DEEP CAUSAL GENERATIVE MODELING FOR TABULAR DATA IMPUTATION AND INTERVENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular data synthesis could overcome the tabular data incompleteness and data availability issue. In most prior works, deep generative models are basically constructed following standard architecture designs. However, these works do not consider the inter-relationships among the features, or the latent variables. To fully leverage these inter-relationships, we develop a novel causal-aware asymmetric variational autoencoder architecture (CAT) for tabular data generation, imputation, and intervention. The developed model, called CAT-MIWAE, learns exogenous causal representation with a pre-defined causal graph in incomplete data context. It provides interpretability for partially observed features and could efficiently address missing value imputation problem. Besides, CAT-MIWAE can sample data from distributions under arbitrary conditions and interventions. This merit enables us to actively generate counterfactual or debiased fair data samples for any subpopulation of interest. To validate the effectiveness of the proposed causally aware models, we conduct extensive experiments on real-world tabular datasets. Experiments show that the proposed models outperform the state-of-the-art models. Moreover, we perform conditional average treatment effect (CATE) estimations to show that CAT-MIWAE model could appropriately extrapolate any conditional or interventional distributions from the original observed data distribution.

1 INTRODUCTION

Tabular data is one of the most common types of data with structured features of heterogeneous types (continuous and discrete). Due to its generic representation capability, on one hand, tabular data has been widely used in many fields, such as medical diagnosis (Ulmer et al., 2020), finance (Tan et al., 2018), recommendation systems (Sun et al., 2019). On the other hand, tabular data suffers from the availability and quality for high cost of data collection or incompleteness of records, which weakens the validity of inferences drawn from estimates and analysis (Tourangeau et al., 2013).

To mitigate this critical issue, tabular data synthesis, which can synthesize tabular data records with high fidelity, is proposed. However, tabular data synthesis has two major concerns: 1) generation: referring to training on the fully observed data and then generating intact records; 2) imputation: referring to online learning on incomplete data and then imputing its missing values. Specifically, imputation can generate data with arbitrary conditions. Substantial deep generative models (DGMs) endeavor to efficiently handle these synthesis tasks, such as generative adversarial network (GAN) (Goodfellow et al., 2014), variational autoencoder (VAE) (Kingma & Welling, 2013), normalizing flow (Dinh et al., 2014), and their several variants Srivastava et al. (2017); Yoon et al. (2018); Li et al. (2018); Ma et al. (2019; 2020); Peis et al. (2022). DGMs learn a mapping function from the latent space (i.e., the latent distribution) to the feature space (i.e., the data distribution), drawing their high fitting capability from deep neural architectures. Several DGMs specialized in tabular data synthesis have been proposed and achieved significant success (Park et al., 2018; Xu et al., 2019; Lee et al., 2021; Kim et al., 2021). However, most of aforementioned works do not explore the correlations among the features, or the latent variables, which may significantly contribute to modeling the target data distribution (Radford et al., 2015). This motivates our work in this paper.

To this end, we leverage causality to benefit tabular data synthesis. Causality is a typical inter-variable relationship, which indicates the exact causal relationship between each feature. Inherently,

causality can represent an individualized generative process for each feature, thereby implying their mutual dependencies. Hence, incorporating causal knowledge in DGMs is capable to capture inter-correlations and model target distribution for synthesizing higher quality data. Following this principle, we leverage causal knowledge to benefit DGMs on imputation tasks. To the best of our knowledge, this is the first work to incorporate causality in imputation tasks. Intuitively, additional causality enables model to better structure and control its online learning on incompletely observed data. Causality can facilitate the target distribution modeling and missing features imputation. We consider tabular data with (partially) incomplete records but available causal knowledge. The causal knowledge could be general domain knowledge provided by experts beforehand, or learned by causal discovery method (Wen et al., 2022).

In this work, we develop a novel "**CAT**" architecture for VAE-based DGMs, standing for: 1) **C**ausal: causality-aware structure; 2) **A**symmetric: single shared encoder with multiple specialized decoders; 3) **T**abular: special treatments for heterogeneous types of features. The developed CAT-based architecture characterizes itself with high flexibility and compatibility, and is capable of efficiently handling both generation and imputation tasks. To show the effectiveness of CAT, we instantiate it with MIWAE (Mattei & Frellsen, 2019) bound and propose CAT-MIWAE for imputation. CAT-MIWAE learns the exogenous causal representation, i.e., a special case of causal representation learning (Scholkopf et al., 2021), for any partially observed data. Besides, due to the combination of imputation and causality, our CAT-MIWAE is able to sample data from distributions given arbitrary conditions and interventions. This critical advantage could potentially generate counterfactual (Karimi et al., 2020), debiasing unfair data (van Breugel et al., 2021), and helping estimate conditional average treatment effect (CATE), etc., exclusively for arbitrary subpopulation (subgroup) of interest, rather than the whole or a single individual(s).

We highlight our contributions as follows.

1. We elaborate the analysis on how to integrate causality into VAE-based generative models, providing theoretical support for our designations.
2. We propose CAT-based architecture with CAT-MIWAE, which can incorporate additional causal knowledge to effectively learn the exogenous causal representation and handle tabular data imputation tasks.
3. We show a significant merit of CAT-MIWAE that it allows to sample data from distributions given arbitrary conditions and interventions (we refer to this as **extrapolation**).
4. We introduce comprehensive and reproducible empirical baselines for comparison. We conduct extensive experiments on real-world datasets, and the results show that CAT-MIWAE outperforms other state-of-the-art solutions.

2 Related Work

DGMs for Tabular Data Choi et al. (2017) firstly applied GAN in electronic health record (EHR) data and proposed MedGAN to generate high-dimensional discrete variables. TableGAN (Park et al., 2018) adopted a well-designed combination of loss functions, striking a balance between model utility and privacy concern. Xu et al. (2019) proposed a special perprocessing on continuous variables for DGMs to more easily model the target distribution. Their proposed models, CTGAN and TVAE, both achieve great performance improvements compared to prior works. Other two works, Lee et al. (2021) and Kim et al. (2021), both followed the preprocessing of Xu et al. (2019), and proposed to augment GAN frameworks with Neural Ordinary Differential Equation (NODE) structure. Though acquiring fair results, the usage of NODE structure incurs huge computational overhead.

DGMs for Missing Value Imputation Works related to imputation basically consider three missing mechanisms setting (Little & Rubin, 2019): 1) missing completely at random (MCAR) : missingness occurs entirely at random; 2) missing at random (MAR): missingness depends only on the observed variables; 3) missing not at random (MNAR): missingness depends on both the observed and the unobserved. For GAN-based DGMs, GAIN (Yoon et al., 2018) and MisGAN (Li et al., 2018) are two representative works, both introducing additional networks to help learn the missing mechanism. However, this may increase optimization difficulty, especially for min-max objective

like GANs. Besides, their theoretical results are only available under MCAR assumption. For VAE-based DGMs, the solution is more decent. In both MCAR and MAR cases, VAE does not consider the missing mechanism. VAE just slightly changes its objective to obtain a lower bound of the likelihood of the observed data (more details in section 3.2 and 3.3). This variant is commonly referred to as partial VAE that has been exploited in Ma et al. (2018; 2019; 2020); Nazabal et al. (2020) to handle imputation task for heterogeneous tabular data. Mattei & Frellsen (2019) proposed MIWAE, which combines partial VAE with IWAE bound (Burda et al., 2016). MIWAE could offer more flexibility and produce more accurate imputed values. Based on MIWAE, Ipsen et al. (2021) proposed not-MIWAE, which can further explicitly model the missing mechanism and handle MNAR case.

DGMs with Causality There are several works integrating causality into DGMs. For GAN-based DGMs, there are CausalGAN (Kocaoglu et al., 2018), GCNN (Goudet et al., 2018), DECAF (van Breugel et al., 2021), and Causal-TGAN (Wen et al., 2022). These works all adopted a generator with causally aware structure. Specifically, CausalGAN aims at generating images from interventional distributions which do not naturally exist. GCNN is targeted at causal discovery. DECAF aims at causally debiasing and generating fair interventional data distribution. The goal of Causal-TGAN is to generate high-quality tabular data. Different from our work, Causal-TGAN limits its scope to generation task and do not provide flexibility or compatibility for imputation. For VAE-based DGMs, Yang et al. (2021) proposed CausalVAE to perform disentangled representation learning for images via causal neural layer. Karimi et al. (2020) used a collection of CVAEs to approximate causal posteriors and generate counterfactuals for algorithmic recourse. Besides, Shen et al. (2020) proposed DEAR, which combines VAE and adversarial training to learn the causal disentangled representation.

