Generating High-Diversity Synthetic Tabular Data via Less-Constrained Prior

Sanghun Park[1,*], **Jaesung Lim**[1], **Jong-June Jeon**[1] and **Seunghwan An**[2,†]
[1]University of Seoul, South Korea
[2]Incheon National University, South Korea
{sh.park, js.lim, jj.jeon}@uos.ac.kr, sh.an@inu.ac.kr

Abstract

Generating high-quality synthetic tabular data, regarding fidelity, diversity, and utility, is crucial for many practical purposes. Recent tabular data synthesis methods, including two-stage generative modeling approaches, have achieved this with nearly perfect fidelity. However, we observe that there remains room for improving diversity, and we show that this limitation arises from an additional constraint imposed on the latent support. Our main contribution is that we effectively eliminate this redundant constraint by directly deriving the objective function from the KL-divergence between the ground-truth density and the generative model used for synthetic sample generation. We empirically demonstrate that our model, which relies on a single prior distribution, significantly improves the quality of the synthetic data, especially in terms of diversity. Our implementation code is available at https://github.com/Optim-Lab/SPT.

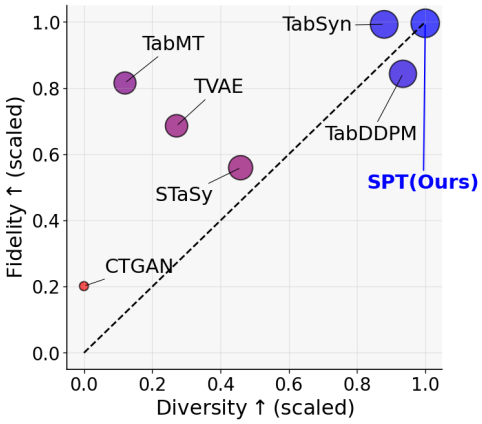

Figure 1: Trade-off between the averaged four fidelity metrics and the averaged three diversity metrics in synthetic tabular data generation, where the circle size denotes the TOPSIS score (overall performance). SPT (Ours) simultaneously achieves high fidelity and strong diversity. All metric scores are min-max scaled.

1 Introduction

Tabular data is one of the most commonly encountered data types across diverse domains [Jordon *et al.*, 2018], and tabular data synthesis is useful for a wide range of purposes [Hernandez *et al.*, 2022; Miletic and Sariyar, 2024]. This has led to extensive research on tabular data synthesis, which requires estimating the multivariate joint distribution while handling complex dependency structures. To address this, the Variational AutoEncoder (VAE) has been widely adopted, which combines the principle of maximum log-likelihood estimation with latent spaces that capture dependency structures in the data space [Xu *et al.*, 2019; An and Jeon, 2023; Liu *et al.*, 2023; Zhang *et al.*, 2024].

In VAE-based approaches, reducing the discrepancy between the prior and the aggregated (i.e., marginalized) posterior distribution is critical for generating high-quality synthetic samples [Li *et al.*, 2020; Aneja *et al.*, 2021]. To achieve this alignment, a promising approach is the *two-stage generative modeling* framework, which reduces the discrepancy by employing a trainable prior while maintaining a stable training process [Li *et al.*, 2015; Dai and Wipf, 2019; Razavi *et al.*, 2019; Rombach *et al.*, 2022]. In this context, [Zhang *et al.*, 2024] proposed TabSyn, a two-stage VAE-based model that demonstrates strong synthetic data generation performance.

On the other hand, the quality of synthetic data is commonly evaluated along three dimensions: (1) fidelity, (2) diversity, and (3) utility [Kim *et al.*, 2023; Liu *et al.*, 2023]. It is well known that fidelity and diversity are regarded as trade-offs, and that most existing methods achieve remarkable statistical fidelity [Hansen *et al.*, 2023]. However, *our empirical results indicate that there remains substantial room for improving diversity while maintaining fidelity*. Diversity is determined by the extent to which the generative model covers the support of the true data distribution [Naeem *et al.*, 2020; Alaa *et al.*, 2022], and that, in latent-variable models, the support of the generative model is also determined by the effective latent support.

In this context, our main contributions are:

1. We show that existing two-stage modeling approaches

*Current address: Gwanak Lab Co., Ltd., South Korea.
†Corresponding author.

impose an additional constraint on the latent support.

2. We propose a new method in which this redundant constraint is effectively eliminated.
3. We demonstrate that our proposed approach significantly improves the quality of synthetic data, particularly in terms of diversity.

Our proposed objective function is directly derived from the KL-divergence between the ground-truth density and the generative model used for synthetic sample generation, relying on a single prior distribution. Our proposed method is termed **SPT** (Single Prior Tabular data synthesis). Note that previous studies have shown that over-regularization by a simple prior can lead to posterior collapse, distort latent structures (e.g., in clustering), and under-represent low-density regions, which degrades sample quality [Bowman *et al.*, 2016; Dilokthanakul *et al.*, 2016; Kim *et al.*, 2024].

As illustrated in Figure 1, which summarizes fidelity, diversity, and overall performance, our proposed model achieves superior fidelity and diversity scores while maintaining high overall performance.

2 Preliminary and Motivation

Notations. A random variable $\mathbf{x}$ denotes an observation with p attributes comprising both continuous and discrete variables (i.e., mixed-type), and its jth coordinate is denoted as $\mathbf{x}_j$. The ground-truth probability density function of $\mathbf{x}$ is represented by p^*. A lower-dimensional latent representation is denoted by $\mathbf{z} \in \mathbb{R}^d$.

Preliminary. The objective function of the conventional two-stage generative model can be written as:

$$\min_{\theta} \quad \mathcal{KL}\Big(p^*(\mathbf{x}) \Big\| \int p(\mathbf{x} \mid \mathbf{z}; \theta)\, p(\mathbf{z})\, d\mathbf{z}\Big) \quad \cdots \text{(VAE)}$$

$$\min_{\eta} \quad \mathcal{KL}\Big(q(\mathbf{z}; \phi) \,\big\|\, p(\mathbf{z}; \eta)\Big), \ \cdots \text{(latent alignment)} \quad (1)$$

where $p(\mathbf{x} \mid \mathbf{z}; \theta)$, $p(\mathbf{z})$, and $p(\mathbf{z}; \eta)$ denote the decoder, fixed prior, and trainable prior distributions, respectively. $q(\mathbf{z}; \phi)$ is the aggregated posterior, defined as $q(\mathbf{z}; \phi) \coloneqq \int q(\mathbf{z} \mid \mathbf{x}; \phi)\, p^*(\mathbf{x})\, d\mathbf{x}$ where $q(\mathbf{z} \mid \mathbf{x}; \phi)$ is the posterior (or encoder), and it serves as the optimal prior that maximizes the ELBO [Tomczak and Welling, 2018].

