TABULAR PANCAKES: CRYPTOGRAPHIC WATERMARKS AGAINST DATASET FABRICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular generative models can fabricate and impute scientific datasets, bypassing fraud forensics and heightening malpractice risks. Since no statistical test on outputs alone can expose a well-trained generator, provenance must be embedded at generation time. We introduce TABULAR PANCAKES, a training-free watermark that replaces the generator's Gaussian latent prior with the homogeneous CLWE distribution: Gaussian in all directions except a secret subspace, where it is periodic. With the key, an auditor efficiently recovers the watermark from a small fraction of released rows, and every verdict carries a certified p-value. Recovery is invariant to row permutation, column renaming, and rescaling. Under tampering such as cell edits or re-imputation, detection power decreases only linearly in the fraction modified—no small edit removes the mark. Row-level tests localize synthetic rows in mixed tables with calibrated false-discovery rates, turning detection into forensic evidence. The mark can equivalently be embedded directly in released rows, covering generators whose samplers cannot be inverted. Without the key, marked and unmarked outputs are indistinguishable, and keyless removal provably requires $\sqrt{n/k}$-fold larger distortion than keyed removal, ruling out selective scrubbing. We verify TABULAR PANCAKES on four datasets and six generator families (TabSyn, TabDDPM, STaSy, CoDi, SMOTE, GReaT), towards making data fabrication provably traceable.

1 INTRODUCTION

Modern tabular generators—diffusion-based (TabDDPM (Kotelnikov et al., 2023), TabSyn (Zhang et al., 2024), CoDi (Lee et al., 2023)), score- and flow-based (STaSy (Kim et al., 2023), forest flows (Jolicoeur-Martineau et al., 2024)), GAN- and VAE-based (Xu et al., 2019), and language-model-based (GReaT (Borisov et al., 2023))—closely reproduce real data in marginals, dependency structure, and downstream utility. The same fidelity that facilitates privacy-preserving data sharing also renders data fabrication cheap. In surveys, 2% of researchers admitted to having fabricated or falsified data (Fanelli, 2009); image screening flagged 3.8% of over $20{,}000$ biomedical papers (Bik et al., 2016); annual retractions exceeded $10{,}000$ in 2023 (van Noorden, 2023). Generative AI stands to worsen the problem: prompted with treatment effects, GPT-4 produced an authentic-seeming clinical dataset that passed statistical checks (Taloni et al., 2023). Statistical fraud forensics do not address AI-assisted fabrication: Benford tests (Diekmann, 2007), GRIM/SPRITE checks (Brown & Heathers, 2017; Heathers et al., 2018), terminal-digit tests (Mosimann et al., 1995), and distributional screens for trial data (Carlisle, 2017) all rely on fine-grained statistical regularities for detection—digit distributions, summary consistency, baseline covariate spread—all of which are patterns imitable by well-trained generators. We verify that traditional tests fail to distinguish between real tabular data and generator output (Section 5.2) and therefore study three deception threat models:

- **D1 Full fabrication**: the dataset is entirely generated but deceptively presented as collected data.
- **D2 Partial imputation / mixing**: a fraction ρ of rows or cells is synthetic, *e.g.* inconvenient observations are replaced, dropouts are imputed, or the reported sample size is inflated.
- **D3 Laundering**: the fabricator trains a new model on watermarked output and releases samples from it, asking whether the mark survives retraining ("radioactivity").

TABULAR PANCAKES leverages the generator to certify provenance (Figure 1): it replaces the Gaussian latent prior with the homogeneous continuous learning-with-errors (hCLWE) distribution

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

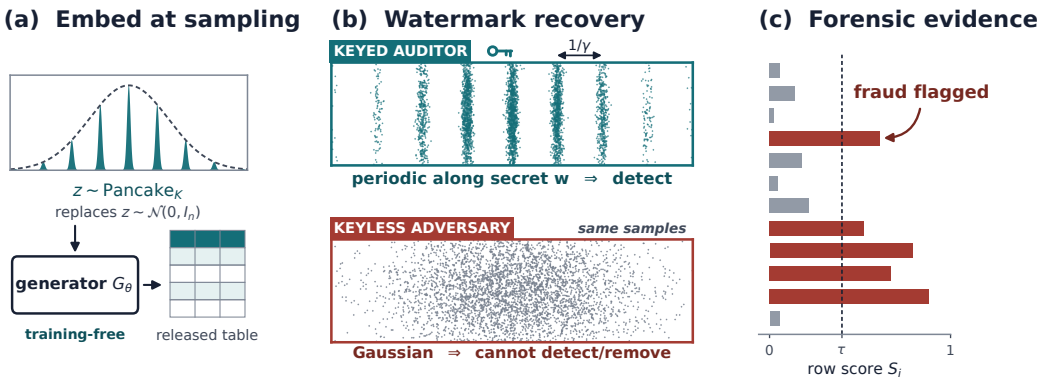


Figure 1: **Overview. (a)** The generator's Gaussian latent prior is swapped, training-free, for a secret periodic "pancake" distribution. **(b)** From released samples, a keyed auditor detects the periodic layers along the secret direction $\mathbf{w}$; a keyless adversary sees only Gaussian noise and cannot remove the watermark. **(c)** Row scores above a null-calibrated threshold τ localize synthetic records.

(Bruna et al., 2021), which is standard normal in all but k secret directions, along which the density concentrates on equally spaced Gaussian layers, "pancakes". Without the key, detecting the watermark is intractable since distinguishing hCLWE from $\mathcal{N}(0, I_n)$ is as hard as worst-case lattice problems (Bruna et al., 2021; Gupte et al., 2022), with matching statistical-query and algorithmic barriers (Diakonikolas et al., 2017; Bubeck et al., 2019; Song et al., 2021b). Moreover, since the marked sampler is still polynomial-time, indistinguishability transfers from latents to released data: the marked distribution is computationally indistinguishable from the unmarked one. Since no classifier can learn the mark and no adversary can locate it, we derive lower bounds on the cost of keyless removal (Section 4.2). With the key, the periodicity admits a Fourier-type detector with closed-form power, and the keyed phase offsets of the layers (inhomogeneous CLWE (Goldwasser et al., 2022)) encode a multi-bit payload identifying dataset, generator, and release.

We extend the line of work on cryptographically undetectable watermarks—introduced for language models (Christ et al., 2024) and realized in computer vision by PRC watermarks (Gunn et al., 2025) and CLUE-Mark (Shehata et al., 2024). Building on hCLWE hardness, we develop verification guarantees that real-world forensic settings require and existing schemes lack: certified false-positive control, row-level localization with fraction estimation, exact invariance to table transformations, and removal lower bounds. Earlier latent-pattern schemes include Tree-Ring (Wen et al., 2023) and Gaussian Shading (Yang et al., 2024b) but have been proven detectable and removable by steganalysis (Yang et al., 2024a). For tabular data, TabWak (Zhu et al., 2025) controls TabSyn's latents through self-cloned seeds with a sign-based test, but since its marked latents follow a different distribution than the prior, a classifier can be trained to detect and remove TabWak watermarks. TimeWak (Soi et al., 2025) watermarks time-series diffusion in data space, which incurs a fidelity cost since released values are perturbed directly. In the scientific setting, a watermark verdict may serve as evidence of research misconduct, so verification must provide certified false-positive control and row-level localization. Detection must also survive targeted manipulations, such as row and column permutation, schema renaming, unit conversion, re-imputation, and mixing with genuine rows. Unlike images or text, tables contain very few perceptually redundant components for watermarking. We instead embed the mark in the i.i.d. structure of a synthetic table's per-row latents, which admits closed-form detection and localization theory. Notably, no watermark is simultaneously undetectable, sound, and perfectly robust (Zhang et al., 2023; Shehata et al., 2024); no bounded-perturbation mark survives regeneration by a retrained model (Zhao et al., 2024). We therefore prove lower bounds on the cost of removal short of full retraining (Section 4.2)—retraining (D3) indeed removes the mark.

