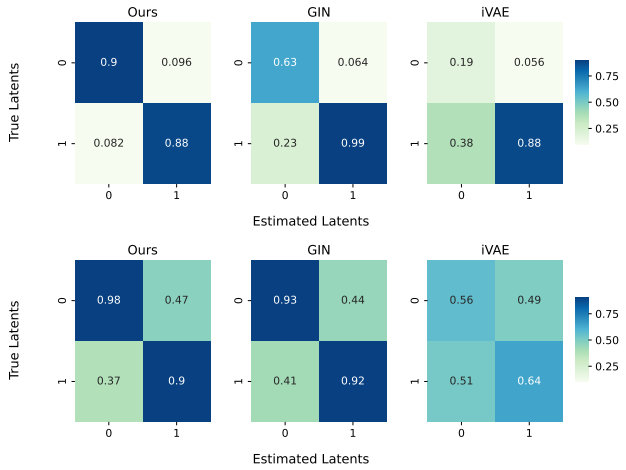


Figure 7: Mean correlation matrices for our model, GIN with selection, and iVAE with selection, matched with the best permutation on the Pendulum (top) and Flow (bottom).

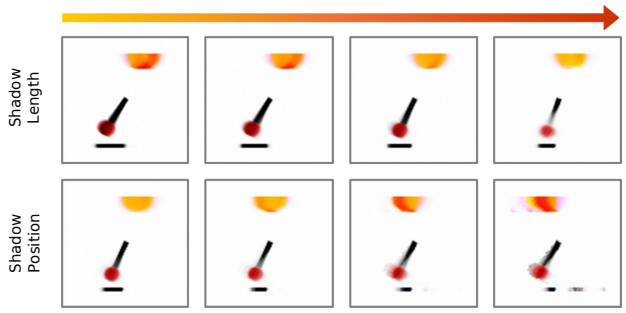


Figure 8: Latent traversals of unobserved sources. Rows vary shadow length and shadow position, respectively.

architecture provides ample capacity for images. The MCC matrices (Fig. 7) also exhibit correlation patterns consistent with disentangled representations of the unobserved source variables. This suggests that the learned representations are broadly consistent with the conditional independence structure of the underlying latent graphs in Figs. 2a and 2b.

However, we also observed that the representations in Flow were more entangled compared to those in Pendulum. In this case, $\mathbf{o}_n$ denotes an observed but unselected source (**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 future work.

5.2 QUALITATIVE ANALYSIS: LATENT TRAVERSAL

To assess disentanglement performance, we extend our model to an image reconstruction task and perform latent traversal. Using the Pendulum dataset (Fig. 2a), we use the pendulum angle and light position as selected internal auxiliaries. To extract features from high-dimensional images, we employ an encoder-decoder architecture with an MSE loss for robust compression and reconstruction.

The encoder first compresses the image into a low-dimensional representation matching the number of causal graph nodes. This representation is then transformed by our model into the structured source coordinates of the same dimensionality. Following a scene-mixture approach, each source coordinate is decoded into an additive image-layer contribution through fully connected layers, and the layers are averaged into the final reconstruction.

Fig. 8 illustrates latent-traversal reconstructions generated by varying individual unobserved source coordinates. The upper row shows the shadow-length coordinate reducing the shadow extent. Similarly, modifying the source coordinate for shadow position shifts the shadow progressively to the right while preserving other factors. This qualitative disentanglement of unobserved sources further supports the model's practical effectiveness.

6 CONCLUSION

In this paper, we investigated an underexplored yet realistic Causal Representation Learning (CRL) setting in which a subset of source variables is observable and enters the mixing mechanism as *internal* inputs. We demonstrated that this internal entanglement breaks the key invariance exploited by standard identifiability arguments based on canceling Jacobian log-determinant terms across auxiliary contexts. To address this challenge, we established an identifiability result for the selected internal-source case in which all observed internal sources are used as conditioning variables. Under a volume-preserving mixing assumption, suitable variability conditions, and additional learned-model alignment conditions, the unobserved source subspaces are identifiable up to subspace-wise invertible transformations and permutations.

Complementing this theoretical result, we proposed a graph-aware variable-selection scheme that leverages the known causal graph to choose observable sources that preserve fine-grained conditional factorizations of the unobserved sources. The identifiability theorem covers the selected-source case $O = C$; when observed-but-unselected sources remain, our method uses a graph-structured objective, and theorem-level guarantees for that case are left open. Empirically, our framework improves disentanglement in the main synthetic settings and is competitive on high-dimensional data against auxiliary-based baselines that treat internal observables as purely external conditioners, thereby reducing spurious solutions that arise from such mismatched assumptions. While the current analysis relies on the causal graph and a volume-preserving structure, our results provide a foundation for CRL under partial observability in physical systems and point to future relaxations of these assumptions.

Acknowledgements

We thank anonymous reviewers for constructive comments to improve the manuscript. This work was supported by the IITP (RS-2022-II220953/50%) and NRF (RS-2023-00211904/50%) grant funded by the Korean government.

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.

Guangyi Chen, Yifan Shen, Zhenhao Chen, Xiangchen Song, Yuewen Sun, Weiran Yao, Xiao Liu, and Kun Zhang. CariNG: Learning temporal causal representation under non-invertible generation process. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=sLZzFTMWSt.

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.

Inwoo Hwang, Yunhyeok Kwak, Yeon-Ji Song, Byoung-Tak Zhang, and Sanghack Lee. On discovery of local independence over continuous variables via neural contextual decomposition. In *Conference on Causal Learning and Reasoning*, pages 448–472. PMLR, 2023.

Inwoo Hwang, Yunhyeok Kwak, Suhyung Choi, Byoung-Tak Zhang, and Sanghack Lee. Fine-grained causal dynamics learning with quantization for improving robustness in reinforcement learning. In *Forty-first International Conference on Machine Learning*, 2024.

Inwoo Hwang, Yushu Pan, and Elias Bareinboim. From black-box to causal-box: Towards building more interpretable models. Technical Report R-127, Columbia CausalAI Laboratory, May 2025. URL https://causalai.net/r127.pdf. Columbia CausalAI Laboratory, Technical Report (R-127).

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, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley, 2001.

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.

Sujin Jeon, Inwoo Hwang, Sanghack Lee, and Byoung-Tak Zhang. Locality-aware concept bottleneck model. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

Shruti Joshi, Théo Saulus, Wieland Brendel, Philippe Brouillard, Dhanya Sridhar, and Patrik Reizinger. Who guards the guardians? the challenges of evaluating identifiability of learned representations, 2026.

