Compliant Generative Diffusion for Finance

Michael Cardei*
University of Virginia
ntr2rm@virginia.edu

Jose M Munoz*
Massachusetts Institute of Technology
josemm@mit.edu

Oscar Barrera
Harvard University
oscarbarrera@g.harvard.edu

Shreyas K Chandrahas
Visa Inc
shrchand@visa.com

Partha Saha
Visa Inc
pasaha@visa.com

Abstract

Generative models in finance face the dual challenge of producing realistic data while satisfying strict regulatory and economic objectives, a requirement that standard tabular diffusion models cannot provide. To address this difficulty, we introduce **Constrained Tabular Diffusion for Finance** (CTDF), a novel integration of sampling-time feasibility operations with mixed-type tabular diffusion in financial applications. By incorporating a training-free feasibility operator into the reverse-diffusion sampling loop, CTDF enforces hard constraints for applications such as simulation, legal compliance, and extrapolation. Experiments on large-scale financial datasets demonstrate zero constraint violations and improvement in scarce data utility. CTDF establishes a robust method for generating trustworthy and compliant synthetic data, enabling compliant generative modeling.

1 Introduction

High-quality synthetic data is essential for the modern financial ecosystem, from global payment networks to digital banking and commerce platforms. Within this domain, synthetic data has become a vital tool, however, the utility of this data is contingent upon two core requirements: it must not only capture the complex statistical patterns of real-world distributions but also adhere to the complex landscape of business logic, economic principles, and regulatory guardrails governing financial operations. Failure to enforce these constraints invalidates model outputs, undermines strategic decision-making, and exposes an institution to significant compliance and operational risk. Conversely, the ability to generate data that is both statistically faithful and structurally sound allows for discovery, enabling more nuanced simulations of market dynamics and deeper insights into the drivers of business performance.

Recent advances in tabular generation have shifted toward diffusion probabilistic models, which yield better mode coverage and sample fidelity. While these approaches provide mixed-type high-fidelity data generation [1–4], their stochastic sampling remains unconstrained, unable to enforce adherence to strict constraints. Additionally, existing approaches to controlling generative output often fall short in the financial domain. Model conditioning, such as classifier-free guidance, can only influence outputs without strictly enforcing them [5]. An alternative, post hoc correction, involves generating a full batch of samples and then filtering out invalid ones or projecting them onto the feasible set in a single final step. This approach is inefficient and distorts the learned distribution [6].

To address these problems, we introduce **Constrained Tabular Diffusion for Finance (CTDF)**, a training-free feasibility operator applied at every reverse-diffusion step. By projecting intermediate

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on Generative AI in Finance.

samples onto the feasible region throughout the denoising trajectory, CTDF ensures hard-constraint compliance without post-hoc filtering. This per-step enforcement keeps synthetic data aligned with the base model's learned joint structure, preserving fidelity while achieving **zero violations** on regulatory and business rules.

2 Related Work

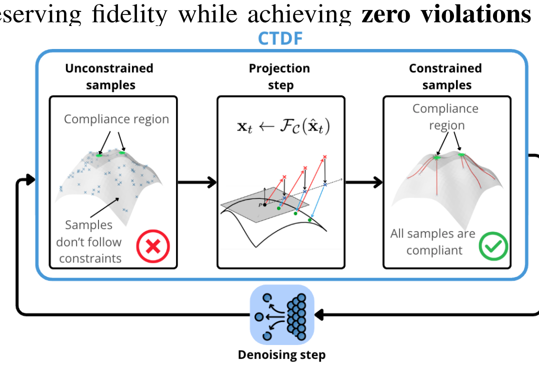


Figure 1: CTDF enforces constraints via the feasibility operation. At each step, unconstrained samples are mapped onto the compliance region.

Diffusion models have emerged as the state-of-the-art approach for tabular synthesis, with methods such as TabDDPM [3], STaSy [7], TabSyn [1], and TabDiff [2] improving fidelity across mixed numerical and categorical data. In finance, FinDiff introduced domain-specific architectures but still lacked mechanisms for strict constraint enforcement. Previous work on guided diffusion (classifier-based or classifier-free) [8, 5] improves conditional alignment but cannot guarantee feasibility. Projection-based approaches such as PDM [9] and NSD [10] enforce constraints during sampling in continuous and symbolic domains. We build on these works with, to our knowledge, the first study to implement and evaluate per-step hard constraint enforcement in mixed-type *tabular* diffusion with a focus on financial datasets and compliance-driven constraints.

3 Methods

CTDF extends a pre-trained tabular diffusion model with a feasibility operator applied at every reverse-diffusion step. Unlike post-hoc filtering, this operation ensures each intermediate state remains within the feasible set, yielding samples that are both statistically faithful and strictly compliant without retraining. This approach relies on diffusion's inherent parallel denoising, modeling all columns simultaneously, updating the current sample jointly at each step. For this implementation, we use TabDiff [2] as the base diffusion model, however, CTDF can be extended to other tabular diffusion architectures.

