General Identifiability and Achievability for Causal Representation Learning

Burak Varıcı[1] **Emre Acartürk**[1] **Karthikeyan Shanmugam**[2] **Ali Tajer**[1]

[1]Rensselaer Polytechnic Institute [2]Google Research India

Abstract

This paper focuses on causal representation learning (CRL) under a general nonparametric latent causal model and a general transformation model that maps the latent data to the observational data. It establishes **identifiability** and **achievability** results using two hard **uncoupled** interventions per node in the latent causal graph. Notably, one does not know which pair of intervention environments have the same node intervened (hence, uncoupled). For identifiability, the paper establishes that perfect recovery of the latent causal model and variables is guaranteed under uncoupled interventions. For achievability, an algorithm is designed that uses observational and interventional data and recovers the latent causal model and variables with provable guarantees. This algorithm leverages score variations across different environments to estimate the inverse of the transformer and, subsequently, the latent variables. The analysis, additionally, recovers the identifiability result for two hard **coupled** interventions, that is when metadata about the pair of environments that have the same node intervened is known. This paper also shows that when observational data is available, additional faithfulness assumptions that are adopted by the existing literature are unnecessary.

1 INTRODUCTION

Consider a causal graph $\mathcal{G}$ with n nodes generating *causal* random variables $Z \triangleq [Z_1, \dots, Z_n]^\top$. These random variables are transformed by a function $g : \mathbb{R}^n \to$ $\mathbb{R}^d$ to generate the d-dimensional *observed* random variables $X \triangleq [X_1, \dots, X_d]^\top$ according to:

$$X = g(Z) \, , \qquad (1)$$

where $\operatorname{im}(g) \triangleq \mathcal{X} \subseteq \mathbb{R}^d$. Causal representation learning (CRL) is the process of using the data X and recovering (i) **the causal structure** $\mathcal{G}$ and (ii) **the latent causal variables** Z. When interventions are viable, the process is referred to as CRL from interventions. Achieving these implicitly involves another objective of recovering the unknown transformation g as well. Addressing CRL consists of two central questions:

- **Identifiability,** which refers to determining the necessary and sufficient conditions under which $\mathcal{G}$ and Z can be recovered. Literature on CRL from interventions commonly assumes that interventional and observational distributions are sufficiently different and inherit faithfulness assumption from causal discovery literature. Note that identifiability can be non-constructive without specifying how to recover $\mathcal{G}$ and Z.

- **Achievability**, which complements identifiability and pertains to designing algorithms that can recover $\mathcal{G}$ and Z while maintaining identifiability guarantees. Achievability hinges on forming reliable estimates for the transformation g.

This paper provides both identifiability and achievability results for CRL under stochastic *hard* interventions when (i) the transformation g can be any function (linear or non-linear) that is a diffeomorphism (i.e., bijective such that both g and g^{-1} are continuously differentiable) onto its image, and (ii) the causal relationships among elements of Z take any arbitrary form (linear or non-linear). Specifically, our main contributions are:

- On identifiability, we show that two *uncoupled* hard interventions per node suffice to guarantee perfect nonparametric identifiability (up to permutation and

Table 1: Comparison of the results to prior studies in different settings. Only the main results from the papers that aim *both* DAG and latent recovery are listed. See Section 2 for exact definitions of perfect DAG and latent recovery. Additional assumptions (*:interventional discrepancy and **: faithfulness) are discussed in Section 3.

Work	Transform	Latent Model	Obs. Data	Interv. Env. (per node)	DAG recovery	Latent recovery
(Squires et al., 2023)	Linear	Lin. Gaussian	Yes	1 Soft	impossibility	impossibility
	Linear	Lin. Gaussian	Yes	1 Hard	Yes	Yes
(Ahuja et al., 2023)	Polynomial	General	Yes	1 *do*	Yes	Yes
	Polynomial	Bounded RV	Yes	1 Soft	Yes	Yes
(Varıcı et al., 2023)	Linear	Non-linear	Yes	1 Soft	Yes	Mixing
	Linear	Non-linear	Yes	1 Hard	Yes	Yes
(Buchholz et al., 2023)	General	Lin. Gaussian	Yes	1 Hard	Yes	Yes
(Zhang et al., 2023)	Polynomial	Non-linear	Yes	1 Soft	Yes	Yes
(von Kügelgen et al., 2023)	General	General	No	2 coupled Hard*	Yes**	Yes
Theorem 1	General	General	Yes	2 uncoupled Hard*	Yes	Yes
Theorem 2	General	General	Yes	2 coupled Hard*	Yes	Yes
Theorem 3	General	General	No	2 coupled Hard*	Yes**	Yes

element-wise transforms). Specifically, we assume the learner does not know which pair of environments intervene on the same node, hence, uncoupled.

- On achievability, we design the first provably correct algorithm that recovers $\mathcal{G}$ and Z under a general transformation and a causal model. This algorithm leverages the variations of the score functions under interventions, and is referred to as **G**eneralized **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (GSCALE-I).

- While establishing identifiability results, we show that faithfulness assumptions are not required when observational data is available in contrast to recent results in the literature that require faithfulness assumptions.

1.1 Background

Learning latent representations from high-dimensional observational data is an important task in machine learning which can significantly improve the generalization and robustness of the models (Bengio et al., 2013). To that end, disentangled representation learning aims to infer a latent representation such that each latent variable corresponds to a meaningful component and the latent variables are statistically independent, similar to independent component analysis (ICA). In a wide range of domains, however, the latent variables are not independent and instead are related causally.

Causal representation learning aims to build realistic models with a causal understanding of the world (Schölkopf et al., 2021). In CRL, the generation of each latent variable is governed by a causal mechanism, i.e., a conditional probability kernel. *Identifiability* is known to be impossible without additional supervision or sufficient statistical diversity among the samples of the observed data X. As shown in (Hyvärinen and Pa-

junen, 1999; Locatello et al., 2019), this is the case even for the simpler settings in which the latent variables are statistically independent (i.e., graph $\mathcal{G}$ has no edges). Hence, identifiability guarantees critically hinge on the additional information on the data generation process and the observed data. A central problem in the literature is investigating the identifiability of CRL when performed in conjunction with *interventions* (Schölkopf et al., 2021). Specifically, leveraging the modularity property of causal models (Pearl, 2009), upon an intervention, *only* the manipulated causal mechanisms related to intervention targets change. This creates proper variations in the observed data, which facilitates identifiability.

1.2 Related Work

In this paper, we address identifiability and achievability in CRL given different interventional environments in which the interventions act on the latent space. The recent studies have been almost entirely focused on the identifiability results. We provide an overview of these existing results and discuss the closely related literature that investigates CRL from interventional data, with the main results summarized in Table 1. We defer the discussion on other approaches to CRL with stronger supervision signals, e.g., counterfactual image pairs and temporal data, to Appendix A.

The recent studies most closely related to the scope of this paper are (Varıcı et al., 2023) and (von Kügelgen et al., 2023). Varıcı et al. (2023) establish an inherent connection between score functions and CRL and leverages that connection to design a score-based CRL framework. Specifically, using *one intervention per node* under a non-linear causal model and a linear transformation, (Varıcı et al., 2023) provides both identifiability and achievability results. We have three major distinctions from (Varıcı et al., 2023) in settings

by assuming nonparametric choices of transformations g, a general latent causal model, and using two hard interventions.

von Kügelgen et al. (2023) consider nonparametric models for g and the causal relationships and shows that two *coupled* hard interventions per node suffice for identifiability. We have two major differences. First, we assume uncoupled interventional environments, whereas (von Kügelgen et al., 2023) focuses on coupled environments. Secondly, the approach in (von Kügelgen et al., 2023) focuses mainly on identifiability (e.g., no algorithm for recovery of the latent variables), whereas we address both identifiability and achievability.

Other related studies that focus on the *parametric* settings include (Squires et al., 2023; Ahuja et al., 2023; Zhang et al., 2023; Buchholz et al., 2023). Specifically, Squires et al. (2023) consider linear causal models and proves identifiability under hard interventions and the impossibility of identifiability under soft interventions. Ahuja et al. (2023) consider a polynomial transform and shows that it can be reduced to an affine transform by an autoencoding process and proves identifiability under *do* interventions or soft interventions on bounded latent variables. Zhang et al. (2023) consider a polynomial transform under non-linear causal models and prove identifiability under soft interventions. Finally, Buchholz et al. (2023) focus on linear Gaussian causal models and extend the results in (Squires et al., 2023) to prove identifiability for general transforms. For *nonparametric* settings, Jiang and Aragam (2023) focus on identifying the latent DAG without recovering latent variables, and show that a restricted class of DAGs can be recovered. Liang et al. (2023) assume that the latent DAG is already known and recover the latent variables under hard interventions.

2 PROBLEM SETTING

Notations. For a vector $a \in \mathbb{R}^m$, the i-the entry is denoted by a_i. Matrices are denoted by bold uppercase letters, e.g., $\mathbf{A}$, where $\mathbf{A}_i$ denotes the i-th row of $\mathbf{A}$ and $\mathbf{A}_{i,j}$ denotes the entry at row i and column j. We denote the indicator function by $\mathbb{1}$, and for a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, we use the convention that $\mathbb{1}\{\mathbf{A}\} \in \{0,1\}^{m \times n}$, where the entries are specified by $[\mathbb{1}\{\mathbf{A}\}]_{i,j} = \mathbb{1}\{\mathbf{A}_{i,j} \neq 0\}$. For a positive integer $n \in \mathbb{N}$, we define $[n] \triangleq \{1, \ldots, n\}$. The permutation matrix associated with any permutation π of $[n]$ is denoted by $\mathbf{P}_\pi$, i.e., $[\pi_1 \ \pi_2 \ \ldots \ \pi_n]^\top = \mathbf{P}_\pi \cdot [1 \ 2 \ \ldots \ n]^\top$. The n-dimensional identity matrix is denoted by $\mathbf{I}_{n \times n}$, and the Hadamard product is denoted by $\odot$. Given a function $f : \mathbb{R}^r \to \mathbb{R}^s$ that has first-order partial derivatives on $\mathbb{R}^r$, we denote the Jacobian of f at $z \in \mathbb{R}^s$ by $J_f(z)$. We use $\text{im}(f)$ to denote the image of f.

Latent causal structure. Consider latent causal random variables $Z \triangleq [Z_1, \ldots, Z_n]^\top$. An *unknown* transformation $g : \mathbb{R}^n \to \mathbb{R}^d$ generates the observable random variables $X \triangleq [X_1, \ldots, X_d]^\top$ from the latent variables Z according to:

$$X = g(Z) \ . \tag{2}$$

We assume that $d \geq n$, and transformation g is continuously differentiable and a diffeomorphism onto its image (otherwise, identifiability is ill-posed). We denote the image of g by $\mathcal{X} \triangleq \text{im}(g) \subseteq \mathbb{R}^d$. The probability density functions (pdfs) of Z and X are denoted by p and p_X, respectively. We assume that p is absolutely continuous with respect to the n-dimensional Lebesgue measure. Subsequently, p_X, which is defined on the image manifold $\text{im}(g)$, is absolutely continuous with respect to the n-dimensional Hausdorff measure rather than d-dimensional Lebesgue measure. The distribution of latent variables Z factorizes with respect to a DAG that consists of n nodes and is denoted by $\mathcal{G}$. Node $i \in [n]$ of $\mathcal{G}$ represents Z_i and p factorizes according to:

$$p(z) = \prod_{i=1}^{n} p_i(z_i \mid z_{\text{pa}(i)}) \ , \tag{3}$$

where $\text{pa}(i)$ denotes the set of parents of node i and $p_i(z_i \mid z_{\text{pa}(i)})$ is the conditional pdf of z_i given the variables of its parents. We also define $\overline{\text{pa}}(i) \triangleq \text{pa}(i) \cup \{i\}$, and use $\text{ch}(i)$ to denote the children of node i. Based on the modularity property, a change in the causal mechanism of node i does not affect those of the other nodes. We also assume that all conditional pdfs $\{p_i(z_i \mid z_{\text{pa}(i)}) : i \in [n]\}$ are continuously differentiable with respect to all z variables and $p(z) \neq 0$ for all $z \in \mathbb{R}^n$.

Score functions. The *score function* associated with a pdf is defined as the gradient of its logarithm. The score function associated with p is denoted by

$$s(z) \triangleq \nabla_z \log p(z) \ . \tag{4}$$

Noting the connection $X = g(Z)$, the density of X under $\mathcal{E}^0$, denoted by p_X, is supported on an n-dimensional manifold $\mathcal{X}$ *embedded* in $\mathbb{R}^d$. Hence, specifying the score function of X requires notions from differential geometry. For this purpose, we denote the tangent space of manifold $\mathcal{X}$ at point $x \in \mathcal{X}$ by $T_x\mathcal{X}$. Furthermore, given a function $f : \mathcal{X} \to \mathbb{R}$, denote its directional derivative at point $x \in \mathcal{X}$ along a tangent vector $v \in T_x\mathcal{X}$ by $D_v f(x)$. The differential of f at point $x \in \mathcal{X}$, denoted by $\mathrm{d}f_x$, is the linear operator mapping tangent vector $v \in T_x\mathcal{X}$ to $D_v f(x)$ (Simon, 2014, p. 57), i.e.,

$$\mathrm{d}f_x : T_x\mathcal{X} \ni v \mapsto D_v f(x) \in \mathbb{R} \ . \tag{5}$$

Let $\mathbf{B} \in \mathbb{R}^{d \times n}$ be a matrix for which the columns of $\mathbf{B}$ form an orthonormal basis for $T_x\mathcal{X}$. Denote the

directional derivative of f along the i-th column of $\mathbf{B}$ by $D_i f$ for all $i \in [n]$. Then, the differential operator can be expressed by the vector

$$Df(x) \triangleq \mathbf{B} \cdot \left[D_1 f(x) \dots D_n f(x) \right]^\top \in \mathbb{R}^d \ , \quad (6)$$

such that $\mathrm{d}f_x(v) = v^\top \cdot Df(x)$ for all $x \in \mathcal{X}$ and $v \in T_x\mathcal{X}$. Note that the differential operator $\mathrm{d}f$ is a generalization of the gradient. Hence, we can generalize the definition of the score function using the differential operator by setting f to the logarithm of pdf. Therefore, the score function of X under $\mathcal{E}^0$ is specified as follows:

$$s_X(x) \triangleq D \log p_X(x) \ , \qquad \forall x \in \mathcal{X} \ . \quad (7)$$

Intervention models. For each node $i \in [n]$, besides the observational mechanism specified by $p_i(z_i \mid z_{\mathrm{pa}(i)})$, we assume that there exist two hard interventional mechanisms specified by $q_i(z_i)$ and $\tilde{q}_i(z_i)$. We assume *interventional discrepancy* (Liang et al., 2023) among the distributions.

Definition 1 (Interventional discrepancy)
Two mechanisms with pdfs $p, q : \mathbb{R} \rightarrow \mathbb{R}$ satisfy interventional discrepancy *if*

$$\frac{\partial}{\partial u} \frac{p(u)}{q(u)} \neq 0 \ , \quad \forall u \in \mathbb{R} \setminus \mathcal{T} \ , \quad (8)$$

where $\mathcal{T}$ is a null set (i.e., has Lebesgue measure zero).

We note that (Liang et al., 2023) shows that for identifiability in the single atomic hard intervention per node setting, even when the latent graph $\mathcal{G}$ is known, it is necessary to have an interventional discrepancy between observational distribution p_i and interventional distribution q_i, for all $z_{\mathrm{pa}(i)} \in \mathbb{R}^{|\mathrm{pa}(i)|}$.