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 2020a.

Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020b.

Jonghwan Kim, Inwoo Hwang, and Sanghack Lee. Causal discovery with deductive reasoning: One less problem. In *Uncertainty in Artificial Intelligence*, pages 1999–2017. PMLR, 2024.

Bohdan Kivva, Goutham Rajendran, Pradeep Kumar Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Murat Kocaoglu, Amin Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. *Advances in neural information processing systems*, 32, 2019.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

Adam Li, Amin Jaber, and Elias Bareinboim. Causal discovery from observational and interventional data across multiple environments. *Advances in Neural Information Processing Systems*, 36:16942–16956, 2023.

Adam Li, Yushu Pan, and Elias Bareinboim. Disentangled representation learning in non-markovian causal systems. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 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.

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.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

Goutham Rajendran, Simon Buchholz, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. From causal to concept-based representation learning. *Advances in Neural Information Processing Systems*, 37: 101250–101296, 2024.

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.

Amrith Setlur, Barnabas Poczos, and Alan W. Black. Nonlinear ISA with auxiliary variables for learning speech representations, 2020. To be presented at Interspeech 2020.

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.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

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.

Matthew Willetts and Brooks Paige. I don't need u: Identifiable non-linear ICA without side information, 2021.

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.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. Spotlight.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17): 1873–1896, 2008.

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.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Wo1HF2wWNZb.

On Causal Representation Learning with Internal 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]Brown University, Providence, Rhode Island, USA
[3]Causal Artificial Intelligence Lab, Columbia University, New York, USA

A DISCUSSION

Reframing the Auxiliary-Variable Framework To handle complex latent dependencies as shown in Fig. 1c, prior works typically split latent sources into singleton independent source coordinates s_i $(i = 1, \dots, n_i)$ and conditionally dependent clusters $\mathbf{s}_{c_j}$ $(j = 1, \dots, d)$, with these pieces jointly covering $\{s_1, \dots, s_n\}$. This leads to the factorization introduced by Zheng and Zhang [2023]:

$$p_{\mathbf{s}|\mathbf{u}}(\mathbf{s}|\mathbf{u}) = \prod_{i=1}^{n_i} p_{s_i}(s_i) \prod_{j=1}^{d} p_{\mathbf{s}_{c_j}|\mathbf{u}}(\mathbf{s}_{c_j}|\mathbf{u}). \tag{8}$$

While Eq. 8 is appropriate for *external* auxiliaries $\mathbf{u}$ that *do not enter the mixing mechanism* (i.e., $\mathbf{u}$ modulates the latent distribution but not the mixing map), it is inapplicable to our setting in which observable sources $\mathbf{o}$ act as *internal* inputs. In Eq. 5, singleton independent variables are absorbed into the partition $\{\mathbf{z}_{c_j}\}_{j=1}^{d}$, yielding the internal-auxiliary factorization

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

This shift from external conditioning to internal entanglement is fundamental. In the internal regime, $\mathbf{o}$ can alter the mixing function's local geometry, rendering prior identifiability proofs—which rely on cancellations of Jacobian log-determinant terms across auxiliary contexts—no longer applicable. Our framework addresses this failure mode by introducing a volume-preserving constraint to stabilize the observation-dependent distortion.

Beyond MCC: DCI-Style Structure Diagnostics While the Mean Correlation Coefficient (MCC) is a widely accepted metric for measuring identifiability [Hyvärinen and Morioka, 2016], our analysis suggests it is insufficient for the internal-

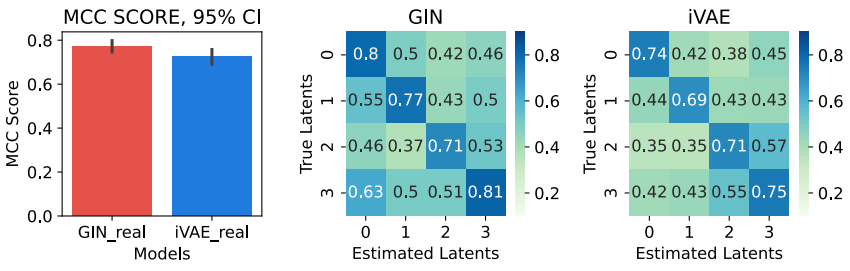


Figure 9: MCC scores and mean correlation matrices of GIN and iVAE matched with the best permutation for the setting of Fig. 1c.

auxiliary setting. The MCC score is computed as:

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

Here, $q = \dim(\mathbf{z})$ is the number of unobserved target dimensions being evaluated. In our scenario, a high MCC score can be achieved spuriously if the estimated target coordinates capture shared information from $\mathbf{o}$ without truly disentangling the unobserved sources $\mathbf{z}$. As illustrated in Fig. 9, although GIN and iVAE achieve MCC scores around 0.7, their correlation matrices reveal significant information leakage across dimensions. This occurs because existing methods do not account for the entanglement induced when the mixing function includes observable sources. Consequently, we report MCC-based DCI-style scores [Eastwood and Williams, 2018] as a diagnostic for whether the representation aligns with the graph-implied conditional independence structure (e.g., with reduced cross-factor leakage) without such spurious correlations.

Internal Sources vs. Raw Observations A clear distinction between observed sources ($\mathbf{o}$) and raw observations ($\mathbf{x}$) is fundamental to our framework. The raw observation represents measurable low-level data (e.g., an image), which is the final output of the generative process $\mathbf{x} = g(\mathbf{s}) = g(\mathbf{o}, \mathbf{z})$. In contrast, $\mathbf{o}$ refers to high-level "causal handles"—intrinsic factors like joint angles in a robotic arm or light positions in a physical system—that are observable yet intertwined with the unobserved sources $\mathbf{z}$ during the mixing process. By treating $\mathbf{o}$ as an internal input rather than an external label, we explicitly model how physical state information shapes the data manifold.

Recovery in Collider Structures In theory, collider structures (e.g., Fig. 1d) present unique challenges for recovery. We address this by incorporating graphical constraints that encourage the learned representation to retain graph-compatible dependencies during training. This grounding can be empirically useful when a collider node serves as an observable source, as it provides a supervised anchor for the latent representation. This highlights the potential of leveraging domain knowledge to *improve recovery* (or *aid learning*) even in complex, non-factorizable latent graphs.

