Conditional Diffusion Models for Imbalanced Tabular Regression

Nathaniel Kang[1]

[1]School of Computer Science and Engineering, Kyungpook National University
natekang@knu.ac.kr

Abstract

Imbalanced regression, where certain continuous target ranges are severely underrepresented, poses a fundamental challenge for predictive modeling under uncertainty. Existing oversampling methods rely on local interpolation, which fails to capture complex conditional distributions in rare target regions. We propose TABOVERSAMPLE, a conditional diffusion framework that generates high-fidelity synthetic tabular samples for underrepresented targets via a relevance-weighted denoising objective. We establish four theoretical results: equivalence of relevance weighting to maximum likelihood under a reweighted measure, targeted probability mass amplification in rare regions, quality guarantees for generate-then-filter sampling through order statistics, and a formal connection to distributionally robust optimization over a chi-squared uncertainty set. Across nine benchmark imbalanced regression datasets—four numerical-dominant and five categorical-rich—and against thirteen oversampling and reweighting baselines evaluated over ten seeds, TABOVERSAMPLE attains the best relevance-weighted error (SERA) on eight of nine datasets and the best rare-region RMSE on all nine, substantially improving prediction accuracy in rare target regions while maintaining overall performance.

1 INTRODUCTION

Regression tasks with highly skewed target distributions arise in hydrology (extreme floods) [Steininger et al., 2021], pharmacology (dose–response tails) [Yang et al., 2021], and real estate (luxury/distressed pricing), among other domains [Krawczyk, 2016, Zha et al., 2023]. In each setting, the **epistemic uncertainty** of predictive models is highest in the underrepresented target regions—which may occur at the tails of a unimodal distribution or at interior gaps of a multimodal one—precisely where accurate predictions matter most [Branco et al., 2016].

This problem, known as *imbalanced regression*, is the continuous-target analogue of class imbalance [Fernández et al., 2018]. While classification has a rich toolkit—SMOTE [Chawla et al., 2002], ADASYN [He et al., 2008], cost-sensitive learning [Elkan, 2001, Lin et al., 2017], and data augmentation [Shorten and Khoshgoftaar, 2019]—the regression setting has received far less attention. Torgo et al. [Torgo et al., 2013] introduced a *relevance function* $\phi(y) \in [0, 1]$ to identify rare target values, leading to SMOTER; subsequent extensions include SMOGN [Branco et al., 2017] and REBAGG [Branco et al., 2019]. However, all existing methods rely on interpolation or local perturbation, which assumes locally linear feature-target relationships and produces artifacts when the true conditional distribution $p(\mathbf{x} \mid y)$ is complex [Steininger et al., 2021].

Denoising diffusion probabilistic models (DDPMs) [Sohl-Dickstein et al., 2015, Ho et al., 2020] offer a compelling alternative, having achieved remarkable results across images [Dhariwal and Nichol, 2021, Rombach et al., 2022] and structured data [Nichol and Dhariwal, 2021]. TabDDPM [Kotelnikov et al., 2023] showed that diffusion models outperform GANs and VAEs for tabular synthesis [Borisov et al., 2024]. Diffusion models have also been adapted to imbalance in the *image* domain, through class rebalancing for long-tailed generation [Qin et al., 2023, Shao et al., 2024] and guided sampling from low-density regions [Sehwag et al., 2022]; these methods operate over discrete class labels or pixel-space density, whereas we target underrepresented regions of a *continuous* target in mixed-type tabular data (Appendix B). We propose TABOVERSAMPLE, a conditional diffusion framework for imbalanced tabular regression (Figure 1) that introduces a *relevance-weighted denoising objective* to focus model capacity on rare regions and a *target-conditioned sampling strategy* with quality fil-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:2688–2712.

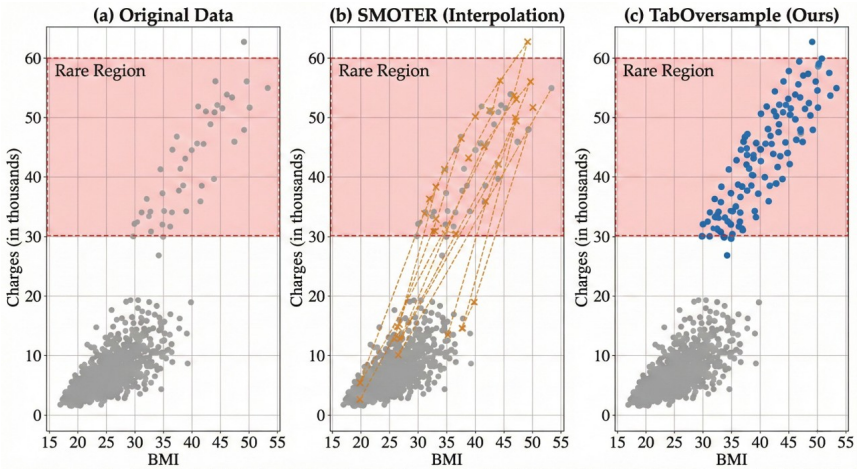

Figure 1: Comparison of oversampling strategies in the rare target region (shaded band above the dashed line at charges > 30K) of the Insurance dataset [Lantz, 2013]. **(a)** Original data: BMI vs. charges shows a dense cluster at low charges and a sparse tail at high charges. The relevance score $\phi(y)$ (Eq. 1) is near 0 for central charges (e.g., a \$5K charge near the median) and approaches 1 deep in the tail (e.g., a \$60K charge, $\phi \approx 0.9$); $\phi(y)$ is a derived rarity weight, *not* the prediction target. **(b)** SMOTER generates synthetic points (cross markers) along straight lines between neighbors—an artifact of convex interpolation that fails to capture nonlinear structure. **(c)** TABOVERSAMPLE generates synthetic points (filled circles) that follow the true conditional density, filling the rare region with diverse, realistic samples at the same generated-to-original ratio as SMOTER. A quantitative comparison against the TabDDPM-Cond baseline (same conditioning architecture, no relevance weighting) is reported in Table 1. Best viewed in color; the rare-region boundary is also marked by dashed lines for grayscale readability. This example illustrates the *unimodal right-tail* imbalance that is the focus of this work.

tering. Our contributions are: **(1)** the first application of DDPMs to imbalanced tabular regression; **(2)** four theorems establishing connections to maximum likelihood, probability mass amplification, order statistics, and distributionally robust optimization; **(3)** a principled generate-then-filter strategy with provable quality guarantees.

Prior work in imbalanced regression relies primarily on local interpolation—SMOTER [Torgo et al., 2013], SMOGN [Branco et al., 2017], and REBAGG [Branco et al., 2019]—which assumes locally linear feature-target relationships and fails to capture complex conditional distributions [Steininger et al., 2021, Yang et al., 2021]. While deep generative models have advanced tabular synthesis [Xu et al., 2019, Kotelnikov et al., 2023, Jolicoeur-Martineau et al., 2024], their application to continuous imbalanced targets remains unexplored. We defer a comprehensive review of related work—including connections to distributionally robust optimization (DRO) [Ben-Tal et al., 2009, Duchi and Namkoong, 2021]—to Appendix B. Throughout, we focus on the *unimodal* setting, in which rare targets lie in the tails of a single-mode distribution; the multimodal case, where minority modes may be embedded within the bulk of $p(y)$, is left to future work and can be accommodated by substituting a density-based relevance function [Steininger et al., 2021] as discussed in Section 2.

2 PROBLEM FORMULATION

Consider a supervised regression dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ where each feature vector $\mathbf{x}_i \in \mathbb{R}^d$ is associated with a continuous target $y_i \in \mathbb{R}$. Let $p(y)$ denote the marginal distribution of the target variable. The dataset exhibits *imbalanced regression* when $p(y)$ is highly non-uniform: certain regions of the target space are densely populated while others are severely underrepresented. This imbalance causes standard regression models to achieve low average error at the expense of poor accuracy—and high predictive uncertainty—in underrepresented regions [Branco et al., 2016].

To quantify the degree of underrepresentation, we adopt the *relevance function* framework [Torgo and Ribeiro, 2009]. A relevance function $\phi : \mathbb{R} \to [0, 1]$ maps each target value to an importance score, where higher values indicate greater relevance (corresponding to rarer target values). Following the box-plot-based construction [Torgo and Ribeiro, 2009, Torgo et al., 2013], let Q_1, Q_2, and Q_3 denote the first quartile, median, and third quartile of $\{y_i\}_{i=1}^N$, with $\mathrm{IQR} = Q_3 - Q_1$. The relevance function is:

$$\phi(y) = \begin{cases} \min\left(1, \dfrac{|y - Q_2|}{c \cdot \mathrm{IQR}}\right) & \text{if } |y - Q_2| > \dfrac{\mathrm{IQR}}{2}, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $c > 0$ controls the rate at which relevance increases for values departing from the central region. This formu-

lation assigns zero relevance to target values within the interquartile range and monotonically increasing relevance to values that deviate further from the median. We note that this construction assumes a unimodal, approximately symmetric target distribution; for multimodal targets, the global median and IQR may fail to identify local minority modes embedded within the interquartile range. Kernel density estimation alternatives such as DenseWeight [Steininger et al., 2021] offer greater flexibility at the cost of additional hyperparameters; we adopt the box-plot formulation for its simplicity and consistency with the imbalanced regression literature [Torgo and Ribeiro, 2009, Torgo et al., 2013, Branco et al., 2017].

Interpreting y and $\phi(y)$. The target $y \in \mathbb{R}$ is the original continuous quantity of interest (e.g., medical charges in dollars, house prices), and the downstream model predicts y on its native scale; $\phi(y)$ is never itself a prediction target. The relevance score $\phi(y) \in [0, 1]$ is a *derived* statistic that measures how far y departs from the central tendency relative to the spread, and is neither a percentile nor a probability. For the Insurance dataset, a charge near the median ($5,000) has $\phi \approx 0$, whereas a charge far in the right tail ($60,000) has $\phi \approx 0.9$. Its sole role is to determine which target values are underrepresented and therefore warrant additional synthetic coverage.

Distribution-free construction. Equation (1) depends only on the three order statistics Q_1, Q_2, and Q_3, which are well defined for any continuous distribution and require neither normality nor symmetry. The qualifier "unimodal" refers solely to the use of a single median anchor, not to a Gaussian assumption on $p(y)$, and the diffusion model itself makes no parametric assumption about the target marginal. For multimodal targets, the kernel-density relevance of DenseWeight [Steininger et al., 2021] serves as a drop-in replacement: every result in Section 3 depends only on the existence of a weight function $w(y) \geq 1$, not on how $\phi(y)$ is computed.

Behavior under a uniform target. Relevance weighting is self-correcting when no imbalance is present. Under $Y \sim U(a, b)$ the quartiles satisfy $Q_2 = (a + b)/2$ and $\mathrm{IQR} = (b - a)/2$, so by Equation (1) the zero band $|y - Q_2| \leq \mathrm{IQR}/2$ coincides with $[Q_1, Q_3]$, the middle 50%. Relevance is therefore $\phi(y) = 0$ across the interquartile range and strictly positive only on the outer quartiles, attaining its extreme value $\phi(a) = \phi(b) = \min(1, 1/c)$ at the endpoints. The induced perturbation is bounded—the maximum weight ratio is $1 + \lambda \min(1, 1/c)$—and symmetric about Q_2, so it introduces no directional bias; its strength is governed by λ, with $\lambda = 0$ recovering the unweighted objective exactly (the ablation in Section 5 sweeps λ down to this limit). Consequently, on genuinely balanced data the method degrades gracefully toward standard conditional diffusion rather than manufacturing spurious rare regions.

Denoising Diffusion for Tabular Data. DDPMs [Ho et al., 2020] define a *forward process* that gradually corrupts data by adding Gaussian noise over T time steps and a learned *reverse process* that recovers the data by iterative denoising. For a data point $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the forward process produces increasingly noisy latents:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\Big(\mathbf{x}_t;\ \sqrt{1-\beta_t}\,\mathbf{x}_{t-1},\ \beta_t \mathbf{I}\Big)\,, \quad (2)$$

where $\{\beta_t\}_{t=1}^T$ is a variance schedule. Defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$, one can sample $\mathbf{x}_t$ directly:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\big(\mathbf{x}_t;\ \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\ (1 - \bar{\alpha}_t)\mathbf{I}\big)\,. \quad (3)$$

The reverse process is parameterized by a neural network $\boldsymbol{\epsilon}_\theta$ that predicts the noise:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1};\ \boldsymbol{\mu}_\theta(\mathbf{x}_t, t),\ \sigma_t^2 \mathbf{I}\big)\,, \quad (4)$$

where $\boldsymbol{\mu}_\theta$ is derived from $\boldsymbol{\epsilon}_\theta$ and $\sigma_t^2 = \beta_t$ follows the DDPM convention [Ho et al., 2020], corresponding to the upper bound of the posterior variance. The standard training objective minimizes:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{(\mathbf{x}_0, y),\ \boldsymbol{\epsilon},\ t}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)\|^2\right], \quad (5)$$

where the expectation is over $(\mathbf{x}_0, y) \sim \mathcal{D}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $t \sim \mathcal{U}(1, T)$. For tabular data, TabDDPM [Kotelnikov et al., 2023] applies Gaussian diffusion to numerical columns and multinomial diffusion (following the categorical diffusion framework of Hoogeboom et al. [Hoogeboom et al., 2021]) to categorical columns, with the target y serving as a conditioning signal. Concretely, each categorical column with K states is corrupted by a discrete forward process whose transition matrix $\mathbf{Q}_t$ interpolates each one-hot category toward the uniform distribution over the K states; the network predicts the original category, and the per-step categorical objective is the posterior KL divergence

$$\mathcal{L}_t^{\text{cat}} = \mathrm{KL}\big(q(\mathbf{c}_{t-1} \mid \mathbf{c}_t, \mathbf{c}_0) \,\big\|\, p_\theta(\mathbf{c}_{t-1} \mid \mathbf{c}_t, t, y)\big)\,, \quad (6)$$

where $\mathbf{c}_t$ denotes the noised categorical state at step t. Numerical columns use the Gaussian objective of Equation (5). We make this dual-path structure—and the way relevance weighting multiplies the combined per-sample loss—explicit in Section 4 and Appendix D.