3 PRELIMINARIES

3.1 GENERATIVE PROCESS WITH CAUSALITY

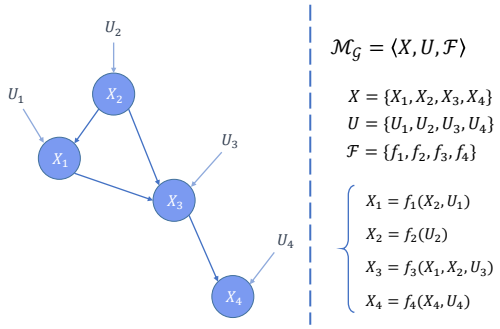

Figure 1: Causal graph $\mathcal{G}$ and SEM $\mathcal{M}_\mathcal{G}$.

Let $X = \{X_i\}_{i=1}^d \in \mathcal{X} \subseteq \mathbb{R}^d$ denote a set of random variables with distribution $p_X(X)$. We adopt structural equation model (SEM)(Pearl, 2009) to causally model the data generative process. We illustrate an example of SEM with 4 variables in Fig. 1. Formally, we formulate SEM as a triplet $\mathcal{M}_\mathcal{G} = \langle X, U, \mathcal{F} \rangle$. $\mathcal{G}$ denotes the underlying causal graph, which covers all causal relationships. X denotes the set of endogenous variables, and $U = \{U_i\}_{i=1}^d$ denotes the set of exogenous variables. Moreover, $\mathcal{F} = \{f_1\}_{i=1}^d$ denotes the set of causal equations. For a certain feature X_i, we can express its generating process as causal equation $X_i = f_i(X_{\mathrm{pa}_\mathcal{G}(i)}, U_i)$, where $X_{\mathrm{pa}_\mathcal{G}(i)}$ is the set of all the endogenous causal patents of X_i in $\mathcal{G}$ (namely, $\mathrm{pa}_\mathcal{G}(i)$ denotes corresponding indices). U_i covers all other unobserved causes of X_i. Note that f_i is actually a deterministic function that places all randomness of the conditional distribution $p(X_i|X_{\mathrm{pa}_\mathcal{G}(i)})$ in the exogenous variable U_i.

In this paper, we consider $\mathcal{M}_\mathcal{G}$ is causally sufficient, i.e., $\{U_i\}_{i=1}^d$ are mutually independent, like other works do (Karimi et al., 2020). In this case, the probability distribution on X satisfies the Markov Condition with respect to $\mathcal{G}$ (Pearl & Verma, 1995). One of the well-known forms is the factorization $p(X) = \prod_{i=1}^d p(X_i|X_{\mathrm{pa}_\mathcal{G}(i)})$, which we may refer to as $\mathrm{MC}_{\mathrm{factorization}}$. Besides, we consider the partial causal knowledge scenario. Namely, causal graph $\mathcal{G}$ is already known or pre-defined as expert knowledge, while the specific structural equations $\mathcal{M}_\mathcal{G}$ is not.

3.2 AUTO-ENCODER WITH AMORTIZED VARIATIONAL INFERENCE

Variational auto-encoder (Kingma & Welling, 2013) is a deep generative model with latent variables. The objective of VAE is to maximize an evidence lower bound (ELBO) on the marginal

log-likelihood, derived from Jensen's inequality as

$$\ell(\theta) = \log p_\theta(\mathbf{x}) = \log \mathbb{E}_{q_\gamma(\mathbf{z}|\mathbf{x})}\left[\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\gamma(\mathbf{z}|\mathbf{x})}\right] \geq \mathbb{E}_{q_\gamma(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\gamma(\mathbf{z}|\mathbf{x})}\right] = \mathcal{L}(\theta,\gamma), \tag{1}$$

where $\mathbf{z}$ is the latent variable. We could factorize the joint distribution $p_\theta(\mathbf{x},\mathbf{z})$ as $p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})$, where $p_\theta(\mathbf{x}|\mathbf{z})$ is the likelihood represented as the generative network (decoder) with parameters θ. $p(\mathbf{z})$ is the prior of latent variable. $q_\gamma(\mathbf{z}|\mathbf{x})$ is the variational posterior represented as the inference network (encoder) with parameters γ, introduced by amortized variational inference (Kingma & Welling, 2013). $q_\gamma(\mathbf{z}|\mathbf{x})$ serves as a proposal distribution to approximate the intractable true posterior $p_\theta(\mathbf{z}|\mathbf{x})$, for practically solving the expectation (integral) of the marginal likelihood. The equal sign in the inequality holds if and only if the variational posterior $q_\gamma(\mathbf{z}|\mathbf{x})$ exactly equals to the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The ELBO $\mathcal{L}(\theta,\gamma)$ could be estimated via Monte Carlo method.

3.3 Missing Value Imputation

When considering a data missing context, we can split each sample into an observed part and a missing part, i.e., $\mathbf{x} = (\mathbf{x}_{\boldsymbol{o}}, \mathbf{x}_{\boldsymbol{m}})$, where $\boldsymbol{o}$ and $\boldsymbol{m}$ denote the indices of observed and missing features, respectively.

In this paper, we consider MCAR and MAR assumptions. Mattei & Frellsen (2019) and Ipsen et al. (2021) stated that the impact of the missing mechanism could be ignored for both MCAR and MAR cases, and the optimization objective of likelihood could simply be pruned to $p_\theta(\mathbf{x}_{\boldsymbol{o}})$. Hence, we could substitute all $\mathbf{x}$ with $\mathbf{x}_{\boldsymbol{o}}$ in Eq. 1 to derive a lower bound of the observed part (partial ELBO). Further, combining with the importance-weighted auto-encoder (IWAE (Burda et al., 2016)), we push the Monte Carlo estimate inside the logarithm with the idea of importance sampling

$$\mathcal{L}^K(\theta,\gamma) = \mathbb{E}_{\mathbf{z}^1,\ldots,\mathbf{z}^K \sim q_\gamma(\mathbf{z}|\mathbf{x}_{\boldsymbol{o}})}\left[\log \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(\mathbf{x}_{\boldsymbol{o}}|\mathbf{z}^k)p(\mathbf{z}^k)}{q_\gamma(\mathbf{z}^k|\mathbf{x}_{\boldsymbol{o}})}\right], \tag{2}$$

which is also known as the MIWAE bound (Mattei & Frellsen, 2019). When $K = 1$, this bound resembles the standard (partial) ELBO. And Burda et al. (2016) proved that the larger K, the tighter the bound, i.e.,

$$\mathcal{L}^1(\theta,\gamma) \leq \ldots \leq \mathcal{L}^K(\theta,\gamma) \xrightarrow[K\to\infty]{} \ell(\theta). \tag{3}$$

With importance sampling, as studied by Cremer et al. (2017); Domke & Sheldon (2018), the variational distribution $q_\gamma(\mathbf{z}|\mathbf{x})$ would be replaced by a more complex distribution q_{IW} which depends both on θ and γ. Thus, IWAE could allow additional flexibility to train a decoder whose posterior does not need to fit assumptions well (Burda et al., 2016; Mattei & Frellsen, 2019).

Note that some pre-imputation shall be performed in advance, thereby filling in the missing features with dummy values in each data point so as to feed it to the encoder. As Mattei & Frellsen (2019) states, even a very rough pre-imputation is acceptable with the usage of MIWAE bound. Therefore, in this work, we conduct pre-imputation by replacing missing values with zeros.

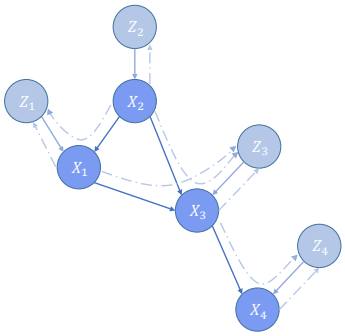

Figure 2: Graphical representation of the standard architecture (C3VAE) with augmented graph $\mathcal{G}'$. Solid arrows denote decoders, while dashdot arrows denote encoders.