B THEORETICAL ANALYSIS

We begin with the standard identifiability-up-to-equivalence formulation used in nonlinear ICA and identifiable latent-variable models [Hyvarinen et al., 2019, Khemakhem et al., 2020a]. Using a Structural Causal Model (SCM; Pearl, 2009), we represent the data-generating process over latent sources as

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

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 2. *(Identifiability). Suppose the observations* $\mathbf{x}$ *are generated by the true source mechanism specified by* $\Theta = (\mathbf{f}, p(\boldsymbol{\epsilon}), \mathbf{g})$ *given in Eqs. 1 and 2. The learned generative model parameterized by* $\hat{\Theta} = \left(\hat{\mathbf{f}}, \hat{p}(\boldsymbol{\epsilon}), \hat{\mathbf{g}}\right)$ *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* $\mathcal{A}$ *be a class of invertible transformations (e.g., subspace-wise transformations and permutations). We say that the model is identifiable up to* $\mathcal{A}$ *if*

$$p_{\hat{\Theta}}(\mathbf{x}) = p_{\Theta}(\mathbf{x}) \;\Rightarrow\; \exists A \in \mathcal{A} \text{ s.t. } \hat{\mathbf{g}} = \mathbf{g} \circ A.$$

Thus, the recovered sources are identifiable up to the same transformation class $\mathcal{A}$.

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

Additional assumptions for subspace identifiability. In addition to the assumptions stated inside each proposition, the subspace-wise arguments use the following technical assumptions. We state them separately because they play different roles: support and smoothness make the differentiated likelihood equations well defined; invariance and learned-side factorization fix the linear system and make the learned-side mixed derivatives vanish; the full block-Hessian variability and block-matching assumptions turn the linear-system solution into subspace-wise identifiability; and the graphical validity assumption justifies using graph-based d-separation to select the factorization.

1. *Common support and smooth positive block densities.* For the auxiliary values used in the rank condition, each conditional block density is positive on a common interior support, and $\log p(\mathbf{z}_{c_j} \mid \mathbf{o}_C)$ or $\log p(\mathbf{z}_{c_j} \mid \mathbf{u})$ is twice continuously differentiable in the source coordinates. This assumption is needed so that the score and all within-block Hessian entries are well defined at the same source point across auxiliary values. Smooth positive densities and differentiable sufficient statistics are standard in auxiliary-variable nonlinear ICA and identifiable VAE analyses [Hyvarinen et al., 2019, Khemakhem et al., 2020a], and the full second-order tensor version is used in nonlinear ISA with multidimensional subspaces [Setlur et al., 2020].

2. *Smooth invertible source-coordinate map.* The true and learned source-coordinate transformations are twice differentiable diffeomorphisms on the relevant support. This is required for the change-of-variables step and for applying the chain rule twice when differentiating the log-density equality with respect to learned coordinates. This follows the usual smooth invertible mixing/reparameterization assumptions in nonlinear ICA and CRL identifiability results [Hyvarinen et al., 2019, Khemakhem et al., 2020a, Zheng and Zhang, 2023].

3. *Common learned-to-true transformation across auxiliary values.* The transformation from learned to true unobserved coordinates is common across the auxiliary values used in the linear system: $\mathbf{z} = r(\hat{\mathbf{z}})$, not $\mathbf{z} = r(\hat{\mathbf{z}}, \mathbf{o}_C)$ or $\mathbf{z} = r(\hat{\mathbf{z}}, \mathbf{u})$. Otherwise, the unknown quantities $z_{a,k}$ and $z_{a,kv}$ would change with the row index and the rank condition could not be applied to a single fixed unknown vector. In the external-auxiliary statement, this property follows from the fixed source-coordinate map induced by two mixing maps that do not take $\mathbf{u}$ as an argument. In Prop. 1, it should be read as an additional invariance requirement on the source-coordinate reparameterization, separate from the VP condition that removes the log-determinant terms.

4. *Correct learned-side block factorization.* The learned source distribution is assumed to satisfy the same selected block factorization as the true source distribution for the proposition under consideration. This condition is needed because the mixed derivative on the learned side is set to zero only when the two learned coordinates belong to different learned blocks. Such model-class matching is standard in identifiable latent-variable results, where the learned model is assumed to belong to the same factorized family as the data-generating model [Hyvarinen et al., 2019, Khemakhem et al., 2020a].

5. *Full block-Hessian sufficient variability.* For a block $\mathbf{z}_{c_j}$ with dimension $m_j > 1$, the variability vector includes the full symmetric block Hessian, not only coordinatewise second derivatives. This is required because differentiating $\log p(\mathbf{z}_{c_j} \mid \cdot)$ through $\mathbf{z} = r(\hat{\mathbf{z}})$ produces within-block terms $\partial_{ab} \log p(\mathbf{z}_{c_j} \mid \cdot)$ for $a \neq b$. The scalar score-curvature rank condition follows the auxiliary-variable nonlinear ICA proof template [Hyvarinen et al., 2019], while the multidimensional derivative-tensor form is aligned with nonlinear ISA [Setlur et al., 2020].

6. *Block-support matching/no-coarsening.* After the full-rank linear system implies that two learned blocks cannot both affect the same true block, we assume a standard block-matching condition: every true block has a nondegenerate Jacobian connection to exactly one learned block of matching dimension, so true blocks are neither merged across several learned blocks nor split into smaller learned blocks. This is needed because invertibility alone rules out a globally singular map but does not by itself turn the block-support exclusion into a one-to-one block permutation. The resulting equivalence class matches the ISA-style conclusion of permutation across subspaces and invertible transformations within subspaces [Theis, 2006, Zheng and Zhang, 2023].

7. *Graphical validity of the selected factorization.* When the partition $\{\mathbf{z}_{c_j}\}_{j=1}^{d}$ is obtained from the known graph by d-separation, we assume the source distribution is Markov and faithful to the latent Bayesian network. This is required to translate graph-implied separations into the conditional factorization used in Prop. 1; without it, the graph search can return a partition that is not a distributional independence statement. This is the standard causal-graph assumption underlying d-separation-based conditional independence reasoning [Geiger et al., 1990, Pearl, 2009].