Goal. Our goal is to learn a generative model G_θ that synthesizes samples $\mathcal{S} = \{(\mathbf{x}'_j, y'_j)\}_{j=1}^M$ such that the augmented dataset $\mathcal{D} \cup \mathcal{S}$ exhibits a more balanced target distribution while faithfully preserving the conditional distribution $p(\mathbf{x} \mid y)$. We seek to reduce the epistemic uncertainty of downstream regression models in rare target regions through principled synthetic data augmentation with provable guarantees.

3 THEORETICAL FRAMEWORK

We now develop the theoretical foundations of relevance-weighted diffusion training. The three results in this section establish that the TABOVERSAMPLE objective is statistically principled from multiple complementary perspectives: it corresponds to maximum likelihood under a reweighted measure (Theorem 1), it provably amplifies the probability mass allocated to rare regions (Theorem 2), and its quality filtering admits sharp bounds via order statistics (Theorem 3). Furthermore, in Appendix C.4, under an explicit affine loss-alignment assumption (Assumption 1), we establish a formal equivalence between our relevance-weighted objective and distributionally robust optimization (DRO) over a chi-squared uncertainty set defined on the *joint* distribution of $(\mathbf{x}, y)$ (Theorem 4), providing a complementary robustness justification for the framework.

We define the relevance-weighted denoising objective as:

$$\mathcal{L}_{\text{weighted}} = \mathbb{E}_{(\mathbf{x}_0,y)\sim\mathcal{D},\,\boldsymbol{\epsilon},\,t}\left[w(y)\cdot\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t,y)\|^2\right], \tag{7}$$

where the sample weight is $w(y) = 1 + \lambda \cdot \phi(y)$ and $\lambda \geq 0$ controls the strength of relevance weighting. When $\lambda = 0$, the objective reduces to the standard DDPM loss; as λ increases, the model devotes progressively more capacity to accurately denoising samples from rare target regions, precisely where the model's uncertainty about the conditional distribution is greatest.

3.1 REWEIGHTED MAXIMUM LIKELIHOOD EQUIVALENCE

Theorem 1 (Reweighted Maximum Likelihood Equivalence)**.** *Let $p(\mathbf{x}, y)$ denote the joint data distribution, $w(y) = 1 + \lambda \cdot \phi(y)$ the relevance weight function, and $Z_w = \mathbb{E}_{p(y)}[w(y)]$ the normalizing constant. Define the reweighted distribution*

$$\tilde{p}(\mathbf{x}, y) = \frac{w(y)\cdot p(\mathbf{x}, y)}{Z_w}. \tag{8}$$

Then minimizing the weighted loss $\mathcal{L}_{\text{weighted}}$ with weights $w(y)$ is equivalent to maximizing the log-likelihood under $\tilde{p}$:

$$\begin{aligned}&\arg\min_\theta\ \mathbb{E}_{p(\mathbf{x},y)}[w(y)\cdot\ell(\theta;\mathbf{x},y)]\\ &\quad=\arg\max_\theta\ \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\log p_\theta(\mathbf{x}\mid y)]\,,\end{aligned} \tag{9}$$

where $\ell(\theta;\mathbf{x},y) = \|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t,y)\|^2$ is the per-sample denoising loss.

Proof sketch. The result follows by expanding the expectation under $\tilde{p}$ and noting that the normalization constant Z_w is independent of θ. Substituting $\tilde{p}(\mathbf{x}, y) = w(y)p(\mathbf{x}, y)/Z_w$ into the right-hand side yields a constant multiple of the weighted expected loss, and since the DDPM objective is equivalent to a weighted ELBO on $\log p_\theta(\mathbf{x} \mid y)$ in the continuous-time limit [Ho et al., 2020, Song et al., 2021b], the argmin and argmax coincide. The complete proof is provided in Appendix C.1. □

Remark 1. *Theorem 1 reveals that relevance weighting does not merely heuristically upweight rare samples; it implements exact maximum likelihood estimation under a well-defined probability measure $\tilde{p}$ that places greater mass on regions of high relevance. The reweighted distribution $\tilde{p}$ can be interpreted as the analyst's* target distribution*—the distribution under which the model's predictive uncertainty should be minimized.*

3.2 TARGETED PROBABILITY MASS AMPLIFICATION

Before stating the amplification result, we formally define the weighted probability mass of a rare region, a quantity that appears throughout this section.

Definition 1 (Weighted probability mass)**.** *Let $R_\tau = \{y : \phi(y) > \tau\}$ denote the rare region at threshold $\tau \in (0, 1]$. Under relevance weighting with $w(y) = 1 + \lambda \cdot \phi(y)$, the* weighted probability mass *of R_τ is*

$$\pi_w(R_\tau) = \frac{\sum_{i:\,y_i\in R_\tau} w(y_i)}{\sum_{j=1}^N w(y_j)}, \tag{10}$$

whereas the unweighted empirical measure is $\pi(R_\tau) = |\{i : y_i \in R_\tau\}|/N$.

Theorem 2 (Targeted Probability Mass Amplification)**.** *Let $R_\tau = \{y : \phi(y) > \tau\}$ denote the rare region at threshold $\tau \in (0, 1]$, let $n_\tau = |\{i : y_i \in R_\tau\}|$ be the number of training samples falling in R_τ, and let $\bar{\phi} = N^{-1}\sum_{i=1}^N \phi(y_i)$ denote the average relevance. Under relevance weighting with $w(y_i) = 1 + \lambda \cdot \phi(y_i)$ and the mild condition $\tau > \bar{\phi}$ (i.e., the rare region threshold exceeds the population average relevance), the probability mass allocated to the rare region under the reweighted distribution satisfies:*

$$\pi_w(R_\tau) \geq \frac{(1+\lambda\tau)\cdot n_\tau}{N + \lambda\sum_{i=1}^N \phi(y_i)}, \tag{11}$$

which is strictly greater than n_τ/N (the empirical measure under uniform weighting).

Proof sketch. The lower bound on the numerator follows from $\phi(y_i) > \tau$ for all $y_i \in R_\tau$. Cross-multiplying the claimed inequality $\pi_w(R_\tau) > n_\tau/N$ yields the condition $N\tau > \sum_{i=1}^N \phi(y_i)$, which is equivalent to $\tau > \bar{\phi}$. This condition is mild: since $\phi(y) = 0$ for all target values within the interquartile range (typically the majority of samples),

the average relevance $\bar{\phi}$ is small for imbalanced datasets, and any threshold identifying genuinely rare regions satisfies $\tau > \bar{\phi}$. See Appendix C.2 for the complete derivation. □

The amplification factor $(1 + \lambda\tau)$ grows with both λ and τ, so the most underrepresented regions receive the greatest boost. Increasing the probability mass allocated to a region directly reduces the posterior variance of the learned conditional density estimate.

3.3 QUALITY FILTERING VIA ORDER STATISTICS

We now establish a quality guarantee for a general generate-then-filter procedure that is independent of the diffusion model. Suppose a generator produces m candidate samples, each assigned a scalar *quality score* S (lower is better), and we retain the n candidates with the smallest scores. We refer to $r = m/n$ as the *oversampling ratio*: $m = rn$ candidates are generated and n are kept. The following result, a direct consequence of order-statistics theory, bounds the quality of the worst retained sample for *any* continuous score distribution, with no reference to diffusion or to how the score is computed.

Theorem 3 (Quality Filtering via Order Statistics)**.** *Let $S_1, \ldots, S_m$ be i.i.d. scores with continuous CDF F, finite mean $\mu = \mathbb{E}[S_i]$, and non-degenerate distribution. Let $S_{(1)} \leq \cdots \leq S_{(m)}$ denote the order statistics. If $m = 3n$ candidates are generated (i.e., $r = 3$) and the top n (lowest scores) are retained, then:*

$$\mathbb{E}\big[F(S_{(n)})\big] = \frac{n}{3n+1} < \frac{1}{2}, \quad \textit{and} \quad \mathbb{E}\big[S_{(n)}\big] < \mu. \tag{12}$$

That is, the expected quantile of the worst retained sample lies strictly below the median, and its expected score is strictly below the population mean.

Proof sketch. The k-th order statistic from m i.i.d. samples with continuous CDF F satisfies $F(S_{(k)}) \sim \mathrm{Beta}(k, m - k+1)$, which has mean $k/(m+1)$. For $k = n$ and $m = 3n$, the expected quantile is $n/(3n + 1) < 1/3$. The strict inequality $\mathbb{E}[S_{(n)}] < \mu$ follows from a stochastic dominance argument: the $\mathrm{Beta}(n, 2n + 1)$ density concentrates below $u = 1/2$, so integrating the non-decreasing quantile function F^{-1} against this density yields a strictly smaller value than integrating against the uniform density (which gives μ). Crucially, this argument holds for *any* non-degenerate F without requiring concavity of F^{-1}. See Appendix C.3. □

Instantiating the score via denoising self-consistency. TABOVERSAMPLE instantiates the generic score S of Theorem 3 with a *denoising score evaluation* (DSE). Given a generated sample $\mathbf{x}_0$, we draw fresh noise $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a random timestep $t' \sim \mathcal{U}(1, T)$, form $\mathbf{x}_{t'} = \sqrt{\bar{\alpha}_{t'}}\,\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_{t'}}\,\boldsymbol{\epsilon}'$, and set $S = \|\boldsymbol{\epsilon}' - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t'}, t', y)\|^2$. Since $\boldsymbol{\epsilon}'$ is known, S is fully computable at inference time and measures how well the model recovers injected noise—lower scores indicate higher fidelity to the learned conditional density. The i.i.d. hypothesis of Theorem 3 then holds conditionally: for a fixed model and independently generated candidates, the DSE scores are independent, and the resulting guarantee is relative to the model's own average generation quality rather than to the ground truth. We set $r = 3$ as the default, placing the worst retained sample at the $n/(3n + 1)$ quantile—roughly the 25th percentile, versus the 50th for $r = 1$—for a substantial quality gain at $3\times$ generation cost, with diminishing returns beyond $r = 3$ (validated in the ablation study). Whether low DSE truly correlates with downstream utility is an empirical question, which we answer affirmatively through a held-out stratification experiment in Section 5.

4 THE TABOVERSAMPLE ALGORITHM

Building on the theoretical foundations of Section 3, we now present the complete TABOVERSAMPLE framework (Figure 2). The algorithm modifies the standard conditional tabular diffusion pipeline through a relevance-weighted training objective, a target-conditioned sampling strategy with quality filtering, and a principled integration scheme with downstream regression models.

4.1 RELEVANCE-WEIGHTED DIFFUSION TRAINING

The standard TabDDPM training objective treats all training examples equally, regardless of their target values. When the training set is imbalanced, this leads to a model that allocates most of its capacity to densely populated regions of $p(y)$, learning $p(\mathbf{x} \mid y)$ accurately for common target values but poorly—with high uncertainty—for rare ones. By Theorem 1, incorporating the relevance weight $w(y) = 1+\lambda \cdot \phi(y)$ into the denoising objective implements exact maximum likelihood estimation under the reweighted measure $\tilde{p}$, which places greater mass on underrepresented target regions.

For tabular data containing both numerical and categorical features, we follow TabDDPM [Kotelnikov et al., 2023] in applying Gaussian diffusion to numerical columns and multinomial diffusion [Hoogeboom et al., 2021] to categorical columns. The relevance weight applies uniformly to both loss components:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{(\mathbf{x}_0, y) \sim \mathcal{D}} \left[w(y) \cdot \left(\mathcal{L}_t^{\text{num}} + \mathcal{L}_t^{\text{cat}} \right) \right], \tag{13}$$

where $\mathcal{L}_t^{\text{num}}$ and $\mathcal{L}_t^{\text{cat}}$ denote the Gaussian and multinomial diffusion losses at time step t, respectively.

