# Differentiable Causal Discovery for Latent Hierarchical Causal Models

**Parjanya Prajakta Prashant**[1]* **Ignavier Ng**[2] **Kun Zhang**[2,3] **Biwei Huang**[1]
[1]University of California San Diego, USA
[2]Carnegie Mellon University, USA
[3]Mohamed bin Zayed University of Artificial Intelligence, UAE

## Abstract

Discovering causal structures with latent variables from observational data is a fundamental challenge in causal discovery. Existing methods often rely on constraint-based, iterative discrete searches, limiting their scalability to large numbers of variables. Moreover, these methods frequently assume linearity or invertibility, restricting their applicability to real-world scenarios. We present new theoretical results on the identifiability of nonlinear latent hierarchical causal models, relaxing previous assumptions in literature about the deterministic nature of latent variables and exogenous noise. Building on these insights, we develop a novel differentiable causal discovery algorithm that efficiently estimates the structure of such models. To the best of our knowledge, this is the first work to propose a differentiable causal discovery method for nonlinear latent hierarchical models. Our approach outperforms existing methods in both accuracy and scalability. We demonstrate its practical utility by learning interpretable hierarchical latent structures from high-dimensional image data and demonstrate its effectiveness on downstream tasks.

## 1 Introduction

Causal discovery, the task of inferring causal relationships from observational data, is fundamental to understanding complex systems across various scientific disciplines. Traditional causal discovery methods often assume the absence of latent confounders, a property known as causal sufficiency [40, 8]. However, this condition is frequently violated in real-world scenarios, where unobserved variables can introduce spurious correlations among observed variables. For instance, in image analysis, latent semantic variables often act as common causes for multiple pixels, creating complex dependencies that are not directly observable.

Recognizing the limitations of causal sufficiency conditions, researchers have developed various approaches to handle latent confounders. Fast Causal Inference (FCI) and its extensions [40, 34, 45, 12, 11, 1] leverage conditional independence information to infer a class of possible causal graphs. While these methods can identify the presence of latent confounders, they do not provide information about causal relationships among the latent variables themselves.

Therefore, another line of research has emerged, focusing on methods that can identify causal relations between latent variables. These approaches typically introduce restrictive parametric conditions, such as linearity or discrete data, to make the problem tractable [38, 26, 19, 43, 14, 37, 42, 13, 2, 23]. Recently, Kong et al. [24] proposed a method for identifying the causal structure of non-linear latent hierarchical models. However, their approach assumes that latent variables and exogenous noise are deterministic functions of measured variables. In this paper, we establish the identifiability of non-linear latent hierarchical models without this condition.

*Corresponding author: pprashant@ucsd.edu

Accepted to the NeurIPS 2024 Workshop on Causal Representation Learning.

Moreover, these methods often employ an iterative procedure, learning local graph structures sequentially. Such iterative approaches often face challenges including scalability issues [9, 32], error propagation, and sensitivity to testing order [39, 12]. To address these empirical issues, researchers have proposed differentiable causal discovery methods [47, 44, 46, 48, 36, 29]. Unlike iterative methods or discrete search-based approaches, these methods formulate the combinatorial search problem of causal discovery as a continuous optimization algorithm, incorporating differentiable algebraic constraints to enforce structural requirements, such as acyclicity. However, these methods typically assume no latent variables. Notable exceptions are the recent work by Bhattacharya et al. [6] and Ma et al. [28] that extended these methods to include latent variables, which assume linearity and do not recover the causal relations between latent variables. To tackle these limitations, we propose a scalable differentiable causal discovery method for non-linear latent hierarchical models.

Our key contributions are as follows:

1. We establish theoretical guarantees for the identifiability of non-linear latent hierarchical models under considerably relaxed conditions. Notably, we eliminate the requirement for latent variables and exogenous noise to be deterministic functions of measured variables.
2. We introduce a novel differentiable causal discovery method for latent hierarchical models. Through comprehensive experiments on synthetic and real datasets, we demonstrate our method's improved performance and scalability compared to existing approaches for latent hierarchical models.

## 2 Non-linear Latent Hierarchical Causal Models

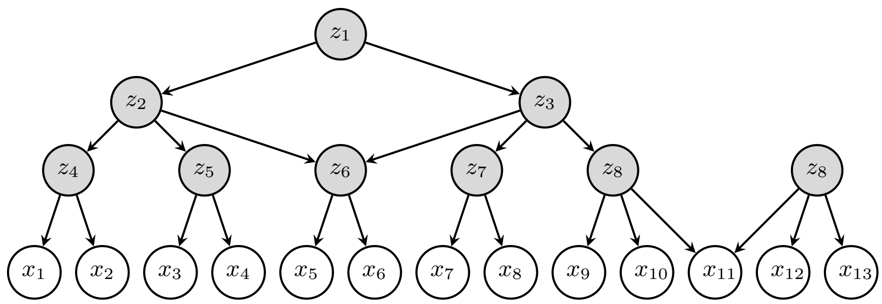



Figure 1: Example of a graph we consider. Note that we allow multiple paths between two nodes and hence generalize trees. The latent variables are shaded.

We consider a latent hierarchical causal model represented by a directed acyclic graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathbb{Z} \cup \mathbb{X}$ comprises latent variables $\mathbb{Z} = \{z_1, z_2, \ldots, z_{n_z}\}$ and measured variables $\mathbb{X} = \{x_1, x_2, \ldots, x_{n_x}\}$, and $\mathcal{E}$ denotes the set of edges representing causal relationships. The variables follow the data-generating procedure:

$$z_j = f_{z_j}(\text{Pa}(z_j), \varepsilon_{z_j}); x_i = f_{x_i}(\text{Pa}(x_i), \varepsilon_{x_i}) \tag{1}$$

where $\text{Pa}(\cdot)$ represents the set of parent variables of a given node in $\mathcal{G}$, and $\text{Pa}(x_i), \text{Pa}(z_j) \subseteq \mathbb{Z}$.

The structure of DAG $\mathcal{G}$ can be characterized by binary matrices $\boldsymbol{M}^z$ and $\boldsymbol{M}^x$, where $\boldsymbol{M}^z_{ij} = 1$ if and only if there is an edge $z_i \to z_j$, and $\boldsymbol{M}^x_{ij} = 1$ if and only if there is an edge $z_i \to x_j$. Without loss of generality, we assume $\boldsymbol{M}^z$ is upper triangular. The binary adjacency matrix $\boldsymbol{M}$ is obtained by horizontally concatenating $\boldsymbol{M}^z$ and $\boldsymbol{M}^x$, i.e., $\boldsymbol{M} = [\boldsymbol{M}^z \mid \boldsymbol{M}^x]$. The goal of this work is to recover the binary matrix $\boldsymbol{M}$ upto the relabeling of the latent variables.

In general, latent hierarchical models are not identifiable. Hence, we require structural conditions on the model to make it identifiable. We consider the following structural conditions:

**Definition 1** (Pure Children). *$v_i$ is a pure child of another variable $v_j$, if $v_j$ is the only parent of $v_i$ in the graph, i.e., Pa$(v_i) = \{v_j\}$.*

**Condition 1.** *(i) Each latent variable has two at least pure children. (ii) For any latent variable $z_i \in \mathbb{Z}$, let $\mathcal{D}_i = De(z_i) \cap \mathbb{X}$ be the set of measured descendants of $z_i$ where De(.) denotes the descendants. Then, for all $x_j, x_k \in \mathcal{D}_i$, $d(z_i, x_j) = d(z_i, x_k)$ where $d(\cdot, \cdot)$ denotes the length of the directed path between two nodes in the graph $\mathcal{G}$.*

We provide additional discussion on the above condition in Appendix D.1.

Define $\mathbb{Z}^l = \{z_i \in \mathbb{Z} : \forall x_j \in \text{De}(z_i) \cap \mathbb{X}, d(z_i, x_j) = l\}$. This denotes the set of latent variables in the $l^{\text{th}}$ layer of the model. Henceforth, we denote the vector obtained by concatenating the elements in $\mathbb{Z}^l$ as $\boldsymbol{z}^l$ and $z_i^l$ as the $i^{\text{th}}$ element of layer $l$. Note that since any node in $\mathbb{Z}^i$ has parents only in $\mathbb{Z}^{i-1}$, the adjacency matrix $\boldsymbol{M}$ can be transformed via suitable column and row permutations to the block upper-triangular structure as shown below:

$$\boldsymbol{M} = \begin{bmatrix} \mathbf{0} & \boldsymbol{M}^k & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{M}^{k-1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{M}^1 \end{bmatrix} \tag{2}$$

where $\boldsymbol{M}^i \in \{0,1\}^{|\mathbb{Z}_i| \times |\mathbb{Z}_{i-1}|}$ are binary matrices that model the causal structure between $\mathbb{Z}^i$ and $\mathbb{Z}^{i-1}$. $\boldsymbol{M}^1 \in \{0,1\}^{|\mathbb{Z}_1| \times |\mathbb{X}|}$ is the binary matrix between $\mathbb{Z}^1$ and $\mathbb{X}$. Henceforth in the paper, we assume $\boldsymbol{M}$ is modeled this way and hence always satisfies condition 1 (ii).

Our structural conditions are fairly general. We reduce the required number of pure measured variables relative to Silva et al. [38] and Kummerfeld & Ramsey [26]. Unlike Choi et al. [10] and Drton et al. [16], we do not restrict children to have only one latent parent. Furthermore, our method imposes no constraints on the neighborhood structure of variables as in [19, 43].

**Additional Notation:** For any matrix $A$, we use $A_{i,:}$ to denote its $i$-th row, $A_{:,j}$ for its $j$-th column, and $A_{i,j}$ for the element at the $i$-th row and $j$-th column. For a set $\mathbb{A}$, $\boldsymbol{a}$ denotes the vector obtained by concatenating all the elements in $\mathbb{A}$ and $\boldsymbol{a}_i$ denotes the $i^{\text{th}}$ element of $\boldsymbol{a}$.

## 3 Identifiability Theory

In this section, we describe the identifiability theory for general latent hierarchical models. Prior work has used rank constraints on the observed distribution as a graphical indicator of latent variables [38, 19, 14] for linear causal graphs. We propose a novel indicator which allows us to determine the number of latent variables which d-separate the two measured sets in the general case.