4 Methodology

4.1 VAE with Causally Aware Architecture

For integrating causal knowledge in VAE-based model, a common intuition is to use the latent variable Z to represent the exogenous variable U. Given a true causal graph $\mathcal{G}$, under causal sufficiency assumption, we can thus manually reorganize Z as a collection of d independent components $\{Z_i\}_{i=1}^d$, where each component Z_i correspondingly accounts for the exogenous cause of X_i. This could also be visually represented as an augmented graph $\mathcal{G}'$ of $\mathcal{G}$, where $X_{\mathrm{pa}_{\mathcal{G}'}(i)} = \{X_{\mathrm{pa}_{\mathcal{G}}(i)}, Z_i\}$.

By this means, the generative process for each feature X_i can be considered as the conditional distribution $p(X_i|X_{\mathrm{pa}_{\mathcal{G}}(i)}, Z_i)$. The augmented graph $\mathcal{G}'$ essentially models the joint distribution $p(X, Z)$. For better illustration, We provide an example in Fig. 2.

With $\mathcal{G}'$, we review the ELBO in Eq. 1. Under extra causal assumptions, we conduct factorization on the feature space, finally deriving a lower bound (see appendix A for more details)

$$\mathcal{L}_{C3VAE}(\theta, \gamma) = \sum_{i=1}^{d} \mathbb{E}_{q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)})} \left[\log \frac{p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i) p(\mathbf{z}_i)}{q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i, \mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)})} \right], \tag{4}$$

where $\theta = \{\theta_i\}_{i=1}^d$ and $\gamma = \{\gamma\}_{i=1}^d$ are collections of the parameters of decoders and encoders, respectively. Each expectation in the above summation respectively corresponds to the ELBO of a conditional VAE (CVAE) modeling conditional distribution $p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)})$ (more interpretations in appendix A.1). This causal collection of total d separate CVAEs (we refer it as C3VAE), is certainly a standard solution for integration of causality. Fig. 2 illustrates an example of the graphical representation of C3VAE. In fact, this architecture has already been exploited by Karimi et al. (2020), which aims at learning the posterior to generate counterfactual distribution for algorithmic recourse.

However, C3VAE has an obvious limitation that it is only suitable for learning in fully observed data context. When some features are missing, the screened-off designation of encoders would not allow to make full use of those observed values.

4.2 CAT-BASED ARCHITECTURE

To construct a more flexible and compatible architecture, we rearrange the intermediate Eq. 14b in the deduction (appendix A) as

$$\mathcal{L}_{CAT\_VAE}(\theta, \gamma) = \sum_{i=1}^{d} \mathbb{E}_{q_{\gamma_i}(\mathbf{z}_i|\mathbf{x})} \left[\log \frac{p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i) p(\mathbf{z}_i)}{q_{\gamma_i}(\mathbf{z}_i|\mathbf{x})} \right]. \tag{5}$$

Thus, we propose to design a shared encoder E_γ, which takes the full data features $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^d$ as input to infer the posteriors of all $\{\mathbf{z}_i\}_{i=1}^d$ simultaneously. Whereas, we still leave multiple specialized decoders topologically structured according to causal graph $\mathcal{G}$, where each decoder D_{θ_i} generates a certain feature $\mathbf{x}_i$. Note that here we slightly abuse the notations. γ_i in Eq. 14b refers to partial parameters of the shared encoder E_γ, which takes charge of inferring $\mathbf{z}_i$.

This is exactly the key idea behind our "CAT" architecture. Though CAT-VAE seemingly does not precisely utilize causal knowledge to infer posteriors, its shared encoder structure can offer more flexibility instead, especially when handling missing values (more details in subsection 4.3). Besides, since neural networks are universal approximators, CAT-VAE could finally arrive at as close convergence as C3VAE, provided that their model capacities (especially for their encoders) are comparable.

Deterministic Structural Function Note that with this modeling, we actually assume a generative process $p(X_i|X_{\mathrm{pa}_{\mathcal{G}}(i)}, Z_i)$, which is probabilistic. To keep in consistency with the deterministic property of each causal equation f_i, we consider a conditional expectation, i.e.,

$$f_i(X_{\mathrm{pa}_{\mathcal{G}}(i)}, Z_i) = \mathbb{E}_{p_{\theta_i}}\left[X_i|X_{\mathrm{pa}_{\mathcal{G}}(i)}, Z_i\right], \tag{6}$$

which will be mainly adopted in the inference stage. In the training stage, we still focus on optimizing the probabilistic form to account for the uncertainty in the estimation.

Modeling Heterogeneous Features We consider two types of tabular features, discrete and continuous, respectively modeling their likelihood with categorical and Gaussian distribution. For more detailed processing and probability assumption, please refer to appendix A.2.

4.3 HANDLING MISSING VALUES

CAT-MIWAE Bound When some data are missing, as MIWAE bound in Eq. 2 suggested, ELBO should only be calculated on those observed features. However, some missing features may be an indispensable input (causal parent) for some generative processes $p(X_i|X_{\mathrm{pa}_{\mathcal{G}}(i)}, Z_i)$. To tackle that,

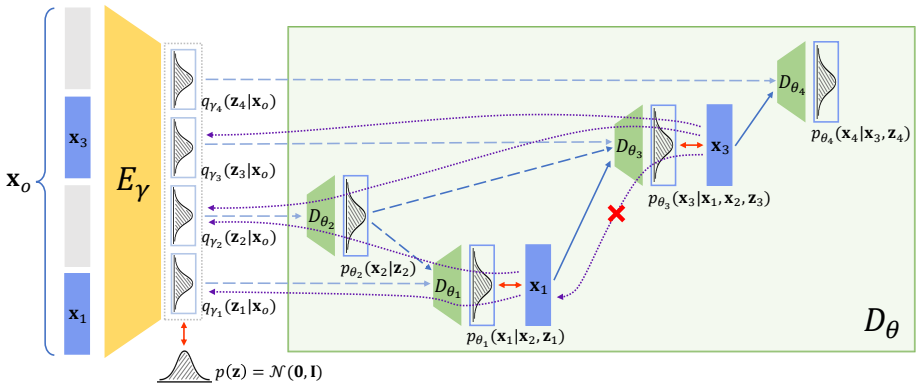

Figure 3: Overall architecture of our CAT-MIWAE with observed features $\mathbf{x}_{\boldsymbol{o}} = \{\mathbf{x}_1, \mathbf{x}_3\}$ and missing features $\mathbf{x}_{\boldsymbol{m}} = \{\mathbf{x}_2, \mathbf{x}_4\}$, based on previous example of Fig. 1. Solid arrows denote feedforward, dashed arrows denote reparameterization and feedforward, dotted arrows denote back propagation, red double arrows denote the calculation of CAT-MIWAE ELBO loss. Red cross denotes that the back propagation path is blocked.

we have to sample those missing values and fill in the vacancies, so as to ensure a thorough flow along the causal graph. In other words, we essentially integrate over both the latent space and the missing part of the data space.

With insights above, our proposed CAT-MIWAE bound could be acquired by reinforcing Eq. 14b with MIWAE, as

$$\mathcal{L}^K_{CAT\text{-}MIWAE}(\theta, \gamma) = \mathbb{E}\left[\log \frac{1}{K}\sum_{k=1}^{K} w_k\right], \tag{7}$$

where w_k is the unnormalized importance weight

$$w_k = \frac{\prod_{i\in\boldsymbol{o}} p_{\theta_i}(\mathbf{x}_i|\widetilde{\mathrm{PA}}^k_{\mathcal{G}}(\mathbf{x}_i), \mathbf{z}^k_i)\prod_{i=1}^d p(\mathbf{z}^k_i)}{\prod_{i=1}^d q_{\gamma_i}(\mathbf{z}^k_i|\mathbf{x}_{\boldsymbol{o}})}, \quad \widetilde{\mathrm{PA}}^k_{\mathcal{G}}(\mathbf{x}_i) = \{\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)\cap\boldsymbol{o}}, \mathbf{x}^k_{\mathrm{pa}_{\mathcal{G}}(i)\cap\boldsymbol{m}}\}, \tag{8}$$

where $(\{\mathbf{z}^1_i\}_{i=1}^d, \{\mathbf{x}^1_i\}_{i\in\boldsymbol{m}}), ..., (\{\mathbf{z}^K_i\}_{i=1}^d, \{\mathbf{x}^K_i\}_{i\in\boldsymbol{m}})$ are K i.i.d. samples from joint distribution via ancestral sampling

$$\prod_{i=1}^{d} q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_{\boldsymbol{o}}) \prod_{i\in\boldsymbol{m}} p_{\theta_i}(\mathbf{x}_i|\widetilde{\mathrm{PA}}_{\mathcal{G}}(\mathbf{x}_i), \mathbf{z}_i), \tag{9}$$

over which the expectation in Eq. 7 is taken. $\mathrm{pa}_{\mathcal{G}}(i) \cap \boldsymbol{o}$ denotes indices of the observed causal parents of i^{th} feature, while $\mathrm{pa}_{\mathcal{G}}(i) \cap \boldsymbol{m}$ refers to the missing ones.