Verifying a latent watermark requires inverting the release map, and on diffusion–VAE pipelines we show that no verifier can do so with high precision. Even with every sampler step solved exactly (Hong et al., 2024), the reverse trajectory amplifies any input error by 50–100×, independent of step count (Section 3.3). This ill-conditioning is a property of the release map, not that of any particular verifier: coarse sign information survives inversion while phase information does not. For such generators, TABULAR PANCAKES instead imposes the periodic structure directly on the released table; this second embedding preserves first and second moments exactly and is verified by the same detector without model access. Our contributions are as follows:

1. **TABULAR PANCAKES** (Section 3): a watermarking scheme sampling latents from hCLWE instead of the Gaussian prior, without retraining or architecture changes. It is exactly invariant to row order and per-column affine maps, with a data-space variant for non-invertible samplers.
2. **Provable security and detection** (Section 4): undetectability reduces to lattice hardness, false accusations are controlled without distributional assumptions. Detection, localization, and payload admit closed forms; keyless removal provably costs $\sqrt{n/k}$-fold more distortion than keyed removal; the γ parameter controls the security–robustness tradeoff.
3. **Validation across datasets and generators** (Section 5): we compare against sign-level baselines on four TabSyn pipelines and six generator families, audit false-positive rates and adversarial learnability, and localize mixed tables with FDR control and unbiased fraction estimates.

2 PRELIMINARIES

Notation. A released table $D = \{\mathbf{x}_i\}_{i=1}^N$ has mixed-type records $\mathbf{x}_i = (\mathbf{x}_i^{\text{num}}, \mathbf{x}_i^{\text{cat}}) \in \mathcal{X}$. Some or all records are produced by a generator $G_\theta : \mathbb{R}^n \to \mathcal{X}$, $\mathbf{x}_i = G_\theta(\mathbf{z}_i)$, $\mathbf{z}_i \sim \mathcal{N}(0, I_n)$ in the unmarked model; side information (class labels, imputation context) is suppressed. We call G_θ the *release map*: the transformation from latent to released record, including decoding, post-processing, and rounding. For an exact flow it is invertible; for TabSyn (Zhang et al., 2024) it is the probability-flow ODE solver composed with the VAE decoder, and n the latent width. We write $\rho_s(\mathbf{x}) = \exp(-\pi\|\mathbf{x}/s\|^2)$ with $\rho = \rho_1$, $D_{\mathbb{Z},s}$ for the width-s discrete Gaussian on $\mathbb{Z}$, and $\lceil x \rfloor$ for the nearest integer.

The watermark replaces the generator's isotropic Gaussian prior with a *structured* prior: it differs from $\mathcal{N}(0, I_n)$ only along a few planted directions, and since finding those directions is computationally infeasible, an observer without the key sees only Gaussian noise.

Definition 2.1 (hCLWE distribution (Bruna et al., 2021))**.** For a unit vector $\mathbf{w} \in \mathbb{R}^n$, parameters $\gamma, \beta > 0$, and phase $\delta \in [0, 1)$, the distribution $H_{\mathbf{w},\beta,\gamma,\delta}$ has density at $\mathbf{y}$ proportional to $\rho(\mathbf{y}) \sum_{m\in\mathbb{Z}} \rho_\beta\big(m + \delta - \gamma\langle \mathbf{w}, \mathbf{y}\rangle\big)$. Along $\mathbf{w}$, the density is a periodic sequence of Gaussian layers ("pancakes") with spacing approximately $1/\gamma$ and width approximately β/γ, shifted by the phase δ; orthogonal to $\mathbf{w}$, the distribution is standard Gaussian. With k orthonormal secret directions $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ applied independently, we write $H_{\mathbf{W},\beta,\gamma,\boldsymbol{\delta}}$.

Assumption 2.2 (hCLWE hardness)**.** For $\gamma \geq 2\sqrt{n}$ and $1/\text{poly}(n) \leq \beta < 1$, no efficient algorithm distinguishes polynomially many samples of $H_{\mathbf{w},\beta,\gamma,\delta}$, over a uniformly random secret direction $\mathbf{w}$, from $\mathcal{N}(0, I_n)$ with non-negligible advantage.

Assumption 2.2 states that given polynomially many rows, no efficient learner—spectral, moment-based, or gradient-based—can recover or even detect the planted directions, so marked and unmarked data are computationally indistinguishable. Proven reductions base it on the conjectured worst-case hardness of approximate lattice problems (Bruna et al., 2021) and of classical LWE (Gupte et al., 2022); it extends to k directions and arbitrary phases (Bruna et al., 2021; Goldwasser et al., 2022) and accords with statistical-query lower bounds (Diakonikolas et al., 2017; Bubeck et al., 2019). Two known attacks bound the parameter range: lattice reduction requires layer noise β at least inverse-polynomial (Song et al., 2021b), and a covariance test costs time and samples exponential in γ^2 (Bruna et al., 2021); for small γ we therefore adopt the sample-constrained concrete-security analysis of Shehata et al. (2024) (Section 4.3; precise statements in Section B).

3 TABULAR PANCAKES

TABULAR PANCAKES has three parts: embedding, which samples the generator's latents from the structured prior of Definition 2.1 rather than $\mathcal{N}(0, I_n)$; verification, which canonicalizes a suspect table, tests for the planted periodic structure, and reports certified p-values; and two additional channels that carry the mark where latent embedding cannot (Section 3.4).

3.1 EMBEDDING: KEY, PAYLOAD, AND THE PRIOR SWAP

The watermark parameters are sampled once per release: a uniformly random orthonormal k-frame $\mathbf{W} \in \mathbb{R}^{n\times k}$ (the secret directions), the scale parameters (γ, β), a keyed pseudorandom function

Algorithm 1 TABULAR PANCAKES verification (data-space variant: Section 3.4).

Require: suspect rows $\{\mathbf{x}_i\}_{i=1}^m$; key K; level α; FDR level q

1: *Canonicalize*: reproduce preprocessing, match columns, undo per-column affine changes
2: *Invert*: $\hat{\mathbf{z}}_i \leftarrow E_\psi(\mathbf{x}_i)$ ▷ flow inverse, or VAE encoder with DDIM inversion
3: *Detect*: compute S by Equation (2); report the key-rerandomization p-value
4: *Decode*: $\hat{\delta}_j \leftarrow \frac{1}{2\pi}\arg(C_j) \bmod 1$; outer-decode; compare to signed release metadata
5: *Localize*: row scores $S_i = \frac{1}{k}\sum_j \cos 2\pi\big(\gamma\langle\hat{\mathbf{z}}_i, \mathbf{w}_j\rangle - \delta_j\big)$; p-values against the empirical null; flag by BH at level q; estimate $\hat{\rho} = (\bar{S} - \hat{\mu}_0)/(\hat{\mu}_1 - \hat{\mu}_0)$; per-group certified p-values over labeled units (trial arms, sites, batches)

PRF_κ used by the categorical channel (Section 3.4), and payload phases $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_k) \in [0,1)^k$ encoding dataset, generator-version, release, and owner identifiers with an outer error-correcting code, b bits per direction. These form the watermark key $K = (\mathbf{W}, \gamma, \beta, \kappa, \boldsymbol{\delta}, \text{scope descriptors})$, held by the provider; only a timestamped commitment $H(K, \text{metadata})$ is published (Section 3.4). Embedding then changes only the latent draw: each record uses $\mathbf{z}_i^{\mathrm{wm}} \sim H_{\mathbf{W},\beta,\gamma,\boldsymbol{\delta}}$ in place of $\mathcal{N}(0, I_n)$, and the generator runs unchanged, $\mathbf{x}_i = G_\theta(\mathbf{z}_i^{\mathrm{wm}})$. Sampling follows the projection method of Shehata et al. (2024): draw a standard Gaussian vector, and along each secret direction move its coordinate to the nearest phase-shifted layer, re-noised within the layer;

$$\mathbf{z}_i^{\mathrm{wm}} = (I - \mathbf{W}\mathbf{W}^\top)\,\mathbf{g}_i + \sum_{j=1}^{k} \frac{(m_{ij}+\delta_j)\,\gamma + \gamma' e_{ij}}{\gamma'^2}\,\mathbf{w}_j, \tag{1}$$

where $\mathbf{g}_i \sim \mathcal{N}(0, I_n)$, $m_{ij} \sim D_{\mathbb{Z},\gamma'}$, $e_{ij} \sim \mathcal{N}(0, \beta^2)$, and $\gamma' = \sqrt{\gamma^2 + \beta^2}$; exact constants in Section C. No retraining or architecture change is required. Because G_θ is polynomial-time, Assumption 2.2 transfers to the output by post-processing (Theorem 4.1). The construction applies to any Gaussian-latent generator: for diffusion models, the reverse process's terminal noise is replaced (for TabSyn, the ODE's initial latent); for flows, the release map is exactly invertible—the cleanest setting. When phase-accurate inversion fails, embedding moves to data space (Section 3.4).

