TABTREEVFM: ENDPOINT VARIATIONAL FLOW MATCHING WITH BOOSTED TREES FOR MIXED-TYPE TABULAR GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic tabular data supports downstream analytics, machine learning, and data sharing when real data is scarce, sensitive, or costly to collect. Mixed-type synthesis must capture joint dependencies while respecting heterogeneous column structure, yet high-fidelity generators can reproduce training records rather than learn the underlying distribution. Gradient-boosted trees suit such heterogeneous data, but existing tree-based flow models regress ambient-space velocities, imposing a Euclidean geometry on inherently multinomial categorical variables. To address this, we propose TabTreeVFM, a tree-boosted endpoint-posterior model with Gaussian regression heads for continuous columns and multinomial classifier heads for categorical variables, in place of direct velocity regression. The resulting posterior means recover the flow-matching velocity analytically, and training decomposes into independent XGBoost subproblems across timesteps and feature groups. Across 10 standard mixed-type benchmarks, TabTreeVFM attains the best overall average rank among 12 generators across fidelity, utility, and empirical privacy. Moreover, we introduce *Mem*, a held-out-calibrated measure of excess full-row reproduction, and find that several strong neural baselines reproduce training rows substantially above the natural collision rate, while TabTreeVFM remains close to the held-out reference. Together, these results show that type-aware endpoint prediction combines the tabular inductive bias of boosted trees with principled flow-matching geometry, yielding strong fidelity and utility alongside a favorable empirical privacy and memorization profile.

1 INTRODUCTION

Tabular data are central to high-stakes domains such as finance (Yeh & Lien, 2009), health (Cormode et al., 2025), e-commerce (Sakar & Kastro, 2018), and public administration (Park et al., 2018). Synthetic tabular data can support benchmarking, model development, data sharing, and privacy-risk assessment when direct release of the original table is undesirable (Sanghi & Haritsa, 2023; Ge et al., 2024; Zhang et al., 2014; Cormode et al., 2025). However, tabular synthesis remains difficult: a single table often mixes continuous, skewed numerical, binary, and categorical variables with strong cross-type dependencies (Fan et al., 2020; Park et al., 2018).

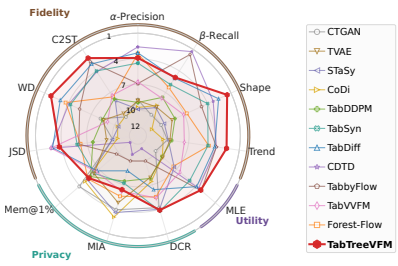


Figure 1: Benchmark profiles of 12 generators. Each spoke shows the average within-dataset rank (lower is better).

Following GAN- and VAE-based generators such as CTGAN and TVAE (Xu et al., 2019), diffusion models including TabDDPM (Kotelnikov et al., 2023), TabSyn (Zhang et al., 2024), TabDiff (Shi et al., 2025), and CDTD (Mueller et al., 2025) handle mixed-type data through latent representations or type-specific processes. Within flow matching, Variational Flow Matching (VFM) learns posteriors over clean-data endpoints (Eijkelboom et al., 2024), while Exponential-family VFM (EF-VFM) assigns Gaussian and multinomial factors to numerical and categorical features, yielding squared-error and cross-entropy objectives, respectively (Guzmán-Cordero et al., 2025). In parallel, ForestDiffusion

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and Forest-Flow use boosted trees, strong predictors for heterogeneous tables (Grinsztajn et al., 2022), to estimate scores or velocities (Jolicoeur-Martineau et al., 2024; Cresswell & Kim, 2024). However, direct tree-flow regression imposes Euclidean geometry on one-hot blocks representing multinomial variables. Type-aware endpoint likelihoods also raise an optimization question: their Fisher geometry is known (Guzmán-Cordero et al., 2025), and NGBoost exploits it through explicit per-sample Fisher inversion (Duan et al., 2020), but how the leaf updates of second-order boosting relate to this geometry has not been characterized. Beyond these modeling limitations, high fidelity can coexist with memorization (Fang et al., 2025; 2026), while standard DCR and MIA indicators can understate re-identification risk and do not directly measure full-row reproduction (Ganev & De Cristofaro, 2025; Yao et al., 2026), leaving fidelity gains insufficient to establish generalization.

We propose TabTreeVFM, a tree-boosted endpoint-posterior model for mixed-type tabular synthesis. Instead of regressing ambient-coordinate velocities, TabTreeVFM estimates type-specific endpoint posteriors using Gaussian regression heads for continuous columns and one multinomial classifier per categorical variable. Their posterior means recover the VFM velocity analytically, decomposing training into independent XGBoost subproblems across timesteps and feature groups. We further prove that the multiclass leaf update of second-order boosting is a diagonal-Fisher-scaled step on the categorical endpoint likelihood, up to an explicit within-leaf dispersion factor. To assess memorization, we introduce *Mem*: existing generators can reproduce training rows despite passing standard privacy tests, and Figure 1 shows that some of the highest-fidelity methods rank near the top on fidelity metrics yet poorly on privacy-related ones. *Mem* measures full-row reproduction relative to a disjoint held-out reference and thereby calibrates for natural collisions. Overall, our contributions are:

- We formulate TabTreeVFM, a boosted-tree EF-VFM model with type-specific endpoint posteriors: Gaussian regression for continuous columns and multinomial classification for categorical variables. Controlled ablations attribute the largest improvements in dependency modeling, coverage, and downstream utility to the multinomial categorical heads, with the endpoint target improving marginal fidelity.
- We prove that the multiclass leaf update of second-order boosting is a diagonal-Fisher-scaled step on categorical endpoint likelihoods, with within-leaf prediction dispersion giving its exact deviation from the representative Fisher step.
- We introduce Mem, a held-out-calibrated measure of full-row reproduction above the natural collision rate, revealing that several strong neural generators exhibit substantial excess reproduction reaching up to 8.37 percentage points.
- Across 10 mixed-type benchmarks, TabTreeVFM attains the best overall average rank, outperforms direct tree-flow regression, and combines leading aggregate fidelity with a favorable empirical privacy and memorization profile.

2 Related Work

Neural tabular generators. CTGAN and TVAE (Xu et al., 2019) adapt adversarial and latent-variable models to tables. STaSy (Kim et al., 2023) uses score-based modeling; TabDDPM (Kotelnikov et al., 2023) and CoDi (Lee et al., 2023) combine numerical and categorical diffusion, whereas TabSyn (Zhang et al., 2024) diffuses learned latents. TabDiff (Shi et al., 2025) and CDTD (Mueller et al., 2025) develop feature-adaptive mixed-type processes. TabbyFlow (Guzmán-Cordero et al., 2025) uses EF-VFM to learn type-specific endpoint posteriors, TabVVFM (Nasution et al., 2026) further studies variational flows, and TabCascade (Mueller et al., 2026) generates mixed-type rows through a cascade of flow-matching stages. TabTreeVFM replaces this amortized network with factorized boosted-tree heads, so training decomposes into independent per-timestep, per-feature problems.

Tree-based generators. ForestDiffusion and Forest-Flow (Jolicoeur-Martineau et al., 2024) fit diffusion scores or flow velocities with XGBoost, with subsequent work improving scalability (Cresswell & Kim, 2024). These ambient-space objectives treat one-hot coordinates as regression targets. TabTreeVFM instead learns factorized, type-specific endpoint posteriors and recovers velocities analytically.

Probabilistic boosting. XGBoost (Chen & Guestrin, 2016) uses second-order loss approximations, whereas NGBoost (Duan et al., 2020) fits predictive distributions using natural gradients.

For categorical endpoint likelihoods, we prove that the second-order multiclass leaf update is a diagonal-Fisher-scaled step whose deviation from the representative Fisher step is governed exactly by within-leaf prediction dispersion, so tree-boosted endpoint learning obtains Fisher-scaled updates at the cost of ordinary boosting.

Memorization assessment. DOMIAS (van Breugel et al., 2023) uses density-based membership inference, and held-out DCR assessments vote, once per synthetic record, on whether it lies closer to training or to unseen records (Platzer & Reutterer, 2021); distance-based indicators summarize aggregate proximity and do not identify reproduced rows (Ganev & De Cristofaro, 2025; Yao et al., 2026). Fang et al. (2025; 2026) document training-row reproduction in tabular diffusion through an absolute nearest-neighbour distance-ratio criterion. Mem counts tolerance-bounded full-row replicas and reports their excess over an exchangeable held-out reference, so it identifies individual reproductions and has a common zero reference across datasets.

3 METHOD

3.1 OVERVIEW

Notation. Given a training table $\mathcal{D} = \{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$ with C numerical columns $\boldsymbol{x}^{\text{num}} \in \mathbb{R}^C$ and G categorical variables, each with K_g categories represented by a one-hot block $\boldsymbol{x}_g^{\text{cat}} \in \{0,1\}^{K_g}$, we learn to generate all columns jointly and unconditionally, including any downstream prediction target. For data endpoints $\boldsymbol{x}_1 \sim p_{\text{data}}$ and Gaussian noise $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, we form states $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_1 + \sigma_t \boldsymbol{x}_0$, $t \in [0,1]$. VFM learns an endpoint posterior $q_t^\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_t)$ whose mean determines the model velocity along this affine path. Categorical posterior means lie on the probability simplex Δ_{K_g-1}, while intermediate ODE states are unconstrained in the ambient space.

Figure 2 summarizes the training and sampling pipeline. At each discretized timestep, TabTreeVFM fits independent boosted-tree heads for Gaussian regression on numerical columns and multinomial classification on categorical variables. Posterior means recover the velocity analytically; sampling integrates this field with a Heun solver and decodes the terminal state into a mixed-type table. Appendix A lists additional notation, and Appendix B reviews FM, VFM, and EF-VFM.

3.2 ENDPOINT POSTERIOR

For a mixed-type table, TabTreeVFM uses the endpoint posterior

$$q_t^\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_t) = q_t^{\theta,\text{num}}(\boldsymbol{x}_1 \mid \boldsymbol{x}_t)\, q_t^{\theta,\text{cat}}(\boldsymbol{x}_1 \mid \boldsymbol{x}_t), \tag{1}$$

$$q_t^{\theta,\text{num}} = \prod_{j=1}^{C} \mathcal{N}\left(x_{1,j}^{\text{num}}; \mu_{t,j}^{\theta,\text{num}}(\boldsymbol{x}_t), \sigma_j^2\right), \qquad q_t^{\theta,\text{cat}} = \prod_{g=1}^{G} \text{Cat}\left(\boldsymbol{x}_{1,g}^{\text{cat}}; \boldsymbol{\pi}_{t,g}^{\theta}(\boldsymbol{x}_t)\right), \tag{2}$$

where $\boldsymbol{\pi}_{t,g}^{\theta}(\boldsymbol{x}_t) = \text{Softmax}(\boldsymbol{\eta}_{t,g}^{\theta}(\boldsymbol{x}_t)) \in \Delta_{K_g-1}$. Numerical columns are quantile-transformed to standard-normal marginals (fitted on the training split), matching the unit scale of the base noise, so we set $\sigma_j^2 = 1$. Since the velocity in (B.8) depends on the posterior only through its mean, σ_j^2 acts as a constant weight on the loss of head j and leaves its minimizer $\mathbb{E}[x_{1,j}^{\text{num}} \mid \boldsymbol{x}_t]$ unchanged. The endpoint likelihood thus decomposes into squared-error regression for numerical columns and cross-entropy classification for categorical variables.

Definition 3.1 (TabTreeVFM Objective)**:** *Let $y_{g,k} = x_{1,g,k}^{\text{cat}}$ be the one-hot endpoint target for category k in group g. Up to constants independent of θ, the endpoint negative log-likelihood is*

$$\mathcal{L}_{\text{EF}}(\theta) = \mathbb{E}_{t,\boldsymbol{x}_1,\boldsymbol{x}_t}\left[\mathcal{L}_{\text{num}} + \mathcal{L}_{\text{cat}}\right], \tag{3}$$

where

$$\mathcal{L}_{\text{num}} = \frac{1}{2}\sum_{j=1}^{C}\left(x_{1,j}^{\text{num}} - \mu_{t,j}^{\theta,\text{num}}(\boldsymbol{x}_t)\right)^2 \qquad \textit{and} \qquad \mathcal{L}_{\text{cat}} = -\sum_{g=1}^{G}\sum_{k=1}^{K_g} y_{g,k} \log \pi_{t,g,k}^{\theta}(\boldsymbol{x}_t).$$

Both terms are Bregman losses of the form (B.9) and hence strictly proper: their population minimizers are exactly the posterior moments that enter the velocity in (B.8) (Gneiting & Raftery, 2007; Gneiting, 2011). The categorical term is the likelihood of a categorical endpoint posterior on the simplex, so the classifier head is itself a factor of the variational model.

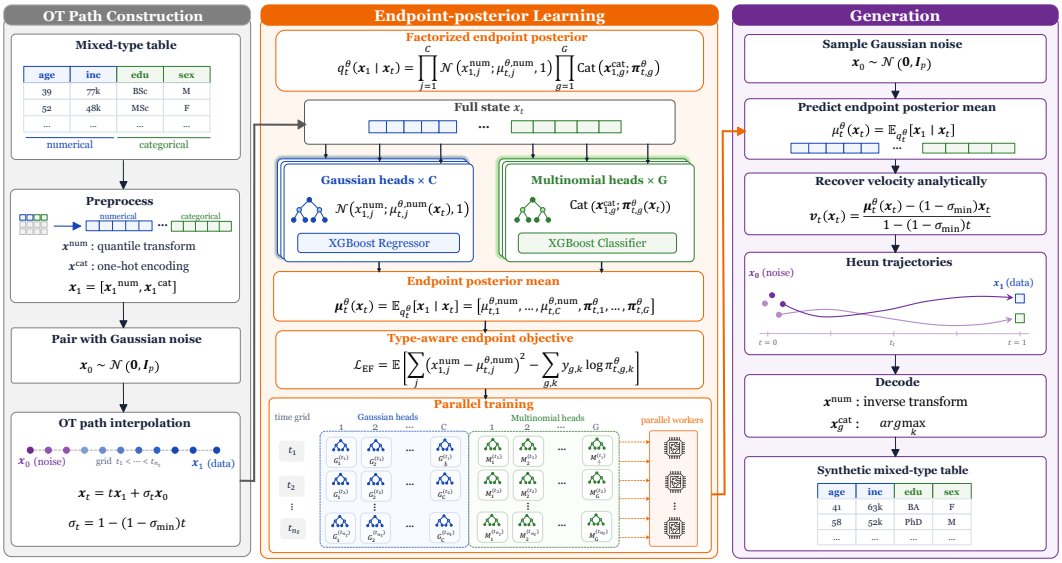


Figure 2: Overview of TabTreeVFM. Mixed-type data and Gaussian noise define a smoothed conditional OT path, on which independent XGBoost regression and multiclass heads learn factorized endpoint posteriors in parallel. Their means analytically define the velocity integrated by a Heun solver to generate mixed-type rows.

Why factorized endpoint prediction? Changing the supervised target from velocity to the clean endpoint creates an opportunity for type-aware modeling, but does not by itself impose the correct output geometry. A direct tree-flow model regresses the ambient velocity under squared error. A naive endpoint variant can likewise fit each coordinate of a one-hot block with an independent regressor,

$$\mathcal{L}_{\text{1hot-MSE}} = \sum_{k=1}^{K_g} \left(y_{g,k} - f_{t,g,k}(\boldsymbol{x}_t)\right)^2, \qquad \mathcal{L}_{\text{cat}} = -\sum_{k=1}^{K_g} y_{g,k} \log \pi^{\theta}_{t,g,k}(\boldsymbol{x}_t). \tag{4}$$

At the population level, both losses can identify the conditional indicator means. With finite boosted-tree estimators, however, the independent regression outputs are neither normalized nor coupled across categories. The multinomial head instead produces a simplex-valued vector at every state through a shared softmax likelihood, so the K_g coordinates form one posterior factor for the raw variable. Section 4.4 compares the supervised target, sampling-time simplex projection of one-hot regression outputs, and the multinomial categorical head in controlled ablations.

Factorization is exact for the velocity. As shown by Eijkelboom et al. (2024), the conditional velocity is affine in $\boldsymbol{x}_1$, so the marginal velocity depends on the endpoint posterior only through its mean $\mathbb{E}[\boldsymbol{x}_1 \mid \boldsymbol{x}_t]$, the concatenation of the per-column conditional means. The factorized posterior of (1) therefore induces the same velocity field as the full joint posterior: the population minimizer of each head in (3) is exactly one of these conditional means (Appendix B), so the heads can be fit independently and still recover the joint-posterior velocity exactly. Cross-column dependence is carried by the conditioning: every head takes the entire noisy row $\boldsymbol{x}_t$ as input, and all heads drive the same evolving state during integration.

3.3 Boosted-Tree Heads

We discretize time as $t_1, \ldots, t_{n_t}$ and take the expectation over t in (3) over $\{t_i\}_{i=1}^{n_t-1}$. At each of these times, sampled pairs $(\boldsymbol{x}_t, \boldsymbol{x}_1)$ define an ordinary supervised learning problem. Continuous endpoint means are fit with XGBoost regressors under squared error. Each raw categorical variable is fit with an XGBoost multiclass classifier under cross-entropy. The training set is formed by duplicating each real row K times and pairing it with fresh base noise, as in prior tree-based diffusion and flow models (Jolicoeur-Martineau et al., 2024; Cresswell & Kim, 2024).

The full training problem thus decomposes into $(n_t - 1) \times (C + G)$ independent fits over timesteps and feature heads, each with its own tree ensemble. These fits are embarrassingly parallel across CPU or GPU workers. Algorithms 1 and 2 give the complete procedure.

Time-grid alignment and Monte Carlo coverage. Training and Heun integration use the same time discretization, so every velocity evaluation calls heads trained at exactly that timestep; no interpolation between tree ensembles is required. For every real endpoint, the duplication factor K supplies K independent base-noise draws and therefore K trajectories through the interpolant. Because every endpoint is replicated equally, this construction leaves the empirical endpoint distribution unchanged and acts only as a Monte Carlo approximation of the base-noise expectation in (3). Increasing K gives denser coverage of the training states with an nK-row table per timestep; Appendices L and M quantify the quality–time–memory trade-off.

3.4 Velocity Recovery and Sampling

At sampling time, the heads trained in the previous subsection are combined into the posterior mean used by VFM: the numerical heads' predicted means and the categorical heads' predicted class-probability vectors are concatenated into

$$\boldsymbol{\mu}_t^\theta(\boldsymbol{x}_t) = \left[\boldsymbol{\mu}_t^{\theta,\mathrm{num}}(\boldsymbol{x}_t), \boldsymbol{\pi}_{t,1}^\theta(\boldsymbol{x}_t), \ldots, \boldsymbol{\pi}_{t,G}^\theta(\boldsymbol{x}_t)\right]. \tag{5}$$

We use the smoothed linear conditional OT path of Lipman et al. (2023), with $\alpha_t = t$ and $\sigma_t = 1 - (1 - \sigma_{\min})t$, where $\sigma_{\min} = 10^{-4}$ is the residual terminal noise scale. Substituting this path into (B.8) gives

$$\boldsymbol{v}_t^\theta(\boldsymbol{x}_t) = \frac{\boldsymbol{\mu}_t^\theta(\boldsymbol{x}_t) - (1 - \sigma_{\min})\boldsymbol{x}_t}{1 - (1 - \sigma_{\min})t}. \tag{6}$$

Sampling starts from $\boldsymbol{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and integrates the ODE with a Heun solver over the same discretized time grid used in training. At each step, the heads predict the endpoint-posterior parameters from the current state, the posterior mean gives the velocity through (6), and the state advances accordingly. Once the solver reaches $t = 1$, the ambient state is decoded into a table: continuous columns are inverse-transformed to the original scale, and each categorical group is decoded by $\arg\max$.

3.5 Endpoint Likelihood Geometry and XGBoost Updates

The endpoint-posterior formulation determines not only the supervised targets of TabTreeVFM, but also the curvature seen by the boosted-tree optimizer. For an exponential-family endpoint factor with natural parameter $\boldsymbol{\eta}$, the sample negative log-likelihood is $\ell(\boldsymbol{\eta}; \boldsymbol{x}_1) = A(\boldsymbol{\eta}) - \boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}_1) + \mathrm{const}$, and therefore

$$\nabla_{\boldsymbol{\eta}} \ell = \boldsymbol{\mu}(\boldsymbol{\eta}) - \boldsymbol{T}(\boldsymbol{x}_1), \qquad \nabla^2_{\boldsymbol{\eta}} \ell = \nabla^2 A(\boldsymbol{\eta}) = \boldsymbol{I}_F(\boldsymbol{\eta}), \tag{7}$$

where $\boldsymbol{I}_F$ denotes the Fisher information (Martens, 2020; Wainwright & Jordan, 2008). The likelihood Hessian of an exponential-family endpoint is thus exactly its statistical curvature. Fixed-variance Gaussian heads have constant curvature and reduce to ordinary squared-error boosting.

For a categorical endpoint head, let $\boldsymbol{p}_i$ denote the class probabilities predicted for sample i, and let $\boldsymbol{y}_i$ be its one-hot endpoint target. XGBoost's `multi:softprob` objective uses, for class k, the gradient and second-order weight

$$g_{ik} = p_{ik} - y_{ik}, \qquad h_{ik} = 2p_{ik}(1 - p_{ik}) = 2[\boldsymbol{I}_F(\boldsymbol{p}_i)]_{kk}. \tag{8}$$

The categorical Fisher matrix is $\boldsymbol{I}_F(\boldsymbol{p}_i) = \mathrm{diag}(\boldsymbol{p}_i) - \boldsymbol{p}_i\boldsymbol{p}_i^\top$. XGBoost uses the diagonal upper bound $2\mathrm{diag}(\boldsymbol{p}_i \odot (1 - \boldsymbol{p}_i))$ for the multinomial loss (Chen & Guestrin, 2016). Its multiclass update aggregates these curvature weights within each class-specific leaf. For a nonempty leaf I_{jk}, the update is

$$\Delta_{jk} = -\nu \frac{\sum_{i \in I_{jk}} g_{ik}}{\sum_{i \in I_{jk}} h_{ik} + \lambda}, \tag{9}$$

where ν is the learning rate and $\lambda > 0$ is the ℓ_2 regularization parameter (Chen & Guestrin, 2016). Equation (9) aggregates per-sample curvature within a leaf; we now characterize exactly how this aggregate relates to the Fisher curvature at a representative prediction of that leaf.

