TABTAILOR: TAIL AND ORDINAL-AWARE MIXED-TYPE DIFFUSION MODEL FOR TABULAR DATA SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models have significantly enhanced the quality of synthetic tabular data and have been applied to diverse real-world domains. However, current tabular generative models lack the ability to preserve tails of numerical features and they consider ordinal features as unordered. Therefore, preserving heavy-tailed numerical distribution and the ordering of ordinal attributes remains challenging. We propose TABTAILOR, a mixed-type hybrid diffusion framework that aims to preserve the tail and ordinal semantics of the tabular data. For numerical variables, TABTAILOR uses an adaptive semiparametric marginal transform to model each feature's tails before the fixed cosine VP diffusion process. The categorical variables employ masking budgets and learnable per feature-and-state-dependent corruption schedules. TABTAILOR also uses a cumulative semantic embedding and a strictly increasing rank coordinate to represent ordinal attributes as an embedding, called CMOE (Cumulative Monotonic Ordinal Embedding). A denoising backbone is implemented using a bidirectional transformer for modelling numerical, nominal and ordinal side information jointly, and a state-consistent mixed-type stochastic-restart sampler is used to mitigate decoding error accumulation. The denoiser and the learnable corruption schedule are trained jointly under one diffusion objective. TABTAILOR performs better than state-of-the-art baselines in seven benchmark datasets and ten metrics while maintaining overall preservation of the numerical tails and ordinal structure.

1 INTRODUCTION

The development of effective generative model for tabular data is a fundamental problem in many data processing and analysis tasks that includes data augmentation [15], data privacy protection [3], data imputation of missing values [55] which are seen in different database types. A well structured synthetic tabular data shares the same statistical characteristics as the actual data. Such synthetic data can be fully replaced in a workflow or used to enrich the data, thereby making it easier to share and use. Also, anonymizing data and increasing sample size without compromise of the overall data quality enables it as suitable candidate for real-world data science tasks. Unlike image data which consist of purely spatial pixel values and local spatial correlation, or text data, which only consist of tokens that share the same dictionary space. Tabular data has got more complex and varied distributions [8] across its features and is difficult to learn the joint probabilities across multiple columns.

Over the past few years, a large number of deep generative models for tabular data generation have been proposed, including autoregressive models and GANs [51]. Although they have significantly enhanced the quality of the generated data compared to traditional machine learning generation methods, like resampling, the quality of the generated data is still insufficient [11], especially because of the limited capacity of the model. In recent years, as diffusion models evolved rapidly such as TabDDPM [22], TabSyn [54] and TabDiff [44], and researchers are extending this domain to build more reliable diffusion frameworks for tabular data synthetsis. For instance, [54] takes various features using different techniques to map all of them to a latent continuous space and apply Gaussian diffusion to the same, whereas [22] implements a combination of the continuous and discrete diffusion processes for numerical and categorical features respectively. However, previously

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.

used approaches are caught in sub-optimal performances either because of extra encoding overhead or because of the lack of accuracy in the discrete-time diffusion modeling, and don't address the problem of distribution heterogeneity across mixed-type frameworks.

The current state-of-the-art deep generative models like TabSyn [54] and TabDiff [44] have become good at preserving distribution heterogeneity across mixed-type tabular data. However we notice two blind spots in current generative models. The first aspect is *tails of the distribution*. Most of the models compresses the tail of a numerical column distribution before the network ever sees them. TabDDPM [22] uses quantile transform for each columns and the Gaussian-copula models of the the Synthetic Data Vault [33] transform each column to a normal scale. It only knows the values observed during the training and doesn't have information about the tails. While nothing in the training objective knows whether the rare regions have been reproduced. and there is no training objective that asks whether rare regions have been reproduced. Heavy-tailed generative models exist, but for other settings like Pareto GAN [19] builds extreme-value theory into a GAN, and heavy-tailed diffusion [32] replaces the Gaussian prior by a multivariate Student-t distribution. But, neither is a mixed-type tabular diffusion model. The second aspect concerns *order*. Categorical columns are usually treated as unordered labels, so few column feature variables might end up having ordering. Therefore treating these variables as a nominal unordered attributes might hamper the pattern they preserve with the other categorical columns; models like TabSyn [54] and TabDiff [44] are good at preserving the overall mixed-type data distribution, without explicitly preserving the ordering of the ordinal feature attributes.

We present TABTAILOR a novel and hybrid mixed-tpye diffsuion framework, whose primary goal is to preserve the eligible tails and ordinal variables structure, while achieving overall better data distribution compared to the state-of-the-art models. Our key innovation is the development of a pre-processing stage,tail-loss function and an ordinal-loss function. The *Adaptive Tail-Invariant (ATI) module* which preserves the real data's own shape where evidence is plentiful and grafts a smooth extreme-value tail [5] [36]. The column is then expressed on a bell-curve scale, where we run the variance-preserving diffusion of DDPM [18] and sample with the deterministic steps of DDIM [45]. We build our categorical mask absorbing diffusion on the generalized masked diffusion of Shi et al. [43] (GenMD4, Google DeepMind), which lets the masking rate depend on the value being hidden. We make this rate learnable for every column and category value, while keeping each column's total masking budget fixed. TABTAILOR introduces a Cumulative Monotonic Ordinal Embedding (CMOE). TABTAILOR uses custom diffusion-conditioned bi-directional self-attention transformer [49] as the denoising network to jointly model both corruptions together, and the sampler can step slightly backwards, as in restart sampling [52], to reconsider decoded values.

TABTAILOR offers three main advantages: (1) our model learns about the tails of eligible numerical features; and (2) the framework uses a masked-state categorical process which respects ordinal order, while jointly modelling the numrical, ordinal and nominal under a single diffusion objective. (3) our model performs better at maintaing column-wise distribution and pair-wise column correlation compared with the state-of-the-art models.

2 RELATED WORK

Synthetic tables and diffusion. CTGAN [51], TVAE [51], GOGGLE [27] and GReaT [9] borrow adversarial, variational, graph-based and language-model tools, respectively. Diffusion models learn to generate by undoing a slow corruption: DDPM [18] removes Gaussian noise, DDIM [45] samples the same network deterministically in fewer steps, and the score-based SDE view [46] unifies them. For tables, STaSy [20] uses score-based generation with self-paced learning, CoDi [24] conditions Gaussian and categorical diffusions on each other, TabDDPM [22] pairs Gaussian and multinomial diffusion, TabSyn [54] diffuses in a learned latent space, and TabDiff [44] diffuses mixed rows in continuous time with learnable per-column noise schedules. TABTAILOR follows this direction but changes how numerical columns are prepared and how category states are corrupted, one level below a per-column schedule.

Masked diffusion. Masking values was introduced for discrete diffusion by D3PM [4] and made simple for text by MDLM [39]. Closest to ours is Simplified and Generalized Masked Diffusion [43] from Google DeepMind, which reduces the objective to weighted cross-entropies and whose generalized model, GenMD4, lets the masking schedule depend on the value being hidden. We build

on it with a schedule that is learnable per column and category value, frequency-informed, and exposure-balanced.

Heavy tails. Exceedances over a high threshold follow a generalized Pareto law (Balkema and de Haan [5]; Pickands [36]), with standard fitting tools [13] and goodness-of-fit tests [12]. Pareto GAN [19] and heavy-tailed diffusion [32] use related ideas in GANs and diffusion models but not for mixed tables, while tabular pipelines map columns through empirical or fitted marginals [33] that cannot describe unseen extremes. ATI combines an empirical body, evidence-gated tails, and the randomized distributional transform [38].

Ordered categories and restarts. Ordinal statistics asks whether a level is at least r: cumulative-link [30] and continuation-ratio models [1], binary threshold decompositions [16], and rank-consistent CORAL [10]. The closest generative work is the ordinal-aware diffusion model for disease images of Kurt et al. [23], which uses an additive severity embedding; CMOE is a follow-up for table columns. Restart sampling [52] alternates noise and denoising to repair early mistakes; we extend it to mixed rows with exact local transitions (see Appendix L for a fuller discussion).

3 Tabular Data Synthesis with TabTailor

3.1 Overview

Mixed-type tabular data. Let $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^{N}$ denote a mixed-type tabular dataset with N observations. We denote the numbers of original numerical and categorical features by M_{num} and M_{cat}, respectively. We write each row as $\mathbf{x}^{(n)} = [\mathbf{x}^{(n)}_{\mathrm{num}}, \mathbf{x}^{(n)}_{\mathrm{cat}}]$, where $\mathbf{x}^{(n)}_{\mathrm{num}} \in \mathbb{R}^{M_{\mathrm{num}}}$ and $\mathbf{x}^{(n)}_{\mathrm{cat}} \in \mathbb{R}^{M_{\mathrm{Cat}}}$. For categorical column j we fit a state $\mathcal{V}_j$ of K_j values on training rows, including a reserved *unseen* value that is never generated, and write $\mathcal{C}_j \subseteq \{0, \dots, K_j - 1\}$ for the admissible clean states, $\boldsymbol{\pi}_j$ for their smoothed training frequencies, and $\mathsf{M}_j = K_j$ for an extra absorbing *mask* state. Sections 3.2–3.5 describe ATI, CMOE with the absorbing schedule, the joint denoiser, and the restart sampler.

3.2 Adaptive Numerical Marginals and Invariant Corruption

ATI gives every numerical column its own "fit" before diffusion starts. On training rows a schema profiler assigns column j a route s_j (continuous, integer-grid, bounded, count, or a zero-inflated variant; Appendix C); the route only decides which distribution function and inverse are used, and the network is the same for every column. Genuine counts stay empirical-discrete, statistically continuous integer columns may receive a tail and are rounded after inversion, and explicit hard bounds switch tail extrapolation off on that side.

Empirical body, selected tails. ATI keeps the empirical CDF $\widehat{F}_j^{\mathrm{emp}}$ where data are dense and grafts a generalized Pareto tail onto a side only when the evidence supports it:

$$\widehat{F}_j(x) = \begin{cases} p_j^- \, \bar{G}_{\xi_j^-,\beta_j^-}\big(\vartheta_j^- - x\big), & x < \vartheta_j^-, \\ \widehat{F}_j^{\mathrm{emp}}(x), & \vartheta_j^- \le x \le \vartheta_j^+, \\ 1 - p_j^+ \, \bar{G}_{\xi_j^+,\beta_j^+}\big(x - \vartheta_j^+\big), & x > \vartheta_j^+. \end{cases} \tag{1}$$

Here $\vartheta_j^{\pm}$ are the selected thresholds, $p_j^{\pm}$ the training fraction of values beyond them, and $\bar{G}_{\xi,\beta}$ the generalized Pareto survival function with shape ξ and scale β (Appendix C); a side without a selected tail keeps the empirical CDF, and $\xi = 0$ is the exponential tail. Rather than accepting the first threshold that passes a test, ATI fits every viable candidate h (a threshold with a family) on training exceedances, keeps those that pass a refitted bootstrap goodness-of-fit test [12], and lets validation rows choose:

$$\widehat{h}_j = \operatorname*{arg\,min}_{h \in \mathcal{H}_j} \Big\{ E_{\mathrm{val}}(h) + \lambda_s \, \mathrm{Stab}(h) + \lambda_u \, \mathrm{Wid}(h) + \lambda_n \, n_h^{-1/2} + \lambda_c \mathbb{1}[h \text{ is GPD}] \Big\}, \tag{2}$$

where $\mathcal{H}_j$ is the set of accepted candidates (empty means the side stays empirical), $\mathrm{Stab}(h)$ is how much the shape estimate changes between neighbouring thresholds, $\mathrm{Wid}(h)$ the width of its bootstrap interval, n_h the number of exceedances, and the last term prefers the simpler exponential

tail. The validation term asks how well h predicts the extreme quantiles of held-out rows,

$$E_{\mathrm{val}}(h) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{\left|\widehat{Q}_h(q) - Q_j^{\mathrm{val}}(q)\right|}{s_j}, \qquad \mathcal{Q} = \{0.99,\, 0.995\}, \tag{3}$$

with $\widehat{Q}_h$ the quantile implied by the candidate tail, Q_j^{val} the held-out empirical quantile, and s_j a robust column scale (mirrored quantiles for the lower side). Validation rows never refit a marginal; they only choose among fits made on training rows (constants in Appendix C).