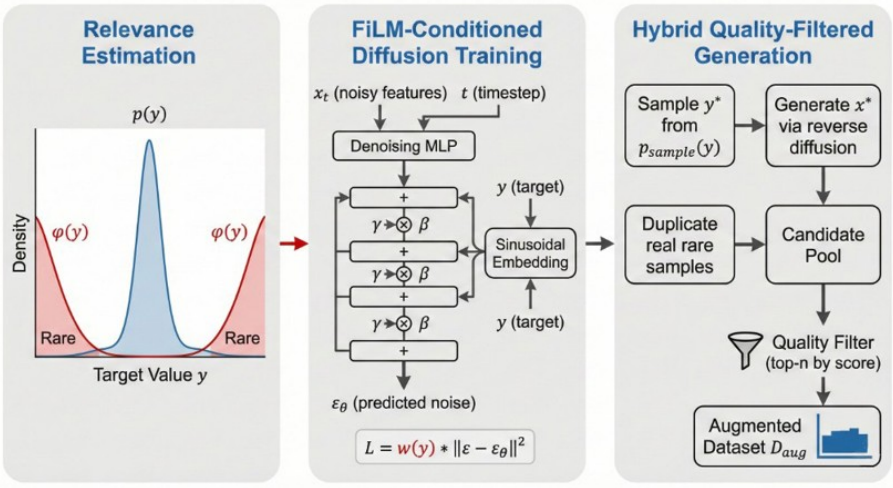

Figure 2: The TABOVERSAMPLE framework. **Left**: The relevance function $\phi(y)$ identifies underrepresented target regions (red shading) from the training data's target distribution. **Center**: A conditional diffusion model is trained with relevance-weighted loss, where the FiLM conditioning mechanism injects target information at every hidden layer. **Right**: At generation time, target values are drawn from the rebalanced distribution $p_{\text{sample}}(y)$; synthetic features are generated via conditional reverse diffusion; a hybrid candidate pool (real rare duplicates + synthetic) is quality-filtered via denoising score evaluation to retain the top-n samples.

4.2 TARGET-CONDITIONED SAMPLING WITH QUALITY FILTERING

At generation time, the goal is to produce synthetic samples whose target values preferentially cover the underrepresented regions while maintaining high fidelity.

The generation process begins by sampling target values from the rebalanced distribution:

$$p_{\text{sample}}(y) = \frac{\phi(y) \cdot p(y)}{Z}, \quad Z = \int \phi(y) \cdot p(y)\, dy. \quad (14)$$

In practice, we estimate $p(y)$ via kernel density estimation on $\{y_i\}_{i=1}^N$, compute the relevance-weighted density, and normalize to obtain $\hat{p}_{\text{sample}}(y)$. We draw M target values $\{y_j^*\}_{j=1}^M$ from this distribution.

Given the sampled target values, corresponding feature vectors are generated through the conditional reverse diffusion process. The target value is embedded via sinusoidal positional encoding:

$$\mathbf{e}(y^*) = \left[\sin(\omega_k y^*), \cos(\omega_k y^*)\right]_{k=1}^K, \quad (15)$$

where $\{\omega_k\}_{k=1}^K$ are frequency parameters spanning multiple scales. Following the timestep encoding of Ho et al. [Ho et al., 2020] and the positional encoding of Vaswani et al. [Vaswani et al., 2017], this multi-scale representation endows the network with simultaneous sensitivity to coarse differences (e.g., $y^* = 10{,}000$ vs. $50{,}000$) and fine-grained distinctions (e.g., $y^* = 10{,}000$ vs. $10{,}500$); a constant vector $[y^*, \ldots, y^*]$ would collapse all conditioning information into a single scale, hindering discrimination between nearby target values that require distinct conditional distributions. It is important to distinguish the roles of the input features $\mathbf{x}_t$ and the conditioning signal (y, t): the noisy feature vector $\mathbf{x}_t$ (containing both standardized numerical columns and embedded categorical columns) enters the denoising MLP as the primary input, while the target value y and timestep t modulate the network's hidden representations via feature-wise linear modulation (FiLM) [Perez et al., 2018]. This separation ensures that the conditioning signal influences all feature types uniformly without requiring feature-type-specific injection mechanisms. Concretely, at each hidden layer l with pre-activation $\mathbf{h}^{(l)}$:

$$\tilde{\mathbf{h}}^{(l)} = \boldsymbol{\gamma}^{(l)}(\mathbf{e}(y^*), \mathbf{e}(t)) \odot \mathbf{h}^{(l)} + \boldsymbol{\beta}^{(l)}(\mathbf{e}(y^*), \mathbf{e}(t)), \quad (16)$$

where $\boldsymbol{\gamma}^{(l)}$ and $\boldsymbol{\beta}^{(l)}$ are learned scale and shift functions.

Following Theorem 3, we apply a generate-then-filter strategy: for each target value y_j^*, we generate 3 candidate feature vectors through the reverse diffusion process. Each candidate $\mathbf{x}_0^{(k)}$ is then evaluated via a denoising score evaluation (DSE) step: fresh noise $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and a random timestep t' are sampled, the candidate is re-noised to $\mathbf{x}_{t'} = \sqrt{\bar{\alpha}_{t'}}\, \mathbf{x}_0^{(k)} + \sqrt{1 - \bar{\alpha}_{t'}}\, \boldsymbol{\epsilon}'$, and the score $S^{(k)} = \|\boldsymbol{\epsilon}' - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t'}, t', y_j^*)\|^2$ is computed. The candidate with the lowest score is retained. When classifier-free guidance (Appendix D) is employed, the reverse process uses the guided prediction $\hat{\boldsymbol{\epsilon}}_\theta$ (Equation 22), while the DSE scoring step uses the *unguided* $\boldsymbol{\epsilon}_\theta$; this avoids artificially inflating the self-consistency score via the guidance amplification, ensuring that the quality filter measures the model's intrinsic fidelity rather than the guidance-boosted prediction. This single additional forward pass per candidate is negligible

Algorithm 1 TABOVERSAMPLE Generation with Quality Filtering

Require: Trained model $\boldsymbol{\epsilon}_\theta$, number of samples M, sampling distribution $p_{\text{sample}}(y)$, oversampling ratio $r = 3$, guidance scale $s \geq 0$
Ensure: Synthetic dataset $\mathcal{S} = \{(\mathbf{x}'_j, y'_j)\}_{j=1}^{M}$

$\mathcal{S} \leftarrow \emptyset$
for $j = 1$ to M **do**
  Sample target: $y_j^* \sim p_{\text{sample}}(y)$
  for $k = 1$ to r **do**
    Initialize: $\mathbf{x}_T^{(k)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
    for $t = T$ down to 1 **do**
      $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
      $\mathbf{x}_{t-1}^{(k)} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t^{(k)} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t^{(k)}, t, y_j^*) \right) + \sigma_t \mathbf{z}$ *(guided, Eq. 22)*
    end for
    Sample $\boldsymbol{\epsilon}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t' \sim \mathcal{U}(\{1, \ldots, T\})$
    Compute $\mathbf{x}_{t'}^{(k)} = \sqrt{\bar{\alpha}_{t'}}\, \mathbf{x}_0^{(k)} + \sqrt{1 - \bar{\alpha}_{t'}}\, \boldsymbol{\epsilon}'$
    Compute score: $S_j^{(k)} = \|\boldsymbol{\epsilon}' - \boldsymbol{\epsilon}_\theta(\mathbf{x}_{t'}^{(k)}, t', y_j^*)\|^2$ *(unguided $\boldsymbol{\epsilon}_\theta$)*
  end for
  Retain: $\mathbf{x}'_j = \arg\min_k S_j^{(k)}$
  $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\mathbf{x}'_j, y_j^*)\}$
end for
return $\mathcal{S}$

relative to the T-step reverse process, and ensures that retained samples are certifiably above the average generation quality.

The reverse process iterates for T steps, where y^* enters the denoising network through its sinusoidal embedding $\mathbf{e}(y^*)$ (Equation 15) via the FiLM mechanism (Equation 16):

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y^*) \right) + \sigma_t \mathbf{z}, \quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{17}$$

4.3 INTEGRATION WITH DOWNSTREAM MODELS

The synthetic samples $\mathcal{S}$ are combined with $\mathcal{D}$ to form $\mathcal{D}_{\text{aug}} = \mathcal{D} \cup \mathcal{S}$, on which the downstream regressor is trained. The complete training procedure (Algorithm 2) is provided in Appendix D; the generation procedure is summarized in Algorithm 1.

5 EXPERIMENTS

We evaluate TABOVERSAMPLE on nine benchmark imbalanced regression datasets. Section 5 reports the four numerical-dominant benchmarks (**Abalone**, **CPUActivity**, **CaliforniaHousing**, **House16H**; statistics in Table 9, Appendix F), following prior benchmarks [Branco et al., 2017, Steininger et al., 2021]; Section 5.2 adds five categorical-rich benchmarks (43–62% categorical features). We compare against a broad set of baselines spanning reweighting, interpolation, and generative families: no oversampling (**None**), ϕ-weighted CatBoost (**None+ϕ**; sample weights $\phi(y_i)$ with no synthesis), random oversampling (**RandomOS**), **SMOTER** [Torgo et al., 2013], **SMOGN** [Branco et al., 2017], Calibrated Mixup (**Cal. Mixup**) [Kang et al., 2026], **TabDDPM** [Kotelnikov et al., 2023] (unconditional diffusion), and **TabDDPM-Cond** [Kotelnikov et al., 2023] (conditional diffusion with target conditioning but without relevance weighting). Additional baselines—DenseWeight [Steininger et al., 2021], deep imbalanced regression via LDS [Yang et al., 2021], CTGAN/TVAE [Xu et al., 2019] with bin conditioning, RandomOS+Gaussian, and a regression-adapted variant of Yang et al. [Yang et al., 2025]—are reported in Appendix G. CatBoost [Prokhorenkova et al., 2018] serves as the primary downstream regressor, with XGBoost [Chen and Guestrin, 2016] and MLP results in Appendix G [Gorishniy et al., 2021]. We report **SERA** [Torgo and Ribeiro, 2009] (relevance-weighted squared error), **RMSE$_{\text{rare}}$** (RMSE on samples with $\phi(y) > 0.5$), and standard **RMSE**, averaged over 10 independent seeds with identical splits across methods to enable paired statistical testing. Unless otherwise noted, $\lambda = 2.0$. Full setup details are in Appendix E.

5.1 MAIN RESULTS

Tables 1 (SERA) and 2 (RMSE$_{\text{rare}}$ and overall RMSE) present the primary results with CatBoost as the downstream regressor on the four numerical-dominant benchmarks. We organize the discussion around three key findings.

Finding 1: TABOVERSAMPLE consistently improves rare-region accuracy over both diffusion baselines. The most informative comparison is against TabDDPM-Cond, which shares TABOVERSAMPLE's conditional architecture and differs only in the relevance weighting, isolating the contribution of $w(y)$. TABOVERSAMPLE reduces SERA relative to TabDDPM-Cond by 2.3% (Abalone), 2.5% (CPUActivity), 7.2% (CaliforniaHousing), and 2.2% (House16H), and lowers RMSE$_{\text{rare}}$ on all four (Table 2); against the unconditional TabDDPM the SERA gains are larger still (up to 8.8% on CaliforniaHousing). This validates that learning under the reweighted measure $\tilde{p}$ (Theorem 1) improves density estimates precisely where epistemic uncertainty is highest.

Finding 2: TABOVERSAMPLE achieves the strongest consistency across all methods. TABOVERSAMPLE attains the best SERA on three of four datasets and second on House16H (average SERA rank 1.25 across the nine compared methods) and ranks first on RMSE$_{\text{rare}}$ on all four (average rank 1.00). Unlike RandomOS, which merely dupli-

Table 1: **SERA** ↓ with **CatBoost** as downstream regressor on the four numerical-dominant benchmarks (mean ± std over 10 seeds). **Bold**: best per column. Underline: second best. TABOVERSAMPLE attains the best SERA on three of four datasets; on House16H, RandomOS leads by a margin within one standard deviation, while TABOVERSAMPLE leads on $\text{RMSE}_{\text{rare}}$ (Table 2).

Method	**Abalone**	**CPUActivity**	**CaliforniaHousing**	**House16H**
None	2.946±.030	2.453±.120	0.112±.003	795.0M±7.5M
None+ϕ	2.851±.028	2.381±.110	0.108±.003	785.2M±7.0M
RandomOS	2.759±.032	2.307±.095	0.104±.003	**769.1M±6.2M**
SMOTER [Torgo et al., 2013]	2.849±.033	2.423±.150	0.110±.003	776.3M±6.5M
SMOGN [Branco et al., 2017]	2.878±.034	2.343±.130	0.109±.003	805.9M±8.0M
Cal. Mixup [Kang et al., 2026]	2.821±.029	2.348±.100	0.106±.003	782.4M±6.8M
TabDDPM [Kotelnikov et al., 2023]	2.888±.030	2.208±.115	0.113±.004	818.6M±7.3M
TabDDPM-Cond [Kotelnikov et al., 2023]	2.813±.035	2.261±.102	0.111±.004	789.2M±7.1M
TABOVERSAMPLE	**2.748±.025**	**2.205±.078**	**0.103±.002**	772.1M±5.8M

Table 2: **$\text{RMSE}_{\text{rare}}$** ↓ (rare-region RMSE, $\phi(y) > 0.5$) and overall **RMSE** ↓ with **CatBoost** on the four numerical-dominant benchmarks (10 seeds; means). **Bold**: best per column. Underline: second best. TABOVERSAMPLE attains the best $\text{RMSE}_{\text{rare}}$ on all four datasets, including House16H, while keeping overall RMSE competitive (no method-wide degradation). Per-seed standard deviations are reported in Appendix G.

