C-iVAE: Causal Structure-Aware Identifiable VAE for Tabular Data with Limited Environments

Zejiang Wang
2023200910@ruc.edu.cn
School of Labor and Human Resources
Renmin University of China
Beijing, China

Feng Zhou*
feng.zhou@ruc.edu.cn
Center for Applied Statistics and School of Statistics
Renmin University of China
Beijing, China

Abstract

Identifiable Variational Autoencoders (iVAEs) provide theoretical guarantees for decoupled representation learning, but their requirement for auxiliary environments grows linearly with latent dimension ($O(dk)$), which is often impractical in real-world scenarios. We propose the Causal Structure-Aware Identifiable Variational Autoencoder (C-iVAE), extending the identifiability framework to static tabular data domains where causal graphs are often known or partially accessible. By leveraging the hierarchical structure of directed acyclic graphs (DAGs), C-iVAE reduces computational demands to $O(rk)$, where $r \ll d$ denotes the number of root nodes. Our core hierarchical identifiability theory proves that non-root nodes can inherit sufficient variability from causal parents, eliminating the need for explicit environmental constraints. Beyond identifiability, C-iVAE's causally aware generation ensures learned representations naturally adhere to structural constraints, yielding logically consistent outputs and enhanced representational capacity. We employ a graph neural network encoder to learn asymmetric structural embeddings, distinguishing nodes sharing identical Markovian blankets. Experiments demonstrate C-iVAE's exceptional robustness in intervention generalization and DAG noise settings, achieving significant performance gains on real-world datasets.

CCS Concepts

• **Computing methodologies** → *Neural networks*; **Learning latent representations**; **Causal reasoning and diagnostics**.

Keywords

Identifiability, Variational Autoencoder, Causal Representation Learning, Low-Resource Learning, Tabular Data

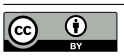
ACM Reference Format:

Zejiang Wang and Feng Zhou. 2026. C-iVAE: Causal Structure-Aware Identifiable VAE for Tabular Data with Limited Environments. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817977

Resource Availability:

The source code and artifacts of this paper have been made publicly available at GitHub (https://github.com/shudiaoh/c-ivae) and archived at Zenodo (https://doi.org/10.5281/zenodo.20497472).

*Corresponding author.



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817977

1 Introduction

Synthesizing high-fidelity data is central to modern data-centric AI, facilitating privacy preservation [7, 18], data augmentation for low-resource domains [12, 13], and fairness within long-tail distributions [2, 34]. As foundational models scale, both the quality [39] and quantity [32] of training corpora increasingly constrain downstream performance and generalization capabilities. Consequently, generating high-quality tabular data, which remains the most ubiquitous modality in enterprise and scientific applications from finance to healthcare, constitutes a critical research imperative for sustainable and trustworthy AI development.

Tabular data modeling presents distinct challenges arising from heterogeneous feature types, multi-modal marginal distributions, and complex higher-order feature interactions that defy simple statistical characterization [4, 14]. State-of-the-art generative models, including TabDDPM [25], TabSyn [38], and TabDiff [29], primarily focus on approximating low-order statistical properties such as marginal distributions and pairwise correlations via density estimation techniques. While effective for replicating training distributions, these approaches often overlook the underlying generative mechanisms and causal structures governing the data. Consequently, the generated samples, despite achieving statistical fidelity on in-distribution metrics, may violate fundamental semantic constraints [5] or fail to capture necessary causal dependencies, thereby limiting their reliability in safety-critical downstream applications such as medical diagnosis or financial risk assessment.

A fundamental limitation is the lack of model identifiability, undermining interpretability and control. Standard deep generative models like Variational Autoencoders (VAEs) [20] assume i.i.d. data, lacking guarantees for recovering true latent factors. Locatello et al. [26] formally established that without proper inductive biases, latent representations are identifiable only up to arbitrary nonlinear transformations. This has practical implications: in scientific discovery, if latent variables corresponding to physical quantities are recovered with unknown transformations, downstream analysis becomes invalid. Such entanglement prevents learning causally meaningful representations, hindering control (e.g., stylized intervention) and invalidating counterfactual reasoning. Furthermore, by optimizing observed data likelihood $P(X)$, these models capture spurious correlations, causing performance collapse under distribution shifts, as reliability requires modeling the underlying Structural Causal Model (SCM) which remains invariant across interventions.

The Identifiable Variational Autoencoder (iVAE) framework [19] addresses this limitation by leveraging auxiliary variables to achieve identifiability in nonlinear independent component analysis (ICA). However, the theoretical condition requires auxiliary variable complexity scaling with $O(dk)$, where d and k denote the latent dimension and sufficient statistic dimension, respectively. Real-world

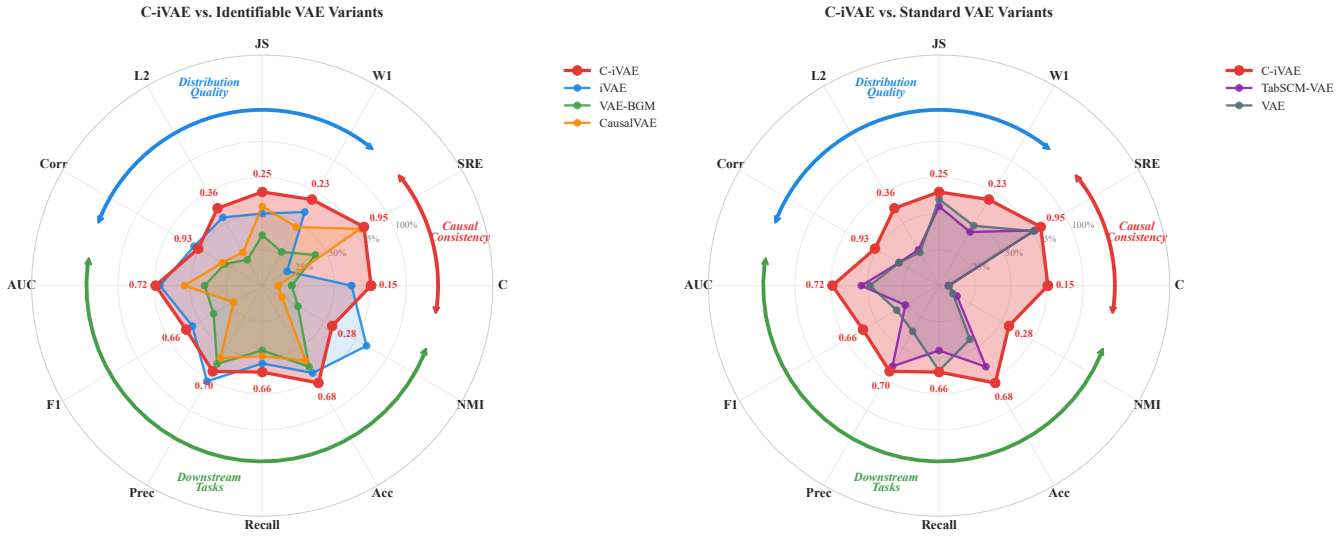


Figure 1: Multi-Dimensional Performance Comparison. C-iVAE (Red) achieves superior or competitive performance across all 12 metrics covering Causal Consistency, Distribution Quality, and Downstream Utility, demonstrating a balanced trade-off compared to VAE, VAE-BGM, and iVAE baselines.

tabular datasets rarely possess such high-dimensional, fine-grained environmental labels, rendering standard iVAE practically inapplicable in environment-scarce settings commonly encountered in industrial and clinical domains.

We propose **Causal-iVAE** (C-iVAE) to resolve this limitation by using the intrinsic causal structure of tabular data as an inductive bias. We model the generative mechanism using Directed Acyclic Graphs (DAGs) and encode topological dependencies into auxiliary embeddings via Graph Neural Networks (GNNs). We establish a hierarchical identifiability theory proving that incorporating causal geometric structure reduces the requirement for external environments from $O(dk)$ to $O(rk)$ (where r is the number of root nodes, and $r \ll d$). This ensures identifiability even when external environmental information is limited. Furthermore, the GNN-based structural encoding provides robustness to DAG misspecification, maintaining strong performance even when up to 30% of edges are erroneous. This approach links deep generative learning with causal discovery by embedding structural constraints. It provides a framework where neural networks are guided by causal diagrams, ensuring generated data preserves valid causal mechanisms alongside statistical fidelity.