Zeros, ties and atoms. If a column has a structural zero atom of mass p_{0j}, ATI separates *whether* the value is zero from *how large* it is when positive: $\widehat{F}_j(0) = p_{0j}$ and $\widehat{F}_j(x) = p_{0j} + (1 - p_{0j})F_{j,+}(x)$ for $x > 0$, where the positive part $F_{j,+}$ receives its own ATI fit. Ties and atoms of any kind are handled by the randomized distributional transform [38],

$$U_j = \widehat{F}_j(X_j^-) + V\{\widehat{F}_j(X_j) - \widehat{F}_j(X_j^-)\}, \qquad V \sim \mathcal{U}(0,1), \qquad Z_{0j} = \Phi^{-1}(U_j), \tag{4}$$

which spreads a structural zero uniformly inside its atom and brings every column to a common Gaussian coordinate Z_{0j} without pretending that the units are comparable.

Invariant corruption. On this shared scale we run the variance-preserving process of DDPM [18] with a fixed cosine schedule [31] in the angular form of Salimans and Ho [41],

$$Z_{tj} = \alpha(t_{\mathrm{N}})Z_{0j} + \sigma(t_{\mathrm{N}})\epsilon_j, \qquad \epsilon_j \sim \mathcal{N}(0,1), \qquad \alpha(t) = \cos\tfrac{\pi t}{2},\ \ \sigma(t) = \sin\tfrac{\pi t}{2}. \tag{5}$$

If the fitted marginal were exact, $Z_{tj} \sim \mathcal{N}(0,1)$ at every time (Proposition 1, Appendix B): the marginal scale of each column stays put while dependence between columns is destroyed and rebuilt. We call this *invariant corruption*; it concerns marginals, not the copula, and lets the numerical schedule stay fixed because ATI absorbs the column heterogeneity.

Raw and calibrated inversion. After reverse diffusion, a raw sample keeps the generated latent magnitude, whereas a rank-calibrated view keeps only the generated order (the rth ordered value among m observed entries is sent to level $(r - \frac{1}{2})/m$):

$$\widehat{X}_{ij}^{\mathrm{raw}} = \widehat{Q}_j\Big(\Phi(\widehat{Z}_{0ij})\Big), \qquad \widehat{X}_{(r)}^{\mathrm{cal}} = \widehat{Q}_j\Big(U_{(r)}^{\mathrm{cal}}\Big), \quad U_{(r)}^{\mathrm{cal}} = \frac{r - \frac{1}{2}}{m}, \quad \widehat{Q}_j = \widehat{F}_j^{-1}. \tag{6}$$

Both come from the same latent table. Calibration discards magnitude, so good calibrated tails show that the fitted marginal is reconstructed, not that tails were learned; only agreement of RAW samples with training and held-out tails shows that the reverse model reaches them.

3.3 Ordinal Structure and Calibrated Categorical Exposure

CMOE: a ladder the network can feel. For an ordinal column j with ordered levels $r \in \{0, \ldots, L_j - 1\}$, the input embedding of level r is

$$e_{jr} = \Big[\, b_j + \sum_{h=1}^{r} \Delta_{jh} \,;\, \gamma_j\, q_{jr} \Big], \qquad q_{jr} = \frac{\sum_{h=1}^{r} d_{jh}}{\sum_{h=1}^{L_j - 1} d_{jh}}, \tag{7}$$

with $d_{jh} > 0$ and $\gamma_j > 0$ enforced by positive transforms. Each level is the previous level plus a learned step Δ_{jh}, the additive idea of Kurt et al. [23], so higher levels "contain" lower ones; one extra coordinate $q_{jr} \in [0,1]$ is guaranteed to increase with the level while the others may learn any semantics. To read the ladder back, a group head first decides whether the clean value is in the ordered group ($\mathrm{G}_j = \mathrm{ord}$) or a special state; given the ordered group, continuation probabilities $\nu_{jh} \in (0,1)$ ("having reached level $h-1$, does it reach level h?") define $\Gamma_{j0} = 1$, $\Gamma_{jr} = \prod_{h \le r} \nu_{jh}$ and $\Gamma_{jL_j} = 0$, and the posterior over clean states is

$$\mu_j(k) = \begin{cases} P_\theta(\mathrm{G}_j = \mathrm{ord} \mid \mathbf{f}_j)\,\big(\Gamma_{j,r(k)} - \Gamma_{j,r(k)+1}\big), & k \text{ ordered, with rank } r(k), \\ P_\theta(\mathrm{G}_j = k \mid \mathbf{f}_j), & k \text{ a special state}, \end{cases} \tag{8}$$

with $\mathbf{f}_j$ the transformer output for column j; special values get no artificial rank.

State-dependent masking. Following the state-dependent schedules of GenMD4 [43], a clean state k of column j is absorbed by the mask at categorical time t_{C} with probability

$$\mathbb{P}(C_{tj} = \mathsf{M} \mid C_{0j} = k) = t_{\mathrm{C}}^{\,w_{jk}}, \qquad \mathbb{P}(C_{tj} = k \mid C_{0j} = k) = 1 - t_{\mathrm{C}}^{\,w_{jk}}. \tag{9}$$

The exponent w_{jk} is where TABTAILOR departs from GenMD4; it is learnable and built from three ingredients,

$$w_{jk} = w_{\min} + (w_{\max} - w_{\min}) \operatorname{sigmoid}\Big(c_j - \eta\, \widetilde{\varrho}_{jk} + u_{jk}\Big). \tag{10}$$

Here $\widetilde{\varrho}_{jk}$ is the *rarity* of state k (a bounded increasing function of its smoothed surprisal $-\log \widetilde{\pi}_{jk}$, centered to zero average over the column); $u_{jk} = \mathrm{rel}_{jk}\, u_{\max} \tanh(\theta_{jk})$ is a learned per-state residual multiplied by a *reliability* $\mathrm{rel}_{jk} = n_{jk}/(n_{jk} + n_0)$ that grows with the training count n_{jk}, so a handful of examples cannot drag the schedule around; and c_j is a per-column intercept fixed by the balance condition below. Rare states get smaller exponents and are hidden more often, which gives the network more practice reconstructing them. To keep the column as a whole from becoming harder, c_j is chosen so that the frequency-weighted masking exposure equals that of a learned column-level baseline exponent $\bar{w}_j$:

$$\sum_k \pi_{jk}\, \Psi(w_{jk}) = \Psi(\bar{w}_j), \qquad \Psi(w) = \mathbb{E}_{T\sim\mathcal{U}(\delta_{\mathrm{C}},1)}\big[T^w\big] = \frac{1-\delta_{\mathrm{C}}^{w+1}}{(1-\delta_{\mathrm{C}})(w+1)}. \tag{11}$$

This *reallocates* exposure from common to rare states; it creates no information absent from the training set.

3.4 COUPLED DENOISER AND LEARNING OBJECTIVE

Decoupled corruption, coupled inference. Numbers and categories are corrupted by different processes with independent clocks, $t_{\mathrm{N}} \sim \mathcal{U}(\delta_{\mathrm{N}}, 1-\delta_{\mathrm{N}})$ and $t_{\mathrm{C}} \sim \mathcal{U}(\delta_{\mathrm{C}}, 1)$:

$$q(Z_{t_{\mathrm{N}}}, C_{t_{\mathrm{C}}} \mid Z_0, C_0) = q_{\mathrm{N}}(Z_{t_{\mathrm{N}}} \mid Z_0, t_{\mathrm{N}})\, q_{\mathrm{C}}(C_{t_{\mathrm{C}}} \mid C_0, t_{\mathrm{C}}), \tag{12}$$

$$p(Z_0, C_0 \mid Z_{t_{\mathrm{N}}}, C_{t_{\mathrm{C}}}) \propto p_{\mathrm{data}}(Z_0, C_0)\, q_{\mathrm{N}}\, q_{\mathrm{C}}. \tag{13}$$

The factorized forward law does not make the columns independent: since the clean row follows a dependent law p_{data}, the reverse posterior generally does not factor. Corruption is type specific but reverse inference is joint, the central design of TABTAILOR. We therefore denoise with one bidirectional transformer [49] with pre-normalization [50]: each column is a token and no causal mask is used, so numerical and categorical tokens attend to each other. Tokens combine the current (noisy, masked, nominal or CMOE) state, column and type identity, missingness, ATI metadata and the clocks, and blocks modulate normalized activations by a per-token scale and shift computed from the clocks, as in FiLM [35] and DiT [34]. A linear head per numerical column predicts the velocity and a nominal or ordinal head per categorical column predicts μ_j (Appendix E).

Numerical objective: count rare rows more. The network predicts the velocity $v = \alpha(t_{\mathrm{N}})\epsilon - \sigma(t_{\mathrm{N}})Z_0$ [41]; L_{base} is its mean squared error over observed columns. Rarity is measured on the *clean* coordinate, never on the noisy input, so noise cannot turn a central value into a false tail example: $p_j^{\mathrm{edge}} = \min\{\Phi(Z_{0j}), 1-\Phi(Z_{0j})\}$. We turn it into a weight, dampen the extra emphasis when the input carries little signal, and renormalize within each column:

$$\omega_j^{\mathrm{raw}} = 1 + (W_{\max}-1)\,\mathrm{clip}\Big(\tfrac{\log(p_0/p_j^{\mathrm{edge}})}{\log(p_0/p_1)}, 0, 1\Big)^{\psi}, \quad g(t_{\mathrm{N}}) = g_0 + (1-g_0)\tfrac{\mathrm{SNR}}{1+\mathrm{SNR}},$$

$$\hat{\omega}_j = 1 + g(t_{\mathrm{N}})\big(\omega_j^{\mathrm{raw}} - 1\big), \qquad \omega_j = \hat{\omega}_j \Big/ \overline{\hat{\omega}}_j, \tag{14}$$

with $\mathrm{SNR} = \alpha(t_{\mathrm{N}})^2/\sigma(t_{\mathrm{N}})^2$ and $\overline{\hat{\omega}}_j$ the mini-batch mean of $\hat{\omega}_j$ over observed entries of column j, so weights redistribute emphasis inside a column without changing its scale. Then

$$L_{\mathrm{tail}} = \frac{\sum_j o_j\, \omega_j\, (\hat{v}_j - v_j)^2}{\max\Big(1, \sum_j o_j\Big)}, \qquad L_{\mathrm{num}} = (1-\lambda_T)L_{\mathrm{base}} + \lambda_T L_{\mathrm{tail}}, \tag{15}$$

where $o_j = 1$ if column j is observed. Observations in the extreme p_0–p_1 range of the clean marginal count up to $W_{\max}$ times more when signal is present; there is no separate tail diffusion process.