Once CAT-MIWAE model has been trained, it can be used to impute missing values. We consider single imputation here and follow Mattei & Frellsen (2019). Using self-normalized importance sampling with proposal distribution in Eq. 9, imputed values would be estimated as

$$\widehat{\mathbf{x}}_{\boldsymbol{m}} = \mathbb{E}\left[\mathbf{x}_{\boldsymbol{m}}|\mathbf{x}_{\boldsymbol{o}}\right] \approx \sum_{k=1}^{K} a_k \mathbb{E}\left[\mathbf{x}_{\boldsymbol{m}}|\mathbf{z}^k\right], \quad a_k = \frac{w_k}{w_1 + ... + w_K}, \tag{10}$$

where a_k is the normalized weight and $w_1, ..., w_K$ are exactly the same as those used in training.

Interpretations of Design An example is illustrated in Fig. 3 for easier comprehension. Basically, we summarize the excellence of our CAT-MIWAE into two aspects. On one hand, the shared global encoder can be always aware of all the data features $\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^d$ when inferring posterior of any $\mathbf{z}_i$. As shown in Fig. 3, when feature $\mathbf{x}_2$ is missing, CAT-MIWAE can instead rely on its observed descendants $\mathbf{x}_1, \mathbf{x}_3$ to infer $\mathbf{z}_2$. However, on the contrary, the screened-off encoder $q_{\gamma_2}(\mathbf{z}_2|\mathbf{x}_2)$ in C3VAE could only shortsightedly perceive $\mathbf{x}_2$ itself. As a consequence, no meaningful information for $\mathbf{z}_2$ could be acquired from its pre-imputed dummy values.

On the other hand, the multiple structured decoders can subtly refine the posteriors through back propagation in accordance with causality. As depicted in Fig. 3, $q_{\gamma_2}(\mathbf{z}_2|\mathbf{x}_{\boldsymbol{o}})$ would be jointly updated by gradients backward from D_{θ_1} and D_{θ_3}. This is achieved by performing reparameterization, sampling $\mathbf{x}_2$, and then feeding forward. Note that the backward path from D_{θ_3} to $q_{\gamma_1}(\mathbf{z}_1|\mathbf{x}_{\boldsymbol{o}})$ is blocked, and this is reasonable since conditioning on $\mathbf{x}_1$ shall screen off all other dependency flows from its descendants to $\mathbf{z}_1$. Besides, for $\mathbf{z}_4$, as neither $\mathbf{x}_4$ nor any descendants are observed, there would be no extra knowledge about its posterior (i.e., posterior shall equal to its prior). As Eq. 8 suggests, the weight related to $\mathbf{z}_4$ is merely $p(\mathbf{z}_4)/q_{\gamma_4}(\mathbf{z}_4|\mathbf{x}_{\boldsymbol{o}})$, which resembles the Kullback-Leibler (KL) divergence, forcing posterior to approach prior. This is exactly what we are expecting.

It might be noticed that in the example of Fig. 3, the true posteriors for $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$ are not supposed to be mutually independent, since conditioning on $\mathbf{x}_1$ and $\mathbf{x}_3$ (but not on $\mathbf{x}_2$) would unblock paths. Thus, the factorized variational posterior $q_{\gamma_1}(\mathbf{z}_1|\mathbf{x}_{\boldsymbol{o}})q_{\gamma_2}(\mathbf{z}_2|\mathbf{x}_{\boldsymbol{o}})q_{\gamma_3}(\mathbf{z}_3|\mathbf{x}_{\boldsymbol{o}})$, which can never model the inter-dependency, might not fit the true joint posterior $p_\theta(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3|\mathbf{x}_{\boldsymbol{o}})$ well. This is exactly when IWAE comes into play, as we stated in section 3.3, IWAE could construct a more complex distribution q_{IW} to alleviate this unfitness and tighten the bound. We empirically show from our experiments that this factorized variational posterior modeling could lead to fair performance.

Extrapolation Let $\mathbf{x}_{\boldsymbol{cond}}$ denote arbitrary conditions and $\mathbf{x}_{\boldsymbol{int}}$ denote arbitrary interventions. Let the remaining features, which are neither conditioned nor intervened, be denoted as $\mathbf{x}_{\boldsymbol{rem}} := \mathbf{x}_{\setminus\{\boldsymbol{cond}\cup\boldsymbol{int}\}}$. In analogy to imputation, by feeding in $\mathbf{x}_{\boldsymbol{cond}}$ as observed values $\mathbf{x}_{\boldsymbol{o}}$, and further performing *do*-operator on $\mathbf{x}_{\boldsymbol{int}}$, we can acquire the extrapolated distribution as

$$p_{\gamma_i,\theta_i}(\mathbf{x}|do(\mathbf{x}_{\boldsymbol{int}}), \mathbf{x}_{\boldsymbol{cond}}) = \int \prod_{i=1}^{d} q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_{\boldsymbol{cond}}) \prod_{i\in\boldsymbol{rem}} p_{\theta_i}(\mathbf{x}_i|\widehat{\mathrm{PA}}_{\mathcal{G}}(\mathbf{x}_i), \mathbf{z}_i)\mathrm{d}\mathbf{z}, \tag{11}$$

where,

$$\widehat{\mathrm{PA}}_{\mathcal{G}}(\mathbf{x}_i) = \{\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)\cap\boldsymbol{int}}, \mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)\cap\{\boldsymbol{cond}\setminus\boldsymbol{int}\}}, \mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)\cap\boldsymbol{rem}}\}. \tag{12}$$

Therefore, $\mathbf{x}_{\boldsymbol{rem}}$ could be sampled ancestrally from the generative process defined inside the integral of Eq. 11. Formally, a newly sampled point is represented as $\mathbf{x} = \{\mathbf{x}_{\boldsymbol{int}}, \mathbf{x}_{\boldsymbol{cond}\setminus\boldsymbol{int}}, \mathbf{x}_{\boldsymbol{rem}}\}$.

Note that the interventions $\mathbf{x}_{\boldsymbol{int}}$ would only take part in $\widehat{\mathrm{PA}}_{\mathcal{G}}(\mathbf{x}_i)$ and affect their descendants, while the conditions $\mathbf{x}_{\boldsymbol{cond}}$ could more than that and they also indirectly influence their ancestors through the posterior $q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_{\boldsymbol{cond}})$. Besides, the interventions $\mathbf{x}_{\boldsymbol{int}}$ shall always have a higher priority than the conditions $\mathbf{x}_{\boldsymbol{cond}}$, where we use $\boldsymbol{cond}\setminus\boldsymbol{int}$ to denote those conditions which have not been overwritten by interventions. When there is an inclusion, i.e., $\boldsymbol{int} \subseteq \boldsymbol{cond}$, this could be interpreted as sampling counterfactuals for the subpopulation $\mathbf{x}_{\boldsymbol{cond}}$.

5 Experiments

We consider imputation and extrapolation for tabular data synthesis. For clear presentation, we defer the dataset description and implementation details to appendix B. Besides, we explore the effectiveness of CAT-based architecture on generation task. Due to space limitation, please refer to appendix C for more results on generation. During the experiments, all reported results is averaged over five random realizations.