Our contributions are summarized as follows:

(1) **Theoretical Guarantee**: We establish Theorem 1 as our core theoretical contribution, proving that integrating causal DAG structure significantly reduces the complexity of auxiliary variables required for identifiability. This aligns with the paradigm of modern causal representation learning (e.g., CITRIS, ILCM), providing rigorous theoretical assurance in environment-scarce scenarios.

(2) **Structural Encoding**: We develop a framework using Markov blankets and GNNs to translate causal topology into node-specific auxiliary embeddings, enabling precise modulation of latent priors.

(3) **Empirical Performance**: Experiments on synthetic and eight real-world datasets demonstrate that C-iVAE achieves superior performance over state-of-the-art baselines, including CausalVAE, in distribution fidelity, downstream utility, causal consistency, and robustness to structural noise.

(4) **Ablation Analysis**: We verify the essential role of structural constraints in enhancing model generalization and robustness against spurious correlations.

2 Related Work

This section reviews related work from three perspectives: tabular data generation, identifiability of latent variable models, and causal representation learning.

2.1 Tabular Data Generation

Generative modeling for mixed-type tabular data faces inherent challenges from complex marginal distributions and categorical-continuous feature interactions. Early works like CTGAN [35] use mode-specific normalization to handle multi-modal continuous features. Diffusion-based models, including TabDDPM [25], TabSyn [38], and TabDiff [29], improve fidelity by modeling heterogeneous features either separately or jointly through specialized noise schedules. However, these methods rely on observation-space statistical fitting, neglecting the underlying data generation mechanisms, which fundamentally limits out-of-distribution generalization. While TabPFN [16] leverages SCM priors for in-context learning and CA-GAN [27] incorporates causal rewards, they suffer from scalability issues or lack explicit latent structural modeling. Recently, Large Language Models (LLMs) have also been adapted for tabular generation [5], yet they often hallucinate correlations absent in the training data. This paper addresses these gaps by

embedding causal structure directly into the latent space to enhance generalizability, robustness against distribution shifts, and computational scalability.

2.2 Identifiability of Latent Variable Models

Identifiability of latent variable models refers to whether model parameters and latent variables can be uniquely determined from observed data, a property critical for understanding the underlying data generation mechanism and enabling valid causal inference. Locatello et al. [26] show that unsupervised learning cannot achieve identifiability without appropriate inductive biases. iVAE [19] proves that, given auxiliary variables, latent variables in nonlinear mixture models are identifiable up to affine transformations. Subsequent studies demonstrate that identifiability can also be achieved without auxiliary variables through mixture model priors [33] or sparsity constraints [21]. Recent work further relaxes identifiability conditions under more general settings [6, 31]. Our work introduces a hierarchical identifiability mechanism that provides more relaxed theoretical guarantees for identifiability specifically tailored to large-scale tabular data generation scenarios where traditional conditions are overly restrictive.

2.3 Causal Representation Learning

Although deep generative models excel at capturing statistical correlations within the training distribution, their generalization degrades significantly on out-of-distribution data due to spurious correlations learned by the model [24]. Causal representation learning seeks to recover low-dimensional representations with causal semantics from complex, high-dimensional observations, thereby enabling counterfactual reasoning and interventional prediction [36]. CausalGAN [22] pioneered this direction by preserving causal structure through conditional GANs but lacked explicit inference mechanisms. Subsequent research, such as CausalVAE [36], explicitly models causal constraints among latent variables using structural causal models. However, CausalVAE relies on a masking layer that behaves more like a soft regularization term than a hard constraint, often leading to 'structural leakage' where intended independencies are violated during training. Furthermore, it lacks the theoretical identifiability guarantees provided by the iVAE framework. Recent advances in modeling graph structures, such as DAG-GNN [37], have demonstrated the efficacy of continuous optimization for discovering and encoding topological relationships. Building on these foundations, we innovatively apply hierarchical constraints directly within the latent space, offering a mathematically rigorous path to identifiability that CausalVAE lacks, achieving stronger guarantees and enhanced robustness compared to standard DAG-based priors.

3 Preliminaries

This section first reviews the theoretical framework of iVAE, establishing the conditions under which latent variables become recoverable, and then introduces Structural Causal Models and their connection to graph structure encoding. These preliminaries lay the theoretical foundation for the proposed C-iVAE model.

3.1 Identifiable Variational Autoencoders

Identifiability of latent variable models refers to whether the parameters of the generative model can be uniquely recovered from observations. For standard VAEs, the prior $p(z)$ does not depend on any auxiliary information. Consider a simple example: let $z \sim \mathcal{N}(0, I)$ and $z' = Qz$ where Q is an orthogonal matrix. Due to the rotational invariance of Gaussian distributions, z' also follows $\mathcal{N}(0, I)$, i.e., $p(z) = p(z')$. Thus, the model cannot distinguish z from z', and learned representations may undergo arbitrary orthogonal transformations without detection. Locatello et al. [26] formally prove this unidentifiability holds for any invertible transformation g.

iVAE [19] breaks this symmetry by introducing auxiliary variables u to modulate the prior distribution. The generative model is

$$p_\theta(x, z|u) = p_\theta(x|z) \cdot p_\theta(z|u), \tag{1}$$

with an exponential family conditional prior:

$$p_\theta(z_i|u) = h(z_i) \exp\left[\mathbf{T}(z_i)^\top \boldsymbol{\lambda}_i(u) - A_i(u)\right], \tag{2}$$

where $\mathbf{T}(z_i)$ denotes sufficient statistics, $\boldsymbol{\lambda}_i(u)$ maps auxiliary variables to natural parameters, and $A_i(u)$ is the normalization constant.

The core theorem of iVAE states that if the mixing function is invertible, sufficient statistics are linearly independent, and there exist enough auxiliary values such that the natural parameter variation matrix has full rank, then true latent variables z^* and learned $\hat{z}$ differ only by an invertible linear transformation, i.e., $\hat{z} = Az^* + b$. The key proof idea is to use the variability of auxiliary variables to constrain the prior parameter space and exclude non-trivial transformations through the full-rank condition. See Appendix for detailed proofs.

However, iVAE has three limitations: (1) it requires $O(dk)$ distinct environments where d is the latent dimension and k is the sufficient statistic dimension, requiring at least 41 environments for $d = 10$. This scaling law makes the method prohibitive for high-dimensional tabular data where such fine-grained environmental meta-data is unavailable or expensive to collect; (2) it assumes $z_i \perp z_j \mid u$, contradicting causal DAG dependencies where latent variables are inherently coupled through structural equations; (3) the global auxiliary variable u cannot achieve node-level modulation, treating all latent dimensions uniformly despite their distinct positions in the causal hierarchy. These limitations motivate C-iVAE to seek a more data-efficient and structurally aware path to identifiability.

3.2 Structural Causal Models and Graph Structure Encoding

A Structural Causal Model (SCM) $\mathcal{M} = (\mathcal{G}, \{f_i\}, P_U)$ consists of a causal DAG $\mathcal{G}$, structural equations $z_i = f_i(\mathrm{Pa}(z_i), u_i)$, and exogenous noise distribution P_U. By the causal Markov condition, the joint distribution factorizes as

$$p(z|\mathcal{G}) = \prod_{i=1}^{d} p(z_i|\mathrm{Pa}(z_i)). \tag{3}$$

The Markov blanket $\mathrm{MB}(i) = \mathrm{Pa}(i) \cup \mathrm{Ch}(i) \cup \mathrm{CoPa}(i)$ is the minimal set rendering z_i conditionally independent of all other variables. For DAGs without non-trivial automorphisms, it uniquely identifies each node.

We encode DAG structure into node-level auxiliary variables u_i via GNNs with positional encodings (random walk, degree, PageRank, topological depth). The embedding $u_i = \text{GNN}(\mathcal{G})[i]$ satisfies $u_i \neq u_j$ for structurally inequivalent nodes.

4 Method

This section presents the complete model design of C-iVAE. We first describe the DAG-structured prior that enables hierarchical identifiability (Section 4.1), then introduce the structure encoder that produces node-specific embeddings (Section 4.2), followed by the causal decoding and generation process (Section 4.3), and finally the variational inference framework (Section 4.4). The overall architecture is illustrated in Figure 2, highlighting the interplay between structural encoding and causal generation to ensure both topological fidelity and generative flexibility.