Algorithm 1: Constrained Tabular Diffusion

Input: Base model μ_θ, feasible set $\mathcal{C}$, steps T
Output: Constraint-compliant sample $\mathbf{x}_0$
Sample $\mathbf{x}_T^{\text{num}} \sim \mathcal{N}(0, I)$; $\mathbf{x}_T^{\text{cat}} \leftarrow m$
for $t = T, \ldots, 1$ **do**
  Compute unconstrained update $\tilde{\mathbf{x}}_{t-1}$
  Apply feasibility operator: $\mathbf{x}_{t-1} \leftarrow \mathcal{F}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t-1})$
return $\mathbf{x}_0$

Feasibility Operator. Let $\mathcal{C} = \{\mathbf{x} \mid h_i(\mathbf{x}) = 0,\ g_j(\mathbf{x}) \leq 0\}$ define the valid region. After each unconstrained update $\tilde{\mathbf{x}}_{t-1}$ from the base model, CTDF applies

$$\mathbf{x}_{t-1} \leftarrow \mathcal{F}_{\mathcal{C}}(\tilde{\mathbf{x}}_{t-1}),$$

mapping samples back into $\mathcal{C}$. Repeated projection keeps the trajectory feasible throughout generation.

Mixed-Type Mapping. Financial data combine numerical and categorical features. For numerics, feasibility reduces to Euclidean projection onto an affine polytope: $\mathcal{F}^{\text{num}}(\hat{\mathbf{x}}) = \arg\min_{\mathbf{x} \in \mathcal{C}^{\text{num}}} \|\mathbf{x} - \hat{\mathbf{x}}\|^2$. For categorical, CTDF redistributes probability mass by zeroing invalid categories and renormalizing, equivalent to minimizing KL divergence to the model's logits. This hybrid mapping preserves fidelity while eliminating violations.

Functional and Symbolic Constraints. Beyond separable rules, many financial constraints couple features (e.g., $\texttt{price} \geq 1000$ if $\texttt{property\_type} = \texttt{Single Family}$). CTDF enforces such conditions by computing constraint boundaries from input features at each step and projecting only the dependent column, supporting both symbolic logic and neural surrogates.

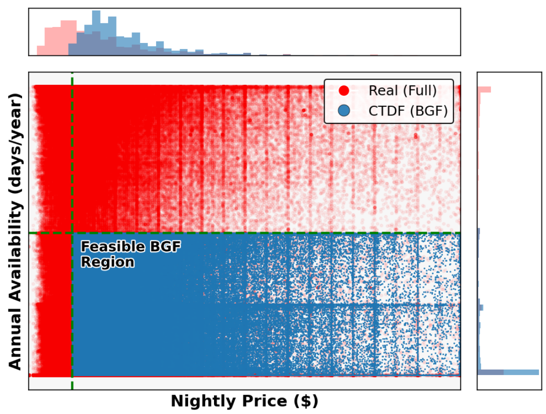
Fidelity evaluations metrics are described in Appendix D.

4 Experimental Results

Before assessing constraint satisfaction, we first validate the fidelity of our underlying unconstrained diffusion model against several established benchmarks for tabular data synthesis. Table 1 summa-

Model	Marginal Fidelity		Correlations	ML Similarity
	Shape Sim. (Avg) ↑	Dist. Dist. (Avg) ↓	Trend Score ↑	Detection AUC → 0.5
CTGAN [11]	0.86483	0.11454	0.79725	0.99999
TVAE [12]	0.84638	0.09050	0.83098	0.99985
FinDiff [4]	0.94472	0.02916	0.87369	0.98654
Tabsyn [1]	0.97646	0.02027	0.94595	0.79479
CTDF (TabDiff [2])	**0.98787**	**0.01116**	**0.95376**	**0.67860**

Table 1: Fidelity Benchmarks on the Housing Market Dataset. Our unconstrained base model (TabDiff) is compared against standard generative models.



Rule	Real (%)	CTDF (%)
Listings < 2	43.7	0.0
Price ≥ 100	35.2	0.0
Min nights ≤ 30	6.5	0.0
Reviews/mo ≥ 1.5	70.2	0.0
Availability ≤ 180	43.7	0.0
All rules	99.9	**0.0**

Statistic	Real $\mu \pm \sigma$	CTDF $\mu \pm \sigma$
Nightly price ($)	207$\pm$109	229$\pm$120
Availability	88$\pm$56	42$\pm$56
Reviews/mo	4.1$\pm$3.0	1.9$\pm$0.7
Listings/host	1.6$\pm$1.0	1.1$\pm$0.3

Figure 2: CTDF vs. real Airbnb listings. *Left:* Scatter plot comparing real NYC listings (red) to CTDF synthetic listings for the Balanced Growth Fund (blue). Green dashed lines mark regulatory thresholds. CTDF samples fall entirely in the compliant lower-right quadrant. *Right:* Constraint-violation rates (top) and descriptive-statistic match (bottom).

rizes the performance across three key dimensions: marginal distribution fidelity, pairwise column correlations, and machine learning utility. The results demonstrate TabDiff's superior performance in all metrics, capturing distributional shapes, pairwise trends, and overall statistical patterns. This provides a high-quality, state-of-the-art starting point for our constrained generation framework. Extended experimental results are in Appendix E.