The two problems in (1) are optimized sequentially, and for the first problem (i.e., stage 1), its upper bound

$$\begin{aligned} & \mathcal{KL}\Big(p^*(\mathbf{x})\, q(\mathbf{z} \mid \mathbf{x}; \phi) \,\big\|\, p(\mathbf{x} \mid \mathbf{z}; \theta)\, p(\mathbf{z})\Big) \\ = \ & \mathbb{E}_{\mathbf{x},\mathbf{z}}\big[-\log p(\mathbf{x} \mid \mathbf{z}; \theta)\big] + \mathbb{E}_{\mathbf{x}}\big[\mathcal{KL}\big(q(\mathbf{z} \mid \mathbf{x}; \phi) \,\|\, p(\mathbf{z})\big)\big] \\ & - H(p^*(\mathbf{x})) \end{aligned} \quad (2)$$

is minimized and is equivalent to the VAE framework, where the expectations of $\mathbb{E}_{\mathbf{x},\mathbf{z}}$ and $\mathbb{E}_{\mathbf{x}}$ are taken over $p^*(\mathbf{x})q(\mathbf{z} \mid \mathbf{x}; \phi)$ and $p^*(\mathbf{x})$, respectively, and $H(\cdot)$ is the entropy.

The second optimization problem of (1) (i.e., stage 2) is *additionally* integrated into the conventional VAE framework to align the aggregated posterior with the trainable prior distribution. The trainable prior distribution can be parameterized with various generative models, such as GANs, VAEs, normalizing flows, or diffusion models [Choi *et al.*, 2017; Dai and Wipf, 2019; Xie *et al.*, 2023; Zhang *et al.*, 2024].

Motivation. In the formulation of (1), there are two distinct prior distributions: $p(\mathbf{z})$ and $p(\mathbf{z}; \eta)$. And the generative model $p(\mathbf{x}; \theta, \eta)$, which is used for synthetic data generation, is defined with $p(\mathbf{z}; \eta)$:

$$p(\mathbf{x}; \theta, \eta) \coloneqq \int p(\mathbf{x}|\mathbf{z}; \theta) p(\mathbf{z}; \eta) d\mathbf{z} \approx \int p(\mathbf{x}|\mathbf{z}; \theta) q(\mathbf{z}; \phi) d\mathbf{z},$$

where this approximation is carried out using the KL-divergence in stage 2, which means that $q(\mathbf{z}; \phi)$ determines the distributional characteristics of $p(\mathbf{z}; \eta)$, including its effective latent support. Moreover, the non-informative prior in stage 1, $p(\mathbf{z})$, serves solely to regularize the representation mapping and is not directly used in the generative model. Based on these observations, our research is guided by the following motivation: *Is the non-trainable, fixed prior distribution $p(\mathbf{z})$ essential for efficient two-stage generative modeling?*

3 Proposal

In this section, we introduce our proposed method, SPT, to address the above research question.

3.1 Two-stage Modeling with Single Prior

To derive a two-stage generative model without a redundant prior distribution, we directly begin with *the single KL-divergence between the ground-truth density $p^*(\mathbf{x})$ and the generative model $p(\mathbf{x}; \theta, \eta)$*, which is the ultimate goal of the generative modeling:

$$\begin{aligned} \min_{\theta,\eta} \quad & \mathcal{KL}\Big(p^*(\mathbf{x}) \,\big\|\, p(\mathbf{x}; \theta, \eta)\Big) \\ = \ & \mathcal{KL}\left(p^*(\mathbf{x}) \,\middle\|\, \int p(\mathbf{x} \mid \mathbf{z}; \theta)\, p(\mathbf{z}; \eta)\, d\mathbf{z}\right) \\ \leq \ & \mathcal{KL}\Big(p^*(\mathbf{x})\, q(\mathbf{z} \mid \mathbf{x}; \phi) \,\big\|\, p(\mathbf{x} \mid \mathbf{z}; \theta)\, q(\mathbf{z}; \phi)\Big) \\ + \ & \mathcal{KL}\Big(q(\mathbf{z}; \phi) \,\big\|\, p(\mathbf{z}; \eta)\Big) \end{aligned} \quad (3)$$

by Jensen's inequality (see Appendix for the detailed derivation). Note that the decoder variance parameter β, which also serves as a scheduling parameter, is included in θ.

Since the first term of the upper bound in (3) does not involve the parameter η, our approach naturally leads to the following two-stage generative modeling framework:

Stage 1: Without redundant constraint

$$\begin{aligned} \min_{\theta,\phi} \quad & \mathcal{KL}\Big(p^*(\mathbf{x})\, q(\mathbf{z} \mid \mathbf{x}; \phi) \,\big\|\, p(\mathbf{x} \mid \mathbf{z}; \theta)\, q(\mathbf{z}; \phi)\Big) \\ = \ & \mathbb{E}_{\mathbf{x},\mathbf{z}}\big[-\log p(\mathbf{x} \mid \mathbf{z}; \theta)\big] - H(p^*(\mathbf{x})) \\ + \ & \mathbb{E}_{\mathbf{x}}\big[\mathcal{KL}\big(q(\mathbf{z} \mid \mathbf{x}; \phi) \,\|\, q(\mathbf{z}; \phi)\big)\big] \end{aligned} \quad (4)$$

In (4), the first term is the reconstruction loss, the second term is constant, and the last term serves as a regularizer for the posterior distribution.

Stage 2: Diffusion models

$$\min_{\eta} \quad \mathcal{KL}\big(q(\mathbf{z}; \phi) \,\big\|\, p(\mathbf{z}; \eta)\big). \quad (5)$$

On the other hand, the objective function of stage 2 in (5) remains the same as that of the existing method described in Section 2. After training our two-stage framework, synthetic data are generated via an ancestral sampling procedure.

3.2 Less-Constrained Prior

From (2) and (4), the KL-divergence regularization imposed on the latent space by existing methods and our proposed method can be compared as follows:

$$\underbrace{\mathbb{E}_{\mathbf{x}}\big[\mathcal{KL}\big(q(\mathbf{z} \mid \mathbf{x};\phi) \,\|\, p(\mathbf{z})\big)\big]}_{\text{in existing methods}} \tag{6}$$

$$= \underbrace{\mathbb{E}_{\mathbf{x}}\big[\mathcal{KL}\big(q(\mathbf{z} \mid \mathbf{x};\phi) \,\|\, q(\mathbf{z};\phi)\big)\big]}_{\text{in our approach}} + \underbrace{\mathcal{KL}\big(q(\mathbf{z};\phi) \,\|\, p(\mathbf{z})\big)}_{\textbf{extra regularization}},$$

where the expectation is taken over $p^*(\mathbf{x})$.