Let $n_{jk} = |I_{jk}|$ and define the leaf means and the within-leaf prediction dispersion

$$\bar{g}_{jk} = \frac{1}{n_{jk}} \sum_{i \in I_{jk}} g_{ik}, \qquad \bar{p}_{jk} = \frac{1}{n_{jk}} \sum_{i \in I_{jk}} p_{ik}, \qquad s^2_{jk} = \frac{1}{n_{jk}} \sum_{i \in I_{jk}} (p_{ik} - \bar{p}_{jk})^2.$$

Consider the representative diagonal-Fisher-scaled step obtained by evaluating the categorical curvature at the leaf-mean prediction,

$$\Delta_{jk}^{\mathrm{rep}} = -\nu \frac{\bar{g}_{jk}}{d_{jk}}, \qquad d_{jk} := 2\bar{p}_{jk}(1-\bar{p}_{jk}) + \frac{\lambda}{n_{jk}}. \tag{10}$$

Proposition 3.1 (Leaf aggregation of diagonal Fisher curvature)**:** *For a nonempty leaf I_{jk} and $\lambda > 0$, define $\rho_{jk} := 2s_{jk}^2/d_{jk}$. Then $0 \le \rho_{jk} < 1$, and the leaf update in Eq.* (9) *satisfies*

$$\Delta_{jk} = \frac{\Delta_{jk}^{\mathrm{rep}}}{1-\rho_{jk}}. \tag{11}$$

Consequently,

$$\left|\Delta_{jk} - \Delta_{jk}^{\mathrm{rep}}\right| = \left|\Delta_{jk}^{\mathrm{rep}}\right| \frac{\rho_{jk}}{1-\rho_{jk}}. \tag{12}$$

In particular, if $\rho_{jk} \le 1/2$, then

$$\left|\Delta_{jk} - \Delta_{jk}^{\mathrm{rep}}\right| \le 2\rho_{jk} \left|\Delta_{jk}^{\mathrm{rep}}\right| = \frac{4s_{jk}^2}{d_{jk}} \left|\Delta_{jk}^{\mathrm{rep}}\right|. \tag{13}$$

The proof is given in Appendix C.

When predictions within a leaf are homogeneous ($s_{jk}^2 = 0$), the leaf update equals the representative diagonal-Fisher-scaled step in Equation (10). For heterogeneous leaves, ρ_{jk} quantifies the relative curvature discrepancy, with larger λ reducing this discrepancy for fixed leaf predictions. Proposition 3.1 thus identifies the multiclass leaf update as a dispersion-corrected diagonal-Fisher step on the categorical endpoint likelihood, with ρ_{jk} a measurable per-leaf correction. Appendix D shows that this curvature lowers the late-time endpoint loss, and the full Fisher matrix adds no further gain.

3.6 Held-out-Calibrated Memorization Criterion

Beyond the endpoint-posterior architecture and training procedure above, we address a failure mode of existing high-fidelity generators. First, synthetic data do not automatically provide privacy: generators can reproduce exact or near-identical training rows, and synthetic datasets can pass standard distance-based privacy tests while remaining vulnerable to membership-inference and other privacy attacks (van Breugel et al., 2023; Ganev & De Cristofaro, 2025; Yao et al., 2026; Fang et al., 2025; 2026). Second, we find that existing generators exhibit a fidelity–privacy trade-off: as previewed in Figure 1 and detailed in Section 4, some of the highest-fidelity methods, such as CDTD (Mueller et al., 2025), rank poorly on privacy-related metrics. Aggregate fidelity metrics can reward a synthetic row that merely reproduces a training record rather than reflecting the underlying distribution, yet a raw replica count conflates such reproduction with natural collisions in the data distribution, especially in small or partly discrete tables. We therefore propose to calibrate reproduction using an exchangeable held-out set. Let $\mathcal{D}_{\mathrm{fit}}$ denote the training records used to fit a generator and $\mathcal{D}_{\mathrm{hold}}$ a disjoint held-out set. For each numerical column j, let r_j be its fixed range, reused for synthetic and held-out queries. We define a replica at tolerance τ by

$$c_\tau(\boldsymbol{x}; \mathcal{D}_{\mathrm{fit}}) = \mathbf{1}\left[\exists\, \boldsymbol{z} \in \mathcal{D}_{\mathrm{fit}} : \substack{x_j = z_j \ \forall j \in \mathrm{cat}, \\ |x_j - z_j| \le \tau r_j \ \forall j \in \mathrm{num}}\right], \tag{14}$$

so $\tau = 0$ detects exact row matches, including on constant columns. For any set S and fixed $\mathcal{D}_{\mathrm{fit}}$, we suppress the reference-set argument in c_τ and R_τ and define

$$R_\tau(S) = \frac{1}{|S|}\sum_{\boldsymbol{x}\in S} c_\tau(\boldsymbol{x}), \quad \Delta_\tau(\widehat{S}) = R_\tau(\widehat{S}) - R_\tau(\mathcal{D}_{\mathrm{hold}}), \quad \mathrm{Mem}_\tau(\widehat{S}) = \max\{\Delta_\tau(\widehat{S}), 0\}. \tag{15}$$

The signed gap preserves the direction relative to held-out reproduction, while Mem records only positive excess reproduction. Table 18 reports signed gaps to distinguish reproduction above and below the held-out reference.

Proposition 3.2 (Held-out calibration)**:** *Suppose $\mathcal{D}_{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{hold}}$ are independent samples from p_{data}, and each row of $\widehat{S}$ is drawn from p_{data} independently of the realized training records. Then $\mathbb{E}[R_\tau(\widehat{S}; \mathcal{D}_{\mathrm{fit}})] = \mathbb{E}[R_\tau(\mathcal{D}_{\mathrm{hold}}; \mathcal{D}_{\mathrm{fit}})]$ and $\mathbb{E}[\Delta_\tau(\widehat{S})] = 0$.*

Proof. Conditioned on $\mathcal{D}_{\mathrm{fit}}$, a released row and a held-out row have the same distribution and are evaluated by the same replica indicator. Averaging over rows and $\mathcal{D}_{\mathrm{fit}}$ gives the result. □

Thus, the held-out rate subtracts replicas induced by natural collisions, and Mem measures excess training-row reproduction. We evaluate this criterion in Section 4.3.

4 EXPERIMENTS

We evaluate TabTreeVFM on 10 mixed-type tabular benchmarks, comparing its distributional fidelity, downstream utility, and empirical privacy indicators with existing generators. We then examine training-row reproduction using the proposed held-out-calibrated Mem criterion and use controlled ablations to examine the supervised target, sampling-time simplex projection, and multinomial categorical modeling.

4.1 EXPERIMENTAL SETUP

Datasets. We evaluate on 10 datasets: Adult, Beijing, Churn, Credit-g, Default, Lending, Magic, News, NMES, and Shoppers. The standard benchmark uses the official Adult train–test split and random 90/10 splits otherwise. Dataset statistics are provided in Appendix F.

Baselines. We compare with 11 generators: GAN- and VAE-based CTGAN (Xu et al., 2019) and TVAE (Xu et al., 2019); diffusion models TabDDPM (Kotelnikov et al., 2023), STaSy (Kim et al., 2023), CoDi (Lee et al., 2023), TabSyn (Zhang et al., 2024), TabDiff (Shi et al., 2025), and CDTD (Mueller et al., 2025); and flow models TabbyFlow (Guzmán-Cordero et al., 2025), TabVVFM (Nasution et al., 2026), and Forest-Flow (Jolicoeur-Martineau et al., 2024).

Evaluation methods. We follow established evaluation protocols for tabular synthesis (Zhang et al., 2024; Shi et al., 2025; Guzmán-Cordero et al., 2025; Nasution et al., 2026; Mueller et al., 2025), covering fidelity, downstream utility, and empirical privacy. For **fidelity**, we report Shape and Trend for marginal and pairwise agreement, Wasserstein Distance (WD) and Jensen–Shannon Distance (JSD) for numerical and categorical marginals, and α-precision and β-recall for fidelity and coverage. C2ST uses an XGBoost discriminator against held-out real test rows with SDMetrics normalization; higher scores indicate less distinguishable samples. For **utility**, Machine Learning Efficiency (MLE) measures the relative test-performance gap between XGBoost predictors trained on synthetic and real data, using AUC for classification and RMSE for regression. For **privacy and memorization**, we report Distance to Closest Record (DCR); a transformed Membership Inference Attack (MIA) score, for which higher values indicate greater attack resistance; and Mem@1%, using $\tau = 1\%$ as defined in Section 3.6. Additional evaluation details are provided in Appendix H.

Implementation details. Following prior work (Zhang et al., 2024; Shi et al., 2025; Guzmán-Cordero et al., 2025), we generate synthetic tables from each fitted generator with 20 different seeds, each matching the training-set size for the corresponding experiment. By default, TabTreeVFM uses $n_t = 50$ and $K = 100$. Training takes 0.7–16 minutes on GPU and 1.3–123 minutes on CPU across the 10 datasets, below the GPU training time of TabDiff, TabbyFlow, TabSyn, and Forest-Flow on every dataset, and $K = 10$ cuts CPU training time by up to $8.8\times$ and peak memory by up to $5.3\times$. Appendix G gives further implementation details, and Appendices L, M, and J report efficiency, sensitivity, and training stability across 10 seeds, respectively.

4.2 OVERALL PERFORMANCE

Table 1 summarizes overall performance using average within-dataset ranks across the applicable benchmark datasets; Avg. is the unweighted mean across metrics. Appendix I reports detailed per-dataset results, with Figure 4 visualizing the means underlying these ranks; red crosses mark method–dataset cases for which no valid synthetic output was available.

Overall ranking. Table 1 shows that TabTreeVFM achieves the best overall average rank of 3.56, ahead of TabDiff at 4.45 and TabSyn at 5.04, and it remains first when the fidelity, utility, and privacy groups are weighted equally. The lead is broad rather than driven by one favorable criterion: TabTreeVFM obtains the best average rank on four fidelity metrics (i.e., Shape, Trend, WD, and C2ST) as well as MLE classification, and the second-best average rank on DCR. Thus, TabTreeVFM provides the strongest aggregate profile among the 12 generators across fidelity, utility, empirical privacy, and memorization.

Fidelity. The fidelity columns show consistent gains at complementary distributional levels. TabTreeVFM achieves average ranks of 1.50 on Shape, 2.50 on Trend, 1.80 on WD, and 2.20 on C2ST; its dataset-averaged C2ST score is 85.26, compared with 79.26 for TabbyFlow, 72.44 for TabDiff, and 20.80 for Forest-Flow. Forest-Flow's low score is amplified by the tree discriminator's sensitiv-

Table 1: Average within-dataset ranks across 10 datasets.

Method	Average Rank ↓												
	Fidelity							Utility		Privacy			Avg.
	Shape	Trend	WD	JSD	C2ST	α-Prec.	β-Rec.	MLE-C	MLE-R	DCR	MIA	Mem@1%	
CTGAN (Xu et al., 2019)	10.00	9.50	8.20	10.10	10.40	9.00	10.40	11.00	9.50	4.20	3.95	**3.90**	8.35
TVAE (Xu et al., 2019)	8.60	9.20	8.60	9.50	10.00	8.90	9.30	8.50	9.75	7.90	4.80	6.20	8.44
STaSy (Kim et al., 2023)	9.60	9.40	10.50	10.10	8.10	10.10	9.30	9.00	5.75	4.00	3.60	5.15	7.88
CoDi (Lee et al., 2023)	11.20	10.00	10.40	10.10	10.20	9.10	7.40	9.67	8.75	7.30	**3.05**	4.80	8.50
TabDDPM (Kotelnikov et al., 2023)	8.22	9.00	8.44	7.78	7.89	9.11	7.78	6.50	12.00	7.89	7.44	5.67	8.14
TabSyn (Zhang et al., 2024)	4.00	4.70	4.30	5.80	4.00	4.50	6.00	5.17	4.75	**3.80**	7.20	6.25	5.04
TabDiff (Shi et al., 2025)	2.60	3.90	3.00	2.90	3.00	3.30	5.10	3.50	4.25	6.40	8.70	6.70	4.45
CDTD (Mueller et al., 2025)	3.30	3.70	4.10	**2.70**	4.10	**2.60**	**1.30**	4.17	**3.00**	11.30	10.70	11.80	5.23
TabbyFlow (Guzmán-Cordero et al., 2025)	4.80	3.70	5.50	6.00	2.70	7.00	1.70	3.33	**3.00**	9.90	9.90	9.70	5.60
TabVVFM (Nasution et al., 2026)	7.10	7.50	9.10	**2.70**	7.40	6.80	8.00	6.67	6.25	5.50	6.20	6.55	6.65
Forest-Flow (Jolicoeur-Martineau et al., 2024)	6.70	4.60	3.70	6.20	7.60	3.40	6.40	7.50	7.50	5.50	5.60	5.25	5.83
TabTreeVFM	**1.50**	**2.50**	**1.80**	3.70	**2.20**	3.90	4.90	**3.00**	3.50	3.90	6.40	5.40	**3.56**

Table 2: Component ablation averaged over 10 datasets.

Variant	Supervised target	Categorical objective	Fidelity							Utility		Privacy		
			Shape ↓	Trend ↓	WD ↓	JSD ↓	C2ST ↑	α-Prec. ↑	β-Rec. ↑	MLE-C gap ↓	MLE-R gap ↓	|DCR − 50| ↓	MIA ↑	Mem@1% ↓
Direct velocity regression	Velocity	Coordinate-wise MSE	2.51	4.89	0.0044	0.0216	78.62	95.68	43.79	3.10	35.76	**1.47**	0.9843	0.07
Endpoint regression	Endpoint	One-hot MSE	1.52	4.44	0.0042	0.0197	80.60	95.65	44.17	3.15	36.33	1.48	**0.9855**	**0.04**
Endpoint regression + simplex projection	Endpoint	One-hot MSE + proj.	1.43	4.18	0.0042	0.0170	81.45	96.48	44.38	2.85	34.77	1.50	0.9837	0.05
TabTreeVFM	Endpoint	Multinomial CE	**1.31**	**2.94**	**0.0039**	**0.0146**	**85.26**	**96.85**	**47.56**	**1.77**	**33.19**	2.48	0.9837	0.14

ity to its min–max representation of numerical columns: under the logistic discriminator of Table 12 it rises to 76.53; controlled ablations in Section 4.4 isolate each modeling choice. TabTreeVFM achieves the strongest overall fidelity profile across the seven metrics; CDTD leads on α-precision and β-recall but has the largest DCR deviation, the lowest MIA score, and the highest Mem@1% of all generators. Appendix K visualizes marginals and pairwise dependency errors.

Utility. Table 1 shows that TabTreeVFM obtains the best MLE classification rank of 3.00 and a regression rank of 3.50. Predictors trained on its synthetic data have an average gap of 1.8% from the train-on-real reference on classification tasks; on regression tasks, the gaps of all generators are dominated by the near-deterministic Lending target, which amplifies small distributional errors. Thus, TabTreeVFM delivers strong downstream utility across both task types.

Empirical privacy. The empirical privacy indicators reveal distinctions not captured by aggregate fidelity. TabTreeVFM achieves the second-best average DCR rank of 3.90 and the smallest dataset-averaged absolute deviation from the 50% reference, at 2.48 percentage points, compared with 26.25 percentage points for CDTD. Its mean MIA score of 0.984 is the highest among the four best-ranked generators overall. Section 4.3 complements these indicators with a direct analysis of training-row reproduction using the held-out-calibrated Mem criterion.

4.3 Training-Row Reproduction

Aggregate fidelity and privacy scores do not establish whether a generator reproduces training rows. We therefore measure full-row reproduction relative to a disjoint held-out reference, applying the Mem criterion of Section 3.6 to the synthetic tables of every method–dataset pair. Mem records excess reproduction of exact or tolerance-bounded replicas above the held-out rate; we read it jointly with coverage, since low excess reproduction is informative for generators that also capture the data distribution. Table 1 reports the Mem@1% ranks of all 12 generators, and Figure 3 plots representative replica-rate curves against the held-out reference. Table 19 repeats the comparison for τ from 0.1% to 5%.

The comparison separates quality from memorization. Averaged across the 10 datasets, CDTD has the highest Mem@1% at 8.37 percentage points, followed by TabbyFlow at 3.51; both are among the strongest generators on fidelity and utility, with CDTD leading α-precision and β-recall. TabTreeVFM has an average Mem@1% of 0.14 percentage points, below TabDiff at 0.30 and TabSyn at 0.31 and thus the lowest value among the four best-ranked generators overall. Its excess stays below one percentage point on every dataset. Generators with lower Mem@1%, such as Forest-Flow at 0.03 and CTGAN at zero, rank lower on coverage and overall, so their low excess reproduction coincides with limited coverage.

The signed gaps in Table 18 give the direction of these differences. On Default, CTGAN and CoDi fall below the held-out reproduction rate by 2.560 and 3.362 percentage points, respectively. Together with their lower β-recall, these gaps are consistent with limited coverage of regions contain-

ing training-row replicas. For TabTreeVFM, every signed gap lies within 1.1 percentage points of the held-out rate.

The per-dataset curves localize these gaps. CDTD shows large excess reproduction on Credit-g, Shoppers, and NMES, while TabbyFlow's Mem@1% is driven by Credit-g. On this financial dataset, CDTD and TabbyFlow reach replica rates of 28.3% and 27.6% at $\tau = 1\%$ against a zero held-out rate; TabTreeVFM also remains at zero. Reproducing applicant profiles at this scale raises record-linkage and membership-exposure concerns in synthetic financial-data releases.

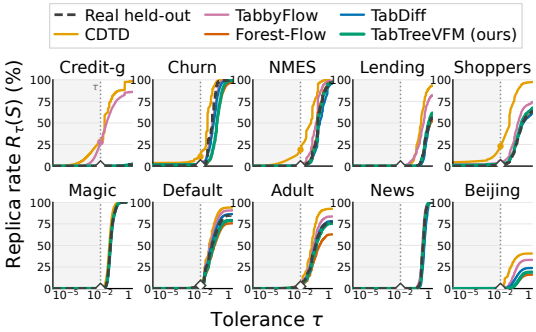


Figure 3: Replica-rate curves $R_\tau(S)$ for representative generators and real held-out data. At $\tau = 1\%$ (dotted line), the vertical gap between a method marker and the held-out diamond is Δ_τ; its positive part is Mem@1%.

Strong synthetic-data quality can therefore coexist with training-row reproduction but does not require it: CDTD and TabbyFlow reproduce specific training rows at substantial rates, whereas TabTreeVFM combines the best aggregate rank with low average excess full-row reproduction.

4.4 Component Ablations

Table 2 compares the supervised target, simplex projection, and categorical heads. All four variants use independent heads and share the probability path, input encoding, solver, and K and depth settings within each dataset–metric comparison. Each is evaluated on 20 tables generated with different seeds; MLE gaps are averaged separately for classification and regression.

The endpoint target improves marginals. Switching from velocity to endpoint MSE reduces Shape error from 2.51 to 1.52 and raises C2ST from 78.62 to 80.60; the remaining gains come from the categorical head.

Simplex projection alone gives smaller gains. Projecting categorical endpoint predictions at sampling time reduces JSD from 0.0197 to 0.0170 and the classification utility gap from 3.15% to 2.85%, while increasing C2ST from 80.60 to 81.45.

Multinomial heads improve quality and utility. Relative to projected regression, multinomial heads reduce Trend from 4.18 to 2.94, raise C2ST from 81.45 to 85.26 and β-recall from 44.38 to 47.56, and narrow the classification utility gap from 2.85% to 1.77%. Trend and β-recall improve on all 10 datasets against both endpoint-MSE controls. Relative to direct velocity regression, the full model improves all seven fidelity metrics and both utility gaps, raising C2ST from 78.62 to 85.26.

Empirical privacy and memorization. The privacy indicators vary little across variants: DCR deviation stays at or below 2.5 percentage points, MIA within 0.002, and Mem@1% below 0.15 percentage points, so the fidelity and utility gains of the multinomial heads are not obtained through memorization.

5 Conclusion

We presented TabTreeVFM, a tree-boosted endpoint-posterior model that aligns boosted-tree prediction with the statistical structure of mixed-type tabular endpoints, and proved that its multiclass leaf updates are diagonal-Fisher-scaled steps up to an explicit within-leaf dispersion factor, so type-aware endpoint learning obtains Fisher-scaled updates from second-order boosting itself. Across 10 benchmarks, TabTreeVFM achieves the best overall average rank among 12 generators, while Mem reveals excess training-row reproduction above the natural collision rate in several strong neural baselines but little for TabTreeVFM. Type-aware endpoint prediction thus makes boosted trees effective, CPU-trainable components of flow-based tabular generators. Finally, because the endpoint posterior factorizes over one-hot blocks, a column with very many levels can be encoded as several low-cardinality categorical factors handled by the same multinomial heads; evaluating TabTreeVFM on such data is a natural direction for future work.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, including language refinement, text condensation, section organization, and LaTeX formatting. We did not use generative AI tools to generate research ideas, theoretical results, experimental results, or scientific claims. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

ETHICS STATEMENT

This work studies synthetic tabular data generation using publicly available benchmark datasets; no new data involving human subjects were collected. Synthetic data should not be assumed to provide privacy merely by virtue of being synthetic: generators may reproduce or reveal information about training records. We explicitly study this risk through membership-inference, distance-based, and held-out-calibrated memorization evaluations. These empirical evaluations do not constitute formal privacy guarantees, and synthetic data generated by the proposed method should therefore be subject to appropriate privacy assessment before deployment in sensitive applications.

REPRODUCIBILITY STATEMENT

An anonymized implementation of TabTreeVFM is provided in the supplementary material. Section 4 and Appendices F–H describe the datasets, preprocessing, train–test splits, model configurations, baseline implementations, and evaluation protocols. Algorithms 1–2 provide the complete training and generation procedures. Additional efficiency and hyperparameter analyses are provided in Appendices L–M.

REFERENCES

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*, pp. 290–306. PMLR, 2022.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. URL `https://doi.org/10.24432/C5XW20`.

R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. URL `https://doi.org/10.24432/C52C8B`.

Song Chen. Beijing PM2.5. UCI Machine Learning Repository, 2015. URL `https://doi.org/10.24432/C5JS49`.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Graham Cormode, Shripad Gade, Samuel Maddock, and Enayat Ullah. Synthetic tabular data: Methods, attacks and defenses. *Proceedings of the VLDB Endowment*, 18(12):5448–5450, 2025.

Jesse C. Cresswell and Taewoo Kim. Scaling up diffusion and flow-based XGBoost models. In *ICML 2024 AI for Science Workshop*, 2024.

Partha Deb and Pravin K. Trivedi. Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics*, 12(3):313–336, 1997.

Tony Duan, Anand Avati, Daisy Yi Ding, Khanh K. Thai, Sanjay Basu, Andrew Ng, and Alejandro Schuler. NGBoost: Natural gradient boosting for probabilistic prediction. In *International Conference on Machine Learning*, pp. 2690–2700. PMLR, 2020.