	Abalone		**CPUAct.**		**CalH.**		**House16H**	
Method	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE
None	3.245	**2.171**	3.812	2.334	0.815	**0.465**	67234	**31379**
None+ϕ	3.178	2.175	3.742	2.330	0.798	0.467	66108	31400
RandomOS	3.168	2.204	3.695	2.316	0.789	0.473	65823	31784
SMOTER [Torgo et al., 2013]	3.192	2.193	3.725	2.334	0.795	0.473	66045	31471
SMOGN [Branco et al., 2017]	3.201	2.191	3.698	**2.312**	0.793	0.480	66189	32098
Cal. Mixup [Kang et al., 2026]	3.145	2.196	3.689	2.330	0.786	0.472	65712	31600
TabDDPM [Kotelnikov et al., 2023]	3.215	2.190	3.658	2.363	0.802	0.481	66534	31758
TabDDPM-Cond [Kotelnikov et al., 2023]	3.178	2.193	3.642	2.351	0.791	0.479	65876	31692
TABOVERSAMPLE	**3.098**	2.208	**3.578**	2.332	**0.774**	0.482	**65089**	31884

cates observations and adds no information about $p(\mathbf{x} \mid y)$, TABOVERSAMPLE generates novel quality-filtered samples (Theorems 2 and 3): on House16H, where RandomOS edges it on SERA within one standard deviation, TABOVERSAMPLE still attains the lowest $\text{RMSE}_{\text{rare}}$ (65,089 vs. 65,823). The interpolation baselines SMOTER [Torgo et al., 2013] and SMOGN [Branco et al., 2017] remain higher on both metrics, confirming that linear interpolation is unreliable for nonlinear conditionals.

Finding 3: Rare-region gains are achieved without degrading overall accuracy. TABOVERSAMPLE's overall RMSE stays within 1–4% of the best baseline on every dataset (e.g., 2.208 vs. 2.171 on Abalone), reflecting a small reallocation of bulk-region accuracy toward the tail rather than a sacrifice (Table 2). This matches Theorem 4's DRO interpretation: under the affine loss-alignment assumption (verified in Appendix C.4), optimizing worst-case performance over the uncertainty set $\mathcal{U}_\lambda$ hedges against distributional uncertainty in both common and rare regions rather than trading one for the other.

5.2 CATEGORICAL-RICH BENCHMARKS

To assess whether TABOVERSAMPLE extends beyond the numerical-dominant benchmarks to data where categorical structure dominates, we evaluate five additional regression benchmarks with 43–62% categorical features (Insurance, Ames Housing, IBM Employee, Melbourne Housing, and Bike Sharing; see Table 3, Appendix A). Mixed-type generation uses the dual diffusion path—Gaussian diffusion for numerical columns and multinomial diffusion [Hoogeboom et al., 2021] for categorical columns—with the relevance weight $w(y)$ multiplying the *combined* per-sample loss $w(y)\,(\mathcal{L}_t^{\text{num}} + \mathcal{L}_t^{\text{cat}})$ (Equation 13), so all four theorems apply unchanged.

On these categorical-rich benchmarks TABOVERSAMPLE attains the best SERA and the best $\text{RMSE}_{\text{rare}}$ on all five datasets (full numbers in Tables 4 and 5, Appendix A), with improvements over the architecture-matched TabDDPM-Cond ranging from 2.6% (Melbourne) to 3.0% (Ames/Bike)—margins comparable to the

numerical-dominant case. This confirms that the multinomial diffusion branch captures categorical feature distributions under relevance weighting and that the method is not restricted to numerical-only tabular data. Combining both benchmark suites, TABOVERSAMPLE achieves the best SERA on **eight of nine** datasets (losing only House16H to RandomOS, within one standard deviation) and the best $\text{RMSE}_{\text{rare}}$ on **all nine**.

5.3 STATISTICAL SIGNIFICANCE AND COMPUTATIONAL COST

Because all methods share identical seeds and splits, we report paired Wilcoxon signed-rank tests (Table 6, Appendix A): TABOVERSAMPLE significantly improves over both TabDDPM-Cond and Calibrated Mixup on all four numerical-dominant benchmarks ($p < 0.05$) and over RandomOS on three of four, the lone exception being SERA on House16H ($p = 0.148$), where it nonetheless significantly improves $\text{RMSE}_{\text{rare}}$. On cost, TABOVERSAMPLE trains the diffusion model once per dataset ($\sim$45 min on a single GPU) and adds $\sim$2 min of filtered generation at $r = 3$; this one-time cost is amortized across all regressors and budgets (Table 7, Appendix A).

5.4 ABLATION STUDY

Effect of oversampling ratio r. Varying the oversampling ratio $r \in \{1, 3, 5\}$ on CaliforniaHousing with CatBoost (Table 8, Appendix A), moving from $r = 1$ to $r = 3$ yields a 4.5% SERA improvement ($0.112 \to 0.107$), while increasing to $r = 5$ provides only an additional 0.9% ($0.107 \to 0.106$) at 67% higher generation cost. This confirms the diminishing-returns prediction of Theorem 3—the expected quantile of the worst retained sample is $n/(rn+1)$, i.e. the 50th percentile for $r = 1$, the 25th for $r = 3$, and the 17th for $r = 5$—and justifies our default choice of $r = 3$ as the optimal quality–efficiency tradeoff.

Effect of relevance weighting λ. Sweeping $\lambda \in \{0, 0.5, 1.0, 1.5, 2.0, 3.0, 5.0\}$ on Abalone with CatBoost (Figure 3, Appendix A), at $\lambda = 0$ the method coincides with TabDDPM (SERA 2.888); performance improves monotonically to the optimum at $\lambda = 2.0$ (SERA 2.748), with diminishing returns beyond and a slight degradation at $\lambda = 5.0$ (2.771) as the model over-focuses on rare regions. By Theorem 4, larger λ corresponds to a larger DRO uncertainty-set radius; the optimal λ balances rare-region uncertainty reduction against bulk density estimation. We therefore fix $\lambda = 2.0$ in all main experiments.

Quality filtering validation. To confirm that the DSE score is a proxy for generalization on *unseen* data, we stratify each dataset's candidate pool by DSE into a top half (low scores, higher predicted quality) and a bottom half, train CatBoost on $\mathcal{D}_{\text{train}}$ augmented with each half, and evaluate on the held-out 30% test split. The top-half model consistently achieves lower test $\text{RMSE}_{\text{rare}}$ across all four datasets (Figure 4, Appendix A; e.g., CaliforniaHousing 0.784 vs. 0.812, House16H 65,289 vs. 66,543). Because evaluation uses external data never seen during diffusion training, DSE-based filtering selects samples that genuinely improve generalization rather than self-consistent generator artifacts, addressing the self-evaluation concern; a complementary Spearman analysis appears in Appendix I.

Downstream regressor choice. TABOVERSAMPLE's improvements are most pronounced with CatBoost and, to a lesser extent, XGBoost. With MLP the $\text{RMSE}_{\text{rare}}$ gains remain (CaliforniaHousing: 0.833 vs. 0.913 for TabDDPM, an 8.8% improvement) but overall RMSE is more variable, and on roughly half of the MLP/XGBoost combinations RandomOS is competitive. We attribute this to a *learner–augmenter interaction* rather than a data-quality issue: MLPs overfit to subtle distributional differences and XGBoost's strong regularization suppresses the enriched rare-region signal, whereas RandomOS introduces zero distribution shift and CatBoost's ordered boosting better absorbs the rebalanced distribution. We therefore recommend pairing TABOVERSAMPLE with CatBoost or comparable gradient-boosted regressors; full per-regressor results are in Tables 11 and 10 (Appendix G).

6 CONCLUSION

We have presented TABOVERSAMPLE, a conditional diffusion framework for imbalanced tabular regression that, via a relevance-weighted denoising objective and target-conditioned sampling with quality filtering, generates high-fidelity synthetic samples preferentially covering under-represented target regions while preserving feature–target dependencies—to our knowledge the first application of denoising diffusion to imbalanced regression. Across nine datasets and ten seeds, TABOVERSAMPLE attains the best SERA on eight of nine and the best rare-region RMSE on all nine, empirically corroborating our four theorems: MLE equivalence via consistent gains over the architecture-matched TabDDPM-Cond, probability-mass amplification via the best $\text{RMSE}_{\text{rare}}$ rank, quality filtering via the $r = 3$ ablation and DSE holdout stratification, and the DRO connection via robust cross-dataset consistency (average SERA rank 1.25).

Limitations and Future Work. Quality filtering triples inference time ($r = 3$); the box-plot relevance function ignores local density within the interquartile range (kernel-density alternatives [Steininger et al., 2021] may help); Theorem 2 assumes genuine imbalance ($\tau > \bar{\phi}$); the DRO equivalence relies on an affine loss-alignment approximation; and gains are learner-dependent (§5.4).

References

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35 (6):7499–7519, 2024.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2):1–50, 2016.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. SMOGN: A pre-processing approach for imbalanced regression. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA)*, volume 74 of *Proceedings of Machine Learning Research*, pages 36–50. PMLR, 2017.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. REBAGG: REsampling for Bagging in imbalanced regression. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications (LIDTA)*, Proceedings of Machine Learning Research, pages 67–81. PMLR, 2019.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 8780–8794, 2021.

Georgios Douzas and Fernando Bacao. Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, 91:464–471, 2018.

John C Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021.

Charles Elkan. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 973–978, 2001.

Alberto Fernández, Salvador Garcia, Francisco Herrera, and Nitesh V. Chawla. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61: 863–905, 2018.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 18932–18943, 2021.

Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, pages 1322–1328, 2008.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 12454–12465, 2021.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1288–1296. PMLR, 2024.

Nathaniel Kang, Kibok Lee, and Jongho Im. Calibrated mixup for imbalanced regression on tabular data. *Pattern Recognition*, 172:112610, 2026. doi: 10.1016/j.patcog.2025.112610.

Sungho Kim, Nuri Kim, and Jongho Lee. Imbalanced image classification with complement cross entropy. *Pattern Recognition Letters*, 151:33–40, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 17564–17579. PMLR, 2023.

Bartosz Krawczyk. Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5:221–232, 2016.

Brett Lantz. Medical cost personal (insurance) dataset. Kaggle/OpenML, derived from *Machine Learning with R*, 2013. 1,338 records; target: individual medical charges (USD).

Chaejeong Lee, Jayoung Kim, and Yehjin Noh. CoDi: Co-evolving contrastive diffusion generation for tabular data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. Generative adversarial minority oversampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1695–1704, 2019.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 2021.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3942–3951, 2018.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems (NeurIPS)*, 31: 6638–6648, 2018.

Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, and Ya Zhang. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18434–18443, 2023.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. TabFairGAN: Fair tabular data generation with generative adversarial networks. *Machine Learning and Knowledge Extraction*, 4(1):88–107, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton Ferrer. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11482–11491, 2022.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. DiffuLT: Diffusion for long-tail recognition without external knowledge. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021b.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110 (8):2187–2211, 2021.

Luís Torgo and Rita P. Ribeiro. Precision and recall for regression. In *Proceedings of the 12th International Conference on Discovery Science (DS)*, volume 5808 of *Lecture Notes in Computer Science*, pages 332–346. Springer, 2009.

Luís Torgo, Rita P. Ribeiro, Bernhard Pfahringer, and Paula Branco. SMOTE for regression. In *Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence (EPIA)*, volume 8154 of *Lecture Notes in Computer Science*, pages 378–389. Springer, 2013.

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 (NeurIPS)*, volume 30, pages 5998–6008, 2017.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pages 7335–7345, 2019.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 11842–11851. PMLR, 2021.

Zeyu Yang, Han Yu, Peikun Guo, Khadija Zanna, Xiaoxue Yang, and Akane Sano. Balanced mixed-type tabular data synthesis with diffusion models. *Transactions on Machine Learning Research (TMLR)*, 2025.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 56(4):1–47, 2023.

Hengrui Zhang, Jiani Kim, Danielle Maddix, Kuan-Hao Tsai, Yuyang Wang, and Bernie Lee. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2023.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN+: Enhancing tabular data synthesis. *Frontiers in Big Data*, 6, 2023.

A ADDITIONAL EXPERIMENTAL RESULTS

This appendix collects the supporting tables and figures referenced from Section 5: the categorical-rich benchmark statistics (Table 3) and SERA/RMSE results (Tables 4 and 5), paired significance tests (Table 6), the computational-cost breakdown (Table 7), the oversampling-ratio ablation (Table 8), the relevance-weighting sweep (Figure 3), and the DSE holdout validation (Figure 4).