This implies that the additional latent regularization term $\mathcal{KL}\big(q(\mathbf{z};\phi) \,\|\, p(\mathbf{z})\big)$ is effectively canceled under our proposed method, unlike in existing approaches. Since $p(\mathbf{z})$ is a non-informative prior, such regularization forces the latent distribution $q(\mathbf{z};\phi)$ to capture less information from the input data, which in turn reduces the diversity of latent variables generated by $p(\mathbf{z};\eta)$ [Zhao *et al.*, 2019; Rezaabad and Vishwanath, 2020].

Theoretical Analysis. (6) further indicates that the stage 1 objective in the existing two-stage generative modeling can be interpreted as a constrained optimization problem based on our proposed stage 1 objective (4), with an additional constraint

$$\mathcal{KL}\big(q(\mathbf{z};\phi) \,\|\, p(\mathbf{z})\big) \le \epsilon. \tag{7}$$

Definition 1. *Suppose that the decoder is Gaussian, $p(\mathbf{x} \mid \mathbf{z};\theta) = \mathcal{N}\big(\mu(\mathbf{z};\theta), \Sigma\big)$ where $\Sigma \succ 0$, and let the corresponding generative model be $\hat{p}(\mathbf{x};\theta,\phi) \coloneqq \int p(\mathbf{x} \mid \mathbf{z};\theta) q(\mathbf{z};\phi) d\mathbf{z}$. Let a measurable set $\mathcal{Z}_0 \subset \mathbb{R}^d$ denote the effective latent support of $q(\mathbf{z};\phi)$, defined as the subset of the latent space with high probability mass. For any $\alpha \in (0,1)$, we define the reachable region induced by the generative model as*

$$\mathcal{T}_\alpha(\mathcal{Z}_0) \coloneqq \bigcup_{\mathbf{z} \in \mathcal{Z}_0} \tau_\alpha(\mathbf{z}),$$

where $c_\alpha \coloneqq F^{-1}_{\chi^2_p}(1-\alpha)$ and the $(1-\alpha)$ high-density region of each decoder is

$$\tau_\alpha(\mathbf{z}) \coloneqq \big\{\mathbf{x} \in \mathbb{R}^p : (\mathbf{x} - \mu(\mathbf{z};\theta))^\top \Sigma^{-1} (\mathbf{x} - \mu(\mathbf{z};\theta)) \le c_\alpha\big\}.$$

Proposition 1. *Suppose that $q(\mathbf{z};\phi)$ is a distribution absolutely continuous with respect to the prior $p(\mathbf{z})$ and assume $\mathcal{KL}\big(q(\mathbf{z};\phi) \,\|\, p(\mathbf{z})\big) \le \epsilon$. Then, for any $t > 0$, it holds that*

$$\mathbb{Q}_\phi\big(\mathbf{z} \in \mathcal{R}_t\big) = \int_{\mathcal{R}_t} q\big(\mathbf{z};\phi\big)\, d\mathbf{z} > 1 - \frac{\epsilon}{t},$$

where the t-region of q relative to p is defined as

$$\mathcal{R}_t \coloneqq \big\{\mathbf{z} \in \mathbb{R}^d : \log\big(q(\mathbf{z};\phi)/p(\mathbf{z})\big) \le t\big\}.$$

Proposition 2. *Under Definition 1 and the conditions of Proposition 1, it holds that*

$$\mathbb{P}_{\theta,\phi}\big(\mathbf{x} \in \mathcal{T}_\alpha(\mathcal{R}_t)\big) = \int_{\mathcal{T}_\alpha(\mathcal{R}_t)} \hat{p}(\mathbf{x};\theta,\phi) d\mathbf{x} > 1 - \frac{\epsilon}{t} - \alpha.$$

Proposition 1 implies that, as $\epsilon \to 0$, the aggregated posterior density concentrates on a subset of the support where the log-likelihood ratio with respect to a fixed prior (e.g., the standard Gaussian) is upper bounded, thereby restricting the effective latent support to $\mathcal{R}_t$. Consequently, by Proposition 2, the generated samples fall into the reachable region with very high probability, whose size is determined by the effective latent support and is therefore constrained by the additional regularization (7). The proofs of Propositions 1 and 2 are provided in Appendix A.1 and A.2, respectively.

Significance. In existing two-stage modeling approaches, the effective support of the latent distribution is constrained by the additional regularization (7). As a result, the image of the decoder outputs in the data space is constrained, which consequently reduces the diversity of the generated synthetic samples. However, we derive a new objective without these constraints on the support by using only a single prior, which leads to significantly higher diversity in synthetic data (see Section 4.3).

3.3 Implementation

Closed-form regularization. We derive an upper bound for the posterior regularization term (the last term in (4)), which can be expressed in closed form under a diagonal Gaussian posterior, $q(\mathbf{z} \mid \mathbf{x};\phi) \coloneqq \mathcal{N}\big(\mathbf{z} \mid \mu(\mathbf{x};\phi), \mathrm{diag}(\sigma^2(\mathbf{x};\phi))\big)$. Here, $\mathrm{diag}(a)$ for $a \in \mathbb{R}^d$ denotes a diagonal matrix with diagonal entries given by a. $\mu(\cdot;\phi) : \mathbb{R}^p \mapsto \mathbb{R}^d$ and $\sigma^2(\cdot;\phi) : \mathbb{R}^p \mapsto \mathbb{R}^d_{++}$ indicate the mean and variance vectors of the Gaussian encoder, respectively. Under this assumption, the posterior regularization term is bounded as

$$\mathbb{E}_{\mathbf{x}}\big[\mathcal{KL}\big(q(\mathbf{z} \mid \mathbf{x};\phi) \| q(\mathbf{z};\phi)\big)\big] \tag{8}$$

$$\le \mathbb{E}_{p^*(\mathbf{x})q(\mathbf{z};\phi)}\left[-\frac{1}{2}\sum_{j=1}^{d}\left(1 - \frac{(\mathbf{z}_j - \mu_j(\mathbf{x};\phi))^2}{\sigma_j^2(\mathbf{x};\phi)}\right)\right]$$

(see Appendix A.5 for details).

This regularization prevents latent variables from capturing information specific to individual observations. Intuitively, the upper bound in (8) can be viewed as the negative coefficient of determination (R^2), which quantifies how much of the latent variability is explained by a single data point. Minimizing this term, therefore, encourages the latent representation to remain uninformative at the instance level.