Floor Eijkelboom, Grigory Bartosh, Christian Andersson Naesseth, Max Welling, and Jan-Willem van de Meent. Variational flow matching for graph generation. *Advances in Neural Information Processing Systems*, 37:11735–11764, 2024.

Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. Relational data synthesis using generative adversarial networks: A design space exploration. *Proceedings of the VLDB Endowment*, 13(11):1962–1975, 2020.

Zhengyu Fang, Zhimeng Jiang, Huiyuan Chen, Xiao Li, and Jing Li. Understanding and mitigating memorization in diffusion models for tabular data. In *International Conference on Machine Learning*, pp. 15976–16005. PMLR, 2025.

Zhengyu Fang, Zhimeng Jiang, Huiyuan Chen, Xiaoge Zhang, Kaiyu Tang, Xiao Li, and Jing Li. A closer look on memorization in tabular diffusion model: A data-centric perspective. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=p2n88DfaXB.

Kelwin Fernandes, Pedro Vinagre, Paulo Cortez, and Pedro Sernadela. Online News Popularity. UCI Machine Learning Repository, 2015. URL https://doi.org/10.24432/C5NS3V.

Georgi Ganev and Emiliano De Cristofaro. The inadequacy of similarity-based privacy metrics: Privacy attacks against "truly anonymous" synthetic datasets. In *IEEE Symposium on Security and Privacy (SP)*, pp. 4007–4025, 2025.

Yunqing Ge, Jianbin Qin, Shuyuan Zheng, Yongrui Zhong, Bo Tang, Yu-Xuan Qiu, Rui Mao, Ye Yuan, Makoto Onizuka, and Chuan Xiao. Privacy-enhanced database synthesis for benchmark publishing. *Proceedings of the VLDB Endowment*, 18(2):413–425, 2024.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35: 507–520, 2022.

Andrés Guzmán-Cordero, Floor Eijkelboom, and Jan-Willem van de Meent. Exponential family variational flow matching for tabular data generation. In *International Conference on Machine Learning*, pp. 21516–21529. PMLR, 2025.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. URL https://doi.org/10.24432/C5NC77.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi S. Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary Markov processes. In *The Thirteenth International Conference on Learning Representations*, 2025.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International Conference on Artificial Intelligence and Statistics*, pp. 1288–1296. PMLR, 2024.

A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi. Improved churn prediction in telecommunication industry using data mining techniques. *Applied Soft Computing*, 24:994–1012, 2014.

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

G. Charbel N. Kindji, Lina M. Rojas-Barahona, Elisa Fromont, and Tanguy Urvoy. Tabular data generation models: An in-depth survey and performance benchmarks with extensive tuning. *Neurocomputing*, 658:131655, 2025.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pp. 17564–17579. PMLR, 2023.

Anton Danholt Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1):6, 2025.

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

Lending Club. Loan Data from Lending Club. OpenIntro data set `loans_full_schema`, 2018. URL `https://www.openintro.org/data/index.php?data=loans_full_schema`.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

Markus Mueller, Kathrin Gruber, and Dennis Fok. Continuous diffusion for mixed-type tabular data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Markus Mueller, Kathrin Gruber, and Dennis Fok. Cascaded flow matching for heterogeneous tabular data with mixed-type features. In *International Conference on Machine Learning*. PMLR, 2026.

Bahrul Ilmi Nasution, Floor Eijkelboom, Mark Elliot, Richard Allmendinger, and Christian A. Naesseth. Flow matching for tabular data synthesis. *Transactions on Machine Learning Research*, 2026. URL `https://openreview.net/forum?id=RdOjoAa66L`.

Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment*, 11(10):1071–1083, 2018.

Michael Platzer and Thomas Reutterer. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in Big Data*, 4:679939, 2021.

Cemal Okan Sakar and Yomi Kastro. Online Shoppers Purchasing Intention Dataset. UCI Machine Learning Repository, 2018. URL `https://doi.org/10.24432/C5F88Q`.

Anupam Sanghi and Jayant R. Haritsa. Synthetic data generation for enterprise DBMS. In *Proceedings of the 39th IEEE International Conference on Data Engineering*, pp. 3585–3588, 2023.

Davide Scassola, Dylan Ponsford, Adrián Javaloy, Sebastiano Saccani, Luca Bortolussi, Henry Gouk, and Antonio Vergari. A sobering look at tabular data generation via probabilistic circuits. In *Conference on Uncertainty in Artificial Intelligence*, pp. 6026–6063. PMLR, 2026.

Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. TabDiff: a mixed-type diffusion model for tabular data generation. In *The Thirteenth International Conference on Learning Representations*, 2025.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. In *International Conference on Artificial Intelligence and Statistics*, pp. 3493–3514. PMLR, 2023.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. *Advances in Neural Information Processing Systems*, 32:7333–7343, 2019.

Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. The DCR delusion: Measuring the privacy risk of synthetic data. In *European Symposium on Research in Computer Security (ESORICS 2025)*, pp. 469–487. Springer, 2026.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2): 2473–2480, 2009.

El-Hacen Zein and Tanguy Urvoy. Tabular data generation: Can we fool XGBoost? In *NeurIPS 2022 First Table Representation Workshop*, 2022.

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.

Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. PrivBayes: Private data release via Bayesian networks. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 1423–1434, 2014.

APPENDIX

A NOTATION

Table 3: Frequently used notations.

Symbol	Description
n, p	Training rows; ambient width $p = C + \sum_g K_g$
C, G	Continuous columns; raw categorical variables
K_g	Number of categories of variable g
$\boldsymbol{x}$	A record, $[\boldsymbol{x}^{\text{num}}, \boldsymbol{x}^{\text{cat}}]$
$\boldsymbol{x}_g^{\text{cat}}$	One-hot block of categorical variable g
Δ_{K_g-1}	Probability simplex over its K_g categories
p_0, p_{data}	Base (Gaussian) and data distributions
$\boldsymbol{x}_t$	Interpolant state, $\alpha_t \boldsymbol{x}_1 + \sigma_t \boldsymbol{x}_0$
α_t, σ_t	Coefficients of the probability path
$\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$	Conditional velocity of the path
$\boldsymbol{v}_t^\theta(\boldsymbol{x}_t)$	Learned marginal velocity field
$q_t^\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_t)$	Variational endpoint posterior
$\boldsymbol{\mu}_t^\theta(\boldsymbol{x}_t)$	Endpoint posterior mean; defines $\boldsymbol{v}_t^\theta$
$\boldsymbol{\pi}_{t,g}^\theta$	Predicted class probabilities for group g
$\boldsymbol{\eta}$, $A(\boldsymbol{\eta})$	Natural parameter; log-partition function
$\boldsymbol{I}_F(\boldsymbol{\eta})$	Fisher information of the endpoint likelihood
n_t, K	Number of time-grid points; duplication factor
$\hat{f}_{t,j}^{\text{num}}$, $\hat{f}_{t,g}^{\text{cat}}$	Boosted-tree endpoint heads at time t
$\hat{\boldsymbol{\pi}}_{t,g}$	Class probabilities returned by $\hat{f}_{t,g}^{\text{cat}}$
$\sigma_{\min}$	Residual terminal noise scale

B BACKGROUND: FLOW MATCHING AND ENDPOINT POSTERIORS

This appendix reviews the flow-matching, variational flow-matching, and exponential-family endpoint-loss formulations on which TabTreeVFM builds (Section 3).

B.1 FLOW MATCHING

Flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025) learns a time-dependent velocity field $\boldsymbol{v}_t^\theta$, $t \in [0,1]$, whose ODE transports samples from a base distribution p_0 toward the data distribution p_{data}. For each data endpoint $\boldsymbol{x}_1 \sim p_{\text{data}}$, conditional flow matching defines a tractable path $p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$ and its velocity $\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$. Averaging over endpoints gives the marginal velocity

$$\boldsymbol{u}_t(\boldsymbol{x}_t) = \int \boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) \frac{p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) p_{\text{data}}(\boldsymbol{x}_1)}{p_t(\boldsymbol{x}_t)} \, d\boldsymbol{x}_1, \tag{B.1}$$

but this integral is generally unavailable. Conditional flow matching therefore minimizes

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\boldsymbol{x}_1,\boldsymbol{x}_t} \left[\left\| \boldsymbol{v}_t^\theta(\boldsymbol{x}_t) - \boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) \right\|_2^2 \right], \tag{B.2}$$

where $\boldsymbol{x}_1 \sim p_{\text{data}}$, $\boldsymbol{x}_t \sim p_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$, and $t \sim \mathcal{U}[0,1]$. We use interpolation paths

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_1 + \sigma_t \boldsymbol{x}_0, \qquad \boldsymbol{x}_0 \sim p_0, \tag{B.3}$$

whose conditional velocity is

$$\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) = \left(\dot{\alpha}_t - \frac{\dot{\sigma}_t}{\sigma_t} \alpha_t \right) \boldsymbol{x}_1 + \frac{\dot{\sigma}_t}{\sigma_t} \boldsymbol{x}_t. \tag{B.4}$$

B.2 VARIATIONAL FLOW MATCHING

Variational flow matching introduces an approximate posterior $q_t^\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_t)$ over the clean endpoint (Eijkelboom et al., 2024) and defines the velocity as the posterior expectation

$$\boldsymbol{v}_t^\theta(\boldsymbol{x}_t) = \mathbb{E}_{q_t^\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_t)} \left[\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1) \right]. \tag{B.5}$$

With a block-factorized endpoint posterior,

$$q_t^\theta(\boldsymbol{x}_1 \mid \boldsymbol{x}_t) = \prod_b q_t^\theta(\boldsymbol{x}_{1,b} \mid \boldsymbol{x}_t), \tag{B.6}$$

the VFM loss is the endpoint negative log-likelihood

$$\mathcal{L}_{\mathrm{MF-VFM}}(\theta) = -\mathbb{E}_{t,\boldsymbol{x}_1,\boldsymbol{x}_t}\left[\sum_b \log q_t^\theta(\boldsymbol{x}_{1,b} \mid \boldsymbol{x}_t)\right] + \text{const.} \tag{B.7}$$

Each block b is either one numerical variable, modeled by a scalar Gaussian factor, or one categorical variable, modeled by a categorical factor supported on the one-hot vertices of Δ_{K_g-1}, whose mean is a probability vector in that simplex. Every factor conditions on the full state $\boldsymbol{x}_t$, so dependence between columns enters through the conditioning: the factorization is placed on the variational posterior, not on the data distribution. Because $\boldsymbol{u}_t(\boldsymbol{x}_t \mid \boldsymbol{x}_1)$ in (B.4) is affine in $\boldsymbol{x}_1$, the marginal velocity depends on the posterior only through the block means $\mathbb{E}[\boldsymbol{x}_{1,b} \mid \boldsymbol{x}_t]$, and the block-factorized family represents these means without restriction. The factorization therefore does not restrict the induced velocity field (Eijkelboom et al., 2024).

Substituting (B.4) into (B.5) gives a closed-form velocity in terms of the posterior endpoint mean $\boldsymbol{\mu}_t^\theta(\boldsymbol{x}_t) = \mathbb{E}_{q_t^\theta(\boldsymbol{x}_1|\boldsymbol{x}_t)}[\boldsymbol{x}_1]$:

$$\boldsymbol{v}_t^\theta(\boldsymbol{x}_t) = \left(\dot{\alpha}_t - \frac{\dot{\sigma}_t}{\sigma_t}\alpha_t\right)\boldsymbol{\mu}_t^\theta(\boldsymbol{x}_t) + \frac{\dot{\sigma}_t}{\sigma_t}\boldsymbol{x}_t. \tag{B.8}$$

Thus VFM converts velocity estimation into endpoint-posterior moment estimation.

B.3 Exponential-Family Endpoint Losses

EF-VFM (Guzmán-Cordero et al., 2025) chooses each endpoint factor from an exponential family. For $p(\boldsymbol{x} \mid \boldsymbol{\eta}) = h(\boldsymbol{x})\exp(\boldsymbol{\eta}^\top \boldsymbol{T}(\boldsymbol{x}) - A(\boldsymbol{\eta}))$, the mean parameter is $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{T}(\boldsymbol{x})] = \nabla A(\boldsymbol{\eta})$. For minimal regular exponential families, the endpoint negative log-likelihood is equivalent, up to constants, to a Bregman divergence in mean-parameter space (Guzmán-Cordero et al., 2025; Banerjee et al., 2005):

$$\mathcal{L}_{\mathrm{VFM}}(\theta) = \mathbb{E}_{t,p_t(\boldsymbol{x}_t)}\left[D_{A^*}\left(\boldsymbol{\mu}_t(\boldsymbol{x}_t) \,\|\, \boldsymbol{\mu}_t^\theta(\boldsymbol{x}_t)\right)\right] + \text{const.} \tag{B.9}$$

For fixed-variance Gaussian endpoint factors this gives squared error. For categorical endpoint factors with one-hot sufficient statistics this gives cross-entropy; the one-hot statistics are affinely dependent, so the correspondence is understood on the family's minimal representation, while the implementation uses the equivalent K_g-logit softmax parameterization. Conditional flow-matching objectives remain valid for losses of this Bregman form in general (Holderrieth et al., 2025).

C Proofs

C.1 Proof of Proposition 3.1

We analyze the leaf update with ℓ_2 regularization, `reg_alpha=0`, and `max_delta_step=0`, using the curvature weights $h_{ik} = 2p_{ik}(1-p_{ik})$ of Eq. (8).

Proof: Using

$$\frac{1}{n_{jk}}\sum_{i\in I_{jk}} p_{ik}(1-p_{ik}) = \bar{p}_{jk}(1-\bar{p}_{jk}) - s_{jk}^2,$$

Equations (8)–(9) give

$$\Delta_{jk} = -\nu\frac{\bar{g}_{jk}}{d_{jk} - 2s_{jk}^2}.$$

Since $\lambda > 0$,

$$d_{jk} - 2s_{jk}^2 = \frac{2}{n_{jk}}\sum_{i\in I_{jk}} p_{ik}(1-p_{ik}) + \frac{\lambda}{n_{jk}} > 0,$$

so $\rho_{jk} < 1$. Since $d_{jk} - 2s_{jk}^2 = d_{jk}(1-\rho_{jk})$, Equation (11) follows immediately. Subtracting $\Delta_{jk}^{\mathrm{rep}}$ gives Equation (12); when $\rho_{jk} \le 1/2$, $1/(1-\rho_{jk}) \le 2$, yielding Equation (13). □

D EMPIRICAL ROLE OF THE FISHER CURVATURE

Section 3.5 relates the second-order weights of XGBoost's multiclass update to the diagonal Fisher information of the categorical endpoint likelihood. This appendix measures how much this curvature contributes relative to a curvature-free update, and whether the full Fisher matrix, as used in natural-gradient boosting (Duan et al., 2020), improves on the diagonal update.

Setup. Across the nine benchmarks other than Magic, whose only categorical variable is binary, we consider all 46 non-binary categorical variables, whose cardinalities range from 3 to 50. Binary variables are excluded because the diagonal and full updates coincide for them. For each variable we select five times from the training grid. At each time we fit the variable's categorical head on the same supervised problem as in Section 3.3: every training row is duplicated K times and paired with fresh base noise, and the head predicts the variable's category from $\boldsymbol{x}_t$. All settings follow Appendix G.1.

By default, XGBoost's multiclass update grows one tree per class in each boosting round, so each class partitions the rows differently. A full-Fisher step couples the classes within a leaf and therefore needs a partition common to all of them. We thus also use *shared trees*, in which a single tree per round splits the rows once and stores one leaf value per class. The four variants differ only in the leaf update or the tree structure:

(i) *Diagonal Fisher* (default): one tree per class, with curvature weights $2p_{ik}(1-p_{ik})$.
(ii) *Constant curvature*: one tree per class, with these weights replaced by their maximum $1/2$.
(iii) *Shared-tree diagonal Fisher*: one shared tree per round, with the same diagonal weights.
(iv) *Shared-tree full Fisher*: shared trees grown in the same way, with each leaf set to the natural-gradient step $-\nu(\sum_i \boldsymbol{F}_i + \lambda \boldsymbol{I})^{-1}\boldsymbol{G}$, where $\boldsymbol{F}_i = \mathrm{diag}(\boldsymbol{p}_i) - \boldsymbol{p}_i\boldsymbol{p}_i^\top$ and $\boldsymbol{G}$ is the leaf gradient sum.

Comparing (iii) with (i) isolates the effect of sharing the tree, and comparing (iv) with (iii) isolates the effect of the full matrix. Each fit is scored by its endpoint cross-entropy on held-out test rows paired with fresh noise.

Table 4: Endpoint cross-entropy relative to the diagonal Fisher update, geometrically averaged over the 46 categorical heads. For the diagonal Fisher update, the last two columns report the row-weighted mean of ρ_{jk} and the percentage of rows in leaves with $\lambda/n_{jk} > 2\bar{p}_{jk}(1-\bar{p}_{jk})$.

t	Loss ratio to diagonal Fisher			ρ_{jk}	λ-dominated (%)
	Constant curvature	Shared-tree diagonal Fisher	Shared-tree full Fisher		
0.24	1.00	1.00	1.01	0.04	1
0.49	1.02	1.02	1.03	0.19	2
0.73	1.20	1.04	1.26	0.57	3
0.90	98.9	1.21	2.07	0.42	26
0.96	108	1.46	2.03	0.41	26

Endpoint-level comparison. As Table 4 shows, with constant curvature the loss is unchanged at early times but grows by about two orders of magnitude for $t \geq 0.90$. There the endpoint posterior concentrates near the vertices of the simplex, the Fisher weights $p_{ik}(1-p_{ik})$ become small, and the Fisher-scaled update adapts its step size accordingly, whereas the constant-curvature update cannot. Sharing the tree across classes raises the loss moderately, mainly at late times. On these shared trees the full Fisher matrix changes the leaf values substantially but raises the loss further at every time.

The last two columns of Table 4 relate these results to Proposition 3.1. The dispersion ratio ρ_{jk} peaks at intermediate times, where leaves mix rows with different predictions; Eq. (11) relates the implemented step to the representative Fisher step through the factor $1/(1-\rho_{jk})$. At late times, where the Fisher curvature $2\bar{p}_{jk}(1-\bar{p}_{jk})$ is small, the ℓ_2 term λ/n_{jk} becomes the larger term in the denominator d_{jk} for about a quarter of the rows and bounds the step.

Effect on generation quality. We trained complete models with each categorical head on Credit-g, Lending, and Shoppers, sharing the numerical heads, and evaluated 20 synthetic tables per variant with the protocol of Appendix H. Table 5 reports the results. Constant curvature changes the categorical marginals most: JSD rises by about half on Lending and Shoppers, together with a higher Shape error, while on Credit-g the main change is a 20% higher Trend error. On Credit-g and Shoppers, the

Algorithm 1: TabTreeVFM Training

Input: Dataset $\{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$; Continuous column indices $\mathcal{C}$; Categorical one-hot groups $\{G_g\}_{g=1}^{G}$; Number of time-grid points n_t; Duplication factor K;

Output: Trained ensembles $\big\{\{\hat{f}^{\mathrm{num}}_{t_i,j}\}_{j\in\mathcal{C}}, \{\hat{f}^{\mathrm{cat}}_{t_i,g}\}_{g=1}^{G}\big\}_{i=1}^{n_t-1}$;

Apply quantile normalization to the continuous columns indexed by $\mathcal{C}$;
Apply one-hot encoding to each categorical group G_g;
$\boldsymbol{X}_1 \leftarrow [\boldsymbol{X}^{\mathrm{num}} \mid \boldsymbol{X}^{\mathrm{cat}}] \in \mathbb{R}^{n\times p}$;
$\boldsymbol{X}_1 \leftarrow \mathrm{tile}(\boldsymbol{X}_1, K) \in \mathbb{R}^{nK\times p}$;
$\boldsymbol{X}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_p)^{nK}$;
$\{t_i\}_{i=1}^{n_t} \leftarrow \mathrm{linspace}(0, 1, n_t)$;
for $i = 1$ **to** $n_t - 1$ **do in parallel**
  $\boldsymbol{X}_{t_i} \leftarrow \alpha_{t_i}\boldsymbol{X}_1 + \sigma_{t_i}\boldsymbol{X}_0$; ▷ Samples along the selected OT path
  foreach $j \in \mathcal{C}$ **do**
    $\hat{f}^{\mathrm{num}}_{t_i,j} \leftarrow \mathrm{XGBRegressor.fit}\big(\boldsymbol{X}_{t_i}, \boldsymbol{X}_1[:, j]\big)$
  for $g = 1$ **to** G **do**
    $\boldsymbol{y}_g \leftarrow \arg\max_k \boldsymbol{X}_1[:, G_g]_k$; ▷ Recover integer labels from one-hot representations
    $\hat{f}^{\mathrm{cat}}_{t_i,g} \leftarrow \mathrm{XGBClassifier.fit}\big(\boldsymbol{X}_{t_i}, \boldsymbol{y}_g\big)$
return $\big\{\{\hat{f}^{\mathrm{num}}_{t_i,j}\}_{j\in\mathcal{C}}, \{\hat{f}^{\mathrm{cat}}_{t_i,g}\}_{g=1}^{G}\big\}_{i=1}^{n_t-1}$;

full-Fisher update and the default are within one standard deviation of each other on every metric. On Lending it has higher JSD and Shape error than the default, and most of this gap already appears with the shared diagonal tree. Differences in C2ST and MLE stay within one standard deviation, except for the lower MLE of constant curvature on Shoppers. Removing the curvature thus degrades the generated tables, whereas the full matrix gives no consistent gain over the diagonal update.

Table 5: Generation quality for the four categorical-head variants.