Interventional environments. We consider two sets of atomic interventional environments denoted by $\mathcal{E} \triangleq \{ \mathcal{E}^m : m \in [n] \}$ and $\tilde{\mathcal{E}} \triangleq \{ \tilde{\mathcal{E}}^m : m \in [n] \}$, for which a single node is intervened in each environment. We denote the node intervened in environment $\mathcal{E}^m$ by $I^m \in [n]$, and similarly denote the intervened node in $\tilde{\mathcal{E}}^m$ by $\tilde{I}^m$. We assume that node-environment pairs are *unspecified*, i.e., the ordered intervention sets $\mathcal{I} \triangleq (I^1, \dots, I^n)$ and $\tilde{\mathcal{I}} \triangleq (\tilde{I}^1, \dots, \tilde{I}^n)$ are two *unknown* permutations of $[n]$. We also adopt the convention that $I^0 = \emptyset$, and define $\mathcal{E}^0$ as the *observational* environment. Next, we define the notion of coupling between the environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$.

Definition 2 (Coupled/Uncoupled environments)
The two environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$ are said to be coupled *if for the unknown permutations $\mathcal{I}$ and $\tilde{\mathcal{I}}$ we know that $\mathcal{I} = \tilde{\mathcal{I}}$, i.e., the same node is intervened in environments $\mathcal{E}^i$ and $\tilde{\mathcal{E}}^i$. The two environment sets are said to be* uncoupled *if $\tilde{\mathcal{I}}$ is an unknown permutation of $\mathcal{I}$.*

We denote the pdfs of Z under the hard interventions in environments $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$, by p^m and $\tilde{p}^m$, respectively, which can be factorized as follows for all $m \in [n]$.

$$\mathcal{E}^m : \ p^m(z) = q_\ell(z_\ell) \prod_{i \neq \ell} p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \ \ell = I^m \ , \quad (9)$$

$$\tilde{\mathcal{E}}^m : \ \tilde{p}^m(z) = \tilde{q}_\ell(z_\ell) \prod_{i \neq \ell} p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \ \ell = I^m \ . \quad (10)$$

Hence, the score functions associated with p^m and $\tilde{p}^m$ are specified as follows.

$$s^m(z) \triangleq \nabla_z \log p^m(z) \ , \ \text{ and } \ \tilde{s}^m(z) \triangleq \nabla_z \log \tilde{p}^m(z) \ . \quad (11)$$

Statement of the objective. The objective of CRL is to use observations X and estimate the true latent variables Z and causal relations among them captured by $\mathcal{G}$. We denote a generic estimator of Z given X by $\hat{Z}(X) : \mathbb{R}^d \rightarrow \mathbb{R}^n$. We also consider a generic estimate of $\mathcal{G}$ denoted by $\hat{\mathcal{G}}$. In order to assess the fidelity of the estimates $\hat{Z}(X)$ and $\hat{\mathcal{G}}$ with respect to the ground truth Z and $\mathcal{G}$, we provide the following identifiability measures. The result in (von Kügelgen et al., 2023, Proposition 3.8) shows that these identifiability measures are the best one can ensure based on interventional data without more direct forms of supervision, e.g., counterfactual data.

Definition 3 (Perfect Identifiability) *To formalize perfect identifiability in CRL we define:*

1. **Perfect DAG recovery:** *DAG recovery is said to be perfect if $\hat{\mathcal{G}}$ is isomorphic to $\mathcal{G}$.*

2. **Perfect latent recovery:** *Given the estimator $\hat{Z}(X)$, latent recovery is said to be perfect if $\hat{Z}(X)$ is an element-wise diffeomorphism of a permutation of Z, i.e., there exists a permutation π of $[n]$ and a set of functions $\{\phi_i : i \in [n]\}$ such that $\phi_i : \mathbb{R} \rightarrow \mathbb{R}$ and we have*

$$\hat{Z}(X) = \mathbf{P}_\pi \cdot \phi(Z) \ , \quad \forall Z \in \mathbb{R}^n \ , \quad (12)$$

where $\phi(Z) \triangleq (\phi_1(Z_1), \dots, \phi_n(Z_n))$.

Recovering the latent causal variables hinges on finding the inverse of g based on the observed data X, which in turn facilitates recovering Z via $Z = g^{-1}(X)$, where g^{-1} denotes the inverse of g. Throughout the rest of this paper, we refer to g^{-1} as the *true encoder*. To formalize the procedures of estimating g^{-1}, we define $\mathcal{H}$ as the set of all possible valid encoders, i.e., candidates for g^{-1}. A function h can be such a candidate if it is invertible; that is, there exists an associated decoder h^{-1} such that $(h^{-1} \circ h)(X) = X$. Hence,

$$\mathcal{H} \triangleq \{ h : \mathcal{X} \rightarrow \mathbb{R}^n \ : \ \exists h^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^d \quad (13)$$

$$\text{such that } \forall X \in \mathcal{X} : \ (h^{-1} \circ h)(X) = X \} \ .$$

Next, corresponding to any pair of observation X and valid encoder $h \in \mathcal{H}$, we define $\hat{Z}(X; h)$ as an *auxiliary* estimate of Z generated by applying the valid encoder h on X, i.e., for all $h \in \mathcal{H}$ and $X \in \mathcal{X}$

$$\hat{Z}(X; h) \triangleq h(X) = (h \circ g)(Z) \ . \tag{14}$$

The estimate $\hat{Z}(X; h)$ inherits its randomness from X, and its statistical model is governed by that of X and the choice of h. To emphasize the dependence on h, we denote the score functions associated with the pdfs of $\hat{Z}(X; h)$ under environments $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$, respectively, by $s_{\hat{Z}}(\cdot; h)$, $s_{\hat{Z}}^m(\cdot; h)$, and $\tilde{s}_{\hat{Z}}^m(\cdot; h)$.

3 IDENTIFIABILITY AND ACHIEVABILITY RESULTS

In this section, we provide the identifiability and achievability results under different sets of assumptions and interpret them vis-á-vis the recent results in the literature. We provide constructive proof for the results by designing CRL algorithms. The details of the CRL algorithm are summarized in Algorithm 1, which is presented in Section 4. Our main result is the following theorem, which establishes perfect identifiability is possible even when the environments corresponding to the same node are not specified in pairs. That is, not only is it unknown what node is intervened in an environment, additionally the learner also does not know which two environments intervene on the same node.

Theorem 1 (Uncoupled Environments) *Using observational data and interventional data from two uncoupled environments for which each pair in $\{p_i, q_i, \tilde{q}_i\}$ satisfies interventional discrepancy, suffices to*

(i) Identifiability: perfectly recover the latent DAG;
(ii) Identifiability: perfectly recover the latent variables;
(iii) Achievability: achieve the above two guarantees (via Algorithm 1).

Theorem 1 shows that using observational data enables us to resolve any mismatch between the uncoupled environment sets and shows identifiability in the setting of uncoupled environments. This generalizes the identifiability result of (von Kügelgen et al., 2023), which requires coupled environments. Importantly, Theorem 1 does not require faithfulness whereas (von Kügelgen et al., 2023) requires that the estimated latent distribution is faithful to the associated candidate graph for all $h \in \mathcal{H}$. Even though a faithfulness assumption does not compromise the identifiability result, it is a strong requirement to verify and poses challenges to devising recovery algorithms. In contrast, we only require observational data, which is generally accessible in practice. Based on this, we design a score-based algorithm, which

is presented and discussed in Section 4. Next, if the environments are coupled, we prove identifiability under weaker assumptions on the interventional discrepancy.

Theorem 2 (Coupled Environments) *Using observational data and interventional data from two uncoupled environments for which the pair (p_i, q_i) satisfies interventional discrepancy for all $i \in [n]$, suffices to*

(i) Identifiability: perfectly recover the latent DAG;
(ii) Identifiability: perfectly recover the latent variables;
(iii) Achievability: achieve the above two guarantees (via Algorithm 1).

In the proof of Theorem 2, we show that the advantage of environment coupling is that it renders interventional data sufficient for perfect latent recovery, and the observational data is only used for recovering the graph. We further tighten this result by showing that for DAG recovery, the observational data becomes unnecessary when we have additive noise models and a weak faithfulness condition holds.

Theorem 3 (No Observational Data) *Using interventional data from two coupled environments for which the pair $(q_i, \tilde{q}_i)$ satisfies interventional discrepancy for all $i \in [n]$, suffices to*

(i) Identifiability: perfectly recover the latent DAG if the latent causal model has additive noise, $p(Z)$ is twice differentiable, and it satisfies the adjacency-faithfulness [1];
(ii) Identifiability: perfectly recover the latent variables
(iii) Achievability: achieve the above two guarantees.

Finally, we remark that the main results in this subsection (i.e., Theorem 1 and 2) hold for *any* latent causal model. The additive noise model assumption is required only for the DAG recovery part of Theorem 3.

4 GSCALE-I ALGORITHM

This section serves a two-fold purpose. First, it provides the constructive proof steps for the identifiability results specified in Theorems 1–2. Secondly, it provides achievability via designing an algorithm that has provable guarantees for perfect recovery of the latent variables and latent DAG for any *general* class of functions (linear and non-linear). We refer to this algorithm as the **G**eneralized **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (GSCALE-I) algorithm. Analysis of the steps involved are provided in Section 5.

---

[1]Adjacency-faithfulness is a weaker version of the faithfulness assumption (Ramsey et al., 2012). It requires that if nodes i and j are adjacent in $\mathcal{G}$, then Z_i and Z_j are dependent conditional on any subset of $\{Z_\ell : \ell \notin \{i, j\}\}$.

A key idea of this score-based algorithm is that the changes in the score functions of the latent variables enable us to find reliable estimates for the inverse of transformation g, which in turn facilitates estimating Z. On the other hand, we do not have access to the latent variables and can compute only the scores of the observed variables X. To circumvent this issue, we establish a relationship between the score *differences* across different pairs. For this purpose, we define s_X, s_X^m, and $\tilde{s}_X^m$ as the score function of the observed variable X under $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$, respectively. Given any valid encoder h, based on (14), the estimated latent variable $\hat{Z}(X; h)$ and X are related through $\hat{Z}(X; h) = h(X)$. We use this relationship to characterize the connection between score differences as follows, which is formalized in Lemma 2 in Section 5.

between $\mathcal{E}^0$ and $\mathcal{E}^m$: $\qquad\qquad\qquad$ (15)

$$s_{\hat{Z}}(\hat{z}; h) - s_{\hat{Z}}^m(\hat{z}; h) = [J_{h^{-1}}(\hat{z})]^\top \cdot [s_X(x) - s_X^m(x)] \ ,$$

between $\mathcal{E}^0$ and $\tilde{\mathcal{E}}^m$: $\qquad\qquad\qquad$ (16)

$$s_{\hat{Z}}(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = [J_{h^{-1}}(\hat{z})]^\top \cdot [s_X(x) - \tilde{s}_X^m(x)] \ ,$$

between $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$: $\qquad\qquad\qquad$ (17)

$$s_{\hat{Z}}^m(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = [J_{h^{-1}}(\hat{z})]^\top \cdot [s_X^m(x) - \tilde{s}_X^m(x)] \ .$$

We will show that among all valid encoders $h \in \mathcal{H}$, the true encoder g^{-1} results in the minimum number of variations between the score estimates $s_{\hat{Z}}^m(\hat{z}; h)$ and $\tilde{s}_{\hat{Z}}^m(\hat{z}; h)$ (see Lemma 3). To formalize these, corresponding to each valid encoder $h \in \mathcal{H}$ we define score change matrices $\mathbf{D}_t(h)$, $\mathbf{D}(h)$, and $\tilde{\mathbf{D}}(h)$ as follows. For all $i, m \in [n]$:

$$[\mathbf{D}_t(h)]_{i,m} \triangleq \mathbb{E}\Big[\big|[s_{\hat{Z}}^m(\hat{Z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{Z}; h)]_i\big|\Big] \ , \quad (18)$$

$$[\mathbf{D}(h)]_{i,m} \triangleq \mathbb{E}\Big[\big|[s_{\hat{Z}}(\hat{Z}; h) - s_{\hat{Z}}^m(\hat{Z}; h)]_i\big|\Big] \ , \quad (19)$$

$$[\tilde{\mathbf{D}}(h)]_{i,m} \triangleq \mathbb{E}\Big[\big|[s_{\hat{Z}}(\hat{Z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{Z}; h)]_i\big|\Big] \ , \quad (20)$$

where expectations are under the measures of latent score functions induced by the probability measure of observational data. The entry $[\mathbf{D}_t(h)]_{i,m}$ will be strictly positive only when there is a set of samples X with a strictly positive measure that renders non-identical scores $s_{\hat{Z}}^m(\hat{z}; h)$ and $\tilde{s}_{\hat{Z}}^m(\hat{z}; h)$. Similar properties hold for the entries of $\mathbf{D}(h)$ and $\tilde{\mathbf{D}}(h)$ for the respective score functions. The GSCALE-I algorithm is summarized in Algorithm 1 and its key steps are described next.

Inputs: The inputs of GSCALE-I are the observed data from the observational and interventional environments, whether environments are coupled/uncoupled, and a set of valid encoders $\mathcal{H}$.

Step 1 – Score differences: We start by computing score differences $(s_X - s_X^m)$, $(s_X - \tilde{s}_X^m)$, and $(s_X^m - \tilde{s}_X^m)$ for all $m \in [n]$.

Algorithm 1 **G**eneralized **S**core-based **Ca**usal **L**atent **E**stimation via **I**nterventions (GSCALE-I)

Input: $\mathcal{H}$, samples of X from environment $\mathcal{E}^0$ and environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$, `is_coupled`.
Output: Latent variable estimate $\hat{Z}$ and latent DAG estimate $\hat{\mathcal{G}}$.
1: **Step 1:** Compute score differences: $(s_X - s_X^m)$, $(s_X - \tilde{s}_X^m)$, and $(s_X^m - \tilde{s}_X^m)$ for all $m \in [n]$.
2: **Step 2:** Identify the encoder by minimizing score variations:
3: **if** `is_coupled` **then**
4: $\quad$ Solve $\mathcal{P}_1$ in (21), select a solution h^*.
5: **else** $\qquad\qquad$ ▷ search for the correct coupling
6: $\quad$ **for all** permutations π of $[n]$ **do**
7: $\qquad$ Temporarily relabel $\tilde{\mathcal{E}}^m$ to $\tilde{\mathcal{E}}^{\pi_m}$ for all $m \in [n]$, and solve $\mathcal{P}_2$ in (22)
8: $\qquad$ If there is a solution, select a solution h^* and break from the loop.
9: $\quad$ **end for**
10: **end if**
11: **Step 3:** Latent estimates: $\hat{Z} = h^*(X)$.
12: **Step 4:** Latent DAG recovery: Construct latent DAG $\hat{\mathcal{G}}$ using (23).
13: **return** $\hat{Z}$ and $\hat{\mathcal{G}}$.

Step 2 – Identifying the encoder: The key property in this step is that the number of variations of the estimated latent score differences is always no less than the number of variations of the true latent score differences. We have two different approaches for coupled and uncoupled settings.

Step 2 (a) – Coupled environments: We solve the following optimization problem

$$\mathcal{P}_1 \triangleq \begin{cases} \min\limits_{h \in \mathcal{H}} & \|\mathbf{D}_t(h)\|_0 \\ \text{s.t.} & \mathbf{D}_t(h) \text{ is a diagonal matrix .} \end{cases} \quad (21)$$

Constraining $\mathbf{D}_t(h)$ to be diagonal enforces that the final estimate $\hat{Z}$ and Z will be related by permutation $\mathcal{I}$ (the intervention order). We select a solution of $\mathcal{P}_1$ in (21) as our encoder estimate and denote it by h^*.