The design of C-iVAE addresses the prohibitive auxiliary variable requirements of existing identifiable VAEs. Our core insight is that dependencies within the causal structure serve as 'internal environments,' reducing reliance on external labels. This is realized through a DAG-structured prior and a specialized structure encoder. The DAG-structured prior exploits causal asymmetry: exogenous root nodes require external environments for symmetry breaking, while non-root nodes condition on their causal parents. As variability propagates from roots to children, each child inherits a unique 'context' acting as a dynamic auxiliary variable, eliminating the need for explicit external environments. This hierarchical propagation reduces environment complexity from $O(dk)$ to $O(rk)$. Complementing this, the GNN-based structure encoder provides node-specific embeddings u_i tailored to each node's topological role. By incorporating positional encodings, the GNN ensures even structurally similar nodes receive identifiable signatures. This combination enables C-iVAE to learn causally meaningful, identifiable representations in complex tabular settings.

C-iVAE achieves hierarchical identifiability by partitioning latent variables into topological layers according to the causal DAG. Root nodes require external environment variation e for symmetry breaking, while non-root nodes inherit sufficient variability from their causal parents $z_{\text{Pa}(i)}$, reducing environmental complexity from $O(dk)$ to $O(rk)$ where $r \ll d$. By recursively conditioning on parents, the prior effectively partitions the latent space into independent causal mechanisms (c-components), where each component is locally identifiable relative to its ancestors. C-iVAE replaces global auxiliary variables with node-level structural embeddings u_i derived from GNNs, enabling fine-grained prior modulation and preserving local structural distinctions that are lost in global conditioning schemes.

4.1 DAG-Structured Prior

The core design of C-iVAE uses the topological structure of the causal DAG $\mathcal{G}$ to factorize the latent prior. Given the DAG, we partition nodes into the root set $L_0 = \{i : \text{Pa}(i) = \varnothing\}$ and the non-root set $V \setminus L_0$. The joint prior is defined as

$$p_\theta(z|\mathcal{G}, e) = \prod_{i \in L_0} p_\theta(z_i|u_i, e) \cdot \prod_{i \notin L_0} p_\theta(z_i|z_{\text{Pa}(i)}, u_i), \tag{4}$$

where $e \in \mathcal{E}$ is the environment label and $u_i = \text{GNN}(\mathcal{G})[i]$ is the structural embedding. Crucially, e represents dynamic exogenous interventions, while u_i encodes the static topological position of node i. Since a single environment e can simultaneously perturb multiple root nodes (a 1:many mapping), it naturally supports do-calculus interventions without requiring a bijection. Root nodes depend on external e for symmetry breaking, while non-root nodes inherit variability from parents $z_{\text{Pa}(i)}$, propagating identifiability without ubiquitous external labels. This distinction is pivotal. Root nodes, being exogenous, have no internal ancestors to anchor their variations, thus necessitating external signals to break rotational symmetry. In contrast, non-root nodes are pinned by the structural constraints of their parents. The variation in parents, propagated through the causal mechanism f_i, acts as a dynamic, node-specific 'environment' for the child, effectively substituting for the external auxiliary information required by standard iVAE theory, provided that the independent causal mechanisms (ICM) assumption holds and parental variability is not degenerated.

For root nodes $i \in L_0$, the prior is an environment-dependent conditional Gaussian:

$$p_\theta(z_i|u_i, e) = \mathcal{N}\left(\mu_\theta^{(r)}(u_i, e), \sigma_\theta^{(r)}(u_i, e)^2\right), \tag{5}$$

where the mean and variance are parameterized by multi-layer perceptrons (MLPs). Root node identifiability requires $|E| \geq 2rk + 1$ distinct environments.

For non-root nodes $i \notin L_0$, the prior depends on sampled parent values:

$$p_\theta(z_i|z_{\text{Pa}(i)}, u_i) = \mathcal{N}\left(\mu_\theta^{(c)}(z_{\text{Pa}(i)}, u_i), \sigma_\theta^{(c)}(z_{\text{Pa}(i)}, u_i)^2\right). \tag{6}$$

Parent aggregation employs an attention mechanism. Variations in parent values, originating from ancestral responses to environments, cascade to child nodes without requiring additional environment labels.

4.2 Structure Encoder

The structure encoder $\text{GNN} : \mathcal{G} \mapsto \{u_i\}_{i=1}^d$ generates node embeddings from the DAG topology and must satisfy $u_i \neq u_j$ for structurally inequivalent nodes $i \neq j$. This asymmetry is critical for identifiability. Standard GNNs often struggle with the 'isomorphism problem,' assigning identical embeddings to structurally symmetric nodes. By integrating positional encodings such as PageRank and topological depth, our Structural Encoder breaks this symmetry, ensuring that even nodes with isomorphic local neighborhoods receive distinct, identifiability-preserving signatures that uniquely characterize their role in the global causal mechanism.

Each node is initialized with a concatenation of topological features:

$$h_i^{(0)} = [\text{RW}_k(i); d_{\text{in}}(i); d_{\text{out}}(i); \text{PR}(i); \text{Level}(i)], \tag{7}$$

including random walk return probabilities, in-degree and out-degree, PageRank centrality, and topological depth. Random walk probabilities (RW_k) capture the local neighborhood geometry and diffusion properties, while PageRank (PR) provides a global measure of node centrality and influence. Topological depth (Level) explicitly encodes the causal hierarchy, distinguishing upstream causes from downstream effects. By combining these multi-scale topological descriptors, the encoder ensures that even nodes in highly symmetric graph structures (e.g., rings or regular grids)

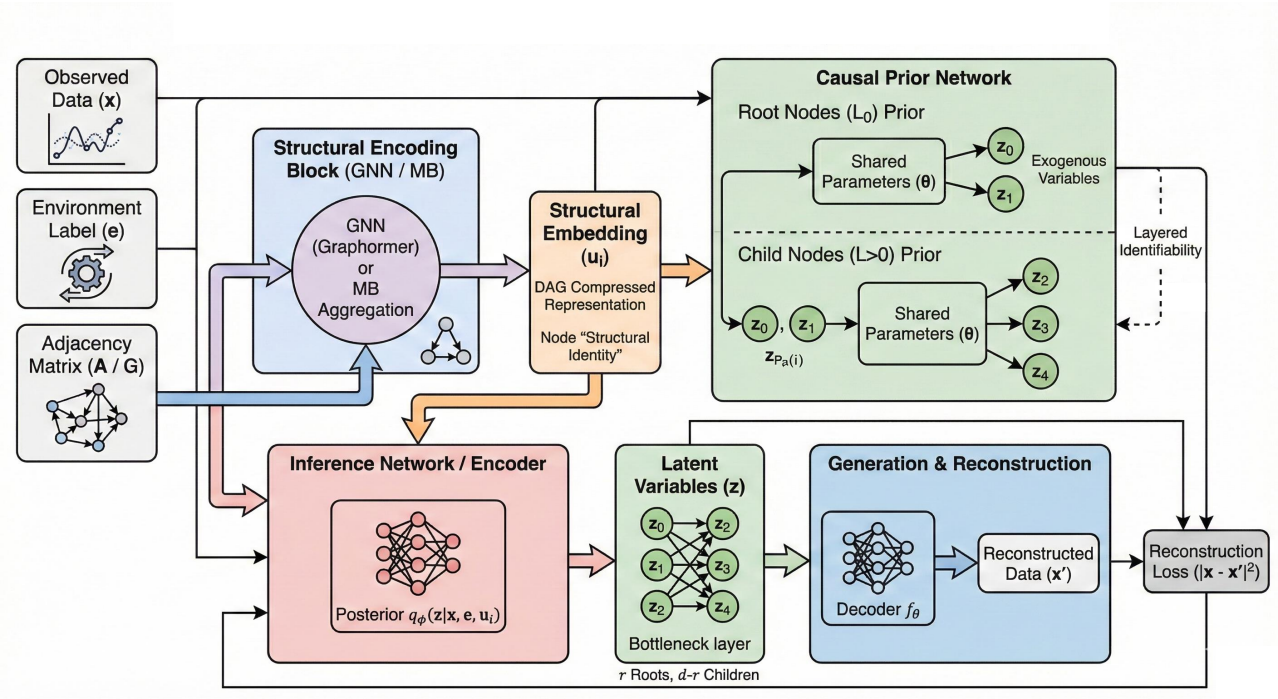


Figure 2: C-iVAE Framework Architecture. The Structural Encoder (left) uses GNNs with positional encodings to generate unique node embeddings u from the DAG topology. The Causal Prior Network (right) leverages these embeddings to modulate hierarchical priors: root nodes rely on external environments for symmetry breaking, while non-root nodes inherit variability from parents, enabling cascading identifiability with reduced environmental complexity.

receive distinct embeddings. This breaks the symmetry that typically confounds standard GNNs (the Weisfeiler-Lehman limitation), thereby satisfying the critical asymmetry assumption (A6) required for our identifiability theorem.