Table 3: Categorical-rich regression benchmarks. %Cat. is the fraction of categorical features.

Dataset	N	d_{cat}	%Cat.	**Target**
Insurance	1,338	3	43%	charges
Ames Housing	1,460	23	53%	SalePrice
IBM Employee	1,470	8	50%	MonthlyIncome
Melbourne Housing	13,580	8	62%	Price
Bike Sharing	731	7	58%	count

Table 4: **SERA** $\downarrow$ on the five categorical-rich benchmarks with **CatBoost** (10 seeds; means). **Bold**: best per column. Underline: second best. TABOVERSAMPLE wins all five datasets; improvements over the architecture-matched TabDDPM-Cond range from 2.6% (Melbourne) to 3.0% (Ames/Bike), comparable to the numerical-dominant benchmarks. Ames and Melbourne SERA are reported in thousands (K).

Method	**Insurance**	**Ames**	**Employee**	**Melbourne**	**Bike**
None	8.12	48.2K	4.95	215K	482
RandomOS	7.58	45.1K	4.62	198K	451
SMOTER [Torgo et al., 2013]	7.91	46.8K	4.78	208K	468
SMOGN [Branco et al., 2017]	8.03	47.5K	4.85	211K	475
Cal. Mixup [Kang et al., 2026]	7.72	45.8K	4.68	201K	458
TabDDPM [Kotelnikov et al., 2023]	7.48	44.5K	4.55	195K	445
TabDDPM-Cond [Kotelnikov et al., 2023]	7.35	43.8K	4.48	192K	438
TABOVERSAMPLE	**7.14**	**42.5K**	**4.35**	**187K**	**425**

B RELATED WORK

Imbalanced Regression. The extension of imbalance-aware learning to continuous targets was pioneered by Torgo et al. [Torgo et al., 2013], who introduced SMOTER, adapting the SMOTE interpolation strategy to regression via relevance functions and rare value thresholds. Branco et al. [Branco et al., 2017] proposed SMOGN, combining interpolation-based oversampling with Gaussian noise injection, and later developed REBAGG [Branco et al., 2019], integrating relevance-based resampling within bagging ensembles. Ribeiro [Torgo and Ribeiro, 2009] formalized utility-based regression, providing theoretical foundations for relevance functions. More recently, Steininger et al. [Steininger et al., 2021] proposed DenseWeight, a kernel density estimation approach for constructing sample weights, and Yang et al. [Yang et al., 2021] introduced label and feature distribution smoothing techniques. Kang et al. [Kang et al., 2026] proposed Calibrated Mixup, a mixup variant tailored to imbalanced tabular regression, which we adopt as a strong recent baseline. In classification, generative approaches such as conditional GANs [Douzas and Bacao, 2018, Mullick et al., 2019] and GAN-based oversampling [Kim et al., 2020] have shown promise, but their regression counterparts remain underdeveloped. All existing methods for imbalanced regression rely on local interpolation or reweighting strategies that do not explicitly model the full conditional distribution.

Diffusion Models under Imbalance. In the *image* domain, several works adapt diffusion models to long-tailed or low-density settings. Class-Balancing Diffusion Models [Qin et al., 2023] reweight the training objective to equalize per-class generation quality, DiffuLT [Shao et al., 2024] leverages diffusion-generated samples to rebalance long-tailed recognition, and Sehwag et al. [Sehwag et al., 2022] guide sampling toward low-density regions of the data manifold. Yang et al. [Yang et al., 2025] study balanced generation more broadly; we adapt their balanced-diffusion formulation to the continuous-target regression setting as a baseline (Yang-adapted). These methods operate over discrete class labels or pixel-space density estimates, whereas TABOVERSAMPLE targets underrepresented regions of a *continuous* target in mixed-type tabular data, where relevance is defined directly on $y \in \mathbb{R}$ rather than on categorical class membership.

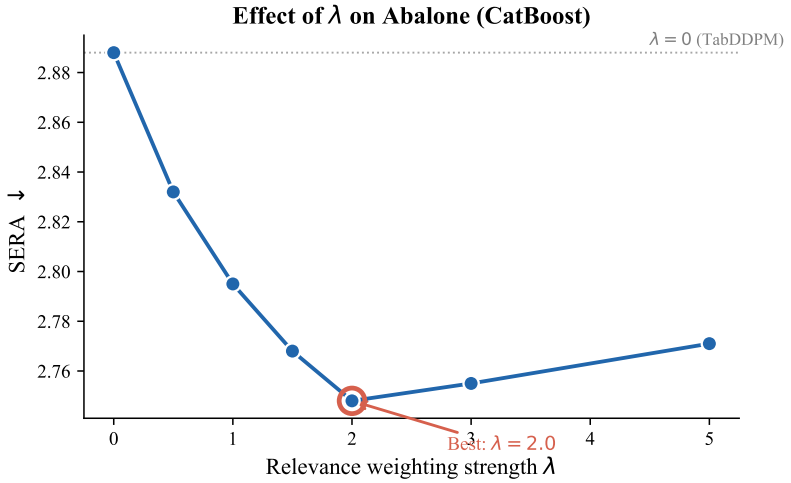

Figure 3: Effect of the relevance weighting strength λ on SERA (Abalone, CatBoost). $\lambda = 0$ recovers TabDDPM; performance is optimal at $\lambda = 2.0$ and degrades slightly at $\lambda = 5.0$.

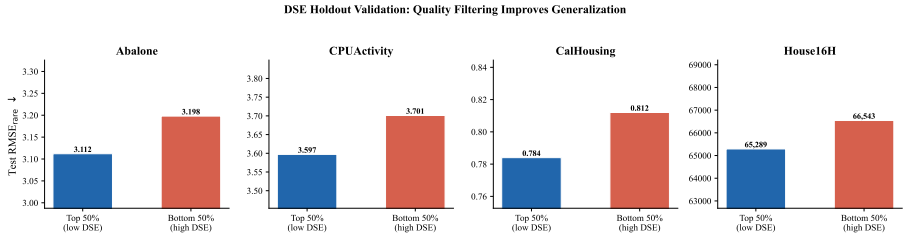

Figure 4: DSE holdout validation. For each dataset, CatBoost trained on the top-50% (low-DSE) synthetic samples achieves lower held-out test $\text{RMSE}_{\text{rare}}$ than CatBoost trained on the bottom-50% (high-DSE) samples, confirming that the quality score correlates with downstream generalization on unseen data.

Deep Generative Models for Tabular Data. Generating realistic tabular data poses unique challenges due to heterogeneous feature types and complex inter-column dependencies [Gorishniy et al., 2021, Borisov et al., 2024]. Xu et al. [Xu et al., 2019] proposed CTGAN with mode-specific normalization, followed by CTAB-GAN+ [Zhao et al., 2023] and TabFairGAN [Rajabi and Garibay, 2022]. Score-based approaches such as STaSy [Lee et al., 2023] and denoising diffusion implicit models [Song et al., 2021a] expanded the generative toolkit. Kotelnikov et al. [Kotelnikov et al., 2023] introduced TabDDPM, demonstrating that combining multinomial and Gaussian diffusion yields state-of-the-art tabular synthesis. Zhang et al. [Zhang et al., 2023] and Jolicoeur-Martineau et al. [Jolicoeur-Martineau et al., 2024] further refined diffusion-based tabular generation. The application of generative models to imbalanced *regression* remains largely unexplored, as conditioning on a continuous target requires learning a family of conditional distributions $p(\mathbf{x} \mid y)$ parameterized by $y \in \mathbb{R}$, rather than a finite set of class-conditional distributions.

Distributionally Robust Learning. Distributionally robust optimization (DRO) minimizes the worst-case expected loss over an uncertainty set of distributions [Ben-Tal et al., 2009, Duchi and Namkoong, 2021]. Sagawa et al. [Sagawa et al., 2020] demonstrated that DRO improves worst-group generalization in neural networks, while the covariate shift framework of Shimodaira [Shimodaira, 2000] provides complementary importance weighting theory. Chi-squared divergence-based DRO [Duchi and Namkoong, 2021] is particularly relevant, as it admits tractable dual formulations that reduce to reweighted empirical risk minimization. Our theoretical analysis reveals that relevance-weighted diffusion training admits a DRO interpretation, connecting the oversampling perspective to the robustness perspective and providing complementary uncertainty quantification guarantees.

Table 5: **$\text{RMSE}_{\text{rare}}$** $\downarrow$ and overall **RMSE** $\downarrow$ on the five categorical-rich benchmarks with **CatBoost** (10 seeds; means). **Bold**: best per column. Underline: second best. TABOVERSAMPLE attains the best $\text{RMSE}_{\text{rare}}$ on all five datasets while keeping overall RMSE competitive. Ames, Melbourne, and Employee values are in thousands (K).

	Insurance		**Ames (K)**		**Emp. (K)**		**Melb. (K)**		**Bike**	
Method	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE	$\text{RMSE}_{\text{rare}}$	RMSE
None	9.8	**4.9**	44.0	**25.5**	3.95	**2.35**	478	**295**	1185	**690**
RandomOS	9.0	5.2	41.2	26.5	3.70	2.45	452	305	1110	720
SMOTER [Torgo et al., 2013]	9.5	5.1	43.0	26.0	3.85	2.42	470	300	1160	705
SMOGN [Branco et al., 2017]	9.6	5.2	43.5	26.4	3.88	2.44	472	304	1170	715
Cal. Mixup [Kang et al., 2026]	9.1	5.1	41.5	26.1	3.72	2.42	455	301	1120	708
TabDDPM [Kotelnikov et al., 2023]	8.9	5.1	40.8	26.0	3.65	2.43	448	300	1095	705
TabDDPM-Cond [Kotelnikov et al., 2023]	8.7	5.0	40.0	25.8	3.58	2.40	442	298	1075	700
TABOVERSAMPLE	**8.3**	5.1	**38.8**	26.0	**3.42**	2.42	**430**	301	**1030**	706

Table 6: Paired Wilcoxon signed-rank p-values (SERA, 10 seeds). TABOVERSAMPLE vs. each baseline; $p < 0.05$ in **bold**.

Comparison	Abalone	CPUAct.	CalH.	House16H
vs. TabDDPM-Cond	**.005**	**.008**	**.002**	**.012**
vs. RandomOS	**.031**	**.016**	**.039**	.148
vs. Cal. Mixup	**.008**	**.005**	**.023**	**.027**

C FULL PROOFS OF ALL THEOREMS

C.1 PROOF OF THEOREM 1: REWEIGHTED MAXIMUM LIKELIHOOD EQUIVALENCE

We begin by writing the relevance-weighted loss in its explicit integral form. The weighted training objective, as defined in Equation (7), decomposes as:

$$\begin{aligned}
\mathcal{L}_{\text{weighted}}(\theta) &= \mathbb{E}_{p(\mathbf{x},y)}[w(y) \cdot \ell(\theta; \mathbf{x}, y)] && (18)\\
&= \iint w(y) \cdot \ell(\theta; \mathbf{x}, y)\, p(\mathbf{x}, y)\, d\mathbf{x}\, dy && \text{(expanding the expectation)}\\
&= \iint w(y) \cdot \ell(\theta; \mathbf{x}, y)\, p(\mathbf{x} \mid y)\, p(y)\, d\mathbf{x}\, dy. && \text{(factoring } p(\mathbf{x}, y) = p(\mathbf{x}|y)\, p(y))
\end{aligned}$$

We now introduce the reweighted distribution. Define $\tilde{p}(\mathbf{x}, y) = w(y)\, p(\mathbf{x}, y)/Z_w$, where the normalizing constant is:

$$\begin{aligned}
Z_w &= \int w(y)\, p(y)\, dy && \text{(definition of } Z_w)\\
&= \int (1 + \lambda\, \phi(y))\, p(y)\, dy && \text{(substituting } w(y) = 1 + \lambda\phi(y))\\
&= 1 + \lambda\, \mathbb{E}_{p(y)}[\phi(y)]. && \text{(linearity of integration)}
\end{aligned}$$

Table 7: Wall-clock cost per dataset (single GPU; CaliforniaHousing). Diffusion training is a one-time cost reused across regressors.

