Score-based Causal Representation Learning from Interventions: Nonparametric Identifiability

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

[1] Rensselaer Polytechnic Institute [2] Google Research

Abstract

This paper focuses on causal representation learning (CRL) under a general non-parametric causal latent model and a general transformation model that maps the latent data to the observational data. It establishes **identifiability** and **achievability** results using two (stochastic) 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 environments). 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 for the algorithm. 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 existing identifiability result for two hard **coupled** interventions, that is when metadata about the pair of environments that have the same node intervened is known. It is noteworthy that the existing results on non-parametric identifiability require assumptions on interventions and additional faithfulness assumptions. This paper shows that when observational data is available, additional faithfulness assumptions are unnecessary.

1 Introduction

Consider a causal graph $\mathcal{G}_Z$ 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:

$$X = g(Z) \ . \tag{1}$$

Causal representation learning (CRL) is the process of using the observed data X and recovering (i) the causal structure $\mathcal{G}_Z$ and (ii) the unknown transformation g. When interventions are viable, the process is referred to as CRL from interventions. Addressing CRL consists of two central questions:

- **Identifiability,** which refers to determining sufficient conditions under which $\mathcal{G}_Z$ and Z can be recovered. Identifiability can be non-constructive without specifying how to recover $\mathcal{G}_Z$ and Z.
- **Achievability**, which pertains to designing algorithms that can recover $\mathcal{G}_Z$ and g, while maintaining identifiability guarantees. Achievability hinges on forming reliable estimates for the function 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 element-wise transforms). Specifically,

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

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	Interv.	Obs. Data	No. of of Intervs.	DAG recovery	Latent recovery
[3]	Linear	Lin. Gaussian	Soft	Yes	1 per node	impossibility	impossibility
	Linear	Lin. Gaussian	Hard	Yes	1 per node	Yes	Yes
[4]	Polynomial	General	*do*	Yes	1 per node	Yes	Yes
	Polynomial	Bounded RV	Soft	Yes	1 per node	Yes	Yes
[1]	Linear	Non-linear	Soft	Yes	1 per node	Yes	Mixing
	Linear	Non-linear	Hard	Yes	1 per node	Yes	Yes
[5]	General	Lin. Gaussian	Hard	Yes	1 per node	Yes	Yes
[6]	Polynomial	Non-linear	Soft	Yes	1 per node	Yes	Yes
[2]	General	General	Hard*	No	2 coupled per node	Yes**	Yes
This work	General	General	Hard*	No	2 coupled per node	Yes**	Yes
	General	General	Hard*	Yes	2 coupled per node	Yes	Yes
	General	General	Hard*	Yes	2 uncoupled per node	Yes	Yes

we assume the learner does not know which pair of environments intervene on the same node, hence, uncoupled.

- On achievability, we design an algorithm that leverages variations of the score functions under interventions for recovering $\mathcal{G}_Z$ and Z under a general transformation and a causal model.
- 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.

Related work. The recent studies most closely related to the scope of this paper are [1] and [2]. [1] establishes an inherent connection between score function 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, [1] 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 $\mathcal{G}_Z$ that have non-linear causal structures. We have three major distinctions from [1] in settings by assuming nonparametric choices of transformations g, a general latent causal model, and using two hard interventions. The study in [2] 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 [2]. First, we assume uncoupled interventional environments, whereas [2] focuses on coupled environments. Secondly, the approach of [2] focuses mainly on identifiability (e.g., no algorithm for recovery of the latent variables), whereas we address both identifiability and achievability. We summarize the main results of other related studies in Table 1 and defer the details to Appendix A.

Notations. For a vector a, the i-the entry is denoted by a_i and $[a]_i$. For a matrix $A \in \mathbb{R}^{m\times n}$, the i-th row is denoted by $[A]_i$, the entry at row i and column j is denoted by $[A]_{i,j}$. I_n denotes the $n \times n$ identity matrix. For a positive integer n, we define $[n] \triangleq \{1, \dots, n\}$. We denote the Jacobian of a function $f : \mathbb{R}^n \to \mathbb{R}^n$ at point $z \in \mathbb{R}^n$ by $J_f(z)$. We denote the indicator function by $\mathbb{1}$. For a matrix $A \in \mathbb{R}^{a\times b}$ we use the convention that $\mathbb{1}\{A\} \in \{0,1\}^{a\times b}$ is defined with entries $[\mathbb{1}\{A\}]_{i,j} = \mathbb{1}\{A_{i,j} \neq 0\}$. We use $\odot$ to denote the Hadamard product.

2 Problem Setting

The data generating process. Consider latent random variables $Z \triangleq [Z_1, \dots, Z_n]^\top$. An unknown deterministic function $g : \mathbb{R}^n \to \mathbb{R}^d$ generates the observable random variables $X \triangleq [X_1, \dots, X_d]^\top$ from the latent variables according to

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

We assume that $d \geq n$, g is continuously differentiable and a diffeomorphism onto its image. Otherwise, identifiability is ill-posed. We denote the probability density function (pdf) of Z by p. For clarity in the analysis, p is assumed to be well-defined.

Latent causal structure. The distribution of latent variables Z factorizes with respect to a DAG represented by $\mathcal{G}_Z$ that consists of n nodes. Node $i \in [n]$ of $\mathcal{G}_Z$ represents Z_i and p factorizes according to

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

where $\mathrm{pa}(i)$ denotes the set of parents of node i. For each node $i \in [n]$, we also define $\overline{\mathrm{pa}}(i) \triangleq \mathrm{pa}(i) \cup \{i\}$. 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 $\{p_i(z_i \mid z_{\mathrm{pa}(i)}) : i \in [n]\}$ are continuously differentiable with respect to all z variables and $p(z) \neq 0$ for all $z \in \mathbb{R}^n$.

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* [7] among the distributions.

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

$$\frac{\partial}{\partial u} \frac{p(u)}{q(u)} \neq 0 \ , \quad \forall u \in \mathbb{R} \setminus \mathcal{T} \ , \tag{4}$$

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

We note that [7] shows that for identifiability in the single atomic hard intervention per node setting, even when the latent graph $\mathcal{G}_Z$ 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 interventional environments denoted by $\mathcal{E} \triangleq \{\mathcal{E}^m : m \in [n]\}$ and $\tilde{\mathcal{E}} \triangleq \{\tilde{\mathcal{E}}^m : m \in [n]\}$. The set of intervention targets in $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ are *unknown*, which we denote by I^m and $\tilde{I}^m$. We focus on the setting of atomic interventions in which each node $i \in [n]$ is intervened in exactly one environment in $\mathcal{E}$ and one environment in $\tilde{\mathcal{E}}$, i.e., $\mathcal{I} \triangleq (I^1, \ldots, I^n)$ and $\tilde{\mathcal{I}} \triangleq (\tilde{I}^1, \ldots, \tilde{I}^n)$ are two *unknown* permutations of $[n]$.

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

We also adopt the convention that $I^0 = \emptyset$, and define $\mathcal{E}^0$ as the *observational* environment. 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. $\forall m \in [n]$:

$$\text{under } \mathcal{E}^m : \qquad p^m(z) = \prod_{i \in I^m} q_i(z_i) \prod_{i \in [n] \setminus I^m} p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \tag{5}$$

$$\text{under } \tilde{\mathcal{E}}^m : \qquad \tilde{p}^m(z) = \prod_{i \in \tilde{I}^m} \tilde{q}_i(z_i) \prod_{i \in [n] \setminus \tilde{I}^m} p_i(z_i \mid z_{\mathrm{pa}(i)}) \ . \tag{6}$$

Score function. Define the *score function* associated with a probability distribution as the gradient of its log pdf. We denote the score functions associated with p, p^m, and $\tilde{p}^m$ by s, s^m, and $\tilde{s}^m$, respectively. Leveraging the causal structure $\mathcal{G}_Z$ and the factorizations in (3), (5), and (6), the score functions in different environments have the following decompositions.

$$\text{under } \mathcal{E}^0 : \qquad s(z) \triangleq \nabla_z \log p(z) = \sum_{i=1}^{n} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \tag{7}$$

$$\text{under } \mathcal{E}^m : \qquad s^m(z) \triangleq \nabla_z \log p^m(z) = \sum_{i \in I^m} \nabla_z \log q_i(z_i) + \sum_{i \notin I^m} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \tag{8}$$

$$\text{under } \tilde{\mathcal{E}}^m : \qquad \tilde{s}^m(z) \triangleq \nabla_z \log \tilde{p}^m(z) = \sum_{i \in \tilde{I}^m} \nabla_z \log \tilde{q}_i(z_i) + \sum_{i \notin \tilde{I}^m} \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ . \tag{9}$$

Statement of the objective. The objective is to use the observational data X and recover the true latent variables Z and causal relations among them. We define $\hat{Z}$ and $\mathcal{G}_{\hat{Z}}$ as estimates of Z and $\mathcal{G}_Z$, respectively. To assess the fidelity of the estimate $\hat{Z}$ with respect to the ground truth Z, we provide the following two measures for identifiability. The result in [2, Proposition 3.8] shows that these two identifiability measures are the best one can ensure based on interventional data without more direct forms of supervision, e.g., counterfactual data.

Definition 3 (Identifiability) *For the identifiability objectives of CRL we define two measures:*

1. **Perfect DAG recovery:** *We have* perfect DAG recovery *if $\mathcal{G}_{\hat{Z}}$ is isomorphic to $\mathcal{G}_Z$.*
2. **Perfect latent recovery:** *We have* perfect latent recovery *if $\hat{Z}(X)$ is an element-wise diffeomorphism of a permutation of Z.*

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 encoder. To estimate g^{-1}, first, we define $\mathcal{H}$ as the set of all possible candidates for it. 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,

$$\mathcal{H} \triangleq \{h : \mathcal{X} \to \mathbb{R}^n \ : \ \exists h^{-1} : \mathbb{R}^n \to \mathbb{R}^d \ \text{ such that } \ \forall X \in \mathcal{X} : \ \ (h^{-1} \circ h)(X) = X\} \ . \tag{10}$$

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

$$\hat{Z}(h) \triangleq h(X) = (h \circ g)(Z) \ , \qquad \forall\, h \in \mathcal{H} \ \ , \ X \in \mathcal{X} \ . \tag{11}$$

$\hat{Z}(h)$ inherits its statistical model from the randomness in X and the choice of h. We denote the score functions of $\hat{Z}(h)$ under environments $\mathcal{E}^0$, $\mathcal{E}^m$, and $\tilde{\mathcal{E}}^m$ by $s_{\hat{Z}}(\cdot; h)$, $s^m_{\hat{Z}}(\cdot; h)$, and $\tilde{s}^m_{\hat{Z}}(\cdot; h)$, respectively.