We employ a Transformer architecture with spatial bias to capture these topological features. However, lightweight GNNs such as GCN or GraphSAGE can also serve as highly efficient and theoretically equivalent alternatives for this structural encoding process:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + B_{\text{spatial}}\right)V, \tag{8}$$

where the spatial bias $B_{\text{spatial}}[i, j]$ depends on the shortest path distance. This spatial bias mechanism allows the attention mechanism to decay with graph distance, effectively focusing the model's capacity on the local causal neighborhood while retaining global context. After L layers, the output $u_i = h_i^{(L)}$ serves as the final embedding.

4.3 Causal Decoding and Generation

The decoding process aims to map latent variables z back to the observation space x while preserving the causal structure of the latent space.

Ancestral Sampling. Given a trained model, data generation follows the causal mechanism defined by the DAG. The generation process proceeds via ancestral sampling:

(1) **Environment Sampling**: Sample an environment label $e \sim P(E)$.
(2) **Root Node Sampling**: Sample root nodes $z_i \sim p(z_i|u_i, e)$ based on the external environment.
(3) **Cascade Sampling**: Sample non-root nodes in topological order $z_i \sim p(z_i|z_{\text{Pa}(i)}, u_i)$ based on their parent values. This ensures the generated latent variables adhere to causal dependencies.
(4) **Observation Decoding**: Generate the observed data $\hat{x} = f_\theta^{-1}(z)$ via the decoder network. The decoder, typically an MLP, maps the causally disentangled latent representation to complex tabular features.

Causal Intervention. Since the generation process explicitly models the DAG structure, C-iVAE supports causal interventions $\text{do}(Z_j = z_j^*)$. During generation, the value of node j is fixed to z_j^*, while the remaining nodes are sampled according to the graph structure (downstream nodes conditional on their potentially intervened parents). This enables the model to answer counterfactual questions and simulate intervention effects.

4.4 Variational Inference

The posterior factorizes in topological order to match the prior structure:

$$q_\phi(z|x, e) = \prod_{i=1}^{d} q_\phi(z_i|x, z_{\text{Pa}(i)}, e). \tag{9}$$

Algorithm 1: C-iVAE Training Procedure

Input: Dataset $\mathcal{D} = \{x^{(n)}, e^{(n)}\}_{n=1}^{N}$, DAG $\mathcal{G}$, Max Iterations T
Output: Trained parameters θ, ϕ

Initialize parameters θ, ϕ;
$\pi \leftarrow \text{TopologicalSort}(\mathcal{G})$;
$L_0, V \setminus L_0 \leftarrow \text{IdentifyRootSet}(\mathcal{G})$;
Precompute spatial bias matrices for GNN Structure Encoder;
while $t < T$ **do**
  Sample batch $\{x, e\}$ from $\mathcal{D}$;
  $u \leftarrow \text{GNN}_{\phi_{str}}(\mathcal{G})$; // Compute structural embeddings
  $\mu_z, \sigma_z \leftarrow \text{Encoder}_{\phi_{enc}}(x, e)$; // Inference
  $z \leftarrow \mu_z + \sigma_z \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$; // Reparameterization
  foreach $i \in L_0$ **do**
    Compute $p(z_i | u_i, e)$; // Root Prior
  end
  foreach i *in order* π ***where*** $i \notin L_0$ **do**
    Gather parents $z_{\text{Pa}(i)}$;
    Compute $p(z_i | z_{\text{Pa}(i)}, u_i)$; // Conditional Prior
  end
  $\mathcal{L}_{recon} \leftarrow -\log p_\theta(x|z)$;
  $\mathcal{L}_{KL} \leftarrow \sum_{i \in L_0} D_{KL}(\text{Root}) + \sum_{i \notin L_0} D_{KL}(\text{Cond})$;
  $\mathcal{L} \leftarrow \mathcal{L}_{recon} + \mathcal{L}_{KL}$;
  Update θ, ϕ using Adam optimizer on $\nabla \mathcal{L}$;
end

Sampling proceeds topologically, starting from root nodes and then proceeding to descendants.

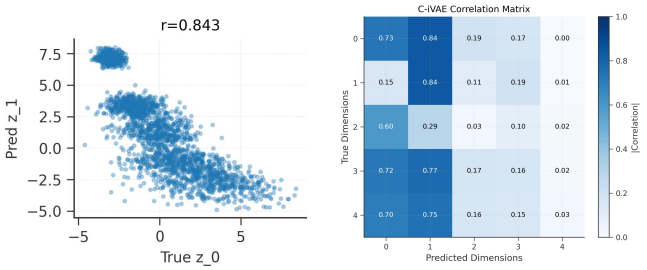
The Evidence Lower Bound (ELBO) decomposes into a reconstruction term and hierarchical KL divergences:

$$\mathcal{L} = \mathbb{E}_q[\log p_\theta(x|z)] - \sum_{i \in L_0} D_{\text{KL}}(q_i \| p_i^{\text{root}}) - \sum_{i \notin L_0} D_{\text{KL}}(q_i \| p_i^{\text{child}}). \quad (10)$$

Both the root and non-root KL terms admit closed-form Gaussian expressions, enabling efficient computation. The optimization of this hierarchical ELBO is remarkably stable compared to standard adversarial training or complex bi-level optimization schemes often used in causal discovery. By grounding the latent space in a fixed, albeit potentially noisy, causal graph, the optimization landscape becomes more convex locally, preventing the mode collapse often seen in GANs or the posterior collapse common in vanilla VAEs with powerful decoders. Furthermore, the GNN encoder acts as a structural regularizer, ensuring that parameters updates for topologically similar nodes are implicitly coupled, which accelerates convergence and improves data efficiency.

5 Theoretical Properties

We establish identifiability under the following assumptions. (A1) The DAG is known and acyclic. (A2) The decoder is invertible. (A3) Priors belong to the exponential family. (A4) The root rank condition requires $|E| \geq 2rk + 1$ environments. (A5) Non-root priors are sensitive to parent values. (A6) Structural embeddings are asymmetric, i.e., $u_i \neq u_j$ for inequivalent nodes.



(a) MCC Visualization **(b) Correlation Matrix**

Figure 3: Overview. (a) Linear alignment in MCC verifies identifiability. (b) Correlation matrix shows structure recovery.

Theorem 5.1 (Hierarchical Identifiability). *Under assumptions A1–A6, the latent variables learned by C-iVAE are identifiable up to DAG automorphisms and element-wise monotonic transformations: $\hat{z}_{\pi(i)} = g_i(z_i^*)$, where $\pi \in Aut(\mathcal{G})$ and each g_i is strictly monotonic.*

Corollary 5.2. *Environment complexity reduces from $O(dk)$ to $O(rk)$, where $r \ll d$.*

The proof proceeds by topological induction over the DAG structure. Root nodes are identified through external environment variation following the standard iVAE theorem. For non-root nodes, parent value variability propagates to their prior parameters, providing internal environments that substitute for external labels. By applying single-variable identifiability arguments recursively in topological order, where each node treats its previously identified parents as conditioning variables, we ensure that every node in the DAG is uniquely recovered up to simple transformations, thus completing the proof. Full details and formal derivations, including the recursive logic and uniqueness checks, appear in the Appendix.