Dataset	Categorical head	Shape ↓	Trend ↓	C2ST ↑	JSD ↓	MLE
Credit-g	Constant curvature	$2.66_{\pm 0.25}$	$5.70_{\pm 0.35}$	$93.27_{\pm 5.92}$	$0.0308_{\pm 0.0037}$	$0.7362_{\pm 0.0457}$
	Shared-tree diagonal Fisher	$2.53_{\pm 0.22}$	$4.92_{\pm 0.34}$	$93.95_{\pm 6.13}$	$0.0283_{\pm 0.0037}$	$0.7370_{\pm 0.0347}$
	Shared-tree full Fisher	$\mathbf{2.51_{\pm 0.23}}$	$4.77_{\pm 0.34}$	$\mathbf{94.14_{\pm 7.32}}$	$\mathbf{0.0277_{\pm 0.0035}}$	$\mathbf{0.7499_{\pm 0.0386}}$
	Diagonal Fisher (default)	$2.58_{\pm 0.22}$	$\mathbf{4.74_{\pm 0.31}}$	$93.57_{\pm 5.41}$	$0.0290_{\pm 0.0034}$	$0.7419_{\pm 0.0366}$
Lending	Constant curvature	$1.78_{\pm 0.09}$	$1.43_{\pm 0.12}$	$66.63_{\pm 2.64}$	$0.0265_{\pm 0.0013}$	$0.0360_{\pm 0.0031}$
	Shared-tree diagonal Fisher	$1.66_{\pm 0.09}$	$1.35_{\pm 0.16}$	$68.22_{\pm 2.89}$	$0.0192_{\pm 0.0011}$	$\mathbf{0.0353_{\pm 0.0028}}$
	Shared-tree full Fisher	$1.70_{\pm 0.09}$	$1.34_{\pm 0.15}$	$67.64_{\pm 2.65}$	$0.0203_{\pm 0.0014}$	$0.0361_{\pm 0.0035}$
	Diagonal Fisher (default)	$\mathbf{1.61_{\pm 0.09}}$	$\mathbf{1.32_{\pm 0.13}}$	$\mathbf{68.83_{\pm 2.61}}$	$\mathbf{0.0177_{\pm 0.0013}}$	$0.0355_{\pm 0.0035}$
Shoppers	Constant curvature	$1.10_{\pm 0.08}$	$1.84_{\pm 0.20}$	$93.70_{\pm 2.44}$	$0.0262_{\pm 0.0014}$	$0.9144_{\pm 0.0040}$
	Shared-tree diagonal Fisher	$0.96_{\pm 0.09}$	$1.52_{\pm 0.23}$	$94.36_{\pm 2.21}$	$0.0183_{\pm 0.0013}$	$0.9186_{\pm 0.0022}$
	Shared-tree full Fisher	$0.93_{\pm 0.08}$	$1.46_{\pm 0.23}$	$\mathbf{94.38_{\pm 2.15}}$	$\mathbf{0.0167_{\pm 0.0014}}$	$\mathbf{0.9195_{\pm 0.0043}}$
	Diagonal Fisher (default)	$\mathbf{0.90_{\pm 0.08}}$	$\mathbf{1.44_{\pm 0.22}}$	$94.21_{\pm 2.60}$	$0.0178_{\pm 0.0014}$	$\mathbf{0.9195_{\pm 0.0034}}$

E Training and Generation Algorithms

Algorithm 1 preprocesses the mixed-type table, pairs each duplicated record with Gaussian noise, and independently fits continuous and categorical endpoint heads at t_i, $i = 1, \ldots, n_t - 1$. These fits share the same interpolated design matrix within a timestep and can be trained in parallel. Algorithm 2 initializes samples from Gaussian noise and applies Heun's method, evaluating the endpoint heads at the current state and the Euler predictor to recover the velocity through Eq. (6). In VELOCITY, regressors return predicted endpoint means and classifiers return class-probability vectors $\hat{\boldsymbol{\pi}}_{t_i,g}(\boldsymbol{X}) \in \Delta^{|G_g|-1}$, the softmax outputs of their `multi:softprob` ensembles. The final interval uses an Euler update rather than a Heun corrector, since the corrector would evaluate the velocity at $t = 1$, where the denominator of Eq. (6) equals $\sigma_{\min}$.

Algorithm 2: TabTreeVFM Generation

Input: Trained ensembles $\big\{\{\hat{f}^{\text{num}}_{t_i,j}\}_{j\in\mathcal{C}}, \{\hat{f}^{\text{cat}}_{t_i,g}\}_{g=1}^{G}\big\}_{i=1}^{n_t-1}$; Number of generated samples n_{gen}; Terminal noise scale $\sigma_{\min}$; Uniform grid $t_i = (i-1)h$ with $h = 1/(n_t-1)$;

Output: Synthetic table $\hat{\boldsymbol{X}}_1$ with n_{gen} rows, C numerical columns, and G categorical columns;

$\boldsymbol{X} \leftarrow \boldsymbol{X}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_p)^{n_{\text{gen}}}$; ▷ Initialize at $t = 0$

for $i = 1$ **to** $n_t - 1$ **do**

  $t \leftarrow t_i$; $t' \leftarrow t_{i+1}$;

  $\boldsymbol{v}^{(1)} \leftarrow \text{Velocity}(\boldsymbol{X}, t, i)$; ▷ First slope in Heun's method

  $\tilde{\boldsymbol{X}} \leftarrow \boldsymbol{X} + h\boldsymbol{v}^{(1)}$; ▷ Euler predictor step

  if $i = n_t - 1$ **then**

    $\boldsymbol{X} \leftarrow \tilde{\boldsymbol{X}}$; ▷ Final Euler step

  else

    $\boldsymbol{v}^{(2)} \leftarrow \text{Velocity}(\tilde{\boldsymbol{X}}, t', i+1)$; ▷ Corrector slope at $t' < 1$

    $\boldsymbol{X} \leftarrow \boldsymbol{X} + \frac{h}{2}\big(\boldsymbol{v}^{(1)} + \boldsymbol{v}^{(2)}\big)$; ▷ Heun update

$\hat{\boldsymbol{X}}_1^{\text{num}} \leftarrow \text{InverseQuantile}\big(\boldsymbol{X}[:, \mathcal{C}]\big)$;

for $g = 1$ **to** G **do**

  $\hat{\boldsymbol{X}}_1^{\text{cat}}[:, g] \leftarrow \arg\max_k \boldsymbol{X}[:, G_g]_k$;

$\hat{\boldsymbol{X}}_1 \leftarrow [\hat{\boldsymbol{X}}_1^{\text{num}} \mid \hat{\boldsymbol{X}}_1^{\text{cat}}]$;

return $\hat{\boldsymbol{X}}_1$;

Function VELOCITY$(\boldsymbol{X}, t, i)$**:**

  $\boldsymbol{\mu}^{\text{num}} \leftarrow \big[\hat{f}^{\text{num}}_{t_i,j}(\boldsymbol{X})\big]_{j\in\mathcal{C}}$; ▷ Predicted endpoint means

  for $g = 1$ **to** G **do**

    $\boldsymbol{\mu}_g^{\text{cat}} \leftarrow \hat{\boldsymbol{\pi}}_{t_i,g}(\boldsymbol{X}) \in \Delta^{|G_g|-1}$; ▷ Class probabilities of $\hat{f}^{\text{cat}}_{t_i,g}$

  $\boldsymbol{\mu}_t \leftarrow [\boldsymbol{\mu}^{\text{num}} \mid \boldsymbol{\mu}_1^{\text{cat}} \mid \cdots \mid \boldsymbol{\mu}_G^{\text{cat}}]$;

  $\boldsymbol{v}_t \leftarrow \dfrac{\boldsymbol{\mu}_t - (1-\sigma_{\min})\boldsymbol{X}}{1-(1-\sigma_{\min})t}$; ▷ OT-path velocity in Eq. (6)

  return $\boldsymbol{v}_t$;

F Dataset Details

Table 6: Statistics of the 10 benchmark datasets. Feature and category counts include the target; Ratio is the minority-to-majority class ratio.

Domain	Dataset	Task	# Rows			# Features		# Categories			Ratio
			Total	Train	Test	Num	Cat	Max	Min	Mean	
Business	Churn (Keramati et al., 2014)	Binary cls.	3,150	2,835	315	9	5	5	2	3	0.16 : 0.84
	Shoppers (Sakar & Kastro, 2018)	Binary cls.	12,330	11,097	1,233	10	8	20	2	8	0.15 : 0.85
Finance	Credit-g (Hofmann, 1994)	Binary cls.	1,000	900	100	7	14	10	2	4	0.30 : 0.70
	Default (Yeh & Lien, 2009)	Binary cls.	30,000	27,000	3,000	14	10	11	2	8	0.22 : 0.78
	Lending (Lending Club, 2018)	Regression	9,182	8,263	919	34	9	50	2	9	—
Healthcare	NMES (Deb & Trivedi, 1997)	Regression	4,406	3,965	441	11	8	4	2	2	—
Science	Beijing (Chen, 2015)	Regression	41,757	37,581	4,176	7	5	31	4	15	—
	Magic (Bock, 2004)	Binary cls.	19,019	17,117	1,902	10	1	2	2	2	0.35 : 0.65
Social	Adult (Becker & Kohavi, 1996)	Binary cls.	48,842	32,561	16,281	6	9	42	2	12	0.24 : 0.76
	News (Fernandes et al., 2015)	Regression	39,644	35,679	3,965	46	2	7	6	6	—

G Implementation Details

Experimental protocol. The standard benchmark uses the official Adult train–test split and random 90/10 splits for the remaining datasets. DCR uses a separate 50/50 split of the full dataset, with every generator retrained only on the fitting half, as described in Appendix H. Following prior work (Zhang et al., 2024; Shi et al., 2025; Guzmán-Cordero et al., 2025), we generate synthetic tables from each fitted generator with 20 different seeds, each matching the training-set size for the corresponding experiment. Experiments are conducted on a server with a 64-core AMD EPYC 9575F processor at 3.30 GHz, 128 GB of RAM, and eight NVIDIA RTX PRO 6000 GPUs.

G.1 TabTreeVFM Implementation

Numerical columns are quantile-transformed to standard Gaussian marginals using a transformation fitted on the training split and inverted after generation. The default configuration uses $n_t = 50$ time-grid points, duplication factor $K = 100$, and residual noise scale $\sigma_{\min} = 10^{-4}$. Each XGBoost head uses 100 boosting rounds, learning rate 0.3, histogram-based tree construction (`tree_method=hist`), subsample 1.0, and ℓ_2 regularization $\lambda = 10$. We select the maximum tree depth from $\{3, 7\}$ on a validation split: we hold out 20% of the training split, train TabTreeVFM with each depth on the remaining 80%, and choose the depth with better marginal and pairwise fidelity between synthetic data and the held-out rows. The final model is retrained with the selected depth on the full training split. This selects depth 3 on Credit-g, NMES, and Lending and 7 on the remaining datasets. Sampling uses a Heun solver on the 50-point grid, with an Euler update on the final interval. Independent heads are trained in parallel across timesteps and features.

G.2 Baseline Implementations

We use official or established benchmark implementations for all baselines. Unless otherwise stated, we retain the architectures and training hyperparameters provided by the corresponding implementation. The implementation sources, key settings, and sampling adjustments are summarized below.

CTGAN and TVAE. CTGAN (Xu et al., 2019) and TVAE (Xu et al., 2019) use the SDV implementations[1] with the configurations provided by the TabSyn benchmark.

STaSy. STaSy (Kim et al., 2023) uses the official implementation[2] with the hyperparameters and network configuration adopted by the TabSyn benchmark.

CoDi. CoDi (Lee et al., 2023) uses the official implementation[3] with the authors' default training configuration.

TabDDPM. TabDDPM (Kotelnikov et al., 2023) uses the official implementation[4] with the recommended MLP architecture and sampling configuration. Its outputs are degenerate on several datasets, as documented in Appendix I.

TabSyn. TabSyn (Zhang et al., 2024) uses the official implementation[5] with its default two-stage VAE and diffusion training configuration. Sampling uses the EDM sampler with 50 discretization steps and second-order correction.

TabDiff. TabDiff (Shi et al., 2025) uses the official implementation[6], trained for 8,000 epochs with a batch size of 4,096. Sampling uses 50 discretization steps with second-order correction.

CDTD. CDTD (Mueller et al., 2025) uses the official implementation[7] with adaptive per-type noise schedules and an MLP backbone. Training runs for 30,000 iterations, followed by deterministic 100-step Euler sampling.

[1] `https://github.com/sdv-dev/SDV`
[2] `https://github.com/JayoungKim408/STaSy`
[3] `https://github.com/ChaejeongLee/CoDi`
[4] `https://github.com/yandex-research/tab-ddpm`
[5] `https://github.com/amazon-science/tabsyn`
[6] `https://github.com/MinkaiXu/TabDiff`
[7] `https://github.com/muellermarkus/cdtd`

TabbyFlow. TabbyFlow (Guzmán-Cordero et al., 2025) uses the authors' implementation[8] with a tokenized Transformer backbone and default training settings. Sampling uses a 50-step Heun solver.

TabVVFM. TabVVFM (Nasution et al., 2026) uses the official implementation[9] with its default MLP architecture and training hyperparameters. Sampling uses a fixed-step Euler solver with 100 steps.

Forest-Flow. Forest-Flow (Jolicoeur-Martineau et al., 2024) uses the scalable implementation of Cresswell & Kim (2024)[10], retaining its recommended XGBoost configuration. Sampling uses a 50-step Heun solver.

H Evaluation Details

Our evaluation follows established tabular synthesis benchmarks (Zhang et al., 2024; Shi et al., 2025; Guzmán-Cordero et al., 2025; Nasution et al., 2026; Mueller et al., 2025). Let $\mathcal{R}$, $\mathcal{T}$, and $\mathcal{S}$ denote the real training set, real test set, and a synthetic table, respectively, and let $\mathcal{J}_{\text{num}}$ and $\mathcal{J}_{\text{cat}}$ index numerical and categorical columns. Marginal and dependency metrics compare $\mathcal{S}$ with $\mathcal{R}$; C2ST and downstream prediction use the held-out set $\mathcal{T}$. The fitting and holdout sets used for DCR and Mem are defined below.

H.1 Shape and Trend

Shape and Trend measure agreement in individual columns and column pairs, respectively, using SDMetrics' Column Shapes and Column Pair Trends properties.[11] We report their complementary error rates, so lower values indicate better agreement.

Shape. For numerical column j, let F_j^R and F_j^S be its empirical cumulative distribution functions in $\mathcal{R}$ and $\mathcal{S}$. For categorical column j, let p_j^R and p_j^S be empirical probability mass functions on the union Ω_j of observed categories, assigning zero mass to absent categories. The column error is the Kolmogorov–Smirnov distance or total variation distance:

$$e_j = \begin{cases} \sup_x \left| F_j^R(x) - F_j^S(x) \right|, & j \in \mathcal{J}_{\text{num}}, \\ \frac{1}{2} \sum_{a \in \Omega_j} \left| p_j^R(a) - p_j^S(a) \right|, & j \in \mathcal{J}_{\text{cat}}. \end{cases} \tag{H.1}$$

We average these errors across columns and report the result as a percentage.

Trend. For two numerical columns, SDMetrics compares their Pearson correlations ρ_{jk}. For categorical pairs, it compares their normalized contingency tables p_{jk}. Mixed numerical–categorical pairs use the same contingency calculation after SDMetrics' default histogram discretization of the numerical column. The pairwise error is

$$e_{jk} = \begin{cases} \frac{1}{2} \left| \rho_{jk}^R - \rho_{jk}^S \right|, & j,k \in \mathcal{J}_{\text{num}}, \\ \frac{1}{2} \sum_{a,b} \left| p_{jk}^R(a,b) - p_{jk}^S(a,b) \right|, & \text{otherwise}. \end{cases} \tag{H.2}$$

We average these errors across valid column pairs and report the result as a percentage. Appendix K visualizes marginal distributions and pairwise dependency errors.

H.2 Wasserstein and Jensen–Shannon Distances

Wasserstein distance (WD). WD averages the first-order Wasserstein distance between corresponding real and synthetic numerical marginals. We apply min–max scaling fitted on the real training data to both tables without clipping synthetic values; a zero column range is replaced by one. Lower WD indicates closer numerical marginals.

[8] https://github.com/andresguzco/ef-vfm

[9] https://github.com/rulnasution/tabular-flow-matching

[10] https://github.com/layer6ai-labs/calo-forest

[11] https://docs.sdv.dev/sdmetrics/data-metrics/quality

Jensen–Shannon distance (JSD). For each categorical column, we compute the Jensen–Shannon distance between empirical real and synthetic category frequencies, aligned over their union of categories. We use the square root of Jensen–Shannon divergence with base-two logarithms and average across categorical columns.[12] JSD lies in $[0, 1]$, and lower values indicate closer categorical marginals.

H.3 α-Precision and β-Recall

Following Alaa et al. (2022), we report α-precision and β-recall, two sample-level metrics assessing the fidelity and coverage of synthetic data: α-precision measures whether synthetic records lie in high-density regions of the real data, and β-recall measures how well synthetic records cover the real data distribution. We compute both with SynthCity 0.2.12, one-hot encoding categorical columns and min–max scaling numerical columns. Both scores are reported as percentages, with higher values indicating better fidelity and coverage, respectively.

H.4 Classifier Two-Sample Test

C2ST evaluates how easily a classifier distinguishes synthetic from real records. We combine $\mathcal{T}$ with an equally sized random subset of $\mathcal{S}$, encode categorical columns using a common one-hot representation, and train an XGBoost discriminator with 300 trees and maximum depth 6 using histogram-based tree construction. We use three-fold stratified cross-validation. This nonlinear discriminator can capture dependencies missed by logistic regression (Zein & Urvoy, 2022; Kindji et al., 2025; Scassola et al., 2026). For comparison, Table 12 repeats the test using SDMetrics' `LogisticDetection`, with its default mean imputation and robust scaling, while keeping all other evaluation settings unchanged. If A_k is the ROC AUC on fold k of M folds, the SDMetrics normalization gives

$$\mathrm{C2ST} = 1 - \frac{1}{M}\sum_{k=1}^{M}\left[2\max\{A_k, 0.5\} - 1\right]. \tag{H.3}$$

The score lies in $[0, 1]$: higher values indicate less distinguishable data under this discriminator, and a score of one corresponds to AUCs at or below chance.

H.5 Machine Learning Efficiency

MLE measures the usefulness of synthetic data for downstream prediction. We train an XGBoost classifier or regressor on each synthetic table and evaluate it on the real test set, applying the same procedure to the real training data as a reference. Each predictor's training table is randomly divided into fitting and validation portions in an 8:1 ratio. We select hyperparameters by validation AUC for classification or validation RMSE for regression, then refit the selected model on the fitting portion before testing. Both real and synthetic predictors use the same grid: 10, 50, or 100 trees; maximum depth 5, 10, or 20; minimum child weight 1 or 10; and split penalty $\gamma \in \{0, 1\}$.

We report test AUC for classification and RMSE for regression, together with the relative percentage decrease in AUC or increase in RMSE compared with the real-data reference. These gaps are computed from mean test scores and averaged separately across classification and regression datasets. Lower gaps indicate better utility; negative values indicate that the synthetic-trained predictor outperforms the real reference.

H.6 Distance to Closest Record

DCR compares the proximity of synthetic records to fitting and holdout references (Zhang et al., 2024; Shi et al., 2025). For this experiment, each full dataset is split 50/50 into $\mathcal{D}_{\mathrm{fit}}$ and $\mathcal{D}_{\mathrm{hold}}$, and each generator is trained on $\mathcal{D}_{\mathrm{fit}}$ to produce synthetic tables of the same size; the holdout records are reserved for evaluation, and the standard evaluator randomly subsamples the larger reference set to match the smaller one for the nearest-neighbor comparison.

[12] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.jensenshannon.html`

Each row is encoded by concatenating its range-scaled numerical values and one-hot categorical values, with numerical ranges computed on $\mathcal{D}_{\text{fit}}$ and the categorical vocabulary shared across the two references. For a synthetic record $\boldsymbol{x}$, let $d(\boldsymbol{x}, \mathcal{D}) = \min_{\boldsymbol{z}\in\mathcal{D}} \|\boldsymbol{x} - \boldsymbol{z}\|_1$ denote its Manhattan distance, under this encoding, to the nearest row of $\mathcal{D}$. The standard evaluator reports the share of synthetic records closer to fitting than to holdout,

$$s_{\text{DCR}} = \frac{1}{|\mathcal{S}|} \sum_{\boldsymbol{x}\in\mathcal{S}} \mathbf{1}[d(\boldsymbol{x}, \mathcal{D}_{\text{fit}}) < d(\boldsymbol{x}, \mathcal{D}_{\text{hold}})] \,, \tag{H.4}$$

with ties contributing zero. We report $100 s_{\text{DCR}}$; values near 50% indicate balanced proximity to the two references, and the cross-dataset summary averages $|100 s_{\text{DCR}} - 50|$ in percentage points.

H.7 Membership Inference Attack

Membership inference tests whether a record's participation in generator training can be inferred from the released synthetic data (Shokri et al., 2017; van Breugel et al., 2023). We use a SynthEval-style classifier attack (Lautrup et al., 2025). The attacker observes synthetic records and an auxiliary set of real nonmembers; it accesses neither generator parameters nor training losses.

Using the standard benchmark split, each attack repetition randomly partitions $\mathcal{T}$ into auxiliary training and evaluation portions in a 3:1 ratio. A LightGBM classifier with default hyperparameters is trained to distinguish an equally sized subset of synthetic records from the auxiliary training portion, then evaluated on a balanced sample of real training records from $\mathcal{R}$, treated as members, and the unused portion of $\mathcal{T}$, treated as nonmembers. Categorical columns share a one-hot encoding with one category dropped per column. We repeat the attack five times with independently seeded splits and average the resulting ROC AUC, $\overline{A}$, then report

$$s_{\text{MIA}} = 1 - \left[2 \max\{\overline{A}, 0.5\} - 1\right] . \tag{H.5}$$

Higher scores indicate less membership information detected by this attack: one corresponds to mean AUC at or below chance, and zero to perfect membership discrimination.

H.8 Held-out-Calibrated Memorization

Mem measures excess full-row reproduction using the criterion in Section 3.6. We use the fitting and holdout partitions of the DCR experiments. At tolerance $\tau = 0.01$, Equation (14) counts a query as a replica if an existing fitting row matches every categorical value and differs by at most 1% of the column range on every numerical value. Constant numerical columns require exact equality. Mem@1% is the positive part of the mean signed reproduction gap, expressed in percentage points. Table 18 reports the signed means, with standard deviations of the per-release gaps in subscripts. Section 4.3 additionally reports replica-rate curves over a range of tolerances, and Table 19 reports Avg Mem and ranks at five tolerances.

Reporting conventions. Shape and Trend are percentage errors; α-precision, β-recall, and DCR are displayed as percentages in the detailed tables, while WD, JSD, C2ST, and MIA retain their native scales. Signed reproduction gaps and Mem are expressed in percentage points. Within-dataset ranks follow the directions specified above, with DCR ranked by absolute deviation from 50%; the overall rank is the unweighted mean of the metric-wise average ranks. Averaging the metric ranks within each group first and then across the three groups gives 3.80 for TabTreeVFM, 4.85 for TabDiff, and 5.16 for TabSyn. Unavailable results are marked by a dash and excluded from the corresponding averages.

I Detailed Per-Dataset Results

Figure 4 shows, for every metric of Table 1, the per-dataset means of all 12 generators; red crosses mark method–dataset cases for which no valid synthetic output was available. TabDDPM's official implementation is known to collapse on some datasets (Zhang et al., 2024; Shi et al., 2025; Mueller et al., 2025; 2026); in our runs, its News tables contain only two distinct rows, which accounts for its Shape error and zero α-precision and β-recall on that dataset, and its Beijing and NMES numerical columns collapse onto the boundaries of the training range.