Step 2 (b) – Uncoupled environments: In this setting, additionally, we need to determine the correct coupling between the interventional environment sets $\mathcal{E}$ and $\tilde{\mathcal{E}}$. To this end, we iterate through permutations π of $[n]$, and temporarily relabel $\tilde{\mathcal{E}}^m$ to $\tilde{\mathcal{E}}^{\pi_m}$ for all $m \in [n]$ within each iteration. Subsequently, we solve the following optimization problem,

$$\mathcal{P}_2 \triangleq \begin{cases} \min\limits_{h \in \mathcal{H}} & \|\mathbf{D}_t(h)\|_0 \\ \text{s.t.} & \mathbf{D}_t(h) \text{ is a diagonal matrix} \\ & \mathbb{1}\{\mathbf{D}(h)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h)\} \\ & \mathbb{1}\{\mathbf{D}(h)\} \odot \mathbb{1}\{\mathbf{D}^\top(h)\} = \mathbf{I}_{n \times n} \ . \end{cases} \quad (22)$$

The constraint $\mathbb{1}\{\mathbf{D}(h)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h)\}$ ensures that a permutation of the correct encoder is a solution to $\mathcal{P}_2$ if the coupling is correct, and the last constraint ensures that $\mathbf{D}(h)$ does not contain 2-cycles. We will show that $\mathcal{P}_2$ is always feasible and, more specifically, admits a solution if and only if π is the correct coupling (see Lemma 5), in which case, we select a solution of $\mathcal{P}_2$ as our encoder estimate and denote it by h^*.

Step 3 – Latent estimates: The latent causal variables are estimated using h^* via $\hat{Z} = h^*(X)$, where X is the observational data.

Step 4 – Latent DAG recovery: We construct DAG $\hat{\mathcal{G}}$ from $\mathbf{D}(h^*)$ by assigning the non-zero coordinates of the i-th column of $\mathbf{D}(h^*)$ as the parents of node i in $\hat{\mathcal{G}}$, i.e.,

$$\hat{\text{pa}}(i) \triangleq \big\{ j \neq i : [\mathbf{D}(h^*)]_{j,i} \neq 0 \big\} , \quad \forall i \in [n] . \quad (23)$$

Remark 1 *For the nonparametric identifiability results, having an oracle that solves the functional optimization problems in (21) and (22), respectively, is sufficient. Solving these two problems in their most general form requires calculus of variations. These two problems, however, for any desired parameterized family of functions $\mathcal{H}$ (e.g., linear, polynomial, and neural networks), reduce to parametric optimization problems.*

5 PROPERTIES OF GSCALE-I

In this section, we analyze the properties and steps of GSCALE-I algorithm. The proofs of the results in this section are provided in Appendices B and C. We start by presenting the key properties of score functions that play pivotal roles throughout the analysis.

5.1 Properties of Score Functions

We investigate the variations of the latent score functions that are caused by the atomic hard interventions. The following lemma delineates the set of coordinates of the score function that are affected under interventions in all relevant cases.

Lemma 1 (Score Changes) *Consider environments $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$ with unknown intervention targets I^m and $\tilde{I}^m$.*

(i) Score functions s and s^m (or $\tilde{s}^m$) differ in their i-th coordinate if and only if node i or one of its children is intervened in $\mathcal{E}^m$ (or $\tilde{\mathcal{E}}^m$), i.e.,

$$\mathbb{E}\Big[\big|[s(Z) - s^m(Z)]_i\big|\Big] \neq 0 \iff i \in \overline{\text{pa}}(I^m) , \quad (24)$$

$$\mathbb{E}\Big[\big|[s(Z)] - \tilde{s}^m(Z)]_i\big|\Big] \neq 0 \iff i \in \overline{\text{pa}}(\tilde{I}^m) . \quad (25)$$

*(ii) **Coupled environments** $I^m = \tilde{I}^m$: In the coupled environment setting, s^m and $\tilde{s}^m$ differ in their i-th coordinate if and only if i is intervened, i.e.,*

$$\mathbb{E}\Big[\big|[s^m(Z) - \tilde{s}^m(Z)]_i\big|\Big] \neq 0 \iff i = I^m . \quad (26)$$

*(iii) **Uncoupled environments** $I^m \neq \tilde{I}^m$: Consider two interventional environments $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ with different intervention targets $I^m \neq \tilde{I}^m$. Consider additive noise models, in which*

$$Z_i = f_i(Z_{\text{pa}(i)}) + N_i , \quad (27)$$

where functions $\{f_i : i \in [n]\}$ are general functions and $\{N_i : i \in [n]\}$ account for noise terms that have pdfs with full support. Given that p is twice differentiable, the score functions s^m and $\tilde{s}^m$ differ in their i-th coordinate if and only if node i or one of its children is intervened,

$$\mathbb{E}\Big[\big|[s^m(Z) - \tilde{s}^m(Z)]_i\big|\Big] \neq 0 \iff i \in \overline{\text{pa}}(I^m, \tilde{I}^m) . \quad (28)$$

In the next lemma, we establish a transformation between the score differences across different environments for *any* injective mapping f from latent to observed space.

Lemma 2 (Score Difference Transformation)
Consider random vectors $Y_1, Y_2 \in \mathbb{R}^r$ and $W_1, W_2 \in \mathbb{R}^s$ that are related through $Y_1 = f(W_1)$ and $Y_2 = f(W_2)$ such that $r \geq s$, probability measures of W_1, W_2 are absolutely continuous with respect to the s-dimensional Lebesgue measure, and $f : \mathbb{R}^s \to \mathbb{R}^r$ is an injective and continuously differentiable function. The difference of the score functions of Y_1 and Y_2, and that of W_1 and W_2 are related as

$$s_{W_1}(w) - s_{W_2}(w) = [J_f(w)]^\top \cdot [s_{Y_1}(y) - s_{Y_2}(y)] , (29)$$

where $y = f(w)$ and $J_f(w) \in \mathbb{R}^s \to \mathbb{R}^r$ is the Jacobian of f at point $w \in \mathbb{R}^s$.

We customize Lemma 2 to two special cases. First, we consider score differences of $\hat{Z}(X; h)$ and X. By setting $f = h^{-1}$, Lemma 2 immediately specifies the score differences of $\hat{Z}(X; h)$ and X under different environment pairs in (15), (16), and (17). Next, we consider score differences of $\hat{Z}(X; h)$ and Z. Note that $\hat{Z}(X; h) = h(X) = (h \circ g)(Z)$. Hence, by defining $\phi_h = h \circ g$ and setting $f = \phi_h^{-1}$, Lemma 2 yields

between $\mathcal{E}^0$ and $\mathcal{E}^m$: $\quad (30)$

$$s_{\hat{Z}}(\hat{z}; h) - s_{\hat{Z}}^m(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} \cdot [s(z) - s^m(z)] ,$$

between $\mathcal{E}^0$ and $\tilde{\mathcal{E}}^m$: $\quad (31)$

$$s_{\hat{Z}}(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} \cdot [s(z) - \tilde{s}^m(z)] ,$$

between $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$: $\quad (32)$

$$s_{\hat{Z}}^m(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} \cdot [s^m(z) - \tilde{s}^m(z)] .$$

Equipped with these results, we analyze the main algorithm steps.

5.2 Analysis of Algorithm Steps

The key idea for identifying the true encoder is that the number of variances between the score estimates $s_{\hat{Z}}^m(\hat{z}; h)$ and $\tilde{s}_{\hat{Z}}^m(\hat{z}; h)$ is minimized under the true encoder g^{-1}. To show that, first, we define the *true* score change matrix $\mathbf{D}_{\mathrm{t}}$ with entries for all $i, m \in [n]$,

$$[\mathbf{D}_{\mathrm{t}}]_{i,m} \triangleq \mathbb{E}\Big[\big|[s^m(Z) - \tilde{s}^m(Z)]_i\big|\Big] . \qquad (33)$$

We start by considering coupled environments. Since the only varying causal mechanism across $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ is the intervened node $I^m = \tilde{I}^m$, based on (25),

$$\mathbb{E}\Big[\big|[s^m(Z) - \tilde{s}^m(Z)]_i\big|\Big] \neq 0 \iff i = I^m , \qquad (34)$$

which implies that $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}\}$ is a permutation matrix. Specifically, $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}\} = \mathbf{P}_{\mathcal{I}}^\top$. We show that the number of variations between the score estimates $s_{\hat{Z}}^m(\hat{z}; h)$ and $\tilde{s}_{\hat{Z}}^m(\hat{z}; h)$, i.e., ℓ_0 norm of $\mathbf{D}_{\mathrm{t}}(h)$, cannot be less than the number of variations under the true encoder g^{-1}, that is $n = \|\mathbf{D}_{\mathrm{t}}\|_0$.

Lemma 3 (Score Change Matrix Density) *For every $h \in \mathcal{H}$, the score change matrix $\mathbf{D}_{\mathrm{t}}(h)$ is at least as dense as the score change matrix $\mathbf{D}_{\mathrm{t}}$ associated with the true latent variables,*

$$\|\mathbf{D}_{\mathrm{t}}(h)\|_0 \geq \|\mathbf{D}_{\mathrm{t}}\|_0 = n . \qquad (35)$$

The rest of the proof of Theorem 2 builds on Lemma 3 and shows that, for any solution h^* to (21), we have $\mathbb{1}\{J_{\phi_{h^*}}^{-1}(z)\} = \mathbf{P}_{\mathcal{I}}^\top$ for all $z \in \mathbb{R}^n$. Finally, using Lemma 1(i) we show that DAG $\hat{\mathcal{G}}$ constructed using $\mathbf{D}(h^*)$ is isomorphic to the true latent DAG $\mathcal{G}$. We defer the complete proof to Appendix C.1.

Next, we consider the uncoupled environments. The proof consists of showing two properties of the optimization problem $\mathcal{P}_2$ specified in (22): (i) it does not have a feasible solution if the coupling is incorrect, and (ii) it has a feasible solution if the coupling is correct, which are given by following Lemma 4 and 5, respectively.

Lemma 4 (Feasibility) *If the coupling is incorrect, i.e., $\pi \neq \mathcal{I}$, the optimization problem in $\mathcal{P}_2$ in (22) does not have a feasible solution.*

The main intuition in the proof of Lemma 4 is that the constraints of $\mathcal{P}_2$ cannot be satisfied simultaneously under an incorrect coupling. We prove it by contradiction. We assume that h^* is a solution, hence, $\mathbf{D}_{\mathrm{t}}(h^*)$ is diagonal and $\mathbb{1}\{\mathbf{D}(h)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h)\}$. Then, by

scrutinizing the *eldest* mismatched node, we show that $\mathbf{D}(h^*) \cdot \mathbf{D}^\top(h^*)$ cannot be a diagonal matrix, which contradicts the premise that h^* is a feasible solution.

Lemma 5 *If the coupling is correct, i.e., $\pi = \mathcal{I}$, $h = \pi^{-1} \circ g^{-1}$ is a solution to $\mathcal{P}_2$ in (22), and yields $\|\mathbf{D}_{\mathrm{t}}(h)\|_0 = n$.*

Lemmas 4 and 5 collectively prove Theorem 1 identifiability as follows. We can search over the permutations of $[n]$ until $\mathcal{P}_2$ admits a solution h^*. By Lemma 4, the existence of this solution means that coupling is correct. Note that when the coupling is correct, the constraint set of $\mathcal{P}_1$ is a subset of the constraints in $\mathcal{P}_2$. Furthermore, the minimum value of $\|\mathbf{D}_{\mathrm{t}}(h)\|_0$ is lower bounded by n (Lemma 3), which is achieved by the solution h^* (Lemma 5). Hence, h^* is also a solution to $\mathcal{P}_1$, and by Theorem 2, it satisfies perfect recovery of the latent DAG and the latent variables.

6 EMPIRICAL EVALUATIONS

We provide empirical assessments of the achievability guarantees. Specifically, we empirically evaluate the performance of the GSCALE-I algorithm for recovering the latent variables Z and the latent DAG $\mathcal{G}$ under *coupled* interventions on synthetic data by solving the optimization problem $\mathcal{P}_1$ in (21). The evaluations pursue a two-fold purpose: (i) evaluating the performance of GSCALE-I, and (ii) showcasing settings for which the existing literature does not have an achievability result (i.e., a constructive algorithm) and provide only identifiability results for them. Hence, the achievability results in this section lack counterparts in the existing literature. We focus on a non-polynomial transform g and a non-linear latent causal model.

Data generation. To generate $\mathcal{G}$ we use the Erdős-Rényi model with density 0.5 and $n \in \{5, 8\}$ nodes. For the observational causal mechanisms, we adopt an additive noise model with

$$Z_i = \sqrt{Z_{\mathrm{pa}(i)}^\top \cdot \mathbf{A}_i \cdot Z_{\mathrm{pa}(i)}} + N_i , \qquad (36)$$

where $\{\mathbf{A}_i : i \in [n]\}$ are positive-definite matrices, and the noise terms are zero-mean Gaussian variables with variances σ_i^2 sampled randomly from $\mathrm{Unif}([0.5, 1.5])$. For the two hard interventions on node i, Z_i is set to $N_{q,i} \sim \mathcal{N}(0, \sigma_{q,i}^2)$ and $N_{\tilde{q},i} \sim \mathcal{N}(0, \sigma_{\tilde{q},i}^2)$. We set $\sigma_{q,i}^2 = \sigma_i^2 + 1$ and $\sigma_{\tilde{q},i}^2 = \sigma_i^2 + 2$. We consider target dimension values $d \in \{5, 8, 25, 100\}$. For each (n, d) pair, we generate 100 latent graphs and n_{s} samples of Z per graph, where we set $n_{\mathrm{s}} = 100$ for $n = 5$ and $n_{\mathrm{s}} = 300$ for $n = 8$. As the transformation, we consider a generalized linear model,

$$X = g(Z) = \tanh(\mathbf{G} \cdot Z) , \qquad (37)$$

Table 2: GSCALE-I for a quadratic causal model with **two coupled hard** interventions.

n	d	perfect scores		noisy scores	
		$\ell(Z, \hat{Z})$	$\text{SHD}(\mathcal{G}, \hat{\mathcal{G}})$	$\ell(Z, \hat{Z})$	$\text{SHD}(\mathcal{G}, \hat{\mathcal{G}})$
5	5	0.03	0.12	1.19	5.1
5	25	0.03	0.04	1.09	4.4
5	100	0.04	0.02	0.86	5.0
8	8	0.16	1.56	0.81	11.9
8	25	0.20	1.55	0.69	10.5
8	100	0.24	1.50	0.77	11.85

in which tanh is applied element-wise, and parameter $\mathcal{G} \in \mathbb{R}^{d \times n}$ is a randomly sampled full-rank matrix.

Score functions. GSCALE-I computes the score differences $(s_X - s_X^m), (s_X - \tilde{s}_X^m)$, and $(s_X^m - \tilde{s}_X^m)$ for all $m \in [n]$ in Step 1. The design of GSCALE-I is agnostic to how Step 1 is performed, i.e., any reliable method for estimating these score differences can be adopted. On the other hand, we note that the *perfect* identifiability guarantees formalized in Theorem 2 rely on having *perfect* score differences. In our experiments, we mainly adopt a score oracle that computes the score differences in Step 1 (see Appendix D for details). Unlike identifiability, for achievability, we need score estimates, which are inevitably noisy. For this purpose, we also adopt a score estimator, sliced score matching with variance reduction (SSM-VR) due to its efficiency and accuracy for downstream tasks (Song et al., 2020).

Candidate encoder and loss function. Leveraging (37), we parameterize valid encoders as

$$\hat{Z} = h(X) = \mathbf{H} \cdot \text{arctanh}(X) . \qquad (38)$$

To use gradient descent to learn parameters $\mathbf{H}$ of h, we relax ℓ_0 norm in (21) and instead minimize the element-wise $\ell_{1,1}$ norm $\|\mathbf{D}_t(h)\|_{1,1}$ computed with empirical expectations. We also add proper regularization terms to ensure that the estimated parameter $\mathbf{H}^*$ will be full-rank.

