TabGeoFlow: A Geometric Flow Matching Model for Tabular Data Synthesis

Jong In Choi
Korea Credit Information Services
jichoi915@yonsei.ac.kr

Abstract

Tabular data synthesis is a key technique for protecting data privacy and addressing class imbalance, yet existing generative models struggle to capture the complex intrinsic structure of the data. To overcome this limitation, we propose TabGeoFlow, a novel geometric flow matching model for tabular data synthesis. The core innovation of TabGeoFlow is the injection of an explicit geometric inductive bias into the conditional flow matching framework. We decompose the learned vector field into local tangent and normal components of the data manifold. By dynamically suppressing the predicted normal component via a controlling loss function, we constrain the generative path to follow the data's intrinsic structure. Implemented with a shared backbone for parameter efficiency, TabGeoFlow achieves competitive or better fidelity and utility, while exhibiting near-random black-box MIA accuracy and DCR ≈ 50%, suggesting reduced memorization without sacrificing quality.

Introduction

Tabular data synthesis is a cornerstone of modern generative modeling, addressing critical challenges in data augmentation, class imbalance mitigation, and privacy preservation. Recent advances, spanning from GAN-based models like CTGAN (Xu et al. 2019) to powerful diffusion-based approaches such as (Kotelnikov, Barsegyan, and Koniaev 2022), TabSyn (Zhang et al. 2023), and (Shi et al. 2025), have demonstrated remarkable success in producing synthetic data that faithfully mirror real-world distributions. These models adeptly capture the intricate correlations across mixed data types—categorical and continuous—thereby substantially improving both the statistical fidelity and the downstream machine learning utility of the generated data.

However, this impressive modeling fidelity introduces a critical vulnerability: an elevated risk to data privacy. The very mechanism that allows a model to accurately approximate a data distribution also enables it to memorize specific training instances, particularly unique or outlier samples. This memorization phenomenon can precipitate significant privacy breaches, where generated samples inadvertently leak sensitive individual information, rendering the model susceptible to privacy attacks like Membership Inference Attacks (MIA) (Shokri et al. 2017). This inherent fidelity-privacy trade-off represents a formidable barrier to the trustworthy and practical deployment of high-quality synthetic tabular data.

To dismantle this barrier, we first hypothesize that privacy leakage is a direct consequence of generative trajectories deviating from the underlying data manifold to overfit specific training instances. Based on this hypothesis, we propose TabGeoFlow, a novel framework that imposes direct geometric constraints on the generative vector field itself to keep the path within the high-density regions of the data distribution. Our approach is founded on the well-established geometric hypothesis that high-dimensional data, including tabular data, are often concentrated near a lower-dimensional manifold embedded within the ambient space (Fefferman, Mitter, and Narayanan 2016). We posit that privacy leakage is a direct consequence of generative trajectories deviating from this underlying data manifold. Specifically, we argue that the learned vector field directs flow into off-manifold regions as it attempts to interpolate or "reach for" memorized training points. In stark contrast to previous regularization techniques that apply indirect constraints—such as on model parameters (e.g., L1/L2 norms) or through auxiliary objectives (e.g., GAN discriminators)—we contend that directly regularizing the vector field, the very space where the generative dynamics unfold, is a more principled and causally direct strategy.

TabGeoFlow operationalizes this geometric intuition by learning to decompose the vector field at any point along a generative path into two orthogonal components: (i) the principal flow, which is aligned with the tangent space of the data manifold and is responsible for capturing the generalized, intrinsic structure of the distribution, and (ii) the normal flow, which captures deviations orthogonal to the manifold and is thus associated with overfitting to specific samples. To mitigate memorization, we introduce Structural Flow Regularization (SFR), a novel loss term designed to

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

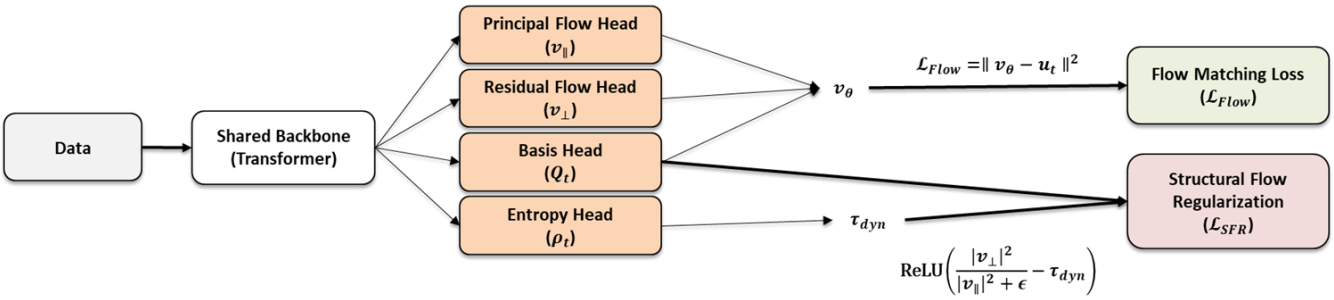

Figure 1: Overview of the TabGeoFlow architecture. The input vector is encoded into a context feature vector through a shared backbone. This vector is transmitted to four parallel heads, each predicting the principal flow($v_\parallel$), the residual flow($v_\perp$), the basis(Q) in the effective tangent space, and the dynamic threshold (τ_{dyn}). The predicted flow components are combined to ensure generation quality through the use of a standard Flow Matching Loss ($\mathcal{L}_{Flow}$) calculation. Simultaneously, these components are used in our core contribution, the calculation of structural flow regularization loss ($\mathcal{L}_{SFR}$), to regulate off-manifold deviation during generation.