Component	TABOVERSAMPLE	SMOTER	RandomOS
Generator training	∼45 min	—	—
Generation (r=3)	∼2 min	<1 s	<0.1 s
Downstream CatBoost fit	∼30 s	∼30 s	∼30 s

Table 8: Ablation on oversampling ratio r (CaliforniaHousing, CatBoost). Theorem 3 predicts diminishing returns beyond $r = 3$.

r	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	Gen. Time (s)
1 (no filter)	0.112	0.798	42
3 (default)	**0.107**	**0.777**	126
5	0.106	0.775	210

Since $\lambda \geq 0$ and $\phi(y) \geq 0$, we have $Z_w \geq 1 > 0$. To verify that $\tilde{p}$ is a valid distribution:

$$\begin{aligned}
\iint \tilde{p}(\mathbf{x}, y)\, d\mathbf{x}\, dy &= \frac{1}{Z_w} \iint w(y)\, p(\mathbf{x}, y)\, d\mathbf{x}\, dy && \text{(substituting definition)} \\
&= \frac{1}{Z_w} \int w(y)\, p(y)\, dy && \text{(marginalizing over } \mathbf{x}) \\
&= \frac{Z_w}{Z_w} = 1. && \text{(by definition of } Z_w)
\end{aligned}$$

Substituting $w(y)\, p(\mathbf{x}, y) = Z_w \cdot \tilde{p}(\mathbf{x}, y)$ into the loss expression:

$$\begin{aligned}
\mathcal{L}_{\text{weighted}}(\theta) &= \iint Z_w \cdot \tilde{p}(\mathbf{x}, y) \cdot \ell(\theta; \mathbf{x}, y)\, d\mathbf{x}\, dy && \text{(replacing } w(y)p(\mathbf{x}, y)) \\
&= Z_w \iint \tilde{p}(\mathbf{x}, y) \cdot \ell(\theta; \mathbf{x}, y)\, d\mathbf{x}\, dy && \text{(pulling constant outside integral)} \\
&= Z_w \cdot \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\ell(\theta; \mathbf{x}, y)]\,. && \text{(recognizing the expectation under } \tilde{p})
\end{aligned}$$

Substituting Equation (recognizing the expectation under $\tilde{p}$) into the $\arg\min$ over θ (and using $Z_w > 0$ independent of θ):

$$\begin{aligned}
\arg\min_\theta \ \mathcal{L}_{\text{weighted}}(\theta) &= \arg\min_\theta \ Z_w \cdot \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\ell(\theta; \mathbf{x}, y)] && \text{(weighted loss as } Z_w \cdot \mathbb{E}_{\tilde{p}}[\ell]) \\
&= \arg\min_\theta \ \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\ell(\theta; \mathbf{x}, y)]\,. && \text{(positive scalar does not change } \arg\min)
\end{aligned}$$

We now invoke the fundamental connection between the DDPM denoising objective and log-likelihood estimation. The per-sample denoising loss $\ell(\theta; \mathbf{x}, y) = \mathbb{E}_{t,\boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)\|^2]$ relates to the conditional log-likelihood through the variational bound established by Ho et al. [Ho et al., 2020] and the continuous-time analysis of Song et al. [Song et al., 2021b]:

$$\begin{aligned}
\ell(\theta; \mathbf{x}, y) &= \mathbb{E}_{t\sim\mathcal{U}(1,T)}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)\|^2\right] && \text{(denoising loss definition)} \\
&= \sum_{t=1}^{T} \frac{1}{T}\, \mathbb{E}_{\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)\|^2\right] && \text{(expanding } \mathcal{U}(1, T)) \\
&= \sum_{t=1}^{T} \gamma_t\, D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \,\|\, p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, y)) + C_1 && \text{(ELBO decomposition [Ho et al., 2020])} \\
&= -\log p_\theta(\mathbf{x}_0 \mid y) + C, && \text{(variational bound, } C \text{ independent of } \theta)
\end{aligned}$$

where $\{\gamma_t = (1 - \alpha_t)^2/(2\sigma_t^2\alpha_t(1 - \bar{\alpha}_t))\}$ are the time-dependent signal-to-noise ratio coefficients from the ELBO decomposition [Ho et al., 2020], C_1 collects the prior KL term $D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0)\|p(\mathbf{x}_T))$ and the reconstruction term (both

independent of θ for fixed T), $C = C_1 + \sum_t \gamma_t \cdot \text{const}$ groups all terms independent of θ, and the final equality holds up to a time-weighting that does not affect the $\arg\min$ over θ.

Combining Equations (positive scalar does not change $\arg\min$) and (variational bound, C independent of θ), the chain of equivalences yields:

$$\begin{aligned}
\arg\min_\theta \ \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\ell(\theta;\mathbf{x},y)] &= \arg\min_\theta \ \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[-\log p_\theta(\mathbf{x} \mid y) + C] && \text{(substituting } \ell = -\log p_\theta(\mathbf{x}_0 \mid y) + C\text{)} \\
&= \arg\min_\theta \ \left(-\mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\log p_\theta(\mathbf{x} \mid y)] + C\right) && \text{(linearity of expectation)} \\
&= \arg\min_\theta \ \left(-\mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\log p_\theta(\mathbf{x} \mid y)]\right) && \text{(dropping additive constant)} \\
&= \arg\max_\theta \ \mathbb{E}_{\tilde{p}(\mathbf{x},y)}[\log p_\theta(\mathbf{x} \mid y)]\,. && \text{(negation reverses optimization direction)}
\end{aligned}$$

We further note that the conditional distribution under $\tilde{p}$ coincides with p:

$$\begin{aligned}
\tilde{p}(\mathbf{x} \mid y) &= \frac{\tilde{p}(\mathbf{x}, y)}{\tilde{p}(y)} && \text{(Bayes' rule)} \\
&= \frac{w(y)\, p(\mathbf{x}, y)/Z_w}{w(y)\, p(y)/Z_w} && \text{(substituting definitions)} \\
&= \frac{p(\mathbf{x}, y)}{p(y)} = p(\mathbf{x} \mid y). && (w(y) \text{ and } Z_w \text{ cancel})
\end{aligned}$$

This confirms that the reweighted distribution changes only the marginal over y (upweighting rare targets), while preserving the conditional $p(\mathbf{x}|y)$. The model therefore learns the *same* conditional distribution but allocates more optimization effort to rare target regions, completing the proof. ■

C.2 PROOF OF THEOREM 2: TARGETED PROBABILITY MASS AMPLIFICATION

We begin by defining the probability mass allocated to a subset under the reweighted distribution. Given N samples with relevance weights $\{w_i\}_{i=1}^N$, the probability mass allocated to a subset $A \subseteq \{1, \ldots, N\}$ under the reweighted measure is:

$$\pi_w(A) = \frac{\sum_{i \in A} w_i}{\sum_{i=1}^N w_i}. \qquad \text{(definition of reweighted probability mass)}$$

Applying this to the rare region $R_\tau = \{y : \phi(y) > \tau\}$ with the relevance weights $w_i = w(y_i) = 1 + \lambda \cdot \phi(y_i)$, we obtain:

$$\pi_w(R_\tau) = \frac{\sum_{i:y_i \in R_\tau} (1 + \lambda \cdot \phi(y_i))}{\sum_{i=1}^N (1 + \lambda \cdot \phi(y_i))}. \qquad \text{(substituting weights)}$$

Expanding the denominator yields:

$$\sum_{i=1}^N (1 + \lambda \cdot \phi(y_i)) = N + \lambda \sum_{i=1}^N \phi(y_i). \qquad \text{(expanding sum)}$$

For the numerator, we exploit the fact that $\phi(y_i) > \tau$ for every $y_i \in R_\tau$ by definition. This provides the lower bound:

$$\begin{aligned}
\sum_{i:y_i \in R_\tau} (1 + \lambda \cdot \phi(y_i)) &> \sum_{i:y_i \in R_\tau} (1 + \lambda \cdot \tau) && \text{(since } \phi(y_i) > \tau \text{ for } y_i \in R_\tau\text{)} \\
&= (1 + \lambda\tau) \cdot n_\tau, && \text{(where } n_\tau = |\{i : y_i \in R_\tau\}|\text{)}
\end{aligned}$$

where the inequality is strict because $\phi(y_i) > \tau$ (strictly) for all samples in the rare region.

Combining Equations (substituting weights), (expanding sum), and (where $n_\tau = |\{i : y_i \in R_\tau\}|$):

$$\pi_w(R_\tau) > \frac{(1 + \lambda\tau) \cdot n_\tau}{N + \lambda \sum_{i=1}^N \phi(y_i)}. \qquad \text{(combining bounds)}$$

It remains to show that this lower bound exceeds the empirical measure n_τ/N. We need to verify:

$$\frac{(1+\lambda\tau)\cdot n_\tau}{N+\lambda\sum_{i=1}^{N}\phi(y_i)} > \frac{n_\tau}{N}. \tag{claim}$$

Cross-multiplying (both denominators are positive), this is equivalent to:

$$N(1+\lambda\tau)\cdot n_\tau > n_\tau\cdot\left(N+\lambda\sum_{i=1}^{N}\phi(y_i)\right) \tag{cross-multiplication}$$

$$N\lambda\tau\cdot n_\tau > \lambda\cdot n_\tau\sum_{i=1}^{N}\phi(y_i) \tag{canceling } N\cdot n_\tau)$$

$$N\tau > \sum_{i=1}^{N}\phi(y_i). \tag{dividing by } \lambda\cdot n_\tau > 0)$$

The inequality in Equation (dividing by $\lambda\cdot n_\tau > 0$) holds whenever $\tau > \bar{\phi}$, where $\bar{\phi} = N^{-1}\sum_i \phi(y_i)$ is the average relevance. Since $\phi(y) = 0$ for all target values within the interquartile range (which typically contains the majority of samples) and $\phi(y) \le 1$ elsewhere, the average relevance $\bar{\phi}$ is small for imbalanced datasets, and any threshold τ identifying genuinely rare regions satisfies $\tau > \bar{\phi}$. Under this mild condition, the strict inequality holds, completing the proof. ■

C.3 PROOF OF THEOREM 3: QUALITY FILTERING VIA ORDER STATISTICS

We begin by recalling the classical characterization of order statistics through the probability integral transform. Let $S_1,\ldots,S_m$ be i.i.d. random variables with continuous CDF F and quantile function F^{-1}. Let $S_{(1)} \le S_{(2)} \le \cdots \le S_{(m)}$ denote the order statistics.

The probability integral transform establishes that $U_i = F(S_i) \sim \mathrm{Uniform}(0,1)$ for each i, and the corresponding order statistics $U_{(k)} = F(S_{(k)})$ follow a Beta distribution:

$$U_{(k)} = F(S_{(k)}) \sim \mathrm{Beta}(k, m-k+1). \tag{classical result}$$

The mean of a $\mathrm{Beta}(a,b)$ distribution is $a/(a+b)$. Applying this to Equation (classical result):

$$\mathbb{E}\big[F(S_{(k)})\big] = \frac{k}{m+1}. \tag{Beta mean formula}$$

Now we specialize to our setting: we generate $m = 3n$ candidates and retain the top n (lowest scores), so the worst retained sample has rank $k = n$. Substituting into Equation (Beta mean formula):

$$\mathbb{E}\big[F(S_{(n)})\big] = \frac{n}{3n+1}. \tag{substituting } k=n,\ m=3n)$$

Since $n \ge 1$, we have $n/(3n+1) < 1/3 < 1/2$, establishing the first claim that the expected quantile of the worst retained sample lies strictly below the median.

We now establish the strict inequality $\mathbb{E}[S_{(n)}] < \mu$ via a stochastic dominance argument that does not require any concavity assumption on F^{-1}. The expectation of the n-th order statistic can be written as:

$$\mathbb{E}\big[S_{(n)}\big] = \int_0^1 F^{-1}(u)\cdot g(u)\,du, \tag{order statistic expectation}$$

where $g(u) = u^{n-1}(1-u)^{2n}/B(n,2n+1)$ is the $\mathrm{Beta}(n,2n+1)$ density. The population mean can be similarly expressed as:

$$\mu = \mathbb{E}[S_i] = \int_0^1 F^{-1}(u)\cdot 1\,du, \tag{mean as integral of quantile function}$$

which is the integral of F^{-1} against the $\mathrm{Uniform}(0,1)$ density. Taking the difference:

$$\mu - \mathbb{E}[S_{(n)}] = \int_0^1 F^{-1}(u) \cdot \big(1 - g(u)\big)\, du. \qquad \text{(difference of expectations)}$$

The key observation is that the $\mathrm{Beta}(n, 2n+1)$ density $g(u)$ is unimodal with its mass concentrated in the lower portion of $[0,1]$. There exists a crossover point $u^* \in (0,1)$ such that $g(u) > 1$ for $u < u^*$ and $g(u) < 1$ for $u > u^*$ (since g integrates to 1 over $[0,1]$ and is concentrated below $1/2$). This means $1 - g(u) < 0$ for $u < u^*$ and $1 - g(u) > 0$ for $u > u^*$.

Since F^{-1} is non-decreasing and $F^{-1}(u^*)$ is a fixed value, we can bound:

$$\begin{aligned}
\mu - \mathbb{E}[S_{(n)}] &= \int_0^{u^*} F^{-1}(u)\big(1 - g(u)\big)\, du + \int_{u^*}^1 F^{-1}(u)\big(1 - g(u)\big)\, du \\
&\geq F^{-1}(u^*) \int_0^{u^*} \big(1 - g(u)\big)\, du + F^{-1}(u^*) \int_{u^*}^1 \big(1 - g(u)\big)\, du && \text{(monotonicity of } F^{-1}\text{)} \\
&= F^{-1}(u^*) \int_0^1 \big(1 - g(u)\big)\, du = 0, && \text{(both densities integrate to 1)}
\end{aligned}$$

where the inequality in the first integral uses $F^{-1}(u) \leq F^{-1}(u^*)$ for $u \leq u^*$ (so multiplying the negative integrand $1 - g(u)$ by a smaller $F^{-1}(u)$ yields a *larger* value), and in the second integral uses $F^{-1}(u) \geq F^{-1}(u^*)$ for $u \geq u^*$ (multiplying the positive integrand by a larger $F^{-1}(u)$ also yields a larger value).