Categorical and ordinal objectives. Let $\mu_j(k) = p_\theta(C_{0j} = k \mid Z_{t_{\mathrm{N}}}, C_{t_{\mathrm{C}}}, t_{\mathrm{N}}, t_{\mathrm{C}})$. For a masked column with true clean state k, the GenMD4 bound [43] gives

$$\ell_j(\mu_j, k) = \sum_h w_{jh}\mu_j(h) - w_{jk}\{1 + \log \mu_j(k)\}. \tag{16}$$

Integrating over the clock contributes a factor $1/t_{\mathrm{C}}$ (for $1 - t^w$ the ratio $\alpha'/(1-\alpha)$ equals $-w/t$), so the Monte Carlo objective (Appendix B) is

$$L_{\mathrm{cat}} = \frac{1}{M_{\mathrm{cat}}} \frac{1-\delta_{\mathrm{C}}}{t_{\mathrm{C}}} \sum_{j:\, C_{t_{\mathrm{C}}j}=\mathsf{M}} \ell_j\big(\mu_j, C_{0j}\big). \tag{17}$$

Ordinal columns add a threshold loss that turns the ladder into "at least r?" questions: L_{ord} averages, over the masked ordinal columns $j \in \mathcal{O}$ with an ordered clean value and over thresholds $r = 1, \ldots, L_j - 1$, the binary cross-entropy between $\mathbb{1}[R_j \geq r]$ and $P_\theta(R_j \geq r \mid \mathrm{G}_j = \mathrm{ord})$. The row objective is

$$L_{\mathrm{row}} = L_{\mathrm{num}} + \lambda_C L_{\mathrm{cat}} + \lambda_O L_{\mathrm{ord}} + \lambda_{\mathrm{term}} L_{\mathrm{term}} + L_{\mathrm{reg}}, \tag{18}$$

where L_{term} is the categorical loss at $t_{\mathrm{C}} = 1$ with every column masked and pure-noise numerics (anchoring the start of sampling) and L_{reg} collects schedule regularizers (residual size, ordinal smoothness of exponents, information leakage, baseline drift). The exponent w_{jk} affects the loss directly through ℓ_j and indirectly because it decides *which* entries are masked; for the latter we draw two masks per row and use a leave-one-out score-function estimator [21], from which numerical loss terms are deliberately excluded so that the numerical branch does not steer the categorical corruption law (Appendix E, Algorithm 1).

3.5 Restart and Posterior-Budgeted Sampling

Sampling starts from Gaussian numerical noise and fully masked categories; numerical values follow a fixed decreasing grid while each row carries its own categorical clock. Before each denoising update the sampler moves both kinds of column slightly *back* toward noise and lets the shared denoiser reconstruct from that state [52], so decoded values can change, and each improved categorical state becomes context for the next numerical step and vice versa.

Numerical restart. At numerical time t with next scheduled time $s < t$ we re-noise by a small fraction ϕ_{N} of what an ordinary step would remove: the restart variance is $\nu_{\mathrm{R}} = \min\{\nu_{\max}, \phi_{\mathrm{N}}(1 - \alpha(t)^2/\alpha(s)^2)\}$ and the restart time $u \geq t$ solves $\alpha(u)^2 = (1-\nu_{\mathrm{R}})\alpha(t)^2$. The exact VP Markov law gives

$$Z_u = \rho Z_t + \sqrt{1-\rho^2}\,\zeta, \qquad \rho = \frac{\alpha(u)}{\alpha(t)}, \qquad \zeta \sim \mathcal{N}(0, I). \tag{19}$$

After one denoiser call at u we recover $\widehat{Z}_0 = \alpha(u) Z_u - \sigma(u)\hat{v}$ and $\hat{\epsilon} = \sigma(u) Z_u + \alpha(u)\hat{v}$ and return to s with the deterministic DDIM step [45],

$$Z_s = \alpha(s)\widehat{Z}_0 + \sigma(s)\hat{\epsilon}, \qquad s < t. \tag{20}$$

Categorical restart. With clean survival $a_{jk}(t) = 1 - t^{w_{jk}}$, moving a decoded state from categorical time t to a noisier $u \geq t$ re-masks it with the exact conditional probability $1 - a_{jk}(u)/a_{jk}(t)$ (Eq. 21 below), and existing masks stay masked. We choose u so that the expected number of fresh re-masks under the fitted forward marginal is a small budget $b_{\mathrm{R}} = \min\{b_{\max}, \phi_{\mathrm{C}} M_{\mathrm{cat}}\}$:

$$\mathbb{P}(\text{re-mask} \mid \text{clean at } t) = 1 - \frac{a_{jk}(u)}{a_{jk}(t)}, \tag{21}$$

$$\sum_j \sum_k \pi_{jk}\big(u^{w_{jk}} - t^{w_{jk}}\big) = b_{\mathrm{R}}, \tag{22}$$

so the restart never depends on which values happen to have been decoded. Given the restarted row and the shared posterior μ_j, a masked field moved back from u to $s \leq u$ follows

$$\mathbb{P}(C_{sj} = k \mid C_{uj} = \mathsf{M}) = \mu_j(k)\,[1 - (s/u)^{w_{jk}}], \tag{23}$$

$$\mathbb{P}(C_{sj} = \mathsf{M} \mid C_{uj} = \mathsf{M}) = \sum_k \mu_j(k)(s/u)^{w_{jk}}. \tag{24}$$

Posterior-budgeted clock. Categorical progress is not forced to share the numerical grid. With $\mathcal{A}$ the currently masked fields, the posterior-expected number of reveals when the clock moves from t to a candidate s is

$$B(s;t) = \sum_{j \in \mathcal{A}} \sum_k \mu_j(k)\,[1 - (s/t)^{w_{jk}}]. \tag{25}$$

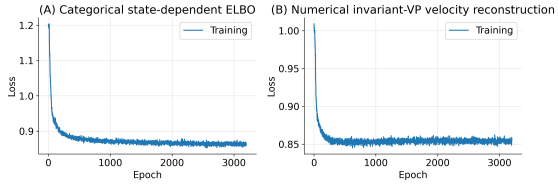


Figure 1: Adult training loss per epoch: (A) categorical state-dependent ELBO; (B) numerical invariant-VP velocity reconstruction.

The sampler picks the next clock s_{C} by solving the monotone equation $B(s_{\mathrm{C}}; u_{\mathrm{C}}) = B^{\star}$ with target

$$B^{\star} = \min\Big\{|\mathcal{A}|,\ \max\big\{\max\{f(\kappa)\,|\mathcal{A}|,\ b_{\min}\} + R,\ \ B(t_{\mathrm{C}}; u_{\mathrm{C}})\big\}\Big\}, \tag{26}$$
$$f(\kappa) = f_{\min} + (f_{\max} - f_{\min})\,\kappa^{\psi_{\mathrm{C}}}.$$

Here $\kappa = \alpha(s_{\mathrm{N}})^2$ is the numerical cleanliness of the next numerical state, R the number of fields the restart just re-masked, and $B(t_{\mathrm{C}}; u_{\mathrm{C}})$ the expected reveals needed merely to return to the pre-restart clock: reveal more categories as the numbers become trustworthy, pay back what the restart took away, and never move the clock backwards. A final rejuvenation pass re-masks a small state-consistent subset once the numerics are complete and updates it once more (Algorithm 2, Appendix E)

4 Evaluation and Results

4.1 Experimental Setup

Data and baselines. We follow the seven benchmark datasets and historical baseline family of TabDiff [44]: Adult [6], Default [53], Shoppers [40], MAGIC [7], Beijing PM2.5 [25], Online News Popularity [14], and Diabetes 130-US Hospitals [47], and add Ann-Thyroid from the UCI Thyroid Disease collection [37], whose anomaly labels are used only for evaluation, as in ADBench [17] (splits in Appendix G). Baselines are CTGAN [51], TVAE [51], GOGGLE [27], GReaT [9], STaSy [20], CoDi [24], TabDDPM [22], TabSyn [54], and TabDiff [44]. Historical numbers are transcribed from TabDiff (Appendix I), not rerun, and cells without a matched experiment stay blank.

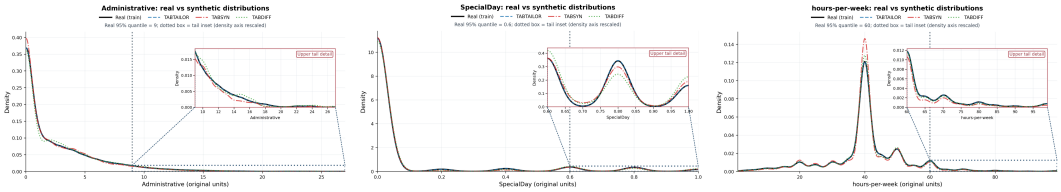
Metrics. Fidelity follows Shape, Trend, α-Precision [2], β-Recall [2] and C2ST [28]; utility is machine-learning efficiency (MLE; AUC or RMSE); privacy is distance to the closest record (DCR), a memorization diagnostic rather than a guarantee [2]; and we add Tail-AD, Tail-JSD and ordinal rank Wasserstein (Section 4.2). Repeated tables come from the best validation checkpoint, so standard deviations quantify generation variability, not training-seed uncertainty.

Training. Implementation details are in Appendix H. Figure 1 shows the Adult training curves: both objectives fall steeply in the first few hundred epochs, then the numerical loss settles while the categorical bound drifts down slowly.

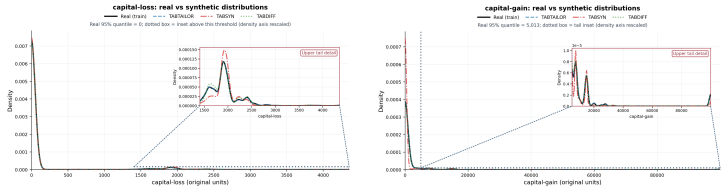
4.2 Tail Fidelity and Ordinal Semantics

Tail-AD applies a two-sample Anderson–Darling statistic [42] to real-defined lower and upper exceedance samples and measures conditional shape inside the tail. **Tail-JSD** splits the line at fixed TRAIN quantiles, with finer tail bins, and takes the Jensen–Shannon divergence [26] between real and synthetic bin masses, reflecting both how much mass enters the tails and where. **Ordinal Rank Wasserstein** compares empirical CDFs over normalized ranks of the ordered subgroup, so a one-step error costs less than moving mass across the whole order (all lower-is-better; Appendix F). We keep four views apart, TRAIN–RAW, TEST–RAW and their calibrated counterparts, because rank calibration repairs one-dimensional marginals and cannot alone show that the denoiser learned extreme magnitudes.