Tables 7–18 provide the corresponding numerical results, with standard deviations in subscripts where available. Following TabDiff and EF-VFM (Shi et al., 2025; Guzmán-Cordero et al., 2025), the tables report Shape and Trend error rates, α-precision, β-recall, and DCR shares as percentages (means and standard deviations multiplied by 100). Signed reproduction gaps and Mem@1% are reported in percentage points, also multiplying means and standard deviations by 100. The remaining table metrics retain their native scales, including C2ST on $[0, 1]$. Figure 4 uses native scales throughout. A dash denotes an unavailable result. Avg is the mean over available datasets, except for DCR, where it is the mean absolute deviation from 50%, expressed in percentage points. Table 18 uses Avg Mem as defined below. In Tables 7–17, bold and underlined entries mark the best and second-best generator values at the displayed precision, respectively. Table 12 repeats C2ST with a logistic-regression discriminator under the protocol of Table 11 (Appendix H.4). TabTreeVFM again has the highest average score, 0.9752 compared with 0.9572 for TabDiff and 0.9379 for CDTD, and is best or second best on 7 of the 10 datasets.

Table 18 reports, for each dataset–method cell, the mean of the signed reproduction gap Δ_τ (Eq. (15)) at $\tau = 1\%$ over the 20 released sample sets, with the per-release standard deviation in subscript. Positive values indicate reproduction above the held-out rate, and negative values indicate reproduction below it. The first row gives the held-out reproduction rate in percent. The final column, Avg Mem, averages the positive parts of the dataset-wise mean gaps over available datasets. Entries are rounded to two decimal places.

Table 19 repeats the Mem comparison at tolerances from 0.1% to 5%, reporting Avg Mem and the average within-dataset rank used in Table 1. CDTD and TabbyFlow have the highest Avg Mem at every tolerance, and TabTreeVFM stays between 0.14 and 0.28 percentage points, below TabDiff and TabSyn throughout. The ranks of the other generators vary more, with Kendall rank correlations between 0.61 and 0.92 relative to the ranking at $\tau = 1\%$.

The largest shifts come from TVAE and CoDi, whose Avg Mem rises from below 0.3 at $\tau = 0.1\%$ to above 7 percentage points at $\tau = 5\%$ as the wider tolerance also counts near-duplicates of training rows.



Figure 4: Per-dataset performance of 12 generators across 10 datasets.

Table 7: Per-dataset Shape error rates (%).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$1.03_{\pm 0.06}$	$1.17_{\pm 0.05}$	$\mathbf{1.48_{\pm 0.20}}$	$\mathbf{2.20_{\pm 0.24}}$	$1.68_{\pm 0.08}$	$4.13_{\pm 0.09}$	$1.09_{\pm 0.10}$	$2.20_{\pm 0.03}$	$2.16_{\pm 0.12}$	$2.24_{\pm 0.09}$	1.94
CoDi	$24.08_{\pm 0.08}$	$17.87_{\pm 0.10}$	$21.86_{\pm 0.40}$	$22.51_{\pm 0.27}$	$23.97_{\pm 0.12}$	$39.19_{\pm 0.12}$	$9.63_{\pm 0.17}$	$26.32_{\pm 0.03}$	$28.44_{\pm 0.19}$	$33.91_{\pm 0.07}$	24.78
CTGAN	$16.55_{\pm 0.08}$	$19.39_{\pm 0.05}$	$7.54_{\pm 0.18}$	$16.52_{\pm 0.28}$	$15.53_{\pm 0.08}$	$7.99_{\pm 0.07}$	$10.41_{\pm 0.08}$	$17.23_{\pm 0.04}$	$9.99_{\pm 0.13}$	$24.50_{\pm 0.09}$	14.56
TabbyFlow	$1.29_{\pm 0.07}$	$1.47_{\pm 0.07}$	$2.43_{\pm 0.21}$	$3.40_{\pm 0.31}$	$2.25_{\pm 0.11}$	$3.04_{\pm 0.07}$	$1.64_{\pm 0.11}$	$3.71_{\pm 0.04}$	$2.43_{\pm 0.13}$	$2.07_{\pm 0.10}$	2.37
Forest-Flow	$9.07_{\pm 0.05}$	$4.73_{\pm 0.04}$	$3.43_{\pm 0.24}$	$12.99_{\pm 0.24}$	$7.31_{\pm 0.08}$	$4.05_{\pm 0.10}$	$1.13_{\pm 0.07}$	$16.03_{\pm 0.04}$	$5.64_{\pm 0.13}$	$13.42_{\pm 0.09}$	7.78
STaSy	$11.59_{\pm 0.06}$	$9.40_{\pm 0.08}$	$20.21_{\pm 0.25}$	$40.67_{\pm 0.31}$	$9.79_{\pm 0.14}$	$14.23_{\pm 4.75}$	$28.14_{\pm 0.19}$	$13.85_{\pm 0.06}$	$13.27_{\pm 0.23}$	$14.86_{\pm 0.09}$	17.60
TVAE	$13.46_{\pm 0.08}$	$16.21_{\pm 0.08}$	$6.28_{\pm 0.29}$	$10.85_{\pm 0.23}$	$14.65_{\pm 0.11}$	$11.36_{\pm 0.11}$	$6.11_{\pm 0.16}$	$16.32_{\pm 0.03}$	$4.84_{\pm 0.18}$	$18.72_{\pm 0.15}$	11.88
TabDDPM	$1.18_{\pm 0.06}$	$33.19_{\pm 0.09}$	$15.10_{\pm 0.58}$	$15.95_{\pm 0.28}$	$1.39_{\pm 0.10}$	–	$1.10_{\pm 0.13}$	$83.08_{\pm 0.07}$	$45.38_{\pm 0.14}$	$2.34_{\pm 0.10}$	22.08
TabDiff	$0.66_{\pm 0.05}$	$1.03_{\pm 0.04}$	$1.94_{\pm 0.17}$	$2.75_{\pm 0.27}$	$\mathbf{1.19_{\pm 0.09}}$	$4.47_{\pm 0.07}$	$\mathbf{0.79_{\pm 0.08}}$	$2.44_{\pm 0.05}$	$1.55_{\pm 0.10}$	$1.57_{\pm 0.11}$	1.84
TabSyn	$0.98_{\pm 0.08}$	$4.75_{\pm 0.10}$	$2.18_{\pm 0.13}$	$5.01_{\pm 0.17}$	$1.25_{\pm 0.05}$	$4.95_{\pm 0.09}$	$1.59_{\pm 0.15}$	$1.95_{\pm 0.04}$	$1.87_{\pm 0.09}$	$1.55_{\pm 0.10}$	2.61
TabVVFM	$1.53_{\pm 0.06}$	$2.09_{\pm 0.05}$	$4.36_{\pm 0.21}$	$5.54_{\pm 0.21}$	$2.27_{\pm 0.12}$	$21.72_{\pm 0.09}$	$2.39_{\pm 0.13}$	$5.52_{\pm 0.08}$	$6.44_{\pm 0.10}$	$4.11_{\pm 0.10}$	5.60
TabTreeVFM	$\mathbf{0.63_{\pm 0.06}}$	$\mathbf{0.88_{\pm 0.05}}$	$1.75_{\pm 0.15}$	$2.56_{\pm 0.17}$	$1.27_{\pm 0.06}$	$\mathbf{1.63_{\pm 0.09}}$	$0.80_{\pm 0.10}$	$\mathbf{1.62_{\pm 0.03}}$	$\mathbf{0.98_{\pm 0.11}}$	$\mathbf{0.95_{\pm 0.09}}$	**1.31**

Table 8: Per-dataset Trend error rates (%).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$2.11_{\pm 0.09}$	$3.42_{\pm 0.40}$	$\mathbf{2.02_{\pm 0.25}}$	$\mathbf{3.87_{\pm 0.31}}$	$1.64_{\pm 0.19}$	$4.17_{\pm 0.20}$	$1.64_{\pm 0.53}$	$5.58_{\pm 0.30}$	$5.21_{\pm 1.16}$	$2.09_{\pm 0.29}$	3.17
CoDi	$31.22_{\pm 0.09}$	$7.36_{\pm 0.33}$	$19.15_{\pm 0.35}$	$24.67_{\pm 1.35}$	$14.54_{\pm 0.12}$	$9.14_{\pm 0.34}$	$5.81_{\pm 0.28}$	$11.48_{\pm 0.28}$	$21.31_{\pm 0.50}$	$12.34_{\pm 0.14}$	15.70
CTGAN	$22.13_{\pm 0.13}$	$46.93_{\pm 0.11}$	$11.02_{\pm 0.59}$	$20.72_{\pm 0.50}$	$16.46_{\pm 0.10}$	$8.88_{\pm 0.09}$	$11.25_{\pm 0.28}$	$5.90_{\pm 0.08}$	$20.55_{\pm 2.82}$	$20.48_{\pm 0.19}$	18.43
TabbyFlow	$2.39_{\pm 0.11}$	$3.51_{\pm 0.48}$	$3.26_{\pm 0.30}$	$4.85_{\pm 0.50}$	$2.01_{\pm 0.12}$	$2.60_{\pm 0.34}$	$1.11_{\pm 0.43}$	$\mathbf{5.20_{\pm 1.20}}$	$6.33_{\pm 0.84}$	$2.12_{\pm 0.21}$	3.34
Forest-Flow	$5.07_{\pm 0.09}$	$4.17_{\pm 0.09}$	$4.40_{\pm 0.24}$	$16.87_{\pm 1.75}$	$4.94_{\pm 0.15}$	$2.37_{\pm 0.46}$	$\mathbf{0.67_{\pm 0.27}}$	$7.36_{\pm 0.44}$	$3.83_{\pm 0.98}$	$1.92_{\pm 0.24}$	5.16
STaSy	$19.60_{\pm 0.07}$	$9.66_{\pm 0.19}$	$26.47_{\pm 0.30}$	$43.98_{\pm 0.54}$	$9.15_{\pm 0.15}$	$7.91_{\pm 7.50}$	$9.05_{\pm 0.23}$	$11.39_{\pm 0.41}$	$19.89_{\pm 0.51}$	$8.37_{\pm 0.18}$	16.55
TVAE	$11.28_{\pm 0.16}$	$15.42_{\pm 0.43}$	$14.02_{\pm 1.07}$	$12.40_{\pm 2.22}$	$11.44_{\pm 0.16}$	$9.64_{\pm 0.88}$	$8.67_{\pm 0.99}$	$12.51_{\pm 0.60}$	$13.37_{\pm 2.36}$	$9.92_{\pm 0.28}$	11.87
TabDDPM	$1.78_{\pm 0.12}$	$70.01_{\pm 0.06}$	$17.37_{\pm 0.56}$	$33.83_{\pm 0.45}$	$6.15_{\pm 0.38}$	–	$1.21_{\pm 0.16}$	$82.28_{\pm 0.00}$	$71.63_{\pm 0.30}$	$6.37_{\pm 0.29}$	32.29
TabDiff	$\mathbf{1.40_{\pm 0.06}}$	$3.26_{\pm 0.24}$	$3.35_{\pm 0.21}$	$5.75_{\pm 0.59}$	$\mathbf{1.32_{\pm 0.12}}$	$16.09_{\pm 0.15}$	$0.82_{\pm 0.22}$	$7.59_{\pm 0.25}$	$7.53_{\pm 0.41}$	$1.89_{\pm 0.19}$	4.90
TabSyn	$2.17_{\pm 0.12}$	$4.98_{\pm 0.21}$	$3.77_{\pm 0.17}$	$7.93_{\pm 1.59}$	$1.47_{\pm 0.05}$	$3.02_{\pm 0.21}$	$1.24_{\pm 0.26}$	$6.91_{\pm 0.27}$	$6.44_{\pm 0.60}$	$1.92_{\pm 0.15}$	3.98
TabVVFM	$1.68_{\pm 0.10}$	$5.73_{\pm 0.49}$	$8.49_{\pm 0.31}$	$9.21_{\pm 0.47}$	$5.43_{\pm 0.99}$	$24.93_{\pm 0.11}$	$5.49_{\pm 0.69}$	$12.81_{\pm 0.90}$	$11.06_{\pm 0.28}$	$9.30_{\pm 0.26}$	9.41
TabTreeVFM	$1.73_{\pm 0.08}$	$\mathbf{2.90_{\pm 0.51}}$	$2.49_{\pm 0.30}$	$4.82_{\pm 0.51}$	$2.61_{\pm 0.09}$	$\mathbf{1.57_{\pm 0.31}}$	$1.06_{\pm 0.28}$	$7.46_{\pm 0.82}$	$\mathbf{3.23_{\pm 0.55}}$	$\mathbf{1.51_{\pm 0.23}}$	**2.94**

Table 9: Per-dataset Wasserstein distance (WD).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$0.0042_{\pm 0.0001}$	$0.0035_{\pm 0.0001}$	$\mathbf{0.0055_{\pm 0.0010}}$	$\mathbf{0.0110_{\pm 0.0026}}$	$0.0022_{\pm 0.0002}$	$0.0114_{\pm 0.0003}$	$0.0030_{\pm 0.0004}$	$0.0052_{\pm 0.0001}$	$0.0085_{\pm 0.0005}$	$0.0060_{\pm 0.0003}$	0.0060
CoDi	$0.0233_{\pm 0.0002}$	$0.0147_{\pm 0.0003}$	$0.0947_{\pm 0.0024}$	$0.0666_{\pm 0.0031}$	$0.0466_{\pm 0.0004}$	$0.0657_{\pm 0.0004}$	$0.0293_{\pm 0.0006}$	$0.0401_{\pm 0.0001}$	$0.0332_{\pm 0.0010}$	$0.0465_{\pm 0.0006}$	0.0461
CTGAN	$0.0172_{\pm 0.0001}$	$0.0341_{\pm 0.0002}$	$0.0245_{\pm 0.0008}$	$0.0505_{\pm 0.0027}$	$0.0081_{\pm 0.0001}$	$0.0097_{\pm 0.0001}$	$0.0230_{\pm 0.0002}$	$0.0095_{\pm 0.0001}$	$0.0152_{\pm 0.0004}$	$0.0168_{\pm 0.0002}$	0.0209
TabbyFlow	$0.0049_{\pm 0.0002}$	$0.0044_{\pm 0.0002}$	$0.0111_{\pm 0.0015}$	$0.0168_{\pm 0.0028}$	$0.0035_{\pm 0.0001}$	$0.0070_{\pm 0.0003}$	$0.0058_{\pm 0.0003}$	$0.0091_{\pm 0.0001}$	$0.0080_{\pm 0.0004}$	$0.0065_{\pm 0.0004}$	0.0077
Forest-Flow	$0.0047_{\pm 0.0002}$	$0.0077_{\pm 0.0002}$	$0.0128_{\pm 0.0011}$	$0.0150_{\pm 0.0016}$	$0.0029_{\pm 0.0001}$	$0.0050_{\pm 0.0002}$	$0.0018_{\pm 0.0002}$	$\mathbf{0.0042_{\pm 0.0001}}$	$0.0047_{\pm 0.0002}$	$0.0038_{\pm 0.0004}$	0.0063
STaSy	$0.0611_{\pm 0.0004}$	$0.0171_{\pm 0.0003}$	$0.1413_{\pm 0.0014}$	$0.3193_{\pm 0.0051}$	$0.0087_{\pm 0.0002}$	$0.0526_{\pm 0.0454}$	$0.0939_{\pm 0.0008}$	$0.0383_{\pm 0.0003}$	$0.0225_{\pm 0.0006}$	$0.0340_{\pm 0.0005}$	0.0789
TVAE	$0.0249_{\pm 0.0003}$	$0.0242_{\pm 0.0003}$	$0.0230_{\pm 0.0011}$	$0.0270_{\pm 0.0021}$	$0.0145_{\pm 0.0002}$	$0.0232_{\pm 0.0003}$	$0.0137_{\pm 0.0003}$	$0.0251_{\pm 0.0001}$	$0.0082_{\pm 0.0005}$	$0.0149_{\pm 0.0006}$	0.0199
TabDDPM	$0.0020_{\pm 0.0001}$	$0.4096_{\pm 0.0009}$	$0.0764_{\pm 0.0029}$	$0.3075_{\pm 0.0040}$	$0.0025_{\pm 0.0002}$	–	$0.0021_{\pm 0.0003}$	$0.5491_{\pm 0.0000}$	$0.3145_{\pm 0.0014}$	$0.0093_{\pm 0.0008}$	0.1859
TabDiff	$\mathbf{0.0017_{\pm 0.0002}}$	$0.0022_{\pm 0.0002}$	$0.0074_{\pm 0.0007}$	$0.0165_{\pm 0.0018}$	$\mathbf{0.0013_{\pm 0.0001}}$	$0.0184_{\pm 0.0002}$	$\mathbf{0.0017_{\pm 0.0002}}$	$0.0056_{\pm 0.0001}$	$0.0062_{\pm 0.0004}$	$0.0027_{\pm 0.0003}$	0.0064
TabSyn	$0.0029_{\pm 0.0002}$	$0.0091_{\pm 0.0003}$	$0.0097_{\pm 0.0009}$	$0.0252_{\pm 0.0013}$	$0.0014_{\pm 0.0001}$	$0.0105_{\pm 0.0003}$	$0.0035_{\pm 0.0005}$	$\mathbf{0.0042_{\pm 0.0001}}$	$0.0054_{\pm 0.0004}$	$0.0030_{\pm 0.0002}$	0.0075
TabVVFM	$0.0084_{\pm 0.0003}$	$0.0090_{\pm 0.0003}$	$0.0272_{\pm 0.0010}$	$0.0735_{\pm 0.0020}$	$0.0126_{\pm 0.0019}$	$0.1434_{\pm 0.0008}$	$0.0099_{\pm 0.0004}$	$0.0218_{\pm 0.0008}$	$0.0352_{\pm 0.0007}$	$0.0222_{\pm 0.0004}$	0.0363
TabTreeVFM	$0.0019_{\pm 0.0001}$	$\mathbf{0.0021_{\pm 0.0002}}$	$0.0070_{\pm 0.0009}$	$0.0125_{\pm 0.0026}$	$\mathbf{0.0013_{\pm 0.0002}}$	$\mathbf{0.0032_{\pm 0.0003}}$	$0.0020_{\pm 0.0004}$	$0.0044_{\pm 0.0001}$	$\mathbf{0.0028_{\pm 0.0003}}$	$\mathbf{0.0022_{\pm 0.0005}}$	**0.0039**

Table 10: Per-dataset Jensen–Shannon distance (JSD).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$0.0145_{\pm 0.0007}$	$0.0101_{\pm 0.0008}$	$0.0097_{\pm 0.0030}$	$0.0217_{\pm 0.0026}$	$0.0162_{\pm 0.0013}$	$0.0245_{\pm 0.0017}$	$0.0024_{\pm 0.0018}$	$0.0094_{\pm 0.0009}$	$\mathbf{0.0057_{\pm 0.0018}}$	$0.0155_{\pm 0.0012}$	**0.0130**
CoDi	$0.2006_{\pm 0.0007}$	$0.0603_{\pm 0.0010}$	$0.1611_{\pm 0.0026}$	$0.1777_{\pm 0.0031}$	$0.1592_{\pm 0.0012}$	$0.1258_{\pm 0.0016}$	$0.1373_{\pm 0.0030}$	$0.1025_{\pm 0.0013}$	$0.0994_{\pm 0.0027}$	$0.1914_{\pm 0.0012}$	0.1416
CTGAN	$0.1373_{\pm 0.0006}$	$0.2108_{\pm 0.0007}$	$0.0820_{\pm 0.0026}$	$0.1247_{\pm 0.0029}$	$0.1916_{\pm 0.0010}$	$0.0788_{\pm 0.0014}$	$0.1144_{\pm 0.0029}$	$0.0886_{\pm 0.0009}$	$0.0898_{\pm 0.0018}$	$0.2150_{\pm 0.0015}$	0.1333
TabbyFlow	$0.0190_{\pm 0.0007}$	$0.0178_{\pm 0.0010}$	$0.0164_{\pm 0.0045}$	$0.0357_{\pm 0.0041}$	$0.0221_{\pm 0.0012}$	$0.0577_{\pm 0.0017}$	$0.0036_{\pm 0.0023}$	$0.0513_{\pm 0.0017}$	$0.0163_{\pm 0.0021}$	$0.0252_{\pm 0.0015}$	0.0265
Forest-Flow	$0.0363_{\pm 0.0010}$	$0.0683_{\pm 0.0008}$	$0.0231_{\pm 0.0027}$	$0.0811_{\pm 0.0028}$	$0.0346_{\pm 0.0012}$	$0.0421_{\pm 0.0010}$	$0.0029_{\pm 0.0023}$	$0.0158_{\pm 0.0017}$	$0.0065_{\pm 0.0014}$	$0.0293_{\pm 0.0011}$	0.0340
STaSy	$0.1103_{\pm 0.0007}$	$0.1402_{\pm 0.0010}$	$0.1495_{\pm 0.0039}$	$0.4094_{\pm 0.0036}$	$0.0806_{\pm 0.0010}$	$0.1532_{\pm 0.0488}$	$0.1015_{\pm 0.0035}$	$0.1877_{\pm 0.0019}$	$0.1735_{\pm 0.0021}$	$0.1280_{\pm 0.0013}$	0.1634
TVAE	$0.1180_{\pm 0.0010}$	$0.2573_{\pm 0.0011}$	$0.0435_{\pm 0.0042}$	$0.0568_{\pm 0.0041}$	$0.1608_{\pm 0.0016}$	$0.1650_{\pm 0.0013}$	$0.0413_{\pm 0.0026}$	$0.1229_{\pm 0.0018}$	$0.0418_{\pm 0.0022}$	$0.1427_{\pm 0.0012}$	0.1150
TabDDPM	$0.0223_{\pm 0.0007}$	$0.1089_{\pm 0.0016}$	$0.0734_{\pm 0.0039}$	$\mathbf{0.0216_{\pm 0.0027}}$	$0.0219_{\pm 0.0011}$	–	$0.0157_{\pm 0.0021}$	$0.8040_{\pm 0.0000}$	$0.3156_{\pm 0.0029}$	$0.0314_{\pm 0.0014}$	0.1572
TabDiff	$\mathbf{0.0089_{\pm 0.0006}}$	$0.0146_{\pm 0.0009}$	$0.0118_{\pm 0.0033}$	$0.0220_{\pm 0.0024}$	$0.0135_{\pm 0.0013}$	$\mathbf{0.0141_{\pm 0.0015}}$	$\mathbf{0.0021_{\pm 0.0014}}$	$0.0369_{\pm 0.0013}$	$0.0068_{\pm 0.0020}$	$0.0172_{\pm 0.0013}$	0.0148
TabSyn	$0.0183_{\pm 0.0008}$	$0.0674_{\pm 0.0009}$	$\mathbf{0.0083_{\pm 0.0020}}$	$0.0668_{\pm 0.0033}$	$0.0220_{\pm 0.0008}$	$0.0757_{\pm 0.0012}$	$0.0281_{\pm 0.0029}$	$0.0230_{\pm 0.0019}$	$0.0090_{\pm 0.0015}$	$0.0278_{\pm 0.0016}$	0.0346
TabVVFM	$0.0095_{\pm 0.0009}$	$0.0098_{\pm 0.0007}$	$0.0096_{\pm 0.0028}$	$0.0279_{\pm 0.0031}$	$\mathbf{0.0100_{\pm 0.0010}}$	$0.2504_{\pm 0.0015}$	$0.0023_{\pm 0.0018}$	$\mathbf{0.0060_{\pm 0.0012}}$	$\mathbf{0.0057_{\pm 0.0011}}$	$\mathbf{0.0131_{\pm 0.0016}}$	0.0344
TabTreeVFM	$0.0126_{\pm 0.0007}$	$\mathbf{0.0088_{\pm 0.0009}}$	$0.0130_{\pm 0.0028}$	$0.0297_{\pm 0.0031}$	$0.0179_{\pm 0.0011}$	$0.0205_{\pm 0.0013}$	$0.0034_{\pm 0.0023}$	$0.0145_{\pm 0.0016}$	$0.0069_{\pm 0.0016}$	$0.0185_{\pm 0.0013}$	0.0146

Table 11: Per-dataset C2ST, using an XGBoost classifier.