3 Identifiability and Achievability Results

In this section, we provide identifiability 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) *By using observational data and interventional data from two* ***uncoupled*** *environments for which each pair of p_i, q_i, and $\tilde{q}_i$ satisfies interventional discrepancy for all $i \in [n]$, identifiability (perfect DAG and latent recovery) is possible. Furthermore, Algorithm 1-*(OPT2) *achieves perfect recovery.*

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 [2], which requires coupled environments. Furthermore, Theorem 1 does not require faithfulness whereas [2] requires that the estimated latent distribution is faithful to the associated candidate graph for all $h \in \mathcal{H}$. This is a strong requirement to verify. Even though it does not compromise the identifiability result, it poses challenges to developing a recovery algorithm. In contrast, we only require access to the observational data, which is generally the case in practice. Based on this, we can develop a concrete recovery algorithm provided in Section 4. Next, if the environments are coupled, we prove identifiability under weaker assumptions on interventional discrepancy.

Theorem 2 (Coupled Environments) *By using interventional data from two* ***coupled*** *environments that satisfy interventional discrepancy for all $i \in [n]$, perfect recovery of the latent variables is possible. Furthermore, perfect DAG recovery is also possible by adding observational data under the relevant interventional discrepancy. Furthermore, Algorithm 1-*(OPT1) *achieves perfect recovery.*

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

Algorithm 1 Generalized Score-based Causal Latent Estimation via Interventions (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 $\mathcal{G}_{\hat{Z}}$.

```
Step 1: Compute score differences: (s_X − s_X^m), (s_X − s̃_X^m), and (s_X^m − s̃_X^m) for all m ∈ [n].
Step 2: Identify the encoder by minimizing score variations:
if is_coupled then
    Solve (OPT1), select a solution h*.
else                                                        ▷ search for the correct coupling
    for all permutations π of [n] do
        Temporarily relabel Ẽ^m to Ẽ^{π_m} for all m ∈ [n], and solve (OPT2)
        If there is a solution, select a solution h* and break from the loop.
    end for
end if
Step 3: Latent estimates: Ẑ = h*(X).
Step 4: Latent DAG recovery: Construct latent DAG G_Ẑ using (16).
return Ẑ and G_Ẑ.
```

for recovering the graph. In the next theorem, we additionally show even for DAG recovery, the observational data becomes unnecessary if we assume additive noise for our causal models and a weak faithfulness condition.

Theorem 3 (Dispensing with Observational Data) *By using interventional data from two* ***coupled*** *environments that satisfy interventional discrepancy for all $i \in [n]$, identifiability (perfect DAG and latent recovery) is possible if the latent causal model has additive noise, p is twice differentiable, and it satisfies the adjacency-faithfulness condition [8] with respect to the original latent graph $\mathcal{G}_Z$.*

4 GSCALE-I Algorithm

This section serves a two-fold purpose. First, it provides for the constructive proof steps for identifiability results specified in Theorems 1–2. Secondly, it provides an algorithm that has provable guarantee 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.

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. The following result is a corollary of Lemma 3 in Appendix B, and establishes that the changes in the score functions of the latent variables can be traced from the changes in the score functions of the observed variables. Specifically, we establish a relationship between the score *differences* across different observational and/or interventional environments. 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 candidate encoder h, based on (11), the estimated latent variable $\hat{Z}(h)$ and X are related through $\hat{Z}(h) = h(X)$. We use this relationship to characterize those between the score differences as formalized next.

Lemma 1 (Score Differences) *Score differences under different environment pairs are related as:*

$$\text{between } \mathcal{E}^0 \text{ and } \mathcal{E}^m : \quad s_{\hat{Z}}(\hat{z}; h) - s_{\hat{Z}}^m(\hat{z}; h) = \left[J_{h^{-1}}(\hat{z})\right]^\top \left(s_X(x) - s_X^m(x)\right) , \tag{12}$$

$$\text{between } \mathcal{E}^0 \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{Z}}(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = \left[J_{h^{-1}}(\hat{z})\right]^\top \left(s_X(x) - \tilde{s}_X^m(x)\right) , \tag{13}$$

$$\text{between } \mathcal{E}^m \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{Z}}^m(\hat{z}; h) - \tilde{s}_{\hat{Z}}^m(\hat{z}; h) = \left[J_{h^{-1}}(\hat{z})\right]^\top \left(s_X^m(x) - \tilde{s}_X^m(x)\right) . \tag{14}$$

We will show that among all candidate encoders $h \in \mathcal{H}$, the ground truth 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 4). To formalize these, corresponding to each candidate encoder $h \in \mathcal{H}$ we define score change matrices

$D_{\mathrm{t}}(h)$, $D_{\mathrm{obs}}(h)$, and $\tilde{D}_{\mathrm{obs}}(j)$ as follows. For all $i, m \in [n]$:

$$[D_{\mathrm{t}}(h)]_{i,m} \triangleq \mathbb{E}\big[\big|[s^m_{\hat{Z}}(\hat{Z};h)]_i - [\tilde{s}^m_{\hat{Z}}(\hat{Z};h)]_i\big|\big] \ , \tag{15}$$

$$[D_{\mathrm{obs}}(h)]_{i,m} \triangleq \mathbb{E}\big[\big|[s_{\hat{Z}}(\hat{Z};h)]_i - [s^m_{\hat{Z}}(\hat{Z};h)]_i\big|\big] \ , \tag{16}$$

$$[\tilde{D}_{\mathrm{obs}}(h)]_{i,m} \triangleq \mathbb{E}\big[\big|[s_{\hat{Z}}(\hat{Z};h)]_i - [\tilde{s}^m_{\hat{Z}}(\hat{Z};h)]_i\big|\big] \ , \tag{17}$$

where expectations are under the measures of latent score functions induced by the probability measure of observational data. The entry $[D_{\mathrm{t}}(h)]_{i,m}$ will be strictly positive only when there is a set of samples X with a strictly positive measure that renders non-identical scores $s^m_{\hat{Z}}(\hat{z};h)$ and $\tilde{s}^m_{\hat{Z}}(\hat{z};h)$. Similar properties hold for the entries of $D_{\mathrm{obs}}(h)$ and $\tilde{D}_{\mathrm{obs}}(h)$ for the respective score functions. The 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 candidate encoders $\mathcal{H}$.

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

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 will have two different approaches for coupled and uncoupled settings.

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

$$\begin{cases} \min\limits_{h \in \mathcal{H}} & \|D_{\mathrm{t}}(h)\|_0 \\ \ \text{s.t.} & D_{\mathrm{t}}(h) \ \ \text{is a diagonal matrix} \ . \end{cases} \tag{OPT1}$$

Constraining $D_{\mathrm{t}}(h)$ to be diagonal enforces that the final estimate $\hat{Z}$ and Z will be related by permutation $\mathcal{I}$. We select a solution of (OPT1) 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 do so, 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. Then, we solve the following optimization problem,

$$\begin{cases} \min\limits_{h \in \mathcal{H}} & \|D_{\mathrm{t}}(h)\|_0 \\ \ \text{s.t.} & D_{\mathrm{t}}(h) \ \ \text{is a diagonal matrix} \\ & \mathbb{1}\{D_{\mathrm{obs}}(h)\} = \mathbb{1}\{\tilde{D}_{\mathrm{obs}}(h)\} \\ & \mathbb{1}\{D_{\mathrm{obs}}(h)\} \odot \mathbb{1}\{D^{\top}_{\mathrm{obs}}(h)\} = I_n \ . \end{cases} \tag{OPT2}$$

The constraint $\mathbb{1}\{D_{\mathrm{obs}}(h)\} = \mathbb{1}\{\tilde{D}_{\mathrm{obs}}(h)\}$ ensures that a permutation of the correct encoder is a solution to (OPT2) if the coupling is correct, and the last constraint ensures that $D_{\mathrm{obs}}(h)$ does not contain 2-cycles. We will show that (OPT2) admits a solution if and only if π is the correct coupling (see Lemma 6), in which case, we select a solution of (OPT2) as our encoder estimate and denoted it by h^*.

Remark 1 *For the nonparametric identifiability results, having an oracle that solves the functional optimization problems in* (OPT1)-(OPT2) *is sufficient. For achievability under any desired class of functions $\mathcal{H}$ (e.g., linear, polynomial, and neural networks) these two problems can be converted to parametric optimization problems.*

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 $\mathcal{G}_{\hat{Z}}$ from $D_{\mathrm{obs}}(h^*)$ by assigning the non-zero coordinates of the i-th column of $D_{\mathrm{obs}}(h^*)$ as the parents of node i in $\mathcal{G}_{\hat{Z}}$, i.e.,

$$\overline{\mathrm{pa}}_{\mathcal{G}_{\hat{Z}}}(i) \triangleq \big\{j : [D_{\mathrm{obs}}(h^*)]_{j,i} \neq 0\big\} \ , \quad \forall i \in [n] \ . \tag{18}$$

5 Empirical Evaluations

We empirically evaluate the performance of the GSCALE-I algorithm for recovering the transformation g and the latent DAG $\mathcal{G}_Z$ under *coupled* interventions on synthetic data by solving an ℓ_1-relaxation of the optimization problem (OPT1). 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. To this end, we focus on a non-polynomial transform g and a non-linear latent model. We elaborate on the implementation details and provide additional results in Appendix D.