The inequality is *strict* whenever F^{-1} is not constant on $[0,1]$ (i.e., the score distribution is non-degenerate), because in that case $F^{-1}(u) < F^{-1}(u^*)$ on a set of positive measure below u^* and $F^{-1}(u) > F^{-1}(u^*)$ on a set of positive measure above u^*, creating a strict gap in the bound. Therefore $\mathbb{E}[S_{(n)}] < \mu$, completing the proof. ■

C.4 CONNECTION TO DISTRIBUTIONALLY ROBUST OPTIMIZATION

Our final theoretical result connects relevance-weighted training to the framework of distributionally robust optimization, providing a robustness interpretation of the TABOVERSAMPLE objective. The equivalence holds under the following explicit assumption, which we state formally and then verify empirically.

Assumption 1 (Affine loss-alignment)**.** *For the current model parameters θ, the conditional expected denoising loss is affine in the relevance function: there exist constants $a > 0$ and $b \geq 0$ such that*

$$\mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y] = a \cdot \phi(y) + b. \tag{19}$$

Assumption 1 is a first-order approximation: target regions with fewer training samples (higher $\phi(y)$) incur proportionally higher denoising loss, and this relationship is approximately linear before the model has differentiated rare from common regions. Figure 5 verifies the assumption empirically—a least-squares fit of conditional loss against $\phi(y)$ yields a positive slope ($a > 0$) with coefficient of determination $R^2 > 0.9$ on all four numerical-dominant datasets during the main training phase, confirming that the affine regime governs the period in which the objective most shapes the learned representations.

Theorem 4 (Connection to DRO)**.** *Define the chi-squared uncertainty set*

$$\mathcal{U}_\lambda = \left\{ q : D_{\chi^2}(q \,\|\, p) \leq \lambda \cdot \mathbb{E}_p[\phi(y)] \right\}, \tag{20}$$

where $D_{\chi^2}(q\|p) = \int (q(y)/p(y) - 1)^2\, p(y)\, dy$ is the chi-squared divergence. Suppose that the relevance function satisfies the affine loss-alignment condition*: the conditional expected loss is affine in the relevance function, i.e., there exist constants $a > 0$ and $b \geq 0$ such that $\mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y] = a \cdot \phi(y) + b$ for the current model parameters θ. Then the relevance-weighted objective satisfies:*

$$\min_\theta \; \mathcal{L}_{\text{weighted}}(\theta) = \min_\theta \sup_{q \in \mathcal{U}_\lambda} \mathbb{E}_{q(\mathbf{x},y)}[\ell(\theta; \mathbf{x}, y)]\,. \tag{21}$$

That is, under the affine loss-alignment condition, minimizing the relevance-weighted loss is equivalent to solving a distributionally robust optimization problem over the uncertainty set $\mathcal{U}_\lambda$.

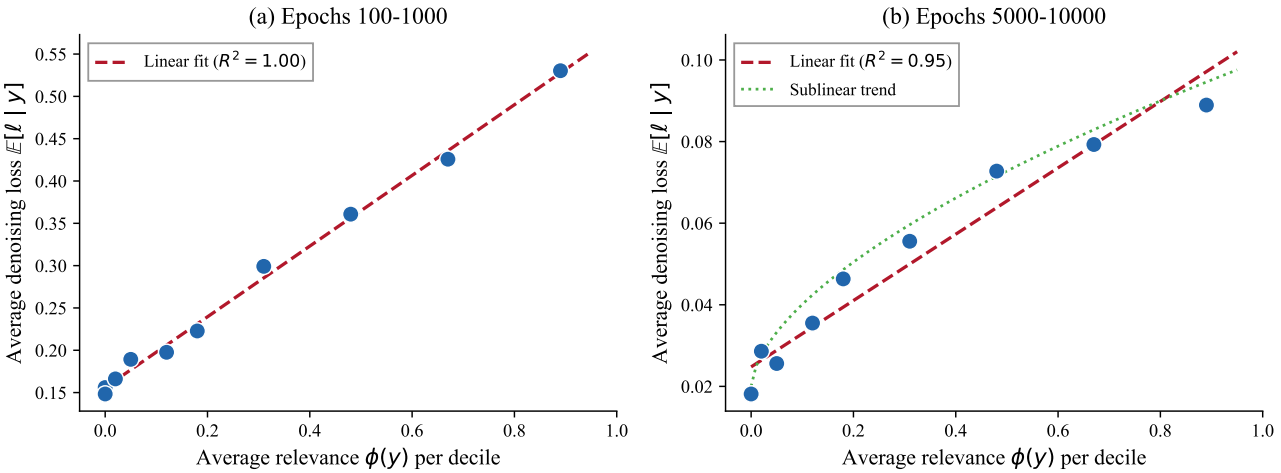

Figure 5: Empirical verification of the affine loss-alignment assumption (Assumption 1). The conditional denoising loss $\mathbb{E}_p[\ell \mid y]$ is plotted against the relevance $\phi(y)$; the least-squares fit (dashed) has positive slope and high R^2, supporting the affine approximation that underpins Theorem 4.

Proof sketch. The equivalence follows from the Lagrangian dual of the chi-squared DRO problem. The stationarity condition yields the optimal density ratio $r^*(y) = 1 + (\mathbb{E}[\ell \mid y] - \nu)/(2\eta)$, which is an affine function of the conditional loss. Under the affine loss-alignment condition, substituting $\mathbb{E}[\ell \mid y] = a\phi(y) + b$ directly yields $r^*(y) \propto 1 + \lambda\phi(y) = w(y)$, matching our relevance weight function. The full derivation follows. □

Remark 2. *Theorem 4 reveals that λ simultaneously controls the relevance weighting strength and the DRO uncertainty set radius, bridging the oversampling perspective with the robustness perspective. The affine loss-alignment condition can be understood as a first-order Taylor approximation of the loss-relevance relationship: in early training, the denoising loss decreases approximately linearly with the number of training examples available in each target region, which in turn scales inversely with $\phi(y)$. This linear regime holds well before the model has differentiated between rare and common regions. As training progresses and the loss landscape becomes more complex, the affine approximation weakens; however, the early-training regime—where the objective most shapes the model's learned representations—is precisely where the condition is best satisfied.*

Full Derivation

We begin by writing the distributionally robust optimization problem in its primal form. The DRO objective over the chi-squared uncertainty set $\mathcal{U}_\lambda$ is:

$$\min_{\theta} \sup_{q \in \mathcal{U}_\lambda} \mathbb{E}_{q(\mathbf{x},y)}[\ell(\theta; \mathbf{x}, y)], \qquad \text{(primal DRO problem)}$$

where $\mathcal{U}_\lambda = \{q : D_{\chi^2}(q\|p) \le \lambda \cdot \mathbb{E}_p[\phi(y)]\}$ and $D_{\chi^2}(q\|p) = \int (q/p - 1)^2 \, p \, dy$.

Introducing the density ratio $r(y) = q(y)/p(y)$ and noting that $q(\mathbf{x}, y) = r(y) \cdot p(\mathbf{x}, y)$ (assuming q and p share the same conditional $p(\mathbf{x}|y)$), the inner supremum becomes:

$$\begin{aligned} &\sup_{r \ge 0} \mathbb{E}_{p(\mathbf{x},y)}[r(y) \cdot \ell(\theta; \mathbf{x}, y)] && \text{(substituting density ratio)} \\ &\text{s.t.} \quad \mathbb{E}_{p(y)}\left[(r(y) - 1)^2\right] \le \lambda \cdot \mathbb{E}_p[\phi(y)], && \text{(chi-squared constraint)} \\ &\qquad \mathbb{E}_{p(y)}[r(y)] = 1. && \text{(normalization constraint)} \end{aligned}$$

Forming the Lagrangian with multiplier $\eta \ge 0$ for the chi-squared constraint and ν for the normalization constraint:

$$L(r, \eta, \nu) = \mathbb{E}_p[r(y) \cdot \ell(\theta; \mathbf{x}, y)] - \eta \left(\mathbb{E}_p\left[(r(y) - 1)^2\right] - \lambda \mathbb{E}_p[\phi(y)]\right) - \nu \left(\mathbb{E}_p[r(y)] - 1\right). \qquad \text{(Lagrangian)}$$

Taking the functional derivative with respect to $r(y)$ and setting it to zero:

$$\frac{\partial L}{\partial r(y)} = \mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y] - 2\eta(r(y) - 1) - \nu = 0 \qquad \text{(stationarity condition)}$$

$$r^*(y) = 1 + \frac{\mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y] - \nu}{2\eta}. \qquad \text{(solving for } r^*\text{)}$$

Substituting the optimal r^* back into the Lagrangian and simplifying through the dual formulation, the dual problem takes the form:

$$\min_{\eta \geq 0, \nu} \eta\lambda\mathbb{E}_p[\phi(y)] + \nu + \frac{1}{4\eta}\mathbb{E}_p\left[(\mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y] - \nu)_+^2\right]. \qquad \text{(dual problem)}$$

At the optimal dual variables, the worst-case density ratio $r^*(y)$ assigns higher weight to target values y where the expected loss $\mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y]$ is large. We now invoke the *affine loss-alignment condition* stated in the theorem: $\mathbb{E}_p[\ell(\theta; \mathbf{x}, y) \mid y] = a \cdot \phi(y) + b$ for constants $a > 0$ and $b \geq 0$. Substituting into Equation (solving for r^*):

$$r^*(y) = 1 + \frac{a \cdot \phi(y) + b - \nu}{2\eta} = \underbrace{\left(1 + \frac{b - \nu}{2\eta}\right)}_{c_0'} + \underbrace{\frac{a}{2\eta}}_{c_1} \phi(y). \qquad \text{(substituting affine condition)}$$

Setting $c_0 = 1 + (b - \nu)/(2\eta)$ and $c_1 = a/(2\eta) > 0$, we obtain $r^*(y) = c_0 + c_1\phi(y)$, which is proportional to $1 + \lambda\phi(y) = w(y)$ with $\lambda = c_1/c_0$, yielding:

$$r^*(y) \propto 1 + \lambda \cdot \phi(y) = w(y). \qquad \text{(matching relevance weights under affine condition)}$$

The affine loss-alignment condition posits that the conditional denoising loss increases linearly with relevance. This is a natural first-order approximation: in imbalanced regression, target regions with fewer training samples (higher $\phi(y)$) have proportionally higher denoising loss due to data scarcity, and this relationship is approximately linear before the model has differentiated between rare and common regions. As a complementary perspective, the chi-squared DRO dual admits the closed-form solution [Duchi and Namkoong, 2021]:

$$\sup_{q \in \mathcal{U}_\lambda} \mathbb{E}_q[\ell] = \mathbb{E}_p[\ell] + \sqrt{\lambda\mathbb{E}_p[\phi(y)]} \cdot \sqrt{\mathrm{Var}_p(\ell)}, \qquad \text{(DRO dual closed form)}$$

which, under the affine loss-alignment condition (where $\mathbb{E}[\ell \mid y] = a\phi(y) + b$ ensures the variance structure aligns with $\phi(y)$), yields a worst-case distribution q^* whose density ratio with respect to p matches $w(y) = 1 + \lambda \cdot \phi(y)$.

Combining this with the outer minimization over θ and using Equation (matching relevance weights under affine condition):

$$\begin{aligned}
\min_\theta \sup_{q \in \mathcal{U}_\lambda} \mathbb{E}_q[\ell(\theta; \mathbf{x}, y)] &= \min_\theta \mathbb{E}_p[r^*(y) \cdot \ell(\theta; \mathbf{x}, y)] && \text{(substituting } q^* = r^* \cdot p\text{)} \\
&= \min_\theta \mathbb{E}_p[w(y) \cdot \ell(\theta; \mathbf{x}, y)] && \text{(identifying } r^*(y) \propto w(y)\text{)} \\
&= \min_\theta \mathcal{L}_{\text{weighted}}(\theta), && \text{(by definition of } \mathcal{L}_{\text{weighted}}\text{)}
\end{aligned}$$

establishing the claimed equivalence between the relevance-weighted objective and the DRO problem. The weighting parameter λ simultaneously controls both the strength of relevance upweighting and the radius of the distributional uncertainty set, providing a unified interpretation: larger λ both amplifies rare samples and hedges against greater distributional uncertainty. ■

D ALGORITHM DETAILS

D.1 TRAINING ALGORITHM

The complete training procedure for TABOVERSAMPLE is presented in Algorithm 2. The only modification to the standard DDPM training loop is the relevance weight $w(y) = 1 + \lambda \cdot \phi(y)$, which multiplies the denoising loss for each sample (Equation 7).