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$0.7215_{\pm 0.0059}$	$0.8537_{\pm 0.0160}$	$\mathbf{0.9288_{\pm 0.0487}}$	$0.9386_{\pm 0.0453}$	$0.7181_{\pm 0.0146}$	$0.5122_{\pm 0.0251}$	$0.8921_{\pm 0.0236}$	$0.4004_{\pm 0.0128}$	$0.8198_{\pm 0.0671}$	$0.7641_{\pm 0.0271}$	0.7549
CoDi	$0.0001_{\pm 0.0001}$	$0.0053_{\pm 0.0008}$	$0.0026_{\pm 0.0021}$	$0.0199_{\pm 0.0113}$	$0.0142_{\pm 0.0012}$	$0.0001_{\pm 0.0000}$	$0.6621_{\pm 0.0147}$	$0.0000_{\pm 0.0000}$	$0.0007_{\pm 0.0005}$	$0.0023_{\pm 0.0005}$	0.0707
CTGAN	$0.0014_{\pm 0.0002}$	$0.0022_{\pm 0.0005}$	$0.4048_{\pm 0.0446}$	$0.0037_{\pm 0.0046}$	$0.0025_{\pm 0.0004}$	$0.0067_{\pm 0.0018}$	$0.0605_{\pm 0.0045}$	$0.0000_{\pm 0.0000}$	$0.0338_{\pm 0.0058}$	$0.0001_{\pm 0.0001}$	0.0516
TabbyFlow	$0.7598_{\pm 0.0043}$	$0.8794_{\pm 0.0145}$	$0.8451_{\pm 0.0486}$	$\mathbf{0.9393_{\pm 0.0631}}$	$0.7326_{\pm 0.0145}$	$\mathbf{0.7361_{\pm 0.0426}}$	$0.9200_{\pm 0.0167}$	$0.4130_{\pm 0.0171}$	$0.8639_{\pm 0.0439}$	$0.8370_{\pm 0.0297}$	0.7926
Forest-Flow	$0.0905_{\pm 0.0025}$	$0.0149_{\pm 0.0015}$	$0.7082_{\pm 0.0599}$	$0.0501_{\pm 0.0184}$	$0.0217_{\pm 0.0017}$	$0.1614_{\pm 0.0146}$	$0.8754_{\pm 0.0220}$	$0.0000_{\pm 0.0001}$	$0.1092_{\pm 0.0195}$	$0.0481_{\pm 0.0056}$	0.2080
STaSy	$0.2509_{\pm 0.0033}$	$0.4985_{\pm 0.0136}$	$0.1294_{\pm 0.0175}$	$0.0027_{\pm 0.0022}$	$0.5115_{\pm 0.0112}$	$0.1410_{\pm 0.0343}$	$0.4198_{\pm 0.0174}$	$0.2128_{\pm 0.0047}$	$0.2883_{\pm 0.0221}$	$0.2372_{\pm 0.0140}$	0.2692
TVAE	$0.0181_{\pm 0.0009}$	$0.0015_{\pm 0.0002}$	$0.5463_{\pm 0.0375}$	$0.0053_{\pm 0.0040}$	$0.0089_{\pm 0.0010}$	$0.0029_{\pm 0.0011}$	$0.4102_{\pm 0.0159}$	$0.0000_{\pm 0.0000}$	$0.0705_{\pm 0.0112}$	$0.0001_{\pm 0.0001}$	0.1064
TabDDPM	$0.8131_{\pm 0.0084}$	$0.0000_{\pm 0.0000}$	$0.5824_{\pm 0.0507}$	$0.0006_{\pm 0.0026}$	$0.7037_{\pm 0.0174}$	–	$0.8925_{\pm 0.0275}$	$0.0000_{\pm 0.0000}$	$0.0016_{\pm 0.0027}$	$0.7993_{\pm 0.0301}$	0.4215
TabDiff	$\mathbf{0.8189_{\pm 0.0075}}$	$\mathbf{0.8913_{\pm 0.0123}}$	$0.8314_{\pm 0.0506}$	$0.9165_{\pm 0.0922}$	$\mathbf{0.8318_{\pm 0.0162}}$	$0.0117_{\pm 0.0023}$	$0.9051_{\pm 0.0259}$	$0.2473_{\pm 0.0124}$	$0.8738_{\pm 0.0464}$	$0.9161_{\pm 0.0261}$	0.7244
TabSyn	$0.7722_{\pm 0.0084}$	$0.6777_{\pm 0.0108}$	$0.7321_{\pm 0.0554}$	$0.7824_{\pm 0.0774}$	$0.6756_{\pm 0.0148}$	$0.3864_{\pm 0.0187}$	$0.8986_{\pm 0.0181}$	$0.5120_{\pm 0.0167}$	$\mathbf{0.8879_{\pm 0.0322}}$	$0.8628_{\pm 0.0308}$	0.7188
TabVVFM	$0.6561_{\pm 0.0090}$	$0.6141_{\pm 0.0149}$	$0.5363_{\pm 0.0458}$	$0.8249_{\pm 0.0629}$	$0.2576_{\pm 0.0120}$	$0.0005_{\pm 0.0002}$	$0.6503_{\pm 0.0247}$	$0.0210_{\pm 0.0021}$	$0.4686_{\pm 0.0353}$	$0.3838_{\pm 0.0200}$	0.4413
TabTreeVFM	$0.8081_{\pm 0.0100}$	$0.8677_{\pm 0.0095}$	$0.8831_{\pm 0.0546}$	$0.9374_{\pm 0.0671}$	$0.8261_{\pm 0.0187}$	$0.6739_{\pm 0.0286}$	$\mathbf{0.9407_{\pm 0.0180}}$	$\mathbf{0.7842_{\pm 0.0122}}$	$0.8574_{\pm 0.0392}$	$\mathbf{0.9474_{\pm 0.0217}}$	**0.8526**

Table 12: Per-dataset C2ST, using a logistic-regression classifier.

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$0.9630_{\pm 0.0066}$	$0.9704_{\pm 0.0143}$	$\mathbf{0.9760_{\pm 0.0233}}$	$0.9513_{\pm 0.0449}$	$0.9553_{\pm 0.0138}$	$0.7392_{\pm 0.0280}$	$0.9860_{\pm 0.0113}$	$0.9259_{\pm 0.0175}$	$0.9607_{\pm 0.0361}$	$0.9513_{\pm 0.0256}$	0.9379
CoDi	$0.1717_{\pm 0.0035}$	$0.7135_{\pm 0.0187}$	$0.3454_{\pm 0.0249}$	$0.2888_{\pm 0.0477}$	$0.2469_{\pm 0.0066}$	$0.0314_{\pm 0.0030}$	$0.7678_{\pm 0.0135}$	$0.1478_{\pm 0.0054}$	$0.5397_{\pm 0.0323}$	$0.1832_{\pm 0.0057}$	0.3436
CTGAN	$0.4363_{\pm 0.0054}$	$0.1726_{\pm 0.0041}$	$0.7803_{\pm 0.0461}$	$0.4305_{\pm 0.0645}$	$0.4383_{\pm 0.0099}$	$0.5958_{\pm 0.0161}$	$0.6689_{\pm 0.0136}$	$0.5818_{\pm 0.0214}$	$0.4813_{\pm 0.0227}$	$0.2944_{\pm 0.0108}$	0.4880
TabbyFlow	$0.9540_{\pm 0.0055}$	$0.9521_{\pm 0.0176}$	$0.9642_{\pm 0.0311}$	$0.9515_{\pm 0.0393}$	$0.9381_{\pm 0.0144}$	$0.7979_{\pm 0.0308}$	$0.9772_{\pm 0.0163}$	$0.8819_{\pm 0.0113}$	$\mathbf{0.9825_{\pm 0.0167}}$	$0.9051_{\pm 0.0225}$	0.9305
Forest-Flow	$0.8675_{\pm 0.0058}$	$0.7932_{\pm 0.0116}$	$0.9694_{\pm 0.0341}$	$0.3736_{\pm 0.0577}$	$0.9315_{\pm 0.0189}$	$0.8019_{\pm 0.0322}$	$\mathbf{0.9868_{\pm 0.0121}}$	$0.1144_{\pm 0.0069}$	$0.8706_{\pm 0.0521}$	$0.9443_{\pm 0.0304}$	0.7653
STaSy	$0.3308_{\pm 0.0067}$	$0.5215_{\pm 0.0099}$	$0.1293_{\pm 0.0081}$	$0.0003_{\pm 0.0005}$	$0.6266_{\pm 0.0095}$	$0.1976_{\pm 0.0490}$	$0.3883_{\pm 0.0115}$	$0.2928_{\pm 0.0071}$	$0.2661_{\pm 0.0172}$	$0.2700_{\pm 0.0109}$	0.3023
TVAE	$0.4970_{\pm 0.0079}$	$0.2141_{\pm 0.0043}$	$0.9327_{\pm 0.0448}$	$0.7998_{\pm 0.0955}$	$0.3778_{\pm 0.0058}$	$0.2614_{\pm 0.0145}$	$0.8791_{\pm 0.0176}$	$0.3892_{\pm 0.0148}$	$0.8012_{\pm 0.0429}$	$0.5071_{\pm 0.0186}$	0.5659
TabDDPM	$0.9407_{\pm 0.0050}$	$0.0726_{\pm 0.0047}$	$0.7172_{\pm 0.0342}$	$0.2978_{\pm 0.0691}$	$\mathbf{0.9801_{\pm 0.0173}}$	–	$0.9851_{\pm 0.0130}$	$0.0000_{\pm 0.0000}$	$0.0071_{\pm 0.0050}$	$0.9218_{\pm 0.0295}$	0.5469
TabDiff	$\mathbf{0.9886_{\pm 0.0077}}$	$0.9670_{\pm 0.0134}$	$0.9757_{\pm 0.0294}$	$0.9522_{\pm 0.0477}$	$0.9794_{\pm 0.0136}$	$0.8378_{\pm 0.0244}$	$0.9846_{\pm 0.0111}$	$0.9384_{\pm 0.0196}$	$0.9737_{\pm 0.0271}$	$0.9749_{\pm 0.0236}$	0.9572
TabSyn	$0.9804_{\pm 0.0084}$	$0.7875_{\pm 0.0131}$	$0.9630_{\pm 0.0300}$	$0.8730_{\pm 0.0910}$	$0.9756_{\pm 0.0147}$	$0.6613_{\pm 0.0230}$	$0.9755_{\pm 0.0138}$	$\mathbf{0.9697_{\pm 0.0125}}$	$0.9748_{\pm 0.0331}$	$0.9443_{\pm 0.0331}$	0.9105
TabVVFM	$0.9537_{\pm 0.0068}$	$0.9486_{\pm 0.0179}$	$0.9569_{\pm 0.0409}$	$0.8786_{\pm 0.0672}$	$0.9473_{\pm 0.0251}$	$0.0907_{\pm 0.0159}$	$0.9748_{\pm 0.0184}$	$0.8129_{\pm 0.0135}$	$0.9362_{\pm 0.0421}$	$0.9501_{\pm 0.0239}$	0.8450
TabTreeVFM	$0.9853_{\pm 0.0068}$	$\mathbf{0.9798_{\pm 0.0151}}$	$0.9748_{\pm 0.0254}$	$\mathbf{0.9711_{\pm 0.0317}}$	$0.9762_{\pm 0.0180}$	$\mathbf{0.9626_{\pm 0.0315}}$	$0.9822_{\pm 0.0171}$	$0.9640_{\pm 0.0128}$	$0.9802_{\pm 0.0264}$	$\mathbf{0.9754_{\pm 0.0225}}$	**0.9752**

Table 13: Per-dataset α-precision (%).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$\mathbf{99.28_{\pm 0.23}}$	$98.44_{\pm 0.33}$	$\mathbf{98.70_{\pm 0.42}}$	$\mathbf{97.99_{\pm 1.01}}$	$\mathbf{99.31_{\pm 0.30}}$	$89.62_{\pm 0.74}$	$98.03_{\pm 0.34}$	$93.85_{\pm 0.33}$	$97.35_{\pm 0.68}$	$\mathbf{99.37_{\pm 0.31}}$	**97.19**
CoDi	$79.26_{\pm 0.35}$	$97.46_{\pm 0.21}$	$80.89_{\pm 0.68}$	$92.04_{\pm 1.89}$	$84.06_{\pm 0.11}$	$58.12_{\pm 0.46}$	$86.33_{\pm 0.35}$	$91.67_{\pm 0.22}$	$64.71_{\pm 0.71}$	$89.61_{\pm 0.55}$	82.41
CTGAN	$81.48_{\pm 0.23}$	$90.50_{\pm 0.15}$	$79.74_{\pm 0.87}$	$80.99_{\pm 1.42}$	$67.80_{\pm 0.22}$	$85.17_{\pm 0.49}$	$85.92_{\pm 0.31}$	$\mathbf{97.92_{\pm 0.17}}$	$82.96_{\pm 0.49}$	$74.11_{\pm 0.39}$	82.66
TabbyFlow	$98.18_{\pm 0.32}$	$97.03_{\pm 0.34}$	$93.33_{\pm 1.37}$	$87.12_{\pm 2.04}$	$96.79_{\pm 0.37}$	$82.18_{\pm 0.78}$	$96.18_{\pm 0.36}$	$83.25_{\pm 0.34}$	$93.76_{\pm 0.75}$	$95.85_{\pm 0.41}$	92.37
Forest-Flow	$99.04_{\pm 0.12}$	$95.11_{\pm 0.30}$	$96.59_{\pm 0.97}$	$93.55_{\pm 1.72}$	$96.16_{\pm 0.14}$	$97.82_{\pm 0.54}$	$99.30_{\pm 0.25}$	$94.65_{\pm 0.31}$	$\mathbf{98.68_{\pm 0.34}}$	$98.12_{\pm 0.35}$	96.90
STaSy	$85.96_{\pm 0.33}$	$93.67_{\pm 0.32}$	$70.91_{\pm 0.86}$	$10.49_{\pm 0.38}$	$91.08_{\pm 0.29}$	$72.03_{\pm 16.84}$	$72.11_{\pm 0.50}$	$82.12_{\pm 0.29}$	$52.16_{\pm 0.66}$	$90.88_{\pm 0.58}$	72.14
TVAE	$80.08_{\pm 0.22}$	$85.50_{\pm 0.31}$	$87.10_{\pm 1.04}$	$87.65_{\pm 1.14}$	$92.74_{\pm 0.29}$	$58.21_{\pm 0.47}$	$90.86_{\pm 0.36}$	$94.15_{\pm 0.16}$	$85.95_{\pm 0.92}$	$67.61_{\pm 0.62}$	82.99
TabDDPM	$94.76_{\pm 0.25}$	$1.01_{\pm 0.02}$	$84.05_{\pm 1.20}$	$43.26_{\pm 1.15}$	$98.27_{\pm 0.39}$	–	$97.79_{\pm 0.32}$	$0.00_{\pm 0.00}$	$4.99_{\pm 0.14}$	$94.26_{\pm 0.52}$	57.60
TabDiff	$98.98_{\pm 0.22}$	$98.11_{\pm 0.23}$	$96.12_{\pm 0.87}$	$97.27_{\pm 1.02}$	$98.62_{\pm 0.30}$	$87.83_{\pm 0.61}$	$\mathbf{99.39_{\pm 0.27}}$	$94.51_{\pm 0.30}$	$98.36_{\pm 0.50}$	$97.36_{\pm 0.51}$	96.66
TabSyn	$98.04_{\pm 0.29}$	$89.52_{\pm 0.24}$	$96.75_{\pm 0.84}$	$82.83_{\pm 1.99}$	$98.63_{\pm 0.16}$	$\mathbf{98.62_{\pm 0.35}}$	$96.81_{\pm 0.46}$	$95.36_{\pm 0.37}$	$98.39_{\pm 0.42}$	$98.04_{\pm 0.62}$	95.30
TabVVFM	$98.79_{\pm 0.25}$	$97.74_{\pm 0.33}$	$95.85_{\pm 0.47}$	$91.18_{\pm 2.06}$	$97.03_{\pm 0.93}$	$18.65_{\pm 0.35}$	$97.84_{\pm 0.42}$	$72.40_{\pm 1.86}$	$92.40_{\pm 0.75}$	$94.27_{\pm 0.48}$	85.62
TabTreeVFM	$98.92_{\pm 0.27}$	$\mathbf{98.63_{\pm 0.25}}$	$95.90_{\pm 0.76}$	$93.67_{\pm 1.72}$	$97.55_{\pm 0.20}$	$96.22_{\pm 0.52}$	$99.13_{\pm 0.27}$	$92.67_{\pm 0.20}$	$98.10_{\pm 0.35}$	$97.69_{\pm 0.44}$	96.85

Table 14: Per-dataset β-recall (%).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$\mathbf{55.14_{\pm 0.31}}$	$\mathbf{67.06_{\pm 0.28}}$	$\mathbf{63.80_{\pm 0.79}}$	$75.98_{\pm 2.25}$	$\mathbf{52.96_{\pm 0.35}}$	$\mathbf{65.93_{\pm 0.56}}$	$\mathbf{58.72_{\pm 0.39}}$	$50.84_{\pm 0.20}$	$\mathbf{70.03_{\pm 0.97}}$	$\mathbf{66.58_{\pm 0.56}}$	**62.70**
CoDi	$8.72_{\pm 0.19}$	$53.14_{\pm 0.23}$	$29.08_{\pm 0.53}$	$28.81_{\pm 1.34}$	$19.68_{\pm 0.25}$	$27.35_{\pm 0.57}$	$50.24_{\pm 0.37}$	$37.15_{\pm 0.29}$	$51.55_{\pm 0.89}$	$18.79_{\pm 0.29}$	32.45
CTGAN	$10.10_{\pm 0.19}$	$3.58_{\pm 0.08}$	$6.45_{\pm 0.42}$	$34.78_{\pm 1.58}$	$13.56_{\pm 0.15}$	$37.82_{\pm 0.45}$	$9.46_{\pm 0.26}$	$31.32_{\pm 0.32}$	$35.87_{\pm 0.99}$	$12.69_{\pm 0.22}$	19.56
TabbyFlow	$55.07_{\pm 0.26}$	$64.21_{\pm 0.26}$	$35.64_{\pm 1.03}$	$\mathbf{81.37_{\pm 1.25}}$	$52.45_{\pm 0.23}$	$61.68_{\pm 0.58}$	$51.50_{\pm 0.42}$	$\mathbf{58.92_{\pm 0.21}}$	$60.40_{\pm 0.70}$	$61.62_{\pm 0.70}$	58.29
Forest-Flow	$43.17_{\pm 0.33}$	$49.85_{\pm 0.21}$	$12.63_{\pm 0.62}$	$35.17_{\pm 1.41}$	$39.79_{\pm 0.26}$	$48.41_{\pm 0.55}$	$47.87_{\pm 0.49}$	$45.96_{\pm 0.27}$	$52.52_{\pm 0.94}$	$48.79_{\pm 0.56}$	42.42
STaSy	$28.80_{\pm 0.26}$	$45.88_{\pm 0.24}$	$12.66_{\pm 0.51}$	$1.61_{\pm 0.27}$	$41.44_{\pm 0.32}$	$24.52_{\pm 5.79}$	$41.12_{\pm 0.35}$	$32.97_{\pm 0.23}$	$33.43_{\pm 0.92}$	$32.39_{\pm 0.35}$	29.48
TVAE	$23.01_{\pm 0.10}$	$1.56_{\pm 0.03}$	$18.17_{\pm 0.66}$	$67.83_{\pm 1.54}$	$14.09_{\pm 0.13}$	$22.14_{\pm 0.25}$	$28.91_{\pm 0.36}$	$21.07_{\pm 0.16}$	$47.29_{\pm 0.63}$	$29.88_{\pm 0.32}$	27.40
TabDDPM	$50.35_{\pm 0.28}$	$6.55_{\pm 0.14}$	$14.97_{\pm 0.83}$	$8.60_{\pm 1.16}$	$47.54_{\pm 0.25}$	–	$48.28_{\pm 0.42}$	$0.00_{\pm 0.00}$	$0.16_{\pm 0.10}$	$51.92_{\pm 0.52}$	25.37
TabDiff	$51.36_{\pm 0.22}$	$59.63_{\pm 0.20}$	$26.49_{\pm 0.68}$	$49.54_{\pm 1.69}$	$51.61_{\pm 0.24}$	$27.47_{\pm 0.55}$	$47.59_{\pm 0.37}$	$41.39_{\pm 0.31}$	$49.41_{\pm 0.76}$	$53.52_{\pm 0.59}$	45.80
TabSyn	$47.88_{\pm 0.25}$	$52.87_{\pm 0.25}$	$18.78_{\pm 0.89}$	$62.94_{\pm 1.27}$	$46.12_{\pm 0.26}$	$39.49_{\pm 0.56}$	$47.62_{\pm 0.39}$	$44.21_{\pm 0.27}$	$49.66_{\pm 0.84}$	$47.97_{\pm 0.48}$	45.75
TabVVFM	$49.77_{\pm 0.25}$	$53.52_{\pm 0.31}$	$10.65_{\pm 0.53}$	$63.91_{\pm 1.72}$	$38.51_{\pm 0.32}$	$1.06_{\pm 0.15}$	$42.98_{\pm 0.48}$	$27.08_{\pm 0.47}$	$43.87_{\pm 0.65}$	$45.09_{\pm 0.47}$	37.64
TabTreeVFM	$48.38_{\pm 0.35}$	$52.92_{\pm 0.26}$	$35.03_{\pm 0.86}$	$49.61_{\pm 2.24}$	$46.25_{\pm 0.33}$	$49.62_{\pm 0.53}$	$48.85_{\pm 0.46}$	$45.01_{\pm 0.36}$	$48.84_{\pm 0.70}$	$51.12_{\pm 0.42}$	47.56

Table 15: Per-dataset MLE: AUC for classification and RMSE for regression (Beijing, Lending, News, and NMES). The Real row is the train-on-real reference.