**Intuition:** Consider the case where a set of latent variables, denoted by $\mathbb{Z}$, d-separates two sets of measured variables, $\mathbb{X}$ and $\mathbb{Y}$. In this scenario, the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ can be expressed as: $p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}$. If the cardinality of $\mathbb{Z}$ is smaller than that of $\mathbb{X}$ or $\mathbb{Y}$, this imposes a constraint on the observable distribution. In most cases, the size of $\mathbb{Z}$ would to the minimum dimension of latent variables such that $p(\boldsymbol{y}|\boldsymbol{x})$ can be written as $\int p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}$. Based on this observation, we formulate a criterion based on the rank of the Jacobian of the function $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$, which allows us to show that the general case holds with probability one.

In order to rigorously prove identifiability results, we introduce some standard differentiability and faithfulness conditions [19, 14].

**Condition 2** (Generalized Faithfulness)**.** *A probability distribution $P$ is faithful to a DAG $\mathcal{G}$ if every rank Jacobian constraint on a pair of set of measured variables that holds in $P$ is entailed by every structural equation model with respect to $\mathcal{G}$.*

Faithfulness conditions are widely used in causal discovery [40, 45, 38, 19]. This is often justified by the fact that the Lebesgue measure of distributions violating faithfulness has been shown to be zero. The following proposition justifies the faithfulness condition for non-linear latent hierarchical models along similar lines.

**Proposition 1.** *The probability of a distribution $P$ generated by a structural model with respect to $\mathcal{G}$ violating Generalized Faithfulness is zero.*

**Condition 3** (Differentiability)**.** *(i) For every pair of measured sets $\mathbb{X}$ and $\mathbb{Y}$, the function $f : \mathbb{R}^{|\mathbb{X}|} \to \mathbb{R}^{|\mathbb{Y}|}$ defined as $f(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$ is continuously differentiable. (ii) For every pair of measured set $\mathbb{X}$ and latent set $\mathbb{Z}$, there exists a continuous differentiable function $g : \mathbb{R}^{|\mathbb{X}|} \to \mathbb{R}^{|\mathbb{Z}|}$ such that $p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|g(\boldsymbol{x}))$.*

Our approach considerably relaxes the constraints compared to existing work. Unlike previous methods that require linear relationships [19, 14] or deterministic functions ($z, \epsilon = f(x)$) [24], our framework accommodates a broader class of non-linear relationships between variables. Also, note that these conditions are sufficient but not necessary. In Section 5 we show we can identify structures even when Condition 3 does not hold.

We introduce a theorem that relates the graphical structure to a constraint of the distribution between two sets of measured variables.

**Theorem 1.** *Let $\mathcal{G}$ be a hierarchical latent causal model satisfying Condition 1. For any two sets of measured variables $\mathbb{X}$ and $\mathbb{Y}$ in $\mathcal{G}$, let $f(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$. Under the faithfulness and differentiability conditions, for any $r < |\mathbb{X}|, |\mathbb{Y}|$, the rank of the Jacobian matrix $\mathbf{J}_f = \frac{\partial f}{\partial \boldsymbol{x}} = r$ if and only if the size of the smallest set of latent variables that d-separates $\mathbb{X}$ from $\mathbb{Y}$ is $r$. Formally,*

$$rank(\mathbf{J}_f) = \min_{\mathbb{Z}} |\mathbb{Z}| \quad \textit{such that} \quad \mathbb{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbb{Y}|\mathbb{Z} \tag{3}$$

*where $\mathbb{Z}$ is a subset of latent variables in $\mathcal{G}$, and $\perp\!\!\!\perp_{\mathcal{G}}$ denotes d-separation in the graph $\mathcal{G}$.*

Henceforth, we use $r(\mathbb{X}, \mathbb{Y})$ to denote the rank of the Jacobian of the function $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$. Moreover, it can be shown that pure descendants of latent variables can be used as a surrogate to calculate d-separation between sets of latent variables as stated in Theorem 2 below. This theorem is partly inspired by Huang et al. [19].

**Theorem 2.** *Let $\mathcal{G}$ be a hierarchical latent causal model satisfying Condition 1. Let $\mathbb{Z}_X, \mathbb{Z}_Y \subseteq \mathbb{Z}$ be two disjoint subsets of latent variables in $\mathcal{G}$, i.e., $\mathbb{Z}_X \cap \mathbb{Z}_Y = \emptyset$. Let $\mathbb{X}$ be the set of measured variables that are d-separated by $\mathbb{Z}_X$ from all other measured variables in $\mathcal{G}$ and let $\mathbb{Y}$ be the set of measured variables that are d-separated by $\mathbb{Z}_Y$ from all other measured variables in $\mathcal{G}$. Then,*

$$r(\mathbb{Z}_X, \mathbb{Z}_Y) = r(\mathbb{X}, \mathbb{Y})$$

We now introduce three lemmas that are instrumental in proving identifiability. These lemmas provide a systematic approach to uncover the latent structure:

**Lemma 1.** *Let $\mathcal{G}$ be a graph satisfying Conditions 1. A set of measured variables $\mathbb{S}$ are pure children of the same parent if and only if for any subset $\mathbb{T} \subseteq \mathbb{S}$, $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$.*

**Lemma 2.** *Let $\mathcal{G}$ be a hierarchical latent causal model satisfying Conditions 1. Let $\mathbb{X} \subset \mathbb{V}$ be the set of measured variables. Under the generalized faithfulness condition, for any measured variable $c \in \mathbb{X}$ and any set of latent variables $\mathbb{P} \subseteq \mathbb{Z}^1$, $c$ is a child of exactly the variables in $\mathbb{P}$ if and only if the following conditions hold:*

1. *For each $\mathbb{S} \subseteq \mathbb{X}$ such that $|\mathbb{S} \cap Ch(z_i)| = 1$ for each $z_i \in \mathbb{P}$, where $Ch(z_i)$ denotes the set of pure children of $z_i$:*

$$r(\mathbb{S}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\})) = r(\mathbb{S} \cup \{c\}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\}))$$

2. *The equality in condition (1) does not hold for any proper subset of $\mathbb{P}$.*

**Lemma 3.** *Let $\mathbb{G}$ be a graph satisfying Conditions 1. A measured variable $c$ has no parent if and only if $r(\{c\}, \mathbb{X} \setminus \{c\}) = 0$.*

**Discussion:** Lemma 1 enables us to identify pure children among the measured variables $\mathbb{X}$. For example, in Figure 1 all subsets of $\{x_1, x_2\}$ are d-separated from the rest of the variables by one variable $\{z_4\}$. However, for the set $\{x_1, x_2, x_3, x_4\}$, the subset $\{x_1, x_3\}$ requires two variables $\{z_4, z_5\}$ to be d-separated from the rest of the variables. Lemma 2 provides a method to determine the parents of non-pure children. For example, in Figure 1, consider $\{x_{12}\}$ whose parents are $\{z_8, z_9\}$. For a set like $\{x_9, x_{12}\}$, which contains exactly one pure child of both parents of $\{x_{12}\}$, $r(\{x_9, x_{12}\}, \mathbb{X} \setminus \{x_9, x_{12}\}) = r(\{x_9, x_{12}, x_{11}\}, \mathbb{X} \setminus \{x_9, x_{12}, x_{11}\})$. However, this is not true for any other set of latent variables. Lemma 3 allows us to identify measured variables that have no latent parents.

Having established the necessary lemmas, we now present the identifiability of the graph structure in hierarchical latent causal models.

**Theorem 3.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hierarchical latent causal model satisfying Condition 1. Let $\boldsymbol{M}$ be the binary adjacency matrix representing the structure of $\mathcal{G}$. Let data $\mathbb{X}$ be generated according to the structural equation model defined in Equation 1. Given a function $r(\mathbb{S}, \mathbb{T})$ which outputs the minimum number of latent variables that d-separate any two measured sets $\mathbb{S}$ and $\mathbb{T}$, $\boldsymbol{M}$ is identifiable up to the permutation of the latent variables.*

The proof for Theorem 3 leverages the preceding lemmas and recursion. We begin by applying Lemmas 1, 2, and 3 to infer the structure between $\mathbb{Z}^1$ and $\mathbb{X}$. Theorem 2 then allows us to relate d-separation between sets in $\mathbb{Z}^1$ to their pure children in $\mathbb{X}$. Thus, this process can be applied recursively to higher levels of the hierarchy, enabling the identification of the entire graph structure.

## 4 Differentiable Causal Discovery Approach

In the previous section, we demonstrated that hierarchical models satisfying Condition 1 yield a unique hierarchical structure for a given distribution of measured variables. To learn the causal structure, two key steps must be performed: (i) matching the observed data distribution, and (ii) enforcing structural constraints on the model.

### 4.1 Matching the Data Distribution

To learn the causal structure, we consider the structural equation models (SEMs) in Equation 1 explicitly parameterized by the binary adjacency matrix $\boldsymbol{M}$.

$$z_j^i = f_j^i(\boldsymbol{M}^{i+1} \odot \boldsymbol{z}^{i+1}, \varepsilon_{z_j^i}); \quad x_j = g_j(\boldsymbol{M}^1 \odot \boldsymbol{z}^1, \varepsilon_{x_j}). \tag{4}$$

We employ a variational autoencoder (VAE) [22] as a generative model to learn the distribution over the measured variables. Let $\theta$ be the parameters of the VAE, and $\boldsymbol{M}$ represent the binary adjacency matrix controlling the structure of the SEM. We aim to maximize the evidence lower bound (ELBO) to approximate the true data distribution.

$$\log p(\boldsymbol{x}; \hat{\theta}, \boldsymbol{M}) \geq -\mathrm{KL}(q(\boldsymbol{\epsilon}|\boldsymbol{x})||p(\boldsymbol{\epsilon}; \hat{\theta})) + \mathbb{E}_q[\log p(\boldsymbol{x}|\boldsymbol{\epsilon}; \hat{\theta}, \boldsymbol{M})] \tag{5}$$

Here, $\boldsymbol{\epsilon}$ represents the latent variable vector obtained by concatenating all individual noise terms $\varepsilon_{z_j^i}$ and $\varepsilon_{x_j}$. The objective of the VAE is to minimize the negative ELBO $\mathcal{L}_{\text{ELBO}}$, where the KL divergence regularizes the variational posterior, and the second term encourages the generative model to match the observed data distribution.