4.1 Housing market

To evaluate CTDF's ability to generate realistic, constraint-compliant tabular data, we use the Airbnb Open Data from major U.S. cities, a large-scale real estate dataset with 450k property listings [13]. This domain is especially well-suited for our setting: listings must satisfy a growing number of local housing regulations, platform-level business rules, and investment constraints.

Compliant Portfolio Simulation To showcase CTDF's scenario-analysis capabilities, we construct the *Balanced Growth Fund (BGF)*, a synthetic portfolio of high-end, legally compliant, entire-home rentals in New York City. For this scenario, we fix the categorical columns to `room_type = Entire home/apt`, `city = New York City`, and impose several hard constraints. To ensure adherence to New York's housing regulations (Local Law 18 [14]), we enforce primary-residency rules by requiring `calculated_host_listings_count < 2` and an annual `availability_365` $\leq$ 180. To reflect an actively managed, non-corporate listing, `reviews_per_month` is constrained between 1.5 and 15. Finally, to align with the fund's investment strategy targeting the upper-tier market, we set a minimum `price` $\geq$ \$100 and `minimum_nights` $\leq$ 30.

Consequently, CTDF enables the simulation of portfolios that are simultaneously realistic, compliant, and strategically aligned, unlocking unprecedented precision for risk modeling and scenario planning in finance. The resulting output of CTDF is shown in Fig. 2, where a comparison between the real housing market and the synthetic sample illustrates the compliance of two vital constraints. This experiment demonstrates how CTDF enables precise simulation of legally compliant and strategically targeted real estate portfolios, empowering financial institutions to model regulatory exposure, test investment hypotheses, and design products in alignment with evolving market rules.

4.2 Loan Analysis

We now turn to the distinct domain of consumer credit to test CTDF in a core financial domain. We use a comprehensive Lending Club dataset, which contains historical data for all peer-to-peer loans issued between 2007 and 2020 [15]. Fidelity results are available in Appendix E.1.

AI-Driven Credit Risk for Small-Business Loans Modern credit risk models rely on large, balanced datasets to train powerful machine learning scoring functions. For small-business lending, severe class imbalance ($< 1\%$ of loans) and regulatory caps on DTI, INT_RATE, and REVOL_UTIL create two AI challenges: (1) data scarcity in the high-risk tail, and (2) bias when unconstrained sampling produces implausible loans. CTDF solves both by generating a large, *compliant* synthetic SB portfolio, enabling AI models to learn robust patterns exactly where real data is weakest. We enforce per-step feasibility operation on every CTDF sample DTI ≤ 0.43, INT_RATE $\leq 25\%$, REVOL_UTIL $\leq 90\%$, `purpose` $=$ `small_business`. These caps align with U.S. underwriting rules and ensure that no training example violates regulatory or economic constraints. By contrast, unconstrained TabDiff generates only 2% valid SB loans at the 25% rate cap, leaving AI models starved of critical high-spread examples.

We evaluate our method by comparing the performance of a CatBoost early-default classifier trained under four data regimes. The Orig-SB regime establishes a baseline using only the limited real small-business loans ($\sim$ 4k rows). The Unconstr-SB regime tests naive data augmentation by adding synthetically generated but non-compliant loans ($\sim$ 10k total rows). The Full-SB regime trains the model on the entire real loan dataset ($\sim$ 350k rows) and tests its performance on the SB segment. Finally, the CTDF-SB regime demonstrates our approach, augmenting the real SB loans with 500K regulatory-compliant synthetic samples generated by CTDF. All models are trained on pre-2018 data, validated on 2018, and evaluated on 2019–2020Q1 SB loans only.

Regime	AUC-PR ↑	F1-score ↑	R.@5% FPR ↑
CTDF–SB	**0.2570(18)**	**0.0305(00)**	**0.3039(13)**
Uncon–SB	0.1820(23)	0.0305(00)	0.2093(68)
Full-SB	0.1704(12)	0.0734(24)	0.2093(21)
Orig–SB	0.1403(05)	0.0329(06)	0.1643(31)

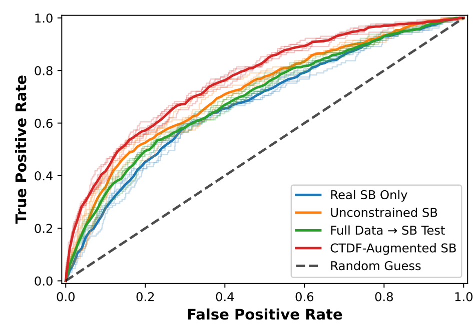


Figure 3: Top: Cross-validated performance metrics on the 2019–2020 Q1 small-business test set. Bottom: Cross-validated ROC curves with shaded ± 1 std across folds.

In the full-data regime, fewer than 1% of training rows have INT_RATE $\geq 20\%$, so the model underfits the high-spread tail where defaults concentrate. CTDF's compliant augmentation supplies abundant, realistic examples in this critical region, sharpening the AI model's boundary and reducing both bias and variance in predicted default risk. Figure 3 demonstrates that CTDF-SB yields the best AI performance on every metric, e.g. a *+0.09* uplift in AUC-PR and a 27% relative increase in Recall@5% FPR versus Full-SB. This demonstrates that *constraint-aware synthetic data* is not just compliance insurance but *AI-grade data augmentation*, enabling financial models to learn from the rare, high-risk patterns that matter most.