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers
Real (reference)	$0.9272_{\pm 0.0004}$	$0.4420_{\pm 0.0069}$	$0.9767_{\pm 0.0034}$	$0.7896_{\pm 0.0281}$	$0.7699_{\pm 0.0038}$	$0.0199_{\pm 0.0011}$	$0.9458_{\pm 0.0020}$	$0.8355_{\pm 0.0172}$	$1.0931_{\pm 0.0403}$	$0.9248_{\pm 0.0020}$
CDTD	$0.9046_{\pm 0.0033}$	$\mathbf{0.5095_{\pm 0.0121}}$	$\mathbf{0.9748_{\pm 0.0058}}$	$\mathbf{0.7924_{\pm 0.0248}}$	$0.7597_{\pm 0.0059}$	$0.0413_{\pm 0.0041}$	$\mathbf{0.9405_{\pm 0.0026}}$	$\mathbf{0.8448_{\pm 0.0228}}$	$1.1347_{\pm 0.0560}$	$0.9063_{\pm 0.0050}$
CoDi	$0.8690_{\pm 0.0077}$	$0.7375_{\pm 0.0490}$	$0.9479_{\pm 0.0099}$	$0.6338_{\pm 0.0401}$	$0.4972_{\pm 0.0078}$	$0.1422_{\pm 0.0192}$	$0.9303_{\pm 0.0020}$	$1.7878_{\pm 0.0643}$	$1.1375_{\pm 0.1068}$	$0.8694_{\pm 0.0079}$
CTGAN	$0.8570_{\pm 0.0065}$	$1.7962_{\pm 0.0345}$	$0.8940_{\pm 0.0101}$	$0.7323_{\pm 0.0513}$	$0.7087_{\pm 0.0065}$	$0.2099_{\pm 0.0080}$	$0.8062_{\pm 0.0052}$	$0.9189_{\pm 0.0218}$	$1.2269_{\pm 0.0470}$	$0.7610_{\pm 0.0260}$
TabbyFlow	$0.9099_{\pm 0.0017}$	$0.5324_{\pm 0.0125}$	$0.9682_{\pm 0.0095}$	$0.7687_{\pm 0.0439}$	$0.7598_{\pm 0.0043}$	$0.0401_{\pm 0.0044}$	$0.9396_{\pm 0.0035}$	$0.8458_{\pm 0.0120}$	$1.1261_{\pm 0.0363}$	$0.9187_{\pm 0.0046}$
Forest-Flow	$0.9093_{\pm 0.0011}$	$0.6486_{\pm 0.0150}$	$0.9206_{\pm 0.0112}$	$0.6253_{\pm 0.0502}$	$0.7567_{\pm 0.0050}$	$0.0434_{\pm 0.0051}$	$0.9353_{\pm 0.0029}$	$3.0545_{\pm 0.1985}$	$1.1388_{\pm 0.0579}$	$0.9144_{\pm 0.0041}$
STaSy	$0.9052_{\pm 0.0015}$	$0.6558_{\pm 0.0104}$	$0.8997_{\pm 0.0142}$	$0.5841_{\pm 0.0943}$	$0.7561_{\pm 0.0036}$	$0.0858_{\pm 0.2558}$	$0.9231_{\pm 0.0042}$	$0.8740_{\pm 0.0284}$	$1.0820_{\pm 0.0487}$	$0.9124_{\pm 0.0031}$
TVAE	$0.8860_{\pm 0.0024}$	$1.1083_{\pm 0.0184}$	$0.9551_{\pm 0.0066}$	$0.7366_{\pm 0.0495}$	$0.7186_{\pm 0.0070}$	$0.3986_{\pm 0.0818}$	$0.9052_{\pm 0.0042}$	$0.9527_{\pm 0.0148}$	$1.2249_{\pm 0.0885}$	$0.8814_{\pm 0.0106}$
TabDDPM	$0.9103_{\pm 0.0009}$	$2.0628_{\pm 0.3136}$	$0.9313_{\pm 0.0110}$	$0.5013_{\pm 0.0896}$	$0.7624_{\pm 0.0042}$	–	$0.9342_{\pm 0.0023}$	–	$1.4108_{\pm 0.1347}$	$0.9115_{\pm 0.0040}$
TabDiff	$0.9121_{\pm 0.0013}$	$0.5637_{\pm 0.0120}$	$0.9637_{\pm 0.0080}$	$0.7293_{\pm 0.0457}$	$0.7620_{\pm 0.0059}$	$0.0899_{\pm 0.0098}$	$0.9373_{\pm 0.0029}$	$0.8629_{\pm 0.0233}$	$1.0759_{\pm 0.0498}$	$\mathbf{0.9214_{\pm 0.0040}}$
TabSyn	$0.9105_{\pm 0.0008}$	$0.6824_{\pm 0.0176}$	$0.9539_{\pm 0.0132}$	$0.7602_{\pm 0.0301}$	$0.7596_{\pm 0.0045}$	$0.0702_{\pm 0.0047}$	$0.9353_{\pm 0.0032}$	$0.8658_{\pm 0.0254}$	$\mathbf{1.0627_{\pm 0.0452}}$	$0.9087_{\pm 0.0049}$
TabVVFM	$0.9094_{\pm 0.0012}$	$0.6363_{\pm 0.0108}$	$0.9280_{\pm 0.0151}$	$0.7349_{\pm 0.0370}$	$\mathbf{0.7625_{\pm 0.0053}}$	$0.4303_{\pm 0.0337}$	$0.9297_{\pm 0.0029}$	$0.9396_{\pm 0.0721}$	$1.0676_{\pm 0.0369}$	$0.9054_{\pm 0.0053}$
TabTreeVFM	$\mathbf{0.9150_{\pm 0.0015}}$	$0.6477_{\pm 0.0181}$	$0.9637_{\pm 0.0068}$	$0.7469_{\pm 0.0446}$	$0.7610_{\pm 0.0031}$	$\mathbf{0.0369_{\pm 0.0033}}$	$0.9377_{\pm 0.0034}$	$0.8511_{\pm 0.0183}$	$1.0848_{\pm 0.0729}$	$0.9195_{\pm 0.0043}$

Table 16: Per-dataset DCR shares (%). Avg $|\mathrm{DCR}-50|$ is the mean absolute deviation from 50%, in percentage points.

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg $\lvert\mathrm{DCR}-50\rvert$
CDTD	$69.40_{\pm 0.27}$	$66.87_{\pm 0.30}$	$72.31_{\pm 1.03}$	$99.79_{\pm 0.18}$	$63.00_{\pm 0.44}$	$89.29_{\pm 0.48}$	$70.35_{\pm 0.46}$	$53.64_{\pm 0.38}$	$86.80_{\pm 0.86}$	$91.05_{\pm 0.36}$	26.25
CoDi	$51.08_{\pm 0.34}$	$56.49_{\pm 0.32}$	$61.91_{\pm 1.20}$	$\mathbf{54.27_{\pm 1.63}}$	$52.29_{\pm 0.38}$	$67.83_{\pm 0.72}$	$65.22_{\pm 0.46}$	$54.03_{\pm 0.31}$	$94.93_{\pm 0.54}$	$51.93_{\pm 0.60}$	11.00
CTGAN	$52.13_{\pm 0.34}$	$46.51_{\pm 0.37}$	$51.97_{\pm 1.13}$	$59.18_{\pm 2.48}$	$52.05_{\pm 0.47}$	$52.66_{\pm 0.90}$	$\mathbf{49.04_{\pm 0.57}}$	$\mathbf{50.54_{\pm 0.28}}$	$54.01_{\pm 1.04}$	$52.01_{\pm 0.64}$	2.90
TabbyFlow	$60.27_{\pm 0.37}$	$61.65_{\pm 0.31}$	$56.68_{\pm 0.78}$	$97.32_{\pm 0.72}$	$57.30_{\pm 0.45}$	$73.34_{\pm 0.65}$	$52.88_{\pm 0.47}$	$64.47_{\pm 0.32}$	$66.23_{\pm 1.07}$	$65.44_{\pm 0.75}$	15.56
Forest-Flow	$51.43_{\pm 0.27}$	$51.67_{\pm 0.40}$	$52.03_{\pm 1.10}$	$58.94_{\pm 2.75}$	$51.34_{\pm 0.42}$	$54.13_{\pm 0.70}$	$54.60_{\pm 0.50}$	$51.98_{\pm 0.35}$	$56.97_{\pm 1.21}$	$52.71_{\pm 0.51}$	3.58
STaSy	$\mathbf{50.39_{\pm 0.31}}$	$\mathbf{51.54_{\pm 0.32}}$	$51.99_{\pm 0.87}$	$69.90_{\pm 2.03}$	$\mathbf{50.50_{\pm 0.34}}$	$55.91_{\pm 0.73}$	$55.66_{\pm 0.48}$	$51.56_{\pm 0.33}$	$53.84_{\pm 0.97}$	$\mathbf{50.60_{\pm 0.74}}$	4.19
TVAE	$56.22_{\pm 0.30}$	$58.13_{\pm 0.25}$	$64.72_{\pm 1.12}$	$88.22_{\pm 1.37}$	$51.82_{\pm 0.34}$	$52.31_{\pm 0.76}$	$51.42_{\pm 0.56}$	$54.78_{\pm 0.30}$	$63.65_{\pm 1.22}$	$57.93_{\pm 0.59}$	9.92
TabDDPM	$50.95_{\pm 0.29}$	$36.02_{\pm 0.25}$	$59.55_{\pm 0.88}$	$81.75_{\pm 1.51}$	$52.51_{\pm 0.35}$	–	$52.87_{\pm 0.46}$	$1.62_{\pm 0.09}$	$48.66_{\pm 1.07}$	$28.95_{\pm 0.55}$	14.71
TabDiff	$56.71_{\pm 0.34}$	$54.88_{\pm 0.34}$	$56.21_{\pm 1.11}$	$63.96_{\pm 1.84}$	$56.94_{\pm 0.40}$	$53.56_{\pm 0.50}$	$52.13_{\pm 0.66}$	$54.30_{\pm 0.30}$	$\mathbf{51.24_{\pm 0.92}}$	$52.24_{\pm 0.70}$	5.22
TabSyn	$50.99_{\pm 0.35}$	$52.59_{\pm 0.36}$	$53.29_{\pm 1.18}$	$80.09_{\pm 2.01}$	$50.90_{\pm 0.40}$	$52.04_{\pm 1.00}$	$52.17_{\pm 0.54}$	$50.63_{\pm 0.35}$	$54.00_{\pm 0.98}$	$51.71_{\pm 0.64}$	4.84
TabVVFM	$52.97_{\pm 0.35}$	$54.04_{\pm 0.40}$	$\mathbf{50.43_{\pm 1.10}}$	$75.18_{\pm 2.33}$	$52.90_{\pm 0.38}$	$\mathbf{51.99_{\pm 0.91}}$	$51.23_{\pm 0.49}$	$52.15_{\pm 0.36}$	$54.86_{\pm 0.86}$	$54.11_{\pm 0.65}$	4.99
TabTreeVFM	$51.21_{\pm 0.24}$	$52.77_{\pm 0.20}$	$51.26_{\pm 0.64}$	$57.04_{\pm 1.81}$	$51.11_{\pm 0.23}$	$52.06_{\pm 0.57}$	$53.77_{\pm 0.41}$	$51.54_{\pm 0.26}$	$51.28_{\pm 0.87}$	$52.75_{\pm 0.43}$	**2.48**

Table 17: Per-dataset MIA score.

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg
CDTD	$0.9614_{\pm 0.0028}$	$0.9904_{\pm 0.0057}$	$0.9377_{\pm 0.0306}$	$0.9175_{\pm 0.0603}$	$0.9683_{\pm 0.0111}$	$0.9793_{\pm 0.0152}$	$0.9819_{\pm 0.0111}$	$0.9817_{\pm 0.0050}$	$0.9519_{\pm 0.0238}$	$0.9630_{\pm 0.0144}$	0.9633
CoDi	$\mathbf{0.9998_{\pm 0.0004}}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9986_{\pm 0.0057}$	$\mathbf{0.9916_{\pm 0.0185}}$	$\mathbf{1.0000_{\pm 0.0000}}$	$\underline{0.9991_{\pm 0.0039}}$	$\mathbf{1.0000_{\pm 0.0001}}$	$0.9935_{\pm 0.0078}$	$0.9804_{\pm 0.0126}$	$\mathbf{1.0000_{\pm 0.0000}}$	**0.9963**
CTGAN	$0.9937_{\pm 0.0033}$	$\underline{0.9997_{\pm 0.0011}}$	$0.9975_{\pm 0.0073}$	$0.9647_{\pm 0.0372}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9920_{\pm 0.0096}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9909_{\pm 0.0081}$	$\underline{0.9998_{\pm 0.0010}}$	$\mathbf{1.0000_{\pm 0.0000}}$	$\underline{0.9938}$
TabbyFlow	$0.9665_{\pm 0.0045}$	$0.9893_{\pm 0.0080}$	$0.9685_{\pm 0.0251}$	$0.9196_{\pm 0.0519}$	$0.9747_{\pm 0.0086}$	$0.9651_{\pm 0.0210}$	$0.9860_{\pm 0.0115}$	$0.9936_{\pm 0.0055}$	$0.9693_{\pm 0.0198}$	$0.9682_{\pm 0.0131}$	0.9701
Forest-Flow	$0.9959_{\pm 0.0027}$	$0.9977_{\pm 0.0027}$	$0.9966_{\pm 0.0120}$	$0.8943_{\pm 0.0405}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9982_{\pm 0.0037}$	$0.9932_{\pm 0.0082}$	$0.9884_{\pm 0.0097}$	$0.9836_{\pm 0.0138}$	$\mathbf{1.0000_{\pm 0.0000}}$	0.9848
STaSy	$\underline{0.9988_{\pm 0.0012}}$	$\mathbf{1.0000_{\pm 0.0000}}$	$\mathbf{0.9998_{\pm 0.0006}}$	$0.8592_{\pm 0.0579}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9888_{\pm 0.0110}$	$\mathbf{1.0000_{\pm 0.0000}}$	$\mathbf{0.9999_{\pm 0.0004}}$	$0.9993_{\pm 0.0022}$	$\underline{0.9979_{\pm 0.0039}}$	0.9844
TVAE	$0.9982_{\pm 0.0019}$	$0.9847_{\pm 0.0047}$	$\underline{0.9992_{\pm 0.0035}}$	$0.9901_{\pm 0.0231}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9933_{\pm 0.0079}$	$\underline{0.9973_{\pm 0.0065}}$	$0.9953_{\pm 0.0063}$	$0.9807_{\pm 0.0106}$	$0.9959_{\pm 0.0075}$	0.9935
TabDDPM	$0.9706_{\pm 0.0038}$	$0.9897_{\pm 0.0086}$	$0.9523_{\pm 0.0267}$	$0.9738_{\pm 0.0365}$	$0.9835_{\pm 0.0086}$	–	$0.9848_{\pm 0.0109}$	$0.9960_{\pm 0.0000}$	$0.9841_{\pm 0.0201}$	$0.9724_{\pm 0.0139}$	0.9786
TabDiff	$0.9688_{\pm 0.0048}$	$0.9886_{\pm 0.0077}$	$0.9563_{\pm 0.0278}$	$0.9270_{\pm 0.0679}$	$0.9859_{\pm 0.0100}$	$0.9755_{\pm 0.0100}$	$0.9937_{\pm 0.0090}$	$0.9869_{\pm 0.0069}$	$0.9738_{\pm 0.0215}$	$0.9793_{\pm 0.0149}$	0.9736
TabSyn	$0.9711_{\pm 0.0051}$	$0.9994_{\pm 0.0012}$	$0.9704_{\pm 0.0249}$	$0.9554_{\pm 0.0443}$	$0.9857_{\pm 0.0104}$	$0.9809_{\pm 0.0138}$	$0.9882_{\pm 0.0082}$	$0.9892_{\pm 0.0053}$	$0.9941_{\pm 0.0128}$	$0.9580_{\pm 0.0099}$	0.9792
TabVVFM	$0.9858_{\pm 0.0032}$	$0.9951_{\pm 0.0052}$	$0.9907_{\pm 0.0125}$	$0.9909_{\pm 0.0180}$	$0.9826_{\pm 0.0093}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9953_{\pm 0.0070}$	$0.9800_{\pm 0.0040}$	$\mathbf{1.0000_{\pm 0.0000}}$	$0.9552_{\pm 0.0111}$	0.9876
TabTreeVFM	$0.9681_{\pm 0.0041}$	$0.9964_{\pm 0.0033}$	$0.9627_{\pm 0.0161}$	$0.9841_{\pm 0.0175}$	$\underline{0.9897_{\pm 0.0063}}$	$0.9929_{\pm 0.0088}$	$0.9763_{\pm 0.0073}$	$\underline{0.9970_{\pm 0.0030}}$	$0.9892_{\pm 0.0111}$	$0.9808_{\pm 0.0129}$	0.9837

Table 18: Per-dataset signed reproduction gaps and average Mem@1% (percentage points).

Method	Adult	Beijing	Churn	Credit-g	Default	Lending	Magic	News	NMES	Shoppers	Avg Mem
Held-out rate	2.50	0.00	1.21	0.00	3.52	0.00	0.02	0.00	0.18	1.33	–
CDTD	$2.12_{\pm 0.09}$	$1.29_{\pm 0.07}$	$9.75_{\pm 0.84}$	$28.26_{\pm 2.24}$	$1.38_{\pm 0.14}$	$0.01_{\pm 0.01}$	$0.36_{\pm 0.08}$	$0.00_{\pm 0.00}$	$18.75_{\pm 0.85}$	$21.73_{\pm 0.47}$	8.37
CoDi	$-2.12_{\pm 0.04}$	$0.00_{\pm 0.00}$	$-0.72_{\pm 0.14}$	$0.00_{\pm 0.00}$	$-3.36_{\pm 0.03}$	$0.00_{\pm 0.00}$	$-0.01_{\pm 0.01}$	$0.00_{\pm 0.00}$	$1.37_{\pm 0.27}$	$-0.51_{\pm 0.09}$	0.14
CTGAN	$-0.94_{\pm 0.07}$	$0.00_{\pm 0.00}$	$-1.20_{\pm 0.02}$	$0.00_{\pm 0.00}$	$-2.56_{\pm 0.09}$	$0.00_{\pm 0.00}$	$-0.02_{\pm 0.00}$	$0.00_{\pm 0.00}$	$-0.11_{\pm 0.06}$	$-1.32_{\pm 0.01}$	0.00
TabbyFlow	$0.71_{\pm 0.15}$	$0.05_{\pm 0.01}$	$2.15_{\pm 0.59}$	$27.59_{\pm 2.23}$	$0.05_{\pm 0.14}$	$0.22_{\pm 0.06}$	$-0.02_{\pm 0.01}$	$0.00_{\pm 0.00}$	$1.51_{\pm 0.27}$	$2.80_{\pm 0.20}$	3.51
Forest-Flow	$-1.05_{\pm 0.06}$	$0.00_{\pm 0.01}$	$-0.69_{\pm 0.18}$	$0.00_{\pm 0.00}$	$-1.90_{\pm 0.12}$	$0.00_{\pm 0.00}$	$-0.02_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.06_{\pm 0.11}$	$0.23_{\pm 0.17}$	0.03
STaSy	$-1.41_{\pm 0.07}$	$0.00_{\pm 0.01}$	$-1.02_{\pm 0.11}$	$0.00_{\pm 0.00}$	$-2.45_{\pm 0.10}$	$0.00_{\pm 0.00}$	$-0.01_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.08}$	$0.39_{\pm 0.17}$	0.04
TVAE	$1.26_{\pm 0.12}$	$0.00_{\pm 0.00}$	$-1.12_{\pm 0.08}$	$9.66_{\pm 1.08}$	$-0.80_{\pm 0.14}$	$0.00_{\pm 0.00}$	$-0.02_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.75_{\pm 0.21}$	$-0.20_{\pm 0.16}$	1.17
TabDDPM	$0.12_{\pm 0.12}$	$0.00_{\pm 0.00}$	$2.39_{\pm 0.37}$	$0.36_{\pm 0.26}$	$-0.80_{\pm 0.13}$	–	$-0.02_{\pm 0.00}$	$0.00_{\pm 0.00}$	$-0.18_{\pm 0.00}$	$-1.32_{\pm 0.01}$	0.32
TabDiff	$0.21_{\pm 0.09}$	$0.01_{\pm 0.01}$	$1.90_{\pm 0.34}$	$0.01_{\pm 0.04}$	$-0.17_{\pm 0.17}$	$0.00_{\pm 0.00}$	$-0.01_{\pm 0.01}$	$0.00_{\pm 0.00}$	$-0.10_{\pm 0.06}$	$0.85_{\pm 0.19}$	0.30
TabSyn	$-0.17_{\pm 0.09}$	$0.01_{\pm 0.00}$	$2.26_{\pm 0.46}$	$0.00_{\pm 0.00}$	$-1.14_{\pm 0.14}$	$0.00_{\pm 0.00}$	$-0.02_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.13}$	$0.84_{\pm 0.16}$	0.31
TabVVFM	$0.25_{\pm 0.09}$	$0.00_{\pm 0.00}$	$0.26_{\pm 0.40}$	$0.06_{\pm 0.09}$	$-1.87_{\pm 0.10}$	$0.00_{\pm 0.00}$	$-0.02_{\pm 0.01}$	$0.00_{\pm 0.00}$	$0.03_{\pm 0.10}$	$0.50_{\pm 0.17}$	0.11
TabTreeVFM	$-0.21_{\pm 0.06}$	$0.00_{\pm 0.00}$	$0.44_{\pm 0.19}$	$0.00_{\pm 0.00}$	$-1.08_{\pm 0.13}$	$0.00_{\pm 0.00}$	$-0.02_{\pm 0.01}$	$0.00_{\pm 0.00}$	$-0.05_{\pm 0.05}$	$0.99_{\pm 0.14}$	0.14

Table 19: Avg Mem and average within-dataset Mem rank at five tolerances.