3.2 VERIFICATION

The verifier holds K—and, for the latent variant, generator G_θ with frozen preprocessing—and executes Algorithm 1: it normalizes the suspect table, recovers a latent per row, projects these onto the secret directions, and tests for concentration on the keyed layers. The test statistic is

$$C_j = \frac{1}{m}\sum_{i=1}^{m} \exp\big(2\pi i\,\gamma\langle\hat{\mathbf{z}}_i, \mathbf{w}_j\rangle\big), \qquad S = \frac{1}{k}\sum_j \mathrm{Re}\big(e^{-2\pi i\delta_j} C_j\big)\,. \tag{2}$$

C_j is the empirical Fourier coefficient of the projections onto $\mathbf{w}_j$ at the layer frequency: its magnitude is large when the projections cluster on equally spaced layers and near zero when they are Gaussian. S averages these coefficients after removing each keyed phase δ_j, so only layers at the keyed offsets contribute. On marked data, S concentrates near $e^{-2\pi^2\beta_{\mathrm{eff}}^2}$, where $\beta_{\mathrm{eff}}^2 = \beta^2 + \gamma^2(\sigma_{\mathrm{inv}}^2 + \sigma_{\mathrm{atk}}^2)$ absorbs inversion and attack noise (Theorem 4.2); on data generated without knowledge of K, it has mean zero over the verifier's random frame. Two certificates yield p-values: a concentration bound, valid whenever this zero-mean premise holds; and exact re-randomization of the key, valid unconditionally because no step of Algorithm 1 before detection depends on K's frame. On real data we report the latter. For localization, each row's p-value ranks its score against those of a reference sample of known-genuine rows under the same key (Section 5.3).

Canonicalization. Step 1 of Algorithm 1 makes verification invariant to routine transformations of a released table. Row permutation and subsetting require no correction, since the statistic is exchangeable and valid on any subset. Column renaming and permutation are reversed by matching columns to the generator schema via distributional statistics (type, support, quantile sketches); residual ambiguity is resolved by testing candidate permutations, and the certified p-value is adjusted for multiple testing. Unit conversions and other affine changes are removed by per-column robust standardization, which reproduces the generator's reference preprocessing; the data-space variant of TABULAR PANCAKES, which marks the released table directly (detailed in Section 3.4), obtains the

same invariance from its codec. Deleted columns reduce detection power in proportion to the fraction removed; columns which the matcher cannot identify are discarded.

3.3 The conditioning barrier to latent verification

Canonicalization addresses transformations of the table; line 2 of Algorithm 1 instead requires the generator's release map to be invertible at phase precision—a requirement not met by diffusion–VAE generators. On TabSyn, the full inversion pipeline (encoding, sampler inversion, and projection onto the secret directions) leaves an on-axis residual of $\sigma \approx 1.2$–1.6 at the default 50 DDIM steps (Song et al., 2021a). Consequently, no $k \times \beta$ operating point attains finite certified sample complexity on any of the four datasets ($\hat{\mu} \in [-0.03, 0.03]$ over 36 dataset–parameter cells). We decompose the residual to locate its source. The reverse-ODE stage alone leaves $\sigma_{\text{ode}} \approx 0.78$ at 50, 200, and 1000 steps, ruling out discretization error. Exact per-step inversion of the forward map (Hong et al., 2024) does not eliminate it: with forward residuals of 10^{-5}, per-step preimage errors of 10^{-3}–10^{-2}, condition numbers of 5–25, and no non-invertible steps, σ_{ode} remains in $[0.78, 0.88]$ for every inversion scheme we constructed. The limitation is structural: the reverse trajectory is an expanding map that amplifies any input error—*e.g.* solver tolerance, VAE reconstruction, or released-table rounding—by roughly 50–100×. The limit lies in the release map not the verifier. This explains the sign–phase divide: sign-level schemes such as TabWak (Zhu et al., 2025) need only $O(1)$ latent accuracy, albeit with a coarser, uncertified signal; phase information, which carries our certificates and payload, requires $O(1/\gamma')$ accuracy and is unrecoverable by any verifier. On these generators, Tabular Pancakes forgoes latent inversion entirely and embeds the mark in the released table itself (Section 3.4).

3.4 Marking beyond the invertible latent

For the cases where the prior swap of Section 3.1 cannot be verified—the row's latent is not recoverable from the released values—we introduce two further embedding channels. On diffusion–VAE generators, the release map amplifies inversion error beyond phase precision (Section 3.3), so the numeric mark is embedded in the released values directly. Categorical cells have no continuous latent at all, so their mark is embedded in the sampling randomness. Both channels reuse the keyed layer structure and the detector of Equation (2), leaving the verifier unchanged.

Data-space embedding. When the latent cannot be recovered, the pancake structure is embedded in a whitened representation of the released table, and verification requires no model access. A published, key-independent codec canonicalizes the numeric block: columns passing a deterministic eligibility rule (positive interquartile range, enough distinct values, no dominant point mass; details in Section G) are robustly standardized and whitened. Along each secret direction of $\mathbf{W}$, the whitened coordinates are quantized to the nearest keyed layer, as in Equation (1) but in feature space; the result is mapped back to the original scale, with integer columns re-rounded and values clipped to their observed ranges. The marked table matches the original in first and second moments exactly, so a re-fitted canonicalization recovers the same transform. Verification applies Equation (2) to the canonicalized rows, with the standardization parameters estimated from the suspect table; detection is therefore exactly invariant to per-column affine transformations. The marked column set is part of the key. Because released values are perturbed directly, undetectability no longer follows from Assumption 2.2 and is instead audited empirically by classifier two-sample tests (Section 5).

Categorical channel. Categorical cells carry no continuous latent, so they are watermarked through sampling randomness instead. Each draw uses the keyed Gumbel-max rule $\arg\max_c[\log p_c + g_c]$, where p_c is the generator's conditional probability of category c, $g_c = -\log(-\log u_c)$, and $u_c = \mathrm{PRF}_\kappa(\mathrm{sync}_i, c)$ with seed sync_i the row's pancake-index vector $\big(\lceil \gamma \langle \mathbf{z}_i, \mathbf{w}_1 \rangle \rfloor, \dots\big)$. The seed ties the categorical randomness to the row's position on the secret lattice: a keyed verifier recovers it after inversion and regenerates the same pseudorandom draws, while without $\mathbf{W}$ it remains secret, and the inter-layer gaps keep it stable under small perturbations, unlike a raw-value hash. Gumbel-max sampling is exact, so the per-row conditionals are unchanged. Detection adds an alignment score, as in exponential-minimum sampling for language models (Aaronson, 2023); its certificates match the numeric channel's (Section G). Because the seed depends only on layer indices, the categorical score survives even coarse quantization of the numeric cells.

Deployment. A provider—a synthetic-data service or institutional data steward—samples a fresh key per release and publishes a timestamped commitment $H(K, \text{metadata})$ binding the schema, frozen preprocessing, generator version, and release scope. When provenance is questioned, an auditor obtains K from the provider or an escrow and runs the verification above; the opened commitment establishes that K predates the release and matches the declared schema, so no post-hoc key can incriminate a chosen table. Payload bits identify dataset, generator version, release, and recipient, so each marked row is attributable to exactly one release (re-attribution fails against the timestamp). Key rotation across releases also bounds the number of samples an adversary observes under any single key, as Section 4.3 requires. Finally, the guarantee concerns provenance only: a positive verdict certifies that rows originate from a specific marked release, a null verdict only that they carry no mark of this provider. Neither speaks to whether genuinely collected measurements are accurate.