Data generation. To generate $\mathcal{G}_Z$ we use the Erdős-Rényi model with density 0.5 and $n \in \{5, 8\}$ nodes, which is generally the size of the latent graphs considered in CRL literature. For the observational causal mechanisms, we adopt an additive noise model with $Z_i = \sqrt{Z_{\mathrm{pa}(i)}^\top A_{p,i} Z_{\mathrm{pa}(i)}} + N_{p,i}$, where $\{A_{p,i} : i \in [n]\}$ are positive-definite matrices, and the noise terms are zero-mean Gaussian variables with variances $\sigma_{p,i}^2$ 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 $\sigma_{q,i}^2 = \sigma_{p,i}^2 + 1$ and $\sigma_{\tilde{q},i}^2 = \sigma_{p,i}^2 + 2$. We consider target dimension values $d \in \{5, 8, 25, 40\}$. For each (n, d) pair, we generate 100 latent graphs and N samples of Z per graph, where we set $N = 100$ for $n = 5$ and $N = 300$ for $n = 8$. As the transformation, we consider a generalized linear model,

$$X = g(Z) = \tanh(T \cdot Z) , \tag{19}$$

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

Score functions. 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 in this section, we adopt a score oracle that computes the score differences in Step 1 by leveraging Lemma 3 and using the ground truth score functions s, s^m and $\tilde{s}^m$ (see Appendix D for details). Subsequently, we can assess the performance of our novel methodology without inheriting the errors of the score estimation procedure.

Evaluation metrics. We assess the recovering latent variables by the closeness of latent variable estimates $\hat{Z}$ (and parameters of the transform, $\hat{T}$) to ground truth Z (and T). We report the mean normalized error rates $\|Z - \hat{Z}\|_2 / \|Z\|_2$ and $\|T - \hat{T}\|_{\mathrm{F}} / \|T\|_{\mathrm{F}}$. For assessing the recovery of the latent DAG, we report structural Hamming distance (SHD) between the estimate $\mathcal{G}_{\hat{Z}}$ and true graph $\mathcal{G}_Z$ as well as the average precision and recall rates of the true edges.

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 Frobenius norm of the error shows that $\hat{T}$ explains more than 80% of T correctly. Note that, $\mathcal{G}_Z$ with $n = 8$ nodes and density 0.5 has an expected number of 14 edges. Hence, having an average SHD of less than 2 edges and precision and recall rates over 0.9 indicate that GSCALE-I yields a high performance at recovering latent causal relationships even when the transformation estimate is reasonable but not perfect. Finally, we 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. This is especially important since the dimension of the observed data is usually much higher than the latent dimension in practice.

Table 2: Recovery of the latent variables and latent DAG using GSCALE-I with a score oracle

n	d	$\frac{\|Z - \hat{Z}\|_2}{\|Z\|_2}$	$\frac{\|T - \hat{T}\|_{\mathrm{F}}}{\|T\|_{\mathrm{F}}}$	SHD	precision	recall
5	5	0.03	0.02	0.12	0.99	0.99
5	25	0.03	0.02	0.04	0.99	0.99
5	40	0.04	0.03	0.09	0.99	0.99
8	8	0.16	0.14	1.56	0.92	0.97
8	25	0.20	0.13	1.55	0.93	0.96
8	40	0.21	0.13	1.14	0.96	0.96

References

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

[2] 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," *arXiv:2306.00542*, 2023.

[3] 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.

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

[5] S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. Ravikumar, "Learning linear causal representations from interventions under general nonlinear mixing," *arXiv:2306.02235*, 2023.

[6] J. Zhang, C. Squires, K. Greenewald, A. Srivastava, K. Shanmugam, and C. Uhler, "Identifiability guarantees for causal disentanglement from soft interventions," *arXiv:2307.06250*, 2023.

[7] W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf, "Causal component analysis," *arXiv preprint arXiv:2305.17225*, 2023.

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

[9] Y. Jiang and B. Aragam, "Learning nonparametric latent causal graphs with unknown interventions," *arXiv:2306.02899*, 2023.

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

A Related Work

In addition to the studies covered in Section 1, other related studies that focus on the *parametric* settings include [3–6]. Specifically, [3] considers linear causal models and proves identifiability under hard interventions and the impossibility of identifiability under soft interventions. The study in [4] 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. [6] builds on the results of [4], considers polynomial transforms under non-linear causal models, and proves identifiability under soft interventions. Finally, [5] focuses on linear Gaussian causal models and extends the results of [3] to prove identifiability for general transforms. Other studies on the *nonparametric* settings include [7, 9]. The study in [9] 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 [7] assumes that the latent DAG is already known and recovers the latent variables under hard interventions.

B Score Function Properties under Interventions

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

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

$$\exists z \in \mathbb{R}^n \;\; f(z) \neq g(z) \quad \iff \quad \mathbb{E}\big[|f(Z) - g(Z)|\big] \neq 0 \, . \tag{20}$$

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, $\mathbb{E}[|f(Z) - g(Z)|] \neq 0$. On the other direction, if $f(z) = g(z)$ for all $z \in \mathbb{R}^n$, then $\mathbb{E}[|f(Z) - g(Z)|] = 0$. ■

Lemma 2 (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 k-th coordinate if and only if node k or one of its children is intervened in $\mathcal{E}^m$ (or $\tilde{\mathcal{E}}^m$), i.e.,

$$\mathbb{E}\big[\big|[s(Z)]_k - [s^m(Z)]_k\big|\big] \neq 0 \iff k \in \overline{\mathrm{pa}}(I^m) \, ,$$
$$\text{and} \;\; \mathbb{E}\big[\big|[s(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0 \iff k \in \overline{\mathrm{pa}}(\tilde{I}^m) \, .$$

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

$$\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0 \iff \{k\} = I^m \, .$$

*(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*

$$Z_i = f_{p,i}(Z_{\mathrm{pa}(i)}) + N_{p,i} \, ,$$

where functions $\{f_{p,i} : i \in [n]\}$ are general functions and $\{N_{p,i} : i \in [n]\}$ 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 k-th coordinate if and only if node k or one of its children is intervened,

$$\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0 \iff k \in \overline{\mathrm{pa}}(I^m, \tilde{I}^m) \, .$$

Proof

Case (i) The statement directly follows from Lemma 4 of [1].

Case (ii) Suppose that $I^m = \tilde{I}^m = \{i\}$. Following (8) and (9), we have

$$s^m(z) = \nabla_z \log q_i(z_i) + \sum_{l \in [n] \setminus \{i\}} \nabla_z \log p_l(z_l \mid z_{\mathrm{pa}(l)}) \ , \tag{21}$$

$$\tilde{s}^m(z) = \nabla_z \log \tilde{q}_i(z_i) + \sum_{l \in [n] \setminus \{i\}} \nabla_z \log p_l(z_l \mid z_{\mathrm{pa}(l)}) \ . \tag{22}$$

Then, subtracting (22) from (21) and looking at k-th coordinate, we have

$$[s^m(z)]_k - [\tilde{s}^m(z)]_k = \frac{\partial \log q_i(z_i)}{\partial z_k} - \frac{\partial \log \tilde{q}_i(z_i)}{\partial z_k} \ . \tag{23}$$

Hence, if $k \neq i$, the right-hand side is zero and we have $[s(z)]_k - [\tilde{s}(z)]_k = 0$ for all z. On the other hand, if $k = i$, $q_i(z_i)$ and $\tilde{q}_i(z_i)$ being distinct implies that there exists $z \in \mathbb{R}^n$ such that $q_i(z_i) \neq \tilde{q}_i(z_i)$, and by Proposition 1, we have $\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0$.

Case (iii) Suppose that $I^m = \{i\}$ and $\tilde{I}^m = \{j\}$, and $i \neq j$. Following (8) and (9), we have

$$s^m(z) = \nabla_z \log q_i(z_i) + \nabla_z \log p_j(z_j \mid z_{\mathrm{pa}(j)}) + \sum_{l \in [n] \setminus \{i,j\}} \nabla_z \log p_l(z_l \mid z_{\mathrm{pa}(l)}) \ , \tag{24}$$

$$\tilde{s}^m(z) = \nabla_z \log q_j(z_j) + \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) + \sum_{l \in [n] \setminus \{i,j\}} \nabla_z \log p_l(z_l \mid z_{\mathrm{pa}(l)}) \ . \tag{25}$$

Hence, s^m and $\tilde{s}^m$ differ in only the causal mechanism of nodes i and j. Subtracting (25) from (24) we have

$$s^m(z) - \tilde{s}^m(z) = \ \nabla_z \log q_i(z_i) + \nabla_z \log p_j(z_j \mid z_{\mathrm{pa}(j)}) - \nabla_z \log q_j(z_j) - \nabla_z \log p_i(z_i \mid z_{\mathrm{pa}(i)}) \ , \tag{26}$$

$$[s^m(z)]_k - [\tilde{s}^m(z)]_k = \frac{\partial \log q_i(z_i)}{\partial z_k} + \frac{\partial \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_k} - \frac{\partial \log q_j(z_j)}{\partial z_k} - \frac{\partial \log p_i(z_i \mid z_{\mathrm{pa}(i)})}{\partial z_k} \ . \tag{27}$$

Proof of $\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0 \implies k \in \overline{\mathrm{pa}}(i,j)$: Note that none of the terms in the right-hand side of (27) is a function of z_k if $k \notin \overline{\mathrm{pa}}(i,j)$. Therefore, all the terms in the right-hand side of (27) are zero, and we have $[s^m(z)]_k - [\tilde{s}^m(z)]_k = 0$ for all z. By Proposition 1, $\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] = 0$. This, equivalently, means that if $\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0$, then $k \in \overline{\mathrm{pa}}(i,j)$.

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

If $k = i$. Then, (27) reduces to

$$0 = [s^m(z)]_i - [\tilde{s}^m(z)]_i = \frac{\partial \log q_i(z_i)}{\partial z_i} - \frac{\partial \log p_i(z_i \mid z_{\mathrm{pa}(i)})}{\partial z_i} \ . \tag{28}$$

If i is a root node, i.e., $\mathrm{pa}(i) = \emptyset$, (27) becomes $(\log q_i)'(z_i) = (\log p_i)'(z_i)$ for all z_i. Integrating, we get $p_i(z_i) = Cq_i(z_i)$ for some constant C. Since both p_i and q_i are pdfs, they both integrate to one, implying $C = 1$ and $p_i(z_i) = q_i(z_i)$, which contradicts the premise that observational and interventional mechanisms are distinct.