Qualitative tail preservation. Figure 2 overlays the real training density with the synthetic densities of TabTailor, TabSyn [54] and TabDiff [44]; insets zoom into the region beyond the real 95%

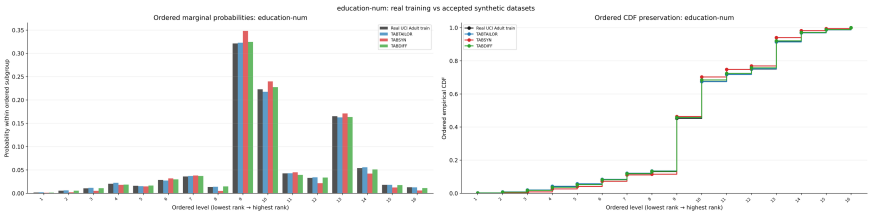


(a) Administrative (Shoppers) **(b)** SpecialDay (Shoppers) **(c)** hours-per-week (Adult)

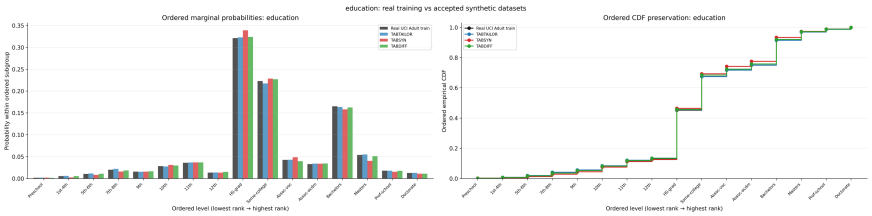


(d) capital-loss (Adult) **(e)** capital-gain (Adult)

Figure 2: Tail preservation on Shoppers and Adult. Real training density (black) against TABTAILOR (blue dashed), TabSyn (red dash-dotted) and TabDiff (green dotted); the boxed inset, with rescaled density axis, zooms into the region above the real 95% quantile.



(a) education-num (integer code, levels 1–16)

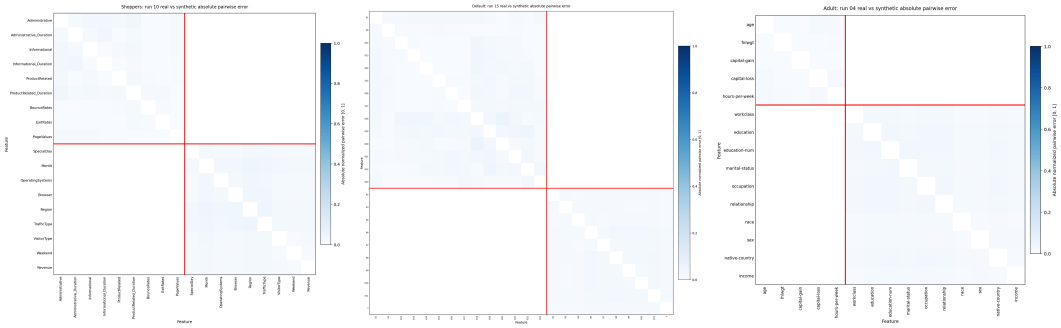


(b) education (text label, Preschool to Doctorate)

Figure 3: Ordinal preservation on Adult. Left of each panel: probability of each ordered level; right: ordered empirical CDF. Real training data (black) against TABTAILOR, TabSyn and TabDiff.

quantile. In *Administrative*, TabSyn falls below the real curve between roughly 13 and 17 pages and TabDiff rises above it around 15. In *SpecialDay*, whose values sit on the grid $\{0, 0.2, \ldots, 1\}$, both baselines flatten the peak at 0.8 and overstate the mass at 1.0. In Adult, TabSyn overshoots the 40-hour mode of *hours-per-week*, undershoots its 60–99 hour tail, and misplaces mass in *capital-gain* and *capital-loss*. The TABTAILOR curve is almost indistinguishable from the real one in every panel, including the small far-right bump of *capital-gain* near 10^5. Closeness to the training density is what a good fit produces but also what memorization would produce, so these panels are read together with held-out (TEST) tail statistics and DCR (Section 4.3); the RAW-versus-calibrated version is reserved for Appendix K.

Qualitative ordinal preservation. Figure 3 shows ordered-level probabilities and ordered CDFs of the Adult education column, stored twice in Adult as an integer code (*education-num*) and as a text label (*education*); TABTAILOR treats both as ordinal. TabSyn puts too much mass on HS-grad and too little on Masters and Doctorate, so its CDF sits above the real one from HS-grad upward (mass moved to lower levels); TabDiff and TABTAILOR are much closer to the real CDF, TABTAILOR nearly on top of it.



(a) Shoppers (run 10) **(b)** Default (run 15) **(c)** Adult (run 04)

Figure 4: Absolute pairwise error between real and synthetic TABTAILOR tables (scale $[0, 1]$, darker is larger). Numerical columns form the upper-left block and categorical columns the lower-right block.

4.3 DATA FIDELITY AND PRIVACY

Shape and trend. Shape averages single-column disagreement (KS-like for numerical, total-variation for categorical columns) and Trend averages pairwise mismatch; historical TabDiff errors [44] are in Tables 4 and 5 (Appendix I), with the Thyroid and TABTAILOR entries open until matched runs exist. Figure 4 shows the pairwise error behind Trend for one synthetic table per dataset: half the absolute difference of real and synthetic Pearson correlations for two numerical columns, and the total-variation distance between contingency tables for two categorical columns, both in $[0, 1]$; mixed pairs are undefined and stay blank. On Shoppers, Default and Adult every defined cell falls in the palest part of the scale, so no column pair is visibly mishandled (the slightly darker cells of Default lie among its numerical columns X18–X20). These are single tables (runs 10, 15 and 04 of 20), so they show typical rather than averaged behaviour.

Privacy. DCR compares each synthetic record with its nearest real training and reference record. A balanced outcome is a useful memorization diagnostic but not differential privacy, so DCR is read jointly with tail fidelity: better rare-event reproduction is not privacy.

4.4 UTILITY AND ABLATION

MLE tests whether a predictor trained on synthetic data transfers to held-out real data; Table 6 (Appendix I) reproduces the historical TabDiff comparison and reserves a row for TABTAILOR. Ann-Thyroid is an unsupervised anomaly benchmark (generator fitted without labels; detector trained on generated rows and scored on held-out real data by AUROC, AUPRC and F1); no matched baseline runs are reported. The ablation protocol toggles one component at a time (no restart, shared clock, empirical-only ATI, no clean-tail weighting, nominal ordinal branch, feature-only categorical schedule) and reports Shape and Trend together with Tail-AD, Tail-JSD and ORW, so that averages cannot improve while targeted structures degrade.

5 CONCLUSION

We presented TABTAILOR, a mixed-type hybrid diffusion framework built around numerical tails and ordinal semantics: ATI maps training-only marginals to a common Gaussian coordinate and separates raw generation from rank calibration; the categorical branch pairs a calibrated state-dependent absorbing process with CMOE; and one bidirectional transformer with restart and posterior-budgeted sampling estimates the joint reverse quantities. The remaining task is to run matched seeds and the full benchmark, fill the reserved TABTAILOR cells, and report generation and training variability.

REFERENCES

[1] Alan Agresti. *Analysis of Ordinal Categorical Data*. Wiley, 2 edition, 2010.

[2] 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*, 2022.

[3] Samuel A Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *Proceedings of the first ACM international conference on AI in finance*, pp. 1–8, 2020.

[4] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[5] August A. Balkema and Laurens de Haan. Residual life time at great age. *The Annals of Probability*, 2(5):792–804, 1974.

[6] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[7] R. K. Bock, A. Chilingarian, M. Gaug, F. Hakl, Th. Hengstebeck, M. Jiřina, J. Klaschka, E. Kotrč, P. Savický, S. Towers, A. Vaiciulis, and W. Wittek. Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. *Nuclear Instruments and Methods in Physics Research Section A*, 516(2–3):511–528, 2004.

[8] Vadim Borisov, Kathrin Seßler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. *arXiv preprint arXiv:2210.06280*, 2022.

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

[10] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

[11] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[12] Vartan Choulakian and Michael A. Stephens. Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, 43(4):478–484, 2001.

[13] Stuart Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[14] Kelwin Fernandes, Pedro Vinagre, and Paulo Cortez. A proactive intelligent decision support system for predicting the popularity of online news. In *Portuguese Conference on Artificial Intelligence*, pp. 535–546, 2015.

[15] Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.

[16] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pp. 145–156, 2001.

[17] Songqiao Han, Xiyang Hu, Hailiang Huang, Minqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[18] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

[19] Todd Huster, Jeremy E. J. Cohen, Zinan Lin, Kevin Chan, Charles Kamhoua, Nandi Leslie, Cho-Yu Jason Chiang, and Vyas Sekar. Pareto GAN: Extending the representational power of GANs to heavy-tailed distributions. In *International Conference on Machine Learning*, 2021.

[20] Jayoung Kim, Chaejeong Lee, and Noseong Park. Stasy: Score-based tabular data synthesis. In *International Conference on Learning Representations*, 2023.

[21] Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019.

[22] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, 2023.

[23] Meryem Mine Kurt, Ümit Mert Çağlar, and Alptekin Temizel. Progressive disease image generation with ordinal-aware diffusion models. *Diagnostics*, 15(20):2558, 2025. doi: 10.3390/diagnostics15202558.

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

[25] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of the Royal Society A*, 471(2182):20150257, 2015.

[26] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

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

[28] David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[29] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *International Conference on Machine Learning*, 2024.

[30] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B*, 42(2):109–142, 1980.

[31] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021.

[32] Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *International Conference on Learning Representations*, 2025.

[33] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. The synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics*, pp. 399–410, 2016.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[35] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018.

[36] James Pickands. Statistical inference using extreme order statistics. *The Annals of Statistics*, 3 (1):119–131, 1975.

[37] Ross Quinlan. Thyroid disease. UCI Machine Learning Repository, 1987.

[38] Ludger Rüschendorf. On the distributional transform, Sklar's theorem, and the empirical copula process. *Journal of Statistical Planning and Inference*, 139(11):3921–3927, 2009.

[39] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[40] C. Okan Sakar, S. Olcay Polat, Mete Katircioglu, and Yomi Kastro. Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10):6893–6908, 2019.

[41] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

[42] Fritz W. Scholz and Michael A. Stephens. K-sample anderson-darling tests. *Journal of the American Statistical Association*, 82(399):918–924, 1987.

[43] Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

[44] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. Tabdiff: A mixed-type diffusion model for tabular data generation. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.20626.

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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

[47] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.

[48] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[50] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 2020.