The encoder of the VAE models the approximate posterior $q(\boldsymbol{\epsilon}|\boldsymbol{x})$. An advantage of modeling $q(\boldsymbol{\epsilon}|\boldsymbol{x})$ over $q(\boldsymbol{z}|\boldsymbol{x})$ is that it simplifies the process of enforcing the independence of each dimension of $\boldsymbol{\epsilon}$. The decoder models the conditional likelihood $p(\boldsymbol{x}|\boldsymbol{\epsilon}; \hat{\theta}, \boldsymbol{M})$, and is designed to follow the SEM equations in Equation 4. This allows the decoder to respect the structural constraints encoded in the binary adjacency matrix $\boldsymbol{M}$, ensuring that the learned distribution also reflects the underlying causal structure.

### 4.2 Enforcing Structural Constraints

In order to enforce structural constraints, we relax the binary adjacency matrix $\boldsymbol{M}$ for gradient-based optimization by using the Gumbel-softmax trick [21]. Here, $\boldsymbol{M} \sim \sigma(\gamma)$, where $\sigma$ represents the softmax function and $\gamma$ is a trainable parameter representing the logits. To ensure the causal structure satisfies Condition 1 (ii), we define $\boldsymbol{M}$ as a block upper-triangular matrix as shown in Equation 2. We now introduce the following lemma to justify the constraint required for Condition 1 (i).

**Lemma 4.** *Consider a DAG $\mathcal{G}$ with a binary adjacency matrix $\boldsymbol{M}$. $\mathcal{G}$ satisfies Condition 1 (i) if and only if:*

$$\left\| \boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:}) \right\|_1 \geq 2 \quad \forall i. \tag{6}$$

To encourage sparsity and avoid learning spurious edges, we apply an $\ell_1$ regularization on $\sigma(\gamma)$, similar to other differentiable causal discovery methods [30, 7]. The optimization objective is formulated as:

$$\max_{\theta, \gamma} \ \mathbb{E}_{\boldsymbol{M} \sim \sigma(\gamma)} \left[ \mathrm{ELBO}(\theta, \boldsymbol{M}) \right] - \lambda \|\sigma(\gamma)\|_1, \tag{7}$$

$$\text{subject to} \quad \|\boldsymbol{M}_{i,:}\|_1 \left\| \|\boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:}) \right\|_1 \geq 2 \quad \forall i. \tag{8}$$

Table 1: Performance of latent hierarchical causal discovery methods on various graphs

| | Ours | | | KONG [24] | | | HUANG [19] | | | GIN [42] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Structure | SHD ↓ | F1 ↑ | Time (s) ↓ | SHD ↓ | F1 ↑ | Time (s) ↓ | SHD ↓ | F1 ↑ | Time (s) ↓ | SHD ↓ | F1 ↑ | Time (s) ↓ |
| Tree | 0.67<br>(1.49) | 0.96<br>(0.08) | 231.40<br>(18.53) | 5.83<br>(2.04) | 0.63<br>(0.09) | 4965.20<br>(2382.16) | 6.00<br>(3.00) | 0.65<br>(0.08) | 2.87<br>(0.04) | 7.50<br>(1.50) | 0.00<br>(0.00) | 3.78<br>(0.06) |
| V-structure | 0.67<br>(1.10) | 0.97<br>(0.05) | 266.23<br>(11.33) | 7.67<br>(4.08) | 0.61<br>(0.14) | 5719.62<br>(2588.82) | 5.50<br>(2.50) | 0.72<br>(0.08) | 5.07<br>(0.06) | 8.00<br>(0.00) | 0.17<br>(0.17) | 6.13<br>(0.193) |

Note: SHD = Structural Hamming Distance (lower is better ↓). F1 scores range from 0 to 1 (higher is better ↑). Time is measured in seconds (lower is better ↓). Standard deviations are reported in parentheses.

Note, that we allow some rows of $\boldsymbol{M}$ to go zero. This allows us to learn the number of latent variables. The above method of using Gumbel softmax to approximate the binary adjacency matrices is inspired by Ng et al. [30], Brouillard et al. [7]. Furthermore, to ensure the independence of the noise terms $\varepsilon$, we introduce the following independence loss, denoted as $\mathcal{L}_{\text{ind}}(\boldsymbol{\epsilon})$, which minimizes the KL divergence between the joint distribution of $\boldsymbol{\epsilon}$ and the product of individual noise distributions $\mathcal{L}_{\text{ind}}(\boldsymbol{\epsilon}) = \text{KL}\Big(p(\boldsymbol{\epsilon}) \| \prod_j \prod_i p(\varepsilon_{z_j^i}) \prod_j p(\varepsilon_{x_j})\Big)$. This can be estimated using the Donsker-Varadhan representation [15, 4]. Therefore, the final loss function is:

$$\begin{aligned}\mathcal{L}_{\text{final}} = &-\mathbb{E}_{\boldsymbol{M}\sim\sigma(\gamma)}\left[\text{ELBO}(\theta, \boldsymbol{M})\right] + \text{KL}(q(\boldsymbol{\epsilon}|\boldsymbol{x})\|p(\boldsymbol{\epsilon}))\\ &+ \lambda_1\mathcal{L}_{\text{ind}}(\boldsymbol{\epsilon}) + \lambda_2\|\sigma(\gamma)\|_1\\ &+ \lambda_3\Big(\sum_i \max(0, \|\boldsymbol{M}_{i,:}\|_1(2 - \|\boldsymbol{M}_{i,:} \odot \prod_{j\neq i}(1 - \boldsymbol{M}_{j,:})\|_1))\Big)^2.\end{aligned} \tag{9}$$

## 5 Experiments

We conduct empirical studies to examine the efficacy of our differentiable causal discovery method. Specifically, we experiment with synthetic data in Section 5.1 and real image data in Section 5.2.

### 5.1 Synthetic Data

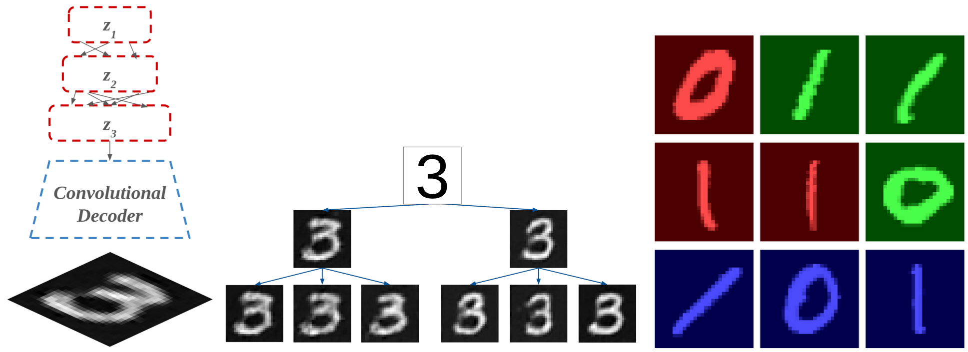
We conduct experiments on four causal structures given in Figure 3 to validate our method. We consider both trees and v-structures. We compare against other methods designed to discover latent hierarchical causal models, namely KONG [24], HUANG [19] and GIN [42]. The structural Hamming distance (SHD) and F1 score are computed for each structure and reported in Table 1. We also report the time taken for each method in seconds. We did not run 1-factor model methods like FOFC [26] since our data does not meet their conditions, and their implementation results in runtime errors. Further experimental details are given in Appendix C.1.

We observe substantial improvement in both the SHD and F1 score compared to the baselines. We note that this improvement is despite the fact that the data does not satisfy Condition 3 since LeakyReLU is not differentiable everywhere. Since other methods are designed for a restrictive class of latent hierarchical models, they are unable to identify the causal graph. Xie et al. [42] does not predict edges for most of the runs resulting in a mean F1 score close to zero.

The linear baselines [19, 42] are faster than the non-linear methods. This is because they do not have to train a non-linear model like a neural network since all relationships are linear. However, we observe that we require a much shorter runtime compared to Kong et al. [24] since we only train one neural network instead of $\mathcal{O}(ln^2)$.

### 5.2 Image data

In this section, we learn a latent causal graph for the MNIST dataset [27]. As shown in Figure 2a, we model a latent hierarchical causal structure followed by a decoder that generates the images. Since our causal discovery approach can be trained end-to-end, incorporating the decoder does not alter our methodology. The convolution decoder allows us to use spatial information and reduce the number of measured variables. Further training details are provided in Appendix C.2.




(a) (b) (c)

Figure 2: Figures for the Image experiments. (a) Latent causal graph for digit images (b) Visualization of subgraph of the learnt latent causal graph on MNIST (c) Samples of CMNIST digits; train set - top row; test set (reverse) - middle row, test set (blue) - bottom row

Table 2: Test Accuracy on the CMNIST dataset. Standard deviations are reported in parentheses.

| | Ours | Autoencoder | VAE | $\beta$-VAE ($\beta = 1\text{e}{-3}$) | $\beta$-VAE ($\beta = 1\text{e}{-1}$) | $\beta$-VAE ($\beta = 10$) |
|---|---|---|---|---|---|---|
| Reverse | 0.979 (0.004) | 0.854 (0.037) | 0.536 (0.000) | 0.843 (0.140) | 0.536 (0.000) | 0.536 (0.000) |
| Blue | 0.753 (0.106) | 0.6492 (0.195) | 0.536 (0.000) | 0.487 (0.215) | 0.536 (0.000) | 0.536 (0.000) |

We initialize the causal model with three layers and learn the underlying structure. The top layer typically captures global features, such as digit identity, while the middle layer learns variations within the same digit. The lowest layer focuses on local features that have minimal impact on the overall digit structure. Figure 2b visualizes a subset of the learned causal graph, highlighting these patterns, which were generated by intervening on different nodes in the subgraph. The top node denotes the concept of a digit three. The lower nodes are visualized upon intervention. The full causal graph, shown in Appendix C.2, has 62 latent variables demonstrating the scalability of our method.