If i is not a root node, consider some $l \in \mathrm{pa}(i)$. Then, taking the derivative of (28) with respect to z_l, we have

$$0 = \frac{\partial^2 \log p_i(z_i \mid z_{\mathrm{pa}(i)})}{\partial z_i \partial z_l} \ . \tag{29}$$

Recall the equation $Z_i = f_{p,i}(Z_{\mathrm{pa}(i)}) + N_{p,i}$ for additive noise models where the noise term $N_{p,i}$ has pdf p_{N_i}. Then, the conditional pdf $p_i(z_i \mid z_{\mathrm{pa}(i)})$ is given by $p_i(z_i \mid z_{\mathrm{pa}(i)}) = p_{N_i}(z_i - f_{p,i}(z_{\mathrm{pa}(i)}))$. Denoting the score function of p_{N_i} by $t_{p,i}$,

$$t_{p,i}(u) \triangleq \frac{\mathrm{d}}{\mathrm{d}u} \log p_{N_i}(u) \ , \tag{30}$$

we have

$$\frac{\partial \log p_i(z_i \mid z_{\mathrm{pa}(i)})}{\partial z_i} = \frac{\partial \log p_{N_i}(z_i - f_{p,i}(z_{\mathrm{pa}(i)}))}{\partial z_i} = t_{p,i}(z_i - f_{p,i}(z_{\mathrm{pa}(i)})) \ . \tag{31}$$

Substituting into (29), we obtain

$$0 = \frac{\partial t_{p,i}(z_i - f_{p,i}(z_{\mathrm{pa}(i)}))}{\partial z_l} \tag{32}$$

$$= -\frac{\partial f_{p,i}(z_{\mathrm{pa}(i)})}{\partial z_l} \cdot (t_{p,i})'(z_i - f_{p,i}(z_{\mathrm{pa}(i)})) \ , \quad \forall z \in \mathbb{R}^n \ . \tag{33}$$

Since l is a parent of i, there exists a fixed $Z_{\mathrm{pa}(i)} = z^*_{\mathrm{pa}(i)}$ realization for which $\partial f_i(z^*_{\mathrm{pa}(i)})/\partial z^*_l$ is non-zero. Otherwise, $f_{p,i}(z_{\mathrm{pa}(i)})$ would not be sensitive to z_l which is contradictory to l being a parent of i. Note that Z_i can vary freely after fixing $Z_{\mathrm{pa}(i)}$. Therefore, for (33) to hold, the derivative of $t_{p,i}$ must always be zero. However, the score function of a valid pdf with full support cannot be constant. Therefore, $[s^m(z)]_i - [\tilde{s}^m(z)]_i$ is not always zero, and we have $[s^m(z)]_i \neq [\tilde{s}^m(z)]_i$.

If $k \neq i$. Then, (27) reduces to

$$0 = [s^m(z)]_k - [\tilde{s}^m(z)]_k = \frac{\partial \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_k} - \frac{\partial \log q_j(z_j)}{\partial z_k} - \frac{\partial \log p_i(z_i \mid z_{\mathrm{pa}(i)})}{\partial z_k} \ . \tag{34}$$

We investigate case by case and reach a contradiction for each case. First, suppose that $k \notin \mathrm{pa}(i)$. Then, we have $k \in \overline{\mathrm{pa}}(j)$, and (34) reduces to

$$0 = [s^m(z)]_k - [\tilde{s}^m(z)]_k = \frac{\partial \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_k} - \frac{\partial \log q_j(z_j)}{\partial z_k} \ . \tag{35}$$

If $k = j$, the impossibility of (35) directly follows from the impossibility of (28). The remaining case is $k \in \mathrm{pa}(j)$. In this case, taking the derivative of the right-hand side of (35) with respect to z_j, we obtain

$$0 = \frac{\partial^2 \log p_j(z_j \mid z_{\mathrm{pa}(j)})}{\partial z_k \partial z_j} \ , \tag{36}$$

which is a realization of (29) for $k \in \mathrm{pa}(j)$ and j in place of $l \in \mathrm{pa}(i)$ and i, which we proved to be impossible previously. Therefore, $k \notin \mathrm{pa}(i)$ is not viable. Finally, suppose that $k \in \mathrm{pa}(i)$. Then, taking the derivative of the right-hand side of (34) with respect to z_i, we obtain

$$0 = \frac{\partial^2 \log p_i(z_i \mid z_{\mathrm{pa}(i)})}{\partial z_k \partial z_i} \ , \tag{37}$$

which is again a realization of (29) for $l = k$, Note that (37) is a specific realization of (29) for $l = k$, which we proved to be impossible previously.

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

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 3 (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 $r \geq s$ and $f : \mathbb{R}^s \to \mathbb{R}^r$ is a differentiable and injective function. The difference of the score functions of Y^1 and Y^2, and W^1 and W^2 are related through*

$$s_{W^1}(w) - s_{W^2}(w) = [J_f(w)]^\top (s_{Y^1}(y) - s_{Y^2}(y)) \ , \tag{38}$$

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

Proof We are given that $Y^1 = f(W^1)$ and $Y^2 = f(W^2)$ where $g : \mathbb{R}^s \to \mathbb{R}^r$ is a differentiable and injective function. The realizations of W^1 and Y^1, and that of W^2 and Y^2, are related through $y = f(w)$. 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

$$[J_f(w)]_{i,j} = \frac{\partial [f(w)]_i}{\partial w_j} = \frac{\partial y_i}{\partial w_j} , \quad \forall i \in [r] , j \in [s] . \tag{39}$$

In this case, the pdfs of W^1 and Y^1 are related through [10]

$$p_{W^1}(w) = p_{Y^1}(y) \cdot \left|\det\left([J_f(w)]^\top \cdot J_f(w)\right)\right|^{1/2} . \tag{40}$$

Next, note that the gradient of f with respect to $w \in \mathbb{R}^s$ is given by

$$[\nabla_w f(y)]_i = \frac{\partial}{\partial w_i} f(y) = \sum_{j=1}^{r} \frac{\partial f(y)}{\partial y_j} \cdot \frac{\partial y_j}{\partial w_i} = \sum_{j=1}^{r} [\nabla_y f(y)]_j \cdot [J_f(w)]_{j,i} . \tag{41}$$

Hence, from (41) and $y = f(w)$, more compactly, we have

$$\nabla_w f(y) = [J_f(w)]^\top \cdot \nabla_y f(y) . \tag{42}$$

Next, given the identities in (40) and (42), we find the relationship between score functions of W^1 and Y^1 as follows.

$$s_{W^1}(w) = \nabla_w \log p_{W^1}(w) \tag{43}$$

$$\overset{(40)}{=} \nabla_w \log p_{Y^1}(y) + \nabla_w \log \left|\det\left([J_f(w)]^\top \cdot J_g(w)\right)\right|^{1/2} \tag{44}$$

$$\overset{(42)}{=} [J_f(w)]^\top \cdot \nabla_y \log p_{Y^1}(y) + \nabla_w \log \left|\det\left([J_f(w)]^\top \cdot J_f(w)\right)\right|^{1/2} \tag{45}$$

$$= [J_f(w)]^\top \cdot s_{Y^1}(y) + \nabla_w \log \left|\det\left([J_f(w)]^\top \cdot J_f(w)\right)\right|^{1/2} . \tag{46}$$

Following the similar steps that led to (46) for W^2 and Y^2, we obtain

$$s_{W^2}(w) = [J_f(w)]^\top \cdot s_{Y^2}(y) + \nabla_w \log \left|\det\left([J_f(w)]^\top \cdot J_f(w)\right)\right|^{1/2} . \tag{47}$$

Subtracting (47) from (46), we obtain the desired result

$$s_{W^1}(w) - s_{W^2}(w) = [J_f(w)]^\top (s_{Y^1}(y) - s_{Y^2}(y)) . \tag{48}$$

■

Using Lemma 3, we immediately prove Lemma 1 as follows.

Proof of Lemma 1 We denote the scores of estimated latent variables $\hat{Z}(h)$ under environments $\mathcal{E}^0$, $\mathcal{E}^m$ and $\tilde{\mathcal{E}}^m$ by $s_{\hat{Z}}(\hat{z}; h)$, $s^m_{\hat{Z}}(\hat{z}; h)$, and $\tilde{s}^m_{\hat{Z}}(\hat{z}; h)$, respectively. Note that $\hat{Z}(h) = h(X)$. Then, by setting $f = h^{-1}$, Lemma 3 yields that, score differences under different environment pairs are related as:

$$\text{between } \mathcal{E}^0 \text{ and } \mathcal{E}^m : \quad s_{\hat{Z}}(\hat{z}; h) - s^m_{\hat{Z}}(\hat{z}; h) = [J_{h^{-1}}(\hat{z})]^\top (s_X(x) - s^m_X(x)) , \tag{49}$$

$$\text{between } \mathcal{E}^0 \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{Z}}(\hat{z}; h) - \tilde{s}^m_{\hat{Z}}(\hat{z}; h) = [J_{h^{-1}}(\hat{z})]^\top (s_X(x) - \tilde{s}^m_X(x)) , \tag{50}$$

$$\text{between } \mathcal{E}^m \text{ and } \tilde{\mathcal{E}}^m : \quad s^m_{\hat{Z}}(\hat{z}; h) - \tilde{s}^m_{\hat{Z}}(\hat{z}; h) = [J_{h^{-1}}(\hat{z})]^\top (s^m_X(x) - \tilde{s}^m_X(x)) . \tag{51}$$

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).