Proposition 2. *(Identifiability with external information) Let $q = \sum_{j=1}^{d} \dim(\mathbf{z}_{c_j}) = \dim(\mathbf{z})$ denote the dimension of the partitioned source coordinates. Let $m_j = \dim(\mathbf{z}_{c_j})$, $\phi_j(\mathbf{z}_{c_j}, \mathbf{u}) = \log p(\mathbf{z}_{c_j} \mid \mathbf{u})$, and $M = q + \sum_{j=1}^{d} m_j(m_j+1)/2$. Here $\mathbf{u}$ denotes an* external *auxiliary that does not enter the mixing, distinct from the internal selected sources $\mathbf{o}_C$ of Prop. 1. Suppose the following assumptions hold:*

1. *The observed data are generated by an invertible target-source mixing map $\mathbf{x} = g(\mathbf{z})$, and both true and learned target-source distributions satisfy $p(\mathbf{z} \mid \mathbf{u}) = \prod_{j=1}^{d} p(\mathbf{z}_{c_j} \mid \mathbf{u})$ and $p(\hat{\mathbf{z}} \mid \mathbf{u}) = \prod_{j=1}^{d} p(\hat{\mathbf{z}}_{c_j} \mid \mathbf{u})$.*

2. *The external auxiliary $\mathbf{u}$ does not enter the mixing: $\mathbf{u}$ is not an argument of the mixing map g (no direct edge $\mathbf{u} \to \mathbf{x}$), so $\partial \mathbf{x}/\partial \mathbf{u} = 0$ at fixed $\mathbf{z}$, and any dependence of $\mathbf{x}$ on $\mathbf{u}$ is mediated through $\mathbf{z}$.*

3. *For every value of* $\mathbf{z}$*, there exist* $M+1$ *values of* $\mathbf{u}$*, such that the* M *vectors* $\mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{u}_i) - \mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{u}_0)$ *are linearly independent, where*

$$\mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{u}_i) = \big(\{\nabla_{\mathbf{z}_{c_j}} \phi_j(\mathbf{z}_{c_j}, \mathbf{u}_i)\}_{j=1}^d, \{\text{vech}(\nabla^2_{\mathbf{z}_{c_j}} \phi_j(\mathbf{z}_{c_j}, \mathbf{u}_i))\}_{j=1}^d\big).$$

Then, under App. Assumptions 1, 2, 4, and 5, together with the block-matching condition in App. Assumption 6, all the unobserved source subspaces $\{\mathbf{z}_{c_i}\}_{i=1}^d$ *are identifiable up to subspace-wise invertible transformations and permutations.*

Proof. Assume auxiliary-conditional observational equivalence between the true and estimated models, i.e., $p_g(\mathbf{x} \mid \mathbf{u}) = p_{\hat{g}}(\mathbf{x} \mid \mathbf{u})$ for every value of $\mathbf{u}$. In the external case, the learned-to-true map is independent of $\mathbf{u}$ because neither mixing map takes $\mathbf{u}$ as an argument. Let $h : \mathbf{z} \to \hat{\mathbf{z}}$ denote the transformation from true sources to estimated sources. Thus, we can write $\hat{g} = g \circ h^{-1}$, equivalently,

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

by repeatedly applying the chain rule. By App. Assumption 2, $\mathbf{J}_h(\mathbf{z})$ must be invertible and have a non-zero determinant because $\mathbf{J}_{\hat{g}}(\hat{\mathbf{z}})$ and $\mathbf{J}_g(\mathbf{z})$ have full column rank. The change-of-variable rule then gives

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

Taking logarithms on both sides yields

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

According to Prop. 2, Assumption 1, App. Assumption 4, and the fact that $\{\mathbf{z}_{c_i}\}_i$ partitions $\mathbf{z}$, the joint log densities can be decomposed as

$$\sum_{j=1}^d \log p(\mathbf{z}_{c_j} \mid \mathbf{u}) + \log |\det(\mathbf{J}_{h^{-1}}(\hat{\mathbf{z}}))| = \sum_{j=1}^d \log p(\hat{\mathbf{z}}_{c_j} \mid \mathbf{u}).$$

Thus, for values $\mathbf{u}_0, \ldots, \mathbf{u}_M$, we have $M+1$ equations. The log-determinant term is independent of $\mathbf{u}$, so subtracting the equation corresponding to $\mathbf{u}_0$ from each equation corresponding to $\mathbf{u}_1, \ldots, \mathbf{u}_M$ removes it and results in M equations:

$$\sum_{i=1}^d (\log p(\mathbf{z}_{c_i} \mid \mathbf{u}_j) - \log p(\mathbf{z}_{c_i} \mid \mathbf{u}_0)) = \sum_{i=1}^d (\log p(\hat{\mathbf{z}}_{c_i} \mid \mathbf{u}_j) - \log p(\hat{\mathbf{z}}_{c_i} \mid \mathbf{u}_0)) \tag{9}$$

We take the derivatives of both sides of Eq. 9 with respect to $\hat{z}_k$ and $\hat{z}_v$ where $k, v \in \{1, \ldots, q\}$ and they are not indices of the same subspace. App. Assumptions 1 and 2 justify these derivatives. In the external case, the fixed mixing maps make $\mathbf{z} = r(\hat{\mathbf{z}})$ independent of the auxiliary rows, so the unknown derivatives remain fixed. The mixed derivative of the right-hand side of Eq. 9 is zero because k and v are not indices of the same subspace and the learned-side factorization in App. Assumption 4 holds. For the left-hand side, write

$$\phi_i^{(j,0)}(\mathbf{z}_{c_i}) = \log p(\mathbf{z}_{c_i} \mid \mathbf{u}_j) - \log p(\mathbf{z}_{c_i} \mid \mathbf{u}_0),$$

and let $\partial_a \phi_i^{(j,0)}$ and $\partial_{ab} \phi_i^{(j,0)}$ denote its first and second derivatives with respect to coordinates z_a, z_b in block c_i. Also write $z_{a,k} = \partial z_a / \partial \hat{z}_k$ and $z_{a,kv} = \partial^2 z_a / (\partial \hat{z}_k \partial \hat{z}_v)$. The mixed derivative yields

$$\sum_{i=1}^d \left[\sum_{a \in c_i} \partial_a \phi_i^{(j,0)} z_{a,kv} + \sum_{a \in c_i} \partial_{aa} \phi_i^{(j,0)} z_{a,k} z_{a,v} + \sum_{\substack{a<b \\ a,b \in c_i}} \partial_{ab} \phi_i^{(j,0)} (z_{a,k} z_{b,v} + z_{b,k} z_{a,v}) \right] = 0, \tag{10}$$

for each $j = 1, \ldots, M$. The unknown vector consists of the entries $z_{a,kv}$ and $z_{a,k} z_{a,v}$ for every coordinate a, together with $z_{a,k} z_{b,v} + z_{b,k} z_{a,v}$ for $a < b$ within each block, and has dimension M.