Evaluation metrics. GSCALE-I ensures perfect latent and DAG recovery. For assessing the recovery of the latent DAG, we report structural Hamming distance (SHD) between $\hat{\mathcal{G}}$ and $\mathcal{G}$. For the recovery of latent variables, we report the normalized ℓ_2 loss, $\ell(Z, \hat{Z}) \triangleq \|Z - \hat{Z}\|_2 / \|Z\|_2$.

Observations. Table 2 shows that by using true score differences $(s_X - s_X^m), (s_X - \tilde{s}_X^m)$, and $(s_X^m - \tilde{s}_X^m)$, we can almost perfectly recover the latent variables and the latent DAG for $n = 5$ nodes. When we consider a larger graph with $n = 8$ nodes, the normalized ℓ_2 loss remains less than 0.25. We note that $\mathcal{G}$ with $n = 8$ nodes and

density 0.5 has an expected number of 14 edges. Hence, having an average SHD of approximately 1.5 edges indicates that GSCALE-I yields a high performance at recovering latent causal relationships even when the transformation estimate is reasonable but not perfect. We also observe that increasing dimension d of the observational data does not degrade the performance, confirming our analysis that GSCALE-I is agnostic to the dimension of observations. Finally, Table 2 shows that GSCALE-I's performance degrades when using noisy scores computed via SSM-VR. This degradation is due to the errors introduced by imperfect score estimates. This performance gap can be improved by the advances in the approaches to score estimation.

7 CONCLUSION

In this paper, we have established identifiability results for latent causal representations from two interventional environments per latent node without restrictions on transformation between latent and observed space, and the causal models. We addressed both *uncoupled* and *coupled* settings. Importantly, the learner does not which pair of environments share the same intervened node in the former setting, which improved upon the recent results in nonparametric CRL identifiability. Furthermore, we have shown that the faithfulness assumption that is required in previous studies in literature can be dispensed with given access to observational data. Finally, we provided an algorithm owing to the constructive proof technique based on variations in score functions under interventions and demonstrated its success at identifying both the inverse transformation and latent DAG via simulations on synthetic data.

One major direction for future work is relaxing the requirement of two atomic hard interventions per node. Partial identifiability guarantees for a non-exhaustive set of interventions can also be useful for making inferences from a reduced number of environments. Similarly, investigating the sufficient conditions for which a set of multi-target interventions guarantee identifiability is a promising research direction. Finally, comparing the experiments with perfect and noisy score estimates indicates that the practical bottleneck of our score-based framework is access to an accurate score difference estimator, which is an interesting future direction itself.

Acknowledgements

References

K. Ahuja, J. S. Hartford, and Y. Bengio. Properties from mechanisms: An equivariance perspective on identifiable representation learning. In *Proc. International Conference on Learning Representations*, virtual, April 2022a.

K. Ahuja, J. S. Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations. In *Proc. Advances in Neural Information Processing Systems*, New Orleans LA, December 2022b.

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional causal representation learning. In *Proc. International Conference on Machine Learning*, Honolulu, Hawaii, July 2023.

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

W. M. Boothby. *An introduction to differentiable manifolds and Riemannian geometry, Revised*, volume 120. Gulf Professional Publishing, 2003.

J. Brehmer, P. De Haan, P. Lippe, and T. S. Cohen. Weakly supervised causal representation learning. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

H. Hälvä and A. Hyvärinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Proc. Conference on Uncertainty in Artificial Intelligence*, virtual, August 2020.

A. Hyvärinen and H. Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Proc. International Conference on Artificial Intelligence and Statistics*, Ft. Lauderdale, FL, April 2017.

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

A. Hyvärinen, H. Sasaki, and R. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *Proc. International Conference on Artificial Intelligence and Statistics*, Naha, Japan, April 2019.

Y. Jiang and B. Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proc. International Conference on Artificial Intelligence and Statistics*, virtual, August 2020a.

I. Khemakhem, R. Monti, D. Kingma, and A. Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2020b.

B. Kivva, G. Rajendran, P. Ravikumar, and B. Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2022.

A. Kumar and B. Poole. On implicit regularization in β-VAEs. In *Proc. International Conference on Machine Learning*, virtual, July 2020.

S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proc. Conference on Causal Learning and Reasoning*, Eureka, CA, April 2022.

W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. Causal component analysis. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. Causal representation learning for instantaneous and temporal effects in interactive systems. In *Proc. International Conference on Learning Representations*, Kigali, Rwanda, May 2023.

Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. v. d. Hengel, K. Zhang, and J. Q. Shi. Identifying weight-variant latent causal models. *arXiv:2208.14153*, 2022.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. International Conference on Machine Learning*, Long Beach, CA, June 2019.

F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *Proc. International Conference on Machine Learning*, virtual, April 2020.

F. Montagna, N. Noceti, L. Rosasco, K. Zhang, and F. Locatello. Scalable causal discovery with score matching. In *Proc. Conference on Causal Learning and Reasoning*, Tübingen, Germany, April 2023.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2009.

J. Ramsey, J. Zhang, and P. L. Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv:1206.6843*, 2012.

P. Rolland, V. Cevher, M. Kleindessner, C. Russell, D. Janzing, B. Schölkopf, and F. Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *Proc. International Conference on Machine Learning*, Baltimore, MD, July 2022.

P. Sanchez, X. Liu, A. Q. O'Neil, and S. A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *International Conference on Learning Representations*, Kigali, Rwanda, May 2023.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021.

R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole. Weakly supervised disentanglement with guarantees. In *Proc. International Conference on Learning Representations*, virtual, May 2020.

L. Simon. Introduction to geometric measure theory. *Tsinghua Lectures*, 2(2):3–1, 2014.

Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proc. Uncertainty in Artificial Intelligence*, virtual, August 2020.

C. Squires, A. Seigal, S. S. Bhate, and C. Uhler. Linear causal disentanglement via interventions. In *Proc. International Conference on Machine Learning*, Honolulu, Hawaii, July 2023.

B. Varıcı, E. Acartürk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning with interventions. *arXiv:2301.08230*, 2023.

J. von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Proc. Advances in Neural Information Processing Systems*, virtual, December 2021.

J. von Kügelgen, M. Besserve, W. Liang, L. Gresele, A. Kekić, E. Bareinboim, D. M. Blei, and B. Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition*, virtual, June 2021.

W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *Proc. International Conference on Learning Representations*, virtual, April 2022.

J. Zhang, C. Squires, K. Greenewald, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Proc. Advances in Neural Information Processing Systems*, New Orleans, LA, December 2023.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes.**

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes.** See Section 5.

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **Yes.** See the supplementary material.

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. **Yes.**

   (b) Complete proofs of all theoretical results. **Yes.** See Section 5 and Appendix C.

   (c) Clear explanations of any assumptions. **Yes.**

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **Yes.** See Appendix D and supplementary material.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Yes.**

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes.**

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Not Applicable.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

    (a) Citations of the creator If your work uses existing assets. **Not Applicable.**

    (b) The license information of the assets, if applicable. **Not Applicable.**

    (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable.**

    (d) Information about consent from data providers/curators. **Not Applicable.**

    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

    (a) The full text of instructions given to participants and screenshots. **Not Applicable.**

    (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**

    (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

General Identifiability and Achievability for Causal Representation Learning: Supplementary Material

A   Related Work

Identifiable representation learning.   One of the primary goals of representation learning is to identify the underlying latent factors responsible for generating observed data. However, as discussed in Section 1, latent factors are not identifiable unless there is auxiliary information or additional structure that explains the data generation process (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Various strategies have been developed to address this issue when there is no inherent causal relationship among the latent factors. Some notable approaches include incorporating posterior regularization (Kumar and Poole, 2020), leveraging knowledge about the mechanisms governing system dynamics (Ahuja et al., 2022a), and using weak supervision along with auxiliary information (Shu et al., 2020). Furthermore, non-linear independent component analysis (ICA) leverages side information in the form of structured time series to exploit temporal information (Hyvärinen and Morioka, 2017; Hälvä and Hyvärinen, 2020), or knowledge of auxiliary variables that make latent variables conditionally independent (Khemakhem et al., 2020a,b; Hyvärinen et al., 2019). In a related context, the identifiability of deep generative models is studied without auxiliary information (Kivva et al., 2022).

Causal representation learning.   In one approach to identifiability of CRL, several studies have investigated the setting when pairs of observations are available – one before and one after a mechanism change (e.g., an intervention) for the same underlying realization of exogenous variables involved (Ahuja et al., 2022b; Locatello et al., 2020; von Kügelgen et al., 2021; Yang et al., 2021; Brehmer et al., 2022). A typical example of this setup is considering images of an object from different angles (Brehmer et al., 2022). From a causal perspective, these pairs can be regarded as counterfactual pairs. Alternatively, some studies use temporal sequences to identify causal variables in the presence of interventions (Lachapelle et al., 2022; Yao et al., 2022; Lippe et al., 2023). However, we note that this paper does not focus on time-series data. Our approach operates under a milder form of supervision, where we can observe data under different interventional distributions in the latent space while the counterfactual instances remain unobservable.

CRL from interventions.   We repeat the related work discussed in Section 1 here. The recent studies most closely related to the scope of this paper are (Varıcı et al., 2023) and (von Kügelgen et al., 2023). (Varıcı et al., 2023) establishes an inherent connection between score functions and CRL, and based on that, designs a score-based CRL framework. Specifically, using *one intervention per node* under a non-linear causal model and a linear transformation, (Varıcı et al., 2023) provides both identifiability and achievability results. It shows that finding the variations of the score functions across different intervention environments is sufficient to recover linear g and $\mathcal{G}$ that have non-linear causal structures. We have three major distinctions from (Varıcı et al., 2023) in settings by assuming nonparametric choices of transformations g, a general latent causal model, and using two hard interventions. The study in (von Kügelgen et al., 2023) considers nonparametric models for g and the causal relationships and shows that two *coupled* hard interventions per node suffice for identifiability. We have two major differences with (von Kügelgen et al., 2023). First, we assume uncoupled interventional environments, whereas (von Kügelgen et al., 2023) focuses on coupled environments. Secondly, the approach of (von Kügelgen et al., 2023) focuses mainly on identifiability (e.g., no algorithm for recovery of the latent variables), whereas we address both identifiability and achievability.

The studies that focus on the *parametric* settings include (Liu et al., 2022; Squires et al., 2023; Ahuja et al., 2023; Zhang et al., 2023; Buchholz et al., 2023). (Liu et al., 2022) aims to learn latent causal graphs and identify latent representations. However, its focus is on linear Gaussian latent models, and its extensions to even non-linear Gaussian models are viable at the expense of restricting the graph structure. (Squires et al., 2023) considers linear causal models and proves identifiability under hard interventions and the impossibility of identifiability

under soft interventions. This setting complements that of (Varıcı et al., 2023), which considers non-linear latent causal models and proves identifiability under hard and milder identifiability guarantees under soft interventions. The study in (Ahuja et al., 2023) considers a polynomial transform and shows that it can be reduced to an affine transform by an autoencoding process and proves identifiability under *do* interventions or soft interventions on bounded latent variables. (Zhang et al., 2023) builds on the results of (Ahuja et al., 2023), considers polynomial transforms under non-linear causal models, and proves identifiability under soft interventions. Finally, (Buchholz et al., 2023) focuses on linear Gaussian causal models and extends the results of (Squires et al., 2023) to prove identifiability for general transforms. Other studies on the *nonparametric* settings include (Jiang and Aragam, 2023; Liang et al., 2023). The study in (Jiang and Aragam, 2023) considers identifying the latent DAG without recovering latent variables, where it is shown that a restricted class of DAGs can be recovered. The study in (Liang et al., 2023) assumes that the latent DAG is already known and recovers the latent variables under hard interventions.

Score functions in causality. The study in (Rolland et al., 2022) uses score-matching to recover non-linear additive Gaussian noise models. The proposed method finds the topological order of causal variables but requires additional pruning to recover the full graph. (Montagna et al., 2023) focuses on the same setting, recovers the full graph from Jacobian scores, and dispenses with the computationally expensive pruning stage. (Sanchez et al., 2023) uses score-matching to learn the topological order as well while significantly improving the training procedure. All of these studies are limited to observed causal variables, whereas in our case, we have a causal model in the latent space.

B   Score Function Properties under Interventions

In this section, we provide the proofs relating to score functions. First, we provide the following facts that will be used repeatedly in the proofs.

Proposition 1 *Consider two continuous functions $f, g : \mathbb{R}^n \to \mathbb{R}$. Then, for any $\alpha > 0$,*

$$\exists z \in \mathbb{R}^n \ \ f(z) \neq g(z) \quad \Longleftrightarrow \quad \mathbb{E}\Big[\big|f(Z) - g(Z)\big|^\alpha\Big] \neq 0 \ . \tag{39}$$

Specifically, for $\alpha = 1$, we have

$$\exists z \in \mathbb{R}^n \ \ f(z) \neq g(z) \quad \Longleftrightarrow \quad \mathbb{E}\Big[\big|f(Z) - g(Z)\big|\Big] \neq 0 \ . \tag{40}$$

Proof: If there exists $z \in \mathbb{R}^n$ such that $f(z) \neq g(z)$, then $f(z) - g(z)$ is non-zero over a non-zero-measure set due to continuity. Then, $\mathbb{E}[|f(Z) - g(Z)|^\alpha] \neq 0$ since p (pdf of Z) has full support. On the other direction, if $f(z) = g(z)$ for all $z \in \mathbb{R}^n$, then $\mathbb{E}[|f(Z) - g(Z)|^\alpha] = 0$. This means that $\mathbb{E}[|f(Z) - g(Z)|^\alpha] \neq 0$ implies that there exists $z \in \mathbb{R}^n$ such that $f(z) \neq g(z)$.

B.1   Proof of Lemma 1

Case (i)   The statement directly follows from Lemma 4 of (Varıcı et al., 2023).

Case (ii) Coupled environments.   Suppose that $I^m = \tilde{I}^m = \ell$. Following (9) and (10), we have

$$s^m(z) = \nabla_z \log q_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \tag{41}$$

$$\text{and} \quad \tilde{s}^m(z) = \nabla_z \log \tilde{q}_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ . \tag{42}$$

Then, subtracting (42) from (41) and looking at i-th coordinate, we have

$$\big[s^m(z) - \tilde{s}^m(z)\big]_i = \frac{\partial \log q_\ell(z_\ell)}{\partial z_i} - \frac{\partial \log \tilde{q}_\ell(z_\ell)}{\partial z_i} \ . \tag{43}$$

If $i \neq \ell$, the right-hand side is zero and we have $\big[s^m(z) - \tilde{s}^m(z)\big]_i = 0$ for all z. On the other hand, if $i = \ell$, since $q_\ell(z_\ell)$ and $\tilde{q}_\ell(z_\ell)$ are distinct, there exists $z \in \mathbb{R}^n$ such that $q_\ell(z_\ell) \neq \tilde{q}_\ell(z_\ell)$. Subsequently, by Proposition 1, we have $\mathbb{E}\big[\big|[s^m(Z) - \tilde{s}^m(Z)]_i\big|\big] \neq 0$.