5.1 Imputation Tasks

For each test, we create missing data of the dataset by randomly dropping out values with a fixed missing rate 50%. We then use this incomplete dataset to both train a DGM and test its imputation performance.

Metrics We follow Nazabal et al. (2020) and use the imputation error, consisting of normalized root mean square error (NRMSE) for continuous variables and classification error for discrete variables.

Comparisons To illustrate the superiority of the proposed model, we consider mean imputation, MICE (Van Buuren & Groothuis-Oudshoorn, 2011), MissForest (Stekhoven & Bühlmann, 2012), and the original MIWAE (Mattei & Frellsen, 2019) as the baselines. For our own models, we test CAT-MIWAE, C3MIWAE (vanilla combination of C3VAE architecture and MIWAE bound), and CAT-MIWAE-Cov (where we consider an unfactorizable posterior $q_\gamma(\mathbf{z}|\mathbf{x}_{\boldsymbol{o}})$, and we model it using

Table 1: Imputation Error on Six Datasets: Mean±Standard Deviation

	Adult	**Dutch**	**Credit**	**Bank**	**Cabs**	**King**
Mean	0.26±0.0	0.242±0.0	0.129±0.004	0.197±0.001	0.259±0.0	0.102±0.0
MICE	0.248±0.002	0.244±0.0	0.129±0.002	0.171±0.0	0.242±0.0	0.082±0.0
MissForest	0.231±0.003	0.225±0.002	0.128±0.003	0.157±0.001	0.224±0.0	0.074±0.002
MIWAE	0.156±0.0	0.161±0.0	0.099±0.002	0.134±0.001	0.175±0.0	**0.06±0.003**
C3MIWAE	0.16±0.003	0.163±0.001	0.1±0.0	0.136±0.0	0.176±0.001	0.064±0.001
CAT-MIWAE-Cov	0.153±0.001	0.162±0.001	0.099±0.0	0.133±0.0	0.175±0.001	0.063±0.001
CAT-MIWAE	**0.151±0.001**	**0.159±0.0**	**0.097±0.0**	**0.132±0.0**	**0.172±0.001**	0.062±0.0

Table 2: CATE Estimation of the Real Data and CAT-MIWAE Extrapolated Data on **Adult**

Conditions	Real	Extrapolated		
		$\widehat{\text{CATE}}$	$\mathbb{E}[Y(0)]$	$\mathbb{E}[Y(1)]$
None	0.281±0.003	0.271±0.006	0.177±0.005	0.448±0.004
$hours\text{-}per\text{-}week = 50$	0.334±0.007	0.334±0.004	0.213±0.003	0.547±0.005
$hours\text{-}per\text{-}week = 40$	0.276±0.005	0.26±0.004	0.164±0.003	0.424±0.003
$relationship = Husband$	0.381±0.004	0.38±0.005	0.278±0.003	0.658±0.003
$relationship = Husband$ & $hours\text{-}per\text{-}week = 50$	0.377±0.014	0.389±0.004	0.325±0.003	0.714±0.004
$relationship = Husband$ & $hours\text{-}per\text{-}week = 40$	0.411±0.004	0.406±0.006	0.290±0.004	0.696±0.003
$relationship = Husband$ & $hours\text{-}per\text{-}week = 50$ & $workclass = Private$	0.418±0.015	0.415±0.003	0.33±0.003	0.745±0.005

multivariate Gaussian with non-diagonal covariance to capture the inter-dependency among the posteriors of exogenous variables (more details in appendix A.2)). For all MIWAE-based models, we set $K = 5$ in the training, while using $K = 500$ for imputation.

Results Table 1 presents the performance for imputation measured by imputation error. We highlight the best results in bold style, while the second best ones with underline. All MIWAE-based models acquire significantly better results than other baselines. Compared to MIWAE, CAT-MIWAE mostly dominates (except for **King**) with a relative performance boost for 1% to 3%. This observation indicates that the introduction of causality could benefit imputation. Moreover, C3MIWAE almost always falls behind others with a relative performance degradation for 1% to 5%, justifying the superiority of CAT architecture for imputation task.

Note that though CAT-MIWAE-Cov seems to be more reasonable, it shows inferior results than CAT-MIWAE. Explicitly modeling the covariance introduces a large number of additional parameters, which might be difficult for optimization instead. Besides, this modeling could be computational-expensive, since its probability requires calculation of determinant. Thus, the factorized variational posterior of CAT-MIWAE would competently be an efficient solution in practice.

5.2 Extrapolation Tasks

To enable CAT-MIWAE better capture the data patterns under arbitrary conditions, we propose to use a dynamic missing mode. Specifically, we would sample a batch of complete data and then dynamically drop out values with missing rate from 10% to 90%, with 10% intervals.

Metrics We consider to compare the consistency of CATEs. We first choose certain interested subpopulation of subjects $\mathbf{x}_{\boldsymbol{cond}}$ and estimate CATE on its real data records. We implement The real CATE estimation by adjusting for confounders (Pearl, 2009) and then fitting a R-learner (Nie & Wager, 2021) with the CausalML library[1]. Then, for the same subpopulation, we follow the extrapolation process in subsection 4.3 and use our trained CAT-MIWAE to generate two interventional distributions respectively as the control group and the treatment group. Here $X_{\boldsymbol{int}}$ is exactly the treatment variable, with the treatment-unassigned value denoted as $\mathbf{x}_{\boldsymbol{int}(0)}$ and the treatment-assigned value as $\mathbf{x}_{\boldsymbol{int}(1)}$. Hence, the extrapolated CATE could be estimated as

$$\widehat{\text{CATE}} = \mathbb{E}[Y(1) - Y(0)|\mathbf{x}_{\boldsymbol{cond}}] = \mathbb{E}[Y|do(\mathbf{x}_{\boldsymbol{int}(1)}), \mathbf{x}_{\boldsymbol{cond}}] - \mathbb{E}[Y|do(\mathbf{x}_{\boldsymbol{int}(0)}), \mathbf{x}_{\boldsymbol{cond}}], \quad (13)$$

[1] https://causalml.readthedocs.io/en/latest/index.html

where Y is the outcome variable, while $Y(0)$ and $Y(1)$ denote the potential outcome for the untreated and the treated, respectively. And the expectation would be taken over the extrapolated distribution given in Eq. 11. For each group, we sample 5000 samples to estimate this expectation with Monte Carlo method.

We mainly rely on **Adult** dataset to conduct extrapolation tasks. The causal graph is given in appendix D. We choose $education$ as the treatment variable with the untreated value $HS\text{-}grad$ (high school graduation, $\mathbf{x}_{\boldsymbol{int}(0)}$) and treated value $Masters$ ($\mathbf{x}_{\boldsymbol{int}(1)}$). And $income$ is the outcome variable, with value 0 (low income, $\leq$ \$50K) and 1 (high income, $>$ \$50K). We choose $relationship = Husband$, $hours\text{-}per\text{-}week = 50$ or 40, $workclass = Private$, and several combinations of these categories as candidate conditions (subpopulation of interest). For real CATE estimation, we always adjust for confounders $race$, age, $native\text{-}country$, and sex.

Results The results of CATE estimation are presented in Table 2. When no condition is given, we estimate ATE, a special full-set case of CATE. We observe that the extrapolated CATE estimations keep in high consistency with the real ones, with absolute errors always less than 2%. Note that for several subpopulations with $hours\text{-}per\text{-}week = 50$, the estimated real CATEs show slightly high variances. This is because of their relatively small population base, making up only less than 8% of all the samples. Resorting to the extrapolation capability of CAT-MIWAE could potentially be an auxiliary solution to mitigate this uncertainty.

Note that though we train CAT-MIWAE with a dynamic missing mode to evaluate its extrapolation capability, this could be made more individualized when considering specific application. Having already defined our subpopulation of interest, we could exclusively design some training tricks (e.g., data augmentation approaches (Yoon et al., 2020)), to further enhance extrapolation performance.