dynamically penalize excessive contributions from this normal flow component. This mechanism actively encourages the generative trajectories to remain within the high-density neighborhood of the data manifold, thereby promoting generalization and reducing the model's reliance on memorized instances.

Our comprehensive empirical results validate that TabGeoFlow significantly enhances resistance to a suite of privacy attacks without compromising sample quality or utility. The primary contributions of this work are threefold:

- Geometric vector field regularization: We propose a novel geometric inductive bias that decomposes the generative vector field and penalizes its normal component, thereby reducing memorization risk.
- Principled motivation and analysis: We provide a theoretical rationale for how the SFR loss constrains the extrinsic curvature of generative paths and limits sample-specific memorization.
- Comprehensive empirical validation: Across seven real-world datasets and eight strong baselines, TabGeoFlow achieves competitive or superior performance in terms of statistical fidelity, downstream utility, and privacy metrics such as DCR and MIA.

Related Work

Tabular data synthesis

The landscape of tabular data synthesis has evolved rapidly, with research coalescing around several dominant modeling paradigms. We categorize existing approaches into GAN-based, VAE-based, autoregressive, and the recently prominent diffusion and flow-based models.

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) were an early and influential approach, leveraging adversarial training between a generator and a discriminator to produce realistic data. Within this family, CTGAN (Xu et al. 2019) achieved a significant breakthrough by introducing conditional generation and mode-specific normalization to handle the complexities of mixed-type data.

Variational Autoencoders (VAEs) (Kingma and Welling 2014)offer a more stable alternative by learning an explicit probabilistic model of the data. TVAE (Xu et al. 2019), for instance, successfully integrated the innovations of CTGAN within a VAE framework, using a Gaussian Mixture Model (GMM) to effectively capture the multi-modal distributions of continuous variables. Although VAEs provide stable training and a well-defined likelihood objective, they often face criticism for producing samples that are "blurry" or overly averaged, failing to capture the sharp, distinct features of the true data distribution.

Autoregressive Models approach the problem by sequentially generating each feature conditioned on the previously generated ones, thereby decomposing the joint probability distribution into a product of conditionals. Models like Naru (Zimmermann et al. 2021) exemplify this strategy. The primary advantage of this approach is the ability to directly optimize data likelihood. However, this comes with significant drawbacks: the generation process is slow due to its sequential nature, and the model's performance can be highly dependent on the arbitrary ordering chosen for the features.

Diffusion and Flow-based Models represent the current state-of-the-art and have garnered significant recent attention. Early work like TabDDPM (Kotelnikov, Barsegyan, and Koniaev 2022) successfully adapted Denoising Diffusion Probabilistic Models (DDPMs) (Golovkin 2023) for tabular data, demonstrating strong potential with a simple MLP-based architecture. Subsequent models like TabSyn (Zhang et al. 2023) and Tabdiff (Shi et al. 2025) have refined this approach, introducing dedicated pipelines for data transformation and more sophisticated architectures to handle categorical and continuous variables separately. While these models have demonstrated an exceptional capacity to learn

complex, high-dimensional distributions with high precision, this power is a double-edged sword. Their slow, iterative sampling process (often requiring hundreds or thousands of steps) is a practical bottleneck. More critically, their very ability to capture fine-grained details paradoxically heightens the risk of memorizing training data, exacerbating privacy concerns.

In parallel, Flow-based Models (Dinh, Krueger, and Bengio 2014) have emerged as a compelling alternative, offering both exact likelihood computation and fast, single-pass sampling via invertible transformations. The recent development of Flow Matching (FM) (Lipman et al. 2023), in particular, has significantly advanced the field by enabling stable, simulation-free training of continuous normalizing flows. Our work, TabGeoFlow, is built upon this efficient FM paradigm. However, it is fundamentally distinguished from prior work by moving beyond the simple application of flow matching. We introduce a novel objective that explicitly controls the intrinsic geometry of the learned vector field, a mechanism designed from the ground up to address the privacy-utility trade-off.

Building a high-quality generative model requires navigating a complex multi-objective landscape involving training stability, sampling speed, and data fidelity. A critical, yet often secondary, consideration is the empirical privacy risk stemming from data memorization. While formal frameworks like Differential Privacy (DP) offer robust theoretical guarantees, they often do so at the cost of a substantial degradation in data utility (Abadi et al. 2016).

Our work takes a different path. Rather than pursuing formal DP guarantees, we aim to enhance the intrinsic structure of the generative model itself. By directly regularizing the geometry of the generative process, TabGeoFlow is designed to suppress sample memorization at its source, thereby achieving excellent empirical privacy and a high sampling rate, all while maintaining state-of-the-art data quality.