Case (iii) Uncoupled environments. Suppose that $I^m = \ell$ and $\tilde{I}^m = j$, and $\ell \neq j$. Following (9), we have

$$s^m(z) = \nabla_z \log q_\ell(z_\ell) + \nabla_z \log p_j(z_j \mid z_{\mathrm{pa}(j)}) + \sum_{k \in [n] \setminus \{\ell, j\}} \nabla_z \log p_k(z_k \mid z_{\mathrm{pa}(k)}) , \qquad (44)$$

$$\text{and} \quad \tilde{s}^m(z) = \nabla_z \log q_j(z_j) + \nabla_z \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)}) + \sum_{k \in [n] \setminus \{\ell, j\}} \nabla_z \log p_k(z_k \mid z_{\mathrm{pa}(k)}) . \qquad (45)$$

Then, subtracting (45) from (44) we have

$$s^m(z) - \tilde{s}^m(z) = \nabla_z \log q_\ell(z_\ell) + \nabla_z \log p_j(z_j \mid z_{\mathrm{pa}(j)}) - \nabla_z \log q_j(z_j) - \nabla_z \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)}) . \qquad (46)$$

Scrutinizing the i-th coordinate, we have

$$\left[s^m(z) - \tilde{s}^m(z) \right]_i = \frac{\partial \log q_\ell(z_\ell)}{\partial z_i} + \frac{\partial \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_i} - \frac{\partial \log q_j(z_j)}{\partial z_i} - \frac{\partial \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})}{\partial z_i} . \qquad (47)$$

Proof of $\mathbb{E}\left[\left| [s^m(Z) - \tilde{s}^m(Z)]_i \right| \right] \neq 0 \implies i \in \overline{\mathrm{pa}}(\ell, j)$: Suppose that $i \notin \overline{\mathrm{pa}}(\ell, j)$. Then, none of the terms in the RHS of (47) is a function of z_i. Therefore, all the terms in the RHS of (47) are zero, and we have $\left[s^m(z) - \tilde{s}^m(z) \right]_i = 0$ for all z. By Proposition 1, $\mathbb{E}\left[\left| [s^m(Z) - \tilde{s}^m(Z)]_i \right| \right] = 0$. This, equivalently, means that if $\mathbb{E}\left[\left| [s^m(Z) - \tilde{s}^m(Z)]_i \right| \right] \neq 0$, then $i \in \overline{\mathrm{pa}}(\ell, j)$.

Proof of $\mathbb{E}\left[\left| [s^m(Z) - \tilde{s}^m(Z)]_i \right| \right] \neq 0 \impliedby i \in \overline{\mathrm{pa}}(\ell, j)$: We prove it by contradiction. Assume that $\left[s^m(z) - \tilde{s}^m(z) \right]_i = 0$ for all z. Without loss of generality, let $\ell \notin \overline{\mathrm{pa}}(j)$.

If $i = \ell$. In this case, (47) is simplified to

$$0 = \left[s^m(z) - \tilde{s}^m(z) \right]_\ell = \frac{\partial \log q_\ell(z_\ell)}{\partial z_\ell} - \frac{\partial \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})}{\partial z_\ell} . \qquad (48)$$

If ℓ is a root node, i.e., $\mathrm{pa}(\ell) = \emptyset$, (48) implies that $(\log q_\ell)'(z_\ell) = (\log p_\ell)'(z_\ell)$ for all z_ℓ. Integrating, we get $p_\ell(z_\ell) = \alpha q_\ell(z_\ell)$ for some constant α. Since both p_ℓ and q_ℓ are pdfs, they both integrate to one, implying $\alpha = 1$ and $p_\ell(z_\ell) = q_\ell(z_\ell)$, which contradicts the premise that observational and interventional mechanisms are distinct. If ℓ is not a root node, consider some $k \in \mathrm{pa}(\ell)$. Then, taking the derivative of (48) with respect to z_k, we have

$$0 = \frac{\partial^2 \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})}{\partial z_\ell \partial z_k} . \qquad (49)$$

Recall the equation $Z_\ell = f_\ell(Z_{\mathrm{pa}(\ell)}) + N_\ell$ for additive noise models specified in (27). Denote the pdf of the noise term N_ℓ by p_{N_ℓ}. Then, the conditional pdf $p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})$ is given by $p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)}) = p_N(z_\ell - f_\ell(z_{\mathrm{pa}(\ell)}))$. Denoting the score function of p_{N_ℓ} by r_ℓ,

$$r_\ell(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log p_N(u) , \qquad (50)$$

we have

$$\frac{\partial \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})}{\partial z_\ell} = \frac{\partial \log p_N(z_\ell - f_\ell(z_{\mathrm{pa}(\ell)}))}{\partial z_\ell} = r_\ell(z_\ell - f_\ell(z_{\mathrm{pa}(\ell)})) . \qquad (51)$$

Substituting this into (49), we obtain

$$0 = \frac{\partial r_\ell(z_\ell - f_\ell(z_{\mathrm{pa}(\ell)}))}{\partial z_k} = -\frac{\partial f_\ell(z_{\mathrm{pa}(\ell)})}{\partial z_k} \cdot r_\ell'(z_\ell - f_\ell(z_{\mathrm{pa}(\ell)})) , \quad \forall z \in \mathbb{R}^n . \qquad (52)$$

Since k is a parent of ℓ, there exists a fixed $Z_{\mathrm{pa}(\ell)} = z_{\mathrm{pa}(\ell)}^*$ realization for which $\partial f_\ell(z_{\mathrm{pa}(\ell)}^*)/\partial z_k$ is non-zero. Otherwise, $f_\ell(z_{\mathrm{pa}(\ell)})$ would not be sensitive to z_k which is contradictory to k being a parent of ℓ. Note that Z_ℓ can vary freely after fixing $Z_{\mathrm{pa}(\ell)}$. Therefore, for (52) to hold, the derivative of $r_{p,\ell}$ must always be zero. However, the score function of a valid pdf with full support cannot be constant. Therefore, $\left[s^m(z)_i - \tilde{s}^m(z) \right]_i$ is not always zero, and we have $\mathbb{E}\left[\left| [s^m(Z) - \tilde{s}^m(Z)]_i \right| \right] \neq 0$.

If $i \neq \ell$. In this case, (47) is simplified to

$$0 = \left[s^m(z) - \tilde{s}^m(z)\right]_i = \frac{\partial \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_i} - \frac{\partial \log q_j(z_j)}{\partial z_i} - \frac{\partial \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})}{\partial z_i} \ . \tag{53}$$

We investigate case by case and reach a contradiction for each case. First, suppose that $i \notin \mathrm{pa}(\ell)$. Then, we have $i \in \overline{\mathrm{pa}}(j)$, and (53) becomes

$$0 = \left[s^m(z) - [\tilde{s}^m(z)\right]_i = \frac{\partial \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_i} - \frac{\partial \log q_j(z_j)}{\partial z_i} \ . \tag{54}$$

If $i = j$, the impossibility of (54) directly follows from the impossibility of (48). The remaining case is $i \in \mathrm{pa}(j)$. In this case, taking the derivative of the right-hand side of (54) with respect to z_j, we obtain

$$0 = \frac{\partial^2 \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_i \partial z_j} \ , \tag{55}$$

which is a realization of (49) for $i \in \mathrm{pa}(j)$ and j in place of $k \in \mathrm{pa}(\ell)$ and ℓ, which we proved to be impossible in $i = \ell$ case. Therefore, $i \notin \mathrm{pa}(\ell)$ is not viable. Finally, suppose that $i \in \mathrm{pa}(\ell)$. Then, taking the derivative of the right-hand side of (53) with respect to z_ℓ, we obtain

$$0 = \frac{\partial^2 \log p_\ell(z_\ell \mid z_{\mathrm{pa}(\ell)})}{\partial z_i \partial z_\ell} \ , \tag{56}$$

which is again a realization of (49) for $k = i$, which we proved to be impossible.

Hence, we showed that $\left[s^m(z) - \tilde{s}^m(z)\right]_i$ cannot be zero for all z values. Then, by Proposition 1 we have $\mathbb{E}\left[\left|[s^m(Z) - \tilde{s}^m(Z)]_i\right|\right] \neq 0$, and the proof is concluded.

B.2 Proof of Lemma 2

Let us recall the setting. Consider random vectors $Y_1, Y_2 \in \mathbb{R}^r$ and $W_1, W_2 \in \mathbb{R}^s$ that are related through $Y_1 = f(W_1)$ and $Y_2 = f(W_2)$ such that $r \geq s$, probability measures of W_1, W_2 are absolutely continuous with respect to the s-dimensional Lebesgue measure and $f : \mathbb{R}^s \to \mathbb{R}^r$ is an injective and continuously differentiable function.

In this setting, the realizations of W_1 and Y_1, and that of W_2 and Y_2, are related through $y = f(w)$. Since f is injective and continuously differentiable, volume element $\mathrm{d}w$ in $\mathbb{R}^s$ gets mapped to $\left|\det\left([J_f(w)]^\top \cdot J_f(w)\right)\right|^{1/2} \mathrm{d}w$ on $\mathrm{im}(f)$. Since W_1 has density p_{W_1} absolutely continuous with respect to the s-dimensional Lebesgue measure, using the area formula (Boothby, 2003), we can define a density for Y_1, denoted by p_{Y_1}, supported only on manifold $\mathcal{M} \triangleq \mathrm{im}(f)$ which is absolutely continuous with respect to the s-dimensional Hausdorff measure:

$$p_{Y_1}(y) = p_{W_1}(w) \cdot \left|\det\left([J_f(w)]^\top \cdot J_f(w)\right)\right|^{-1/2} \ , \quad \text{where} \quad y = f(w) \ . \tag{57}$$

Densities p_{Y_2} and p_{W_2} of Y_2 and W_2 are related similarly. Subsequently, score functions of $\{W_1, W_2\}$ and $\{Y_1, Y_2\}$ are specified similarly to (4) and (7), respectively. Denote the Jacobian matrix of f at point $w \in \mathbb{R}^s$ by $J_f(w)$, which is an $r \times s$ matrix with entries given by

$$\left[J_f(w)\right]_{i,j} = \frac{\partial \left[f(w)\right]_i}{\partial w_j} = \frac{\partial y_i}{\partial w_j} \ , \quad \forall i \in [r] \, , j \in [s] \ . \tag{58}$$

Next, consider a function $\phi \colon \mathcal{M} \to \mathbb{R}$. Since the domain of ϕ is a manifold, its differential, denoted by $D\phi$, is defined according to (6). By noting $y = f(w)$, we can also differentiate ϕ with respect to $w \in \mathbb{R}^s$ as (Simon, 2014, p. 57)

$$\nabla_w \phi(y) = \nabla_w (\phi \circ f)(w) = \left[J_f(w)\right]^\top \cdot D\phi(y) \ . \tag{59}$$

Next, given the identities in (57) and (59), we find the relationship between score functions of W_1 and Y_1 as follows.

$$s_{W_1}(w) = \nabla_w \log p_{W_1}(w) \tag{60}$$

$$\overset{(57)}{=} \nabla_w \log p_{Y_1}(y) + \nabla_w \log \left| \det\left([J_f(w)]^\top \cdot J_g(w)\right) \right|^{1/2} \tag{61}$$

$$\overset{(59)}{=} [J_f(w)]^\top \cdot D \log p_{Y_1}(y) + \nabla_w \log \left| \det\left([J_f(w)]^\top \cdot J_f(w)\right) \right|^{1/2} \tag{62}$$

$$= [J_f(w)]^\top \cdot s_{Y_1}(y) + \nabla_w \log \left| \det\left([J_f(w)]^\top \cdot J_f(w)\right) \right|^{1/2} . \tag{63}$$

Following the similar steps that led to (63) for W_2 and Y_2, we obtain

$$s_{W_2}(w) = [J_f(w)]^\top \cdot s_{Y_2}(y) + \nabla_w \log \left| \det\left([J_f(w)]^\top \cdot J_f(w)\right) \right|^{1/2} . \tag{64}$$

Subtracting (64) from (63), we obtain the desired result

$$s_{W_1}(w) - s_{W_2}(w) = [J_f(w)]^\top \cdot \left[s_{Y_1}(y) - s_{Y_2}(y)\right] . \tag{65}$$

Corollary 1 *Under the same setting and the assumptions as Lemma 2, we have*

$$s_{Y_1}(y) - s_{Y_2}(y) = \left[[J_f(w)]^\dagger\right]^\top \cdot \left[s_{W_1}(w) - s_{W_2}(w)\right] , \quad where \ y = f(w) . \tag{66}$$

Proof: Multiplying (65) from left with $\left[[J_f(w)]^\dagger\right]^\top$, we obtain

$$\left[[J_f(w)]^\dagger\right]^\top \cdot \left[s_{W_1}(w) - s_{W_2}(w)\right] = \left[[J_f(w)]^\dagger\right]^\top \cdot [J_f(w)]^\top \cdot \left[s_{Y_1}(y) - s_{Y_2}(y)\right] . \tag{67}$$

Note that

$$\left[[J_f(w)]^\dagger\right]^\top \cdot [J_f(w)]^\top = J_f(w) \cdot [J_f(w)]^\dagger . \tag{68}$$

Note that, by properties of the Moore-Penrose inverse, for any matrix $\mathbf{A}$, we have $\mathbf{A} \cdot \mathbf{A}^\dagger \cdot \mathbf{A} = \mathbf{A}$. This means that $\mathbf{A} \cdot \mathbf{A}^\dagger$ acts as a left identity for vectors in the column space of $\mathbf{A}$. By definition, s_{Y_1} and s_{Y_2} have values in $T_w \operatorname{im}(f)$, the tangent space of the image manifold f at point w. This space is equal to the column space of matrix $J_f(w)$. Therefore, $J_f(w) \cdot [J_f(w)]^\dagger$ acts as a left identity for $s_{Y_1}(y)$ and $s_{Y_2}(y)$, and we have

$$J_f(w) \cdot [J_f(w)]^\dagger \cdot \left[s_{Y_1}(y) - s_{Y_2}(y)\right] = s_{Y_1}(y) - s_{Y_2}(y) . \tag{69}$$

Substituting (68) and (69) into (67) completes the proof.

C   Proofs of Identifiability Results

In this section, we first prove identifiability in the coupled environments along with the observational environment case (Theorem 2). Then, we show that the result can be extended to coupled environments without observational environment (Theorem 3) and uncoupled environments (Theorem 1).

For convenience, we recall the following equations from the main paper. For each $h \in \mathcal{H}$ we define $\phi_h \triangleq h \circ g$. Then, $\hat{Z}(X; h)$ and Z are related as

$$\hat{Z}(X; h) = h(X) = (h \circ g)(Z) = \phi_h(Z) . \tag{70}$$

Then, by setting $f = \phi_h^{-1}$, Lemma 2 yields

$$\text{between } \mathcal{E}^0 \text{ and } \mathcal{E}^m : \quad s_{\hat{Z}}(\hat{z}; h) - s_{\hat{Z}}^m(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} \cdot [s(z) - s^m(z)] , \tag{71}$$

$$\text{between } \mathcal{E}^0 \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{Z}}(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} \cdot [s(z) - \tilde{s}^m(z)] , \tag{72}$$

$$\text{between } \mathcal{E}^m \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{Z}}^m(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} \cdot [s^m(z) - \tilde{s}^m(z)] . \tag{73}$$

C.1 Proof of Theorem 2

First, we investigate the perfect recovery of latent variables.

Recovering the latent variables. We recover the latent variables using only the coupled interventional environments $\{(\mathcal{E}^m, \tilde{\mathcal{E}}^m) : m \in [n]\}$. Let ρ be the permutation that maps $\{1, \ldots, n\}$ to $\mathcal{I}$, i.e., $I^{\rho_i} = i$ for all $i \in [n]$ and $\mathbf{P}_\rho$ to denote the permutation matrix that corresponds to ρ, i.e.,

$$[\mathbf{P}_\rho]_{i,m} = \begin{cases} 1 , & m = \rho_i , \\ 0 , & \text{else} . \end{cases} \tag{74}$$

Since we consider coupled atomic interventions, the only varying causal mechanism between $\mathcal{E}^{\rho_i}$ and $\tilde{\mathcal{E}}^{\rho_i}$ is that of the intervened node in $I^{\rho_i} = \tilde{I}^{\rho_i} = i$. Then, by Lemma 1(ii), we have

$$\mathbb{E}\left[\left|\left[s^m(Z) - \tilde{s}^m(Z)\right]_k\right|\right] \neq 0 \quad \Longleftrightarrow \quad k = i . \tag{75}$$

We recall the definition of true score change matrix $\mathbf{D}_t$ in (33) with entries for all $i, m \in [n]$,

$$[\mathbf{D}_t]_{i,m} \triangleq \mathbb{E}\left[\left|\left[s^m(Z) - \tilde{s}^m(Z)\right]_i\right|\right] . \tag{76}$$

Then, using (75), we have

$$\mathbb{1}\{\mathbf{D}_t\} = \mathbf{P}_\rho = \mathbf{P}_\mathcal{I}^\top . \tag{77}$$

Next, we show that the number of variations between the score estimates $s_{\hat{Z}}^m(\hat{z}; h)$ and $\tilde{s}_{\hat{Z}}^m(\hat{z}; h)$ cannot be less than the number of variations under the true encoder g^{-1}, that is $n = \|\mathbf{D}_t\|_0$. First, we provide the following linear algebraic property.