Interpretations of Exogenous Representation In Table 3, we present KL divergence (KLD) between posteriors and priors given conditions "$relationship = Husband$, $workclass = Private$, and $hours\text{-}per\text{-}week = 50$". We highlight KLD higher than 0.05 in bold style. The exogenous posteriors of $workclass$ and $hours\text{-}per\text{-}week$ both show striking distortions with KLD of 0.17 and 1.4, thereby providing extra information for their reconstructions. It seems unexpected that $relationship$ stays insusceptible with only a KLD of 0.02. However, this could be understood by tracing back to its parents $marital\text{-}status$ and sex, whose posteriors are distorted with KLD of 0.22 and 0.16. We observe that these two exogenous posteriors could respectively produce values of $marital\text{-}status = Married\text{-}civ\text{-}spouse$ and $sex = Male$, which are sufficient to serve as the parental conditions to yield $relationship = Husband$. This allows the posterior of $relationship$ itself to be not that informative. Besides, the posteriors of age and $education$ are also affected to some extent. This easily makes sense since the role of a husband requires legal age of marriage, while workclass and working hours may require certain education levels. The above analysis shows that the exogenous causal representations learned by CAT-MIWAE could provide interpretablity for partially observed features.

Table 3: KL divergence of CAT-MIWAE on **Adult**, when conditioning on "$relationship = Husband$, $workclass = Private$, and $hours\text{-}per\text{-}week = 50$"

Feature	KL
age	**0.05**
$workclass$	**0.17**
$education$	**0.06**
$marital\text{-}status$	**0.22**
$occupation$	0.02
$relationship$	0.02
$race$	0.00
sex	**0.16**
$hours\text{-}per\text{-}week$	**1.40**
$native\text{-}country$	0.01
$income$	0.01

6 CONCLUSION

In this paper, we propose a novel "CAT" (Causal, Asymmetric, and Tabular) architecture for VAE-based DGMs. We instantiate CAT as CAT-MIWAE, which enables to learn the exogenous causal representation with a pre-defined causal graph in incomplete data context. Our proposed model provides interpretability for partially observed features and could efficiently address missing value imputation problem. We further show that CAT-MIWAE is able to extrapolate distributions given arbitrary conditions and interventions. We conduct extensive experiments on real-world tabular datasets to demonstrate the effectiveness of our proposed approach on generation, imputation, and extrapolation. Furthermore, CAT-MIWAE can be potentially applied to areas of counterfactual explanations, causal debiasing, etc., especially when interested in certain subpopulation of subjects.

REFERENCES

Yuri Burda, Roger B Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *ICLR (Poster)*, 2016.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pp. 286–305. PMLR, 2017.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

Justin Domke and Daniel R Sheldon. Importance weighting and variational inference. *Advances in neural information processing systems*, 31, 2018.

Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. In *Proceedings of NAACL-HLT*, pp. 240–250, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, pp. 39–80. Springer, 2018.

Niels Bruun Ipsen, Pierre-Alexandre Mattei, and Jes Frellsen. not-miwae: Deep generative modelling with missing not at random data. In *ICLR 2021-International Conference on Learning Representations*, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*. OpenReview. net, 2017.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in neural information processing systems*, 33:265–277, 2020.

Jayoung Kim, Jinsung Jeon, Jaehoon Lee, Jihyeon Hyeong, and Noseong Park. Oct-gan: Neural ode-based conditional tabular gans. In *Proceedings of the Web Conference 2021*, pp. 1506–1515, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Jihyeon Hyeong, Jinsung Jeon, Noseong Park, and Jihoon Cho. Invertible tabular gans: Killing two birds with one stone for tabular data synthesis. *Advances in Neural Information Processing Systems*, 34:4263–4273, 2021.

Steven Cheng-Xian Li, Bo Jiang, and Benjamin Marlin. Misgan: Learning from incomplete data with generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Chao Ma, Wenbo Gong, José Miguel Hernández-Lobato, Noam Koenigstein, Sebastian Nowozin, and Cheng Zhang. Partial vae for hybrid recommender system. In *NIPS Workshop on Bayesian Deep Learning*, volume 2018, 2018.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, Jose Miguel Hernandez-Lobato, Sebastian Nowozin, and Cheng Zhang. Eddi: Efficient dynamic discovery of high-value information with partial vae. In *International Conference on Machine Learning*, pp. 4234–4243. PMLR, 2019.

Chao Ma, Sebastian Tschiatschek, Richard Turner, José Miguel Hernández-Lobato, and Cheng Zhang. Vaem: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 33:11237–11247, 2020.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pp. 4413–4423. PMLR, 2019.

Alfredo Nazabal, Pablo M Olmos, Zoubin Ghahramani, and Isabel Valera. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10), 2018.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pp. 789–811. Elsevier, 1995.

Ignacio Peis, Chao Ma, and José Miguel Hernández-Lobato. Missing data imputation and acquisition with deep hierarchical models and hamiltonian monte carlo. *arXiv preprint arXiv:2202.04599*, 2022.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109:612–634, 2021.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *ArXiv*, abs/2010.02637, 2020.

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

Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Baohua Sun, Lin Yang, Wenhan Zhang, Michael Lin, Patrick Dong, Charles Young, and Jason Dong. Supertml: Two-dimensional word embedding for the precognition on structured tabular data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

Fei Tan, Xiurui Hou, Jie Zhang, Zhi Wei, and Zhenyu Yan. A deep learning approach to competing risks representation in peer-to-peer lending. *IEEE transactions on neural networks and learning systems*, 30(5):1565–1574, 2018.

Roger Tourangeau, T.J. Plewes, Division Education, and National Council. *Nonresponse in social science surveys: A research agenda*. 10 2013. doi: 10.17226/18293.

Dennis Ulmer, Lotta Meijerink, and Giovanni Ciná. Trust issues: Uncertainty estimation does not enable reliable ood detection on medical tabular data. In *ML4H@NeurIPS*, 2020.

Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, 34:22221–22233, 2021.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Bingyang Wen, Yupeng Cao, Fan Yang, Koduvayur Subbalakshmi, and Rajarathnam Chandramouli. Causal-tgan: Modeling tabular data using causally-aware gan. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32, 2019.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9588–9597, 2021.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

Jinsung Yoon, Yao Zhang, James Jordon, and Mihaela van der Schaar. Vime: Extending the success of self-and semi-supervised learning to tabular domain. *Advances in Neural Information Processing Systems*, 33:11033–11043, 2020.

Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017.

A DEDUCTION FOR VAE ELBO WITH CAUSALITY

We review ELBO in Eq. 1. With augmented graph $\mathcal{G}'$ modeling, we conduct factorization on the feature space

$$\mathbb{E}_{q_\gamma(\mathbf{z}|\mathbf{x})}\left[\log\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\gamma(\mathbf{z}|\mathbf{x})}\right] \tag{14}$$

$$= \mathbb{E}_{q_\gamma(\mathbf{z}|\mathbf{x})}\left[\log\frac{\prod_{i=1}^d p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)\prod_{i=1}^d p(\mathbf{z}_i)}{q_\gamma(\mathbf{z}|\mathbf{x})}\right] \tag{14a}$$

$$= \mathbb{E}_{\prod_{i=1}^d q_{\gamma_i}(\mathbf{z}_i|\mathbf{x})}\left[\log\frac{\prod_{i=1}^d p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)\prod_{i=1}^d p(\mathbf{z}_i)}{\prod_{i=1}^d q_{\gamma_i}(\mathbf{z}_i|\mathbf{x})}\right] \tag{14b}$$

$$= \mathbb{E}_{\prod_{i=1}^d q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i,\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})} \left[\log\frac{\prod_{i=1}^d p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)\prod_{i=1}^d p(\mathbf{z}_i)}{\prod_{i=1}^d q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i,\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})}\right] \tag{14c}$$

$$= \sum_{i=1}^d \mathbb{E}_{q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i,\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})}\left[\log\frac{p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)p(\mathbf{z}_i)}{q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i,\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})}\right]. \tag{14d}$$

Eq. 14a could be derived by

1. conducting $\mathrm{MC}_{\mathrm{factorization}}$ on the joint distribution $p_\theta(\mathbf{x},\mathbf{z})$ with respect to graph $\mathcal{G}'$;
2. substituting all $\mathbf{x}_{\mathrm{pa}_{\mathcal{G}'}(i)}$ with $\{x_{\mathrm{pa}_\mathcal{G}(i)}, z_i\}$.