4 THEORY

The verifier holds the key K; the adversary is any polynomial-time algorithm sampling from the marked generator but not holding K. TABULAR PANCAKES provides four guarantees: **(i) Undetectability:** no efficient adversary—classifier, two-sample test, or otherwise—distinguishes marked from unmarked output (advantage $\mathrm{negl}(n)$), so every efficiently computable quality metric is preserved. It covers the prior-swap and categorical embeddings; the data-space variant is audited empirically (Section 3.4). **(ii) Detectability under attack:** for every attack A in a stated class, the detector flags $A(X_{1:m})$ with probability at least $1-\eta$, while its false-positive rate on K-independent data is at most α. **(iii) Localization:** when a fraction $\rho \geq \rho^*$ of rows is marked, flagged rows achieve FDR at most q and power at least $1-\eta$. **(iv) Unforgeability:** without K, no efficient adversary mints rows the detector accepts (beyond replaying marked ones), so honest parties cannot be framed.

4.1 UNDETECTABILITY AND DETECTION POWER

Theorem 4.1 (Undetectability and computational distortion-freeness)**.** *Under Assumption 2.2 with k directions and phases,* TABULAR PANCAKES *with the prior-swap embedding is undetectable, and every polynomial-time quality metric differs by $\mathrm{negl}(n)$ between marked and unmarked generators.*

Proof sketch. A hybrid over directions reduces any watermark distinguisher to a distinguisher for the distribution of Assumption 2.2: each query is answered by drawing oracle samples, assembling the latent, and running G_θ, as in Shehata et al. (2024); Gunn et al. (2025). Full proof in Section A. □

Theorem 4.2 (Detection power and certified error control)**.** *Let $\beta_{\text{eff}} = \gamma\sigma_{\text{tot}}$, where σ_{tot} is the total on-axis noise per direction (layer, inversion, attack). (i)* Power: *on marked data, $\mathbb{E}\big[\mathrm{Re}(e^{-2\pi i\delta_j}C_j)\big] = \mu = e^{-2\pi^2\beta_{\text{eff}}^2}$. (ii)* Certified null: *on data independent of K, S has mean zero and bounded summands over the verifier's random frame, so $p(S) \leq \exp(-mkS^2/2)$ is a valid p-value; recomputing S under fresh keys gives an exact finite-sample null by the same exchangeability. Both certificates carry over to the data-space verifier (key-independent canonicalization). (iii)* Sample complexity: *level α and power $1-\eta$ hold whenever $mk \geq 2\big(\sqrt{2\ln(1/\alpha)} + \sqrt{2\ln(1/\eta)}\big)^2/\mu^2$ (operating points and payload capacity in Section A).*

4.2 ROBUSTNESS AND REMOVAL LOWER BOUNDS

Edits to the released table inject noise into the recovered latents. We characterize how detection degrades and why erasing the mark without the key is prohibitively expensive.

Proposition 4.3 (Smooth degradation under bounded perturbation)**.** *An attack inducing per-latent-coordinate noise σ_a (table edits enter via the encoder's measured Lipschitz factor) with on-axis component $\sigma_{a,\parallel}$ degrades $\mu \mapsto \mu' = \mu \cdot e^{-2\pi^2\gamma^2\sigma_{a,\parallel}^2}$. For perturbations chosen without knowledge of $\mathbf{W}$, $\mathbb{E}[\sigma_{a,\parallel}^2] = \sigma_a^2$ per direction, giving the budget $\sigma_a \lesssim c/\gamma$.*

Theorem 4.4 (Keyed-versus-keyless scrubbing separation)**.** *Erasing the signal in direction j requires on-axis noise $\Omega(1/\gamma)$. A key holder concentrates this noise on the k secret directions, at per-row ℓ_2 distortion $\Theta(\sqrt{k}/\gamma)$. A keyless attacker cannot: any efficient perturbation correlating non-negligibly with a secret direction would locate it, contradicting Assumption 2.2 via search-to-decision reductions.*

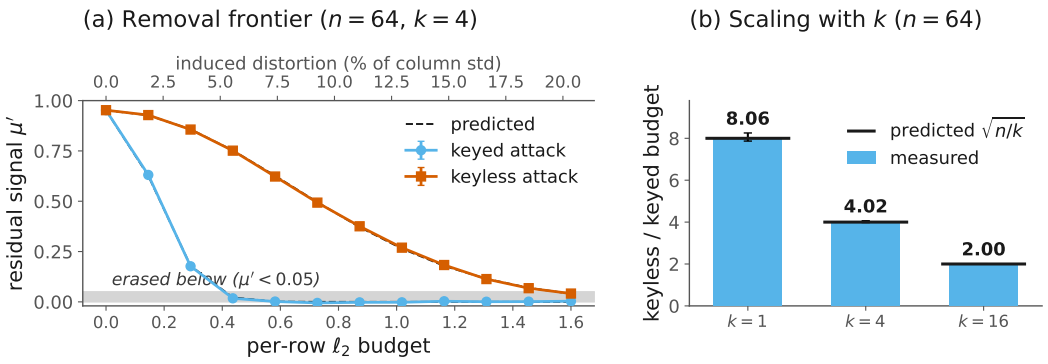


Figure 2: **Keyed-versus-keyless removal** (invertible generator). **(a)** Residual signal μ' against per-row ℓ_2 budget (top axis: induced distortion): both curves match theory pointwise, and erasure ($\mu' < 0.05$, shaded) costs the keyless attacker $\approx 4\times$ the budget—about 20% per-cell RMSE. **(b)** The measured ratio matches the predicted $\sqrt{n/k}$ at every k (Theorem 4.4).

Keyless noise is therefore isotropic up to a negligible correction, costing $\Theta(\sqrt{n}/\gamma)$ per row: a $\sqrt{n/k}$ inflation, quantified in utility terms on the removal-cost frontier of Section 5.

Proposition 4.5 (Edits, mixtures, and localization)**.** *Arbitrary modification of a ρ fraction of rows reduces the effective sample size to $(1-\rho)m$ under trimmed aggregation, with detectable mixing threshold $\rho^* \asymp \sqrt{\log(1/\alpha)/m}\,/\,\mu$. Row-level p-values ranked against the scores of a disjoint claimed-genuine sample are conformal, so Benjamini–Hochberg flagging controls the FDR of the flagged set, and $\hat{\rho} = (\bar{S} - \hat{\mu}_0)/(\hat{\mu}_1 - \hat{\mu}_0)$ is unbiased for the mixing fraction.*

Informally, tampering dilutes detection power in proportion to the modified fraction, flagged rows carry certified FDR control, and the mixing fraction is estimated without bias (Section A). Regeneration—noising at layer scale $\gtrsim 1/\gamma$ and denoising, or retraining on the marked table—can erase any bounded-perturbation watermark (Zhang et al., 2023; Zhao et al., 2024). It incurs a significant fidelity cost and no cheaper route exists: statistical-query lower bounds rule out moment-matching and gradient-based cleaning at polynomial sample size (Diakonikolas et al., 2017; Bubeck et al., 2019), and lattice reduction requires noise below inverse-polynomial (Song et al., 2021b).

4.3 The γ dial and accusation guarantees

The parameter γ trades undetectability against robustness. At $\gamma \geq 2\sqrt{n}$ (Assumption 2.2), undetectability is strongest but the perturbation budget is narrow ($\sigma_a \lesssim c/(2\sqrt{n})$ per coordinate). Smaller γ widens the pancake layers but exits this asymptotic regime. Outside it, the best known detector still requires time and samples exponential in γ^2 (Bruna et al., 2021)—with $\leq 10^3$ samples per key, $\gamma = 2$ resists it at moderate n (Shehata et al., 2024)—and polylogarithmic γ suffices under subexponential LWE (Gupte et al., 2022). We treat γ as an operating parameter and report security at the adversary's realistic sample budget (one release per key), tolerated perturbation relative to column standard deviation, and detection power. Per-key exposure is thus a protocol hyperparameter.