Method	$\tau = 0.1\%$		$\tau = 0.5\%$		$\tau = 1\%$		$\tau = 2\%$		$\tau = 5\%$	
	Mem	Rank	Mem	Rank	Mem	Rank	Mem	Rank	Mem	Rank
CDTD	2.45	10.25	5.79	11.25	8.37	11.80	17.30	11.70	29.80	11.00
CoDi	0.00	4.75	0.00	4.50	0.14	4.80	2.34	6.10	7.72	6.95
CTGAN	0.00	4.75	0.00	4.20	0.00	3.90	0.00	3.90	0.01	3.95
TabbyFlow	0.43	9.60	1.79	9.15	3.51	9.70	6.54	10.15	12.06	9.80
Forest-Flow	0.00	4.75	0.00	4.75	0.03	5.25	0.18	5.00	0.33	5.30
STaSy	0.04	5.55	0.05	5.50	0.04	5.15	0.04	4.70	0.13	4.35
TVAE	0.28	6.00	0.87	6.60	1.17	6.20	3.91	7.45	7.23	8.35
TabDDPM	0.20	5.56	0.26	5.50	0.32	5.67	0.38	4.61	0.55	4.94
TabDiff	0.22	7.20	0.25	6.95	0.30	6.70	0.66	6.60	0.96	6.00
TabSyn	0.26	7.05	0.29	6.60	0.31	6.25	0.40	5.75	0.58	5.50
TabVVFM	0.08	6.10	0.09	6.70	0.11	6.55	0.23	6.40	0.96	6.45
TabTreeVFM	0.19	5.80	0.18	5.65	0.14	5.40	0.14	4.90	0.28	4.70

J Stability Across Training Seeds

To confirm that the results of Section 4 do not depend on a particular training run, we train TabTreeVFM with 10 random seeds on every dataset, keeping the data splits and hyperparameters fixed. The training seed sets both the base-noise draws that define every interpolated training table and the random seed of XGBoost. Each model generates 20 synthetic tables that are evaluated under the main protocol, and Table 20 reports the mean and standard deviation of the resulting scores over the 10 models.

Table 20: Stability of TabTreeVFM across 10 training seeds. Shape, Trend, C2ST, α-precision, and β-recall are in %; $|\mathrm{DCR} - 50|$ and Mem@1% are in percentage points; MLE is AUC for classification and RMSE for regression datasets.

Dataset	Fidelity							Utility	Privacy		
	Shape ↓	Trend ↓	WD ↓	JSD ↓	C2ST ↑	α-Prec. ↑	β-Rec. ↑	MLE	$\lvert\mathrm{DCR}-50\rvert$ ↓	MIA ↑	Mem@1% ↓
Adult	$0.64_{\pm 0.02}$	$1.53_{\pm 0.02}$	$0.0019_{\pm 0.0001}$	$0.0110_{\pm 0.0002}$	$81.22_{\pm 0.29}$	$98.81_{\pm 0.05}$	$48.57_{\pm 0.08}$	$0.9167_{\pm 0.0004}$	$1.37_{\pm 0.07}$	$0.9677_{\pm 0.0013}$	$0.00_{\pm 0.00}$
Beijing	$0.85_{\pm 0.01}$	$3.18_{\pm 0.15}$	$0.0021_{\pm 0.0000}$	$0.0086_{\pm 0.0001}$	$86.91_{\pm 0.17}$	$98.62_{\pm 0.07}$	$52.86_{\pm 0.06}$	$0.6451_{\pm 0.0025}$	$2.72_{\pm 0.08}$	$0.9940_{\pm 0.0021}$	$0.00_{\pm 0.00}$
Churn	$1.80_{\pm 0.02}$	$2.58_{\pm 0.07}$	$0.0072_{\pm 0.0002}$	$0.0133_{\pm 0.0007}$	$88.94_{\pm 0.73}$	$95.75_{\pm 0.16}$	$34.25_{\pm 0.25}$	$0.9617_{\pm 0.0031}$	$0.87_{\pm 0.41}$	$0.9543_{\pm 0.0059}$	$0.45_{\pm 0.09}$
Credit-g	$2.52_{\pm 0.04}$	$4.64_{\pm 0.10}$	$0.0121_{\pm 0.0004}$	$0.0287_{\pm 0.0006}$	$92.95_{\pm 0.89}$	$94.25_{\pm 0.39}$	$48.64_{\pm 0.26}$	$0.7427_{\pm 0.0070}$	$8.28_{\pm 0.41}$	$0.9803_{\pm 0.0072}$	$0.00_{\pm 0.00}$
Default	$1.28_{\pm 0.01}$	$2.62_{\pm 0.03}$	$0.0013_{\pm 0.0000}$	$0.0188_{\pm 0.0002}$	$82.40_{\pm 0.40}$	$97.43_{\pm 0.03}$	$46.28_{\pm 0.09}$	$0.7603_{\pm 0.0009}$	$1.17_{\pm 0.08}$	$0.9883_{\pm 0.0030}$	$0.00_{\pm 0.00}$
Lending	$1.58_{\pm 0.02}$	$1.44_{\pm 0.10}$	$0.0032_{\pm 0.0000}$	$0.0177_{\pm 0.0003}$	$69.84_{\pm 0.86}$	$96.16_{\pm 0.15}$	$49.68_{\pm 0.15}$	$0.0354_{\pm 0.0006}$	$2.11_{\pm 0.15}$	$0.9955_{\pm 0.0019}$	$0.00_{\pm 0.00}$
Magic	$0.78_{\pm 0.02}$	$1.08_{\pm 0.08}$	$0.0019_{\pm 0.0001}$	$0.0032_{\pm 0.0004}$	$93.66_{\pm 0.55}$	$99.28_{\pm 0.05}$	$48.80_{\pm 0.08}$	$0.9372_{\pm 0.0004}$	$4.55_{\pm 0.09}$	$0.9766_{\pm 0.0035}$	$0.00_{\pm 0.00}$
News	$1.63_{\pm 0.01}$	$7.50_{\pm 0.24}$	$0.0044_{\pm 0.0000}$	$0.0140_{\pm 0.0004}$	$78.87_{\pm 0.35}$	$92.69_{\pm 0.06}$	$45.03_{\pm 0.07}$	$0.8549_{\pm 0.0043}$	$1.54_{\pm 0.07}$	$0.9965_{\pm 0.0014}$	$0.00_{\pm 0.00}$
NMES	$1.00_{\pm 0.03}$	$3.23_{\pm 0.15}$	$0.0029_{\pm 0.0001}$	$0.0073_{\pm 0.0005}$	$87.75_{\pm 0.93}$	$98.03_{\pm 0.08}$	$48.92_{\pm 0.16}$	$1.0912_{\pm 0.0129}$	$1.57_{\pm 0.41}$	$0.9920_{\pm 0.0053}$	$0.00_{\pm 0.00}$
Shoppers	$0.93_{\pm 0.02}$	$1.49_{\pm 0.09}$	$0.0022_{\pm 0.0001}$	$0.0184_{\pm 0.0005}$	$94.15_{\pm 0.45}$	$97.67_{\pm 0.08}$	$51.09_{\pm 0.14}$	$0.9192_{\pm 0.0009}$	$2.73_{\pm 0.14}$	$0.9855_{\pm 0.0030}$	$0.95_{\pm 0.03}$
Avg	$1.30_{\pm 0.00}$	$2.93_{\pm 0.05}$	$0.0039_{\pm 0.0000}$	$0.0141_{\pm 0.0001}$	$85.67_{\pm 0.19}$	$96.87_{\pm 0.04}$	$47.41_{\pm 0.04}$	–	$2.69_{\pm 0.07}$	$0.9831_{\pm 0.0011}$	$0.14_{\pm 0.01}$

Results are stable across training seeds. The dataset-averaged scores vary little across seeds: their standard deviation is below 0.01 for Shape, 0.05 for Trend, 0.19 for C2ST, and 0.04 for both α-precision and β-recall, all in percentage points, and the C2ST standard deviation stays below one percentage point on every dataset. The dataset-averaged C2ST margin of TabTreeVFM over the next best generator, about six percentage points, is more than 30 times the seed-to-seed standard deviation of the average. The privacy diagnostics are equally stable: the average MIA score varies by 0.001 and the average $|\mathrm{DCR} - 50|$ by 0.07 percentage points, and Mem@1% is nonzero only on Churn and Shoppers, as in the main benchmark.

K Qualitative Analysis

Figures 5 and 6 localize the distributional errors summarized by the aggregate fidelity metrics of Section 4.2.

Univariate marginals. Figure 5 plots the same representative numerical and categorical columns for the selected methods. On Lending, for example, TabTreeVFM follows the multiple modes of the interest-rate attribute, whereas TabDiff smooths over much of this structure. Across the categorical panels, TabTreeVFM also tracks both dominant and lower-frequency categories without the systematic frequency shifts visible for several baselines. These examples agree with the Shape, WD, and JSD results.

Pairwise dependencies. Figure 6 reports pairwise dependency errors across all 10 datasets, using correlation error for numerical pairs and contingency-table distance otherwise. Forest-Flow's large error bands on Adult and Credit-g are largely absent for TabTreeVFM, whose remaining errors are lighter and more localized. This visual contrast is consistent with the aggregate Trend improvement and the endpoint-posterior ablation in Section 4.4.



Figure 5: Univariate marginals across 10 datasets.

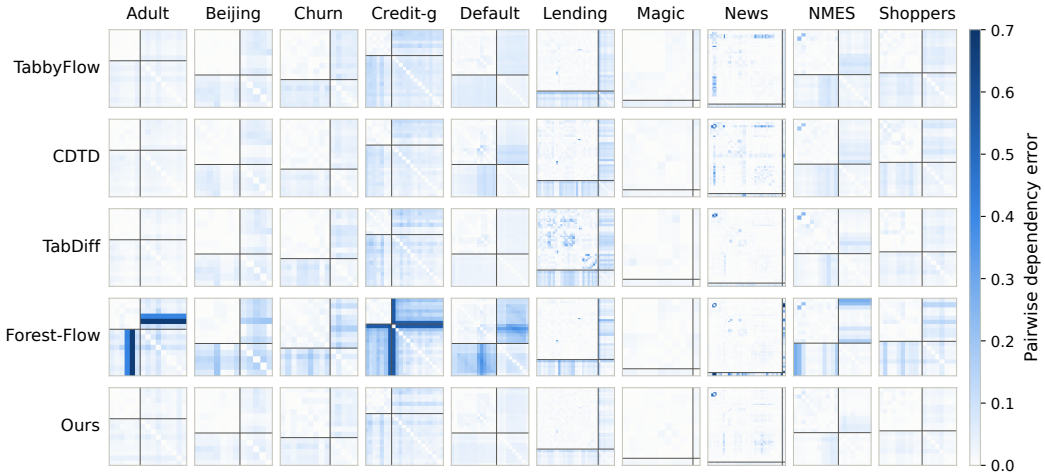


Figure 6: Pairwise dependency errors across 10 datasets. Columns denote datasets and rows denote methods.

L Efficiency and Scalability

We evaluate TabTreeVFM in terms of training time, sampling time, memory use, and scalability. As in prior tree-based diffusion and flow models (Cresswell & Kim, 2024), training uses nK interpolated examples; we therefore evaluate how training time and memory scale with both n and K.

The measurements follow the same benchmark protocol as the main experiments, with sampling cost reported as seconds per 1,000 generated rows. We first characterize TabTreeVFM itself and then compare it with representative neural baselines and the tree-based Forest-Flow. Because the per-head fits require neither backpropagation through timesteps nor a shared optimizer state, the same estimator trains in CPU-only or GPU-accelerated mode without modification. We report both modes to separate the intrinsic algorithmic cost from hardware acceleration.

We report row and K scaling of training time and memory in Figures 7 and 8, and cross-method training and sampling costs in Tables 21 and 22.

Scaling in the number of rows. Figure 7 reports CPU and GPU training-time scaling at $K{=}100$ and $K{=}10$. Once fixed overheads are amortized, training time grows approximately linearly with

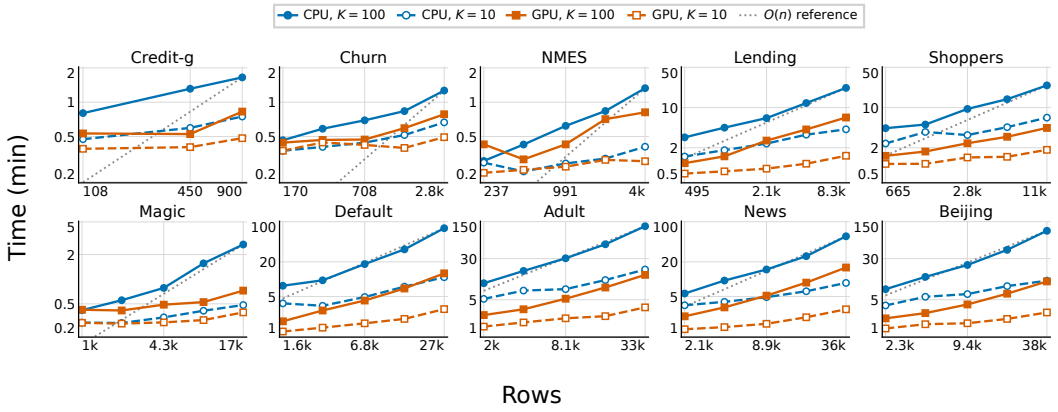


Figure 7: Training-time scaling.

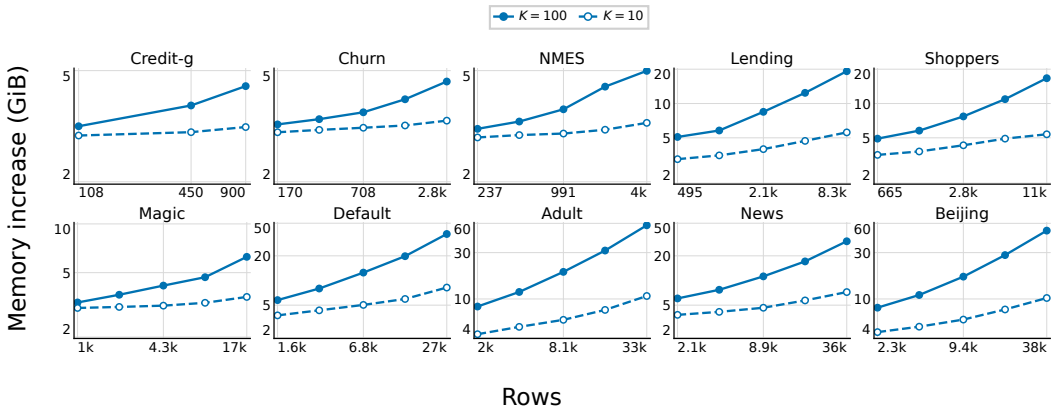


Figure 8: CPU memory scaling.

the number of rows. The benefit of GPU acceleration depends on workload size and feature composition. The CPU–GPU gap is modest on the smaller datasets, whereas GPU training at full scale is $6.1\times$ faster on Default, $8.3\times$ faster on Adult, and $9.1\times$ faster on Beijing. This pattern is consistent with larger, categorical-heavy workloads providing sufficient parallel work to amortize accelerator overhead.

Reducing K shifts both the CPU and GPU curves downward by shrinking each timestep's training table of nK rows. On CPU, $K{=}10$ reduces full-scale training time by $6.6\times$ on Adult, $8.8\times$ on Beijing, $7.1\times$ on Default, and $6.5\times$ on News relative to $K{=}100$.

Memory. Figure 8 shows that peak memory use during CPU training grows approximately linearly with the number of rows and decreases substantially with K. At full scale, reducing K from 100 to 10 lowers peak memory by 1.4–$5.3\times$ across the 10 datasets, from 4.4–57.3 GiB to 3.1–10.7 GiB. These measurements show that K provides direct control over training memory.

Training cost. Table 21 shows that, at $K{=}100$, TabTreeVFM trains in 0.7–16 minutes on GPU and 1.3–123 minutes on CPU across the 10 datasets. Its GPU training time is lower than that of TabDiff, TabbyFlow, TabSyn, and Forest-Flow across the benchmark, while CDTD's fixed-iteration schedule keeps its GPU time at 2.3–3.0 minutes on every dataset, a level TabTreeVFM reaches at $K{=}10$. Reducing K to 10 further lowers training time to 0.3–3.6 minutes on GPU and 0.4–19 minutes on CPU, demonstrating the direct cost control provided by K.

Sampling cost. Table 22 shows that TabTreeVFM achieves competitive sampling throughput while evaluating all boosted heads at every solver step. On GPU, it samples at 0.01–0.09 seconds per $1{,}000$ rows on nine of the 10 datasets, faster than TabDiff and TabbyFlow everywhere, tied for fastest on News and Magic, and fastest on Lending. On Credit-g, whose 900 rows leave the fixed per-step work unamortized, sampling takes 0.21 seconds per $1{,}000$ rows. CPU sampling stays within 0.02–0.30 seconds per $1{,}000$ rows throughout.

Table 21: Training time (minutes).

Method	HW	Dataset									
		Credit-g	Churn	NMES	Lending	Shoppers	Magic	Default	Adult	News	Beijing
TabDiff	GPU	63	63	73	84	82	85	101	104	113	101
TabbyFlow	GPU	60	61	72	85	80	84	100	104	110	100
TabSyn	GPU	18	20	20	25	25	29	31	29	40	35
CDTD	GPU	2.4	2.3	2.9	3.0	2.9	2.5	2.8	3.0	2.3	2.6
Forest-Flow	CPU	1.1	1.2	1.0	16	20	4.6	63	99	55	80
Forest-Flow	GPU	4.7	4.2	2.2	17	27	3.5	51	74	36	55
TabTreeVFM (K=100)	CPU	1.6	1.3	1.3	22	24	2.6	77	123	56	101
TabTreeVFM (K=100)	GPU	0.8	0.8	0.8	6.7	4.4	0.7	13	15	16	11
TabTreeVFM (K=10)	CPU	0.8	0.7	0.4	4.2	6.7	0.5	11	19	8.6	11
TabTreeVFM (K=10)	GPU	0.5	0.5	0.3	1.4	1.8	0.4	3.0	3.6	3.0	2.9

Table 22: Sampling time (seconds per 1,000 generated rows).

Method	HW	Dataset									
		Credit-g	Churn	NMES	Lending	Shoppers	Magic	Default	Adult	News	Beijing
TabDiff	GPU	1.33	0.32	0.27	0.21	0.18	0.09	0.14	0.14	0.17	0.10
TabbyFlow	GPU	1.10	0.30	0.24	0.17	0.15	0.09	0.12	0.10	0.17	0.08
TabSyn	GPU	0.05	0.05	0.05	0.09	0.05	0.05	0.06	0.05	0.09	0.04
CDTD	GPU	0.44	0.14	0.10	0.08	0.06	0.04	0.04	0.04	0.04	0.03
Forest-Flow	CPU	0.16	0.05	0.02	0.08	0.12	0.01	0.13	0.15	0.07	0.11
Forest-Flow	GPU	0.15	0.05	0.02	0.09	0.14	0.02	0.15	0.15	0.08	0.12
TabTreeVFM	CPU	0.30	0.09	0.06	0.15	0.13	0.02	0.14	0.16	0.10	0.12
TabTreeVFM	GPU	0.21	0.05	0.03	0.07	0.08	0.01	0.08	0.09	0.04	0.07

Overall, training time and memory grow approximately linearly with the number of rows and decrease as K is reduced. Every dataset trains on CPU alone, GPU training provides higher throughput, and sampling is faster than TabDiff and TabbyFlow on every dataset. Together, these results show that TabTreeVFM is practical on both CPU and GPU and exposes K as a direct quality–cost control.

M Hyperparameter Sensitivity

We examine the sensitivity of TabTreeVFM to its two main training hyperparameters: the duplication factor $K \in \{10, 20, 30, 50, 100\}$ and maximum tree depth in $\{3, 5, 7, 9\}$. Figure 9 summarizes their effects on six representative fidelity and utility metrics across the benchmark datasets.

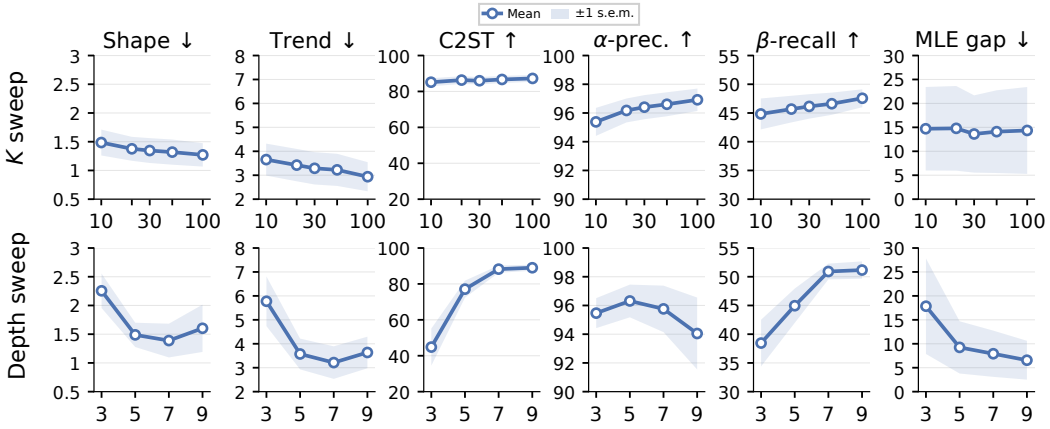


Figure 9: Sensitivity of TabTreeVFM to K (top) and tree depth (bottom), aggregated across datasets.

The top row of Figure 9 shows the performance of TabTreeVFM as the duplication factor K varies. Across the tested range, increasing K improves the aggregate performance profile. From K=10 to K=100, average Shape decreases from approximately 1.5 to 1.3, Trend decreases from about 3.6

to 2.9, β-recall rises from roughly 45 to 48, and the MLE gap stays within one percentage point of its $K{=}100$ value; C2ST and α-precision improve by smaller margins. The gains are gradual, with β-recall improving most near the upper end of the sweep. This behavior is consistent with additional noise pairings providing denser coverage of the interpolated states used to learn the endpoint predictors, while expanding each timestep's training table to nK rows. $K{=}100$ gives the strongest aggregate results and serves as our main benchmark setting, while $K{=}10$ offers a 1.9–$8.8\times$ cheaper CPU operating point (Figure 7) with a gradual loss in fidelity; several metrics still improve at the upper end of the sweep, leaving headroom for larger K.

The bottom row of Figure 9 shows a capacity trade-off. On average, increasing depth from 3 to 7 lowers Trend from about 5.8 to 3.3 and raises β-recall from about 39 to 51; depth 9 trades Shape, Trend, and α-precision for small further gains in C2ST, β-recall, and MLE, so depth 7 is the balanced default. The optimal depth depends on dataset complexity: depth 3 provides the best or near-best marginal and dependency fidelity on Credit-g, NMES, and Lending, whereas the more complex datasets benefit from deeper trees.