Methodology

In this section, we first provide a mathematical formulation of Conditional Flow Matching (CFM), the theoretical foundation of our model. We then detail the core components of TabGeoFlow, our proposed method for geometrically regularized tabular data synthesis.

Conditional Flow Matching

Flow-based generative models learn a transformation from a simple prior distribution to a complex data distribution. This transformation is defined by a time-dependent probability path $p_t(x)$ for $t \in [0,1]$ which connects a prior p_0 (e.g., a standard Gaussian $\mathcal{N}(0, I)$) to the target data distribution p_1. The evolution of samples along this path is governed by a time-dependent vector field $u_t(x)$.

Flow Matching (FM) (Lipman et al. 2023) provides a highly efficient, simulation-free method for learning this vector field directly through a regression objective. Conditional Flow Matching (CFM) extends this framework to incorporate conditional information y(e.g., class labels). CFM defines a deterministic path between a point $x_0 \sim p_0$ and data point $x_1 \sim p_1$ via simple linear interpolation.

Given $x_0 \sim p_0$ and a data sample $(x_1, y) \sim p_{data}(x, y)$, the conditional path at time $t \in [0,1]$ is a delta distribution centered at the point x_t as follows:

$$x_t \coloneqq tx_1 + (1-t)x_0 \quad (1)$$

The corresponding target vector field for this path is obtained by differentiating with respect to time.

The Target Vector Field u_t that generates the path is constant along the path as follows:

$$u_t(x_t|x_1, x_0) = \frac{dx_t}{dt} = x_1 - x_0 \quad (2)$$

The objective of CFM is to train the neural network v_θ by minimizing the following L2 regression loss, which measures the expected squared error between the predicted vector field and the target vector field:

$$\mathcal{L}_{\mathcal{CFM}} = E_{t,p(x_1,y),p_{\mathbb{0}}(x_{\mathbb{0}})}[|v_\theta(tx_1 + (1-t)x_0, t, y) - (x_1 - x_0)|^2] \quad (3)$$

Here, the expectation is taken over uniform time steps $t \in U(0,1)$, data points (x_1, y) from the true distribution, and prior samples x_0. This objective frames the learning problem as a direct regression, which can be optimized efficiently without requiring costly numerical simulations. This simulation-free training process is a key advantage, leading to exceptional stability and computational efficiency.

Once the model is trained, we treat v_θ as the dynamics of an ordinary differential equation (ODE). Starting from a random sample $x(0)$ drawn from the prior distribution p_0, we integrate the following ODE from $t = 0$ to $t = 1$: to obtain a new sample

$$\frac{dx}{dt} = v_\theta(x_t, t, y), \quad \text{with initial condition } x(0) = x_0 \quad (4)$$

While CFM is stable and efficient, it imposes no structural constraint on v_θ, leaving the model free to fit idiosyncratic outliers and potenetially memorize individual samples. This motivates TabGeoFLow, which wquips CFM with an explicit geometric regularizer on the learned vector field.

TabGeoFlow

The core innovation of TabGeoFlow is the integration of a Geometric Structure-Aware Module into the CFM framework. This module is tasked with semantically decomposing the learned vector field and regulating its components to align with the intrinsic geometry of the data. Our approach is grounded in the manifold hypothesis (Fefferman, 2016.), which posits that high-dimensional data is concentrated near

Algorithm 1: TabGeoFlow Learning Algorithm

Require: dataset $\mathcal{D}$, model parameter θ, number of training iterations N_{iter}, batch size B, regulatory strength λ_{SFR}
Output: Trained parameters θ

for iter = 1 to N_{iter} do
  Sample minibatch $(x_1, y) \sim \mathcal{D}$ of size B
  Sample $x_0 \sim \mathcal{N}(0, I)$
  Sample $t \sim Uniform(0,1)$
  Compute interpolated input: $x_t \leftarrow x_1 + (1-t)x_0$
  Compute target vector field: $u_t \leftarrow x_1 - x_0$
  Perform Forward pass-through model: $v_\theta, Q_t, \rho_t \leftarrow \text{TabGeoFlow}_\theta(x_t, t, y)$
  Decompose by projection: $v_\perp \leftarrow v_\theta - v_\parallel$
  Compute Flow matching loss: $\mathcal{L}_{Flow} \leftarrow |v_\theta - u_t|^2$
  Compute Dynamic threshold: $\tau_{dyn} \leftarrow \tau_{end} + (\tau_{start} - \tau_{end}) \cdot \rho_t$
  Compute SFR loss: $\mathcal{L}_{SFR} \leftarrow \text{ReLU}\left(\frac{|v_\perp|^2}{|v_\parallel|^2+\epsilon} - \tau_{dyn}\right)$
  Compute total loss: $\mathcal{L}_{Total} \leftarrow E[\mathcal{L}_{Flow} + \lambda_{SFR}\mathcal{L}_{SFR}]$
  Update model parameters: $\theta \leftarrow \theta - \eta\nabla_\theta \mathcal{L}_{Total}$
end for
return θ

a low-dimensional intrinsic manifold $\mathcal{M}$. An ideal generative path should evolve along this manifold, meaning its velocity vector v_θ ought to lie within the manifold's tangent space $T_{x_t\mathcal{M}}$
at every point x_t.