We use the following equations in the proof of all theorems. For each $h \in \mathcal{H}$ we define $\phi_h \triangleq h \circ g$. Then, $\hat{Z}(h)$ and Z are related as

$$\hat{Z}(h) = h(X) = (h \circ g)(Z) = \phi_h(Z) . \tag{52}$$

Then, by setting $f = \phi_h^{-1}$, Lemma 3 yields

$$\text{between } \mathcal{E}^0 \text{ and } \mathcal{E}^m : \quad s_{\hat{Z}}(\hat{z}; h) - s^m_{\hat{Z}}(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} (s(z) - s^m(z)) , \tag{53}$$

$$\text{between } \mathcal{E}^0 \text{ and } \tilde{\mathcal{E}}^m : \quad s_{\hat{Z}}(\hat{z}; h) - \tilde{s}^m_{\hat{Z}}(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} (s(z) - \tilde{s}^m(z)) , \tag{54}$$

$$\text{between } \mathcal{E}^m \text{ and } \tilde{\mathcal{E}}^m : \quad s^m_{\hat{Z}}(\hat{z}; h) - \tilde{s}^m_{\hat{Z}}(\hat{z}; h) = [J_{\phi_h}(z)]^{-\top} (s^m(z) - \tilde{s}^m(z)) . \tag{55}$$

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 takes $(1, \dots, n)$ to $\mathcal{I}$, i.e., $I^{\rho_i} = i$ for all $i \in [n]$ and P_ρ to denote the permutation matrix that corresponds to ρ, i.e.,

$$[P_\rho]_{i,m} = \begin{cases} 1\,, & m = \rho_i\,, \\ 0\,, & \text{else}\,. \end{cases} \tag{56}$$

Since we consider coupled atomic interventions, the only varying causal mechanism across $\mathcal{E}^{\rho_i}$ and $\tilde{\mathcal{E}}^{\rho_i}$ is that of the intervened node in $I^{\rho_i} = \tilde{I}^{\rho_i} = \{i\}$. Then, by Lemma 2(ii), we have

$$\mathbb{E}\big[\big|[s^m(Z)]_k - [\tilde{s}^m(Z)]_k\big|\big] \neq 0 \quad \iff \quad k = i\,. \tag{57}$$

We define the *true* score change matrix $D_{\rm t}$ with entries for all $i, m \in [n]$,

$$[D_{\rm t}]_{i,m} \triangleq \mathbb{E}\big[\big|[s^m(Z)]_i - [\tilde{s}^m(Z)]_i\big|\big]\,. \tag{58}$$

Then, we have $\mathbb{1}\{[D_{\rm t}]_{:,\rho_m}\} = e_m$ (where e_m denotes m-th standard basis vector in $\mathbb{R}^n$), and $\mathbb{1}\{D_{\rm t}\} = P_\rho$. Next, we show that the number of variations between the score estimates $s^m_{\hat{Z}}(\hat{z}; h)$ and $\tilde{s}^m_{\hat{Z}}(\hat{z}; h)$ cannot be less than the number of variations under the true encoder g^{-1}, that is $n = \|D_{\rm t}\|_0$.

Lemma 4 *For every $h \in \mathcal{H}$, the score change matrix $D_{\rm t}(h)$ is at least as dense as the score change matrix D associated with the true latent variables,*

$$\|D_{\rm t}(h)\|_0 \geq \|D_{\rm t}\|_0 = n\,. \tag{59}$$

Proof Recall the definition of score change matrix $D_{\rm t}(h)$ in (15). Using (55), we can write entries of $D_{\rm t}(h)$ equivalently as

$$[D_{\rm t}(h)]_{i,m} = \mathbb{E}\big[\big|[J^{-\top}_{\phi_h}(Z)]_i(s^m(Z) - \tilde{s}^m(Z))\big|\big]\,, \quad \forall\, i, m \in [n]\,. \tag{60}$$

Since $\phi_h = h \circ g$ is a diffeomorphism, $[J^{-\top}_{\phi_h}(z)]$ is full rank for all $(h, z) \in \mathcal{H} \times \mathbb{R}^n$. Using Proposition 7 of [1], for all (h, z), there exists a permutation $\pi(h, z)$ of $[n]$ with permutation matrix $P_1(h, z)$ such that $P_1(h, z) J^{-\top}_{\phi_h}(z)$ has non-zero entries in its diagonal.

Next, recall that interventional discrepancy means that, for each $i \in [n]$, there exist 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 $z_i \in \mathbb{R} \setminus \mathcal{T}_i$ (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:

$$\mathcal{Z} \triangleq \{z \in \mathbb{R}^n : \ \ [s^{\rho_i}(z)]_i \neq [\tilde{s}^{\rho_i}(z)]_i \ \ \forall i \in [n]\}\,. \tag{61}$$

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

$$[D_{\rm t}(h)]_{\pi_i(h,z),\rho_i} = \mathbb{E}\big[\big([J^{-\top}_{\phi_h}(Z)]_{\pi_i(j,z)}(s^{\rho_i}(Z) - \tilde{s}^{\rho_i}(Z))\big|\big] \tag{62}$$

$$= \mathbb{E}\big[\big|[J^{-\top}_{\phi_h}(Z)]_{\pi_i(h,z),i}([s^{\rho_i}(Z)]_i - [\tilde{s}^{\rho_i}(Z)]_i)\big|\big]\,. \tag{63}$$

By definition of $\pi(h, z)$, for any $z \in \mathcal{Z}$, we know that $[J^{-\top}_{\phi_h}(z)]_{\pi_i(h,z),i} \neq 0$. Furthermore, by definition of $\mathcal{Z}$, we have $[s^{\rho_i}(z)]_i - [\tilde{s}^{\rho_i}(z)]_i \neq 0$. Then, we have $[D_{\rm t}(h)]_{\pi_i(h,z),\rho_i} \neq 0$. This implies that

$$\mathbb{1}\{D_{\rm t}(h)\} \succcurlyeq P_1^\top(h, z) P_\rho\,, \quad \forall h \in \mathcal{H}, \forall z \in \mathcal{Z}\,. \tag{64}$$

Therefore, $\|D_{\rm t}(h)\|_0 \geq \|P_\rho\|_0 = n$ for any candidate h, and the proof is concluded since we have $\mathbb{1}\{D_{\rm t}\} = P_\rho$. ∎

The lower bound for ℓ_0 norm is achieved if and only if $\mathbb{1}\{D_{\rm t}(h)\} = P_\rho$, which is an unknown permutation matrix. Since the only diagonal permutation matrix is I_n, the solution set of the constrained optimization problem in (OPT1) is given by

$$\mathcal{H}_1 \triangleq \{h \in \mathcal{H} : \mathbb{1}\{D_{\rm t}(h)\} = I_n\}\,. \tag{65}$$

Now, consider some fixed solution $h^* \in \mathcal{H}_1$. Since $\mathbb{1}\{D_{\rm t}(h^*)\} = I_n \succcurlyeq P_1^\top(h^*, z)P_\rho$ due to (64), we must have $P_1(h^*, z) = P_\rho$ for all $z \in \mathcal{Z}$. Then, $\pi_i(h^*, z) = \rho_i$ for all $i \in [n]$. We will show that $[J^{-1}_{\phi_{h^*}}(z)]_{j,\rho_i} = 0$ for all $z \in \mathbb{R}^n$ if $j \neq i$. To show this, consider $i \neq j$, which implies $[D_{\rm t}(h^*)]_{\rho_i,\rho_j} = 0$ since $\mathbb{1}\{D_{\rm t}(h^*)\} = I_n$. Then, using (60), $\mathbb{1}\{D_{\rm t}(h^*)\} = I_n$ and Lemma 2(ii), we have

$$0 = [D_{\rm t}(h^*)]_{\rho_i,\rho_j} = \mathbb{E}\Big[\big|[J^{-\top}_{\phi_{h^*}}(Z)]_{\rho_i}(s^{\rho_j}(Z) - \tilde{s}^{\rho_j}(Z))\big|\Big] \tag{66}$$

$$= \mathbb{E}\Big[\big|[J^{-1}_{\phi_{h^*}}(Z)]_{j,\rho_i}[s^{\rho_j}(Z)]_j - [\tilde{s}^{\rho_j}(Z)]_j\big|\Big] \tag{67}$$

Note that $[s^{\rho_j}(z)]_j - [\tilde{s}^{\rho_j}(z)]_j \neq 0$ for all $z \in \mathcal{Z}$. Hence, if $[J^{-1}_{\phi_{h^*}}(z)]_{j,\rho_i}$ was non-zero over a non-zero-measure set within $\mathcal{Z}$, $[D_{\rm t}(h^*)]_{\rho_i,\rho_j}$ would be non-zero. Therefore, $[J^{-1}_{\phi_{h^*}}(z)]_{j,\rho_i} = 0$ on a set of measure 1. Since $J^{-1}_{\phi_{h^*}}$ is a continuous function, this implies that $[J^{-1}_{\phi_{h^*}}(z)]_{j,\rho_i} = 0$ for all $z \in \mathbb{R}^n$. To see this, suppose that $[J^{-1}_{\phi_{h^*}}(z^*)]_{j,\rho_i} > 0$ for some $z^* \in \mathcal{Z}$. Due to continuity, there exists an open set including z^* for which $[J^{-1}_{\phi_{h^*}}(z^*)]_{j,\rho_i} > 0$, and since open sets have non-zero measure, we reach a contradiction. Therefore, if $i \neq j$, $[J^{-1}_{\phi_{h^*}}(z)]_{j,\rho_i} = 0$ for all $z \in \mathbb{R}^n$. Since $J^{-1}_{\phi_{h^*}}(z)$ must be full rank for all $z \in \mathbb{R}^n$, we have $[J^{-1}_{\phi_{h^*}}(z)]_{i,\rho_i} \neq 0$ for all $z \in \mathbb{R}^n$, $\forall i \in [n]$.

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

$$[\hat{Z}(h^*)]_{\rho_i} = \phi_{h^*}(Z_i) \ , \quad \forall i \in [n] \ , \tag{68}$$

which concludes the proof.