Incorporating discrete columns. To handle heterogeneous tabular data, we define the decoder as a Gaussian distribution for continuous columns and a categorical distribution for discrete columns, both conditioned on the latent variable. The final objective function for stage 1, combining this decoder with our closed-form regularization of the posterior distribution in (8), is provided in the Appendix.

Stage 2. As in TabSyn [Zhang *et al.*, 2024], we adopt score-based diffusion models [Song *et al.*, 2021; Karras *et al.*, 2022] to minimize the KL-divergence term in (5), with ϕ fixed as the parameter fitted during stage 1. Due to space constraints, we refer the reader to the Appendix for the detailed optimization and training procedure of stage 2.

Controllable privacy. Since stage 2 of our framework generates latent representations from noise using a diffusion-based generative model, privacy can be enhanced at inference time by adjusting the noise level during latent sampling (i.e., by controlling the noise scaling factor c), without requiring any model re-training. Specifically, sampling the latent variable $\mathbf{z}$ at the initial time t_N from $\mathcal{N}(\mathbf{0}, c^2 \cdot \sigma^2(t_N)\mathbf{I})$, where $c = 1$ serves as the baseline. This mechanism enables a direct and interpretable trade-off between privacy and sample quality. Moreover, perturbing latent variables is more effective than injecting noise into the generated data, as it jointly affects both continuous and categorical features through the compact latent representation of the model.

4 Experiments

We empirically evaluate SPT across a range of benchmark datasets. Our experiments[1] are designed to address the following key research questions[2]:

RQ1. *Does SPT achieve state-of-the-art performance in synthetic tabular data generation?*

RQ2. *To what extent does a less-constrained prior contribute to the synthetic data generation performance of SPT?*

RQ3-1. *Ablation study: How does the encoder variance affect the performance of SPT in synthetic data generation?*

RQ3-2. *Ablation study: How does the adaptive regularization influence the synthetic data quality generated by SPT?*

Datasets. To evaluate the performance of our method, we conduct experiments on 12 real-world tabular datasets from Kaggle and the UCI Machine Learning Repository[3]. These datasets include variation in the number of continuous and categorical variables and in the total sample sizes. A detailed description of each dataset is provided in the Appendix.

Baseline models. We compare our proposed method with the following 6 tabular data synthesizers. Details of the hyperparameter configuration for each baseline model are provided in the Appendix.

- *GAN-based*: CTGAN [Xu *et al.*, 2019]
- *VAE-based*: TVAE [Xu *et al.*, 2019]
- *Transformer-based*: TabMT [Gulati and Roysdon, 2023]
- *Diffusion-based*: TabDDPM [Kotelnikov *et al.*, 2023], STaSy [Kim *et al.*, 2023], TabSyn [Zhang *et al.*, 2024]

4.1 Evaluation Metrics

To comprehensively evaluate the quality of the generated synthetic data, we consider three dimensions: (1) statistical fidelity, (2) statistical diversity, and (3) synthetic data utility.

[1]All experiments are conducted on an NVIDIA RTX A6000 GPU using implementations in PyTorch.

[2]We conduct an additional experiment to evaluate the privacy preservability of SPT in Subsection 4.5.

[3]https://archive.ics.uci.edu/, https://www.kaggle.com/datasets/

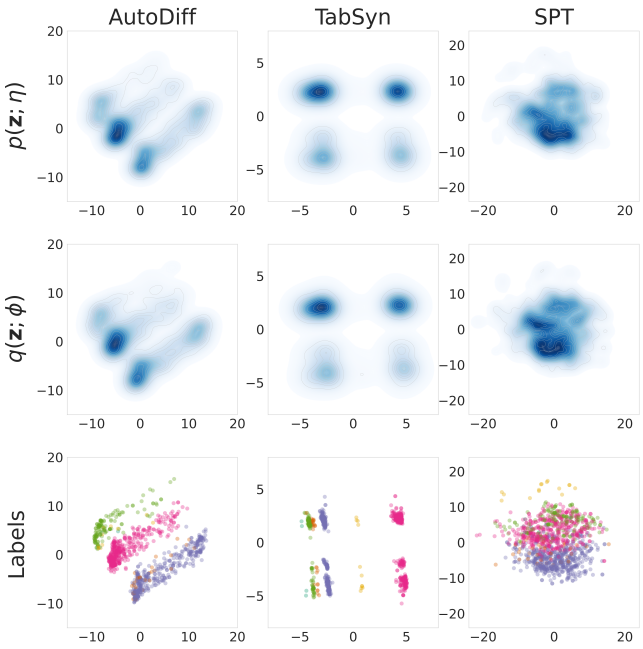

Figure 2: **RQ2: Qualitative comparison**. Latent space visualization using principal component analysis in the redwine dataset. Top row: trainable prior $p(\mathbf{z};\eta)$, Middle row: aggregated posterior $q(\mathbf{z};\phi)$, Bottom row: aggregated posterior $q(\mathbf{z};\phi)$ colored by latent class labels.

Statistical fidelity. Statistical fidelity quantifies the extent to which the synthetic data replicate the distributional properties of the original dataset. We assess this aspect using four metrics: Goodness-of-Fit (`GoF`), Maximum Mean Discrepancy (`MMD`), Pairwise Correlation Difference (`PCD`), and α`-precision` [Alaa *et al.*, 2022].

Statistical diversity. Statistical diversity evaluates how well the synthetic data captures the variability and support of the real data distribution. We assess this aspect using three metrics: β`-recall` [Alaa *et al.*, 2022], `Coverage` [Naeem *et al.*, 2020], and Clipped Coverage (`CC`) [Salvy *et al.*, 2025].

Synthetic data utility. Synthetic data utility evaluates how effectively the synthetic data can be used for the downstream task. Following [Hansen *et al.*, 2023], we adopt the three metrics: classification accuracy (`Acc`), model selection consistency (`Model`), and feature selection consistency (`Feature`).

Overall performance. To ensure fair evaluation across diverse metrics and to identify the best hyperparameter configurations, we adopt the TOPSIS method [Tzeng and Huang, 2011], a widely used multi-criteria decision-making approach across various domains [Behzadian *et al.*, 2012]. Detailed descriptions of the evaluation procedure are in the Appendix.