Considering Prop. 2, Assumption 3 and the full block-Hessian role made explicit in App. Assumption 5, the resulting $M \times M$ coefficient matrix has full rank. Hence the only solution of Eq. 10 is

$$z_{a,kv} = 0, \quad z_{a,k} z_{a,v} = 0, \quad z_{a,k} z_{b,v} + z_{b,k} z_{a,v} = 0,$$

where the first two equalities hold for every coordinate a, and the third for all $a < b$ in the same true block. Therefore, for each true block, the block-gradient vectors with respect to $\hat{z}_k$ and $\hat{z}_v$ cannot both be nonzero. App. Assumption 6, together with the invertibility of the source-coordinate transformation, then yields a block permutation and invertible within-block transformations. Equivalently, the inverse mapping from $\hat{\mathbf{z}}$ to $\mathbf{z}$ remains block-wise invertible. □

We now prove the selected-source internal-auxiliary identifiability statement in Prop. 1, where the selected auxiliary directly influences the observation $\mathbf{x}$ through the mixing function.

Proposition 1 (Identifiability with internal observed sources)**.** *Let $q = \sum_{j=1}^{d} \dim(\mathbf{z}_{c_j}) = \dim(\mathbf{z})$ denote the dimension of the partitioned unobserved coordinates. For each block, let $m_j = \dim(\mathbf{z}_{c_j})$, $\phi_j(\mathbf{z}_{c_j}, \mathbf{o}_C) = \log p(\mathbf{z}_{c_j} \mid \mathbf{o}_C)$, and define $M = q + \sum_{j=1}^{d} m_j(m_j+1)/2$. Suppose the following assumptions hold:*

1. *We consider the selected-source case $O = C$, so $\mathbf{o}_n = \emptyset$ and the source vector in the change-of-variables argument is $\mathbf{s} = (\mathbf{o}_C, \mathbf{z})$. The learned model is restricted to the observed-coordinate preserving class: for all relevant $(\mathbf{o}_C, \mathbf{z})$, $\hat{g}^{-1}(g(\mathbf{o}_C, \mathbf{z})) = (\mathbf{o}_C, \hat{\mathbf{z}})$, and the marginal law of $\mathbf{o}_C$ is shared by the true and learned models. The observed data and sources are generated from Eq. 1, and both true and learned target-source distributions satisfy the conditional factorization in Eq. 5, i.e., given the selected conditioning set $\mathbf{o}_C$.*
2. *The true and learned mixing functions are volume-preserving over these source coordinates, i.e., $|\det(\mathbf{J}_g(\mathbf{s}))| = |\det(\mathbf{J}_{\hat{g}}(\hat{\mathbf{s}}))| = 1$, where $\mathbf{s} = (\mathbf{o}_C, \mathbf{z})$ and $\hat{\mathbf{s}} = (\mathbf{o}_C, \hat{\mathbf{z}})$.*
3. *For every value of $\mathbf{z}$, there exist M values of $\mathbf{o}_C$ such that the M full score-curvature vectors $\mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{o}_C^{(i)})$ are linearly independent, where*

$$\mathbf{w}_{\text{full}}(\mathbf{z}, \mathbf{o}_C^{(i)}) = \big(\{\nabla_{\mathbf{z}_{c_j}} \phi_j(\mathbf{z}_{c_j}, \mathbf{o}_C^{(i)})\}_{j=1}^{d}, \{\text{vech}(\nabla^2_{\mathbf{z}_{c_j}} \phi_j(\mathbf{z}_{c_j}, \mathbf{o}_C^{(i)}))\}_{j=1}^{d}\big).$$

Then, under the standing assumptions in App. B—in particular the conditioning-independent reparameterization in App. Assumption 3 and the block-matching condition in App. Assumption 6—the differentiated likelihood equations imply block-support exclusion: two learned blocks cannot both affect the same true block, and all the unobserved source subspaces $\{\mathbf{z}_{c_i}\}_{i=1}^{d}$ are identifiable up to subspace-wise invertible transformations and permutations.

Proof. In Prop. 1, we restrict to $O = C$ and condition on the selected set $\mathbf{o}_C$. When the selected factorization is obtained from the graph, App. Assumption 7 justifies reading d-separation as a distributional factorization. We use the full-density change-of-variables identity for the invertible map $g : (\mathbf{o}_C, \mathbf{z}) \mapsto \mathbf{x}$, not a Lebesgue density $p(\mathbf{x} \mid \mathbf{o}_C)$ on $\mathbb{R}^n$. Assume observational equivalence of the marginal law of $\mathbf{x}$ and the observed-coordinate consistency stated in Prop. 1, Assumption 1. In particular, the learned inverse preserves the selected observed coordinate, $\hat{g}^{-1}(g(\mathbf{o}_C, \mathbf{z})) = (\mathbf{o}_C, \hat{\mathbf{z}})$, on the support. Let $\mathbf{s} = (\mathbf{o}_C, \mathbf{z})$ and $\hat{\mathbf{s}} = (\mathbf{o}_C, \hat{\mathbf{z}})$, where the selected conditioning coordinates are fixed to the measured value $\mathbf{o}_C$. By App. Assumption 2, the change-of-variable rule gives

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

Since $p(\mathbf{s}) = p(\mathbf{z} \mid \mathbf{o}_C) \cdot p(\mathbf{o}_C)$ and $\hat{p}(\hat{\mathbf{s}}) = \hat{p}(\hat{\mathbf{z}} \mid \mathbf{o}_C) \cdot \hat{p}(\mathbf{o}_C)$, we equivalently have

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

The factor for $\mathbf{o}_C$ is common, $p(\mathbf{o}_C) = \hat{p}(\mathbf{o}_C)$, because the selected observed source vector is shared by the true and learned models under the observed-coordinate preserving restriction.