Proposition 2 *If $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a full-rank matrix, then there exists a permutation matrix $\mathbf{P}$ such that the diagonal elements of $\mathbf{P} \cdot \mathbf{A}$ are non-zero.*

Proof See Appendix C.6.

Lemma 3 (Score Change Matrix Density) *For every $h \in \mathcal{H}$, the score change matrix $\mathbf{D}_t(h)$ is at least as dense as the score change matrix $\mathbf{D}_t$ associated with the true latent variables,*

$$\|\mathbf{D}_t(h)\|_0 \geq \|\mathbf{D}_t\|_0 = n . \tag{78}$$

Proof: Recall the definition of score change matrix $\mathbf{D}_t(h)$ in (18). Using (73), we can write entries of $\mathbf{D}_t(h)$ equivalently as

$$\left[\mathbf{D}_t(h)\right]_{i,m} = \mathbb{E}\left[\left|\left[J_{\phi_h}^{-\top}(Z)\right]_i \cdot \left[s^m(Z) - \tilde{s}^m(Z)\right]\right|\right] , \quad \forall i, m \in [n] . \tag{79}$$

Since $\phi_h = h \circ g$ is a diffeomorphism, $\left[J_{\phi_h}^{-\top}(z)\right]$ is full-rank for all $(h, z) \in \mathcal{H} \times \mathbb{R}^n$. Using Proposition 2, for all (h, z), there exists a permutation $\pi(h, z)$ of $[n]$ with permutation matrix $\mathbf{P}_1(h, z)$ such that

$$\left[\mathbf{P}_1(h, z) \cdot J_{\phi_h}^{-\top}(z)\right]_{i,i} \neq 0 , \quad \forall i \in [n] , \quad \text{where} \quad \mathbf{P}_1(h, z) \triangleq \mathbf{P}_{\pi(h,z)} . \tag{80}$$

Next, recall that interventional discrepancy means that, for each $i \in [n]$, there exists a null set $\mathcal{T}_i \subset \mathbb{R}$ such that $[s^{\rho_i}(z)]_i \neq [\tilde{s}^{\rho_i}(z)]_i$ for all $z_i \in$
$mcK \setminus \mathcal{T}_i$ (regardless of the value of other coordinates of z). Then, there exists a null set $\mathcal{T} \subset \mathbb{R}^n$ such that $[s^{\rho_i}(z)]_i \neq [\tilde{s}^{\rho_i}(z)]_i$ for all $i \in [n]$ for all $z \in \mathbb{R}^n \setminus \mathcal{T}$. We denote this set $\mathbb{R}^n \setminus \mathcal{T}$ by $\mathcal{Z}$ as follows:

$$\mathcal{Z} \triangleq \{z \in \mathbb{R}^n : \left[s^{\rho_i}(z)\right]_i \neq \left[\tilde{s}^{\rho_i}(z)\right]_i , \quad \forall i \in [n]\} . \tag{81}$$

Then, for all $z \in \mathcal{Z}$, $h \in \mathcal{H}$, and $i \in [n]$, we have

$$\left[\mathbf{D}_{\mathrm{t}}(h)\right]_{\pi_i(h,z),\rho_i} = \mathbb{E}\left[\left|\left[J_{\phi_h}^{-\top}(Z)\right]_{\pi_i(h,z)} \cdot \left[s^{\rho_i}(Z) - \tilde{s}^{\rho_i}(Z)\right]\right|\right] \tag{82}$$

$$= \mathbb{E}\left[\left|\left[J_{\phi_h}^{-\top}(Z)\right]_{\pi_i(h,z),i} \cdot \left[s^{\rho_i}(Z)_i - \tilde{s}^{\rho_i}(Z)\right]_i\right|\right] , \tag{83}$$

in which $\pi_i(h,z)$ denotes the i-th element of the permutation $\pi(h,z)$. By the definition of $\pi(h,z)$, for any $z \in \mathcal{Z}$, we know that $\left[J_{\phi_h}^{-\top}(z)\right]_{\pi_i(h,z),i} \neq 0$. Furthermore, by the definition of $\mathcal{Z}$, we have $\left[s^{\rho_i}(z) - \tilde{s}^{\rho_i}(z)\right]_i \neq 0$ for $z \in \mathcal{Z}$. Then, we have $\left[\mathbf{D}_{\mathrm{t}}(h)\right]_{\pi_i(h,z),\rho_i} \neq 0$, which implies

$$\mathbb{1}\left\{\mathbf{D}_{\mathrm{t}}(h)\right\} \succcurlyeq \mathbf{P}_1^\top(h,z) \cdot \mathbf{P}_\rho , \quad \forall h \in \mathcal{H}, \ \forall z \in \mathcal{Z} . \tag{84}$$

Therefore, $\|\mathbf{D}_{\mathrm{t}}(h)\|_0 \geq \|\mathbf{P}_\rho\|_0 = n$ for any $h \in \mathcal{H}$, and the proof is concluded since we have $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}\} = \mathbf{P}_\rho$. ∎

The lower bound for ℓ_0 norm is achieved if and only if $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}(h)\} = \mathbf{P}_\rho$, which is an unknown permutation matrix. Since the only diagonal permutation matrix is $\mathbf{I}_{n\times n}$, the solution set of the constrained optimization problem in (21) is given by

$$\mathcal{H}_1 \triangleq \left\{h \in \mathcal{H} : \mathbb{1}\{\mathbf{D}_{\mathrm{t}}(h)\} = \mathbf{I}_{n\times n}\right\} . \tag{85}$$

Next, consider some fixed solution $h^* \in \mathcal{H}_1$. Due to (84), we have

$$\mathbb{1}\{\mathbf{D}_{\mathrm{t}}(h^*)\} = \mathbf{I}_{n\times n} \succcurlyeq \mathbf{P}_1^\top(h^*,z) \cdot \mathbf{P}_\rho , \tag{86}$$

which implies that we must have $\mathbf{P}_1(h^*,z) = \mathbf{P}_\rho$ for all $z \in \mathcal{Z}$. Then, $\pi_i(h^*,z) = \rho_i$ for all $i \in [n]$. We will show that for all $i \neq j$, we have

$$\left[J_{\phi_{h^*}}^{-1}(z)\right]_{j,\rho_i} = 0 , \quad \forall z \in \mathbb{R}^n \tag{87}$$

To show (87), first consider $i \neq j$, which implies $\left[\mathbf{D}_{\mathrm{t}}(h^*)\right]_{\rho_i,\rho_j} = 0$ since $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}(h^*)\} = \mathbf{I}_{n\times n}$. Then, using (79), $\mathbb{1}\{\mathbf{D}_{\mathrm{t}}(h^*)\} = \mathbf{I}_{n\times n}$ and Lemma 1(ii), we have

$$\left[\mathbf{D}_{\mathrm{t}}(h^*)\right]_{\rho_i,\rho_j} = \mathbb{E}\left[\left|\left[J_{\phi_{h^*}}^{-\top}(Z)\right]_{\rho_i} \cdot \left[s^{\rho_j}(Z) - \tilde{s}^{\rho_j}(Z)\right]\right|\right] \tag{88}$$

$$= \mathbb{E}\left[\left|\left[J_{\phi_{h^*}}^{-1}(Z)\right]_{j,\rho_i} \cdot \left[s^{\rho_j}(Z) - \tilde{s}^{\rho_j}(Z)\right]_j\right|\right] \tag{89}$$

$$= 0 . \tag{90}$$

Note that $\left[s^{\rho_j}(z) - \tilde{s}^{\rho_j}(z)\right]_j \neq 0$ for all $z \in \mathcal{Z}$. Hence, if $\left[J_{\phi_{h^*}}^{-1}(z)\right]_{j,\rho_i}$ is non-zero over a non-zero-measure set within $\mathcal{Z}$, then $\left[\mathbf{D}_{\mathrm{t}}(h^*)\right]_{\rho_i,\rho_j}$ would not be zero. Therefore, $\left[J_{\phi_{h^*}}^{-1}(z)\right]_{j,\rho_i} = 0$ on a set of measure 1. Since $J_{\phi_{h^*}}^{-1}$ is a continuous function, this implies that

$$\left[J_{\phi_{h^*}}^{-1}(z)\right]_{j,\rho_i} = 0 , \quad z \in \mathbb{R}^n . \tag{91}$$

To see this, suppose that $\left[J_{\phi_{h^*}}^{-1}(z^*)\right]_{j,\rho_i} > 0$ for some $z^* \in \mathcal{Z}$. Due to continuity, there exists an open set including z^* for which $\left[J_{\phi_{h^*}}^{-1}(z^*)\right]_{j,\rho_i} > 0$, and since open sets have non-zero measure, we reach a contradiction. Therefore, if $i \neq j$, $\left[J_{\phi_{h^*}}^{-1}(z)\right]_{j,\rho_i} = 0$ for all $z \in \mathbb{R}^n$. Since $J_{\phi_{h^*}}^{-1}(z)$ must be full-rank for all $z \in \mathbb{R}^n$, we have

$$\left[J_{\phi_{h^*}}^{-1}(z)\right]_{i,\rho_i} \neq 0 , \quad \forall z \in \mathbb{R}^n , \forall i \in [n] . \tag{92}$$

Then, for any $h^* \in \mathcal{H}_1$, $[\hat{Z}(X;h^*)]_{\rho_i} = [\phi_{h^*}(Z)]_{\rho_i}$ is a function of only Z_i, and we have

$$[\hat{Z}(X;h^*)]_{\rho_i} = \phi_{h^*}(Z_i) , \quad \forall i \in [n] , \tag{93}$$

which concludes the proof.

Recovering the latent graph   Consider the selected solution $h^* \in \mathcal{H}_1$. We construct the graph $\hat{\mathcal{G}}$ as follows. We create n nodes and assign the non-zero coordinates of ρ_j-th column of $\mathbf{D}(h^*)$ as the parents of node ρ_j in $\hat{\mathcal{G}}$, i.e.,

$$\hat{\text{pa}}(\rho_j) \triangleq \left\{ \rho_i \neq \rho_j : [\mathbf{D}(h^*)]_{\rho_i, \rho_j} \neq 0 \right\}, \quad \forall j \in [n]. \tag{94}$$

Using (19) and (71), we have

$$\hat{\text{pa}}(\rho_j) \overset{(19)}{=} \left\{ \rho_i \neq \rho_j : \mathbb{E}\left[\left| \left[s_{\hat{Z}}(\hat{Z}; h^*) - s_{\hat{Z}}^{\rho_j}(\hat{Z}; h^*) \right]_{\rho_i} \right| \right] \neq 0 \right\} \tag{95}$$

$$\overset{(71)}{=} \left\{ \rho_i \neq \rho_j : \mathbb{E}\left[\left| \left[J_{\phi_{h^*}}^{-\top}(Z) \right]_{\rho_i} \cdot \left[s(Z) - \tilde{s}^{\rho_j}(Z) \right] \right| \right] \neq 0 \right\} \tag{96}$$

$$= \left\{ \rho_i \neq \rho_j : \mathbb{E}\left[\left| \left[J_{\phi_{h^*}}^{-\top}(Z) \right]_{\rho_i, i} \cdot \left[s(Z) - \tilde{s}^{\rho_j}(Z) \right]_i \right| \right] \neq 0 \right\}. \tag{97}$$

Since $[J_{\phi_{h^*}}^{-\top}(z)]_{\rho_i, i} \neq 0$ for all $z \in \mathbb{R}^n$, we have

$$\hat{\text{pa}}(\rho_j) = \left\{ \rho_i \neq \rho_j : \mathbb{E}\left[\left| \left[s(Z) - \tilde{s}^{\rho_j}(Z) \right]_i \right| \right] \neq 0 \right\}. \tag{98}$$

By Lemma 1(i), $\mathbb{E}\left[\left| [s(Z) - \tilde{s}^{\rho_j}(Z)]_i \right| \right] \neq 0$ if and only if $i \in \overline{\text{pa}}(j)$. Therefore, (94) implies that $\rho_i \in \hat{\text{pa}}(\rho_j)$ if and only if $i \in \text{pa}(j)$, which shows that $\mathcal{G}$ and $\hat{\mathcal{G}}$ are related through a graph isomorphism by permutation ρ, which was defined as $\mathcal{I}^{-1}$.

C.2   Proof of Theorem 3

In the proof of Theorem 2, we showed that coupled hard interventions (without using observational environment) are sufficient for recovering the latent variables. Then, in this proof, we just focus on recovering the latent graph. Specifically, we will show that if p (pdf of Z) is adjacency-faithful to $\mathcal{G}$ and the latent causal model is an additive noise model, then we can recover $\mathcal{G}$ without having access to observational environment $\mathcal{E}^0$. By Lemma 1(iii), true latent score changes across $\{\mathcal{E}^{\rho_i}, \tilde{\mathcal{E}}^{\rho_j}\}$ gives us $\overline{\text{pa}}(i,j)$ for $i \neq j$. First, we use the perfect latent recovery result to show that Lemma 1(iii) also applies to estimated latent score changes. Specifically, using (73) and $\mathbb{1}\{J_{\phi_{h^*}}^{-1}\} = \mathbf{P}_\rho$, we have

$$\left[s_{\hat{Z}}^{\rho_i}(\hat{z}; h^*) - \tilde{s}_{\hat{Z}}^{\rho_j}(\hat{z}; h^*) \right]_{\rho_k} = \left[J_{\phi_{h^*}}^{-\top}(z) \right]_{\rho_k} \cdot \left[s^{\rho_i}(z) - \tilde{s}^{\rho_j}(z) \right] \tag{99}$$

$$= \left[J_{\phi_{h^*}}^{-\top}(z) \right]_{\rho_k, k} \cdot \left[s^{\rho_i}(z) - \tilde{s}^{\rho_j}(z) \right]_k. \tag{100}$$

Recall that we have found $\left[J_{\phi_{h^*}}^{-\top}(z) \right]_{\rho_k, k} \neq 0$ for all $z \in \mathbb{R}^n$ in (92). Then, we have

$$\mathbb{E}\left[\left| \left[s_{\hat{Z}}^{\rho_i}(\hat{Z}; h^*) - \tilde{s}_{\hat{Z}}^{\rho_j}(\hat{Z}; h^*) \right]_{\rho_k} \right| \right] \neq 0 \quad \Longleftrightarrow \quad \mathbb{E}\left[\left| \left[s^{\rho_i}(Z) - \tilde{s}^{\rho_j}(Z) \right]_k \right| \right] \neq 0. \tag{101}$$

Hence, by Lemma 1(iii),

$$\mathbb{E}\left[\left| \left[s_{\hat{Z}}^{\rho_i}(\hat{Z}; h^*) - \tilde{s}_{\hat{Z}}^{\rho_j}(\hat{Z}; h^*) \right]_{\rho_k} \right| \right] \neq 0 \quad \Longleftrightarrow \quad k \in \overline{\text{pa}}(i,j). \tag{102}$$

Let us denote the graph $\mathcal{G}_\rho$ that is related to $\mathcal{G}$ by permutation ρ, i.e., $i \in \text{pa}(j)$ if and only if $\rho_i \in \text{pa}_\rho(\rho_j)$ for which $\text{pa}_\rho(\rho_j)$ denotes the parents of node ρ_j in $\mathcal{G}_\rho$. Using (102), we have

$$\mathbb{E}\left[\left| \left[s_{\hat{Z}}^{\rho_i}(\hat{Z}; h^*) - \tilde{s}_{\hat{Z}}^{\rho_j}(\hat{Z}; h^*) \right]_{\rho_k} \right| \right] \neq 0 \quad \Longleftrightarrow \quad \rho_k \in \overline{\text{pa}}_\rho(\rho_i, \rho_j). \tag{103}$$

In the rest of the proof, we will show how to obtain $\{\text{pa}_\rho(i) : i \in [n]\}$ using $\{\overline{\text{pa}}_\rho(i,j) : i,j \in [n], \ i \neq j\}$. Since $\mathcal{G}_\rho$ is a graph isomorphism of $\mathcal{G}$, this problem is equivalent to obtaining $\{\text{pa}(i) : i \in [n]\}$ using $\{\overline{\text{pa}}(i,j) : i,j \in [n], \ i \neq j\}$.