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

[52] Yilun Xu, Mingyang Deng, Xiang Cheng, Yonglong Tian, Ziming Liu, and Tommi Jaakkola. Restart sampling for improving generative processes. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

[53] 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.

[54] 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 *International Conference on Learning Representations*, volume 2024, pp. 52829–52857, 2024.

[55] Shuhan Zheng and Nontawat Charoenphakdee. Diffusion models for missing value imputation in tabular data. *arXiv preprint arXiv:2210.17128*, 2022.

A Appendix

A.1 Notation

Table 1: Core notation used throughout the paper.

Symbol	Meaning
$\mathbf{x} = [\mathbf{x}_{\text{num}}, \mathbf{x}_{\text{cat}}]$	One mixed-type table row; M_{num} numerical and M_{cat} categorical columns (the latter includes numerical-missingness columns).
$X_j,\ \widehat{F}_j,\ \widehat{Q}_j$	Original numerical field, fitted training CDF, and generalized inverse.
$\vartheta_j^{\pm},\ p_j^{\pm},\ \xi,\ \beta$	Selected lower/upper tail thresholds, training fraction beyond them, generalized Pareto shape and scale.
$\mathcal{H}_j,\ E_{\text{val}},\ s_j$	Accepted tail candidates, held-out tail-quantile error, robust scale of the column.
$U_j,\ Z_{0j}$	Randomized probability coordinate and clean ATI Gaussian coordinate.
$Z_t;\ \alpha(t), \sigma(t)$	Numerical latent after VP corruption; signal and noise coefficients with $\alpha^2 + \sigma^2 = 1$.
$t_{\text{N}}, t_{\text{C}};\ \delta_{\text{N}}, \delta_{\text{C}}$	Numerical and categorical clocks and their lower cut-offs.
$\mathcal{V}_j,\ K_j,\ \mathcal{C}_j,\ \mathsf{M}$	Vocabulary, its size, admissible clean states, and the absorbing mask state.
C_{tj}	Categorical state of column j at categorical time t.
$w_{jk},\ \bar{w}_j,\ c_j$	Masking exponent of state k in column j, column-level baseline exponent, and balancing intercept.
$\Psi(w)$	Mean masking exposure $\mathbb{E}_{T\sim\mathcal{U}(\delta_{\text{C}},1)}[T^w]$ of exponent w.
$\pi_{jk},\ n_{jk}$	Smoothed training frequency and raw training count of state k.
$\widetilde{\varrho}_{jk},\ \text{rel}_{jk},\ u_{jk}$	Centered rarity, reliability gate, and gated learned residual of state k.
$\mu_j(k)$	Denoiser posterior probability that the clean categorical state is k.
$R_j,\ L_j$	Ordinal rank and number of ordered levels of column j.
$e_{jr},\ \nu_{jh},\ \Gamma_{jr}$	CMOE embedding of rank r; continuation probability; cumulative survival $\prod_{h\le r}\nu_{jh}$.
$v,\ \hat{v}$	VP velocity target and transformer prediction.
$p_j^{\text{edge}},\ o_j,\ \omega_j$	Clean edge probability, observed-entry indicator, and normalized tail weight.
$L_{\text{base}}, L_{\text{tail}}, L_{\text{num}}$	Base velocity loss, clean-PIT tail-weighted velocity loss, and blended numerical loss.
$L_{\text{cat}}, L_{\text{ord}}, L_{\text{term}}$	Categorical absorbing-diffusion loss, ordinal threshold loss, terminal all-mask loss.
$u,\ \rho,\ \zeta$	Restart time, restart correlation $\alpha(u)/\alpha(t)$, and restart noise.
$\mathcal{A},\ B(s;t),\ B^\star,\ \kappa$	Currently masked fields, posterior-expected reveals, reveal target, numerical cleanliness $\alpha(s_{\text{N}})^2$.
Tail-AD, Tail-JSD, ORW	Tail Anderson–Darling, tail-focused Jensen–Shannon, and ordinal rank Wasserstein metrics.

B Analytical Properties of TabTailor

This appendix moves the structural analysis out of the main paper. The statements below describe the mathematical construction under idealized assumptions; they are not empirical performance claims.

B.1 MARGINAL INVARIANCE UNDER ATI

Proposition 1 (ATI marginal invariance). *If $X_j \sim F_j$, the randomized distributional transform is exact, and $Z_{0j} = \Phi^{-1}(U_j)$, then $Z_{0j} \sim \mathcal{N}(0,1)$. Under Eq. (5), $Z_{tj} \sim \mathcal{N}(0,1)$ for every t.*

Proof. The randomized distributional transform makes $U_j \sim \mathcal{U}(0,1)$ even in the presence of atoms. Hence $\Phi^{-1}(U_j)$ is standard normal. If Z_0 and ϵ are independent standard normals, $\alpha Z_0 + \sigma\epsilon$ is normal with variance $\alpha^2 + \sigma^2 = 1$. □

This proposition is coordinate-wise. It neither asserts independence among numerical columns nor preservation of the clean copula under noising. Finite-data fitting, probability clipping, rounding, and support corrections create small departures from the idealized statement.

B.2 RANK CALIBRATION AND WHY IT IS NOT TAIL-LEARNING EVIDENCE

Proposition 2 (Midpoint quantile grid). *Let $U_{(r)} = (r - 1/2)/m$, $r = 1, \ldots, m$, and $X_{(r)} = Q(U_{(r)})$ for an exact generalized inverse Q of CDF F. The empirical CDF G_m of $\{X_{(r)}\}$ approximates F with quantile-grid error bounded by $1/(2m)$ away from generalized-inverse ambiguities.*

The construction depends only on generated order, not generated magnitude. Consequently two latent columns with identical rankings but very different extrema produce the same calibrated sorted marginal. This is why TABTAILOR treats Eq. (6) as evidence about learned tail reach and the calibrated form in the same equation as a marginal-calibration diagnostic.

B.3 UNIQUE INTEGRATED-EXPOSURE CALIBRATION

For positive prior weights π_{jk} and finite state offsets, $\Psi(w)$ is strictly decreasing in w. The sigmoid parameterization in Eq. (10) makes every w_{jk} strictly increasing in the intercept c_j. Therefore the prior-weighted left side of Eq. (11) is continuous and strictly decreasing in c_j, giving a unique root whenever the target baseline exposure $\Psi(\bar{w}_j)$ lies inside the attainable range. The implementation finds the root by bisection and propagates gradients through it with the implicit-function theorem, $\mathrm{d}c_j/\mathrm{d}\phi = -(\partial F/\partial\phi)/(\partial F/\partial c_j)$ for any parameter ϕ of the balance function F, away from numerical saturation.

B.4 ORIGIN OF THE $1/t_C$ WEIGHT AND OF THE RESTART BUDGET

For the schedule $\alpha_{jk}(t) = 1 - t^{w_{jk}}$ the ratio $\alpha'_{jk}(t)/(1-\alpha_{jk}(t))$ equals $-w_{jk}/t$, so the continuous-time bound of GenMD4 [43] is an integral of Eq. (16) against $\mathrm{d}t/t$ over $[\delta_C, 1]$. With $t_C \sim \mathcal{U}(\delta_C, 1)$ an unbiased Monte Carlo estimate multiplies each masked term by $(1 - \delta_C)/t_C$, which is Eq. (17). For the restart budget of Eq. (22), under the fitted forward marginal a column is clean at time t in state k with probability $\pi_{jk}(1 - t^{w_{jk}})$ and is then re-masked by time u with probability $(u^{w_{jk}} - t^{w_{jk}})/(1 - t^{w_{jk}})$, so the expected number of fresh re-masks is $\sum_{j,k} \pi_{jk}(u^{w_{jk}} - t^{w_{jk}})$.

B.5 PROPER CATEGORICAL AND ORDINAL TARGETS

For a fixed corrupted context and true clean categorical posterior p, the population excess risk of Eq. (16) can be written as a positive weighted sum of Bregman-like terms,

$$R(\mu) - R(p) = \sum_k w_k \left[\mu_k - p_k + p_k \log \frac{p_k}{\mu_k}\right] \geq 0, \tag{27}$$

with equality at $\mu = p$. Conditional ordinal threshold cross-entropy is likewise proper for each event $\{R \geq r\}$, and the continuation-ratio construction yields a normalized distribution over ranks.

B.6 EXACT RESTART KERNELS VERSUS APPROXIMATE JOINT SAMPLING

Equations (19) and (21) are exact forward conditional transitions for their respective processes. Equation (24) is exact for a single field if μ_j is the exact clean posterior. The implementation samples multiple categorical fields conditionally independently given a shared transformer context

and uses finite numerical steps, so these local facts do not imply an exact finite-step joint reverse sampler.

C ATI Details: Schema Routing, Marginal Fitting, and Tail Selection

C.1 Schema routes

The profiler is fitted on TRAIN only. Each field is assigned a `FeatureSpec` containing branch, semantic type, confidence/reason, observed and unique counts, integer-likeness, zero mass, explicit bounds and provenance, integer-restoration flag, zero-inflation flag, tail eligibility, review flag, and optional ordered values. Numerical semantic routes in the current implementation are: continuous, integer-grid continuous, bounded quantitative, count, bounded count, zero-inflated continuous, and zero-inflated integer-grid. Categorical routes are ordinal, binary, and nominal.

The generalized Pareto survival function used throughout Section 3.2 is

$$\bar{G}_{\xi,\beta}(y) = \begin{cases} (1+\xi y/\beta)^{-1/\xi}, & \xi \neq 0, \\ \exp(-y/\beta), & \xi = 0, \end{cases} \qquad y \geq 0, \quad \beta > 0,\ 1+\xi y/\beta > 0, \tag{28}$$

the limit law of threshold exceedances established by Balkema and de Haan [5] and Pickands [36]. Estimation and diagnostics follow standard practice [13].

The routing principle is semantic rather than dtype-based. An integer-grid continuous field may use an empirical body and eligible GPD/exponential tail, then be rounded after inversion. Counts and bounded counts are conservatively empirical-discrete. Known hard bounds are enforced only when explicitly supplied or justified by schema metadata. Zero-inflated fields split off the zero atom and fit a separate positive ATI component.

The automatic rules use the following thresholds, all computed on TRAIN only and overridable by explicit metadata. A column is *zero-inflated* if its non-zero values are all positive, its zero mass is at least 5%, and it has at least 40 non-zero observations with at least 12 distinct non-zero values. Otherwise an integer-valued column is a *count* if it is non-negative, its name carries a count hint, it has at most 96 distinct values and at least 20% of the integers between its minimum and maximum occur; it is *bounded quantitative* if explicit support information is supplied; it is *integer-grid continuous* if it has at least 16 distinct values; and it is *bounded count* otherwise. Non-integer columns are continuous, or bounded quantitative if bounds are supplied. Only continuous, integer-grid, bounded-quantitative and zero-inflated routes are tail-eligible. Top-level branch assignment sends non-numeric columns and columns with at most 15 distinct values to the categorical branch unless dataset metadata says otherwise, and every decision carries a confidence and a reason, with decisions below confidence 0.75 flagged for review.