5 Conclusion

We introduce Constrained Tabular Diffusion for Finance (CTDF), a training-free framework that integrates mapping-based feasibility operations into each sampling step of a mixed-type diffusion model. CTDF enforces hard affine and logical constraints with zero violations while preserving distributional fidelity and downstream utility. Our experiments on large-scale datasets serve as illustrative case studies, showcasing CTDF's strengths in scenario analysis, stress testing, and risk modeling. More broadly, the ability to generate compliant synthetic data opens new avenues across finance and beyond, laying the foundation for deploying constraint-aware generative modeling in diverse financial applications.

References

[1] Hengrui Zhang, Jiani Zhang, et al. Mixed-type tabular data synthesis with score-based diffusion in latent space, 2024. URL https://arxiv.org/abs/2310.09656.

[2] Juntong Shi, Minkai Xu, et al. Tabdiff: a mixed-type diffusion model for tabular data generation, 2025. URL https://arxiv.org/abs/2410.20626.

[3] Akim Kotelnikov, Dmitry Baranchuk, et al. Tabddpm: modelling tabular data with diffusion models, 2023.

[4] Timur Sattarov, Marco Schreyer, and Damian Borth. Findiff: Diffusion models for financial tabular data generation, 2023. URL https://arxiv.org/abs/2309.01472.

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

[6] Christian P Robert, George Casella, and George Casella. *Monte Carlo statistical methods*, volume 2. Springer, 1999.

[7] Jayoung Kim, Chaejeong Lee, and Noseong Park. Stasy: Score-based tabular data synthesis, 2023. URL https://arxiv.org/abs/2210.04018.

[8] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. URL https://arxiv.org/abs/2105.05233.

[9] Jacob K Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models, 2024. URL https://arxiv.org/abs/2402.03559.

[10] Jacob K. Christopher, Michael Cardei, et al. Neuro-symbolic generative diffusion models for physically grounded, robust, and safe generation, 2025. URL https://arxiv.org/abs/2506.01121.

[11] Lei Xu, Maria Skoularidou, et al. Modeling tabular data using conditional gan, 2019. URL https://arxiv.org/abs/1907.00503.

[12] Haque Ishfaq, Assaf Hoogi, and Daniel Rubin. Tvae: Triplet-based variational autoencoder using metric learning, 2023. URL https://arxiv.org/abs/1802.04403.

[13] Kritik Seth. U.S. Airbnb Open Data. https://www.kaggle.com/datasets/kritikseth/us-airbnb-open-data, October 2020. Accessed: 2025-07-13.

[14] NY Mayor's Office of Special Enforcement. Registration law. https://www.nyc.gov/site/specialenforcement/registration-law/registration.page, 2025. Accessed: 2025-07-12.

[15] ethon0426 (Kaggle user). Lending club 2007–2020 q1. https://www.kaggle.com/datasets/ethon0426/lending-club-20072020q1, 2025. Accessed: 2025-07-15.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

[17] Jascha Sohl-Dickstein, Eric A. Weiss, et al. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

[18] Yang Song, Jascha Sohl-Dickstein, et al. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

[19] Yixin Liu, Kai Zhang, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models, 2024. URL https://arxiv.org/abs/2402.17177.

[20] Robin Rombach, Andreas Blattmann, et al. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[21] Jacob Austin, Daniel D. Johnson, et al. Structured denoising diffusion models in discrete state-spaces, 2023. URL https://arxiv.org/abs/2107.03006.

[22] Subham Sekhar Sahoo, Marianne Arriola, et al. Simple and effective masked diffusion language models, 2024. URL https://arxiv.org/abs/2406.07524.

[23] Yair Schiff, Subham Sekhar Sahoo, et al. Simple guidance mechanisms for discrete diffusion models, 2025. URL https://arxiv.org/abs/2412.10193.

[24] Kyungeun Lee, Ye Seul Sim, et al. Binning as a pretext task: Improving self-supervised learning in tabular domains, 2024. URL https://arxiv.org/abs/2405.07414.

[25] Shuhan Zheng and Nontawat Charoenphakdee. Diffusion models for missing value imputation in tabular data, 2023. URL https://arxiv.org/abs/2210.17128.

[26] Liudmila Prokhorenkova, Gleb Gusev, et al. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[27] U.S. Department of Justice, Civil Rights Division. The equal credit opportunity act. https://www.justice.gov/crt/equal-credit-opportunity-act-3, jan 2025. Accessed: July 18, 2025.