Likewise, Eq. 14b and Eq. 14c are derived as a factorization of the exogenous posterior. Eq. 14b holds due to the fact that observing all features X ensures independence among all exogenous variables. Eq. 14c holds due to the fact that conditioning on X_i and $X_{\mathrm{pa}_\mathcal{G}(i)}$ would sufficiently block (screen off) all dependency flows to Z_i (thus, knowing X_i and $X_{\mathrm{pa}_\mathcal{G}(i)}$ is enough to infer its posterior). Note that this factorization should have inherently applied to the true posterior $p_\theta(\mathbf{z}|\mathbf{x})$. However, as we expect the variational posterior $q_\gamma(\mathbf{z}|\mathbf{x})$ to be a good approximator of $p_\theta(\mathbf{z}|\mathbf{x})$, it is reasonable to assume that $q_\gamma(\mathbf{z}|\mathbf{x})$ shall also follow the same factorization.

A.1 C3VAE ARCHITECTURE

In Eq. 14d, each expectation in the summation exactly corresponds to the ELBO of a conditional VAE (CVAE) model with decoder $D_{\theta_i} := p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)$, encoder $E_{\gamma_i} := q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}_i,\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})$, and prior $p(\mathbf{z}_i)$. As a matter of fact, this CVAE ultimately attempts to maximize the conditional log-likelihood

$$p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)}) = \int p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)p(\mathbf{z}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})\mathrm{d}\mathbf{z} = \int p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i)p(\mathbf{z}_i)\mathrm{d}\mathbf{z}, \tag{15}$$

where Eq. 15 holds due to the independence between $\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)}$ and $\mathbf{z}_i$ (unless conditioned on $\mathbf{x}_i$).

Note that essentially, this CVAE ELBO could be achieved in another equivalent but more explicit way. By directly conducting $\mathrm{MC}_{\mathrm{factorization}}$ on the marginal likelihood of Eq. 1 w.r.t. $\mathcal{G}$, we get $\ell(\theta) = \log p_\theta(\mathbf{x}) = \sum_{i=1}^d \log p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)})$. We still tend to present the more detailed deduction in Eq. 14 for ease of explanations on our CAT-based model designation.

A.2 PROBABILISTIC ASSUMPTION AND FEATURE PROCESSING

For the latent space, we assume each prior as standard multivariate Gaussian, $p(\mathbf{z}_i) = \mathcal{N}(\mathbf{0},\mathbf{I})$, and each posterior as multivariate Gaussian with diagonal covariance matrix, $q_{\gamma_i}(\mathbf{z}_i|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\gamma_i}(\mathbf{x}),\boldsymbol{\Lambda}_{\gamma_i}(\mathbf{x}))$. $\boldsymbol{\mu}_{\gamma_i}$ and $\boldsymbol{\Lambda}_{\gamma_i}$ denote the mappings of encoder for mean and diagonal covariance, respectively.

For the feature space, we note that tabular data consists of mixed types of variables, which shall be treated separately. We model each discrete variable as categorical distribution $p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_\mathcal{G}(i)},\mathbf{z}_i) =$

$Cat(\boldsymbol{\alpha}_{\theta_i}(\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i))$, where $\boldsymbol{\alpha}_{\theta_i}$ denotes the mapping of decoder for outputing discrete probability vector as its parameters. For each continuous variable, we standardize its values into [-1, 1] and model with Gaussian distribution $p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i) = \mathcal{N}(\mu_{\theta_i}(\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i), \sigma^2_{\theta_i}(\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i))$, where μ_{θ_i} and $\sigma^2_{\theta_i}$ denote the mappings of decoder for scalar mean and variance, respectively.

For generation and extrapolation tasks, we follow the mode-specific normalization of Xu et al. (2019) to capture more patterns and produce richer diversity of continuous variables. Concretely, a variational Gaussian mixture model (VGM) is fit so that each continuous value would be transformed into: 1) a one-hot vector indicating which component it belongs to; 2) a single value normalized in [-1, 1] using parameters of that Gaussian component. Thus, we finally model each continuous variable as a joint distribution

$$p_{\theta_i}(\mathbf{x}_i|\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i) = \mathcal{N}(\mu_{\theta_i}(\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i), \sigma^2_{\theta_i}(\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i))Cat(\boldsymbol{\alpha}_{\theta_i}(\mathbf{x}_{\mathrm{pa}_{\mathcal{G}}(i)}, \mathbf{z}_i)). \tag{16}$$

For CAT-MIWAE-Cov, we model the joint variational posterior as multivariate Gaussian with non-diagonal cavariance, as

$$q(\mathbf{z}|\mathbf{x_o}) = \mathcal{N}(\boldsymbol{\mu}_\gamma(\mathbf{x_o}), \boldsymbol{\Sigma}_\gamma(\mathbf{x_o})), \tag{17}$$

where $\boldsymbol{\Sigma}_\gamma(\mathbf{x_o})$ is a positive definite matrix yielded by

$$\boldsymbol{\Sigma}_\gamma(\mathbf{x_o}) = \boldsymbol{L}_\gamma(\mathbf{x_o})\boldsymbol{L}^T_\gamma(\mathbf{x_o}) + \epsilon\mathbf{I}. \tag{18}$$

$\boldsymbol{L}_\gamma(\mathbf{x_o})$ is the raw output matrix of encoder. ϵ is a small positive real number to ensure the positive definiteness of $\boldsymbol{\Sigma}_\gamma(\mathbf{x_o})$ to serve as covariance.

B Experimental Configuration

B.1 Datasets

We choose six real-world tabular datasets to test the performance of our proposed approach, which are listed as follow

- **`Adult`**[2]: composed of diverse demographic information in the U.S. from the 1994 Census Survey. The task is to predict two classes of high ($>$ \$50K) and low ($\leq$ \$50K) income.
- **`Dutch`**[3]: extracted from The Dutch Virtual Census of 2001 program conducted by Statistics Netherlands, where only anomynised values are available. The task is to predict occupation status.
- **`Credit`**[4]: for prediction of the future loan status.
- **`Bank`**[5]: personal loans data provided by Zhongyuan bank in the CCF Big and Computing Intelligence Contest, for predicting if the loan is default.
- **`Cabs`**[6]: collected by an Indian cab aggregator service company for predicting the types of customers.
- **`King`**[7]: house sale data for King County in Seattle between May 2014 and May 2015, for predicting house price.

We remove irrelevant columns (e.g., "id", "date") in some of the datasets, and select a portion of all records to serve as the entire set for our experiments. A summary is presented in Table 4, where **C** denotes classification, **R** denotes regression.

Regarding the pre-defined causal graph $\mathcal{G}$, for **`Adult`** and **`Dutch`**, we use the graphs discovered and presented by Zhang et al. (2017). For the other datasets, we use PC algorithm Spirtes et al. (2000) with the pycausal library[8] to estimate their causal graphs.

[2] `http://archive.ics.uci.edu/ml/datasets/adult`
[3] `https://sites.google.com/site/faisalkamiran/`
[4] `https://www.kaggle.com/zaurbegiev/my-dataset`
[5] `https://www.datafountain.cn/competitions/530/datasets`
[6] `https://www.kaggle.com/arashnic/taxi-pricing-with-mobility-analytics`
[7] `https://www.kaggle.com/harlfoxem/housesalesprediction`
[8] `http://www.phil.cmu.edu/tetrad/index.html`

Table 4: Summary of the Datasets used in our Experiments

Name	# records	# continuous	# discrete	Task
Adult	30k	3	8	**C**
Dutch	60k	0	12	**C**
Credit	29k	12	5	**C**
Bank	10k	17	12	**C**
Cabs	41k	6	7	**C**
King	21k	12	4	**R**

B.2 Implementation Details

We use Python 3.7 with PyTorch 1.8.1 to implement our CAT-MIWAE models. We construct each individualized decoder as two-layer fully connected network, where the dimensions of the hidden layers $\dim\_\text{hidden}(D_{\theta_i})$ are set to 32. For the encoder, we instantiate it with a multi-headed network architecture. The first two layers are shared by all posteriors, with dimension adaptively set as $\dim\_\text{hidden}(D_{\theta_i}) \times d$ (hopefully to match the network capacities with those decoders). Then, the following structure of encoder branches into d heads where each head consists of two-layer with the same dimension with decoders $\dim\_\text{hidden}(D_{\theta_i})$. We set the dimension of each latent variable $\dim(\mathbf{z}_i)$ to 2. We use Gumbel softmax Jang et al. (2017) to ensure reparameterization for discrete missing variables in the training stage. We train all our models using Adam optimizer Kingma & Ba (2015) with learning rate 10^{-3}.