Since the true data manifold $\mathcal{M}$ is unknown and may have a complex structure (e.g., multiple clusters, varying local dimensionality), we do not attempt to recover its exact geometry. Instead, we aim to approximate an operational tangent space that captures the principal directions of data variation at each point x_t. TabGeoFlow's Tangent Basis Head(f_{basis}) takes a feature vector h (derived from the shared backbone for the current state(x_t, t, y)) and outputs an orthonormal basis $Q_t \in R^{D\times k}$ that spans a k-dimensional subspace of the D-dimensional ambient space:

$$Q_t = \text{orthonormal\_basis}\big(f_{basis}(h)\big) \tag{5}$$

Here $k < D$ s an estimate of the manifold's intrinsic dimension. Informed by studies suggesting that the intrinsic dimensionality of real-world tabular data is often low (Facco et al. 2017.), (Levina and Bickel 2004), we set $k = 30$ for all experiments, demonstrating the model's robustness to this choice in our ablation study (Section 5.5).

Crucially, the Tangent Basis Head is trained implicitly, without any direct supervision on the accuracy of the tangent plane. The entire model is trained end-to-end to minimize the total loss $\mathcal{L}_{Total}$. This process incentivizes the model to discover the most efficient geometric decomposition on its own. The model learns to project the majority of the target vector field's energy onto the principal component $v_\parallel$ to minimize the flow matching loss, while simultaneously keeping the residual component $v_\perp$small to avoid the $\mathcal{L}_{SFR}$ penalty. This naturally encourages the Tangent Basis Head to align its basis vectors with the dominant directions of the flow field, which serves as a proxy for the operational tangent plane.

Using the learned basis Q_t, we can define a projection operator $P_{Q_t} = Q_t Q_t^T$ onto the approximated tangent space. We then decompose the model's predicted vector field v_θ into a principal component $v_\parallel$(parallel to the tangent space) and a normal component $v_\perp$(orthogonal to it).

$$v_\parallel = P_{Q_t} v_\theta = Q_t Q_t^T v_\theta \tag{6}$$

$$v_\perp = \big(I - P_{Q_t}\big)v_\theta = v_\theta - v_\parallel \tag{7}$$

By definition, $v_\parallel \in span(Q_t)$, $v_\perp \perp span(Q_t)$, and their dot product is zero. TabGeoFlow uses separate neural network heads to predict components that are then projected to form $v_\parallel$ and $v_\perp$, whose sum constitutes the final vector field v_θ.

Simply decomposing the vector field is insufficient; we must actively discourage the generative path from straying off-manifold by suppressing the normal flow $v_\perp$. To this end, we introduce our core contribution, the Structural Flow Regulation (SFR) loss. TabGeoFlow's full objective function is a weighted sum of the standard flow matching loss and our new regularization term:

$$\mathcal{L}_{Total} = \mathcal{L}_{Flow} + \lambda_{SFR}\mathcal{L}_{SFR} \tag{8}$$

where $\mathcal{L}_{Flow}$ is the CFM loss from Eq. (3), and λ_{SFR} is a hyperparameter controlling the regularization strength.

$$\mathcal{L}_{SFR} = E_{t,p(x_1,y),p_0(x_0)}\left[\text{ReLU}\left(\frac{|v_\perp|^2}{|v_\parallel|^2+\epsilon} - \tau_{dyn}\right)\right] \tag{9}$$

Here, ϵ is a small constant for numerical stability. This loss function acts as an adaptive penalty. The ReLU ensures that a penalty is incurred only when the energy ratio of the normal flow to the principal flow exceeds a dynamic threshold τ_{dyn}.

This threshold, τ_{dyn}, is not fixed but is dynamically adjusted based on the generative process's state. It is controlled by a scalar value $\rho_t = f_{proxy}(h) \in [0,1]$, predicted by an Uncertainty Proxy Head. This head learns to output a high value when the sample x_t is in a low-density or high-uncertainty region (e.g., near the noisy prior at $t \approx 0$) and a low value when it is close to the data manifold (at $t \approx 1$).

$$\tau_{dyn} = \tau_{end} + (\tau_{start} - \tau_{end}) \cdot \rho_t \tag{10}$$

where $\tau_{start} > \tau_{end}$ are hyperparameters. This schedule yields weaker regularization early (high τ_{dyn} near $t \approx 0$ when ρ_t is high) and stronger regularization late (lower τ_{dyn} as ρ_t decreases), consistent with Appendix C.2. This annealing-like schedule improves both learning stability and overall efficiency.

The key motivation for SFR is to prevent sample memorization by suppressing generative paths that deviate from the data manifold. Geometrically, this has the effect of limiting the path's extrinsic curvature, preventing it from "bending" sharply towards specific, isolated training samples. A more detailed geometric argument is provided in Appendix A. The complete learning process is summarized in Algorithm 1.