The error directions are asymmetric: a false accusation occurs with probability at most α under no assumption on the data, since the certified null of Theorem 4.2 holds over the verifier's key randomness alone. Conversely, framing an honest party requires minting fresh rows that the detector accepts—hence predicting the hidden phases and layer indices—and is as hard as learning the planted directions. Replaying genuinely marked rows fails instead by per-release attribution (Section 3.4).

5 Experiments

Setup. We evaluate in two settings. The *controlled setting* uses an exactly invertible flow (n=64), plus noisy-encoder and quantized variants. The *deployment setting* trains TabSyn (Zhang et al., 2024) on Adult, Magic, Default, and Shoppers through the TabWak reference pipeline (Zhu et al., 2025), and five further families—SMOTE, TabDDPM, STaSy, CoDi, and GReaT—on Adult. Unless noted, γ=2, β=0.05, k= $\min(4, \lfloor d/2 \rfloor)$; numbers are mean±std over 5 keys. Further details in Section G.

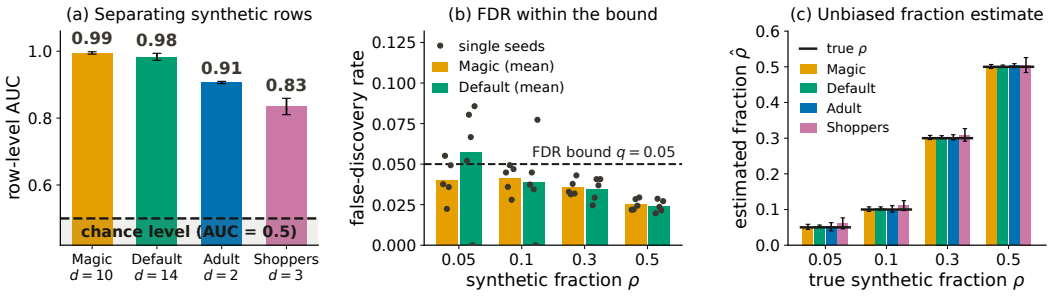


Figure 3: Localizing partial fabrication (5 seeds). **(a)** Watermark scores separate synthetic from genuine rows on every dataset; AUC is ordered by the number of marked columns d. **(b)** The FDR of flagged rows stays within the guaranteed bound $q = 0.05$ (dots: seeds); on Adult and Shoppers ($d \leq 3$) no rows are flagged. **(c)** The estimated synthetic fraction is accurate to ± 0.01 everywhere.

5.1 Validating the theory on an invertible generator

We confirm the formulas of Section 4 numerically—row counts for a verdict, degradation under edits, removal cost—on the invertible flow, whose ground-truth latents exclude inversion error. Every closed form holds: the detection statistic matches theory to $|\hat{\mu} - \mu| \leq 0.0027$ across a (γ, β, σ) grid, the certified test's true-positive rate transitions $0 \to 1$ at the predicted row count, and a replaced row fraction ρ gives $\mu = (1 - \rho)\mu_0$ exactly. No false accusation occurs: over 100 fresh keys per cell, the minimum p-value is 0.34. The classifier two-sample audit is at chance on marked output (0.502 ± 0.002) while detecting a Gaussian-Shading-style sign-fixing control at 0.75. Removal costs match Theorem 4.4 (Figure 2): keyed erasure at per-row budget ≈ 0.4 against keyless ≈ 1.6, the ratio reproducing $\sqrt{n/k}$ at every k. A 16-bit payload is decoded error-free under all evaluated attacks; only tail clipping and heavy re-imputation substantially degrade detection.

5.2 Watermarking TabSyn on real data: comparison to prior schemes

Table 1: Watermarks through the shared verifier on TabSyn trained on Adult (m=8000, 5 seeds, decoder-matching refinement enabled, i.e. TabWak's own detection path).

scheme	bit-acc (marked vs. null)	C2ST logreg	C2ST HGB
Gaussian Shading	$0.558_{\pm.027}$ vs $0.530_{\pm.024}$	0.730	0.864
TabWak	$0.514_{\pm.062}$ vs $0.507_{\pm.061}$	0.511	0.540
TabWak*	$0.268_{\pm.041}$ vs $0.259_{\pm.047}$	0.513	0.558
Tabular Pancakes (latent)	$S = -0.000_{\pm.004}$ (Section 3.3)	0.504	0.502
Tabular Pancakes (data-space)	$S = 0.868_{\pm.041}$, $p_{\mathrm{MC}} \leq 2.7 \times 10^{-3}$	0.501	0.962

We stress-test whether a deployed mark is detectable without the key: TabSyn is trained on Adult and released marked tables are audited against unmarked output, alongside prior schemes through a shared verifier. The prior swap is invisible in released tables (marked-vs-unmarked C2ST 0.497 ± 0.005; same-vs-same 0.499 ± 0.003), and real and marked census records fail Benford and terminal-digit forensics *identically*. Under the same verifier (Table 1), sign-based baselines survive sampler inversion only as small per-bit shifts (certifiable from $\sim 10^4$ rows), Gaussian Shading visibly distorts the table, and every phase-based latent statistic is ≈ 0 (Section 3.3), while data-space Tabular Pancakes certifies from $\sim$60 rows at 3.0% cell RMSE, exactly invariant to unit changes. Adult is hardest: only two columns are continuous enough to mark, and a full-strength mark in two dimensions is learnable by gradient-boosted trees (0.962; logistic regression at chance). The (γ, β) dial trades detection for learnability monotonically (Figure 4), and with 64 markable columns full strength is undetectable ($0.498/0.500$ at $S = 0.949$)—undetectability improves with more continuous columns.

5.3 Detecting partially fabricated tables

We simulate D2: a fraction ρ of a genuine table's rows is replaced by marked generator output. The verifier applies the published codec, computes per-row conformal p-values against a disjoint genuine

sample under the true key, and flags rows by Benjamini–Hochberg. Calibrating under the realized key is crucial: a fresh-key null is valid only on average over keys, with per-seed FDR up to 0.16 on integer-valued data. Figure 3 summarizes all four datasets. Confident *per-row* accusations require several marked columns (power $0.97/0.78$ at ρ=.3 on Magic/Default, zero on the $d \leq 3$ datasets), while ranking suspect rows and estimating the fabricated fraction succeed everywhere.

Table 2: Data-space TABULAR PANCAKES on TabSyn trained on four datasets (5 seeds, m=6000; certified p by MC key rerandomization; localization in Figure 3).

dataset	d_{mark}/d	S (marked)	$p_{\text{MC}} \leq$	S (null)	C2ST lr/HGB	RMSE
Adult	2/6	$0.868 \pm .041$	2.7×10^{-3}	-0.010	.501/.962	3.0%
Magic	10/10	$0.833 \pm .028$	10^{-4}	0.001	.497/.746	8.5%
Default	14/14	$0.561 \pm .090$	10^{-4}	0.019	.508/.993	6.6%
Shoppers	3/10	$0.561 \pm .055$	3.4×10^{-3}	0.116	.504/.920	1.9%

5.4 GENERALIZATION ACROSS DATASETS AND GENERATORS

Across the four datasets (Table 2), marked tables certify at the default operating point, with exact invariance to unit changes and degradation following Proposition 4.3. Released tables are indistinguishable from unmarked ones and forensically match real data. The shallow $(1, 0.3)$ point shows the weakest usable mark strength is dataset-dependent: it certifies on Magic, Adult, and Default, but on Shoppers' three markable columns unmarked coherence already matches the mark's ($S_0 = 0.26$ vs. $S = 0.27$), so the operating point must be chosen per release; the certified test reports this limitation. The six generator families (Table 3) show consistent results: detection certifies, logistic-regression C2ST stays at chance, invariance is exact, and localization with unbiased $\hat{\rho}$ holds throughout.