4.2 RQ1: Synthetic Data Generation.

Overall results. As shown in Table 1, our proposed model SPT outperforms all baseline methods by achieving the best performance on 8 out of 10 evaluation metrics. Across statistical fidelity, statistical diversity, and synthetic data util-

Model	*Statistical fidelity*				*Statistical diversity*			*Synthetic data utility*			TOPSIS↑
	GoF↓	MMD↓	PCD↓	α-prec.↑	β-rec.↑	Cov.↑	CC↑	Acc↑	Model↑	Feat.↑	
CTGAN	$.537_{\pm.120}$	$.118_{\pm.020}$	$2.864_{\pm.958}$	$.647_{\pm.051}$	$.063_{\pm.029}$	$.375_{\pm.080}$	$.208_{\pm.057}$	$.564_{\pm.086}$	$.411_{\pm.080}$	$.471_{\pm.077}$	0.441
TVAE	$.102_{\pm.017}$	$.049_{\pm.010}$	$1.293_{\pm.319}$	$.807_{\pm.035}$	$.152_{\pm.031}$	$.583_{\pm.071}$	$.395_{\pm.077}$	$.694_{\pm.069}$	$.523_{\pm.092}$	$.585_{\pm.053}$	0.734
TabMT	$.079_{\pm.019}$	$.011_{\pm.002}$	$2.092_{\pm.260}$	$\mathbf{.963}_{\pm.003}$	$.153_{\pm.034}$	$.427_{\pm.065}$	$.263_{\pm.054}$	$.673_{\pm.069}$	$.595_{\pm.080}$	$.507_{\pm.042}$	0.723
STaSy	$2.650_{\pm2.538}$	$.047_{\pm.008}$	$.782_{\pm.190}$	$.882_{\pm.030}$	$.290_{\pm.034}$	$.664_{\pm.079}$	$.466_{\pm.071}$	$.687_{\pm.082}$	$.683_{\pm.055}$	$.743_{\pm.042}$	0.798
TabDDPM	$.037_{\pm.009}$	$.015_{\pm.004}$	$1.617_{\pm.683}$	$.952_{\pm.008}$	$\underline{.485}_{\pm.036}$	$\underline{.939}_{\pm.010}$	$\underline{.877}_{\pm.037}$	$\underline{.738}_{\pm.069}$	$\mathbf{.869}_{\pm.022}$	$.823_{\pm.039}$	0.915
TabSyn	$\underline{.025}_{\pm.005}$	$\underline{.010}_{\pm.002}$	$\underline{.469}_{\pm.138}$	$\mathbf{.963}_{\pm.009}$	$.428_{\pm.027}$	$.910_{\pm.023}$	$.843_{\pm.050}$	$.726_{\pm.069}$	$.818_{\pm.040}$	$\underline{.826}_{\pm.032}$	$\underline{0.934}$
SPT	$\mathbf{.020}_{\pm.004}$	$\mathbf{.008}_{\pm.001}$	$\mathbf{.441}_{\pm.133}$	$\underline{.962}_{\pm.006}$	$\mathbf{.526}_{\pm.033}$	$\mathbf{.953}_{\pm.009}$	$\mathbf{.923}_{\pm.037}$	$\mathbf{.741}_{\pm.068}$	$\underline{.867}_{\pm.035}$	$\mathbf{.830}_{\pm.033}$	**0.969**
Improve(%)	**+18.9%**	**+18.7%**	**+5.9%**	−0.1%	**+8.6%**	**+1.5%**	**+5.3%**	**+0.3%**	−0.2%	**+0.5%**	**+3.7%**

Table 1: **RQ1: Synthetic data generation performance.** The means and their standard errors across 12 real-world datasets are reported over 5 random seeds. ↑ (↓) denotes that higher (lower) is better. The best value is **bolded**, and the second best is underlined. 'Improve(%)' refers to the relative change ratio compared to the second best, and the **blue** highlights the positive improvement.

ity, SPT consistently demonstrates strong and well-balanced performance, whereas baseline models tend to favor specific aspects of the trade-off. Based on the TOPSIS results reported in Appendix D.1, which summarizes overall performance across all metrics, SPT achieves the highest TOPSIS score on 9 out of 12 datasets and ranks second on the remaining 3 datasets, indicating consistent performance across diverse real-world tabular datasets. According to the detailed experimental results presented in Appendix D, SPT achieves the best metric values in 53 out of 120 settings (12 datasets × 10 metrics), compared to 32 for TabDDPM and 30 for TabSyn.

Statistical fidelity. As shown in Table 1, SPT attains the best values for GoF (0.020), MMD (0.008), and PCD (0.441) outperforming the second best baseline, TabSyn, with relative improvements of 18.9%, 18.7%, and 5.9%, respectively. Although SPT ranks second in α-precision (0.962), its performance remains comparable to the best-performing baselines. Overall, these results demonstrate that SPT preserves the statistical properties of real data more faithfully than other synthesizers.

Statistical diversity. As shown in Table 1, SPT achieves the best performance on all diversity-related metrics, including β-recall, Coverage, and CC. Compared to second best baseline, SPT improves β-recall by 8.6%, indicating substantially enhanced coverage of rare or underrepresented modes in the data. Moreover, SPT achieves the highest Coverage (0.953) and CC (0.923), reflecting its ability to generate diverse samples while maintaining coherent joint distributions. These results suggest that the proposed approach effectively addresses the fidelity-diversity trade-off by allowing the model to have a less-constrained prior and a less-constrained effective latent support.

Synthetic data utility. In terms of synthetic data utility, SPT consistently matches or outperforms existing baselines in Table 1. SPT achieves the highest downstream classification accuracy (Acc: 0.741), improving upon the second best method by 0.3%. For feature-based and model-based utility metrics, SPT ranks first in Feature and second in Model. SPT maintains high fidelity and utility while achieving substantially improved diversity, highlighting its ability to generate synthetic data that is both realistic and practically useful for downstream tasks.

	AutoDiff	TabSyn (%)	SPT (%)
$p(\mathbf{z})$	✗	*w/*	*w/o*
GoF↓	.033	.025 (+24.2%)	**.020 (+39.4%)**
MMD↓	.011	.010 (+9.1%)	**.008 (+27.3%)**
PCD↓	.523	.469 (+1.3%)	**.441 (+15.7%)**
α-prec.↑	.955	**.963 (+0.8%)**	.962 (+0.7%)
β-rec.↑	.415	.428 (+3.1%)	**.526 (+26.7%)**
Cov.↑	.894	.910 (+1.8%)	**.953 (+6.6%)**
CC↑	.806	.843 (+4.6%)	**.923 (+14.5%)**
Acc↑	.721	.726 (+0.7%)	**.740 (+2.6%)**
Model↑	.767	.818 (+6.6%)	**.867 (+13.0%)**
Feat.↑	.796	.826 (+3.8%)	**.830 (+4.3%)**
TOPSIS↑	.922	.934 (+1.3%)	**.969 (+5.1%)**

Table 2: **RQ2: Effect of less-constrained prior.** Means across 12 real-world datasets are reported over 5 random seeds. ↑ (↓) indicates higher (lower) is better. The best value is **bolded**. Values in parentheses denote the relative improvement over AutoDiff. **Blue** (or **red**) highlights positive improvements.