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

Taking logarithms on both sides yields

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

By Prop. 1, Assumption 2, the log-determinant terms are zero. According to Prop. 1, Assumption 1, App. Assumption 4, and the fact that $\{\mathbf{z}_{c_i}\}_i$ partitions $\mathbf{z}$, the joint log densities can then be decomposed as

$$\sum_{j=1}^{d} \log p(\mathbf{z}_{c_j} \mid \mathbf{o}_C) = \sum_{j=1}^{d} \log \hat{p}(\hat{\mathbf{z}}_{c_j} \mid \mathbf{o}_C).$$

Thus, for values $\mathbf{o}_C^{(1)}, \ldots, \mathbf{o}_C^{(M)}$, we have M equations. Unlike Prop. 2, no baseline subtraction is needed here because the VP assumption has already set the log-determinant terms to zero. We take the derivatives of both sides of the above equation with respect to $\hat{z}_k$ and $\hat{z}_v$ where $k, v \in \{1, \ldots, q\}$ and they are not indices of the same subspace. App. Assumptions 1 and 2 justify these derivatives, and App. Assumption 3 keeps the unknown derivatives of $\mathbf{z} = r(\hat{\mathbf{z}})$ fixed across the selected-conditioning rows. The right-hand side has zero mixed derivative because k and v are not indices of the same subspace and the learned-side factorization in App. Assumption 4 holds. For the left-hand side, write $\phi_i^{(t)}(\mathbf{z}_{c_i}) = \log p(\mathbf{z}_{c_i} \mid \mathbf{o}_C^{(t)})$, let $\partial_a \phi_i^{(t)}$ and $\partial_{ab}\phi_i^{(t)}$ denote first and second derivatives with respect to coordinates z_a, z_b in block c_i, and write $z_{a,k} = \partial z_a / \partial \hat{z}_k$ and $z_{a,kv} = \partial^2 z_a / (\partial \hat{z}_k \partial \hat{z}_v)$. The mixed derivative gives the row

$$\sum_{i=1}^{d} \left[\sum_{a \in c_i} \partial_a \phi_i^{(t)} z_{a,kv} + \sum_{a \in c_i} \partial_{aa} \phi_i^{(t)} z_{a,k} z_{a,v} + \sum_{\substack{a<b \\ a,b \in c_i}} \partial_{ab} \phi_i^{(t)} (z_{a,k} z_{b,v} + z_{b,k} z_{a,v}) \right] = 0, \tag{11}$$

where $t = 1, \ldots, M$. The unknown vector consists of $z_{a,kv}$ and $z_{a,k} z_{a,v}$ for every coordinate a, together with $z_{a,k} z_{b,v} + z_{b,k} z_{a,v}$ for $a < b$ within each block, so its dimension is M, matching Prop. 1, Assumption 3.

Considering Prop. 1, Assumption 3 and the full block-Hessian role made explicit in App. Assumption 5, the resulting $M \times M$ coefficient matrix has full rank. The only solution of Eq. 11 is

$$z_{a,kv} = 0, \quad z_{a,k} z_{a,v} = 0, \quad z_{a,k} z_{b,v} + z_{b,k} z_{a,v} = 0,$$

where the first two equalities hold for every coordinate a, and the third for all $a < b$ in the same true block. Hence, for any pair of learned coordinates from different learned subspaces, the corresponding block-gradient vectors cannot both be nonzero within the same true block. By App. Assumption 6 and the invertibility of the source-coordinate transformation, the learned partition must align with the true partition up to a block permutation and invertible within-block transformations. Equivalently, the inverse mapping from $\hat{\mathbf{z}}$ to $\mathbf{z}$ remains block-wise invertible. □

C BAYES-BALL ALGORITHM

A standard graphical criterion for conditional independence is *d-separation* [Geiger et al., 1990]. We aim to find clusters with intra-cluster d-connectedness and inter-cluster d-separation.

We employ the *Bayes-ball* algorithm to examine the conditional independence of two node sets on the given graph $\mathcal{G}$. In Alg. 1, $Partition(G, T, O)$ applies the Bayes-ball routine to each unassigned node in $O^- = [n] \setminus O$ and forms connected components of the pairwise d-connection relation given T. The routine below returns the set of nodes d-connected to an input seed set R, excluding conditioned and observed nodes from the returned set.

D EXPERIMENTAL DETAILS

The implementation of the experiments is based on Liang et al. [2023]. Table 2 shows the hyperparameters for training our model, GIN, and iVAE.

Evaluation matching protocol. For each synthetic graph, all metrics are computed on the unobserved target sources, i.e., the coordinates left after removing the selected and observed sources (e.g., $\{0, 1\}$ for graph c_real). Because the graph constraint confines each target source to the corresponding learned target block in our model, we align the true target block with the same-indexed learned target block, resolving only the within-block permutation. For the GIN and iVAE baselines, whose learned coordinates are not pre-aligned, we instead align the true target block with the *full* learned representation and let the metric select the best-correlated coordinates—a strictly larger matching search space, and thus a conservative choice that favors the baselines.

E ADDITIONAL ABLATION STUDIES

We report additional experiments on nonlinear source mechanisms, nonlinear evaluation metrics, the role of the VP assumption in the data-generating mixing, and the coefficient range used in the synthetic SCM. These runs use 25 seeds and 20 training epochs unless otherwise stated.

Algorithm 2: 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 directed states $(node, direction)$
Initialize an empty queue Q
for each node n in R **do**
  Add (n, up) to Q
while Q is not empty **do**
  $(node, direction) \leftarrow Q.pop()$
  if $(node, direction) \in V$ **then**
    continue
  Add $(node, direction)$ to V
  if $node \in C$ **and** $direction \neq$ down **then**
    continue
  if $direction =$ up **then**
    for each $parent$ of $node$ in G **do**
      Add $(parent, \text{up})$ to Q
    for each $child$ of $node$ in G **do**
      Add $(child, \text{down})$ to Q
  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
    if $node \in C$ **or** $check =$ true **then**
      for each $parent$ of $node$ in G **do**
        Add $(parent, \text{up})$ to Q
    else
      for each $child$ of $node$ in G **do**
        Add $(child, \text{down})$ to Q
  if $node \notin C$ **and** $node \notin O$ **then**
    Add $node$ to R
return R

The nonlinear SCM ablation in Table 3 changes how the latent sources are generated, not the mixing map. MCC-based DCI-style disentanglement remains within cross-seed variation across SCM choices, while MCC can drop because element-wise correlation is sensitive to within-block rotation and scaling. The nonlinear regression metrics in Table 4 show recovery under predictor-based nonlinear evaluation as well, with R^2 remaining in the .53–.66 range across the six settings. Table 5 separates the theoretical VP assumption from empirical recovery under VP-method training: recovery degrades consistently under non-VP data-generating mixing, which we interpret as finite-sample sensitivity of this pipeline to the VP property rather than a claim that VP is necessary for identifiability in general.