Experiments

In this section, we present a comprehensive empirical evaluation of TabGeoFlow. Our goal is to rigorously assess its performance against a wide array of state-of-the-art (SOTA) models. The evaluation is structured along three critical axes: Fidelity (statistical similarity to real data), Utility (performance on downstream machine learning tasks), and Privacy (resilience to membership inference attacks)

Dataset

We selected seven widely used public datasets that span a diverse range of sizes, domain types, and task complexities (both classification and regression). This diversity allows us to test the robustness and generalizability of our model. Detailed statistics for each dataset are provided in Table 1.

Dataset	# Train	# Validation	# Test	Task
Adult	28,943	3,618	16,281	Classification
Default	24,000	3,000	3,000	Classification
Shoppers	9,864	1,233	1,233	Classification
Magic	15,215	1,902	1,902	Classification
Beijing	35,058	4,383	4,383	Regression
News	31,714	3,965	3,965	Regression
Diabetes	61,059	20,353	20,354	Classification

Table 1: Overview of datasets used in the evaluation

Experimental Setup

To ensure a fair and comprehensive comparison, our experimental setup, including dataset selection, preprocessing, and evaluation metrics, closely follows the robust protocol established by recent SOTA works, particularly TabSyn (Zhang et al. 2023) and TabDiff (Shi et al. 2025).

We adhere to the standard preprocessing pipeline to ensure a fair comparison with baseline models. Continuous variables are transformed using a QuantileTransformer to approximate a Gaussian distribution, a standard technique for improving model stability. While we acknowledge that such transformations can alter the original data's manifold structure, we operate under the assumption that the core topological and correlational features are sufficiently preserved. Categorical variables are integer-encoded via a LabelEncoder and subsequently one-hot encoded before being fed into the models. This includes variables with very high cardinality, such as those in the Diabetes dataset, which serves as a stress test for handling high-dimensional, sparse inputs. Missing values were imputed using the mean (for continuous) or mode (for categorical) of their respective columns.

We compare TabGeoFlow against a strong and diverse set of eight SOTA models: CTGAN, TVAE, GOGGLE, GReaT, STaSy, CoDi, TabDDPM, and TabDiff. This selection provides a comprehensive benchmark, covering GAN, VAE, Autoregressive, and Diffusion-based paradigms. For all baseline models, we utilized their officially published codebases and adopted the optimal hyperparameter settings recommended in their respective papers. All reported metrics are averaged over three independent runs, with standard deviations provided to ensure statistical robustness. A detailed description of the evaluation metrics for Fidelity, Utility, and Privacy is provided in the Appendix.

TabGeoFlow is implemented using Transformer-based architecture as its shared backbone, with the AdamW optimizer for training. Key hyperparameters, such as learning rate, weight decay, and dropout, were tuned for each dataset using its corresponding validation set. To ensure full reproducibility of our work, we will make our source code publicly available upon publication. All specific hyperparameter configurations used for our model and the baselines are detailed in the Appendix.

We follow the standard preprocessing pipeline used in TabSyn (Zhang et al. 2023). Continuous variables are transformed into Gaussian distributions using QuantileTransformer. Although these transformations are standard techniques to increase the learning stability of the model, we recognize that the manifold structure of the original data can be changed. However, we assume that this transformation preserves key features of the topology and correlation structure of the data. Categorical variables are transformed into integers through the Label Encoder, followed by one-hot encoding upon model input. The same one-hot encoding was applied for variables with very high cardinality, such as the Diabetes dataset, which serves as a stress test to evaluate our model's ability to handle high-dimensional sparse data. All datasets were replaced with mean/mode values for missing values before the experiment.

We compare TabGeoFlow with eight state-of-the-art and SOTA models: CTGAN, TVAE, GOGGLE, GReaT, STaSy, CoDi, TabDDPM, and Tabdiff. This ensures fair comparison, including GAN, VAE, Autoregressive, Diffusion, and other models. All comparison models were run with the optimal hyperparameters proposed in each paper, using officially published codes.

We follow the comprehensive evaluation protocol proposed in TabSyn (Zhang et al. 2023). It consists of Fidelity, which evaluates statistical characteristics of data, Utility, which evaluates downstream work performance, and Privacy, which evaluates privacy leak risk. All experiments were run in triplicate to report mean and standard deviation. Details of the hyperparameter setting are provided in the appendix.