C.2 Tail-candidate diagnostics

For each eligible side ATI considers a fixed set of training quantile thresholds. Candidate exceedances must satisfy minimum sample and distinctness requirements. Generalized Pareto candidates are estimated by likelihood; their goodness of fit is assessed using a parametric bootstrap in which synthetic exceedance samples are generated from the fitted candidate, refitted, and compared with the observed KS discrepancy. The shape parameter ξ receives a bootstrap uncertainty interval. Candidates with excessive $|\xi|$, excessive interval width, or unstable neighboring-threshold estimates are rejected. An exponential candidate is included as the $\xi = 0$ alternative. A validation split contributes only the predictive tail-quantile term in Eq. (2); it never refits the training marginal.

The candidate thresholds are the training quantiles 0.80, 0.85, 0.90, 0.92, 0.94, 0.95, 0.96, 0.97, 0.975 and 0.98 for the upper side and their mirror images for the lower side, keeping only those with at least 20 exceedances (at least 8 distinct values); a tail model is attempted only for columns with at least 300 training values. The bootstrap goodness-of-fit test uses 128 refitted replicates at level 0.05 and the shape interval uses 256 bootstrap resamples (2.5%–97.5% quantiles). A generalized Pareto candidate is rejected if $|\xi| > 1$, if its interval is wider than 1, or if $\mathrm{Stab}(h) > 0.25$, where $\mathrm{Stab}(h)$ is the median absolute difference between its ξ and those of the two neighbouring accepted

thresholds on each side. In Eq. (2) we use $\lambda_s = 0.5$, $\lambda_u = 0.05$, $\lambda_n = 0.1$ and $\lambda_c = 0.01$. The held-out term E_{val} of Eq. (3) uses $\mathcal{Q} = \{0.99, 0.995\}$ (mirrored for the lower side), only quantiles beyond the candidate threshold, and only when the validation sample contains at least 8 tail rows; $s_j = \max\{\text{IQR}, \text{sd}\}$ of the training column. If no candidate survives, the side stays empirical.

C.3 CLEAN-PIT TAIL WEIGHTING

Eq. (14) uses $p_0 = 0.025$ (tail start), $p_1 = 0.001$ (rarity floor), $W_{\max} = 6$, power $\psi = 0.75$, and SNR gate floor $g_0 = 0.35$; the blend is $\lambda_T = 0.65$. The edge probability is clamped below at p_1, so the depth is 0 for $p^{\text{edge}} \geq p_0$ and 1 at the floor. Normalizing by the observed-coordinate mini-batch mean of each column means weighting redistributes emphasis within a feature instead of changing its average scale. This is the reason all schema-routed numerical variables can share Eq. (15): schema changes the map $X_j \mapsto Z_{0j}$, while the neural objective operates in a common probability geometry.

D CATEGORICAL SCHEDULE AND ORDINAL DETAILS

D.1 FREQUENCY-AWARE EXPOSURE

Let n_{jk} be the training count of state k in column j, let $K_j^{\text{c}} = |\mathcal{C}_j|$ and $N_j = \sum_k n_{jk}$, and let $\widetilde{\pi}_{jk} = (n_{jk} + a_0)/(N_j + a_0 K_j^{\text{c}})$ be its Dirichlet-smoothed frequency with $a_0 = 0.5$. The rarity of Eq. (10) is

$$\varrho_{jk} = \left(\epsilon_\varrho + (1 - 2\epsilon_\varrho)\frac{s_{jk}}{s_{jk} + \tau}\right)^{p_\varrho}, \quad s_{jk} = -\log \widetilde{\pi}_{jk}, \qquad \widetilde{\varrho}_{jk} = \varrho_{jk} - \sum_l \pi_{jl}\varrho_{jl}, \tag{29}$$

with $\epsilon_\varrho = 0.05$, $\tau = 2$, $p_\varrho = 0.5$ and rarity strength $\eta = 0.75$. The residual is $u_{jk} = \text{rel}_{jk}\, u_{\max} \tanh(\widetilde{\theta}_{jk})$, where $\widetilde{\theta}_{jk}$ is the learned raw residual centered under $\boldsymbol{\pi}_j$, $u_{\max} = 0.6$ and $\text{rel}_{jk} = n_{jk}/(n_{jk} + n_0)$ with $n_0 = 32$. Exponents are bounded in $[w_{\min}, w_{\max}] = [1, 6]$ and the column baseline starts at $w = 1.5$. Smaller w_{jk} means larger $t^{w_{jk}}$ for $0 < t < 1$, hence more masked reconstruction targets for a rare state. The intercept c_j is solved by 36 bisection steps on the balance condition of Eq. (11).

D.2 SCHEDULE REGULARIZERS

The regularizer L_{reg} of Eq. (18) has four parts, each averaged over the relevant columns and states: (i) a residual prior $\lambda_{\text{res}}\, \overline{\widetilde{\theta}^2}/(2 \cdot 0.75^2)$; (ii) an ordinal-smoothness term λ_{sm} times the mean squared difference of the exponents of adjacent ordered levels; (iii) an information-leakage term λ_{leak} times a 16-point Gauss–Legendre approximation of $\overline{\int I(K_j; B_j \mid t)\, \mathrm{d}t}$, the mutual information between the clean state and its mask indicator; and (iv) a baseline anchor λ_{anc} times the mean squared drift of the column-level baselines from their initial values. We use $\lambda_{\text{res}} = 5 \times 10^{-4}$, $\lambda_{\text{sm}} = 2 \times 10^{-4}$, $\lambda_{\text{leak}} = 2 \times 10^{-3}$ and $\lambda_{\text{anc}} = 2 \times 10^{-4}$. Schedule-only regularizers are diagnostics and are excluded from checkpoint ranking.

D.3 CMOE AND CONTINUATION RATIO

The CMOE encoder in Eq. (7) guarantees ordering only in one dedicated coordinate; the semantic coordinates remain unconstrained. This avoids imposing an artificial Euclidean metric on ordinal categories. Special values such as "unknown" can be represented outside the ordered group. The posterior first allocates mass among the ordered group and special states, then uses continuation probabilities to split the ordered-group mass across ranks. The implementation also supports a nominal-ablation mode in which an otherwise ordinal field is treated as ordinary categorical data.

E HYBRID DENOISER, OBJECTIVES, AND SCHEDULE GRADIENT

E.1 DENOISER

The bidirectional transformer [49] receives one token per feature. Numerical tokens sum a linear projection of the current latent value, a linear projection of the (standardized) 22-dimensional ATI metadata vector of the column, column and type embeddings, a three-way numerical-missingness embedding, and 0.25 times the conditioning vector. Categorical tokens sum the nominal or CMOE state embedding, column and type embeddings, and 0.25 times the conditioning vector. Conditioning uses fixed Fourier features [48] (12 sine/cosine pairs with log-spaced frequencies from 1 to 1000, followed by a linear layer) of the noise level $\frac{1}{4}\,\mathrm{clip}\big(\log(\alpha^2/\sigma^2), -20, 20\big)$ and of t_{C}: numerical tokens receive both, categorical tokens receive t_{C} plus the mean of the numerical noise-level embeddings. Each block is pre-norm [50] multi-head self-attention followed by a GELU feed-forward network, with layerwise scale and shift computed per token from the conditioning vector by a zero-initialized linear layer and applied after each normalization, in the manner of FiLM [35] and DiT [34] (without gates). Nominal heads are initialized at the log training frequencies and the continuation logits of ordinal heads at the empirical continuation ratios.

Algorithm 1 Hybrid diffusion training

Require: Training row $(x^{\mathrm{num}}, x^{\mathrm{cat}})$, fitted ATI maps, categorical schedules
1: Map observed numerical values to Z_0 using schema-specific ATI
2: Sample independent clocks $t_{\mathrm{N}} \sim \mathcal{U}(\delta_{\mathrm{N}}, 1-\delta_{\mathrm{N}})$ and $t_{\mathrm{C}} \sim \mathcal{U}(\delta_{\mathrm{C}}, 1)$
3: Sample $\epsilon \sim \mathcal{N}(0, I)$ and form $Z_{t_{\mathrm{N}}}$ by Eq. (5)
4: Sample absorbing categorical state $C_{t_{\mathrm{C}}}$ by Eq. (9)
5: $(\hat{v}, \mu) \leftarrow \textsc{JointDenoiser}(Z_{t_{\mathrm{N}}}, C_{t_{\mathrm{C}}}, t_{\mathrm{N}}, t_{\mathrm{C}})$
6: Compute $L_{\mathrm{num}}, L_{\mathrm{cat}}, L_{\mathrm{ord}}$ and auxiliary regularizers
7: Update transformer and the schedule through its direct (pathwise) derivatives
8: Update categorical sampling-law parameters with the RLOO score term
9: Update EMA parameters; validate according to the checkpoint protocol

Algorithm 2 State-consistent restart and posterior-budgeted sampling

1: Initialize $Z \sim \mathcal{N}(0, I)$ and all categorical fields to M
2: **while** numerical grid unfinished or any categorical mask remains **do**
3: Choose next numerical time s_{N}
4: Optionally apply VP restart $Z_t \to Z_u$ by Eq. (19)
5: Optionally re-mask decoded categorical states by Eqs. (21) and (22)
6: $(\hat{v}, \mu) \leftarrow \textsc{JointDenoiser}(Z_u, C_u, u_{\mathrm{N}}, u_{\mathrm{C}})$
7: Return numerics to s_{N} by Eq. (20)
8: Set the reveal target $B^\star$ by Eq. (26)
9: Solve $B(s_{\mathrm{C}}; u_{\mathrm{C}}) = B^\star$ for the next categorical clock
10: Sample each masked field with Eqs. (23) and (24)
11: **end while**
12: Apply terminal state-consistent categorical rejuvenation conditioned on final numerics
13: Decode one RAW table and one rank-calibrated view by Eq. (6)
14: **return** RAW sample and calibrated diagnostic view

Algorithm 3 Joint bidirectional denoiser

Require: $Z_t, C_s, t_{\mathrm{N}}, t_{\mathrm{C}}$, ATI metadata, schema and ordinal specifications
1: **for** each numerical field j **do**
2: Build token from Z_{tj}, column/type IDs, missingness, ATI metadata and both clock conditions
3: **end for**
4: **for** each categorical field j **do**
5: Encode nominal/special state or CMOE ordinal state; add column/type IDs and both clock conditions
6: **end for**
7: Concatenate all feature tokens into one row sequence
8: **for** each bidirectional pre-norm transformer block **do**
9: Apply clock-conditioned normalization, self-attention, then conditioned feed-forward update
10: **end for**
11: Predict numerical velocities with direct linear heads
12: Predict categorical clean-state posteriors with nominal or ordinal heads
13: **return** $(\hat{v}, \mu)$

E.2 Objectives

The numerical target is $v = \alpha\epsilon - \sigma Z_0$ [41]. The tail term is a reweighted copy of the same squared velocity error, not a second diffusion target. Numerical missing values are excluded from the velocity loss and represented through categorical missingness states. The categorical objective uses Eq. (16) only where a field is masked and includes the time-measure factor of Eq. (17). The clocks are drawn as $t_{\mathrm{N}} \sim \mathcal{U}(\delta_{\mathrm{N}}, 1-\delta_{\mathrm{N}})$ with $\delta_{\mathrm{N}} = 10^{-4}$ and $t_{\mathrm{C}} \sim \mathcal{U}(\delta_{\mathrm{C}}, 1)$ with $\delta_{\mathrm{C}} = 10^{-3}$; the categorical clock is stratified across the mini-batch (one jittered point per equal-probability stratum, randomly permuted) to reduce clustering. The ordinal auxiliary averages threshold BCE over active ordinal targets, rather than dividing by the total number of categorical columns. The terminal loss L_{term} evaluates Eq. (16) on all-mask input with $t_{\mathrm{C}} = 1$ and pure numerical noise, on at most 512 rows every fourth step (with its weight multiplied by 4 on those steps). The coefficients are $\lambda_C = 1$, $\lambda_O = 0.20$ and $\lambda_{\mathrm{term}} = 0.05$.