Table 3: Data-space TABULAR PANCAKES on six generator families trained on Adult (5 seeds).

model	family	d	S (marked)	$p_{\text{MC}} \leq$	C2ST lr/HGB	D2 AUC / $\hat{\rho}$@.3
TabSyn	VAE+diffusion	2	$0.868 \pm .041$	2.7×10^{-3}	.501/.962	0.907 / .302
SMOTE	interpolation	2	$0.915 \pm .022$	10^{-3}	.495/.951	0.916 / .298
TabDDPM	diffusion	2	$0.896 \pm .065$	6×10^{-4}	.500/.948	0.914 / .298
STaSy	score-based	2	$0.788 \pm .113$	1.8×10^{-3}	.507/.957	0.917 / .300
CoDi	co-evol. diff.	5	$0.546 \pm .198$	2×10^{-4}	.507/.891	0.965 / .299
GReaT	LLM (GPT-2)	2	$0.846 \pm .114$	2.6×10^{-3}	.499/.851	0.898 / .296

6 DISCUSSION AND LIMITATIONS

Laundering—retraining on marked output—removes the mark, as it does any bounded-perturbation watermark (Section 4.2); we instead lower-bound removal cost. On diffusion–VAE generators, where latent verification is infeasible (Section 3.3), we mark the released table directly and audit undetectability empirically in place of the hardness reduction. Tables with few continuous columns constrain any numeric-space watermark: higher learnability, weaker certified evidence, and reduced per-row power; the (γ, β) dial manages this trade-off within dataset-dependent limits. Verification reveals the key: once $\mathbf{W}$ is disclosed, data marked under that key can be unmarked, so keys are escrowed and rotated, and zero-knowledge detection is a natural direction for future work.

Conclusion. We leverage the recent Gaussian-pancakes hardness result to build a provenance mechanism for synthetic tabular data. TABULAR PANCAKES samples the generator's latents from hCLWE instead of the Gaussian prior, requires no training, and is provably undetectable without the key. A keyed auditor obtains certified detection, attribution, and row-level localization of partial fabrication, while keyless removal inherits lower bounds from worst-case lattice problems. Both variants—latent and data-space—sustain these guarantees across four datasets and six generator families. The same generative fidelity that defeats statistical forensics can thus certify provenance.

AI USE STATEMENT

We used generative AI tools to aid and polish writing; for the retrieval and discovery of related work; to draft sections of the paper from human-authored outlines, theoretical results and empirical findings. All submitted text has been reviewed and revised by the authors. AI coding assistants were used to support implementation, refactoring and documentation of our codebase, grounded in author-written pseudocode and specifications. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The proposed scheme provides a provenance and tamper-evidence mechanism intended to deter data fabrication. Its principal risk is false-positive accusations; we mitigate this in four ways. First, a document is flagged only against an explicit, assumption-free null model, so the test never presumes data is suspect. Second, we adopt a conservative significance level α, trading detection sensitivity to minimize false alarms. Third, when deploying TABULAR PANCAKES for auditing, all decision thresholds should be fixed in advance, so they cannot be adjusted after inspecting a case. Fourth, we report calibrated p-values as *investigative leads*—a reason to look closer—rather than as verdicts of misconduct. Our method is probabilistic: the presence/absence of a watermark should not be regarded as proof that underlying measurements are genuine (or otherwise). A flag should trigger human review, not be acted on alone.

REPRODUCIBILITY STATEMENT

All constructions are fully specified in the paper: the embedding with exact sampling constants (Section 3, Equation (1), Section C), the verification procedure as pseudocode (Algorithm 1, Equation (2)), and the canonicalization, certification, and localization protocols with all parameter values—codec eligibility rule, Monte-Carlo draw counts, calibration-set sizes, and BH level—in Section G. Complete proofs of every stated result are given in Section A, with hardness statements in Section B. Datasets are standard public benchmarks and all generators are trained through public upstream pipelines with default configurations (Section G); all results report mean±std over seeds with per-seed values retained in the released result files. Our code forks the public TabSyn/TabWak pipeline and is available anonymously at https://anonymous.4open.science/r/tabular-pancakes-DF2F/; a snapshot is also packaged as a zip in the supplementary material. The release includes scripts reproducing every table and figure; watermark keys are excluded and generated by each deployment.

REFERENCES

Scott Aaronson. Watermarking of large language models, August 2023. URL https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17. Video lecture.

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149 – 178, 2023. doi: 10.1214/22-AOS2244. URL https://doi.org/10.1214/22-AOS2244.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188, 2001. doi: 10.1214/aos/1013699998. URL https://doi.org/10.1214/aos/1013699998.

Elisabeth M. Bik, Arturo Casadevall, and Ferric C. Fang. The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3):10.1128/mbio.00809–16, 2016. doi: 10.1128/mbio.00809-16. URL https://journals.asm.org/doi/abs/10.1128/mbio.00809-16.

Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language models are realistic tabular data generators. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=cEygmQNOeI.

Nicholas JL Brown and James AJ Heathers. The grim test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4):363–369, 2017.

Joan Bruna, Oded Regev, Min Jae Song, and Yi Tang. Continuous LWE. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2021, pp. 694–707, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380539. doi: 10.1145/3406325.3451000. URL https://doi.org/10.1145/3406325.3451000.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 831–840. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/bubeck19a.html.

J. B. Carlisle. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*, 72(8):944–952, 2017. doi: https://doi.org/10.1111/anae.13938. URL https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/abs/10.1111/anae.13938.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Statistical query lower bounds for robust estimation of high-dimensional Gaussians and Gaussian mixtures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 73–84. IEEE, 2017.

Andreas Diekmann. Not the first digit! Using Benford's law to detect fraudulent scientific data. *Journal of Applied Statistics*, 34(3):321–329, 2007. doi: 10.1080/02664760601004940. URL https://doi.org/10.1080/02664760601004940.

Daniele Fanelli. How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. *PLOS ONE*, 4(5):1–11, 05 2009. doi: 10.1371/journal.pone.0005738. URL https://doi.org/10.1371/journal.pone.0005738.

Shafi Goldwasser, Michael P Kim, Vinod Vaikuntanathan, and Or Zamir. Planting undetectable backdoors in machine learning models. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 931–942. IEEE, 2022.

Samuel Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=jlhBFm7T2J.

Aparna Gupte, Neekon Vafa, and Vinod Vaikuntanathan. Continuous LWE is as hard as LWE & applications to learning Gaussian mixtures. Cryptology ePrint Archive, Paper 2022/437, 2022. URL https://eprint.iacr.org/2022/437.

James Heathers, Jordan Anaya, Tim van der Zee, and Nick Brown. Recovering data from summary statistics: Sample parameter reconstruction via iterative techniques (sprite), 05 2018.

Seongmin Hong, Kyeonghyun Lee, Suh Yoon Jeon, Hyewon Bae, and Se Young Chun. On exact inversion of dpm-solvers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7069–7078. IEEE, 2024.

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

Jayoung Kim, Chaejeong Lee, and Noseong Park. STaSy: Score-based tabular data synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1mNssCWt_v.

Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling tabular data with diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17564–17579. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kotelnikov23a.html.

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

James E. Mosimann, Claire V. Wiseman, and Ruth E. Edelman. Data fabrication: Can people generate random digits? *Accountability in Research*, 4:31–55, 1995. URL https://api.semanticscholar.org/CorpusID:120643236.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Kareem Shehata, Aashish Kolluri, and Prateek Saxena. CLUE-Mark: Watermarking diffusion models using CLWE. *arXiv preprint arXiv:2411.11434*, 2024.

Zhi Wen Soi, Chaoyi Zhu, Fouad Abiad, Aditya Shankar, Jeroen Galjaard, Huijuan Wang, and Lydia Chen. Timewak: Temporal chained-hashing watermark for time series data. *Advances in Neural Information Processing Systems*, 38:137917–137958, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Min Jae Song, Ilias Zadik, and Joan Bruna. On the cryptographic hardness of learning single periodic neurons. *Advances in Neural Information Processing Systems*, 34:29602–29615, 2021b.