TabGeoFlow uses a Transformer-based architecture and is trained with AdamW optimizer. Major hyperparameters

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Average
CTGAN	20.23(1.20)	26.95(0.93)	13.08(0.16)	7.00(0.19)	22.95(0.08)	5.37(0.05)	18.95(0.34)	16.36
TVAE	14.15(0.88)	19.50(0.95)	18.67(0.38)	5.82(0.49)	18.01(0.08)	6.17(0.09)	32.74(0.26)	16.44
GOGGLE	45.29(0.00)	21.94(0.00)	23.90(0.00)	9.47(0.00)	45.94(0.00)	23.19(0.00)	27.56(0.00)	28.18
GReaT	17.59(0.22)	70.02(0.12)	45.16(0.18)	10.23(0.40)	59.60(0.55)	OOM	OOM	44.24
STaSy	14.51(0.25)	5.96(0.26)	8.49(0.15)	6.61(0.53)	8.00(0.10)	3.07(0.04)	OOM	7.77
CoDi	22.49(0.08)	68.41(0.05)	17.78(0.11)	6.53(0.25)	7.07(0.15)	11.10(0.04)	29.21(0.12)	23.21
TabDDPM	3.01(0.25)	4.89(0.10)	6.61(0.16)	1.70(0.22)	2.71(0.09)	13.16(0.01)	51.54(0.05)	11.95
TabSyn	1.85(0.18)	2.70(0.32)	6.33(0.12)	1.54(0.18)	**2.45(0.09)**	**2.13(0.11)**	3.57(0.05)	2.94
TabDiff	**1.65(0.19)**	3.13(0.53)	2.48(0.10)	**0.91(0.17)**	3.17(0.18)	**2.01(0.23)**	**2.65(0.09)**	2.29
TabGeoFlow	1.84(0.21)	**2.23(0.27)**	**2.10(0.08)**	1.41(0.20)	**2.58(0.15)**	3.05(0.05)	3.15(0.11)	2.34

Table 2: Trend Similarity Error Rate (%). The lower, the better.

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Average
CTGAN	16.84(0.03)	16.83(0.04)	21.15(0.10)	9.81(0.08)	21.39(0.05)	16.09(0.02)	9.82(0.08)	15.99
TVAE	14.22(0.08)	16.83(0.05)	24.51(0.06)	8.25(0.06)	19.16(0.06)	16.62(0.03)	18.86(0.13)	15.97
GOGGLE	16.97(0.00)	17.02(0.00)	22.33(0.00)	1.90(0.00)	16.93(0.00)	25.32(0.00)	24.92(0.00)	17.91
GReaT	12.12(0.04)	19.94(0.06)	14.51(0.12)	16.16(0.09)	8.25(0.12)	OOM	OOM	14.20
STaSy	11.29(0.06)	5.77(0.06)	9.37(0.09)	6.29(0.13)	6.71(0.03)	6.89(0.03)	OOM	7.72
CoDi	21.38(0.06)	15.77(0.07)	31.84(0.05)	11.56(0.26)	16.94(0.02)	32.27(0.04)	21.13(0.25)	21.55
TabDDPM	1.75(0.03)	1.57(0.08)	2.72(0.13)	1.01(0.09)	1.30(0.03)	78.75(0.01)	31.44(0.05)	16.93
TabSyn	**0.69(0.04)**	**0.85(0.02)**	1.33(0.07)	1.00(0.13)	**1.32(0.05)**	**2.13(0.02)**	1.74(0.03)	**1.29**
TabDiff	**0.65(0.03)**	1.13(0.03)	1.48(0.06)	**0.91(0.08)**	**1.17(0.05)**	**2.22(0.03)**	**1.15(0.33)**	**1.24**
TabGeoFlow	0.98(0.02)	**0.93(0.05)**	**1.23(0.05)**	**0.95(0.11)**	1.72(0.06)	2.35(0.01)	**1.25(0.07)**	**1.34**

Table 3: Shape Similarity Error Rate (%). The lower, the better. OOM stands for out of memory.

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Avg. Gap
Real	.927(.000)	.770(.005)	.926(.001)	.946(.001)	.423(.003)	.842(.002)	.704(.002)	0.0
CTGAN	.886(.002)	.696(.005)	.875(.009)	.855(.006)	.902(.019)	.880(.016)	.569(.004)	23.7
TVAE	.878(.004)	.724(.005)	.871(.006)	.887(.003)	.770(.011)	1.01(.016)	.594(.009)	20.2
GOGGLE	.778(.012)	.584(.005)	.658(.052)	.654(.024)	1.09(.025)	.877(.002)	.475(.008)	42.1
GReaT	.913(.003)	.755(.006)	.902(.005)	.888(.008)	.653(.013)	OOM	OOM	13.3
STaSy	.906(.001)	.752(.006)	.914(.005)	.934(.003)	.656(.014)	.871(.002)	OOM	10.9
CoDi	.871(.006)	.525(.006)	.865(.006)	.932(.003)	.818(.021)	1.21(.005)	.505(.004)	30.2
TabDDPM	.907(.001)	.758(.004)	.918(.005)	.935(.002)	.592(.011)	4.86(3.04)	.521(.008)	11.95
TabSyn	.909(.001)	.758(.005)	.918(.003)	**.938(.004)**	.580(.010)	**.862(.020)**	**.684(.007)**	6.82
TabDiff	**.912(.001)**	**.761(.003)**	**.921(.004)**	.936(.003)	**.561(.012)**	**.865(.021)**	**.688(.014)**	6.02
TabGeoFlow	**.913(0.01)**	**.761(.002)**	**.919(.003)**	**.941(.003)**	.585(0.14)	.873(0.33)	.695(0.11)	6.75

Table 4: Machine Learning Efficiency (TSTR Paradigm). Classification is AUC(↑) and regression is RMSE(↓).

such as learning rate, weight attenuation, and dropout rate were optimized using the validation set of each dataset.