E.3 Schedule gradient

The state-dependent masking law contains stochastic discrete mask draws. To propagate a schedule gradient through that sampling law, the implementation draws two independent categorical corruptions B_1, B_2 of the same row and adds the two-sample leave-one-out (RLOO) score-function surrogate [21],

$$L_{SF} = \tfrac{1}{2}\,\mathrm{stopgrad}(A_1 - A_2)\{\log q_\phi(B_1) - \log q_\phi(B_2)\}, \tag{30}$$
$$\log q_\phi(B) = \sum_j \Big[b_j \log t_{\mathrm{C}}^{\,w_{jC_{0j}}} + (1-b_j)\log\big(1 - t_{\mathrm{C}}^{\,w_{jC_{0j}}}\big)\Big],$$

where $b_j \in \{0,1\}$ indicates that column j is masked, ϕ denotes the schedule parameters, and $A_i = \lambda_C L_{\mathrm{cat}} + \lambda_O L_{\mathrm{ord}}$ under draw i. Direct derivatives through w_{jk} in ℓ_j remain active. Numerical loss is deliberately absent from A_i, so this term should not be described as an unbiased score estimator for the entire mixed objective. The schedule is frozen for the first 149 epochs and then trained with its own AdamW optimizer (learning rate 10^{-4}, 50-epoch warm-up, cosine decay to 10%, gradient-norm clip 0.25).

F Evaluation Metrics

F.1 Tail-AD

For numerical feature j, let TRAIN define lower and upper thresholds $T_j^L = Q_j^{\mathrm{train}}(q_L)$ and $T_j^U = Q_j^{\mathrm{train}}(q_U)$. For a real reference sample R_j and synthetic sample S_j, define the lower exceedance

samples $R_j^L = \{x \in R_j : x < T_j^L\}$, $S_j^L = \{x \in S_j : x < T_j^L\}$ and upper exceedance samples analogously. We compute the standardized two-sample Anderson–Darling statistic [42] separately on valid lower and upper pairs and average across feature-tail comparisons. Because it is applied after conditioning on being in the tail, Tail-AD primarily measures tail shape; exceedance probability itself is assessed separately.

F.2 TAIL-JSD

TRAIN quantiles define a seven-bin partition with finer resolution in the two tails and a coarse central region. Let p_b and s_b be real and synthetic probabilities in bin b, $m_b = (p_b + s_b)/2$. Then

$$\text{Tail-JSD}_j = \frac{1}{2}\sum_b p_b \log_2 \frac{p_b}{m_b} + \frac{1}{2}\sum_b s_b \log_2 \frac{s_b}{m_b}. \tag{31}$$

Zero terms contribute zero. This is a standard bounded Jensen–Shannon divergence [26] applied to a tail-focused, TRAIN-frozen partition. Unlike Tail-AD, it is not renormalized within the tail and therefore reflects both tail mass and local allocation.

F.3 ORDINAL RANK WASSERSTEIN

For an ordinal feature with $K \geq 2$ ordered levels, let $F_R(r)$ and $F_S(r)$ be empirical CDFs over rank r after conditioning on membership in the ordered subgroup. The normalized one-dimensional Wasserstein distance is

$$\text{ORW}_j = \frac{1}{K-1}\sum_{r=0}^{K-2} |F_R(r) - F_S(r)|. \tag{32}$$

It lies in normalized rank units and is reported with real and synthetic ordered-group coverage. The implementation verifies this discrete-CDF formula against a standard 1-Wasserstein computation on normalized ranks.

F.4 TRAIN FIDELITY AND HELD-OUT GENERALIZATION

For a tail probability p, TRAIN defines $T_{j,p}^U = Q_j^{\text{train}}(1-p)$ and $T_{j,p}^L = Q_j^{\text{train}}(p)$. The same thresholds are used everywhere. Upper-tail fitting and generalization errors are

$$E_{j,p}^{\text{fit}} = \left|\mathbb{P}_{\text{syn}}(X_j > T_{j,p}^U) - \mathbb{P}_{\text{train}}(X_j > T_{j,p}^U)\right|, \tag{33}$$

$$E_{j,p}^{\text{gen}} = \left|\mathbb{P}_{\text{syn}}(X_j > T_{j,p}^U) - \mathbb{P}_{\text{test}}(X_j > T_{j,p}^U)\right|, \tag{34}$$

with analogous lower-tail quantities and scaled quantile errors. The evaluator also computes numerical co-exceedance errors, categorical distributions conditional on numerical tail events (TVD), and ordinal conditional-rank differences. Four reports are kept separate: RAW–TRAIN, RAW–TEST, calibrated–TRAIN, and calibrated–TEST.

G DATASETS AND BENCHMARK PROTOCOL

Table 2: Benchmark datasets

Dataset	Rows	Num.	Cat.	Train	Val.	Test / task
Adult	48,842	6	9	28,943	3,618	16,281 / class.
Default	30,000	14	11	24,000	3,000	3,000 / class.
Shoppers	12,330	10	8	9,864	1,233	1,233 / class.
MAGIC	19,019	10	1	15,215	1,902	1,902 / class.
Beijing	43,824	7	5	35,058	4,383	4,383 / regr.
News	39,644	46	2	31,714	3,965	3,965 / regr.
Diabetes	101,766	9	27	61,059	20,353	20,354 / class.

The Ann-Thyroid anomaly convention follows the commonly used 7,200-row, six-feature benchmark with 534 anomalies (7.42%) reported by ADBench [17], while the source collection is UCI Thyroid Disease [37]. Because this dataset is absent from the current notebook registry, the final camera-ready protocol must record the exact train/validation/test partition and random seed before any scores are filled.

H Implementation Details and Hyperparameters

Table 3: Key current implementation settings.

Component	Current setting
Transformer width / layers / heads	64 / 6 / 4
Feed-forward width / dropout	256 / 0.05
Fourier features (noise level and t_{C})	24
Training epochs / batch size	4,000 / 4,096
Model / categorical-schedule learning rate	3×10^{-4} / 10^{-4}
Optimizer / weight decay / EMA decay	AdamW, warm-up 300 steps then cosine to 5% / 10^{-4} / 0.999
Gradient-norm clip (model / schedule)	1.0 / 0.25
Best-checkpoint competition	starts at epoch 3,001, validation loss only
Numerical loss tail mixture λ_T	0.65
Categorical / ordinal coefficients	1.0 / 0.20
Tail rarity start p_0 / floor p_1	0.025 / 0.001
Maximum raw tail weight $W_{\max}$ / power ψ / gate floor g_0	6 / 0.75 / 0.35
Clock cut-offs δ_{N} / δ_{C}	10^{-4} / 10^{-3}
Exponent range / initial baseline	$[1, 6]$ / 1.5
Numerical reverse steps	64
Numerical restart ϕ_{N} / cap $\nu_{\max}$	0.25 / 0.05
Categorical restart ϕ_{C} / cap $b_{\max}$	0.01 / 0.50
Reveal fraction $f_{\min}$–$f_{\max}$ / power ψ_{C} / floor $b_{\min}$	0.03–0.18 / 0.75 / 0.15
Terminal rejuvenation rounds / fraction / cap	2 / 0.03 / 1.0
Maximum categorical steps / generation batch	256 / 10,000
Primary numerical decoding	RAW; calibrated view also saved
Repeated synthetic tables	20 from one selected fitted generator

The ATI candidate grid, selection coefficients and schema-routing thresholds are listed in Appendix C, and the categorical-schedule constants in Appendix D; the restart and reveal-budget constants of Section 3.5 are in Table 3. Validation contributes only to tail-candidate selection and checkpoint choice. The implementation records total training wall time and per-synthetic-table generation time with GPU synchronization around timing boundaries.

For registered Adult data, the schema overrides include ordinal treatment of education-number, integer-grid continuous treatment of final sampling weight, zero-inflated integer-grid treatment of capital gain/loss, and bounded quantitative treatment of age and hours-per-week. Other datasets use the same profiler with optional explicit metadata overrides when domain semantics are known.

I Detailed Historical Fidelity and Privacy Tables

The following values are historical results reported by TabDiff [44]. They provide context only; they are not reruns from the supplied TabTailor notebook.

Table 4: Shape error rates (%, lower is better)[44].

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Avg.
CTGAN	16.84±.03	16.83±.04	21.15±.10	9.81±.08	21.39±.05	16.09±.02	9.82±.08	15.99
TVAE	14.22±.08	10.17±.05	24.51±.06	8.25±.06	19.16±.06	16.62±.03	18.86±.13	15.97
GOGGLE	16.97	17.02	22.33	1.90	16.93	25.32	24.92	17.91
GReaT	12.12±.04	19.94±.06	14.51±.12	16.16±.09	8.25±.12	OOM	OOM	14.20
STaSy	11.29±.06	5.77±.06	9.37±.09	6.29±.13	6.71±.03	6.89±.03	OOM	7.72
CoDi	21.38±.06	15.77±.07	31.84±.05	11.56±.26	16.94±.02	32.27±.04	21.13±.25	21.55
TabDDPM	1.75±.03	1.57±.08	2.72±.13	1.01±.09	1.30±.03	78.75±.01	31.44±.05	16.93
TabSyn	0.81±.05	1.01±.08	1.44±.07	1.03±.14	1.26±.05	2.06±.04	1.85±.02	1.35
TabDiff	0.63±.05	1.24±.07	1.28±.09	0.78±.08	1.03±.05	2.35±.03	0.89±.23	1.17
TABTAILOR	0.42±0.06	0.24±0.03	0.41±0.04	0.14±.02	0.76±0.04	1.67±0.95	0.56±0.78	

Table 5: Trend error rates (%, lower is better) [44].