Additionally, we observe that in both C3VAE and our CAT-based architecture, applying the naive training procedure is vulnerable to KL vanishing, leading to uninformative posteriors. This may result from the concatenation of latent variable and high dimensional conditions vector (causal parents) at the input of decoder. Additional information provided by the causal parents could make the latent variable prone to be ignored and degrade to prior. Thus, we use the cyclical annealing schedule (Fu et al., 2019), which cyclically adjusts the weight of KL term during training, to alleviate this issue. We implement it by setting a scale factor β as the exponent of likelihood in Eq. 8, i.e.,

$$w_k = \frac{\prod_{i \in \boldsymbol{o}} p_{\theta_i}(\mathbf{x}_i | \widetilde{\text{PA}}_{\mathcal{G}}^k(\mathbf{x}_i), \mathbf{z}_i^k)^\beta \prod_{i=1}^d p(\mathbf{z}_i^k)}{\prod_{i=1}^d q_{\gamma_i}(\mathbf{z}_i^k | \mathbf{x}_{\boldsymbol{o}})}. \tag{19}$$

In the t^{th} epoch, β is set as

$$\beta = \begin{cases} 2 - \sin \frac{\pi(t \bmod \tau)}{2\Theta} & if\ t \bmod \tau < \Theta \\ 1 & otherwise \end{cases}, \tag{20}$$

where τ is the period, and Θ is the threshold. In the experiments, we set τ to 200 and Θ to 100.

C Experimental Results for Generation Tasks

For each test, we run five-fold cross validation, where each fold consists of 80% data as training set and the rest 20% as test set (validated also on training set).

Metrics We follow the evaluation protocol of machine learning efficacy (MLE) in Xu et al. (2019). We first use trained DGM to generate fake tabular data, on which multiple classifiers (Decision Tree, AdaBoost, MLP, and Losgistic Regression) or regressors (Linear Regression and MLP) are trained. Then, we evaluate the performance of these learners using the test set with several metrics.

Comparisons We choose three typical models for tabular data synthesis, CTGAN, TVAE (Xu et al., 2019), and Causal-TGAN (Wen et al., 2022). For our own models, we test both CAT-VAE and C3VAE.

Results Table 5 and 6 present the performance for generation measured by MLE. We highlight the best results in bold style, while the second best with underline.

In all 14 cases(total 14 evaluation metrics for 6 datasets), C3VAE or CAT-VAE could ensure at least the second best performance, while achieving the best for 11 cases (for F1 on **Credit**, even outperform the real). This demonstrates the effectiveness of our proposed approaches on generation task.

Table 5: Machine Learning Efficacy on Binary Classification Datasets: Mean±Standard Deviation

	Adult		**Dutch**		**Credit**		**Bank**	
	ROCAUC	F1	ROCAUC	F1	ROCAUC	F1	ROCAUC	F1
Real	0.864±0.011	0.738±0.012	0.873±0.051	0.803±0.047	0.712±0.069	0.635±0.06	0.802±0.079	0.622±0.075
CTGAN	0.849±0.012	0.692±0.041	0.859±0.048	0.78±0.047	0.671±0.063	0.603±0.061	0.765±0.081	0.593±0.073
TVAE	0.848±0.011	0.709±0.022	0.848±0.053	0.776±0.053	0.697±0.06	0.639±0.045	**0.794±0.079**	0.597±0.073
Causal-TGAN	0.837±0.01	0.72±0.01	0.859±0.051	0.784±0.051	0.69±0.063	0.627±0.045	0.754±0.112	**0.613±0.074**
C3VAE	0.851±0.019	0.715±0.03	**0.867±0.05**	**0.796±0.049**	0.699±0.056	0.642±0.028	0.793±0.08	0.602±0.073
CAT-VAE	**0.860±0.011**	**0.721±0.021**	**0.867±0.05**	0.794±0.045	**0.704±0.061**	**0.642±0.034**	0.793±0.08	0.599±0.077

Table 6: Machine Learning Efficacy on Multi-class Classification and Regression Datasets: Mean±Standard Deviation

	Cabs			**King**		
	Accuracy	Macro F1	Micro F1	R2	MSE	Exp.Var
Real	0.728±0.007	0.701±0.009	0.725±0.008	0.595±0.053	0.044±0.011	0.596±0.053
CTGAN	0.657±0.034	0.621±0.041	0.645±0.042	0.402±0.17	0.053±0.015	0.451±0.13
TVAE	0.615±0.029	0.568±0.036	0.607±0.033	0.334±0.415	0.054±0.014	0.357±0.375
Causal-TGAN	0.699±0.018	**0.671±0.017**	0.697±0.018	0.459±0.05	0.051±0.012	0.462±0.049
C3VAE	0.704±0.012	0.665±0.022	**0.698±0.013**	**0.464±0.066**	0.051±0.012	**0.467±0.065**
CAT-VAE	**0.707±0.023**	0.65±0.061	0.697±0.034	0.463±0.061	**0.051±0.011**	0.465±0.061

Note that C3VAE and CAT-VAE always show consistently high performance with only imperceptible gaps (always less than 0.01). This observation almost verifies our analysis in 4.2 that though CAT-based architecture seemingly do not make full use of causal knowledge to infer posteriors, it could still converge to similar points and learn similar SEMs as C3VAE architecture.

We can conclude that the introduction of extra causal knowledge could indeed help improve the quality of generated tabular data, since Causal-TGAN also achieves fair results on almost all F1 metrics, especially Macro F1 on **Cabs**. This indicates that Causal-TGAN may better pay attention to the data imbalance of **Cabs**, while as a contrast, TVAE fails to handle this, showing a performance degradation even greater than 0.1. Besides, TVAE suffers extremely high variance on **King**, this results from negative R2 scores of certain folds.

D CAUSAL GRAPH FOR ADULT DATASET

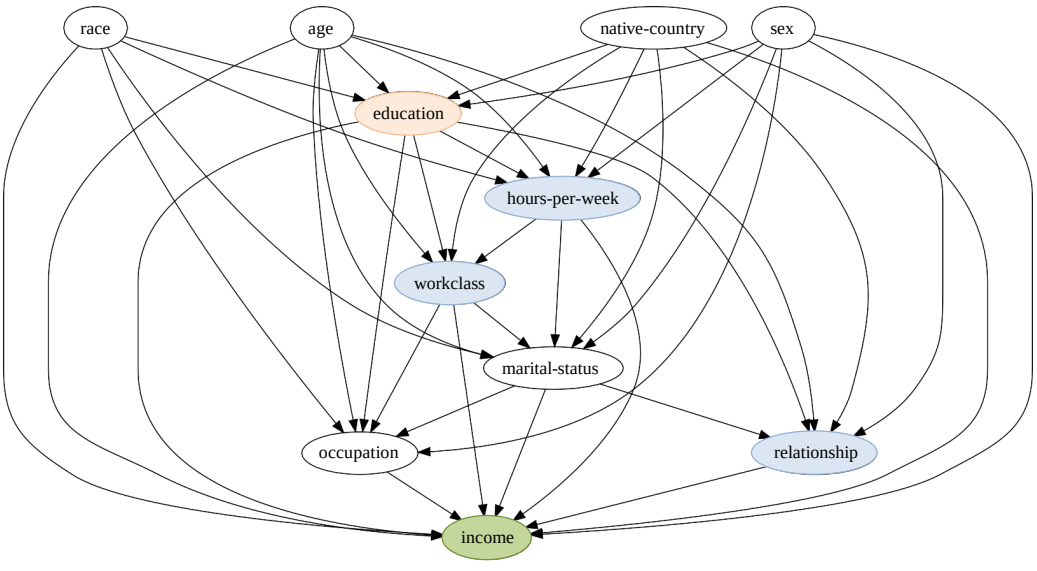

Figure 4: Causal graph used for **Adult** dataset. The treatment variable is in orange ($education$), the outcome variable in green ($income$), and the potential conditional variables in blue ($hours\text{-}per\text{-}week$, $workclass$, $relationship$).