Causal representations often contribute to better generalization and transfer learning due to the transfer of causal relations [35]. To demonstrate the effectiveness of our learned representations in domain transfer, we evaluate them on the CMNIST dataset [3]. Figure 2c shows our train and test sets. Further details on the data generating process are given in Appendix C.2.

To predict the digit labels, we first learn a latent causal structure from the dataset. We then train a logistic regression classifier on these latent representations, applying L1 regularization to encourage sparsity in the model weights. This regularization promotes the use of features within the Markov blanket of the label. We evaluate the model on the test set, and the results are presented in Table 2.

Table 2 compares performance on the two test sets: ‘Reverse,’ where the color-digit relationship is reversed in the test set, and ‘Blue,’ where all digits in the test set are blue.

## 6 Conclusion

In this work, we introduce a differentiable causal discovery method for recovering the structure of latent hierarchical causal graphs under rather mild conditions. Our approach significantly outperforms existing baselines and is scalable to high-dimensional datasets such as images. Additionally, we establish novel identifiability results without imposing restrictive assumptions on the structural equation models. Notably, our result that provides graphical information based on rank of the Jacobian matrix may inspire future work in this area.

## References


[1] Akbari, S., Mokhtarian, E., Ghassami, A., and Kiyavash, N. Recursive causal structure learning in the presence of latent variables and selection bias. *Advances in Neural Information Processing Systems*, 34:10119–10130, 2021.

[2] Anandkumar, A., Hsu, D., Javanmard, A., and Kakade, S. Learning linear bayesian networks with latent variables. In *International Conference on Machine Learning*, pp. 249–257. PMLR, 2013.

[3] Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[4] Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

[5] Bellot, A. and van der Schaar, M. Deconfounded score method: Scoring dags with dense unobserved confounding. *arXiv preprint arXiv:2103.15106*, 2021.

[6] Bhattacharya, R., Nagarajan, T., Malinsky, D., and Shpitser, I. Differentiable causal discovery under unmeasured confounding. In *International Conference on Artificial Intelligence and Statistics*, pp. 2314–2322. PMLR, 2021.

[7] Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., and Drouin, A. Differentiable causal discovery from interventional data. *Advances in Neural Information Processing Systems*, 33:21865–21877, 2020.

[8] Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

[9] Chickering, M., Heckerman, D., and Meek, C. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

[10] Choi, M. J., Tan, V. Y., Anandkumar, A., and Willsky, A. S. Learning latent tree graphical models. *The Journal of Machine Learning Research*, 12:1771–1812, 2011.

[11] Claassen, T., Mooij, J., and Heskes, T. Learning sparse causal models is not np-hard. *arXiv preprint arXiv:1309.6824*, 2013.

[12] Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pp. 294–321, 2012.

[13] Cui, R., Groot, P., Schauer, M., and Heskes, T. Learning the causal structure of copula models with latent variables. 2018.

[14] Dong, X., Huang, B., Ng, I., Song, X., Zheng, Y., Jin, S., Legaspi, R., Spirtes, P., and Zhang, K. A versatile causal discovery framework to allow causally-related hidden variables. *arXiv preprint arXiv:2312.11001*, 2023.

[15] Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212, 1983.

[16] Drton, M., Lin, S., Weihs, L., and Zwiernik, P. Marginal likelihood and model selection for gaussian latent tree and forest models. 2017.

[17] Gu, Y. and Dunson, D. B. Bayesian pyramids: Identifiable multilayer discrete latent structure models for discrete data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):399–426, 2023.

[18] Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017.

[19] Huang, B., Low, C. J. H., Xie, F., Glymour, C., and Zhang, K. Latent hierarchical causal structure discovery with rank constraints. *Advances in neural information processing systems*, 35:5549–5561, 2022.

[20] Hyvarinen, A., Sasaki, H., and Turner, R. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 859–868. PMLR, 2019.

[21] Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[22] Kingma, D. P. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[23] Kivva, B., Rajendran, G., Ravikumar, P., and Aragam, B. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34:18087–18101, 2021.

[24] Kong, L., Huang, B., Xie, F., Xing, E., Chi, Y., and Zhang, K. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36:2010–2032, 2023.

[25] Kong, L., Chen, G., Huang, B., Xing, E. P., Chi, Y., and Zhang, K. Learning discrete concepts in latent hierarchical models. *arXiv preprint arXiv:2406.00519*, 2024.

[26] Kummerfeld, E. and Ramsey, J. Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1655–1664, 2016.

[27] LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[28] Ma, P., Ding, R., Fu, Q., Zhang, J., Wang, S., Han, S., and Zhang, D. Scalable differentiable causal discovery in the presence of latent confounders with skeleton posterior (extended version). *arXiv preprint arXiv:2406.10537*, 2024.

[29] Nazaret, A., Hong, J., Azizi, E., and Blei, D. Stable differentiable causal discovery. *arXiv preprint arXiv:2311.10263*, 2023.

[30] Ng, I., Zhu, S., Fang, Z., Li, H., Chen, Z., and Wang, J. Masked gradient-based causal structure learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pp. 424–432. SIAM, 2022.

[31] Ng, I., Huang, B., and Zhang, K. Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning*, pp. 71–105. PMLR, 2024.

[32] Niu, W., Gao, Z., Song, L., and Li, L. Comprehensive review and empirical evaluation of causal discovery algorithms for numerical data. *arXiv preprint arXiv:2407.13054*, 2024.

[33] Pearl, J. Probabilistic reasoning in intelligent systems; network of plausible inference. *Morgan Kaufmann, 1988*, 1988.

[34] Pearl, J. et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

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

[36] Sethuraman, M. G., Lopez, R., Mohan, R., Fekri, F., Biancalani, T., and Hütter, J.-C. Nodags-flow: Nonlinear cyclic causal structure learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 6371–6387. PMLR, 2023.

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

[38] Silva, R., Scheines, R., Glymour, C., Spirtes, P., and Chickering, D. M. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2), 2006.

[39] Spirtes, P. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

[40] Spirtes, P., Glymour, C., and Scheines, R. *Causation, prediction, and search*. MIT press, 2001.

[41] Vahdat, A. and Kautz, J. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[42] Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., and Zhang, K. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

[43] Xie, F., Huang, B., Chen, Z., He, Y., Geng, Z., and Zhang, K. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pp. 24370–24387. PMLR, 2022.

[44] Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International conference on machine learning*, pp. 7154–7163. PMLR, 2019.

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

[46] Zhang, M., Jiang, S., Cui, Z., Garnett, R., and Chen, Y. D-vae: A variational autoencoder for directed acyclic graphs. *Advances in neural information processing systems*, 32, 2019.

[47] Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[48] Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pp. 3414–3425. Pmlr, 2020.

[49] Zheng, Y., Ng, I., and Zhang, K. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in neural information processing systems*, 35:16411–16422, 2022.

## A Related Work

**Causal Discovery for Latent Hierarchical Models:** Prior work has focused extensively on the linear case [38, 2, 26, 42, 19, 43, 14]. Silva et al. [38] and Kummerfeld & Ramsey [26] utilize tetrad conditions—the rank of each $2 \times 2$ sub-covariance matrix—to discover latent variables. However, they require further structural conditions, such as trees and three measured pure children for each latent variable. Anandkumar et al. [2] employ matrix factorization to identify latent variables based on decomposing the covariance matrix. Xie et al. [42, 43] extend the independent noise condition to latent models with non-Gaussian noise.

Huang et al. [19] use rank deficiency constraints on pairs of measured variable sets to identify the minimum number of latent variables that d-separate the two sets. Dong et al. [14] extend this framework to allow measured variables to be parents of other variables as well. Both these methods use an iterative search procedure with rank deficiency test to discover the latent graph, with a time complexity at least quadratic in the number of variables.

Kong et al. [24] introduce identifiability results for the non-linear case when the latent variables and exogenous noise are a differentiable invertible function of the measured variables ($z, \epsilon = f(x)$). Similar to other methods, they rely on an iterative search procedure that requires training at least $\mathcal{O}(ln^2)$ generative models, where $n$ is the number of measured variables and $l$ is the number of layers in the model.

Another line of work focuses on learning hierarchical structures for discrete latent variables [33, 10, 17]. However, these approaches assume the measured variables are discrete, which often does not hold in many real-world scenarios, such as images. Kong et al. [25] allows continuous variables to be adjacent to latent discrete variables. However, causal relationships and the hierarchical structure are only learned for discrete variables. Moreover, latent variables are assumed to be a deterministic invertible function of the measured variables, similar to Kong et al. [24].

**Differentiable Causal Discovery:** NOTEARS introduced a continuous optimization-based algorithm to learn linear directed acyclic graphs (DAG) by formulating graphical constraints into differentiable constraints [47]. Subsequent work parameterized DAGs using neural networks [44, 46, 30]. [48] extended these approaches to non-parametric DAGs. While these approaches scale well, they usually assume the absence of any latent variables in the DAG.

More recent differentiable causal discovery algorithms have been developed for handle latent variables [6, 5]. However, these methods do not recover the causal relations between latent variables and assume linearity. [28] make similar conditions but use a supervised approach for causal discovery.

**Non-linear ICA:** Non-linear Independent Component Analysis methods aim to recover independent latent sources from a set of measured variables. However, they do not consider generic dependence between latent variables and rely on additional assumptions such as existence of auxiliary variables [20] or sparsity [49].

## B Proofs

### B.1 Proof of Proposition 1

**Proposition 1.** *The probability of a distribution P generated by a structural model with respect to G violating Generalized Faithfulness is zero.*

*Proof.* We begin our proof with the following lemma.

**Lemma 5.** *Let $A \in \mathbb{R}^{m \times p}$ and $B \in \mathbb{R}^{p \times n}$ be random matrices drawn from a continuous distribution with $m, n \geq p$, whose entries are drawn from continuous distributions. Let $C = AB$ be their product. Then,*

$$\mathbb{P}(\mathit{rank}(C) < p) = 0$$

*with respect to the Lebesgue measure on $\mathbb{R}^{m \times n}$.*

*Proof.* The lebesgue measure of non-full rank matrices is zero. Therefore $\mathbb{P}(\text{rank}(A) = p) = 1$ and $\mathbb{P}(\text{rank}(B) = p) = 1$.

Since $C = AB$, $rank(C) \leq min(rank(A), rank(B)) = p$. For $rank(C) < p$, the vectors $Ab_{:,i}$ would have to be linearly dependent. This adds hard constraints on the matrices which has lebesgue measure zero.