Method	Adult	Default	Shoppers	MAGIC	Beijing	News	Diabetes	Avg.
CTGAN	20.23±1.20	26.95±.93	13.08±.16	7.00±.19	22.95±.08	5.37±.05	18.95±.34	16.36
TVAE	14.15±.88	19.50±.95	18.67±.38	5.82±.49	18.01±.08	6.17±.09	32.74±.26	16.44
GOGGLE	45.29	21.94	23.90	9.47	45.94	23.19	27.56	28.18
GReaT	17.59±.22	70.02±.12	45.16±.18	10.23±.40	59.60±.55	OOM	OOM	44.24
STaSy	14.51±.25	5.96±.26	8.49±.15	6.61±.53	8.00±.10	3.07±.04	OOM	7.77
CoDi	22.49±.08	68.41±.05	17.78±.11	6.53±.25	7.07±.15	11.10±.01	29.21±.12	23.21
TabDDPM	3.01±.25	4.89±.10	6.61±.16	1.70±.22	2.71±.09	13.16±.11	51.54±.05	11.95
TabSyn	1.93±.07	2.81±.48	2.13±.10	0.88±.18	3.13±.34	1.52±.03	3.90±.04	2.33
TabDiff	1.49±.16	2.55±.75	1.74±.08	0.76±.12	2.59±.15	1.28±.04	2.20±.16	1.80
TABTAILOR	1.26±0.10	1.17±0.05	1.27±0.14	1.54±0.07	1.45±0.02	0.83±0.76	2.13±0.43	

Table 6: Machine-learning efficiency. AUC ↑ is used for Adult, Default, Shoppers, MAGIC, and Diabetes; RMSE ↓ for Beijing and News.[44].

Method	Adult	Default	Shoppers	Magic	Beijing	News	Diabetes	Avg. gap (%)
Real	.927±.000	.770±.005	.926±.001	.946±.001	.423±.003	.842±.002	.704±.002	0.0
CTGAN	.886±.002	.696±.005	.875±.009	.855±.006	.902±.019	.880±.016	.569±.004	23.7
TVAE	.878±.004	.724±.005	.871±.006	.887±.003	.770±.011	1.01±.016	.594±.009	20.2
GOGGLE	.778±.012	.584±.005	.658±.052	.654±.024	1.09±.025	.877±.002	.475±.008	42.1
GReaT	.913±.003	.755±.006	.902±.005	.888±.008	.653±.013	OOM	OOM	13.3
STaSy	.906±.001	.752±.006	.914±.005	.934±.003	.656±.014	.871±.002	OOM	10.9
CoDi	.871±.006	.525±.006	.865±.006	.932±.003	.818±.021	1.21±.005	.505±.004	30.2
TabDDPM	.907±.001	.758±.004	.918±.005	.935±.003	.592±.011	4.86±3.04	.521±.008	11.95
TabSyn	.909±.001	.763±.002	.914±.004	.937±.002	.580±.009	.862±.024	.684±.002	6.78
TabDiff	.912±.002	.763±.005	.921±.004	.936±.003	.555±.013	.866±.021	.689±.016	5.76
TABTAILOR	.910±.001	.775±.003	.924±.004	.921±.003	.598±0.011	.811+0.05	.682±0.04	

Table 7: Historical α-Precision (%, higher is better).

Method	Adult	Default	Shop.	MAGIC	Beijing	News	Diabetes	Avg
CTGAN	77.74	62.08	76.97	86.90	96.27	96.96	79.89	82.40
TVAE	98.17	85.57	58.19	86.19	97.20	86.41	19.24	75.85
GOGGLE	50.68	68.89	86.95	90.88	88.81	86.41	23.09	70.81
GReaT	55.79	85.90	78.88	85.46	98.32	OOM	OOM	80.87
STaSy	82.87	90.48	89.65	86.56	89.16	94.76	OOM	88.91
CoDi	77.58	82.38	94.95	85.01	98.13	87.15	64.80	84.29
TabDDPM	96.36	97.59	88.55	98.59	97.93	0.00	28.35	72.48
TabSyn	99.39	98.65	98.36	99.42	97.51	95.05	96.61	97.86
TabDiff	99.02	98.49	99.11	99.47	98.06	97.36	95.69	98.22
TABTAILOR	99.81	98.76	99.26	99.31	98.76	98.43	98.56	

Table 8: Historical β-Recall (%, higher is better).

Method	Adult	Default	Shop.	MAGIC	Beijing	News	Diabetes	Avg
CTGAN	30.80	18.22	31.80	11.75	34.80	24.97	9.42	23.11
TVAE	38.87	23.13	19.78	32.44	28.45	29.66	4.92	25.32
GOGGLE	8.80	14.38	9.79	9.88	19.87	2.03	3.74	9.78
GReaT	49.12	42.04	44.90	34.91	43.34	OOM	OOM	43.34
STaSy	29.21	39.31	37.24	53.97	54.79	39.42	OOM	42.32
CoDi	9.20	19.94	20.82	50.56	52.19	34.40	2.70	27.12
TabDDPM	47.05	47.83	47.79	48.46	56.92	0.00	0.03	35.44
TabSyn	47.92	46.45	49.10	48.03	59.15	43.01	33.72	46.77
TabDiff	51.64	51.09	49.75	48.01	59.63	42.10	41.74	49.40
TABTAILOR	49.12	48.79	49.80	47.14	52.12	39.6	42.34	48.87

Table 9: Historical C2ST scores (higher is better under the TabDiff convention).

Method	Adult	Default	Shop.	MAGIC	Beijing	News	Diabetes	Avg
CTGAN	.5949	.4875	.7488	.6728	.7531	.6947	.5593	.6444
TVAE	.6315	.6547	.2962	.7706	.8659	.4076	.0487	.5250
GOGGLE	.1114	.5163	.1418	.9526	.4779	.0745	.0912	.3380
GReaT	.5376	.4710	.4285	.4326	.6893	OOM	OOM	.5118
STaSy	.4054	.6814	.5482	.6939	.7922	.5287	OOM	.6083
CoDi	.2077	.4595	.2784	.7206	.7177	.0201	.0008	.3435
TabDDPM	.9755	.9712	.8349	.9998	.9513	.0002	.1980	.7044
TabSyn	.9910	.9826	.9662	.9960	.9528	.9255	.5953	.9156
TabDiff	.9950	.9774	.9843	.9989	.9781	.9308	.9865	.9787
TABTAILOR	.9981	.9988	.9971	.9999	.9786	.9964	.9545	.9756

J EFFICIENCY AND ROBUSTNESS STUDY

J.1 HISTORICAL EFFICIENCY CONTEXT

The current TABTAILOR implementation records full training wall time, best-checkpoint epoch, and per-table generation time. We recommend reporting both seconds per synthetic table and rows per second, together with GPU model and generation batch size.

J.2 TAIL AND ORDINAL ROBUSTNESS PROTOCOL

We propose three robustness axes: (i) independent training seeds; (ii) reverse-step count and restart strength; and (iii) fitted-marginal perturbations that vary eligible tail thresholds within the validated region. For every setting we report Shape/Trend together with Tail-AD, Tail-JSD, ORW, TRAIN–RAW and TEST–RAW tail errors. The key question is whether a method's targeted gains persist without requiring rank calibration.

J.3 SAMPLING-STEP SENSITIVITY

The historical TabDiff study reports that both TabSyn and TabDiff improve substantially when moving from very small reverse-step budgets to moderate ones, after which Shape and Trend stabilize. For TABTAILOR, the corresponding sensitivity study should additionally track Tail-AD and ORW, because a low average error could coexist with degraded extreme or ordered behavior. We leave the final curves open until the same hardware and data splits are used.

Table 10: Historical DCR nearest-record diagnostic (%, values near 50 are balanced). MAGIC was not included in the cited DCR table.

Method	Adult	Default	Shoppers	Beijing	News	Diabetes
STaSy	50.33	50.23	51.53	50.59	50.59	OOM
CoDi	49.92	51.82	51.06	50.87	50.79	51.12
TabDDPM	51.14	52.15	63.23	80.11	79.31	37.76
TabSyn	50.94	51.20	52.90	50.37	50.85	50.62
TabDiff	50.10	51.11	50.24	50.50	51.04	50.43
TABTAILOR	50.27±0.89	50.07±0.92	50.40±0.78	50.01±0.01	50.63±0.16	

K ADDITIONAL DIAGNOSTICS

L EXTENDED RELATED WORK

This appendix gives the fuller related-work discussion; Section 2 contains a condensed version.

Synthetic tables and diffusion. The first deep tabular generators borrowed tools from images. CTGAN [51] trains a generator against a critic and conditions it so that rare category values are seen more often, TVAE [51] compresses each row into a small code, GOGGLE [27] learns a graph of relations between columns, and GReaT [9] fine-tunes a language model on rows written as sentences. Diffusion models learn to generate by undoing a slow corruption. DDPM [18] corrupts data with Gaussian noise and trains a network to remove it, DDIM [45] samples the same network deterministically in far fewer steps, and the score-based SDE view [46] places both in one continuous-time family whose variance-preserving member is our numerical branch. For tables, STaSy [20] applies score-based generation with self-paced learning; CoDi [24] runs Gaussian and categorical diffusions that condition on each other; TabDDPM [22] pairs Gaussian diffusion for numbers with multinomial diffusion for categories; TabSyn [54] diffuses inside a learned latent space; and TabDiff [44] diffuses mixed rows in continuous time with learnable noise schedules per column. TABTAILOR follows this mixed-type direction but changes how numerical columns are prepared and how category states are corrupted, going one level below a per-column schedule.

Masked diffusion. Replacing a value by a mask symbol was introduced for discrete diffusion by D3PM [4] and made simple and effective for text by MDLM [39]; SEDD [29] instead estimates ratios of probabilities. Closest to ours is Simplified and Generalized Masked Diffusion [43] from Google DeepMind, which shows that the objective reduces to a weighted sum of cross-entropy losses and whose generalized model, GenMD4, lets the masking schedule depend on the value being hidden. Our categorical branch builds on this state-dependent formulation and adds a schedule that is learnable per column and category value, frequency-informed, and balanced so that a column's overall masking exposure does not change.

Heavy tails. Beyond a high threshold, exceedances behave like a generalized Pareto distribution, a result due to Balkema and de Haan [5] and Pickands [36], with standard fitting tools [13] and goodness-of-fit tests [12]. Pareto GAN [19] and heavy-tailed diffusion [32] use this kind of idea in GANs and diffusion models, respectively, but not for mixed tables, while tabular pipelines typically map each column to a normal scale through an empirical or fitted marginal [33], which cannot describe values beyond those seen in training. ATI keeps the empirical body, adds a tail only when evidence supports it, and uses the randomized distributional transform [38] so that ties and structural zeros are mapped exactly.

Ordered categories and restarts. Statistics models an ordered level by asking whether it is at least r: the cumulative-link model [30] and continuation-ratio models [1] do so, Frank and Hall [16] turned it into binary threshold questions, and CORAL [10] made those questions rank-consistent for neural networks. The closest generative work is the ordinal-aware diffusion model for disease images of Kurt et al. [23], which conditions on severity through an additive embedding; CMOE is a follow-up for table columns that adds a strictly increasing coordinate and a continuation-ratio head compatible with masked diffusion. Finally, restart sampling [52] alternates a little added noise

with denoising so that early mistakes can be repaired; we extend it to mixed rows, with exact local transitions for both numerical and categorical columns.