A Extended Related Works

Diffusion models have recently emerged as a dominant approach for tabular data synthesis. TabDDPM introduced separate kernels for numerical and categorical features [3], while subsequent models such as STaSy proposed self-paced curricula to enhance sample fidelity [7]. To improve efficiency, TabSyn performed diffusion in a compressed latent space learned via a VAE encoder [1]. TabDiff further addressed feature heterogeneity by employing feature-wise learnable noise schedules, enabling a single continuous-time model to adapt to the distributional characteristics of each column [2]. In the financial domain, state-of-the-art diffusion models have demonstrated growing promise. However, their stochastic sampling procedures provide no guarantees of adherence to the stringent business rules and regulatory requirements inherent to financial data. To mitigate this, FinDiff incorporated domain-specific architectures and financial pretraining to improve synthesis quality on structured financial datasets [4]. However, it still lacks mechanisms for enforcing strict constraints and cannot guarantee regulatory compliance during sampling.

To ensure outputs follow specific attributes, recent work has focused on integrating guidance directly into the diffusion process. Diffusion-based guidance methods steer samples toward desired attributes by modifying the score. Classifier guidance augments the score with the gradient of a pretrained classifier to bias generation toward a target class [8], while classifier-free guidance interpolates between conditional and unconditional scores removing the need for an external classifier [5]. Such guidance improves alignment with conditioning variables but does *not* provide hard, per-sample feasibility guarantees. Other work handles this with the integration of constraint optimization within the reverse diffusion process by recasting sampling as a constrained optimization problem. Projected Diffusion Models (PDM) introduced this idea for continuous data, using a projection operator after each denoising step to keep samples within a feasible set [9]. This projection-based approach was generalized across data types with the Neurosymbolic Diffusion framework of NSD [10]. We build on these works with, to our knowledge, the first study to implement and evaluate per-step hard constraint enforcement in mixed-type *tabular* diffusion with a focus on financial datasets and compliance-driven constraints.

B Preliminaries

Diffusion Models. Generative diffusion models [16–18] have revolutionized data synthesis, generating high-fidelity, state-of-the-art image and video samples [19, 20]. Diffusion models consist of a *forward* Markov chain that gradually corrupts a clean sample $\mathbf{x}_0 \sim p_{\text{data}}$ to approach the noisy distribution $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, followed by a learned *reverse* chain that reconstructs $\mathbf{x}_0$ from $\mathbf{x}_T$. Formally, the forward process $\{\mathbf{x}_t\}_{t=0}^{T}$ is determined by the kernel $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t \mathbf{I}\big)$, where the variance schedule $\{\beta_t\}$ increases monotonically. The reverse process is guided by a neural network $s_\theta(\mathbf{x}_t, t)$ trained to predict the score, $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$, where p_t is the marginal data distribution at time t. This score function defines a deterministic generative process via the probability flow ODE. To generate a sample, one solves this ODE backwards in time from $t = T$ to $t = 0$, typically with a numerical solver. A representative update step is:

$$\mathbf{x}_{t-\Delta t} = \mathbf{x}_t + g(t)s_\theta(\mathbf{x}_t, t)\Delta t$$

where $g(t)$ is a function of the noise schedule. This procedure deterministically transforms a noise vector $\mathbf{x}_T$ into a data sample $\mathbf{x}_0$ that approximates the true data distribution p_{data}.

Discrete Diffusion Models. Although diffusion was initially developed for continuous data, recent works extend this concept from continuous vectors to discrete token sequences [21, 22]. Each token is a one-hot vector $\mathbf{x} \in \{0, 1\}^V$ over a vocabulary of size V, where a sample is a length–L sequence $\mathbf{x}_0 = (\mathbf{x}_0^{(1)}, \ldots, \mathbf{x}_0^{(L)})$. The forward Markov chain replaces tokens with noise at a time-dependent rate β_t given by: $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = (1 - \beta_t)\, \mathbf{x}_{t-1} + \beta_t\, \nu$. Here, ν is either a uniform distribution [23] or a dedicated `[MASK]` symbol [22]. The reverse process is modeled as

$$x_{t-\Delta} = \begin{cases} \text{Cat}\left(x_{t-\Delta};\, x_t\right), & \text{if } x_t \neq \nu \\ \text{Cat}\left(x_{t-\Delta};\, \frac{\beta(t-\Delta)\, \nu + (\beta(t) - \beta(t-\Delta))\, s_\theta(x_t, t)}{\beta(t)}\right), & \text{if } x_t = \nu \end{cases}$$

where the learned denoiser, $s_\theta(x_t, t)$ approximates the sample x_0 and x_t is the probability distribution over V for each token in the sequence.

Mixed-Type Tabular Diffusion Models. Tabular data combines numerical $\mathbf{x}^{(\text{num})} \in \mathbb{R}^N$ and categorical $\mathbf{x}^{(\text{cat})} \in \prod_{k=1}^{C} \Delta_{K_k}$ features. Purely continuous diffusion fails to capture categorical structure, and purely discrete diffusion cannot represent real-valued geometry. Early tabular diffusion approaches tackled this by running separate processes for each feature type [3, 24], or by first embedding every row into a continuous latent space via a VAE and then diffusing there [1, 25], bypassing the need to handle categorical noise kernels directly. More recent work (TABDIFF) [2] models an entire table row $\mathbf{x}_t$ as a trajectory of a continuous-time diffusion process and assigns to each column j an independent, learnable noise schedule $\alpha_j(t) \in [0, 1]$. A shared Transformer encoder, conditioned on the partially corrupted row $\mathbf{x}_t$ and a time embedding $\tau(t)$, produces a time-aware representation $h_t = f_\theta(\mathbf{x}_t, \tau(t))$. Two specialized heads then predict (i) Gaussian score estimates for numerical features and (ii) categorical logits for discrete features. During generation, a reverse step denoises the numerical features while categorical tokens are resampled from the predicted logits, enabling a single model to reconstruct mixed-type tables jointly.