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

$$\overline{\rm pa}_{\mathcal{G}_{\hat{Z}}}(\rho_j) \triangleq \big\{\rho_i : [D_{\rm obs}(h^*)]_{\rho_i,\rho_j} \neq 0\big\} \ , \quad \forall j \in [n] \ . \tag{69}$$

Using (16) and (53), we have

$$\overline{\rm pa}_{\mathcal{G}_{\hat{Z}}}(\rho_j) \overset{(16)}{=} \{\rho_i : \mathbb{E}\big[\big|[s_{\hat{Z}}(\hat{Z}; h^*)]_{\rho_i} - [s^{\rho_j}_{\hat{Z}}(\hat{Z}; h^*)]_{\rho_i}\big|\big] \neq 0\} \tag{70}$$

$$\overset{(53)}{=} \{\rho_i : \mathbb{E}\big[\big|[J^{-\top}_{\phi_{h^*}}(Z)]_{\rho_i}(s(Z) - \tilde{s}^{\rho_j}(Z))\big|\big] \neq 0\} \tag{71}$$

$$= \{\rho_i : \mathbb{E}\big|\big([J^{-\top}_{\phi_{h^*}}(Z)]_{\rho_i,i}([s(Z)]_i - [\tilde{s}^{\rho_j}(Z)]_i)\big|\big] \neq 0\} \ . \tag{72}$$

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

$$\overline{\rm pa}_{\mathcal{G}_{\hat{Z}}}(\rho_j) = \{\rho_i : \mathbb{E}\big[\big|[s(Z)]_i - [\tilde{s}^{\rho_j}(Z)]_i\big|\big] \neq 0\} \ . \tag{73}$$

From Lemma 2(i), $\mathbb{E}\big[\big|[s(Z)]_i - [\tilde{s}^{\rho_j}(Z)]_i\big|\big] \neq 0$ if and only if $i \in \overline{\rm pa}(j)$. Therefore, (69) implies that $\rho_i \in \overline{\rm pa}_{\mathcal{G}_{\hat{Z}}}(\rho_j)$ if and only if $i \in \overline{\rm pa}(j)$, which shows that $\mathcal{G}_Z$ and $\mathcal{G}_{\hat{Z}}$ 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 is adjacency-faithful to $\mathcal{G}_Z$ and the latent causal model is an additive noise model, we can still recover $\mathcal{G}_Z$ without having access to observational environment $\mathcal{E}^0$. By Lemma 2(iii), true latent score changes across $\{\mathcal{E}^{\rho_i}, \tilde{\mathcal{E}}^{\rho_j}\}$, $i \neq j$ pairs gives us $\overline{\rm pa}(i, j)$ for every (i, j) pair. First, we will use the perfect latent recovery result to show that Lemma 2(iii) also applies to estimated latent score changes. Specifically, using (55) and $\mathbb{1}\{J^{-1}_{\phi_{h^*}}\} = P_\rho$, we have

$$[s^{\rho_i}_{\hat{Z}}(\hat{z}; h^*)]_{\rho_k} - [\tilde{s}^{\rho_j}_{\hat{Z}}(\hat{z}; h^*)]_{\rho_k} = [J^{-\top}_{\phi_{h^*}}(z)]_{\rho_k}\left(s^{\rho_i}(z) - \tilde{s}^{\rho_j}(z)\right) \tag{74}$$

$$= [J^{-\top}_{\phi_{h^*}}(z)]_{\rho_k,k}\left([s^{\rho_i}(z)]_k - [\tilde{s}^{\rho_j}(z)]_k\right) \ . \tag{75}$$

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

$$\mathbb{E}\big[\big|[s_{\hat{Z}}^{\rho_i}(\hat{Z};h^*)]_{\rho_k} - [\tilde{s}_{\hat{Z}}^{\rho_j}(\hat{Z};h^*)]_{\rho_k}\big|\big] \neq 0 \quad \iff \quad \mathbb{E}\big[\big|[s^{\rho_i}(Z)]_k \neq [\tilde{s}^{\rho_j}(Z)]_k\big|\big] \neq 0 \,. \tag{76}$$

Hence, by Lemma 2(iii),

$$\mathbb{E}\big[\big|[s_{\hat{Z}}^{\rho_i}(\hat{Z};h^*)]_{\rho_k} - [\tilde{s}_{\hat{Z}}^{\rho_j}(\hat{Z};h^*)]_{\rho_k}\big|\big] \neq 0 \quad \iff \quad k \in \overline{\mathrm{pa}}(i,j) \,. \tag{77}$$

Let us define the graph $\mathcal{G}_\rho$ that is related to $\mathcal{G}_Z$ by permutation ρ, i.e., $i \in \mathrm{pa}(j)$ if and only if $\rho_i \in \mathrm{pa}_{\mathcal{G}_\rho}(\rho_j)$. By (77), we have

$$\mathbb{E}\big[\big|[s_{\hat{Z}}^{\rho_i}(\hat{Z};h^*)]_{\rho_k} - [\tilde{s}_{\hat{Z}}^{\rho_j}(\hat{Z};h^*)]_{\rho_k}\big|\big] \neq 0 \quad \iff \quad \rho_k \in \overline{\mathrm{pa}}_{\mathcal{G}_\rho}(\rho_i,\rho_j) \,. \tag{78}$$

In the rest of the proof, we will show how to obtain $\{\mathrm{pa}_{\mathcal{G}_\rho}(i) : i \in [n]\}$ using $\{\overline{\mathrm{pa}}_{\mathcal{G}_\rho}(i,j) : i,j \in [n],\ i \neq j\}$. Since $\mathcal{G}_\rho$ is a graph isomorphism of $\mathcal{G}_Z$, it is equivalent to obtaining $\{\mathrm{pa}(i) : i \in [n]\}$ using $\{\overline{\mathrm{pa}}(i,j) : i,j \in [n],\ i \neq j\}$. Note that $\hat{Z}_i$ (which corresponds to node i in $\mathcal{G}_\rho$) is intervened in environments $\mathcal{E}^i$ and $\tilde{\mathcal{E}}^i$.

Define $B_i \triangleq \cap_{j\neq i}\overline{\mathrm{pa}}(i,j)$ and use $R \triangleq \{i \in [n] : \mathrm{pa}(i) = \emptyset\}$ to denote the set of root nodes. Note that $\overline{\mathrm{pa}}(i) \subseteq B_i$. Hence, i is a root node if $|B_i| = 1$. Construct the set $B \triangleq \{i : |B_i| = 1\}$. We investigate the graph recovery in 3 cases.

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

$$\overline{\mathrm{pa}}(i) \subseteq B_i \subseteq \bigcap_{j\in R\setminus\{i\}} \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(i) \cup (\bigcap_{j\in R\setminus\{i\}} \{j\}) = \overline{\mathrm{pa}}(i) \,. \tag{79}$$

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

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

$$\overline{\mathrm{pa}}(i) \subseteq B_i \subseteq \bigcap_{j\in R} \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(i) \cup (\bigcap_{j\in R}\{j\}) = \overline{\mathrm{pa}}(i) \,. \tag{80}$$

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

3. $|B| \leq 1$: 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, 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 $B_i = \overline{\mathrm{pa}}(i)$ 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

$$\overline{\mathrm{pa}}(j) \subseteq B_j \subseteq \overline{\mathrm{pa}}(i,j) = \overline{\mathrm{pa}}(j) \cup \{i\} \,. \tag{81}$$

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

■

C.3 Proof of Theorem 1

Recall that $\tilde{\mathcal{I}} = \{\tilde{I}^1, \dots, \tilde{I}^n\}$ is the permutation of intervened nodes in $\tilde{\mathcal{E}}$, so coupling π considered in (OPT2) is just equal to $\tilde{\mathcal{I}}$. Similarly to definition of ρ for $\mathcal{I}$ in the Proof of Theorem 2, let $\tilde{\rho}$ be the permutation that takes $(1, \dots, n)$ to $\tilde{\mathcal{I}}$, i.e., $I^{\tilde{\rho}_i} = i$ for all $i \in [n]$ and $P_{\tilde{\rho}}$ denote the permutation matrix for the intervention order of the environments $\{\tilde{\mathcal{E}}^1, \dots, \tilde{\mathcal{E}}^n\}$,

$$[P_{\tilde{\rho}}]_{i,j} = \begin{cases} 1 \,, & j = \tilde{\rho}_i \,, \\ 0 \,, & \text{else} \,. \end{cases} \tag{82}$$

First, we show that if the coupling is incorrect, i.e., $\rho \neq \tilde{\rho}$, the optimization problem in (OPT2) does not have a feasible solution.

Lemma 5 *If the coupling is incorrect, i.e., $\rho \neq \tilde{\rho}$, the following optimization problem does not have a feasible solution.*

$$\begin{cases} \min\limits_{h \in \mathcal{H}} & \|D_{\rm t}(h)\|_0 \\ \textit{s.t.} & D_{\rm t}(h) \ \textit{ is a diagonal matrix} \\ & \mathbb{1}\{D_{\rm obs}(h)\} = \mathbb{1}\{\tilde{D}_{\rm obs}(h)\} \\ & \mathbb{1}\{D_{\rm obs}(h)\} \odot \mathbb{1}\{D_{\rm obs}^\top(h)\} = I_n \end{cases} \tag{OPT2}$$

Proof We will it prove by contradiction. Suppose that h^* is a solution to (OPT2). Note that $[J_{\phi_{h^*}}^{-\top}(z)]$ is full rank for all $z \in \mathbb{R}^n$ and for any $m \in [n]$, the score difference *vector* $(s^{\rho_i}(z) - \tilde{s}^{\rho_i}(z))$ is not identically zero. Then, from (60) and Proposition 1, $D_{\rm t}(h^*)$ cannot have any zero columns. Subsequently, $\|D_{\rm t}(h^*)\|_0 \geq n$, and since $D_{\rm t}(h^*)$ is diagonal, we have $\mathbb{1}\{D_{\rm t}(h^*)\} = I_n$. We use $J^* \triangleq J_{\phi_{h^*}}^{-\top}$ as shorthand. If $\rho_i = \tilde{\rho}_i$ for some $i \in [n]$, using $D_{\rm t}(h^*) = I_n$ and Lemma 2(ii), for $j \neq i$, we have

$$0 = [D_{\rm t}(h^*)]_{\rho_j,\rho_i} = \mathbb{E}\Big[\big|[J^*(Z)]_{\rho_j}(s^{\rho_i}(Z) - \tilde{s}^{\rho_i}(Z))\big|\Big] = \mathbb{E}\Big[\big|[J^*(Z)]_{\rho_j,i}([s^{\rho_i}(Z)]_i - [\tilde{s}^{\rho_i}(Z)]_i)\big|\Big] \,. \tag{83}$$

Recall that $[s^{\rho_i}(z)]_i - [\tilde{s}^{\rho_i}(z)]_i \neq 0$ except for a null set. Then, (83) 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

$$\forall z \in \mathbb{R}^n \ \ [J^*(z)]_{j,i} \neq 0 \quad \iff \quad j = \rho_i \,. \tag{84}$$

Now, consider the set of *mismatched nodes*

$$\mathcal{A} \triangleq \{i \in [n] : \rho_i \neq \tilde{\rho}_i\} \,. \tag{85}$$