Results

Empirical Evaluation

A successful generative model must produce data that is both statistically faithful (high fidelity) and practically useful for downstream tasks (high utility). Our first set of experiments evaluates TabGeoFlow on these two fundamental criteria.

The results presented in Tables 2, 3, and 4 collectively demonstrate that TabGeoFlow achieves a level of performance on par with current state-of-the-art models. In terms of both statistical fidelity (Shape and Trend similarity) and machine learning utility (TSTR), TabGeoFlow's performance is highly competitive with, and in several cases exceeds, that of leading diffusion models like TabDiff.

It indicates that the geometric constraints introduced by our Structural Flow Regulation(SFR) do not hinder the model's ability to learn and reproduce the complex, high-dimensional distributions characteristic of real-world tabular data. The generated data is not only statistically sound but also preserves the intricate predictive relationships necessary for downstream applications. This establishes a strong baseline: TabGeoFlow is a high-quality generative model, capable of competing at the highest level, before even considering its privacy implications.

Having established its high data quality, we now turn to the central hypothesis of our work: that by directly regularizing the geometry of the generative process, we can mitigate privacy leakage without sacrificing performance. We

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Average
STaSy	**50.33(0.19)**	**50.23(0.09)**	51.53(0.16)	**50.47(0.18)**	50.59(0.29)	**50.59(0.14)**	OOM	50.62
CoDi	49.92(0.18)	51.82(0.26)	51.06(0.18)	51.13(0.32)	50.87(0.11)	50.79(0.23)	51.12(0.19)	50.96
TabDDPM	51.14(0.18)	52.15(0.20)	63.23(0.25)	51.90(0.13)	80.11(2.68)	79.31(0.29)	37.76(0.23)	59.37
TabSyn	50.94(0.17)	51.20(0.18)	52.90(0.22)	50.99(0.14)	50.37(0.13)	50.85(0.33)	50.62(0.28)	51.12
TabDiff	52.38(0.26)	53.13(0.23)	51.18(0.19)	50.87(0.25)	52.04(0.32)	53.52(0.36)	51.33(0.16)	52.06
TabGeoFlow	**50.37(0.25)**	**50.19(0.19)**	**49.46(0.21)**	**50.48(0.20)**	**50.49(0.25)**	**50.59(0.26)**	**50.48(0.22)**	**50.45**

Table 5: DCR score (%). The closer you get to 50%, the better (privacy protected).

assess performance by using the Distance to Closest Record (DCR) metric, where a score near the ideal 50% suggests robust protection against sample memorization.

We adopt black-box threat model: the attacker does not access the generator. Instead, it observes logits of a TSTR classifier trained on synthetic data (shadow setting) and predicts membership for queried records. We report attack accuracy (50% = random). TabGeoFlow remains near-random across datasets (see App. C.3.), aligning with our DCR ≈ 50%.

The privacy evaluation results in Table 5 highlight the distinct advantage of our proposed approach. TabGeoFlow consistently yields DCR scores remarkably close to the ideal 50%, achieving an average of 50.45%.

This result should be viewed in contrast to other high-fidelity models. Models, which excel in utility and fidelity, show notable deviations from the 50% mark (52.06% and 51.12%, respectively), suggesting a higher risk of privacy leakage. TabGeoFlow, however, effectively addresses this common trade-off. By encouraging generative paths to remain on the data manifold, our SFR mechanism provides a principled method for reducing the model's tendency to memorize specific training instances.

Ablation Study

To verify that the observed benefits are indeed attributable to our specific design choices, we conducted a systematic ablation study. The results are summarized in Table 6.

Effectiveness of SFR: The most critical experiment is the removal of the Structural Flow Regulation ($\lambda_{SFR} = 0$). The result is clear: a marked increase in privacy risk with only a marginal change in utility. This provides strong evidence that SFR is the key component driving the privacy-preserving properties of TabGeoFlow.

Method	Adult		Default	
	AUC	DCR	AUC	DCR
w/o SFR	.898	51.11	.745	48.46
w/o Dyn	.903	50.99	.760	49.13
K = 10	.905	50.67	.751	50.41
K = 20	.909	50.82	.748	49.84
K = 40	.903	51.28	.755	50.20
TabGeoFlow	**.913**	**50.37**	**.761**	**50.19**

Table 6: Ablation Study on the Adult and Default datasets.

Replacing the dynamic threshold τ_{dyn} with a fixed value led to a general decrease in performance. This suggests that the adaptive nature of our regularization—applying stricter constraints early in the process and relaxing them later—is important for achieving stable and effective training.

The model's performance remains stable across a reasonable range of values for the estimated intrinsic dimension, k. This demonstrates a degree of robustness, suggesting that our method does not require extensive, dataset-specific tuning of this hyperparameter to be effective

Conclusion

In this paper, we introduced TabGeoFlow, a novel framework for tabular data synthesis that addresses the critical privacy-utility trade-off by directly regularizing the geometry of the generative process. Our core contribution is to decompose the learned vector field into a principal component aligned with the data manifold and a normal component associated with sample memorization. By introducing Structural Flow Regulation (SFR), we selectively suppress this normal flow, encouraging the generative paths to adhere to the data's intrinsic structure.