C Feasibility Operation for Mixed-Type Financial Data.

As financial data frequently contains both numerical and categorical features, the choice of the distance metric D is critical and must be tailored to the mixed-data-type nature. The distance metric is a sum of modality-specific distances:

$$D(\mathbf{x}, \hat{\mathbf{x}}) = D_{\text{num}}(\mathbf{x}^{\text{num}}, \hat{\mathbf{x}}^{\text{num}}) + D_{\text{cat}}(\mathbf{x}^{\text{cat}}, \hat{\mathbf{x}}^{\text{cat}}).$$

Numerical constraints in finance are typically affine, defining a convex polytope $\mathcal{C}^{\text{num}} = \{\mathbf{x} \in \mathbb{R}^N \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$. The natural distance metric for real-valued vectors is the squared Euclidean distance, $D_{\text{num}}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|^2$. The mapping is thus:

$$\mathcal{F}_{\mathcal{C}^{\text{num}}}(\hat{\mathbf{x}}^{\text{num}}) = \underset{\mathbf{x} \in \mathcal{C}^{\text{num}}}{\operatorname{argmin}} \|\mathbf{x} - \hat{\mathbf{x}}^{\text{num}}\|^2.$$

For categorical features, the diffusion model outputs a vector of logits for each column, which gets interpreted as an unnormalized log-probability distribution over the possible categories. Let $\hat{\mathbf{p}}_k = \text{softmax}(\hat{\mathbf{z}}_k)$ be the predicted probability vector for the k-th categorical feature.

The appropriate distance metric for probability distributions is the Kullback-Leibler (KL) divergence. The mapping finds a new distribution $\mathbf{q}_k$ supported only on the valid categories Ω_k that is closest to the model's prediction $\hat{\mathbf{p}}_k$:

$$\mathcal{F}_{\mathcal{C}_k^{\text{cat}}}(\hat{\mathbf{p}}_k) = \underset{\mathbf{q}_k \in \Delta_{K_k}, \text{supp}(\mathbf{q}_k) \subseteq \Omega_k}{\operatorname{argmin}} \text{KL}(\mathbf{q}_k \| \hat{\mathbf{p}}_k).$$

The solution to this KL operation has a simple and intuitive closed form: set the probabilities of all invalid categories to zero and re-normalize the probabilities of the valid ones. This preserves the relative likelihoods assigned by the model within the valid subset.

D Fidelity Evaluation

We evaluate generated data fidelity using four metrics. Marginal distribution quality is measured by average Shape Similarity; the average of Kolmogorov-Smirnov (KS) for continuous columns, and Total Variation Distance (TVD) for categorical columns. Distributional Distance consists of the average 1D-Wasserstein distance for continuous columns and Jensen-Shannon (JS) divergence for categorical columns. Bivariate dependencies are assessed via the Trend Score, which compares the pairwise Pearson's column correlations between the real and synthetic datasets for continuous columns and TVD for categorical columns. Finally, overall similarity is measured by the Detection AUC, the ROC-AUC score of a CatBoost classifier [26], which is trained to distinguish real from synthetic data, where 0.5 indicates perfect indistinguishability. The baseline benchmarks for all models were performed using their default parameters and using the same 90/10 train-test split, where the fidelity metrics were computed using the test split. All the experiments were performed using NVIDIA A30 GPUs.

E Extended Experimental Results

E.1 Lending Data Fidelity

We further evaluate the Fidelity of TabDiff for the Lending club dataset in Table 2. We observe improved marks in all fidelity sub-categories. This large-scale dataset, with nearly 3 million records, provides a granular view of borrowers' financial health, loan characteristics, and credit history. After preprocessing, our working dataset features 41 columns, representing a complex mixed-type environment. This includes 22 numerical features capturing both continuous and discrete values, such as loan amount and interest rates, alongside 19 categorical attributes that detail loan status and purpose.

Model	**Marginal Fidelity**		**Correlations**	**ML Similarity**
	Shape Sim. (Avg) ↑	Dist. Dist. (Avg) ↓	Trend Score ↑	Detection AUC → 0.5
CTGAN [11]	0.84188	0.10975	0.80008	0.99997
TVAE [12]	0.79720	0.00867	0.87228	0.99845
FinDiff [4]	0.96711	0.00637	0.94790	0.93186
Tabsyn [1]	0.99402	0.01338	0.96397	0.86901
CTDF (TabDiff [2])	**0.99345**	**0.00002**	**0.96519**	**0.70553**

Table 2: Quantitative Fidelity Benchmarks on the Lending Club Dataset. CTDF's base model (TabDiff) is compared against standard generative baselines. Arrows indicate the preferred direction for model quality.