Andrea Taloni, Vincenzo Scorcia, and Giuseppe Giannaccare. Large language model advanced data analysis abuse to create a fake data set in medical research. *JAMA Ophthalmology*, 141 (12):1174–1175, 12 2023. ISSN 2168-6165. doi: 10.1001/jamaophthalmol.2023.5162. URL https://doi.org/10.1001/jamaophthalmol.2023.5162.

Richard van Noorden. More than 10,000 research papers were retracted in 2023 — a new record. *Nature*, 624:479–481, 2023. URL https://api.semanticscholar.org/CorpusID:266223634.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust. *arXiv preprint arXiv:2305.20030*, 2023.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. Steganalysis on digital watermarking: Is your defense truly impervious? *arXiv preprint arXiv:2406.09026*, 2024a.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12162–12171. IEEE, 2024b.

Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*, 2023.

Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-type tabular data synthesis with score-based diffusion in latent space. In *International Conference on Learning Representations*, volume 2024, pp. 52829–52857, 2024.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai. *Advances in Neural Information Processing Systems*, 37: 8643–8672, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/10272bfd0371ef960ec557ed6c866058-Paper-Conference.pdf.

Chaoyi Zhu, Jiayi Tang, Jeroen M. Galjaard, Pin-Yu Chen, Robert Birke, Cornelis Bos, and Lydia Y. Chen. TabWak: A watermark for tabular diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=71pur4y8gs.

A PROOFS

Throughout this appendix, $\gamma' = \sqrt{\gamma^2 + \beta^2}$, $f = \gamma'^2/\gamma$ denotes the detection frequency, and $\sigma_0 = \beta\gamma'/\gamma$ the intrinsic phase jitter in cycle units; Section C derives these constants from the exact sampler. Since $\gamma'/\gamma = 1 + O(\beta^2/\gamma^2)$, the main text writes γ for f and β for σ_0; at the deployed point $(\gamma, \beta) = (2, 0.05)$ the relative difference is 3×10^{-4}.

A.1 PROOF OF THEOREM 4.1

Lemma A.1 (Post-processing)**.** *Let (P_n) and (Q_n) be ensembles such that every probabilistic polynomial-time (PPT) distinguisher has advantage $\mathrm{negl}(n)$, and let g be any PPT map. Then every PPT distinguisher between $g(P_n)$ and $g(Q_n)$ has advantage $\mathrm{negl}(n)$.*

Proof. If $\mathcal{D}$ distinguishes $g(P_n)$ from $g(Q_n)$ with advantage $\varepsilon(n)$, then $\mathcal{D} \circ g$ is PPT and distinguishes P_n from Q_n with the same advantage. □

Proof of Theorem 4.1. Fix a PPT adversary $\mathcal{A}$ that observes $N = \mathrm{poly}(n)$ released rows. Consider three generation processes, which differ only in the latent law and in the categorical randomness:

H_0: $\mathbf{z}_i \overset{\text{iid}}{\sim} \mathcal{N}(0, I_n)$; numeric block $\mathbf{x}_i = G_\theta(\mathbf{z}_i)$; categorical cells drawn by Gumbel-max with fresh i.i.d. uniforms.

H_1: $\mathbf{z}_i \overset{\text{iid}}{\sim} H_{\mathbf{W},\beta,\gamma,\boldsymbol{\delta}}$; the rest as in H_0.

H_2: the marked generator: $\mathbf{z}_i \overset{\text{iid}}{\sim} H_{\mathbf{W},\beta,\gamma,\boldsymbol{\delta}}$ and categorical uniforms $u_c = \mathrm{PRF}_\kappa(\mathrm{sync}_i, c)$.

Gumbel-max sampling with fresh uniform noise draws each categorical cell exactly from its model conditional, so H_0 is precisely the unmarked generator and H_2 precisely the marked one.

H_0 *versus* H_1. The released table is a PPT map of the latent sample set $(\mathbf{z}_1, \ldots, \mathbf{z}_N)$: apply G_θ per row and run the (fresh-randomness) categorical sampler. By Assumption 2.2 in its k-direction, arbitrary-phase form (Section B), the two latent sample sets are computationally indistinguishable for $N = \mathrm{poly}(n)$ rows, and Lemma A.1 transfers this to the released tables: $\left|\Pr[\mathcal{A}(\mathsf{H}_0) = 1] - \Pr[\mathcal{A}(\mathsf{H}_1) = 1]\right| \le \mathrm{negl}(n)$.

H_1 *versus* H_2. First replace PRF_κ by a uniformly random function F; since κ is secret and the adversary never queries the PRF directly, this changes any PPT adversary's acceptance probability by at most the PRF advantage, $\mathrm{negl}(n)$. Now condition on the event D_N that the N sampled rows have pairwise distinct sync vectors: on D_N the values $F(\mathrm{sync}_i, c)$ are jointly i.i.d. uniform, so the conditional output law equals that of H_1. Hence the residual advantage is at most $\Pr[D_N^c] \le \binom{N}{2} q^k$, where $q = \sum_{m \in \mathbb{Z}} \Pr[M = m]^2 \approx (2\sqrt{\pi}\,\gamma')^{-1}$ is the per-direction layer-collision probability of the discrete-Gaussian index. This term is $\mathrm{negl}(n)$ whenever $k = \Omega(\log N / \log \gamma')$.

Combining the two steps proves indistinguishability of marked and unmarked output. For the second claim, any polynomial-time quality metric Q is a PPT map of the released table; hence the distribution of Q differs by $\mathrm{negl}(n)$ between the two worlds, and for bounded Q so does its expectation. In particular no adversarial classifier separates marked from unmarked output—a classifier is a distinguisher—the property probed empirically by the two-sample audits of Section 5. □

Remark A.2 (The categorical coupling at deployed parameters). The per-row conditionals are preserved *exactly* in H_2 regardless of collisions, since Gumbel-max is a faithful sampler for any noise realization drawn once per (sync, c) pair. What collisions induce is a cross-row coupling: rows sharing a sync vector share their Gumbel variables. At deployed parameters ($k \le 4$, $m \approx 6000$) colliding pairs exist, but the coupling is confined to collision classes of the secret index map $\mathbf{z} \mapsto \mathrm{sync}(\mathbf{z})$, which are hidden without $\mathbf{W}$; identifying even one colliding pair amounts to partial recovery of the planted subspace. We therefore treat categorical-channel invisibility at deployed parameters as empirical, covered by the C2ST audits of Section 5, with the theorem giving the asymptotic guarantee for $k = \Omega(\log N)$.

A.2 Proof of Theorem 4.2

Proof of (i): detection power. By the exact mixture representation (Section C), the marked on-axis coordinate along $\mathbf{w}_j$ is $t_{ij} = (M_{ij} + \delta_j)\gamma/\gamma'^2 + (\beta/\gamma')\varepsilon_{ij}$ with M_{ij} an integer and $\varepsilon_{ij} \sim \mathcal{N}(0,1)$, so at the detection frequency $f = \gamma'^2/\gamma$,

$$f\,\langle \hat{\mathbf{z}}_i, \mathbf{w}_j \rangle \;=\; M_{ij} + \delta_j + \xi_{ij}, \qquad \xi_{ij} \sim \mathcal{N}\big(0,\; \sigma_0^2 + f^2(\sigma_{\text{inv}}^2 + \sigma_{\text{atk}}^2)\big) \;=\; \mathcal{N}(0, \beta_{\text{eff}}^2),$$

where $\sigma_{\text{inv}}, \sigma_{\text{atk}}$ are the independent on-axis inversion and attack noises in latent units. Because M_{ij} is an integer it vanishes in $e^{2\pi i f \langle \hat{\mathbf{z}}_i, \mathbf{w}_j \rangle}$, and the Gaussian characteristic function gives, *exactly*,

$$\mathbb{E}\big[e^{2\pi i f \langle \hat{\mathbf{z}}_i, \mathbf{w}_j \rangle}\big] = e^{2\pi i \delta_j}\, e^{-2\pi^2 \beta_{\text{eff}}^2}, \qquad \text{hence} \qquad \mathbb{E}\big[\mathrm{Re}\big(e^{-2\pi i \delta_j} C_j\big)\big] = e^{-2\pi^2 \beta_{\text{eff}}^2} = \mu.$$