4.3 RQ2: Effect of Less-Constrained Prior

To investigate the effect of our regularization term, we compare three models with different forms of latent regularization: (1) AutoDiff [Suh *et al.*, 2023], which applies no regularization to the latent space; (2) TabSyn, which imposes a KL-divergence between the aggregated posterior and a fixed Gaussian prior; and (3) SPT, which regularizes the posterior distribution using a self-regularization scheme. Note that the same latent dimensionality ($d = 4$) is used for AutoDiff, TabSyn, and SPT.

Latent space analysis. Figure 2 visualizes the resulting latent distributions $p(\mathbf{z}; \eta)$ and learned posteriors $q(\mathbf{z}; \phi)$, and together with the corresponding class labels on the redwine dataset. The middle row in Figure 2 shows that the additional latent regularization (6) constrains the effective support of the learned posterior, creating large near-zero-density regions (i.e., holes) in latent space, consistent with our theoretical coverage bound in Section 3.2.

AutoDiff, which applies no explicit regularization, yields

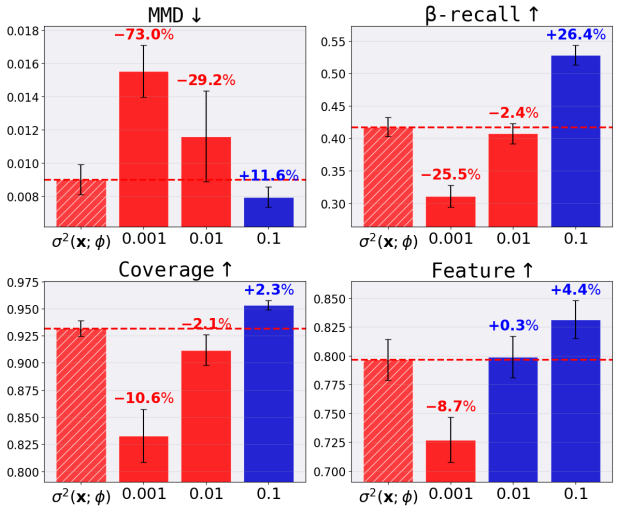

Figure 3: **RQ3-1: Ablation study on encoder variance**. $\sigma^2(\mathbf{x};\phi)$ denotes a trainable variance, and $\sigma^2 \in \{0.001, 0.01, 0.1\}$ is a fixed variance. The means and their standard errors across 12 real-world datasets are reported over 5 random seeds. ↑ (↓) denotes that higher (lower) is better. Blue text indicates improvement over the trainable variance setting, while red text indicates degradation.

a latent space in which latent variables corresponding to the same class label form clearly separated clusters. TabSyn, due to the additional regularization in (6), produces much more compact class-wise clusters than AutoDiff, resulting in a strongly constrained latent support where large regions of the latent space carry almost no probability mass. In contrast, because the additional latent-support constraint induced by a fixed prior is eliminated, SPT yields a latent space in which class-wise clusters are connected without near-zero-density regions. Consistently, PCA scores in Figure 2 indicate that TabSyn exhibits substantially smaller values than SPT due to KL-divergence regularization toward a fixed standard Gaussian prior. Therefore, our proposed method enables smoother transitions within the effective latent support and thereby facilitates the generation of more diverse samples.

Comparison of regularization methods. We investigate how different regularization strategies influence synthetic data generation performance, as summarized in Table 2. Compared to AutoDiff (no latent regularization), TabSyn (KL regularization to a fixed Gaussian prior) improves performance across all evaluation dimensions, indicating that introducing a structured latent constraint can enhance synthetic data generation quality. However, additionally replacing the fixed-prior regularization with our self-regularized posterior (SPT), we observe substantially larger gains on nearly all metrics. SPT achieves the best values on 9 out of 10 metrics and yields the overall improvement under TOPSIS ($0.922 \rightarrow 0.969$, $+5.1\%$). Specifically, SPT yields substantial gains in diversity-related metrics (β-recall, Coverage, and CC), indicating improved coverage of minority modes, with only a slight drop in α-precision, reflecting a mild trade-off toward broader coverage.

β-annealing	*w/o*	*$w/$*	Δ
GoF↓	$.070_{\pm.008}$	$.020_{\pm.004}$	**+71.4%**
MMD↓	$.016_{\pm.002}$	$.008_{\pm.001}$	**+50.0%**
PCD↓	$1.179_{\pm.159}$	$.446_{\pm.145}$	**+62.2%**
α-prec.↑	$.851_{\pm.017}$	$.964_{\pm.006}$	**+13.3%**
β-rec.↑	$.343_{\pm.019}$	$.528_{\pm.036}$	**+53.9%**
Cov.↑	$.765_{\pm.021}$	$.951_{\pm.009}$	**+24.3%**
CC↑	$.639_{\pm.029}$	$.916_{\pm.039}$	**+43.3%**
Acc↑	$.709_{\pm.031}$	$.755_{\pm.073}$	**+6.5%**
Model↑	$.611_{\pm.041}$	$.868_{\pm.039}$	**+42.1%**
Feat.↑	$.754_{\pm.019}$	$.833_{\pm.035}$	**+10.5%**

Table 3: **RQ3-2: Ablation study on adaptive regularization.** 'w/o' denotes SPT without β-annealing, and '$w/$' denotes SPT with β-annealing. The means and their standard errors across 12 real-world datasets are reported over 5 random seeds. ↑ (↓) indicates that higher (lower) is better. Δ denotes the relative change ratio compared to 'w/o'. **Blue** indicates positive improvement.

These results support our argument that a fixed prior is not essential for effective two-stage generative modeling and that eliminating the additional constraint on latent support leads to improved synthetic data quality.

4.4 RQ3: Ablation Studies

Encoder variance. During stage 1, we fix the encoder variance to a constant for training stability, as the regularization term in (8) allows variance inflation to reduce the objective, potentially leading to unstable dynamics; empirically, treating the encoder variance as a non-trainable hyperparameter also consistently improves performance.