E.2 Equitable Pricing Auditing and Algorithmic Gentrification Correction Simulation

Metric	Real	Unconstrained	CTDF
Shape ↑	–	0.988	0.983
Dist. Dist ↓	–	0.011	0.022
Trend ↑	–	0.948	0.946
Constraint Viol. (%)	42.64	42.34	**0.0**

Table 3: Constraint violation and fidelity results for unconstrained generation and CTDF.

Housing price models can absorb and amplify location effects tied to historical demographic patterns, leading to systematic over- or under-pricing across neighborhoods, referred to as algorithmic gentrification. In this work, we use CTDF to (i) measure these disparities and (ii) *simulate* corrective policies inside a controlled generative setting relevant to financial risk, revenue, and fairness. We train a regression network f_θ on *intrinsic* listing attributes (amenities, size, host / review features), deliberately excluding `latitude`, `longitude`, `neighbourhood_group`, and `neighbourhood`. Its prediction $\hat{p}^{\text{IFP}} = f_\theta(x^{\text{intr}})$ estimates a location-neutral (intrinsic) price. The inequity score is $e = p - \hat{p}^{\text{IFP}}$, with $e > 0$ interpreted as an over-pricing premium and $e < 0$ as discounted. Aggregating e over neighborhoods exposes systematic disparities.

An unconstrained TabDiff model is trained on the NYC subset over the full joint of intrinsic attributes, neighborhood indicators, and observed price p. To simulate the removal of premiums, we enforce the hard constraint $e \leq 0$. During reverse diffusion, each denoising step applies a per-step mapping on the price column, leaving all other attributes untouched. Samples already satisfying

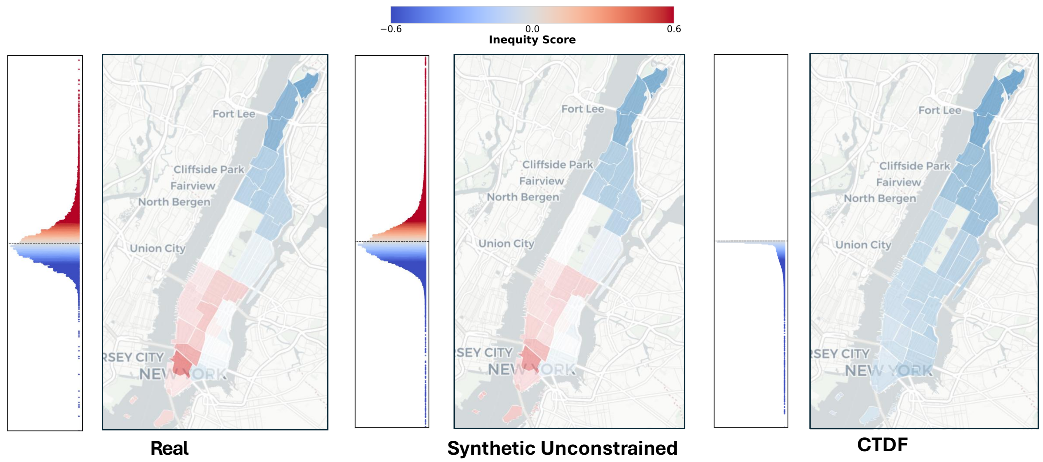


Figure 4: CTDF eliminates location-based price inequity in the NYC housing market. The inequity score (red = overpriced, blue = underpriced) reflects the difference between a listing's actual price and a location-blind "fair" price. *Left:* The real market shows significant price inflation in certain neighborhoods. *Center:* An unconstrained generative model learns and reproduces this geographical bias. *Right:* By constraining generated prices to match fair prices, CTDF creates a synthetic dataset purged of this bias, demonstrating its utility for algorithmic fairness and auditing.

$e \leq 0$ pass through unchanged; only over-priced listings are mapped to their intrinsic surrogate. For this application, both f_θ and the constraint are restricted to listings with `neighbourhood_group` = MANHATTAN; other NYC borough listings are excluded from fitting the surrogate and from constrained sampling.

Figure 4 (Real, Synthetic Unconstrained, CTDF) shows that the unconstrained model faithfully reproduces the spatial over-pricing pattern (positive tail in e and red neighborhoods), indicating the disparity is learnable. Under CTDF, the positive tail is cleanly truncated while neighborhood structure and variability remain. Table 3 confirms that constraint violations drop to 0% while marginal, distributional, and dependency fidelity metrics stay near the unconstrained baseline (a small divergence increase reflects removal of the premium tail). This yields a plausible counterfactual "fair-price" synthetic market stripped of location-driven markups. This simulation demonstrates CTDF's ability to offer a transparent mechanism to audit and then simulate alternative pricing regimes via systematic constraint enforcement.

The fidelity benchmarks presented in Table 2 confirm that the underlying TabDiff model again establishes a state-of-the-art performance.