Let $a \in \mathcal{A}$ be a non-descendant of all the other nodes in $\mathcal{A}$. There exists $b, c \in \mathcal{A}$, not necessarily distinct, such that $\rho_a = \tilde{\rho}_b$ and $\rho_c = \tilde{\rho}_a$. In four steps, we will show that $D_{\rm obs}(h^*)_{\rho_a,\rho_c} \neq 0$ and $D_{\rm obs}(h^*)_{\rho_c,\rho_a} \neq 0$, which violates the constraint $\mathbb{1}\{D_{\rm obs}(h)\} \odot \mathbb{1}\{D_{\rm obs}^\top(h)\} = I_n$ 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

$$\mathbb{E}\big[|f(Z)|\big] \neq 0 \quad \iff \quad \mathbb{E}\Big[\big|f(Z)([s(Z)]_a - [s^{\rho_a}(Z)]_a)\big|\Big] \neq 0 \,, \tag{86}$$

$$\text{and} \quad \mathbb{E}\big[|f(Z)|\big] \neq 0 \quad \iff \quad \mathbb{E}\Big[\big|f(Z)([s(Z)]_a - [\tilde{s}^{\rho_c}(Z)]_a)\big|\Big] \neq 0 \,. \tag{87}$$

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 p_a and q_a, there exists an open set within Ψ for which $[s^{\rho_a}(Z)]_a - [s(Z)]_a \neq 0$. This implies that $\mathbb{E}\Big[\big|f(Z)([s^{\rho_a}(Z)]_a - [s(Z)]_a)\big|\Big] \neq 0$. For the other direction, suppose that $\mathbb{E}\Big[\big|f(Z)([s^{\rho_a}(Z)]_a - [s(Z)]_a)\big|\Big] \neq 0$, which implies that there exists an open set Ψ for which both $f(z)$ and $[s^{\rho_a}(z)]_a - [s(z)]_a$ are non-zero. Then, $\mathbb{E}\big[|f(Z)|\big] \neq 0$, and we have (86). Similarly, due to $\rho_c = \tilde{\rho}_a$ and interventional discrepancy between p_a and $\tilde{q}_a$, we obtain (87).

Step 1: Show that $\mathbb{E}\big[|[J^*(Z)]_{\rho_a,a}|\big] \neq 0$. First, using (60) and Lemma 2(i), we have

$$[D_{\rm obs}(h^*)]_{\rho_a,\rho_a} = \mathbb{E}\big[\big|[J^*(Z)]_{\rho_a}(s(Z) - s^{\rho_a}(Z))\big|\big] \tag{88}$$

$$= \mathbb{E}\big[\big|\sum_{j\in\overline{\rm pa}(a)} [J^*(Z)]_{\rho_a,j}([s(Z)]_j - [s^{\rho_a}(Z)]_j)\big|\big] \ . \tag{89}$$

Note that $\overline{\rm pa}(a) \cap \mathcal{A} = \{a\}$ since a is non-descendant of the other nodes in $\mathcal{A}$. Consider $j \in {\rm pa}(a)$, which implies that $j \notin \mathcal{A}$ and $\rho_j = \tilde{\rho}_j$. By (84), we have $[J^*(Z)]_{\rho_a,j} = 0$. Then, the equation above reduces to

$$[D_{\rm obs}(h^*)]_{\rho_a,\rho_a} = \mathbb{E}\big[|[J^*(Z)]_{\rho_a,a}([s(Z)]_a - [s^{\rho_a}(Z)]_a)|\big] \neq 0 \ , \tag{90}$$

since diagonal entries of $D_{\rm obs}(h^*)$ are non-zero due to the last constraint in (OPT2). Then, (86) implies that $\mathbb{E}[|[J^*(Z)]_{\rho_a,a}|] \neq 0$.

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

$$[\tilde{D}_{\rm obs}(h^*)]_{\rho_a,\rho_c} = \mathbb{E}\big[|[J^*(Z)]_{\rho_a}(s(Z) - \tilde{s}^{\rho_c}(Z))|\big] \tag{91}$$

$$= \mathbb{E}\big[\big|\sum_{j\in\overline{\rm pa}(a)} [J^*(Z)]_{\rho_a,j}([s(Z)]_j - [\tilde{s}^{\rho_c}(Z)]_j)\big|\big] \tag{92}$$

$$= \mathbb{E}\big[\big|[J^*(Z)]_{\rho_a,a}([s(Z)]_a - [\tilde{s}^{\rho_c}(Z)]_a)\big|\big] \tag{93}$$

Using (87) and Step 1 result, we have $[\tilde{D}_{\rm obs}(h^*)]_{\rho_a,\rho_c} \neq 0$.

Step 3: Show that $\mathbb{E}\big[|[J^*(Z)]_{\rho_c,a}|\big] \neq 0$. Using (60) and Lemma 2(i), we have

$$[\tilde{D}_{\rm obs}(h^*)]_{\rho_c,\rho_c} = \mathbb{E}\big[\big|[J^*(Z)]_{\rho_c}(s(Z) - \tilde{s}^{\rho_c}(Z))\big|\big] \tag{94}$$

$$= \mathbb{E}\big[\big|\sum_{j\in\overline{\rm pa}(a)} [J^*(Z)]_{\rho_c,j}([s(Z)]_j - [\tilde{s}^{\rho_c}(Z)]_j)\big|\big] \tag{95}$$

$$= \mathbb{E}\big[\big|[J^*(Z)]_{\rho_c,a}([s(Z)]_a - [\tilde{s}^{\rho_c}(Z)]_a)\big|\big] \tag{96}$$

Since $\mathbb{1}\{D_{\rm obs}(h^*)\} = \mathbb{1}\{\tilde{D}_{\rm obs}(h^*)\}$, the diagonal entry $[\tilde{D}_{\rm obs}(h^*)]_{\rho_c,\rho_c}$ is non-zero. Then, using (87) we have $\mathbb{E}\big[|[J^*(Z)]_{\rho_c,a}|\big] \neq 0$.

Step 4: Show that $[D_{\rm obs}(h^*)]_{\rho_c,\rho_a} \neq 0$. Next, we use $\rho_c = \tilde{\rho}_a$ and Lemma 2(i) to obtain

$$[\tilde{D}_{\rm obs}(h^*)]_{\rho_c,\rho_a} = \mathbb{E}\big[\big|[J^*(Z)]_{\rho_c}(s(Z) - s^{\rho_a}(Z))\big|\big] \tag{97}$$

$$= \mathbb{E}\big[\big|\sum_{j\in\overline{\rm pa}(a)} [J^*(Z)]_{\rho_c,j}([s(Z)]_j - [s^{\rho_a}(Z)]_j)\big|\big] \tag{98}$$

$$= \mathbb{E}\big[\big|[J^*(Z)]_{\rho_c,a}([s(Z)]_a - [s^{\rho_a}(Z)]_a)\big|\big] \tag{99}$$

Using (86) and Step 3 result, we have $[D_{\rm obs}(h^*)]_{\rho_c,\rho_a} \neq 0$.

However, using the constraint $\mathbb{1}\{D_{\rm obs}(h^*)\} = \mathbb{1}\{\tilde{D}_{\rm obs}(h^*)\}$, we have $[D_{\rm obs}(h^*)]_{\rho_a,\rho_c} \neq 0$. Then, $[D_{\rm obs}(h^*) \odot D_{\rm obs}^\top(h^*)]_{\rho_a,\rho_c} \neq 0$, which violates the last constraint in (OPT2). Therefore, if the pairing is incorrect, optimization problem (OPT2) has no feasible solution. ■

Next, we show that if the pairing is correct, i.e., $\rho = \tilde{\rho}$, there exists a solution to (OPT2).

Lemma 6 *If the pairing is correct, i.e., $\rho = \tilde{\rho}$, $h = \rho^{-1} \circ g^{-1}$ is a solution to* (OPT2)*, and yields $\|D_{\rm t}(h)\|_0 = n$.*

Proof 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 (OPT2). First, note that $\phi_h = \rho^{-1} \circ g^{-1} \circ g = \rho^{-1}$, a simple permutation, and $J_{\phi_h}^{-\top}$ becomes permutation matrix $P_\rho^\top$. Then, for all $i, m \in [n]$ we have

$$[D_{\rm t}(h)]_{\rho_i,m} = \mathbb{E}\big[\big|[P_\rho^\top]_{\rho_i}(s^m(Z) - \tilde{s}^m(Z))\big|\big] \tag{100}$$

$$= \mathbb{E}\big[\big|[s^m(Z)]_i - [\tilde{s}^m(Z)]_i\big|\big] \ . \tag{101}$$

Then, by Lemma 2(ii), we have $[D_{\rm t}(h)]_{\rho_i,m} \neq 0$ if and only if $i = I^m$, which means $m = \rho_i$ and $D_{\rm t}(h)$ is a diagonal matrix. Hence, h satisfies the first constraint. Next, consider $D_{\rm obs}(h)$. For all $i, j \in [n]$, we have

$$[D_{\rm obs}(h)]_{\rho_i,\rho_j} = \mathbb{E}\big[\big|[P_\rho^\top]_{\rho_i}(s(Z) - s^m(Z))\big|\big] = \mathbb{E}\big[\big|[s(Z)]_i - [s^m(Z)]_i\big|\big] \ . \tag{102}$$

By Lemma 2(i), we have $[D_{\rm obs}(h)]_{\rho_i,\rho_j} \neq 0$ if and only if $i = \overline{\rm pa}(j)$. Since $\rho = \tilde{\rho}$, similarly, we have $[\tilde{D}_{\rm obs}(h_\rho)]_{\rho_i,\rho_j} \neq 0$ if and only if $i = \overline{\rm pa}(j)$. Therefore, we have $\mathbb{1}\{D_{\rm obs}(h)\} = \mathbb{1}\{\tilde{D}_{\rm obs}(h)\}$, $D_{\rm obs}(h)$ 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 (OPT2) since it satisfies all constraints and the diagonal matrix $D_{\rm t}(h)$ has $\|D_{\rm t}(h)\|_0 = n$, which is the lower bound established. ■

Lemmas 5 and 6 collectively prove identifiability as follows. We can search over the permutations of $[n]$ until (OPT2) admits a solution h^*. By Lemma 5, the existence of this solution means that pairing is correct. Note that, when the pairing is correct, the constraint set of (OPT1) is a subset of the constraints in (OPT2). Furthermore, the minimum value of $\|D_{\rm t}(h)\|_0$ is lower bounded by n (Lemma 4), which is achieved by the solution h^* (Lemma 6). Hence, h^* is also a solution to (OPT1), and perfect recovery of latent variables and the latent DAG follows from the proof of Theorem 2.