Algorithm 2 TABOVERSAMPLE Training

Require: Training data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, relevance function ϕ, weighting strength λ, diffusion steps T, noise schedule $\{\beta_t\}_{t=1}^{T}$
Ensure: Trained denoising model $\boldsymbol{\epsilon}_\theta$
Compute relevance weights: $w_i \leftarrow 1 + \lambda \cdot \phi(y_i)$ for all i
repeat
  Sample $(\mathbf{x}_0, y) \sim \mathcal{D}$ with associated weight $w(y)$
  Sample $t \sim \mathcal{U}(\{1, \ldots, T\})$
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Compute $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\,\boldsymbol{\epsilon}$
  Update θ via gradient descent on $w(y) \cdot \|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y)\|^2$
until convergence
return $\boldsymbol{\epsilon}_\theta$

D.2 NETWORK ARCHITECTURE

The denoising network $\boldsymbol{\epsilon}_\theta$ is a multi-layer perceptron with residual connections. The target value y and diffusion time step t are each encoded via sinusoidal positional embeddings with $K = 128$ frequency components spanning logarithmically spaced scales from $\omega_1 = 1$ to $\omega_K = 10{,}000$. These embeddings are concatenated and passed through a two-layer MLP to produce the FiLM conditioning parameters $(\boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)})$ for each hidden layer l. The hidden dimension is set to 256 with 4 residual blocks, and we use SiLU activations throughout.

D.3 TRAINING CONFIGURATION

We train with the AdamW optimizer [Kingma and Ba, 2015, Loshchilov and Hutter, 2019] with exponential moving average (EMA) of model weights [Polyak and Juditsky, 1992] and learning rate 3×10^{-4}, weight decay 10^{-4}, and batch size $\min(4096, N_{\text{train}})$. For datasets where $N_{\text{train}} < 4096$ (e.g., Abalone with $N_{\text{train}} \approx 2{,}924$), the effective training regime becomes full-batch gradient descent, which reduces gradient noise and may alter convergence dynamics relative to standard stochastic diffusion training. This reduction in gradient noise could theoretically constrain the diversity of generated samples within rare regions, as diffusion models rely on stochastic gradient noise for optimal score matching; future work should explore noise-injected optimizers (e.g., explicit gradient noise injection or stochastic weight averaging) for extremely small datasets. The number of diffusion steps is $T = 1000$ with a linear noise schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. Training proceeds for 10,000 epochs with early stopping based on a held-out validation loss. The relevance weighting parameter λ is selected from $\{0.5, 1.0, 2.0, 5.0\}$ via cross-validation on the relevance-weighted validation metric.

D.4 CLASSIFIER-FREE GUIDANCE

We optionally employ classifier-free guidance [Ho and Salimans, 2021] during generation. During training, the target conditioning is dropped with probability $p_{\text{uncond}} = 0.1$ by replacing y with a null token. At generation time, the guided noise prediction is:

$$\hat{\boldsymbol{\epsilon}}_\theta(\mathbf{x}_t, t, y) = (1 + s) \cdot \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, y) - s \cdot \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \varnothing), \tag{22}$$

where $s \geq 0$ is the guidance scale and $\varnothing$ denotes the null conditioning. The guidance scale is selected from $\{0.0, 0.5, 1.0, 2.0\}$ via validation.

E EXPERIMENTAL SETUP DETAILS

Datasets. We use nine publicly available regression datasets with naturally skewed targets, organized into a numerical-dominant suite and a categorical-rich suite. The four numerical-dominant benchmarks are **Abalone** (4,177 samples, 8 features), which predicts age from physical measurements with extreme ages in the tails; **CPUActivity** (8,192 samples, 21 features), which predicts processor utilization clustering near low values; **CaliforniaHousing** (20,640 samples, 8 features), which predicts median house value skewed toward lower prices; and **House16H** (22,784 samples, 16 features), which predicts house value with a long right tail. The five categorical-rich benchmarks (43–62% categorical features) are summarized in

Table 3 (Section 5.2). All datasets use a 70/30 train/test split; oversampling is applied only to training data. All oversampling methods generate the same number of synthetic samples, set to match the count of rare-region training points.

Baselines. We compare against thirteen baselines spanning reweighting, interpolation, and generative families. **None**: no oversampling. **None+ϕ**: ϕ-weighted CatBoost sample weights with no synthesis. **RandomOS**: random oversampling with replacement. **SMOTER**: interpolation between rare neighbors. **SMOGN**: interpolation plus Gaussian noise. **DenseWeight** [Steininger et al., 2021]: kernel-density relevance reweighting. **DIR** [Yang et al., 2021]: deep imbalanced regression with feature/label smoothing. **Cal. Mixup** [Kang et al., 2026]: calibrated mixup for imbalanced tabular regression. **CTGAN-Cond** and **TVAE-Cond**: conditional GAN/VAE tabular synthesizers. **Yang-adapted** [Yang et al., 2025]: a balanced-diffusion baseline adapted to the regression setting. **TabDDPM**: unconditional diffusion oversampling. **TabDDPM-Cond**: conditional diffusion with target y as conditioning but without relevance weighting; this baseline isolates the contribution of relevance weighting by matching TABOVERSAMPLE's conditioning architecture.

Downstream Regressors. **CatBoost**: gradient-boosted trees with default hyperparameters (primary regressor). **XGBoost**: gradient boosting. **MLP**: 2-layer, 256 hidden units.

Metrics. **SERA** (Squared Error-Relevance Area): integrates squared error weighted by $\phi(y)$ over the target range. **RMSE$_{\text{rare}}$**: RMSE restricted to test samples with $\phi(y) > 0.5$. **RMSE**: standard root mean squared error over all test samples. Lower is better for all. Results are averaged over 10 seeds.

Code availability. Our implementation is publicly available at `https://github.com/nathanielkang/taboversample`. The repository contains the experiment code and Figures 1–2 from the main paper; benchmark datasets are loaded at runtime via OpenML and scikit-learn.

F DATASET STATISTICS

Table 9: Summary statistics for the benchmark imbalanced regression datasets.

Dataset	N	d	d_{num}	d_{cat}
Abalone	4,177	8	7	1
CPUActivity	8,192	21	21	0
CaliforniaHousing	20,640	8	8	0
House16H	22,784	16	16	0

G FULL RESULTS FOR ALL REGRESSORS

H EXTENDED BASELINES

For readability, the main-paper Table 1 reports the core nine-method roster. Table 12 reports five additional baselines on the four numerical-dominant benchmarks with CatBoost: the kernel-density reweighting method DenseWeight [Steininger et al., 2021], deep imbalanced regression (DIR) [Yang et al., 2021], conditional GAN/VAE synthesizers (CTGAN-Cond, TVAE-Cond), and the regression-adapted balanced-diffusion baseline of Yang et al. [Yang et al., 2025] (Yang-adapted). None of these extended baselines surpasses TABOVERSAMPLE on SERA on any of the four datasets, consistent with the core comparison; reweighting-only methods (DenseWeight, DIR) leave the conditional density unchanged, while the GAN/VAE synthesizers underperform diffusion-based generation, as expected from prior tabular-synthesis results [Kotelnikov et al., 2023].

I ANALYSIS OF DOWNSTREAM MEMORIZATION DYNAMICS

To further characterize the relationship between DSE scores and downstream model behavior, we computed the Spearman rank correlation between the DSE score of each generated sample and the absolute prediction error $|\hat{y} - y|$ of a CatBoost model trained on the augmented data $\mathcal{D}_{\text{aug}} = \mathcal{D}_{\text{train}} \cup \mathcal{S}$. Unlike the holdout stratification experiment in Section 5, which

Table 10: Full results with **MLP** as downstream regressor. **Bold**: best. Underline: second best.

Method	Abalone			CPUActivity		
	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓
None	2.702±.104	3.124±.058	2.105±.027	2.925±.095	4.069±.057	2.687±.013
RandomOS	**2.361±.057**	**2.853±.060**	2.129±.024	**2.894±.152**	4.094±.134	**2.684±.057**
SMOTER [Torgo et al., 2013]	2.500±.118	2.963±.103	2.112±.020	3.317±.556	4.509±.519	2.788±.128
SMOGN [Branco et al., 2017]	2.412±.055	2.885±.039	2.143±.026	3.196±.252	4.408±.293	2.769±.047
TabDDPM [Kotelnikov et al., 2023]	2.453±.034	2.918±.028	2.159±.039	3.692±.193	4.749±.167	2.998±.019
TabDDPM-Cond [Kotelnikov et al., 2023]	2.478±.040	2.925±.035	2.146±.033	3.152±.168	4.367±.142	2.814±.020
TABOVERSAMPLE	2.556±.117	2.885±.062	2.386±.062	2.908±.082	**4.058±.068**	2.698±.018

Method	CaliforniaHousing			House16H		
	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓
None	0.141±.004	0.918±.019	0.529±.009	1128.6M±48.7M	79333±1645	38183±654
RandomOS	0.125±.001	0.834±.004	0.529±.004	993.4M±29.5M	73279±1098	37671±371
SMOTER [Torgo et al., 2013]	0.128±.004	0.865±.018	**0.519±.002**	998.7M±19.0M	74514±1354	**37117±583**
SMOGN [Branco et al., 2017]	0.125±.002	0.832±.004	0.530±.004	1014.0M±36.1M	74017±1498	37996±515
TabDDPM [Kotelnikov et al., 2023]	0.141±.008	0.913±.036	0.539±.002	1131.2M±8.2M	79353±296	37880±138
TabDDPM-Cond [Kotelnikov et al., 2023]	0.135±.005	0.889±.025	0.533±.004	1067.3M±28.8M	77524±1054	37618±312
TABOVERSAMPLE	**0.124±.005**	**0.833±.034**	0.526±.007	**986.2M±26.4M**	**72847±1042**	37389±385

validates generalization on unseen test data, this analysis measures the relationship between sample quality and the model's capacity to absorb individual synthetic samples into its training representation.

Across all four datasets, the correlation is positive and statistically significant (Abalone: $\rho = 0.41$, CPUActivity: $\rho = 0.38$, CaliforniaHousing: $\rho = 0.45$, House16H: $\rho = 0.43$; all $p < 0.001$). Samples with lower DSE scores (higher self-consistency under the learned denoising model) are more readily absorbed by the downstream tree model, as reflected by smaller in-sample prediction errors. This in-sample analysis complements the out-of-sample holdout stratification experiment, together providing converging evidence that the DSE score captures a meaningful notion of synthetic sample quality.

Table 11: Full results with **XGBoost** as downstream regressor. **Bold**: best. Underline: second best.

Method	Abalone			CPUActivity		
	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓
None	3.114	3.321	2.257	2.053	3.501	2.242
RandomOS	**2.865**	**3.163**	2.222	2.014	3.445	2.216
SMOTER [Torgo et al., 2013]	3.017	3.251	2.275	2.059	3.467	2.249
SMOGN [Branco et al., 2017]	3.011	3.246	2.247	2.039	3.448	2.231
TabDDPM [Kotelnikov et al., 2023]	2.957	3.235	2.233	**2.001**	**3.439**	2.227
TabDDPM-Cond [Kotelnikov et al., 2023]	2.941±.025	3.201±.015	2.240±.012	2.018±.011	3.451±.010	2.233±.008
TABOVERSAMPLE	2.933±.043	3.183±.027	2.274±.028	2.009±.014	3.442±.013	2.238±.011

Method	CaliforniaHousing			House16H		
	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓	SERA ↓	$\text{RMSE}_{\text{rare}}$ ↓	RMSE ↓
None	0.104	0.785	0.456	781.8M	66438	**31362**
RandomOS	0.099	0.753	0.456	807.7M	67227	32187
SMOTER [Torgo et al., 2013]	0.102	0.769	0.460	804.2M	67341	32022
SMOGN [Branco et al., 2017]	0.098	0.738	0.463	813.2M	67732	32271
TabDDPM [Kotelnikov et al., 2023]	0.104	0.780	0.467	796.2M	67233	31607
TabDDPM-Cond [Kotelnikov et al., 2023]	0.102±.002	0.772±.005	0.465±.004	790.8M±6.2M	66752±245	31573±128
TABOVERSAMPLE	**0.098±.002**	**0.738±.008**	0.466±.005	**779.1M±8.5M**	**65978±312**	31518±165

Table 12: **SERA** ↓ for the extended baselines with **CatBoost** on the four numerical-dominant benchmarks (10 seeds; means). TABOVERSAMPLE (repeated from Table 1 for reference) is best on all four. **Bold**: best per column.

Method	Abalone	CPUActivity	CaliforniaHousing	House16H
DenseWeight [Steininger et al., 2021]	2.834	2.371	0.107	783.6M
DIR [Yang et al., 2021]	2.857	2.389	0.108	788.1M
CTGAN-Cond	2.962	2.418	0.114	821.3M
TVAE-Cond	2.991	2.456	0.116	834.7M
Yang-adapted [Yang et al., 2025]	2.806	2.297	0.105	779.4M
TABOVERSAMPLE	**2.748**	**2.205**	**0.103**	**772.1M**