6 Experiments

In this section, we rigorously validate the theoretical claims of C-iVAE through comprehensive experiments. We first utilize synthetic data with ground-truth DAGs to empirically verify our structure-identifiability hypothesis, demonstrating that C-iVAE achieves identifiability even under limited environmental diversity where standard baselines fail. Subsequently, we extend this evaluation to eight diverse real-world datasets, establishing that our model delivers superior performance across downstream utility, distribution quality, and causal consistency, effectively balancing the trade-offs inherent in prior works. Finally, we analyze the model's robustness to structural noise and out-of-distribution shifts using controlled interventions. Detailed dataset descriptions and supplementary results are provided in Appendix A and B, ensuring reproducibility and providing further insights into hyperparameter sensitivity. The experimental results serve not just to benchmark performance, but to stress-test the theoretical bounds of our hierarchical identifiability theorem under non-ideal conditions.

Algorithm 2: Synthetic Data Generation Process

Input: DAG G, mechanisms f_i, noise distributions P_ϵ
Output: Dataset X
for $i \in \mathit{TopologicalOrder}(G)$ **do**
  Sample noise $\epsilon_i \sim P_\epsilon$;
  $x_i \leftarrow f_i(x_{pa(i)}) + \epsilon_i$;
end
return $X = \{x_1, \ldots, x_d\}$;

6.1 Synthetic Data

Datasets and Baselines. Following the protocol of iVAE [19], we construct synthetic datasets with known DAG structures. To strictly mimic real-world causal processes, we employ a non-linear SCM. Specifically, for each node j in the DAG, its value is generated as $x_j = f_j(x_{\mathrm{Pa}(j)}) + \epsilon_j$, where f_j is a randomly initialized MLP with LeakyReLU activation ensuring non-linearity, and ϵ_j is noise drawn from Gaussian distributions. The key variable is the **number of environments** E. We compare against **VAE** [20], **iVAE** [19], and **CausalVAE** [36].

Evaluation Metrics. We employ the Mean Correlation Coefficient (MCC) as the primary metric to quantify identifiability. MCC measures the Pearson correlation between the learned latent variables and the ground truth generative factors, computed after finding the optimal permutation to match latent dimensions. Higher MCC indicates better disentanglement and structural recovery.

Result Analysis. As shown in Figure 4(a-b), C-iVAE achieves a Mean Correlation Coefficient (MCC) exceeding **0.85** even in the challenging regime of minimal environmental diversity ($E \geq 2$), significantly outperforming iVAE and standard VAE. This empirical result strongly validates our "Structure-Identifiability" hypothesis: the internal variability inherited from structural dependencies effectively acts as a surrogate for external environment labels, providing the necessary auxiliary information for disentanglement. Furthermore, while iVAE performance eventually recovers as the number of environments E increases, C-iVAE consistently maintains lower variance and faster convergence across all settings, demonstrating superior sample efficiency and optimization stability.

Notably, all models exhibit declining MCC beyond $E > 3$. This phenomenon arises because excessive environment partitioning reduces per-environment sample size, leading to unstable variance estimation in the prior distribution. Additionally, fine-grained environment labels may introduce spurious correlations that conflict with true causal structure. Our structural constraints mitigate this degradation, as evidenced by C-iVAE's relatively stable MCC (0.74–0.76) even at $E = 20$. We further evaluate DAG misspecification and encoder sensitivity in Appendices C–D, confirming that C-iVAE remains robust under noisy structural inputs.

6.2 Real-world Data

Datasets and Baselines. We evaluate on **8 real-world datasets**: *Sachs* [28], *IHDP* [15], *Adult* [23], *German* [11], *ACS* [10], *Heart* [17], *Wine* [9], and *Covertype* [3]. These span causal inference, social science, and medical/nature domains with sample sizes from 747 to 581k. For DAG specification, we use domain-expert knowledge where available (e.g., the experimentally validated protein network in Sachs [28]); otherwise, we apply standard causal discovery algorithms (PC [30] or GES [8]) on held-out data, following standard practice in causal representation learning. Full statistics are in Appendix A. We compare against **5 baselines**: **VAE** [20], **VAE-BGM** [1], **iVAE** [19], **CausalVAE** [36], and a **TabSCM-VAE** implementation combining VAE inference with SCM-style topological generation. Implementation details are in Appendix A.

Evaluation Metrics. We employ **12 metrics** in three categories (see Appendix A for formal definitions): (1) **Causal Consistency**: SRE (structural reconstruction error) and C (latent-true correlation); (2) **Distribution Quality**: Wasserstein distance (W_1), Jensen-Shannon divergence (JS), L_2 error, and correlation preservation (Corr); and (3) **Downstream Utility**: AUC, F1-Score, Precision, Recall, Accuracy, and NMI for clustering. We also evaluate encoder ablations and efficiency in Appendix D.

Result Analysis. C-iVAE consistently achieves the lowest Structural Reconstruction Error (SRE) across most datasets, confirming its ability to capture ground truth dependencies (Figures 3(b) and 4(c)). C-iVAE ranks top two in 85% of cases. Figure 1 illustrates distinct trade-offs in baselines: VAE-BGM excels in distribution fitting (W_1, JS) but lacks structural consistency, creating a 'mimicry' without understanding the skeleton. This phenomenon highlights a critical danger in tabular generation: the 'illusion of fidelity'. A model can perfectly replicate marginal distributions while hallucinating the joint dependency structure, leading to potential disasters if such ostensibly 'accurate' data is used for causal decision-making. iVAE shows disentanglement potential but struggles with complex distributions. In contrast, C-iVAE exhibits no significant weaknesses, balancing causal structure, distribution quality, and downstream utility. By prioritizing the skeleton, C-iVAE ensures generated data *behaves* correctly under manipulation. Notably, on the large-scale **Covertype** dataset, C-iVAE achieves a premier F1-Score of **0.856**, demonstrating that causal structure regularization enhances generalization.

6.3 Robustness Analysis

OOD Generalization (Exp 2). Under significant distributional shifts $\delta \in \{0, 1, 2\}$, standard VAE performance degrades rapidly due to its inability to disentangle causal mechanisms. While iVAE leads at $\delta = 0$, its error spikes dramatically by 5.4× at $\delta = 2$, indicating a failure to generalize beyond the training support. In sharp contrast, C-iVAE remains remarkably stable and superior even at large shifts (Table 1). This stability under shift ($\delta = 2$) suggests that C-iVAE learns the invariant causal mechanism $P(X|\mathrm{Pa}(X))$, contrasting with baselines that merely fit the joint distribution $P(X, Y)$.

(See Appendix D for Table 8 and detailed ablation analysis).

Structural Noise Robustness (Figure 5). We rigorously tested the model's tolerance to imperfect DAG priors by introducing both random edge noise and systematic discovery-style errors. Even with 30% edge perturbation, C-iVAE maintains performance levels comparable to the iVAE baseline in Figure 5. Appendix Table 6 further evaluates PC-inferred graphs, edge reversals, deletions, additions, omitted confounders, and feature swaps. We attribute this