Note that $\hat{Z}_i$ (which corresponds to node i in $\mathcal{G}_\rho$) is intervened in environments $\mathcal{E}^i$ and $\tilde{\mathcal{E}}^i$. We denote the set of root nodes by

$$\mathcal{K} \triangleq \{i \in [n] : \mathrm{pa}(i) = \emptyset\} , \tag{104}$$

and also define

$$\mathcal{B}_i \triangleq \bigcap_{j \neq i} \overline{\mathrm{pa}}(i,j) , \quad \forall i \in [n] , \quad \text{and} \quad \mathcal{B} \triangleq \{i : |\mathcal{B}_i| = 1\} . \tag{105}$$

Note that $\overline{\mathrm{pa}}(i) \subseteq \mathcal{B}_i$. Hence, $|\mathcal{B}_i| = 1$ implies that i is a root node. We investigate the graph recovery in three cases.

1. $|\mathcal{B}| \geq 3$: For any node $i \in [n]$, we have

$$\overline{\mathrm{pa}}(i) \subseteq \mathcal{B}_i \subseteq \bigcap_{j \in \mathcal{K} \setminus \{i\}} \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(i) \cup \Big\{ \bigcap_{j \in \mathcal{K} \setminus \{i\}} \{j\} \Big\} = \overline{\mathrm{pa}}(i) . \tag{106}$$

Note that, the last equality is due to $\cap_{j \in \mathcal{K} \setminus \{i\}} \{j\} = \emptyset$ since there are at least two root nodes excluding i. Then, $\mathcal{B}_i = \overline{\mathrm{pa}}(i)$ for all $i \in [n]$ and we are done.

2. $|\mathcal{B}| = 2$: The two nodes in $\mathcal{B}$ are root nodes. If there were at least three root nodes, we would have at least three nodes in $\mathcal{B}$. Hence, the two nodes in $\mathcal{B}$ are the only root nodes. Subsequently, every $i \notin \mathcal{B}$ is also not in $\mathcal{K}$ and we have

$$\overline{\mathrm{pa}}(i) \subseteq \mathcal{B}_i \subseteq \bigcap_{j \in \mathcal{K}} \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(i) \cup \Big\{ \bigcap_{j \in \mathcal{K}} \{j\} \Big\} = \overline{\mathrm{pa}}(i) . \tag{107}$$

Hence, $\mathcal{B}_i = \overline{\mathrm{pa}}(i)$ for every non-root node i and we already have the two root nodes in $\mathcal{B}$, which completes the graph recovery.

3. $|\mathcal{B}| \leq 1$: First, consider all (i,j) pairs such that $|\overline{\mathrm{pa}}(i,j)| = 2$. For such an (i,j) pair, at least one of the nodes is a root node, otherwise $\overline{\mathrm{pa}}(i,j)$ would contain a third node. Using these pairs, we identify all root nodes as follows. Note that a hard intervention on node i makes Z_i independent of all of its non-descendants, and all conditional independence relations are preserved under element-wise diffeomorphisms such as ϕ_{h^*}. Then, using the adjacency-faithfulness assumption, we infer that

   - if $\hat{Z}_i \perp\!\!\!\perp \hat{Z}_j$ in $\mathcal{E}^i$ and $\hat{Z}_i \perp\!\!\!\perp \hat{Z}_j$ in $\tilde{\mathcal{E}}^j$, then both i and j are root nodes.
   - if $\hat{Z}_i \not\perp\!\!\!\perp \hat{Z}_j$ in $\mathcal{E}^i$, then $i \to j$ and i is a root node.
   - if $\hat{Z}_i \not\perp\!\!\!\perp \hat{Z}_j$ in $\tilde{\mathcal{E}}^j$, then $j \to i$ and j is a root node.

   This implies that by using at most two independence tests, we can determine whether i and j nodes are root nodes. Hence, by at most n independence tests, we identify all root nodes. We also know that there are at most two root nodes. If we have two root nodes, then $\mathcal{B}_i = \overline{\mathrm{pa}}(i)$ for all non-root nodes, and the graph is recovered. If we have only one root node i, then for any $j \neq i$ we have

$$\overline{\mathrm{pa}}(j) \subseteq \mathcal{B}_j \subseteq \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(j) \cup \{i\} . \tag{108}$$

   Finally, if $\hat{Z}_j \perp\!\!\!\perp \hat{Z}_i \mid \{\hat{Z}_\ell : \ell \in \mathcal{B}_j \setminus \{i\}\}$ in $\tilde{\mathcal{E}}^j$, we have $i \notin \overline{\mathrm{pa}}(j)$ due to adjacency-faithfulness. Otherwise, we conclude that $i \in \overline{\mathrm{pa}}(j)$. Hence, an additional $(n-1)$ conditional independence tests ensure the recovery of all $\overline{\mathrm{pa}}(j)$ sets, and the graph recovery is complete.

C.3 Proof of Lemma 4

We will prove it by contradiction. Suppose that h^* is a solution to the optimization problem $\mathcal{P}_2$ specified in (22). Using the fact that $[J_{\phi_{h^*}}^{-\top}(z)]$ is full-rank for all $z \in \mathbb{R}^n$, and the score difference vector $[s^{\rho_i}(z) - \tilde{s}^{\rho_i}(z)]$ is not identically zero, (79) and Proposition 1 imply that $\mathbf{D}_t(h^*)$ does not have any zero columns. Subsequently, $\|\mathbf{D}_t(h^*)\|_0 \geq n$, and since $\mathbf{D}_t(h^*)$ is diagonal, we have $\mathbb{1}\{\mathbf{D}_t(h^*)\} = \mathbf{I}_{n \times n}$. We use $J^* \triangleq J_{\phi_{h^*}}^{-\top}$ as shorthand. If

$\rho_i = \tilde{\rho}_i$ for some $i \in [n]$, using $\mathbf{D}_{\mathrm{t}}(h^*) = \mathbf{I}_{n \times n}$ and Lemma 1(ii), for $j \neq i$, we have

$$0 = \left[\mathbf{D}_{\mathrm{t}}(h^*) \right]_{\rho_j, \rho_i} \tag{109}$$

$$= \mathbb{E}\left[\left| \left[J^*(Z) \right]_{\rho_j} \cdot \left[s^{\rho_i}(Z) - \tilde{s}^{\rho_i}(Z) \right] \right| \right] \tag{110}$$

$$= \mathbb{E}\left[\left| \left[J^*(Z) \right]_{\rho_j, i} \cdot \left[s^{\rho_i}(Z) - \tilde{s}^{\rho_i}(Z) \right]_i \right| \right] . \tag{111}$$

Recall that, by interventional discrepancy, $[s^{\rho_i}(z) - \tilde{s}^{\rho_i}(z)]_i \neq 0$ except for a null set. Then, (109) implies that we have $[J^*(z)]_{\rho_j, i} = 0$ except for a null set. Since J^* is continuous, this implies that $[J^*(z)]_{\rho_j, i} = 0$ for all $z \in \mathbb{R}^n$. Furthermore, since $J^*(z)$ is invertible for all z, none of its columns can be a zero vector. Hence, for all $z \in \mathbb{R}^n$, $[J^*(z)]_{\rho_i, i} = 0$. To summarize, if $\rho_i = \tilde{\rho}_i$, we have

$$\forall z \in \mathbb{R}^n \ \left[J^*(z) \right]_{j, i} \neq 0 \quad \Longleftrightarrow \quad j = \rho_i . \tag{112}$$

Now, consider the set of *mismatched nodes*

$$\mathcal{A} \triangleq \{ i \in [n] : \rho_i \neq \tilde{\rho}_i \} . \tag{113}$$

Let $a \in \mathcal{A}$ be a non-descendant of all the other nodes in $\mathcal{A}$. There exist nodes $b, c \in \mathcal{A}$, not necessarily distinct, such that

$$\rho_a = \tilde{\rho}_b , \quad \text{and} \quad \rho_c = \tilde{\rho}_a . \tag{114}$$

In four steps, we will show that $\mathbf{D}(h^*)_{\rho_a, \rho_c} \neq 0$ and $\mathbf{D}(h^*)_{\rho_c, \rho_a} \neq 0$, which violates the constraint $\mathbb{1}\{\mathbf{D}(h)\} \odot \mathbb{1}\{\mathbf{D}^\top(h)\} = \mathbf{I}_{n \times n}$ and will conclude the proof by contradiction. Before giving the steps, we provide the following argument which we repeatedly use in the rest of the proof. For any continuous function $f : \mathbb{R}^n \to \mathbb{R}$, we have

$$\mathbb{E}\left[\left| f(Z) \right| \right] \neq 0 \quad \Longleftrightarrow \quad \mathbb{E}\left[\left| f(Z) \cdot \left[s(Z) - s^{\rho_a}(Z) \right]_a \right| \right] \neq 0 , \tag{115}$$

$$\text{and} \quad \mathbb{E}\left[\left| f(Z) \right| \right] \neq 0 \quad \Longleftrightarrow \quad \mathbb{E}\left[\left| f(Z) \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z) \right]_a \right| \right] \neq 0 . \tag{116}$$

First, suppose that $\mathbb{E}\big[|f(Z)|\big] \neq 0$. Then, there exists an open set $\Psi \subseteq \mathbb{R}^n$ for which $f(z) \neq 0$ for all $z \in \Psi$. Due to interventional discrepancy between the pdfs $p_a(z_a)$ and $q_a(z_a)$, there exists an open set within Ψ for which $[s^{\rho_a}(Z) - s(Z)]_a \neq 0$. This implies that

$$\mathbb{E}\left[\left| f(Z) \cdot \left[s(Z) - s^{\rho_a}(Z) \right]_a \right| \right] \neq 0 . \tag{117}$$

For the other direction, suppose that $\mathbb{E}\big[|f(Z) \cdot [s^{\rho_a}(Z) - s(Z)]_a|\big] \neq 0$, which implies that there exists an open set Ψ for which both $f(z)$ and $[s^{\rho_a}(z) - s(z)]_a$ are non-zero. Then, $\mathbb{E}\big[|f(Z)|\big] \neq 0$, and we have (115). Similarly, due to $\rho_c = \tilde{\rho}_a$ and interventional discrepancy between p_a and $\tilde{q}_a$, we obtain (116).

Step 1: Show that $\mathbb{E}\left[\left| [J^*(Z)]_{\rho_a, a} \right| \right] \neq 0$. First, using (79) and Lemma 1(i), we have

$$\left[\mathbf{D}(h^*) \right]_{\rho_a, \rho_a} = \mathbb{E}\left[\left| \left[J^*(Z) \right]_{\rho_a} \cdot \left[s(Z) - s^{\rho_a}(Z) \right] \right| \right] \tag{118}$$

$$= \mathbb{E}\left[\left| \sum_{j \in \overline{\mathrm{pa}}(a)} \left[J^*(Z) \right]_{\rho_a, j} \cdot \left[s(Z) - s^{\rho_a}(Z) \right]_j \right| \right] . \tag{119}$$

Note that $\overline{\mathrm{pa}}(a) \cap \mathcal{A} = \{a\}$ since a is non-descendant of the other nodes in $\mathcal{A}$. Consider $j \in \mathrm{pa}(a)$, which implies that $j \notin \mathcal{A}$ and $\rho_j = \tilde{\rho}_j$. By (112), we have $[J^*(Z)]_{\rho_a, j} = 0$. Then, (119) becomes

$$\left[\mathbf{D}(h^*) \right]_{\rho_a, \rho_a} = \mathbb{E}\left[\left| \left[J^*(Z) \right]_{\rho_a, a} \cdot \left[s(Z) - s^{\rho_a}(Z) \right]_a \right| \right] \neq 0 , \tag{120}$$

since diagonal entries of $\mathbf{D}(h^*)$ are non-zero due to the last constraint in (22). Then, (115) implies that $\mathbb{E}\big[|[J^*(Z)]_{\rho_a, a}|\big] \neq 0$.

Step 2: Show that $\left[\tilde{\mathbf{D}}(h^*)\right]_{\rho_a,\rho_c} \neq 0.$ Next, we use $\rho_c = \tilde{\rho}_a$ and Lemma 1(i) to obtain

$$\left[\tilde{\mathbf{D}}(h^*)\right]_{\rho_a,\rho_c} = \mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_a} \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z)\right]\right|\right] \tag{121}$$

$$= \mathbb{E}\left[\left|\sum_{j \in \overline{\mathrm{pa}}(a)} \left[J^*(Z)\right]_{\rho_a,j} \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z)\right]_j\right|\right] \tag{122}$$

$$= \mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_a,a} \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z)\right]_a\right|\right]. \tag{123}$$

Using (116) and Step 1 result, we have $\left[\tilde{\mathbf{D}}(h^*)\right]_{\rho_a,\rho_c} \neq 0.$

Step 3: Show that $\mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_c,a}\right|\right] \neq 0.$ Using (79) and Lemma 1(i), we have

$$\left[\tilde{\mathbf{D}}(h^*)\right]_{\rho_c,\rho_c} = \mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_c} \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z)\right]\right|\right] \tag{124}$$

$$= \mathbb{E}\left[\left|\sum_{j \in \overline{\mathrm{pa}}(a)} \left[J^*(Z)\right]_{\rho_c,j} \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z)\right]_j\right|\right] \tag{125}$$

$$= \mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_c,a} \cdot \left[s(Z) - \tilde{s}^{\rho_c}(Z)\right]_a\right|\right]. \tag{126}$$

Since $\mathbb{1}\{\mathbf{D}(h^*)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h^*)\}$, the diagonal entry $\left[\tilde{\mathbf{D}}(h^*)\right]_{\rho_c,\rho_c}$ is non-zero. Then, using (116) we have $\mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_c,a}\right|\right] \neq 0.$

Step 4: Show that $\left[\mathbf{D}(h^*)\right]_{\rho_c,\rho_a} \neq 0.$ Next, we use $\rho_c = \tilde{\rho}_a$ and Lemma 1(i) to obtain

$$\left[\tilde{\mathbf{D}}(h^*)\right]_{\rho_c,\rho_a} = \mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_c} \cdot \left[s(Z) - s^{\rho_a}(Z)\right]\right|\right] \tag{127}$$

$$= \mathbb{E}\left[\left|\sum_{j \in \overline{\mathrm{pa}}(a)} \left[J^*(Z)\right]_{\rho_c,j} \cdot \left[s(Z) - s^{\rho_a}(Z)\right]_j\right|\right] \tag{128}$$

$$= \mathbb{E}\left[\left|\left[J^*(Z)\right]_{\rho_c,a} \cdot \left[s(Z) - s^{\rho_a}(Z)\right]_a\right|\right]. \tag{129}$$

Using (115) and Step 3 result, we have $\left[\mathbf{D}(h^*)\right]_{\rho_c,\rho_a} \neq 0.$

However, using the constraint $\mathbb{1}\{\mathbf{D}(h^*)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h^*)\}$, we have $\left[\mathbf{D}(h^*)\right]_{\rho_a,\rho_c} \neq 0$. Then, $\left[\mathbf{D}(h^*) \odot \mathbf{D}^\top(h^*)\right]_{\rho_a,\rho_c} \neq 0$, which violates the last constraint in (22). Therefore, if the coupling is incorrect, optimization problem $\mathcal{P}_2$ has no feasible solution.