In Figure 3, we present empirical results comparing the performance of SPT with a trainable encoder variance to that of models using fixed variance values. When the encoder variance is fixed at $\sigma^2 = 0.1$, SPT achieves an 11.6% improvement in MMD, indicating enhanced statistical fidelity. In terms of diversity, β-recall and Coverage improve by 26.4% and 2.3% under the same fixed variance setting, respectively. Synthetic data utility metrics also benefit from fixing the variance at $\sigma^2 = 0.1$, with Feature increasing by 4.4%. We further observe consistent performance gains for $\sigma^2 = 0.1$ across the remaining evaluation metrics (the complete experimental results are reported in Appendix D.3). Consequently, the TOPSIS-based evaluation consistently selects the fixed-variance setting with $\sigma^2 = 0.1$ across all datasets.

Adaptive regularization scheduling. For KL-divergence regularization in stage 1, we adopt a scheduling strategy for the decoder variance parameter β. The detailed regularization scheduling algorithm and its implementation details are provided in Algorithm 1 in the Appendix.

As shown in Table 3, we observe that β-annealing does not merely improve one aspect (e.g., fidelity) at the expense of another (e.g., diversity); instead, it shifts SPT toward a better operating point where fidelity, diversity, and utility improve simultaneously. The most pronounced effects appear in fidelity metrics (GoF, MMD, PCD) and diversity metrics

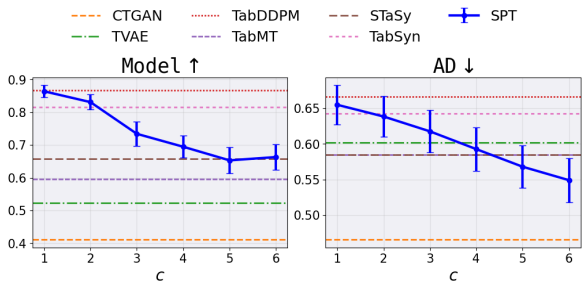

Figure 4: **Trade-off between synthetic data quality and privacy preservability**. c denotes the noise scaling factor in latent diffusion sampling. The means and their standard errors across 12 real-world datasets are reported over 5 random seeds. ↑ (↓) denotes that higher (lower) is better.

(β-recall, CC). These results highlight the importance of gradually relaxing the regularization strength during training, especially in variational models where structured latent representations play a key role. The annealing process effectively balances reconstruction fidelity and latent space regularity, thereby improving the quality of generated samples.

4.5 Privacy Preservability

We evaluate our model's privacy preservability. For privacy evaluation, we measure k-anonymity property (k-anonymity) [Sweeney, 2002], Attribute Disclosure (AD) [Choi *et al.*, 2017], and Distance to Closest Records (DCR) [Park *et al.*, 2018].

Trade-off between privacy and quality. Figure 4 illustrates this trade-off by reporting the synthetic data quality metric Model and privacy preservability metric AD under varying noise strengths. We observe that the Model metric remains competitive as the noise scaling factor c increases from 1 to 2, indicating that moderate noise injection can improve privacy without significantly compromising Model. However, when c exceeds this range, Model gradually deteriorates relative to other baselines, even though AD remains consistently strong. These results demonstrate that our framework enables a controllable trade-off between privacy and synthetic data quality. Descriptions of the evaluation metrics and additional experiments are provided in the Appendix.

5 Related Works

Two-stage generative modeling. In this framework, a first-stage model extracts a compact latent representation and a second-stage model learns its distribution [Li *et al.*, 2015; Engel *et al.*, 2018].[Dai and Wipf, 2019] showed that using a VAE in the first stage and a second VAE to model the latent distribution yields higher-quality samples than a single VAE. [Van Den Oord *et al.*, 2017; Razavi *et al.*, 2019] further demonstrated that discretizing the latent space with vector quantization in the first stage and utilizing an autoregressive generative model in the second stage produces a more meaningful representation and enables high-fidelity image synthesis. Notably, the latent diffusion model of [Rombach *et al.*, 2022] achieves high-resolution image generation more efficiently than pixel-space DDPMs [Ho *et al.*, 2020], while preserving details and enabling flexible conditioning. This two-stage strategy has also shown strong performance in synthetic tabular data generation [Choi *et al.*, 2017; Apellániz *et al.*, 2024; Zhang *et al.*, 2024]. Motivated by these successes, we also propose a two-stage generative framework that extends this line of work.

Over-regularization. [Bowman *et al.*, 2016] showed that using a simple prior can cause posterior collapse, leading the decoder to ignore the latent representation and reducing the diversity of generated samples. [Dilokthanakul *et al.*, 2016] demonstrated that over-regularization of the latent space in unsupervised clustering leads to the collapse of distinct cluster structures, where the model fails to separate data into meaningful groups and instead merges all data points into a single large cluster. [Kim *et al.*, 2024] also showed that an overly simple prior imposes excessive regularization on the latent space, under-representing low-density regions and thereby degrading the quality of synthesized data.

Tabular data synthesis. Before the emergence of diffusion models and Transformers, GANs and VAEs dominated tabular data synthesis [Park *et al.*, 2018; Xu *et al.*, 2019; Zhao *et al.*, 2023; An and Jeon, 2023]. Recently, Transformer-based approaches have gained attention. GReaT [Borisov *et al.*, 2023] reformulates each table row as a text input and models its distribution using GPT-2. Other methods, such as TabMT [Gulati and Roysdon, 2023] and MaCoDE [An *et al.*, 2024], adopt Transformer encoder architectures with masked language modeling. Furthermore, diffusion models have advanced significantly, with improved architectures and training protocols, and have often been highlighted for their ability to capture better multi-modal distributions, which is closely related to sample diversity. TabDDPM [Kotelnikov *et al.*, 2023], STaSy [Kim *et al.*, 2023], and CoDi [Lee *et al.*, 2023] have adopted diffusion-based approaches. AutoDiff [Suh *et al.*, 2023] and TabSyn [Zhang *et al.*, 2024] employ a diffusion model in a latent space.

6 Conclusion and Limitation

We introduce a new two-stage generative modeling objective with a less-constrained effective latent support. This objective is derived directly from a single KL-divergence between the ground-truth distribution and the generative model, and our proposed approach relies on a single prior distribution, unlike existing methods. Through extensive experiments, we empirically demonstrate that relaxing the constraint on the effective latent support leads to significantly improved diversity in synthetic data, while improving or maintaining other quality measures.

Among many hyperparameters, we observed that the synthetic data quality of the model is particularly sensitive to the encoder variance. The dependency on the choice of this hyperparameter can be considered a limitation of our proposed model. Moreover, the diversity of generated samples is crucial not only in the tabular domain but also in other domains such as images [Razavi *et al.*, 2019; Naeem *et al.*, 2020]. An important direction for future work is to investigate whether the proposed less-constrained prior can similarly enhance synthetic sample diversity in other domains.