□

Using the proof of Theorem 1, we know $J_{\boldsymbol{h}}(\boldsymbol{x}) = J_f(g(\boldsymbol{x}))\cdot J_g(\boldsymbol{x})$ where $J_f(g(\boldsymbol{x})) \in \mathbb{R}^{|\mathbb{Y}|\times|\mathbb{Z}|}$, $J_g \in \mathbb{R}^{|\mathbb{Z}|\times|\mathbb{X}|}$. Condition 3 ensures the Jacobian matrices are continuous. By Lemma 5, $rank(J_{\boldsymbol{h}}(\boldsymbol{x})) = |\mathbb{Z}|$ with probability one. Therefore, the probability of $rank(J_{\boldsymbol{h}}(\boldsymbol{x})) < |\mathbb{Z}|$ for all $\boldsymbol{x}$ is zero. □

### B.2 Proof of Theorem 1

**Theorem 1.** *Let $\mathcal{G}$ be a hierarchical latent causal model satisfying Condition 1. Under the faithfulness and differentiability conditions, for any two sets of measured variables $\mathbb{X}$ and $\mathbb{Y}$ in $\mathcal{G}$, the rank of the Jacobian matrix $J_f = \frac{\partial f}{\partial \boldsymbol{x}} = r < |\mathbb{X}|, |\mathbb{Y}|$ (where $f(\boldsymbol{x}) = \mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$) if and only if the size of the smallest set of latent variables that d-separates $\mathbb{X}$ from $\mathbb{Y}$ is $r$. Formally,*

$$rank(J_f) = \min_{\mathbb{Z}} |\mathbb{Z}| \quad \textit{such that} \quad \mathbb{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbb{Y} \mid \mathbb{Z} \tag{10}$$

*where $\mathbb{Z}$ is a subset of latent variables in $\mathcal{G}$, and $\perp\!\!\!\perp_{\mathcal{G}}$ denotes d-separation in the graph $\mathcal{G}$.*

*Proof.* Let $\mathbb{X}$ and $\mathbb{Y}$ be two sets of measured variables in $\mathcal{G}$. By the structure of the hierarchical latent causal model and Condition 1, there are no direct edges between measured variables. Therefore, any d-separation between $\mathbb{X}$ and $\mathbb{Y}$ must be mediated through a set of latent variables.

Let $\mathbb{Z}$ be the minimal set of latent variables that d-separates $\mathbb{X}$ from $\mathbb{Y}$, i.e.,

$$\mathbb{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbb{Y} \mid \mathbb{Z}.$$

This implies that the conditional distribution satisfies

$$p(\boldsymbol{y}|\boldsymbol{x}) = \int p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}.$$

Taking expectations, we obtain

$$\begin{aligned}
\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] &= \int \boldsymbol{y}\, p(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} && (11)\\
&= \int \boldsymbol{y} \left( \int p(\boldsymbol{y}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} \right) d\boldsymbol{y} && (12)\\
&= \int \left( \int \boldsymbol{y}\, p(\boldsymbol{y}|\boldsymbol{z})d\boldsymbol{y} \right) p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z} && (13)\\
&= \int \mathbb{E}[\boldsymbol{y}|\boldsymbol{z}]\, p(\boldsymbol{z}|\boldsymbol{x})d\boldsymbol{z}. && (14)
\end{aligned}$$

By Condition 3, there exists a differentiable function $g : \mathbb{R}^{|\mathbb{X}|} \to \mathbb{R}^{|\mathbb{Z}|}$ such that

$$p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z}|g(\boldsymbol{x})).$$

Substituting this into the expectation, we have

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \int \mathbb{E}[\boldsymbol{y}|\boldsymbol{z}]\, p(\boldsymbol{z}|g(\boldsymbol{x}))d\boldsymbol{z} = h(g(\boldsymbol{x})),$$

where we define the function $h : \mathbb{R}^{|\mathbb{Z}|} \to \mathbb{R}^{|\mathbb{Y}|}$ by

$$f(\boldsymbol{w}) = \int \mathbb{E}[\boldsymbol{y}|\boldsymbol{z}]\, p(\boldsymbol{z}|\boldsymbol{w})d\boldsymbol{z}.$$

Applying the chain rule to the composition of functions, the Jacobian of $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}]$ with respect to $\boldsymbol{x}$ is

$$J_f(\boldsymbol{x}) = J_h(g(\boldsymbol{x})) \cdot J_g(\boldsymbol{x}),$$

where $J_h(g(\boldsymbol{x}))$ is an $|\mathbb{Y}| \times |\mathbb{Z}|$ matrix and $J_g(\boldsymbol{x})$ is an $|\mathbb{Z}| \times |\mathbb{X}|$ matrix.

Using the rank inequality for matrix multiplication, we have

$$\text{rank}(J_f(\boldsymbol{x})) \leq \min\left(\text{rank}(J_h(g(\boldsymbol{x}))), \text{rank}(J_g(\boldsymbol{x}))\right).$$

Therefore,

$$\text{rank}(J_f(\boldsymbol{x})) \leq \min(|\mathbb{Z}|, |\mathbb{X}|, |\mathbb{Y}|)$$

Proposition 1 implies that the Jacobian achieves its maximal possible rank almost everywhere. Thus, using Condition 2,

$$\text{rank}(J_f(\boldsymbol{x})) = \min(|\mathbb{Z}|, |\mathbb{X}|, |\mathbb{Y}|)$$

Therefore, $J_f = \frac{\partial f}{\partial \boldsymbol{x}} = r < |\mathbb{X}|, |\mathbb{Y}|$ if and only if $|\mathbb{Z}| = r$. This completes the proof.

**Linear Case:** We show that linear latent hierarchical models ([19]) are a special case of this theorem. If the causal relationsip between $y$ and $z$ is linear, we have: $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}] = \mathbb{E}[\boldsymbol{y}|\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}]]$. Therefore, $\mathbb{E}[\boldsymbol{z}|\boldsymbol{x}]]$ being a continuous differentiable function of $\boldsymbol{x}$ suffices and we do not require Condition 3 (ii).

□

### B.3 Proof of Theorem 2

**Theorem 2.** *Let $\mathcal{G}$ be a hierarchical latent causal model satisfying Condition 1. Let $\mathbb{Z}_X, \mathbb{Z}_Y \subseteq \mathbb{Z}$ be two disjoint subsets of latent variables in $\mathcal{G}$, i.e., $\mathbb{Z}_X \cap \mathbb{Z}_Y = \emptyset$. Let $\mathbb{X}$ be the set of observed variables that are d-separated by $\mathbb{Z}_X$ from all other observed variables in $\mathcal{G}$ and let $\mathbb{Y}$ be the set of observed variables that are d-separated by $\mathbb{Z}_Y$ from all other observed variables in $\mathcal{G}$. Then,*

$$r(\mathbb{Z}_X, \mathbb{Z}_Y) = r(\mathbb{X}, \mathbb{Y})$$

*Proof.* Since $\mathbb{Z}_X$ and $\mathbb{Z}_Y$ d-separate $\mathbb{X}$ and $\mathbb{Y}$ from all other variables, we know that $\mathbb{X}$ are the measured pure descendants of $\mathbb{Z}_X$, and $\mathbb{Y}$ are the measured pure descendants of $\mathbb{Z}_Y$. Therefore, if $\mathbb{Z}$ d-separates $\mathbb{Z}_X$ and $\mathbb{Z}_Y$, if $\mathbb{Z}$ d-separates $\mathbb{X}$ and $\mathbb{Y}$. Moreover, if $\mathbb{Z}$ d-separates $\mathbb{X}$ and $\mathbb{Y}$ then given our structure, it must d-separate $\mathbb{Z}_X$ and $\mathbb{Z}_Y$.

This completes the proof. □

### B.4 Proof of Lemma 1

**Lemma 1.** *Let $\mathcal{G}$ be a graph satisfying Conditions 1. A set of measured variables $\mathbb{S}$ are pure children of the same parent if and only if for any subset $\mathbb{T} \subseteq \mathbb{S}$, $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$.*

*Proof.* ($\Rightarrow$) Suppose the measured variables in $\mathbb{S}$ are pure children of the same parent node $\boldsymbol{p}$. For any subset $\mathbb{T} \subseteq \mathbb{S}$, $\mathbb{T}$ and $\mathbb{X} \setminus \mathbb{T}$ are d-separated by node $\boldsymbol{p}$. Therefore, we have $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$.

($\Leftarrow$) We prove the contrapositive. Suppose the union of parents of measured variables in $\mathbb{S}$ contains more than one node. Then there exist distinct parent nodes $\boldsymbol{p}$ and $\boldsymbol{p}'$ such that at least one of their respective children is in $\mathbb{S}$. We can choose $\mathbb{T}$ such that both $\mathbb{T}$ and $\mathbb{X} \setminus \mathbb{T}$ contains at least one pure child of $\boldsymbol{p}$ and one pure child of $\boldsymbol{p}'$. This choice is possible due to Condition 1, which states that all latent variables have at least two pure children.

For this choice of $\mathbb{T}$, we have $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) \geq 2$, as both $\boldsymbol{p}$ and $\boldsymbol{p}'$ are needed to d-separate the two sets. This contradicts the condition that $r(\mathbb{T}, \mathbb{S} \setminus \mathbb{T}) = 1$ for all $\mathbb{T} \subseteq \mathbb{S}$. □

### B.5 Proof of Lemma 2

**Lemma 2.** *Let $\mathcal{G}$ be a hierarchical latent causal model satisfying Conditions 1. Let $\mathbb{X} \subset \mathbb{V}$ be the set of observed variables. Under the faithfulness condition, for any observed variable $c \in \mathbb{X}$ and any set of latent variables $\mathbb{P} \subseteq \mathbb{Z}^1$, $c$ is a child of exactly the variables in $\mathbb{P}$ if and only if the following conditions hold:*

1. *$\forall \mathbb{S} \subseteq \mathbb{X}$ such that $|\mathbb{S} \cap Ch(z_i)| = 1$ for each $z_i \in \mathbb{P}$, where $Ch(z_i)$ denotes the set of pure children of $z_i$:*
$$r(\mathbb{S} \cup \{c\}, \mathbb{X} \setminus (\mathbb{S} \cup \{c\})) = \mathbb{P}$$
2. *The equality in condition (1) does not hold for any proper subset of $\mathbb{P}$.*

*Proof.* We will prove both directions of the if and only if statement.