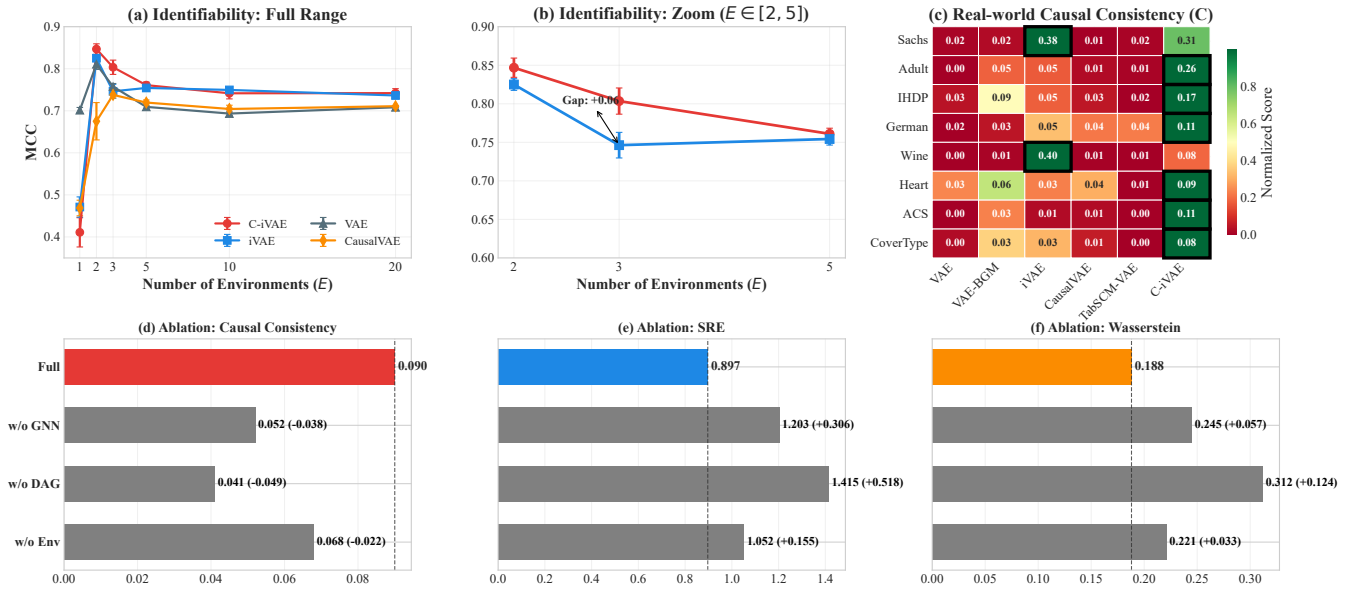


Figure 4: Comprehensive Experimental Analysis. (a-b) Identifiability on Synthetic Data: MCC vs. Environments (Full Range & Zoom). (c) Real-world Consistency: Heatmap of Causal Consistency metric (C) across 8 datasets. (d-f) Ablation Study: Impact of components on C, SRE, and W_1.

Table 1: OOD Generalization (Metric: MSE ↓). Values are Mean ± SEM.

Shift δ	VAE	iVAE	CausalVAE	**C-iVAE**
0.0	0.56 ± 0.35	**0.38** ± 0.00	0.68 ± 0.01	0.39 ± 0.18
1.0	0.49 ± 0.40	0.57 ± 0.00	0.74 ± 0.03	**0.38** ± 0.10
2.0	0.89 ± 0.83	2.05 ± 0.00	1.81 ± 0.00	**0.58** ± 0.08

resilience to the GNN encoder's aggregation over Markov Blanket neighborhoods, which provides a natural smoothing mechanism that filters local structural noise while preserving useful topology.

6.4 Ablation Study

To quantify the contribution of each architectural component, we performed a comprehensive stepwise ablation study (Figure 4(c-e) and Appendix D Table 8) to disentangle the three levels of causal feature learning. First, the hierarchical DAG prior forms our core theoretical foundation for identifiability; removing the DAG reduces the model to the iVAE setting. Second, the GNN structural encoder translates this topology into node-aware embeddings; the corrected ablation label is "w/o GNN Structure Encoder," not "w/o DAG structure." Third, the Markov Blanket mask enforces strict local independence; removing it causes a substantial increase in SRE, highlighting the necessity of differentiating causal mechanisms across nodes.

Furthermore, Appendix Table 8 reports the efficiency trade-off across Transformer, GCN, and GraphSAGE encoders. While the full Transformer encoder yields strong structural reconstruction,

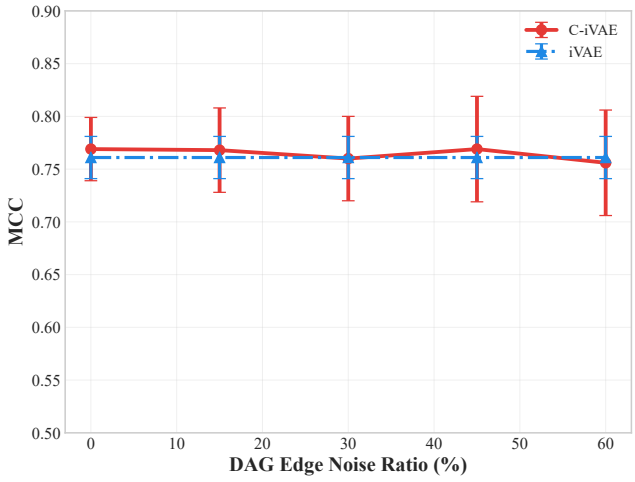


Figure 5: Structural Noise Robustness Analysis. C-iVAE maintains high MCC even with 30% redundant/missing edges.

simpler GNNs can offer competitive MCC at lower inference cost, suggesting a practical trade-off for real-world deployment.

7 Conclusion

We presented C-iVAE, a framework connecting identifiable representation learning and tabular data generation by using causal DAGs. By encoding topological dependencies via GNNs, C-iVAE reduces environmental complexity from $O(dk)$ to $O(rk)$, enabling identifiability in environment-scarce regimes. Validation results

confirm C-iVAE outperforms baselines in fidelity and causal consistency, while exhibiting robustness to structural noise. However, we acknowledge that our reliance on the ICM assumption may present limitations in scenarios with strong unobserved confounders or highly adversarial distribution shifts. Future work will explore dynamic graphs, simultaneous causal discovery, and relaxing the ICM constraints. Additionally, integrating Large Language Models (LLMs) with C-iVAE presents a notable direction. LLMs could provide initial causal priors or interpret structural embeddings, enhancing automation. C-iVAE thus serves as a step towards Causal AI systems capable of reasoning in complex environments.

Acknowledgments

This work was supported by the NSFC Project (No.62576346), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), the fundamental research funds for the central universities, and the research funds of Renmin University of China (24XNKJ13), and Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing.

References

[1] P. A. Apellaniz, J. Parras, and S. Zazo. 2024. An Improved Tabular Data Generator with VAE-GMM Integration. In *EUSIPCO*.
[2] Enrico Barbierato et al. 2022. A Methodology for Controlling Bias and Fairness in Synthetic Data Generation. *Applied Sciences* (2022).
[3] Jock A Blackard and Denis J Dean. 1999. Comparative accuracies of artificial neural networks and discriminant analysis for remote sensing classification. *Computers and Electronics in Agriculture* (1999).
[4] Vadim Borisov et al. 2022. Deep Neural Networks and Tabular Data: A Survey. *IEEE TNNLS* (2022).
[5] Vadim Borisov et al. 2023. Language Models are Realistic Tabular Data Generators. In *ICLR*.
[6] Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. 2023. Learning Linear Causal Representations from Interventions under General Nonlinear Mixing. In *NeurIPS*.
[7] Richard J Chen et al. 2021. Synthetic data in machine learning for medicine and healthcare. In *Nature Biomedical Engineering*.
[8] David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research* 3, Nov (2002), 507–554.
[9] Paulo Cortez et al. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* (2009).
[10] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *NeurIPS*.
[11] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml.
[12] Cristobal Esteban et al. 2017. Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs. In *arXiv*.
[13] Maayan Frid-Adar et al. 2018. GAN-based Synthetic Medical Image Augmentation for increased CNN Performance in Liver Lesion Classification. In *ISBI*.
[14] Yury Gorishniy et al. 2021. Revisiting Deep Learning Models for Tabular Data. In *NeurIPS*.
[15] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* (2011).
[16] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. 2023. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *ICLR*.
[17] Andras Janosi et al. 1988. Heart disease data set. *UCI Machine Learning Repository* (1988).
[18] James Jordon et al. 2019. PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In *ICLR*.
[19] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. 2020. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *AISTATS*.
[20] Diederik P Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *arXiv* (2013).
[21] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. 2022. Identifiability of Deep Generative Models without Auxiliary Information. In *NeurIPS*.
[22] Murat Kocaoglu et al. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *ICLR*.
[23] Ron Kohavi. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*.
[24] Aneesh Komanduri et al. 2024. From Identifiable Causal Representations to Controllable Counterfactual Generation: A Survey on Causal Generative Modeling. *Transactions on Machine Learning Research (TMLR)* (2024).
[25] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: Modeling Tabular Data with Diffusion Models. In *ICML*.
[26] Francesco Locatello et al. 2019. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *ICML*.
[27] Thach Nguyen et al. 2024. CAGAN: A Novel GAN Approach to Augment Limited Tabular Data for Short-Term Substance Use Prediction. In *IJCAI*.
[28] Karen Sachs, Omar Perez, Dana Pe'er, et al. 2005. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308, 5721 (2005), 523–529.
[29] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. 2025. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation. In *ICLR*.
[30] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, Prediction, and Search*. MIT press.
[31] Boyang Sun, Ignavier Ng, Guangyi Chen, Yifan Shen, Qirong Ho, and Kun Zhang. 2024. Continual Learning of Nonlinear Independent Representations. *arXiv preprint arXiv:2408.05788* (2024).
[32] Pablo Villalobos et al. 2022. Will we run out of data? Limits of LLM scaling based on human-generated data. *arXiv* (2022).
[33] Julius von Kügelgen et al. 2021. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. (2021).
[34] Lei Xu et al. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. In *IEEE Big Data*.
[35] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling Tabular Data using Conditional GAN. In *NeurIPS*.
[36] Mengyue Yang et al. 2021. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *CVPR*.
[37] Yue Yu et al. 2019. DAG-GNN: DAG Structure Learning with Graph Neural Networks. In *ICML*.
[38] Hengrui Zhang et al. 2024. TabSyn: Modeling Tabular Data with Structure-based Diffusion in Latent Space. In *ICLR*.
[39] Chunting Zhou et al. 2023. LIMA: Less Is More for Alignment. *arXiv* (2023).