Our comprehensive experiments demonstrate that this geometric approach is highly effective. TabGeoFlow achieves state-of-the-art performance in terms of both statistical fidelity and downstream machine learning utility, while simultaneously providing substantially improved empirical privacy protection compared to other high-performing models. This work suggests that directly controlling the geometry of the vector field is a principled and effective strategy for mitigating the common trade-off between data quality and privacy.

Building on these promising results, our future work will advance in several key directions. We plan to integrate our SFR mechanism with formal privacy frameworks like Differential Privacy (DP) to achieve certifiable guarantees. Concurrently, we will extend the TabGeoFlow architecture to enhance its practical utility, focusing on specialized modules for high-cardinality categorical variables and robust missing data imputation. Finally, we will conduct more extensive security validations against a broader suite of advanced privacy attacks to further establish the robustness of our geometric approach.

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep Learning with Differential Privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318. Vienna, Austria: ACM.

Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; and Kasneci, G. 2023. Language Models Are Realistic Tabular Data Generators. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR).

Chen, R. T. Q.; and Lipman, Y. 2024. Flow Matching on General Geometries. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR).

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. NICE: Non-linear Independent Components Estimation. arXiv preprint arXiv: 1410.8516.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In Proceedings of the Third conference on Theory of Cryptography, 265–284. New York, NY: Springer-Verlag.

Facco, E.; d'Errico, M.; Rodriguez, A.; and Laio, A. 2017. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific reports, 7(1): 12140.

Fefferman, C.; Mitter, S.; and Narayanan, H. 2016. Testing the Manifold Hypothesis. Journal of the American Mathematical Society 29(4): 983–1049.

Ganev, G.; and De Cristofaro, E. 2025. On the Inadequacy of Similarity-based Privacy Metrics. In 2025 IEEE Symposium on Security and Privacy (SP).

Golovkin, D. 2023. Tabular Data Synthesis with Latent Diffusion Models. In Proceedings of the ICML 2023 Workshop on The Symbiosis of Deep Learning and Differential Equations. Honolulu, HI.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, 2672–2680. Montreal, Canada.

Guzmán-Cordero, G.; et al. 2025. Exponential-Family Variational Flow Matching for Tabular Data Generation. arXiv preprint arXiv:2501.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada.

Jolicoeur-Martineau, A.; et al. 2024. Generating and Imputing Tabular Data via Diffusion and Flow-based Gradient-Boosted Trees. arXiv preprint arXiv:2402.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Proceedings of the Second International Conference on Learning Representations. Banff, Canada.

Kotelnikov, A.; Barsegyan, D.; and Koniaev, A. 2022. TabDDPM: Modelling Tabular Data with Diffusion Models. arXiv preprint arXiv:2209.15421.

Levina, E.; and Bickel, P. 2004. Maximum Likelihood Estimation of Intrinsic Dimension. In Advances in Neural Information Processing Systems (NeurIPS), 17.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; and Behrmann, J. 2023. Flow Matching for Generative Modeling. In Proceedings of the Eleventh International Conference on Learning Representations. Kigali, Rwanda.

Park, N.; Mohammadi, M.; Gorde, K.; Jajodia, S.; Park, H.; and Kim, Y. 2018. Data Synthesis based on Generative Adversarial Networks. Proceedings of the VLDB Endowment 11(10): 1071–1083.

Scassola, E.; et al. 2025. Graph Conditional Flow Matching for Relational Data Generation. arXiv preprint arXiv:2502.

Shi, J.; Xu, M.; Hua, H.; Zhang, H.; Ermon, S.; and Leskovec, J. 2025. TabDiff: A Mixed-Type Diffusion Model for Tabular Data Generation. arXiv preprint arXiv:2410.20626.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership Inference Attacks Against Machine Learning Models. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), 3–18. San Jose, CA: IEEE.

Tang, S.; Wu, Z. S.; Aydore, S.; Kearns, M.; and Roth, A. 2024. Membership Inference Attacks on Diffusion Models via Quantile Regression. In Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR.

Xie, T.; Zhu, Y.; Yu, L.; Yang, T.; Cheng, Z.; Zhang, S.; Zhang, X.; and Zhang, C. 2024. Reflected Flow Matching. In Proceedings of the 41st International Conference on Machine Learning (ICML), PMLR.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling Tabular Data Using Conditional GAN. In Advances in Neural Information Processing Systems (NeurIPS), 32.

Zhang, H.; Zhang, J.; Shen, Z.; Srinivasan, B.; Qin, X.; Faloutsos, C.; Rangwala, H.; and Karypis, G. 2024. Mixed-Type Tabular Data Synthesis with Score-Based Diffusion in Latent Space. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR).

Zimmermann, T.; Troullinou, E.; Zeginis, D.; and Kondylakis, H. 2021. Naru: A new deep learning model for tabular data. In Proceedings of the 24th International Conference on Extending Database Technology (EDBT). Nicosia, Cyprus.