($\Rightarrow$) Suppose $c$ is a child of exactly the variables in $\mathbb{P}$.

Let $\mathbb{S} \subseteq \mathbb{X}$ be any set such that $|\mathbb{S} \cap \text{Ch}(z_i)| = 1$ for each $z_i \in \mathbb{P}$, and let $\mathbb{T} = \mathbb{X} \setminus (\mathbb{S} \cup \{c\})$.

By the structure of the graph, $\mathbb{P}$ d-separates $\mathbb{S}$ from $\mathbb{T}$, and $\mathbb{P}$ also d-separates $\mathbb{S} \cup \{c\}$ from $\mathbb{T}$. This is because $c$ is a child of all nodes in $\mathbb{P}$, so including it with $\mathbb{S}$ doesn't create any new paths to $\mathbb{T}$ that aren't blocked by $\mathbb{P}$. Therefore, we have $r(\mathbb{S}, \mathbb{T}) = r(\mathbb{S} \cup \{c\}, \mathbb{T}) = |\mathbb{P}|$.

This satisfies condition (1). Condition (2) is satisfied because no proper subset of $\mathbb{P}$ contains all parents of $c$, so no proper subset of $\mathbb{P}$ can d-separate $\mathbb{S} \cup \{c\}$ from $\mathbb{T}$.

($\Leftarrow$) Now suppose conditions (1) and (2) hold. We will prove that $c$ is a child of exactly the variables in $\mathbb{P}$.

First, we show that $\mathbb{P}$ d-separates $c$ from all other variables in $\mathbb{X}$ not in $\mathbb{S} \cup \{c\}$.

Let $\mathbb{S}$ be as defined in condition (1), and $\mathbb{T} = \mathbb{X} \setminus (\mathbb{S} \cup \{c\})$. The equality in condition (1) implies:

$$r(\mathbb{S} \cup \{c\}, \mathbb{T}) = |\mathbb{P}|$$

Note that, $\mathbb{P}$ d-separates $\mathbb{S}$ and $\mathbb{T}$. Hence, $r(\mathbb{S}, \mathbb{T}) = |\mathbb{P}|$. Any set d-separating $\mathbb{S} \cup \{c\}$ from $\mathbb{T}$ must contain $\mathbb{P}$. Therefore, $r(\boldsymbol{s} \cup \{c\}, \boldsymbol{t}) \geq |\mathbb{P}|$. However, since they are equal, $\mathbb{P}$ d-separates $c$ from $\mathbb{T}$

This implies that the parents of $c$ must be a subset of $\mathbb{P}$, as any path from $c$ to $\mathbb{T}$ not going through $\mathbb{P}$ would violate the d-separation.

Now, condition (2) states that this equality doesn't hold for any proper subset of $\mathbb{P}$. This means that every variable in $\mathbb{P}$ is necessary for the d-separation. If any variable in $\mathbb{P}$ were not a parent of $c$, then we could remove it and still maintain the d-separation, contradicting condition (2).

Therefore, $c$ must be a child of exactly the variables in $\mathbb{P}$. □

### B.6 Proof of Lemma 3

**Lemma 3.** *Let $\mathbb{G}$ be a graph satisfying Conditions 1. A measured variable $c$ has no parent if and only if $r(\{c\}, \mathbb{X} \setminus \{c\}) = 0$.*

*Proof.* We will prove both directions of the if and only if statement.

($\Rightarrow$) Suppose $c$ has no parent.

In this case, $c$ is independent of all other variables in the graph. Therefore, $r(\{c\}, \mathbb{X} \setminus \{c\}) = 0$.

($\Leftarrow$) Suppose $r(\{c\}, \mathbb{X} \setminus \{c\}) = 0$.

This means that no latent variables are needed to render $c$ independent of all other observed variables. In other words, $c$ is already independent of all other observed variables.

Now, suppose for the sake of contradiction that $c$ has a parent $z$. By Condition 1, $z$ must have at least two pure children, one of which could be $c$, and let's call the other one $x$. Then $c$ and $x$ would be dependent through their common parent $z$, contradicting the independence of $c$ from all other observed variables.

Therefore, $c$ cannot have any parent. □

### B.7 Proof of Theorem 3

**Theorem 3.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hierarchical latent causal model satisfying Condition 1. Let $\boldsymbol{M}$ be the binary adjacency matrix representing the structure of $\mathcal{G}$. Let data $\mathbb{X}$ be generated according* *to the structural equation model defined in Equation 1. Given a function* $r(\mathbb{S}, \mathbb{T})$ *which outputs the minimum number of latent variables which d-separate any two sets measured sets* $\mathbb{S}$ *and* $\mathbb{T}$*,* $\boldsymbol{M}$ *is identifiable up to the permutation of the latent variables.*

*Proof.* We prove Theorem 3 using the principle of mathematical induction on the number of layers. We begin with three lemmas that we require for our proof.

We prove this theorem by induction on the number of layers in the hierarchical latent causal model.

**Base case:** Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hierarchical latent causal model with one latent layer, i.e., $\mathbb{Z} = \mathbb{Z}^1$.

We begin by identifying the pure children of each latent variable in $\mathbb{Z}^1$. By Lemma 1, for any set of measured variables $\mathbb{S} \subseteq \mathbb{X}$, if $r(\mathbb{T}, \mathbb{X} \setminus \mathbb{T}) = 1$ for all $\mathbb{T} \subseteq \mathbb{S}$, then all variables in $\mathbb{S}$ are pure children of the same parent. For all $|\mathbb{T}| = 1$, this trivially holds. Using Theorem 1, we can exhaustively check this condition for all subsets $|\mathbb{T}| > 1$ of $\mathbb{X}$ to identify all sets of pure children.

Next, we identify the parents of non-pure children using Lemma 2. For each observed variable $c \in \mathbb{X}$ that is not identified as a pure child in the previous step, we determine its parents by verifying the conditions stated in Lemma 2 for all possible subsets of $\mathbb{Z}^1$. For most cases, $|\mathbb{S} \cup \{c\}|, |\mathbb{X} \setminus (\mathbb{S} \cup \{c\})| > |\mathbb{P}|$ which allows us to use Theorem 1. For cases, where this does not hold, it implies $\mathbb{P} = \mathbb{Z}^1$. In this case, Lemma 2 can be applied for all other subsets of $\mathbb{Z}^1$ and if none of them satisfy the conditions, the set of parents has to be $\mathbb{Z}^1$.

Through these two steps, we fully identify the structure between $\mathbb{Z}^1$ and $\mathbb{X}$, thus recovering the binary adjacency matrix $\boldsymbol{M}$ for the one-layer model.

**Inductive step:** Assume the theorem holds for all models with $L - 1$ layers, where $L > 1$. We will prove it holds for models with $L$ layers.

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a hierarchical latent causal model with $L$ layers. We first identify the structure between $\mathbb{Z}^1$ and $\mathbb{X}$ using the same process as in the base case, applying Lemmas 1 and 2.

Let $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ be the subgraph of $\mathcal{G}$ obtained by removing all measured variables $\mathbb{X}$. We claim that $\mathcal{G}'$ satisfies Condition 1. Each latent variable in $\mathbb{Z}^2$ has at least two pure children in $\mathcal{G}'$, and these belong to $\mathbb{Z}^1$. Moreover, the equal path length condition is preserved since the path from any latent to any variable in $\mathbb{Z}^1$ is one less than the path to any variable in $\mathbb{X}$. Some variables $\mathbb{Z}^1$ may not have any parent. We can identify those using Lemma 3.

To determine the d-separation relations between variables in $\mathbb{Z}^1$, we utilize Theorem 2. For any two subsets $\mathbb{Z}_X, \mathbb{Z}_Y \subseteq \mathbb{Z}^1$, let $\mathbb{X}$ and $\mathbb{Y}$ be sets of their respective pure children. By Theorem 2, we have $r(\mathbb{Z}_X, \mathbb{Z}_Y) = r(\mathbb{X}, \mathbb{Y})$, allowing us to infer the d-separation relations within $\mathbb{Z}^1$.

Now, $\mathcal{G}'$ is a hierarchical latent causal model with $L - 1$ layers that satisfies the conditions of the theorem. By the induction hypothesis, we can identify the structure of $\mathcal{G}'$, recovering the corresponding part of the binary adjacency matrix $\boldsymbol{M}$.

Therefore, by the principle of mathematical induction, the theorem holds for hierarchical latent causal models with any number of layers.

□

### B.8 Proof of Lemma 4

**Lemma 4.** *Consider a DAG* $\mathcal{G}$ *with a binary adjacency matrix* $\boldsymbol{M}$*.* $\mathcal{G}$ *satisfies Condition 1 (i) if and only if:*

$$\|\boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:})\|_1 \geq 2 \quad \forall i. \tag{15}$$

*Proof.* To prove this lemma, we need to demonstrate that the given condition on the adjacency matrix $\boldsymbol{M}$ holds if and only if each latent variable has at least two pure children, as stated in Condition 1.

Let's first analyze the term inside the $\ell_1$ norm:

$$\boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:})$$

The $k$th element of this vector is given by:

$$M_{ik} \cdot \prod_{j \neq i} (1 - M_{jk})$$

This product equals 1 if and only if $M_{ik} = 1$ and $M_{jk} = 0$ for all $j \neq i$. In other words, this term is 1 if and only if the vertex $v_k$ is a child of $v_i$ and not a child of any other $v_j$ ($j \neq i$). Hence, this product identifies whether $v_k$ is a pure child of $v_i$.

The $\ell_1$ norm, $\|\cdot\|_1$, sums up all these terms, meaning it counts the number of pure children of $v_i$.

Now, according to Condition 1, each latent variable $v_i$ must have at least 2 pure children. Therefore, the condition:

$$\|\boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:})\|_1 \geq 2 \quad \forall i$$

ensures that each $v_i$ has at least 2 pure children, satisfying Condition 1.