Acknowledgments

Sanghun Park and Jaesung Lim were supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (No. RS-2022-NR068754). Jong-June Jeon was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT) (No. RS-2022-NR068754, No. RS-2025-23525436). Seunghwan An was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25480090). Moreover, the authors acknowledge the Urban Big data and AI Institute of the University of Seoul supercomputing resources (http://ubai.uos.ac.kr) made available for conducting the research reported in this paper.

Contribution Statement

Sanghun Park and Jaesung Lim are co-first authors.

References

[Alaa *et al.*, 2022] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pages 290–306. PMLR, 2022.

[An and Jeon, 2023] Seunghwan An and Jong-June Jeon. Distributional learning of variational autoencoder: Application to synthetic data generation. *Advances in Neural Information Processing Systems*, 36:57825–57851, 2023.

[An *et al.*, 2024] Seunghwan An, Gyeongdong Woo, Jaesung Lim, ChangHyun Kim, Sungchul Hong, and Jong-June Jeon. Masked language modeling becomes conditional density estimation for tabular data synthesis. In *AAAI Conference on Artificial Intelligence*, 2024.

[Aneja *et al.*, 2021] Jyoti Aneja, Alex Schwing, Jan Kautz, and Arash Vahdat. A contrastive learning approach for training variational autoencoder priors. *Advances in neural information processing systems*, 34:480–493, 2021.

[Apellániz *et al.*, 2024] Patricia A Apellániz, Juan Parras, and Santiago Zazo. An improved tabular data generator with vae-gmm integration. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1886–1890. IEEE, 2024.

[Behzadian *et al.*, 2012] Majid Behzadian, S Khanmohammadi Otaghsara, Morteza Yazdani, and Joshua Ignatius. A state-of the-art survey of topsis applications. *Expert Systems with applications*, 39(17):13051–13069, 2012.

[Borisov *et al.*, 2023] Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2023.

[Bowman *et al.*, 2016] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In Stefan Riezler and Yoav Goldberg, editors, *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics.

[Choi *et al.*, 2017] 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 Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 18–19 Aug 2017.

[Dai and Wipf, 2019] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2019.

[Dilokthanakul *et al.*, 2016] Nat Dilokthanakul, Pedro AM Mediano, Marta Garnelo, Matthew CH Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[Engel *et al.*, 2018] Jesse Engel, Matthew Hoffman, and Adam Roberts. Latent constraints: Learning to generate conditionally from unconditional generative models. In *International Conference on Learning Representations*, 2018.

[Gulati and Roysdon, 2023] Manbir S Gulati and Paul F Roysdon. Tabmt: Generating tabular data with masked transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Hansen *et al.*, 2023] Lasse Hansen, Nabeel Seedat, Mihaela van der Schaar, and Andrija Petrovic. Reimagining synthetic tabular data generation through data-centric AI: A comprehensive benchmark. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

[Hernandez *et al.*, 2022] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jordon *et al.*, 2018] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[Karras *et al.*, 2022] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[Kim *et al.*, 2023] Jayoung Kim, Chaejeong Lee, and Noseong Park. STasy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

[Kim *et al.*, 2024] Juno Kim, Jaehyuk Kwon, Mincheol Cho, Hyunjong Lee, and Joong-Ho Won. t^3-variational autoencoder: Learning heavy-tailed data with student's t and power divergence. In *The Twelfth International Conference on Learning Representations*, 2024.

[Kotelnikov *et al.*, 2023] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.

[Lee *et al.*, 2023] Chaejeong Lee, Jayoung Kim, and Noseong Park. Codi: Co-evolving contrastive diffusion models for mixed-type tabular synthesis. In *International Conference on Machine Learning*, pages 18940–18956. PMLR, 2023.

[Li *et al.*, 2015] Yujia Li, Kevin Swersky, and Richard S. Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, 2015.

[Li *et al.*, 2020] Ruizhe Li, Xiao Li, Guanyi Chen, and Chenghua Lin. Improving variational autoencoder for text modelling with timestep-wise regularisation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2381–2397, 2020.

[Liu *et al.*, 2023] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. GOGGLE: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*, 2023.

[Miletic and Sariyar, 2024] Marko Miletic and Murat Sariyar. Challenges of using synthetic data generation methods for tabular microdata. *Applied Sciences*, 2024.

[Naeem *et al.*, 2020] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pages 7176–7185. PMLR, 2020.

[Park *et al.*, 2018] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.*, 11:1071–1083, 2018.

[Razavi *et al.*, 2019] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

[Rezaabad and Vishwanath, 2020] Ali Lotfi Rezaabad and Sriram Vishwanath. Learning representations by maximizing mutual information in variational autoencoders. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2729–2734. IEEE, 2020.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Salvy *et al.*, 2025] Nicolas Salvy, Hugues Talbot, and Bertrand Thirion. Enhanced generative model evaluation with clipped density and coverage. *arXiv preprint arXiv:2507.01761*, 2025.

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[Suh *et al.*, 2023] Namjoon Suh, Xiaofeng Lin, Din-Yin Hsieh, Mehrdad Honarkhah, and Guang Cheng. Autodiff: combining auto-encoder and diffusion model for tabular data synthesizing. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*, 2023.

[Sweeney, 2002] Latanya Sweeney. k-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl. Based Syst.*, 10:557–570, 2002.

[Tomczak and Welling, 2018] Jakub Tomczak and Max Welling. Vae with a vampprior. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1214–1223. PMLR, 09–11 Apr 2018.

[Tzeng and Huang, 2011] Gwo-Hshiung Tzeng and Jih-Jeng Huang. *Multiple attribute decision making: methods and applications*. CRC press, 2011.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[Xie *et al.*, 2023] Jianwen Xie, Yaxuan Zhu, Yifei Xu, Dingcheng Li, and Ping Li. A tale of two latent flows: learning latent space normalizing flow with short-run langevin flow for approximate inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10499–10509, 2023.

[Xu *et al.*, 2019] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

[Zhang *et al.*, 2024] Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *The Twelfth International Conference on Learning Representations*, 2024.

[Zhao *et al.*, 2019] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *AAAI Conference on Artificial Intelligence*, 2019.

[Zhao *et al.*, 2023] Zilong Zhao, Aditya Kunar, Robert Birke, Hiek Van der Scheer, and Lydia Y Chen. Ctab-gan+: enhancing tabular data synthesis. *Frontiers in Big Data*, 6:1296508, 2023.