E.3 Simulating a Compliant Prime Lending Portfolio

In consumer lending, lenders must prove that every product assignment is anchored in legitimate, measurable credit risk factors and never in

Metric	Real	Unconstrained	CTDF
Shape ↑	–	0.99344	0.94231
Dist. Dist. ↓	–	0.01651	0.11876
Trend ↑	–	0.95996	0.92126
Viol. (%)	90.51	90.22	**0.0**

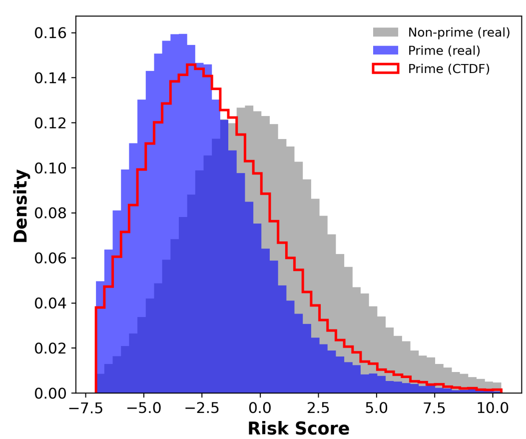


Figure 5: **Top:** Quantitative metrics (shape, distance, trend fidelity, and constraint violations). **Bottom:** Distribution of the composite Credit Risk Score. Lower values indicate safer borrowers. Non-prime loans (grey) cluster above 0; prime loans (blue) lie below; CTDF's synthetic prime loans (red outline) track the real prime profile.

prohibited attributes under laws such as the Equal Credit Opportunity Act (ECOA) [27]. To illustrate how a generative model can help satisfy a specific, policy-driven risk appetite, we tasked CTDF with synthesizing a portfolio drawn exclusively from the highest credit-quality segment of the market, a slice so small in the real data that traditional sampling is impractical for stress testing, capital planning, or training specialized risk models.

CTDF enforces four hard constraints on every synthetic loan: (1) `FICO_RANGE_LOW` $\geq$ 720 to guarantee prime-tier credit; (2) `DTI` $\leq$ 36 to cap leverage; (3) `VERIFICATION_STATUS` $\in$ {Verified, Source Verified} to reduce fraud risk; and (4) `SUB_GRADE` restricted to the A and B buckets. Only 10% of the original dataset meets all four rules simultaneously, yet CTDF generates 300k samples with zero violations, demonstrating that the model can learn and obey complex, multi-column business logic inside a narrowly defined risk band.

To gauge overall creditworthiness beyond the imposed boundaries, we construct a simple composite credit risk score by standardizing each feature, then summing the risk-increasing factors: interest rate, credit inquiries in the past six months, revolving-credit utilization, number of active bank-card lines, total revolving accounts, number of trade lines opened in the past 12 months, presence of a debt-settlement flag, and shorter time since the last delinquency, while subtracting the risk-reducing factors: annual income, percentage of accounts that have never been delinquent, and length of employment. Lower scores mark safer profiles. We then plot the distribution of this composite score for three sets: the broad non-prime population, the observed prime slice that meets all four hard constraints in the real data, and the synthetic prime loans produced by CTDF. Figure 5 displays these distributions, showing how the synthetic prime curve closely tracks the real prime histogram while remaining well-separated from the non-prime mass.

Feature (Mean)	Real (All)	Real (High DTI)	Synth. Hyp.
Loan Amount ($)	15,611.26	17,727.61	17,502.75
Interest Rate (%)	13.00	15.00	13.00
DTI	19.79	52.05	41.05
FICO Score (Low)	702.86	706.76	705.58

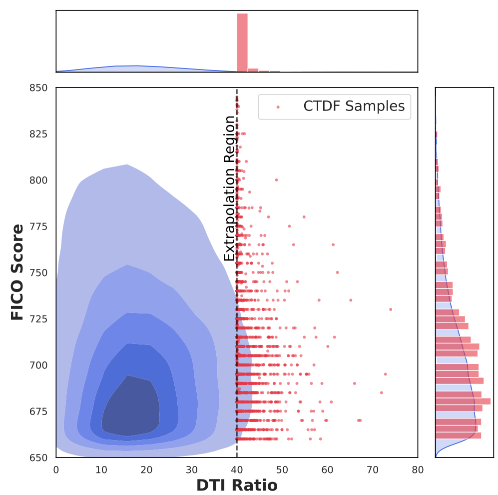


Figure 6: Top: Descriptive statistics Bottom: Realistic extrapolation of the DTI/FICO relationship; real joint density (blue) and synthetic constrained samples (red)

E.4 Hypothetical Scenario Planning and Extrapolation

Financial models must often reason about scenarios absent from historical data. Our analysis of the Lending Club dataset revealed one such "data desert": no records for mortgage-holders with both a high annual income (>$300k) and a high debt-to-income ratio (DTI > 40). To test if a generative model could fill this gap, we used CTDF for extrapolating 100k synthetic loans under these exact constraints.

The table results in Figure 6 reveal a sophisticated, non-obvious profile. While the synthetic borrowers sought large loans similar to the real high-DTI segment, the model assigned a moderate 13% interest rate instead of a punitive one. This indicates that CTDF learned a realistic economic trade-off: the borrowers' extreme income acts as a powerful risk mitigator, offsetting their high leverage and qualifying them for more favorable terms.

By ensuring extrapolated scenarios are grounded in learned economic principles, even uncovering non-naive trends. Hence, it builds the necessary trust for financial institutions to use synthetic data for strategic decision-making.