Conversely, if each latent variable $v_i$ has at least 2 pure children, the sum $\left\|\boldsymbol{M}_{i,:} \odot \prod_{j \neq i} (1 - \boldsymbol{M}_{j,:})\right\|_1$ must be at least 2 for each $i$, proving the equivalence.

Thus, the lemma is proven. □

# C Experimental Details

In this section, we provide a detailed explanation of our experimental setup, model architecture, training procedure, hyperparameter settings, and evaluation metrics used in the paper.

## C.1 Synthetic Data

We follow the same procedure and hyperparameter values for all four graphs.

**Data generation:** For the ground truth graphs, the functions in Equation 1 are modeled using a linear transformation of the input followed by a LeakyReLU activation function with $\alpha = 0.2$. The weights for the linear transformation are uniformly sampled from $[-5, -2] \cup [2, 5]$. Exogenous noise is sampled from $[-\alpha, \alpha]$ where $\alpha$ is sampled from $[-3, -1] \cup [1, 3]$.

**Model architecture:** We use a VAE to learn our causal structure as described in Section 4. The model consists of several components. The VAE encoder is a two-hidden-layer fully connected neural network with 64 and 32 hidden neurons, followed by ReLU activations. The encoder outputs both the mean ($\mu$) and the log variance ($\log \sigma^2$) for the latent variables. For the decoder, each function in Equation 4 is modeled using a one-hidden-layer fully connected neural network with 32 hidden neurons. The masking matrices have shape $\left\lfloor \frac{|\mathbb{X}|}{2^{i+1}} \right\rfloor \times \left\lfloor \frac{|\mathbb{X}|}{2^{i}} \right\rfloor$ since Condition 1 allows a maximum of $\left\lfloor \frac{|\mathbb{X}|}{2^{i}} \right\rfloor$ latent variables in $\mathbb{Z}^i$. We use ReLU activation for all neural networks.

**Training Procedure:** We use mean squared error as the reconstruction loss. We calculate the $\mathcal{L}_{\text{ind}}(\boldsymbol{\epsilon}) = \text{KL}(p(\boldsymbol{\epsilon}) \| \prod_j \prod_i p(\varepsilon_{z_j^i}) \prod_j p(\varepsilon_{x_j}))$ using a method similar to MINE ([4]). We warm up the MINE model for 100 epochs before training. Our model is trained using the Adam optimizer with a learning rate of $1 \times 10^{-3}$ for 400 epochs. We use a batch size of 32. For Gumbel-Softmax, we set the temperature to 1.0 throughout training. The coefficient for $\mathcal{L}_{\text{ind}}(\boldsymbol{\epsilon})$ is set to 10. $\lambda_2$ is set to 1e-4 and $\lambda_3$ is set as $10^{-3+\frac{\text{epoch}}{100}}$. Since the training objective is non-convex, we may get suboptimal solutions [31]. Therefore, we run the model with 10 random initializations and select the one with the lowest loss.

**Baselines:** For [24], we use the code shared with us by the authors. For [19] and [42], we use the publicly available implementation. Default hyperparameters were used for all methods. We attempted to use FOFC [26] as a baseline, however an error occurred for our data since it does not satisfy the conditions for their code to run.

**Evaluation Metrics:** We evaluate our model using the Structural Hamming Distance (SHD) and F1 score between the learned adjacency matrix and the ground truth. We train each run three times for different seed values and report the mean and standard deviation across the three runs of two graphs. Since the latent graph may only be recovered up to a permutation of the latent variables, we calculate the SHD over all possible $|\mathbb{Z}|!$ permutations of the estimated graph and select the lowest SHD. Since the evaluation time is $O(n!)$ in the number of latent variables, evaluating methods becomes intractable for very large graphs. The time was calculated in seconds. For the standard deviation of time, we report the mean of the standard devation of each graph since time can vary a lot based on the size of the graph.

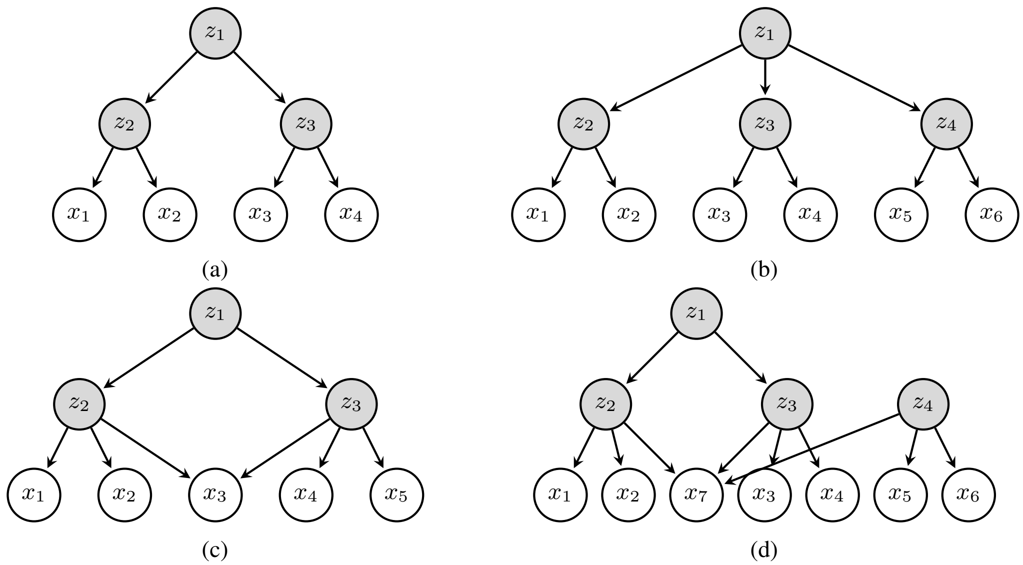



Figure 3: Ground truth causal graphs for Synthetic Experiments. (a) and (b) are trees (only one path between any two nodes). (c) and (d) allow v-structures (multiple paths between two nodes)

### C.2 Image Data

**Model Architecture:** The proposed Hierarchical VAE model consists of a convolutional encoder, a hierarchical latent structure, and a transposed convolutional decoder. The convolutional encoder has two convolution layers (32 and 64 filters, 3x3 kernels, stride 2) followed by a fully connected layer. Each structural equation (Equation 4) is modeled using a neural network with three hidden layer. The first two layers are shared to reduce the number of parameters. The convolutional decoder reconstructs images from the final latent layer. The decoder consists of a fully connected layer that maps the final latent representation of dimension 49 to a 1568-dimensional space. This output is then reshaped to a 32-channel 7x7 feature map. We then have two transposed convolutional layers with 32 and 16 filters respectively (both using 3x3 kernels, stride 2, padding 1, and output padding 1), and a final transposed convolutional layer that reconstructs the image. We initialize $\boldsymbol{M}$ with three layers with 10, 20 and 49 nodes in each of the three layers. We use ReLU activation for all neural networks.

**Training:** We train the model on a subset of 10,000 images. The model was trained for 300 epochs. We use batch size of 256 and do not use MINE to enforce independence between the exogenous variables. We find it makes little difference to the final output. The temperature for gumbel softmax starts at 100 and exponentially decreases to 0.1 at 120 epochs and then stays constant. $\lambda_2$ is 0.03 and $\lambda_3$ is exponentially increased from $10^{-3}$ to 10 at 100 epochs and then stays constant. We used a Adam optimizer with learning rate 1e-3.

**Visualization:** Figure 4 displays the complete causal graph constructed from the MNIST dataset. Note that due to non-convexity, we could not achieve zero loss for the pure children constraint term. Hence, the learnt graph does not exactly satisfy Conditions 1. To interpret the semantics of each latent feature, we perform targeted interventions designed to isolate their individual effects. Specifically, for each latent variable at the topmost layer, we set its ancestral nodes to values 5 standard deviations above or below their means, while keeping the remaining variables at their mean values. We then intervene on the current node by setting its value to the mean, effectively neutralizing its direct influence, and observe the resulting changes in the generated images. This procedure allows us to visualize and understand the specific contribution of each latent variable to the overall image structure.

By comparing the images before and after the intervention, we can discern the unique effects attributable to each latent feature. Table 3 presents these visualizations for each feature. For each feature, the first image shows the output when the top latent variable is set 5 standard deviations below the mean; the second image shows the result when, in addition, the target feature is intervened upon and set to the mean; the third image displays the output when the top latent variable is set 5 standard deviations above the mean; and the fourth image shows the result when the target feature is intervened upon and set to the mean under this condition.

For Figure 2b, we adopt a different methodology to visualize the influence of latent variables. Starting from the topmost layer of the causal graph, we traverse downward through each subsequent layer. At each node, we randomly assign its value to be either five standard deviations above or below its mean. This stochastic intervention allows us to observe the cumulative effects of these variations as they propagate through the graph.

**CMNIST details:** For the coloured MNIST dataset, we have around 12,000 training samples and 2,000 test samples. In this dataset, the training set consists of digits 0 and 1, colored either red or green. The color acts as a cause for the digit, with $P(\text{digit} = 0|\text{color} = \text{red}) = 0.9$ and $P(\text{digit} = 0|\text{color} = \text{green}) = 0.1$ in the training set. These probabilities are reversed in the test set. We further evaluate on an additional test set where all digits are colored blue. Samples from these sets are shown in Figure 2c. Although the correlation between color and digit varies across datasets, the causal relationship between the digit's image features and its label remains unchanged.

Since we do not aim to visualize the images, we downsample them to 14x14 and train our model. We train our model for 50 epochs with early stopping with patience 3. After training the latent hierarchical model, we train a logistic regression classifier to predict the digit from the latent representations. The coefficient of the L1 regularization is 10. For all the baselines, we train the model for 50 epochs with early stopping with patience 3. For all models, we used a Adam optimizer with learning rate 1e-3.

We compare our method with standard Autoencoders, Variational Autoencoders (VAEs) [22], and $\beta$-VAEs [18] across different values of $\beta$. For each baseline, a logistic regression classifier is trained on the learned latent representations. All methods are evaluated over three random trials, with the mean and standard deviation reported to assess performance consistency.

# D Discussion

## D.1 Condition 1

In Figure 5, we see two examples of causal graphs which violate Condition 1. Figure 5a violates the pure children condition since each latent does not have two pure children. Figure 5b violates the Condition 1(ii) since $d(z_1, x_4) = 2 \neq 1 = d(z_1, x_3)$. While Condition 1(ii) may not hold in all cases, it is a reasonable assumption to make for image data. Several prior works Vahdat & Kautz [41], Kong et al. [25] have effectively modeled images using a multi-level latent hierarchical structure.