We also ablated the SCM coefficient range because the paper setting $\beta_{ij} \sim \text{Uniform}[0.5, 1.0]$ is positive-only. Holding $\varepsilon_i \sim \mathcal{N}(0, 1)$ fixed, we compared the paper setting with a sign-varying setting $\beta_{ij} \sim \text{Uniform}[-1, 1]$ under $|\beta_{ij}| \geq 0.5$ rejection sampling and a small-magnitude setting $\beta_{ij} \sim \text{Uniform}[0.1, 0.3]$. The sign-varying setting is comparable to the paper setting on the four-node ablation graph c (DCI-style .32 vs. .30) and gives a modest drop on c_real (DCI-style .26 vs. .33); the small-magnitude setting is strongest on both graphs, consistent with weaker parent-child coupling approaching an easier near-ICA regime.

F GRAPH-STRUCTURED PREDICTION PRIOR

The graphical constraint in Sec. 4.3 is implemented by a structural prediction prior using a fixed graph-indexed slot convention. For each observed-but-unselected coordinate, the prior constructs a node-specific predictor that reads the

Table 2: 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
Coupling blocks	4+4+2
Optimizer	Adam
Batch size	1024
Epochs	Flow: 50, Pend.: 80

(d) High-dimensional data

GIN	
LR scheduler	-
Learning rate	0.001
Coupling blocks	4+4+2
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

Table 3: Linear vs. location-scale nonlinear SCM ablation. The mixing g is kept as a VP flow; only the latent SCM changes.

Graph	SCM	DCI-style disent.	MCC
a	linear	$.220 \pm .079$	$.687 \pm .076$
a	location-scale	$.268 \pm .123$	$.513 \pm .134$
b	linear	$.176 \pm .097$	$.627 \pm .107$
b	location-scale	$.272 \pm .105$	$.582 \pm .126$
c_real	linear	$.334 \pm .189$	$.611 \pm .123$
c_real	location-scale	$.288 \pm .236$	$.564 \pm .188$

representation coordinates at the graph-predecessor indices together with the target-coordinate slot, so its input dimension is the parent count plus one. Here, $C^- = [n] \setminus C$ denotes the indices not selected for conditioning.

G EXTENDED DISCUSSIONS, LIMITATIONS, AND FUTURE WORK

To improve readability and reduce possible ambiguity in interpreting our theoretical and empirical results, we provide additional clarifications on our core assumptions, the scope of our framework, and discuss the limitations alongside promising future directions.

Scope relative to prior auxiliary-based identifiability settings. Our setting is specifically designed for cases in which observable sources act as *internal inputs* to the data-generating mechanism, rather than as externally provided side information. This distinction fundamentally alters the structure of the identifiability argument. Accordingly, our results are best viewed as complementary to prior external-auxiliary results, which correspond to a simpler regime in which the observation-dependent distortion issue does not arise.

The Known Causal Graph Assumption and Prospects for Relaxation. The primary objective of this work is not latent graph discovery, but rather to study how graphical information can be leveraged to improve the recoverability of latent sources. In some applied domains (e.g., engineered physical systems or robotics), where internal sources (e.g., joint angles) are governed by known mechanics or design constraints, assuming a known causal graph can be a practical modeling premise. Under this scope, the known-graph assumption allows us to focus on critical questions regarding optimal conditioning to refine the independence structure among unobserved sources. The graph-guided selection step reads conditional independences from d-separation, so it should be understood under the usual Markov and faithfulness conditions for the latent Bayesian network. Nevertheless, there is meaningful potential to relax this assumption in future work. While

Table 4: Nonlinear metric validation using GBR-based DCI and kernel-ridge R^2.

Graph	SCM	DCI disent. (GBR)	R^2 (kernel-ridge)
a	linear	$.47 \pm .12$	$.61 \pm .07$
a	location-scale	$.30 \pm .20$	$.66 \pm .13$
b	linear	$.37 \pm .10$	$.62 \pm .11$
b	location-scale	$.35 \pm .15$	$.65 \pm .15$
c_real	linear	$.35 \pm .22$	$.57 \pm .13$
c_real	location-scale	$.36 \pm .31$	$.53 \pm .19$

Table 5: VP vs. non-VP data-generating mixing ablation. The non-VP setting replaces the VP GIN coupling block with a Glow-style affine coupling block while keeping the depth, subnet, permutations, optimizer, hyperparameters, and SCM fixed.

Graph	VP DCI-style disent.	non-VP DCI-style disent.	VP MCC	non-VP MCC
a	$.22 \pm .08$	$.18 \pm .06$	$.69 \pm .08$	$.52 \pm .19$
b	$.18 \pm .10$	$.13 \pm .08$	$.63 \pm .11$	$.45 \pm .17$
c_real	$.33 \pm .19$	$.21 \pm .15$	$.61 \pm .12$	$.52 \pm .12$

our current theory targets the purely observational regime, interventional approaches such as Li et al. [2024] illustrate how graphical modeling can be combined with stronger supervision. More broadly, integrating partial graph knowledge or data-driven structure learning [Pearl, 2009, Spirtes et al., 2000, Zhang, 2008, Kocaoglu et al., 2019, Hwang et al., 2023, Li et al., 2023, Kim et al., 2024, Hwang et al., 2024] into our framework could enable more flexible structural settings.

Role and Practicality of the Volume-Preserving (VP) Assumption. Our identifiability result is stated under a volume-preserving assumption ($|\det(\mathbf{J}_g(\mathbf{s}))| = 1$) on the mixing function. Rather than suggesting that this condition is strictly necessary in general, we adopt it as a *clean sufficient condition* that makes the new technical difficulty of the internal-auxiliary setting explicit. When observable sources $\mathbf{o}$ directly participate in the mixing process, the effective geometric distortion can depend on the observed values, which prevents the direct application of standard proof templates. The VP assumption removes this obstruction, allowing us to isolate the identifiability mechanism in a principled manner. This is a modeling assumption, not a data-side claim that can be certified by the likelihood of a VP estimator; a VP model can fit some non-VP targets well without identifying the true mechanism. In practice, we approximate the theoretical condition by restricting the encoder architecture to volume-preserving flows (e.g., GIN), which removes log-determinant terms from the objective and yields a tractable implementation. VP structure is natural in several dynamics-level settings, including Hamiltonian systems via Liouville's theorem, incompressible flows with $\nabla \cdot v = 0$, rigid-body kinematics, symplectic integrators, and HMC kernels. Pendulum and Flow in our experiments are physically grounded synthetic image-rendering settings rather than direct real-world measurements, so systematic empirical validation of VP plausibility on real data remains future work. Empirically, while enforcing VP restricts architectural choices to flow-based models, these models can remain expressive in practice. For instance, empirical findings from Appendix C of Yang et al. [2022] indicate that the GIN framework can perform implicit dimensionality reduction by shrinking the variance of redundant dimensions. This suggests practical flexibility beyond the strict theoretical requirement, though further exploration into alternative constraints or relaxation of this assumption is warranted.