A Experimental Setup

Tables 2–3 collect the reproducibility details needed to interpret the main experiments. The environment variable e is observed side information that modulates root-node priors; for non-root nodes, environmental variation is inherited through the parent-conditioned prior.

Table 2: Datasets and environment variables.

Dataset	Samples	Task	Environment e
Sachs	7,466	Structure	Intervention
Adult	48,842	Class.	Age bins
IHDP	747	Causal effect	Treatment
German	1,000	Class.	Credit purpose
Wine	6,497	Regression	Wine type
Heart	920	Class.	Medical center
ACS	15,000	Class.	Region
Covertype	581,012	Multi-class	Wilderness area

Table 3: Shared implementation and evaluation protocol.

Item	Setting
Optimizer	Adam, learning rate 10^{-3}, batch size 256
Schedule	100-epoch KL warm-up, same splits across methods
Backbone	3-layer MLP encoder/decoder; C-iVAE adds a 2-layer graph encoder
Environment use	e modulates roots; children condition on parent latents
Causal metrics	Structural reconstruction error (SRE), latent-true correlation C
Quality/utility	W_1, JS, L_2, Corr; AUC, F1, precision, recall, accuracy, NMI

B Full Proof and ELBO

Table 4: Assumptions used in the hierarchical identifiability proof.

ID	Assumption	Role
A1	Known acyclic DAG $\mathcal{G}$	Defines topological layers and parents
A2	Invertible decoder f_θ	Transfers equality of $p(x \mid e)$ to latents
A3	Exponential-family priors	Enables natural-parameter rank arguments
A4	Root rank, $\lvert E\rvert \geq 2rk+1$	Identifies r root nodes using external e
A5	Parent sensitivity of λ_i	Turns parent variation into child auxiliary signal
A6	Asymmetric structural embeddings u_i	Removes non-automorphic label switching

Theorem B.1 (Hierarchical Identifiability). *Under A1–A6, if two C-iVAE parameterizations induce the same conditional density $p_\theta(x \mid e) = p_{\hat\theta}(x \mid e)$ for all x, e, then there exists a DAG automorphism $\pi \in \mathrm{Aut}(\mathcal{G})$ and component-wise invertible monotone maps $\{g_i\}_{i=1}^d$ such that*

$$\hat z_{\pi(i)} = g_i(z_i^*) \quad \text{for all } i.$$

Proof. Let $L_0, L_1, \ldots, L_K$ be the topological layers of $\mathcal{G}$, where L_0 contains roots and every parent of a node in L_k lies in an earlier layer. The conditional prior for a node is exponential-family:

$$\log p_\theta(z_i \mid c_i) = T_i(z_i)^\top \lambda_i(c_i) - A_i(c_i) + \log h_i(z_i),$$

where $c_i = (u_i, e)$ for $i \in L_0$ and $c_i = (z_{\mathrm{Pa}(i)}, u_i)$ otherwise.

Root layer. For $i \in L_0$, the prior factorizes as $\prod_{i\in L_0} p_\theta(z_i \mid u_i, e)$. By A2, equality of observed conditionals implies equality of the induced latent densities up to an invertible reparameterization. By A3 and the iVAE rank argument, if the natural-parameter contrast matrix

$$\mathbf{L}_0 = \big[\lambda(u,e_1) - \lambda(u,e_0), \ldots, \lambda(u,e_m) - \lambda(u,e_0)\big]$$

has full rank over the rk root sufficient statistics, then $T(\hat z_{L_0}) = AT(z^*_{L_0}) + b$. A4 ensures the needed rank with $\lvert E\rvert \geq 2rk+1$. For the univariate sufficient statistics used by each root, the affine relation in sufficient-statistic space corresponds to a permutation and component-wise monotone transformation of the root variables.

Propagation lemma. Consider a non-root node j and define $\Lambda_j(v, u_j) = \lambda_j(v, u_j)$ with $v = z_{\mathrm{Pa}(j)}$. If the parents vary and A5 holds, then the child natural parameters vary. Around $v_0 = \mathbb{E}[v]$,

$$\Lambda_j(v) \approx \Lambda_j(v_0) + J_{\Lambda_j}(v_0)(v - v_0),$$

and therefore

$$\mathrm{Cov}[\Lambda_j(v)] \approx J_{\Lambda_j}(v_0)\,\mathrm{Cov}[v]\,J_{\Lambda_j}(v_0)^\top.$$

The covariance is non-degenerate in the directions where parents vary, while A5 makes the Jacobian full rank almost everywhere. Thus parent variation supplies valid auxiliary contrast for node j.

Induction step. Assume layers $L_0, \ldots, L_{k-1}$ are identified. For any $j \in L_k$, all parents are in earlier layers, so $z_{\mathrm{Pa}(j)}$ is known up to already identified component-wise maps. Conditioning on these parents gives the local conditional prior $p(z_j \mid z_{\mathrm{Pa}(j)}, u_j)$. By the propagation lemma, $z_{\mathrm{Pa}(j)}$ plays the role of an internal auxiliary variable. Applying the one-node iVAE argument to z_j yields $T_j(\hat z_j) = a_j T_j(z_j^*) + b_j$, hence z_j is identifiable up to a monotone component map.

Automorphism ambiguity. A6 assigns distinct structural embeddings to structurally inequivalent nodes, so non-automorphic node swaps would change the conditional prior family and cannot preserve $p(x \mid e)$. If the graph itself has an automorphism preserving all parent-child relations and structural embeddings, no observational distribution can distinguish nodes inside that symmetry class. Hence the only remaining permutation is $\pi \in \mathrm{Aut}(\mathcal{G})$. Induction over all layers proves the theorem. □

Rank condition and monotone recovery. The proof uses the same identifiability mechanism as iVAE only at the points where a variable has a sufficiently rich conditioning signal. For roots, the signal is the external environment e. For descendants, the signal is the parent vector after it has already been identified by induction. The exponential-family form makes this precise: variation in c_i changes the natural parameter vector $\lambda_i(c_i)$, and the full-rank contrast excludes alternative latent coordinates that would keep all conditional densities unchanged. In the scalar latent setting used here, an affine relation between sufficient statistics corresponds to an invertible component-wise transformation of the latent coordinate. We state the result as monotone recovery because the ordering of each scalar factor is preserved by the identifiable sufficient-statistic parameterization; arbitrary nonlinear mixing across nodes is excluded by the local conditional factorization.

Why children do not need new environment labels. Standard iVAE treats all d latent coordinates as requiring external auxiliary variation. In C-iVAE, the root layer is the only layer without internal causes, so external environments are needed only there. Once roots are identified, their descendants receive varying parent values through the structural equations. These parent values are not observed labels, but they are valid internal auxiliaries because the child prior is explicitly conditioned on them. The induction argument therefore trades a global external-rank requirement for a sequence of local conditional-rank requirements. This is the mathematical reason that the external environment count scales with r, the number of roots, rather than with the full latent dimension d.