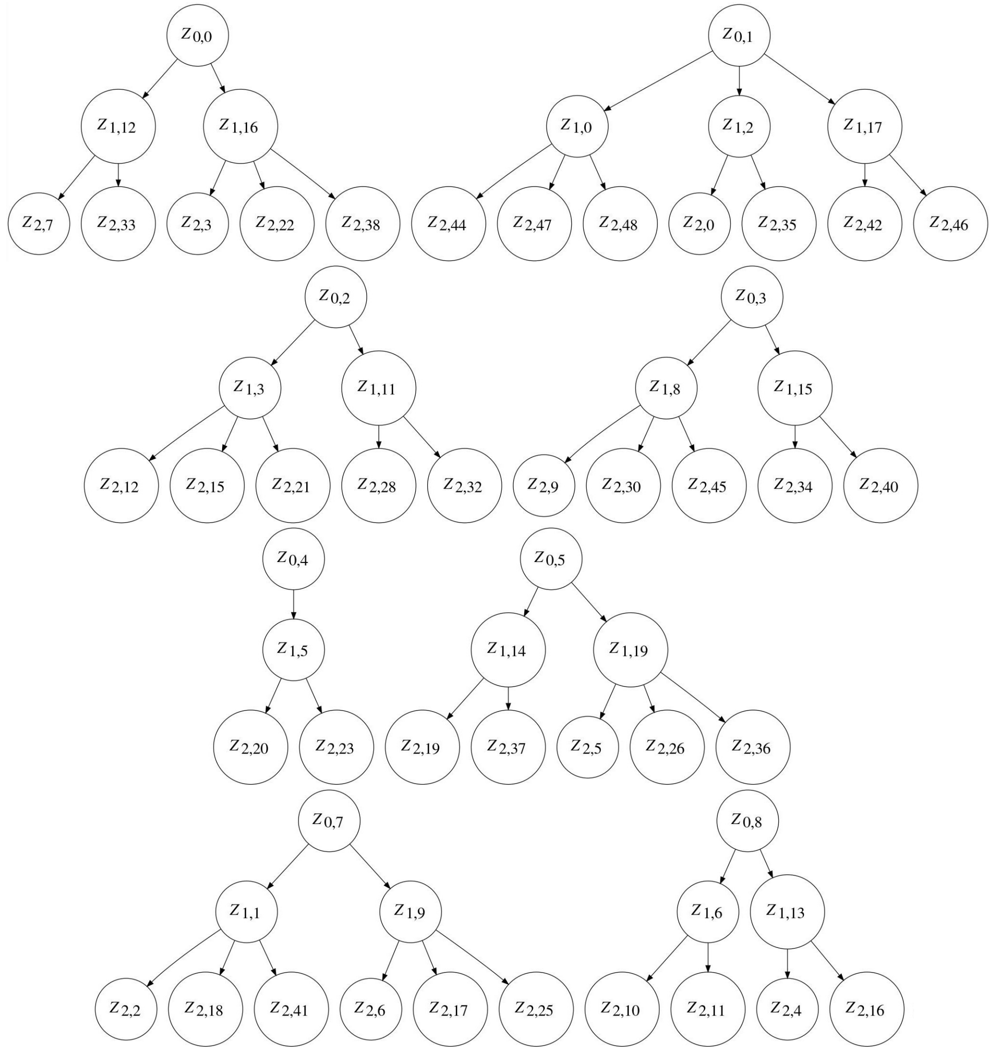
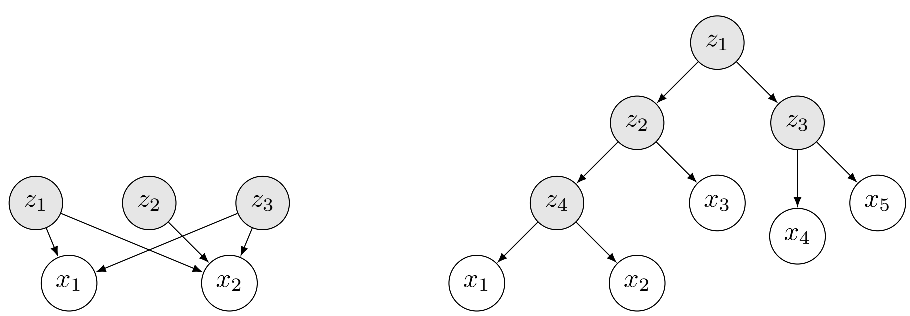



Figure 4: Latent causal graph for the MNIST Dataset. $z_{i,j}$ denotes the $j^{\text{th}}$ latent variable in $\mathbb{Z}^i$.

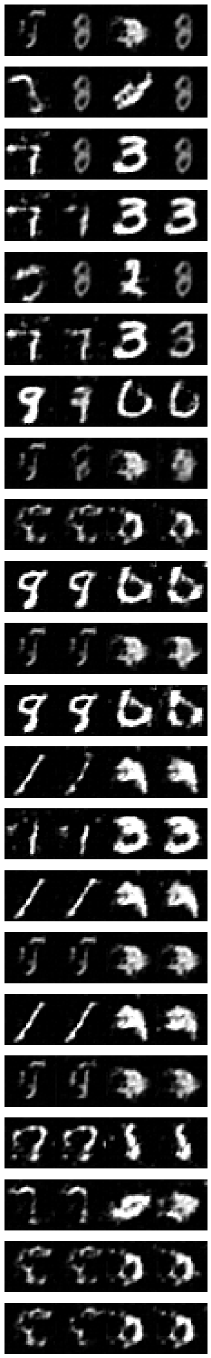
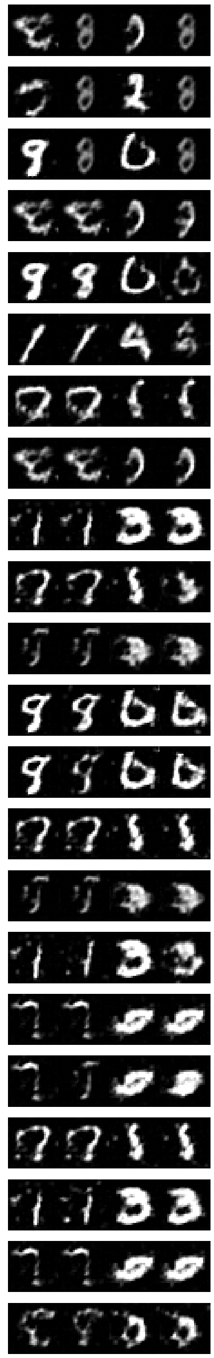
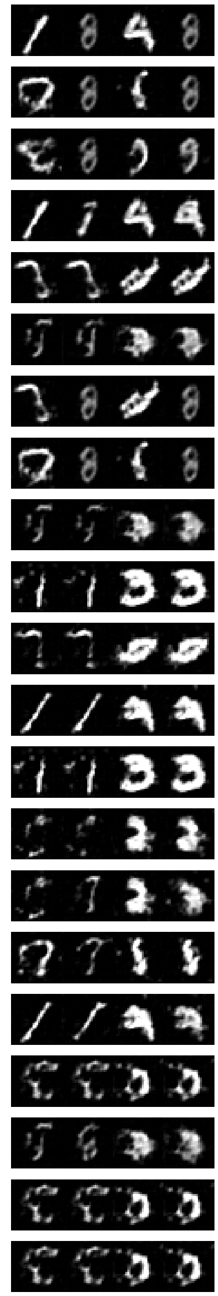



(a) Causal structure which violates Condition 1(i)

(b) Causal structure which violates Condition 1(ii)

Figure 5: Two examples of causal structures which violate Condition 1

Table 3: Visualization of MNIST layer dimensions. For each feature, he first image shows the output when the top latent variable is set 5 standard deviations below the mean; the second image shows the result when, in addition, the target feature is intervened upon and set to the mean; the third image displays the output when the top latent variable is set 5 standard deviations above the mean; and the fourth image shows the result when the target feature is intervened upon and set to the mean under this condition.

| **Dim** | **Image** | **Dim** | **Image** | **Dim** | **Image** |
|---|---|---|---|---|---|
| $z_0^0$ |  | $z_1^0$ |  | $z_2^0$ |  |
| $z_3^0$ | | $z_4^0$ | | $z_5^0$ | |
| $z_7^0$ | | $z_8^0$ | | $z_0^1$ | |
| $z_1^1$ | | $z_2^1$ | | $z_3^1$ | |
| $z_5^1$ | | $z_6^1$ | | $z_8^1$ | |
| $z_9^1$ | | $z_{11}^1$ | | $z_{12}^1$ | |
| $z_{13}^1$ | | $z_{14}^1$ | | $z_{15}^1$ | |
| $z_{16}^1$ | | $z_{17}^1$ | | $z_{19}^1$ | |
| $z_0^2$ | | $z_2^2$ | | $z_3^2$ | |
| $z_4^2$ | | $z_5^2$ | | $z_6^2$ | |
| $z_7^2$ | | $z_8^2$ | | $z_9^2$ | |
| $z_{10}^2$ | | $z_{11}^2$ | | $z_{12}^2$ | |
| $z_{15}^2$ | | $z_{16}^2$ | | $z_{17}^2$ | |
| $z_{18}^2$ | | $z_{19}^2$ | | $z_{20}^2$ | |
| $z_{21}^2$ | | $z_{22}^2$ | | $z_{23}^2$ | |
| $z_{24}^2$ | | $z_{25}^2$ | | $z_{26}^2$ | |
| $z_{28}^2$ | | $z_{30}^2$ | | $z_{32}^2$ | |
| $z_{33}^2$ | | $z_{34}^2$ | | $z_{35}^2$ | |
| $z_{36}^2$ | | $z_{37}^2$ | | $z_{38}^2$ | |
| $z_{40}^2$ | | $z_{41}^2$ | | $z_{42}^2$ | |
| $z_{44}^2$ | | $z_{45}^2$ | | $z_{46}^2$ | |
| $z_{47}^2$ | | $z_{48}^2$ | | | |