D Details of Simulations

We perform experiments for the coupled environments setting and using a regularized, ℓ_1-relaxed version of the optimization problem (OPT1). Specifically, in Step 2 of GSCALE-I, we solve the following optimization problem:

$$\min_{h\in\mathcal{H}} \|D_{\rm t}(h)\|_{1,1} + \lambda_1 \mathbb{E}\|h^{-1}(h(X)) - X\|_2^2 + \lambda_2 \mathbb{E}\|h(X)\|_2^2 \ . \tag{OPT3}$$

In this section, we describe data generation, computation of the ground truth score differences for X, justification of the optimization problem in (OPT3) and other implementation details.

Data generation details. To generate $\mathcal{G}_Z$ 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 $Z_i = \sqrt{Z_{{\rm pa}(i)}^\top A_{p,i} Z_{{\rm pa}(i)}} + N_{p,i}$, where $\{A_{p,i} : i \in [n]\}$ are positive-definite matrices and the noise terms are zero-mean Gaussian variables with variances $\sigma_{p,i}^2$ sampled randomly from ${\rm Unif}([0.5, 1.5])$. We construct the positive-definite matrix $A_{p,i}$ by generating a matrix $B_{p,i} \in \mathbb{R}^{|{\rm pa}(i)|\times|{\rm pa}(i)|}$ by sampling its entries from ${\rm Unif}([0, 1])$ and setting $A_{p,i} = B_{p,i}^\top B_{p,i}$.

For two hard interventions on node i, Z_i is set to $N_{q,i} \sim \mathcal{N}(0, \sigma_{q,i}^2)$ and $N_{\tilde{q},i} \sim \mathcal{N}(0, \sigma_{\tilde{q},i}^2)$, respectively. We set $\sigma_{q,i}^2 = \sigma_{p,i}^2 + 1$ and $\sigma_{\tilde{q},i}^2 = \sigma_{p,i}^2 + 2$. 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 samples of Z per environment per graph, where we set $N = 100$ for $n = 5$ and $N = 300$ for $n = 8$. As the transformation, we consider a generalized linear model,

$$X = g(Z) = \tanh(T \cdot Z) \ , \tag{103}$$

$$Z = g^{-1}(X) = T^+ \cdot {\rm arctanh}(X) \ , \tag{104}$$

in which $\tanh$ and ${\rm arctanh}$ is applied element-wise, and the ground truth parameters $T \in \mathbb{R}^{d\times n}$ is a randomly sampled full-rank matrix.

Score function of the quadratic model. Score functions can be computed using (7), (8), and (9). For additive noise models, all the terms in these equations have closed-form expressions. Specifically, using (31), we have

$$[s(z)]_i = t_{p,i}(n_{p,i}) - \sum_{j\in{\rm ch}(i)} \frac{\partial f_{p,j}}{\partial z_i}(z_{{\rm pa}(j)}) \cdot t_{p,j}(n_{p,j}) \ . \tag{105}$$

Since the model under investigation The model we investigate is an additive noise model with hard interventions given by

$$f_{p,i}(z_{\mathrm{pa}(i)}) = \sqrt{z_{\mathrm{pa}(i)}^\top A_{p,i} z_{\mathrm{pa}(i)}}\ , \quad \text{and} \quad N_{p,i} \sim \mathcal{N}(0, \sigma_{p,i}^2)\ , \tag{106}$$

$$f_{q,i}(z_{\mathrm{pa}(i)}) = 0\ , \quad \text{and} \quad N_{q,i} \sim \mathcal{N}(0, \sigma_{q,i}^2)\ , \tag{107}$$

$$f_{\tilde{q},i}(z_{\mathrm{pa}(i)}) = 0\ , \quad \text{and} \quad N_{\tilde{q},i} \sim \mathcal{N}(0, \sigma_{\tilde{q},i}^2)\ . \tag{108}$$

which implies

$$t_{p,i}(n_{p,i}) = -\frac{n_{p,i}}{\sigma_{p,i}^2}\ , \quad \text{and} \quad \frac{\partial f_{p,j}}{\partial z_i}(z_{\mathrm{pa}(j)}) = \frac{[A_{p,j}]_i \cdot z_{\mathrm{pa}(j)}}{\sqrt{z_{\mathrm{pa}(j)}^\top A_{p,j} z_{\mathrm{pa}(j)}}}\ , \tag{109}$$

$$t_{q,i}(n_{q,i}) = -\frac{n_{q,i}}{\sigma_{q,i}^2}\ , \quad \text{and} \quad \frac{\partial f_{q,j}}{\partial z_i}(z_{\mathrm{pa}(j)}) = 0\ , \tag{110}$$

$$t_{\tilde{q},i}(n_{q,i}) = -\frac{n_{\tilde{q},i}}{\sigma_{\tilde{q},i}^2}\ , \quad \text{and} \quad \frac{\partial f_{\tilde{q},j}}{\partial z_i}(z_{\mathrm{pa}(j)}) = 0\ . \tag{111}$$

Score functions s^m and $\tilde{s}^m$ can be computed similarly. Subsequently, using Lemma 3, we can compute the score differences $(s_X - s_X^m), (s_X - \tilde{s}_X^m)$, and $(s_X^m - \tilde{s}_X^m)$ for all $m \in [n]$ via the equations

$$[J_g^\top(z)]^+(s(z) - s^m(z)) = s_X(x) - s_X^m(x)\ , \tag{112}$$

$$[J_g^\top(z)]^+(s(z) - \tilde{s}^m(z)) = s_X(x) - \tilde{s}_X^m(x)\ , \tag{113}$$

$$[J_g^\top(z)]^+(s^m(z) - \tilde{s}^m(z)) = s_X^m(x) - \tilde{s}_X^m(x)\ . \tag{114}$$

Implementation and evaluation steps. Similar to the ground truth, we parameterize the candidate transformations h as

$$\hat{Z} = h(X) = U^+ \cdot \operatorname{arctanh}(X)\ , \tag{115}$$

$$\hat{X} = h^{-1}(\hat{Z}) = \tanh\Big(U \cdot \hat{Z}\Big)\ , \tag{116}$$

with parameter $U \in \mathbb{R}^{d \times n}$. Note that given this parameterization, the function $\phi_h(z) = (h \circ g)(z)$ is given by

$$\hat{Z} = \phi_h(Z) = U^+ \cdot T \cdot Z\ . \tag{117}$$

Subsequently, the only element-wise diffeomorphism between Z and $\hat{Z}$ is element-wise scaling, which corresponds to scaling the columns of the candidate parameter U. Thus, to eradicate the scaling effect and compare Z and $\hat{Z}$ directly, we normalize columns of the ground truth parameters T and the candidate U such that each of their columns has a unit-norm.

We use N samples from the observational environment to compute empirical expectations. Since the candidate transform h is parameterized by U, we use gradient descent to learn parameters U. To do so, we relax ℓ_0 norm in (OPT1) and instead minimize element-wise $\ell_{1,1}$ norm $\|D_{\mathrm{t}}(h)\|_{1,1}$. Note that, scaling up $\hat{Z}(h)$ by a constant c scales down the score differences by $1/c$. Hence, to prevent the vanishing of the score difference loss trivially, we add the following regularization term to the optimization objective.

$$\mathbb{E}\|\hat{Z}(h)\|_2^2 = \mathbb{E}\|h(X)\|_2^2\ . \tag{118}$$

We also add a reconstruction loss to ensure that h is an invertible transform,

$$\mathbb{E}\|h^{-1}(h(X)) - X\|_2^2\ . \tag{119}$$

In the end, we minimize the objective function

$$\|D_{\mathrm{t}}(h)\|_{1,1} + \lambda_1 \mathbb{E}\|h^{-1}(h(X)) - X\|_2^2 + \lambda_2 \mathbb{E}\|h(X)\|_2^2\ , \tag{120}$$

and denote the final parameter estimate by $\hat{T}$. Note that we do not enforce the diagonality constraint upon $D_{\mathrm{t}}(h)$. Since we learn the latent variables up to permutation, we change this constraint to a post-processing step. Specifically, we permute the columns of $\hat{T}$ to make $D_{\rm t}(h^*)$ as close to as diagonal, i.e., $\|{\rm diag}(D_{\rm t}(h^*) \odot I_n)\|_1$ is maximized.

We set $\lambda_1 = 10^{-4}$ and $\lambda_2 = 1$, and solve (OPT3) using RMSprop optimizer with learning rate 10^{-3} for 3×10^4 steps for $n = 5$ and 4×10^4 steps for $n = 8$. Finally, we normalize both T and the estimate $\hat{T}$ to compare them at the same scale. To do so, we normalize each column of T and $\hat{T}$ by the ℓ_2-norm of the respective columns.

Recall that the latent graph estimate $\mathcal{G}_{\hat{Z}}$ is constructed using $\mathbb{1}\{D_{\rm obs}(h^*)\}$. We use a threshold $\lambda_{\mathcal{G}}$ to obtain the graph from the upper triangular part of $D_{\rm obs}(h^*)$ as follows.

$$\mathrm{pa}_{\mathcal{G}_{\hat{Z}}}(i) = \{j : j < i \quad \text{and} \quad [D_{\rm obs}(h^*)]_{j,i} \geq \lambda_{\mathcal{G}}\} \, , \qquad \forall i \in [n] \, . \tag{121}$$

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

Increasing observed data dimension d. In Table 2 of Section 5, we have provided evaluations up to $d = 40$. In addition, we repeat the similar experiments for $d = 100$ with $n \in \{5, 8\}$ nodes for 100 latent graphs and $N = 100$ samples. Table 3 demonstrates the similar results to Table 2. Hence, we show the scalability of our methodology with respect to the dimensionality of the observed data.

Table 3: Recovery of the latent variables and latent DAG using GSCALE-I with a score oracle

n	d	$\frac{\|Z-\hat{Z}\|_2}{\|Z\|_2}$	$\frac{\|T-\hat{T}\|_{\rm F}}{\|T\|_{\rm F}}$	SHD	precision	recall
5	100	0.04	0.03	0.02	1.00	0.99
8	100	0.24	0.19	1.50	0.93	0.97