Corollary B.2 (Environment Complexity). *Only root nodes require external environment rank. If each node uses k sufficient statistics and the DAG has r roots, C-iVAE needs rank over rk root statistics rather than dk statistics for all nodes. The external requirement therefore scales as $O(rk)$ instead of $O(dk)$; child nodes obtain their auxiliary variation from identified parents.*

ELBO derivation. For one observation, the conditional marginal likelihood satisfies Jensen's inequality:

$$\log p_\theta(x \mid e, \mathcal{G}) = \log \int q_\phi(z \mid x, e, \mathcal{G}) \frac{p_\theta(x, z \mid e, \mathcal{G})}{q_\phi(z \mid x, e, \mathcal{G})} dz$$
$$\geq \mathbb{E}_{q_\phi}\left[\log \frac{p_\theta(x, z \mid e, \mathcal{G})}{q_\phi(z \mid x, e, \mathcal{G})}\right] \equiv \mathcal{L}(x, e). \quad (11)$$

The DAG prior and topological posterior are

$$p_\theta(z \mid e, \mathcal{G}) = \prod_{i \in L_0} p_\theta(z_i \mid u_i, e) \prod_{i \notin L_0} p_\theta(z_i \mid z_{\mathrm{Pa}(i)}, u_i),$$

$$q_\phi(z \mid x, e, \mathcal{G}) = \prod_{i=1}^{d} q_\phi(z_i \mid x, e, z_{\mathrm{Pa}(i)}, u_i).$$

Define the local terms

$$R_i = D_{\mathrm{KL}}(q_i(z_i \mid x, e) \| p_i(z_i \mid u_i, e))$$

for roots and

$$C_i(v) = D_{\mathrm{KL}}(q_i(z_i \mid x, e, v) \| p_i(z_i \mid v, u_i))$$

for children with $v = z_{\mathrm{Pa}(i)}$. Substituting the factorizations gives

$$\mathcal{L}(x, e) = \mathbb{E}_{q_\phi}[\log p_\theta(x \mid z)] - \sum_{i \in L_0} R_i - \sum_{i \notin L_0} \mathbb{E}_{q(v \mid x, e)}[C_i(v)].$$

When $q = \mathcal{N}(\mu_q, \sigma_q^2)$ and $p = \mathcal{N}(\mu_p, \sigma_p^2)$,

$$D_{\mathrm{KL}}(q \| p) = \log \frac{\sigma_p}{\sigma_q} + \frac{\sigma_q^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} - \frac{1}{2}.$$

For child nodes, μ_p and σ_p depend on sampled parents, so the outer expectation is estimated with reparameterized posterior samples. This is the objective optimized by Algorithm 1.

Estimator used in training. In implementation, root KL terms are computed directly from the encoder and root-prior Gaussian parameters. For child nodes, we sample $v = z_{\mathrm{Pa}(i)}$ using the reparameterization trick, evaluate the child prior network at (v, u_i), and average the resulting Gaussian KL over the minibatch. This estimator is unbiased for the Monte Carlo approximation of the ELBO and remains differentiable with respect to encoder, decoder, graph encoder, and prior-network parameters. The GNN embeddings u_i are shared across samples for a fixed graph, while the parent samples vary by observation and environment, which is exactly the internal variation required by the proof.

Connection between ELBO and identifiability. The theorem is a population statement about the model family: when the conditional distribution is matched exactly, the latent coordinates are unique up to the stated ambiguities. The ELBO supplies the tractable optimization objective used to approach that population optimum. Its KL decomposition is important because it matches the proof structure: root KL terms enforce the externally modulated root mechanisms, while child KL terms enforce parent-conditioned mechanisms in topological order. Thus the training objective is not merely a VAE regularizer; it mirrors the assumptions used in the identifiability argument.

Table 5: How the proof components correspond to model components.

Proof component	Model component	Consequence
Root rank	$p(z_i \mid u_i, e)$ for $i \in L_0$	Identifies root sufficient statistics
Propagation	$p(z_j \mid z_{\mathrm{Pa}(j)}, u_j)$	Converts parent variation into auxiliary contrast
Induction	Topological sampling order	Identifies descendants layer by layer
Asymmetry	GNN structural embedding u_i	Blocks non-automorphic label switching
Gaussian KL	Closed-form local terms	Enables stable end-to-end optimization

C Robustness to DAG Misspecification

Table 6 targets the rebuttal concern that the supplied DAG may be wrong. The iVAE baseline is 0.8741±0.013 MCC; C-iVAE stays above it across inferred, edited, and node-index-corrupted DAGs. The largest degradation occurs under severe deletion or node swapping, which is expected because these operations remove or relocate causal parents rather than adding redundant local context.

The robustness experiment is designed to separate three realistic failure modes. First, discovery noise is represented by replacing the oracle graph with a PC-inferred graph. Second, edge-level errors are represented by reversals, deletions, and additions. Third, semantic mismatch is represented by omitted confounders and node-index swaps. The consistent positive MCC gap indicates that C-iVAE does not require a perfect graph to improve over an unstructured iVAE baseline, although the proof guarantee itself assumes the correct DAG.

Table 6: Robustness under inferred and misspecified DAGs. Gap is C-iVAE minus iVAE.

Condition	DAG error	C-iVAE MCC	Gap	p
Ground truth	none	0.8928 ± 0.013	0.0187	0.031
PC inferred	discovered graph	0.8981 ± 0.011	0.0241	0.063
Flip 20%	reverse edges	0.9019 ± 0.009	0.0278	0.031
Flip 50%	reverse edges	0.8975 ± 0.009	0.0234	0.031
Delete 20%	remove true edges	0.8928 ± 0.012	0.0187	0.031
Delete 50%	remove true edges	0.8828 ± 0.007	0.0087	0.156
Add 20%	add spurious edges	0.8956 ± 0.012	0.0215	0.063
Add 50%	add spurious edges	0.8925 ± 0.008	0.0184	0.031
Confounder 50%	omit confounders	0.8912 ± 0.009	0.0171	0.031
Swap 20%	swap node indices	0.8876 ± 0.011	0.0136	0.219

Table 7: Interpretation of systematic DAG errors used in robustness tests.

Error family	What changes	Failure mode tested
PC inferred	Replace oracle graph with discovered graph	Realistic discovery noise
Flip	Reverse a fraction of directed edges	Wrong causal orientation
Delete	Remove true parent-child links	Missing mechanisms
Add	Insert spurious directed links	Over-connected topology
Confounder	Remove common causes and connect children	Latent confounding proxy
Swap	Exchange node indices before encoding	Feature-to-node mismatch

D Ablation and Efficiency

The earlier label "w/o Structure" only removed the GNN structure encoder; the DAG hierarchy and parent-conditioned prior were still present. Table 8 separates the no-DAG iVAE baseline from topology-aware encoders on the harder three-root topology. Removing the full hierarchy collapses C-iVAE to no-DAG iVAE, whereas replacing the structural encoder preserves the topology-aware prior but changes the cost-accuracy trade-off.

Table 8: Encoder ablation and efficiency on the three-root topology.

Variant	Structure used	MCC ↑	SRE ↓	Train s ↓	Infer ms ↓	MB
iVAE	no DAG	0.737 ± 0.015	0.523 ± 0.119	12.7	1.3	23.7
Transformer	full GNN encoder	0.745 ± 0.007	0.159 ± 0.249	71.0	28.0	24.4
GCN	2-layer GCN	**0.751 ± 0.020**	**0.075 ± 0.048**	60.8	23.1	23.8
GraphSAGE	2-layer GraphSAGE	0.747 ± 0.024	0.181 ± 0.173	52.0	9.2	23.8

Table 9 summarizes what each ablation removes. The important distinction is that the no-DAG baseline removes the internal auxiliary variables used in the proof, while encoder substitutions only change how node-specific structural embeddings are computed. This explains why GCN and GraphSAGE can retain competitive MCC: local message passing still preserves parent-child context, even though it is cheaper than full attention.

Table 9: Component-level interpretation of ablation results.

Component	Removed in	Interpretation
DAG hierarchy	iVAE baseline	Loses parent-conditioned internal environments
GNN structure encoder	"w/o Structure"	Loses node-aware topology embeddings, not the DAG prior
Transformer attention	GCN/GraphSAGE	Lower cost with similar topology aggregation
Markov blanket mask	main ablation	Higher SRE from weaker local mechanism separation

Overall, the ablation supports the theoretical decomposition: identifiability benefits mainly come from the hierarchical DAG prior and the root-to-child propagation of auxiliary variation, while the specific graph encoder controls efficiency and structural smoothness.