C.4   Proof of Lemma 5

We consider the true encoder g^{-1} under the permutation ρ^{-1}, that is $h = \rho^{-1} \circ g^{-1}$, and show that it is a solution to $\mathcal{P}_2$ specified in (22). First, note that $\phi_h = \rho^{-1} \circ g^{-1} \circ g = \rho^{-1}$, which is just a permutation. Hence, $J_{\phi_h}^{-\top}$ becomes the permutation matrix $\mathbf{P}_\rho^\top$. Then, for all $i, m \in [n]$ we have

$$\left[\mathbf{D}_{\mathrm{t}}(h)\right]_{\rho_i,m} = \mathbb{E}\left[\left|\left[\mathbf{P}_\rho^\top\right]_{\rho_i} \cdot \left[s^m(Z) - \tilde{s}^m(Z)\right]\right|\right] = \mathbb{E}\left[\left|\left[s^m(Z) - \tilde{s}^m(Z)\right]_i\right|\right]. \tag{130}$$

Then, by Lemma 1(ii), we have $\left[\mathbf{D}_{\mathrm{t}}(h)\right]_{\rho_i,m} \neq 0$ if and only if $i = I^m$, which means $m = \rho_i$ and $\mathbf{D}_{\mathrm{t}}(h)$ is a diagonal matrix. Hence, h satisfies the first constraint. Next, consider $\mathbf{D}(h)$. For all $i, j \in [n]$, we have

$$\left[\mathbf{D}_{\mathrm{t}}(h)\right]_{\rho_i,\rho_j} = \mathbb{E}\left[\left|\left[\mathbf{P}_\rho^\top\right]_{\rho_i} \cdot \left[s(Z) - s^{\rho_j}(Z)\right]\right|\right] = \mathbb{E}\left[\left|\left[s^{\rho_j}(Z) - \tilde{s}^{\rho_j}(Z)\right]_i\right|\right]. \tag{131}$$

By Lemma 1(i), we have $[\mathbf{D}(h)]_{\rho_i,\rho_j} \neq 0$ if and only if $i = \overline{\mathrm{pa}}(j)$. Since $\rho = \tilde{\rho}$, similarly, we have $[\tilde{D}(h_\rho)]_{\rho_i,\rho_j} \neq 0$ if and only if $i \in \overline{\mathrm{pa}}(j)$. Therefore, we have $\mathbb{1}\{\mathbf{D}(h)\} = \mathbb{1}\{\tilde{\mathbf{D}}(h)\}$, $\mathbf{D}(h)$ has full diagonal and it does not have non-zero values in symmetric entries. Hence, h satisfies the second and third constraints. Therefore, h is a solution to $\mathcal{P}_2$ since it satisfies all constraints and the diagonal matrix $\mathbf{D}_\mathrm{t}(h)$ has $\|\mathbf{D}_\mathrm{t}(h)\|_0 = n$, which is the lower bound established.

C.5 Proof of Theorem 1

Recall that $\tilde{\mathcal{I}} = \{\tilde{I}^1, \ldots, \tilde{I}^n\}$ is the permutation of intervened nodes in $\tilde{\mathcal{E}}$, so coupling π considered in (22) is just equal to $\tilde{\mathcal{I}}$. Similarly to the definition of ρ for $\mathcal{I}$ in the proof of Theorem 2, let $\tilde{\rho}$ be the permutation that maps $\{1, \ldots, n\}$ to $\tilde{\mathcal{I}}$, i.e., $I^{\tilde{\rho}_i} = i$ for all $i \in [n]$. Then, $\mathbf{P}_{\tilde{\rho}}$ denotes the permutation matrix for the intervention order of the environments $\{\tilde{\mathcal{E}}^1, \ldots, \tilde{\mathcal{E}}^n\}$, i.e.,

$$\left[\mathbf{P}_{\tilde{\rho}}\right]_{i,j} = \begin{cases} 1 , & j = \tilde{\rho}_i , \\ 0 , & \text{else} . \end{cases} \tag{132}$$

Lemma 4 shows that if the coupling is incorrect, i.e., $\pi \neq \mathcal{I}$ or equivalently $\rho \neq \tilde{\rho}$, the optimization problem in (22) does not have a feasible solution. Next, Lemma 5 shows that if the coupling is correct, i.e., $\rho = \tilde{\rho}$, there exists a solution to $\mathcal{P}_2$. Lemmas 4 and 5 collectively prove identifiability as follows. We can search over the permutations of $[n]$ until $\mathcal{P}_2$ admits a solution h^*. By Lemma 4, the existence of this solution means that coupling is correct. Note that when the coupling is correct, the constraint set of $\mathcal{P}_1$ is a subset of the constraints in $\mathcal{P}_2$. Furthermore, the minimum value of $\|\mathbf{D}_\mathrm{t}(h)\|_0$ is lower bounded by n (Lemma 3), which is achieved by the solution h^* (Lemma 5). Hence, h^* is also a solution to $\mathcal{P}_1$, and by Theorem 2, it satisfies perfect recovery of the latent DAG and the latent variables.

C.6 Proof of Proposition 2

Denote the set of all permutations of $[n]$ by $\mathcal{S}_n$. From Leibniz formula for matrix determinants, for a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ we have

$$\det(\mathbf{A}) = \sum_{\pi \in \mathcal{S}_n} \mathrm{sgn}(\pi) \cdot \prod_{i=1}^n \mathbf{A}_{i,\pi_i} \tag{133}$$

where $\mathrm{sgn}(\pi)$ for a permutation π of $[n]$ is $+1$ and -1 for even and odd permutations, respectively. $\mathbf{A}$ is invertible if and only if $\det(\mathbf{A}) \neq 0$, which implies that

$$\exists \pi \in \mathcal{S}_n \; : \; \mathrm{sgn}(\pi) \cdot \prod_{i=1}^n \mathbf{A}_{i,\pi_i} \neq 0 . \tag{134}$$

By the definition of $\mathbf{P}_\pi$, we have $[\mathbf{P}_\pi \cdot \mathbf{A}]_{i,i} = \mathbf{A}_{i,\pi_i}$. Then,

$$\exists \pi \in \mathcal{S}_n \; : \; \mathrm{sgn}(\pi) \cdot \prod_{i=1}^n [\mathbf{P}_\pi \cdot \mathbf{A}]_{i,i} \neq 0 , \tag{135}$$

which holds if and only if $[\mathbf{P}_\pi \cdot \mathbf{A}]_{i,i} \neq 0$ for all $i \in [n]$.

D Details of Simulations

We perform experiments for the coupled environments setting and using a regularized, ℓ_1-relaxed version of the optimization problem (21). Specifically, in Step 2 of GSCALE-I, we solve the following optimization problem:

$$\min_{h \in \mathcal{H}} \|\mathbf{D}_\mathrm{t}(h)\|_{1,1} + \lambda_1 \mathbb{E}\|h^{-1}(h(X)) - X\|_2^2 + \lambda_2 \mathbb{E}\|h(X)\|_2^2 . \tag{136}$$

In this section, we describe data generation, computation of the ground truth score differences for X, justification of the optimization problem in (136) and other implementation details.

Data generation details. To generate $\mathcal{G}$ we use the Erdős-Rényi model with $n \in \{5, 8\}$ nodes and density 0.5. For the observational causal mechanisms, we adopt an additive noise model with

$$Z_i = \sqrt{Z_{\mathrm{pa}(i)}^\top \cdot \mathbf{A}_i \cdot Z_{\mathrm{pa}(i)}} + N_i \ , \tag{137}$$

where $\{\mathbf{A}_i : i \in [n]\}$ are positive-definite matrices, and the noise terms are zero-mean Gaussian variables with variances σ_i^2 sampled randomly from $\mathrm{Unif}([0.5, 1.5])$. We construct the positive-definite matrix $\mathbf{A}_i$ by generating a matrix $\mathbf{B}_i \in \mathbb{R}^{|\mathrm{pa}(i)| \times |\mathrm{pa}(i)|}$ by sampling its entries from $\mathrm{Unif}([0,1])$ and setting $\mathbf{A}_i = \mathbf{B}_i^\top \mathbf{B}_i$.

For two hard interventions on node i, Z_i is set to $N_{q,i} \sim \mathcal{N}(0, \sigma_{q,i}^2)$ and $N_{\tilde{q},i} \sim \mathcal{N}(0, \sigma_{\tilde{q},i}^2)$. We set $\sigma_{q,i}^2 = \sigma_i^2 + 1$ and $\sigma_{\tilde{q},i}^2 = \sigma_i^2 + 2$. We consider target dimension values $d \in \{5, 25, 40\}$ for $n = 5$ and $d \in \{8, 25, 40\}$ for $n = 8$. For each (n, d) pair, we generate 100 latent graphs, and n_{s} samples per environment per graph, where we set $n_{\mathrm{s}} = 100$ for $n = 5$ and $n_{\mathrm{s}} = 300$ for $n = 8$. As the transformation, we consider a generalized linear model,

$$X = g(Z) = \tanh(\mathbf{G} \cdot Z) \ , \tag{138}$$

in which tanh is applied element-wise, and the parameter $\mathbf{G} \in \mathbb{R}^{d \times n}$ is a randomly sampled full-rank matrix.

Score function of the quadratic model. Following (9), score functions s^m and $\tilde{s}^m$ are decomposed as follows.

$$s^m(z) = \nabla_z \log q_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \tag{139}$$

$$\text{and} \quad \tilde{s}^m(z) = \nabla_z \log \tilde{q}_\ell(z_\ell) + \sum_{i \neq \ell} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ . \tag{140}$$

For additive noise models, the terms in (139) and (140) have closed-form expressions. Specifically, using (51) and denoting the score functions of the noise terms $\{N_i : i \in [n]\}$ by $\{r_i : i \in [n]\}$, we have

$$[s(z)]_i = r_i(n_i) - \sum_{j \in \mathrm{ch}(i)} \frac{\partial f_i(z_{\mathrm{pa}(j)})}{\partial z_i} \cdot r_j(n_j) \ . \tag{141}$$

In the quadratic latent model we consider in (36), we have

$$f_i(z_{\mathrm{pa}(i)}) = \sqrt{z_{\mathrm{pa}(i)}^\top \cdot \mathbf{A}_i \cdot z_{\mathrm{pa}(i)}} \ , \quad \text{and} \quad N_i \sim \mathcal{N}(0, \sigma_i^2) \ , \tag{142}$$

which implies

$$\frac{\partial f_j(z_{\mathrm{pa}(j)})}{\partial z_i} = \frac{[\mathbf{A}_j]_i \cdot z_{\mathrm{pa}(j)}}{\sqrt{z_{\mathrm{pa}(j)}^\top \cdot \mathbf{A}_j \cdot z_{\mathrm{pa}(j)}}} \ , \quad \text{and} \quad r_i(n_i) = -\frac{n_i}{\sigma_i^2} \ , \quad \forall i \in [n] \ . \tag{143}$$

Components of the score functions s^m and $\tilde{s}^m$ can be computed similarly. Subsequently, using Corollary 1 of Lemma 2, we can compute the score differences of observed variables as follows.

$$s_X(x) - s_X^m(x) = \left[[J_g(z)]^\dagger \right]^\top \cdot \left[s(z) - s^m(z) \right] \ , \tag{144}$$

$$s_X(x) - \tilde{s}_X^m(x) = \left[[J_g(z)]^\dagger \right]^\top \cdot \left[s(z) - \tilde{s}^m(z) \right] \ , \tag{145}$$

$$s_X^m(x) - \tilde{s}_X^m(x) = \left[[J_g(z)]^\dagger \right]^\top \cdot \left[s^m(z) - \tilde{s}^m(z) \right] \ . \tag{146}$$

Implementation and evaluation steps. Leveraging (37), we parameterize valid encoders h with parameter $\mathbf{H} \in \mathbb{R}^{n \times d}$.

$$\hat{Z}(X; h) = h(X) = \mathbf{H} \cdot \mathrm{arctanh}(X) \ , \tag{147}$$

$$\hat{X} = h^{-1}(\hat{Z}(X; h)) = \tanh\left(\mathbf{H}^\dagger \cdot \hat{Z}(X; h)\right) \ , \tag{148}$$

Note that given this parameterization, the function $\phi_h(z) = (h \circ g)(z)$ is given by

$$\hat{Z} = \phi_h(Z) = \mathbf{H} \cdot \mathbf{G} \cdot Z \ . \tag{149}$$

Subsequently, the only element-wise diffeomorphism between Z and $\hat{Z}$ is an element-wise scaling. Hence, we can evaluate the estimated latent variables against the scaling consistency metric. To this end, we use normalized ℓ_2 loss specified in Section 6.

We use n_{s} samples from the observational environment to compute empirical expectations. Since encoder h is parameterized by $\mathbf{H}$, we use gradient descent to learn this matrix. To do so, we relax ℓ_0 norm in (21) and instead minimize element-wise $\ell_{1,1}$ norm $\|\mathbf{D}_{\mathrm{t}}(h)\|_{1,1}$. Note that, scaling up $\hat{Z}(h)$ by a constant factor scales down the score differences by the same factor. Hence, to prevent the vanishing of the score difference loss trivially, we add the following regularization term to the optimization objective.

$$\mathbb{E}\left[\left\|\hat{Z}(h)\right\|_2^2\right] = \mathbb{E}\left[\left\|h(X)\right\|_2^2\right] \ . \tag{150}$$

We also add the following reconstruction loss to ensure that h is an invertible transform.

$$\mathbb{E}\left[\left\|h^{-1}(h(X)) - X\right\|_2^2\right] \ . \tag{151}$$

In the end, we minimize the objective function

$$\|\mathbf{D}_{\mathrm{t}}(h)\|_{1,1} + \lambda_1 \mathbb{E}\left[\left\|h^{-1}(h(X)) - X\right\|_2^2\right] + \lambda_2 \mathbb{E}\left[\left\|h(X)\right\|_2^2\right] \ . \tag{152}$$

We denote the final parameter estimate by $\mathbf{H}^*$ and the encoder parameterized by $\mathbf{H}^*$ by h^*. Note that we do not enforce the diagonality constraint upon $\mathbf{D}_{\mathrm{t}}(h)$. Since we learn the latent variables up to permutation, we change this constraint to a post-processing step. Specifically, we permute the rows of $\mathbf{H}^*$ to make $\mathbf{D}_{\mathrm{t}}(h^*)$ as close to as diagonal, i.e., $\|\mathrm{diag}(\mathbf{D}_{\mathrm{t}}(h^*) \odot \mathbf{I}_n)\|_1$ is maximized.

In the simulation results reported in Section 6, we set $\lambda_1 = 10^{-4}$ and $\lambda_2 = 1$, and minimize (152) using RMSprop optimizer with learning rate 10^{-3} for 3×10^4 steps for $n = 5$ and 4×10^4 steps for $n = 8$. Recall that the latent graph estimate $\hat{\mathcal{G}}$ is constructed using $\mathbb{1}\{\mathbf{D}(h^*)\}$. We use a threshold $\lambda_{\mathcal{G}}$ to obtain the graph from the upper triangular part of $\mathbf{D}(h^*)$ as follows.

$$\hat{\mathrm{pa}}(i) = \{j : j < i \quad \text{and} \quad [\mathbf{D}(h^*)]_{j,i} \geq \lambda_{\mathcal{G}}\} \ , \qquad \forall i \in [n] \ . \tag{153}$$

We set $\lambda_{\mathcal{G}} = 0.1$ for $n = 5$ and $\lambda_{\mathcal{G}} = 0.2$ for $n = 8$.