Interpretation of MCC, Structure-Sensitive Evaluation, and Baselines. While the Mean Correlation Coefficient (MCC) is informative as a summary of pairwise alignment, it may not fully reflect whether the learned representation respects the desired factorization structure in the internal-auxiliary regime. For example, a representation can attain strong diagonal correlations (and thus a high MCC) while still exhibiting non-negligible cross-factor leakage; conversely, when the target equivalence class allows within-subspace rotations, element-wise MCC can understate valid subspace-level recovery. This matches the broader warning of Joshi et al. [2026] that identifiability metrics can produce false positives or false negatives when their implicit assumptions do not match the theoretical equivalence class. Therefore, we additionally inspect correlation patterns and report MCC-based DCI-style scores, with predictor-based nonlinear checks in App. E, to better capture disentanglement quality in a manner aligned with the graph-implied structure. Furthermore, to isolate the effect of modeling internal auxiliaries, our empirical comparisons provide baseline methods with the same selected auxiliary variables whenever applicable, thereby reducing confounding due to differences in input information.

Algorithm 3: GraphPrior structural prediction constraint

Require: Learned representation $\hat{\mathbf{s}}$, observed-but-unselected set $O \setminus C$, adjacency matrix A, node-wise predictors $\{\hat{f}_i\}$

$\mathcal{L}_{\text{graph}} \leftarrow 0$
for $i \in O \setminus C$ **do**
  $d_i \leftarrow |\{j : A_{ji} = 1\}|$ {number of parents of source node i}
  $\hat{\mathbf{s}}_i^{\text{slot}} \leftarrow d_i$ graph-predecessor coordinate slots plus the target-coordinate slot from $\hat{\mathbf{s}}_{C^-}$
  $\hat{o}_i \leftarrow \hat{f}_i(\hat{\mathbf{s}}_i^{\text{slot}})$
  $\mathcal{L}_{\text{graph}} \leftarrow \mathcal{L}_{\text{graph}} + \log p(o_i \mid \hat{o}_i)$
return $\mathcal{L}_{\text{graph}}$

Theoretical vs. Practical Optimization Limitations. While our theory provides population-level guarantees, practical training is performed with finite samples, finite-capacity models, and multi-term objectives. As a result, residual entanglement may remain even when the theoretical conditions suggest recoverability. Prop. 1 also relies on a common auxiliary-independent reparameterization $\mathbf{z} = r(\hat{\mathbf{z}})$ as stated in App. Assumption 3. Because the mixing slice $g(\mathbf{o}_C, \cdot)$ varies with $\mathbf{o}_C$, this is a genuine restriction that rules out $\mathbf{o}_C$-dependent gauge transformations. Relaxing this restriction is left to future work. Prop. 1 applies directly when $O = C$, so the source vector in the proof is $(\mathbf{o}_C, \mathbf{z})$. When observed-but-unselected sources $(\mathbf{o}_n)$ also enter the mixing, applying the same proof to $\mathbf{z}$ alone would omit density terms such as $p(\mathbf{o}_n \mid \mathbf{z}, \mathbf{o}_C)$. One conservative extension is to enlarge the unconditioned source vector to $\mathbf{y} = (\mathbf{z}, \mathbf{o}_n)$ and require a conditional block factorization of $p(\mathbf{y} \mid \mathbf{o}_C)$; however, the resulting blocks may be coarser than the desired fine-grained partition of $\mathbf{z}$, especially when $\mathbf{o}_n$ couples multiple unobserved source blocks. In the present framework, the graphical constraint is therefore used to model these dependencies and guide finer recovery, while a population-level theorem for this setting is left open. Furthermore, the identifiability guarantees rely on the assumption that source variables can be partitioned into conditionally independent clusters given auxiliary variables. When this assumption fails—such as when all sources form a single dependent cluster—identifiability is no longer guaranteed. The strengthened sufficient-variability condition can be diagnosed only heuristically in practice, for example by rank-testing the full block score and block-Hessian vectors appearing in the assumption across selected values of $\mathbf{o}_C$. Because the required number of rows grows as $M = q + \sum_j m_j(m_j+1)/2$, persistent rank deficiency, especially when $\mathbf{o}_C$ is low-dimensional or weakly informative, indicates a violation. We view such residual entanglement and edge cases as important practical considerations for optimization and future model design. Characterizing graphical conditions that permit non-trivial recovery in such edge cases, including theorem-level uses of the graphical constraint for observed-but-unselected sources, remains an important direction for extending the applicability of the framework.

Connection to Partial Observability. Our setting can be viewed as a structured partial-observability problem. Writing each setting as (observed, unobserved), a standard POMDP-style setting is $(\mathbf{x}, \mathbf{z})$, where the latent state $\mathbf{z}$ is unobserved and the relation between $\mathbf{x}$ and $\mathbf{z}$ is otherwise unspecified. With external auxiliaries, the split is $((\mathbf{x}, \mathbf{u}), \mathbf{z})$, where $\mathbf{u}$ is outside the latent generative chain and does not enter g. In our internal-auxiliary setting, the split is $((\mathbf{x}, \mathbf{o}), \mathbf{z})$: $\mathbf{o}$ is both a pre-mixing source coordinate and an input to g, while $\mathbf{z}$ remains unobserved and must be recovered using assumptions on $\mathbf{o}$, the known graph, and VP mixing. Thus the contribution is not generic partial observability, but identifiability under the additional structure that the observable state components are internal to the mixing process.

Connection to Concept-based Methods. A recent approach, called concept-based representation learning [Rajendran et al., 2024], aims to recover high-level human-interpretable concepts, instead of true generative factors. We consider extending our framework and results to concept-based representation learning as a promising future direction. Another closely related line of work is concept bottleneck models [Koh et al., 2020, Yuksekgonul et al., 2023, Jeon et al., 2024, Oikarinen et al., 2023, Hwang et al., 2025], where the goal is to build inherently interpretable models that make predictions based on such high-level concepts. Viewing partially observable sources as *concept-like* labels, we believe that incorporating our framework into concept bottleneck models is another promising avenue for future research.