No folding correction arises: the exact sampler places the layer index as an integer at frequency f. The main-text display $\beta_{\text{eff}}^2 = \beta^2 + \gamma^2(\sigma_{\text{inv}}^2 + \sigma_{\text{atk}}^2)$ follows with relative error $O(\beta^2/\gamma^2)$. □

Proof of (ii): certified null. Under the null the suspect rows are generated without reference to K; all randomness below is the verifier's key draw, $\mathbf{W}$ Haar and $\delta_j \overset{\text{iid}}{\sim} \mathrm{Unif}[0,1)$. Write $s_{ij} = \cos\big(2\pi(f\langle \hat{\mathbf{z}}_i, \mathbf{w}_j \rangle - \delta_j)\big)$ and $T_j = \frac{1}{m}\sum_i s_{ij}$, so $S = \frac{1}{k}\sum_j T_j$. Three certificates hold, in increasing strength of premise:

(a) Phase certificate, assumption-free. Condition on $\mathbf{W}$ and on the data. The T_j depend on disjoint independent phases δ_j, hence are independent; each has mean exactly zero, since $\int_0^1 \cos(2\pi(\theta - \delta))\, d\delta = 0$ for every θ; and $|T_j| \le 1$. Hoeffding's inequality gives $\Pr[S \ge s] \le e^{-ks^2/2}$ for *any* fixed dataset, degenerate or adversarial.

(b) Row certificate. Suppose additionally the suspect rows are i.i.d. from some law P independent of K. Conditional on K, the per-row scores $s_i = \frac{1}{k}\sum_j s_{ij} \in [-1,1]$ are i.i.d. with conditional mean $\nu(K) = \frac{1}{k}\sum_j \mathrm{Re}\big(e^{-2\pi i \delta_j}\psi_P(\mathbf{w}_j)\big)$, where $\psi_P(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim P} e^{2\pi i f \langle \hat{\mathbf{z}}(\mathbf{x}), \mathbf{w} \rangle}$ is the population characteristic function of the canonicalized data along $\mathbf{w}$. Hoeffding then gives $\Pr[S \ge s] \le e^{-m(s - b_P)^2/2}$ for $s > b_P := \sup_{\mathbf{w}} |\psi_P(\mathbf{w})|$. For whitened, continuously distributed data, $\langle \hat{\mathbf{z}}, \mathbf{w} \rangle$ under Haar $\mathbf{w}$ is approximately $\mathcal{N}(0, \|\hat{\mathbf{z}}\|^2/n)$, so $b_P \approx e^{-2\pi^2 f^2}$—astronomically small at $f \ge 1$. On heavily discretized data b_P can be macroscopic: it is precisely the *lattice coherence* measured as nonzero unmarked S_0 on Shoppers and Default (Section E), and the reason certified verdicts never rely on this certificate alone.

(c) Product regime. When, in addition, the k direction scores of each row are independent—exact on the marked side, where the layer noises ξ_{ij} are independent across j, and accurate under the null for $k \ll n$, where the Haar projections $(\langle \hat{\mathbf{z}}_i, \mathbf{w}_j \rangle)_{j \le k}$ of a fixed row are asymptotically i.i.d. Gaussian—the mk summands s_{ij} are independent, mean-zero, and bounded, giving the displayed bound $p(S) \le e^{-mkS^2/2}$ up to the bias b_P and $O(k^2/n)$ corrections. This is the regime verified against ground truth on the invertible generator in Section 5.1.

Exact Monte-Carlo certificate. Let $K_1, \ldots, K_B$ be fresh keys drawn i.i.d. from the key law. Under the null, K is independent of the data and identically distributed with each K_b, so $\big(S(K), S(K_1), \ldots, S(K_B)\big)$ is exchangeable conditional on the data. Hence $p = \big(1 + \#\{b : S(K_b) \ge S(K)\}\big)/(B+1)$ satisfies $\Pr[p \le \alpha] \le \alpha$ for every α and every finite B (Phipson & Smyth, 2010). This certificate is assumption-free and is the reported verdict on real data.

Data-space verifier. The codec is fit without reference to K—by the published eligibility-and-standardization rule, or refit on the suspect table—so the composite map table $\mapsto \hat{\mathbf{z}}$ is a fixed measurable function of the data. All four certificates above are statements about a fixed function of the data composed with the random frame, and apply verbatim. □

Proof of (iii): sample complexity. In the product regime of (ii)(c), both tails are sub-Gaussian at rate mk: choosing the threshold $\tau = \sqrt{2\ln(1/\alpha)/mk}$ makes the false-positive rate at most α, and $\Pr[S \le \tau] \le e^{-mk(\mu - \tau)^2/2} \le \eta$ whenever $\mu - \tau \ge \sqrt{2\ln(1/\eta)/mk}$. Both hold when

$$mk \;\ge\; \big(\sqrt{2\ln(1/\alpha)} + \sqrt{2\ln(1/\eta)}\big)^2/\mu^2,$$

and the theorem's factor-2 margin absorbs the bias and product-approximation terms of (ii). □

Numerical operating points and payload capacity. At $\beta_{\text{eff}} = 0.25$ ($\mu = e^{-2\pi^2 \cdot 0.0625} \approx 0.29$) and $\alpha = \eta = 10^{-6}$, the bound of (iii) with its margin gives $mk \approx 2600$, i.e. $m \approx 330$ rows at $k = 8$. For the payload, write $C_j = e^{2\pi i \delta_j}(\mu + \epsilon)$ with ϵ the empirical fluctuation; the delta method gives

$$\operatorname{std}(\hat{\delta}_j) \;\approx\; \frac{1}{2\pi\mu}\sqrt{\frac{1-\mu^4}{2m}},$$

using $\operatorname{Var}\big(\sin 2\pi\xi\big) = \frac{1}{2}\big(1 - \mathbb{E}\cos 4\pi\xi\big) = \frac{1}{2}(1-\mu^4)$. Encoding b bits per direction as $\delta_j \in \{0, 2^{-b}, \dots\}$, nearest-codeword decoding errs only if $|\hat{\delta}_j - \delta_j| > 2^{-(b+1)}$, with probability at most $2\bar{\Phi}\big(2^{-(b+1)}/\operatorname{std}(\hat{\delta}_j)\big)$. At the point above ($\mu \approx 0.29$, $m = 400$), $\operatorname{std}(\hat{\delta}_j) \approx 0.019$, supporting 2–3 bits per direction before outer coding; at the strong data-space point ($\mu \approx 0.95$, $m = 6000$), $\operatorname{std}(\hat{\delta}_j) \approx 7 \times 10^{-4}$, supporting 6–7. The deployed 16-bit payload (2 bits over $k = 8$ directions) decodes with $0/80$ bit errors under the full attack suite (Section D).

A.3 Proof of Theorem 4.4

Proof. (a) On-axis noise is necessary and sufficient. By Proposition 4.3, noise orthogonal to $\mathbf{w}_j$ leaves the phase along $\mathbf{w}_j$ unchanged, while on-axis noise of standard deviation $\sigma_\parallel$ degrades $\mu \mapsto \mu\, e^{-2\pi^2\gamma^2\sigma_\parallel^2}$, strictly monotonically. Driving the per-direction signal from μ below a detection floor $\mu_{\min}$ therefore requires $\sigma_\parallel \geq c_0/\gamma$ with $c_0 = \sqrt{\ln(\mu/\mu_{\min})}/(\pi\sqrt{2})$, and this much suffices.

(b) The keyed cost. A key holder applies independent $\mathcal{N}(0, c_0^2/\gamma^2)$ noise along each of the k secret directions and nothing elsewhere: expected squared per-row distortion kc_0^2/γ^2, i.e. ℓ_2 cost $\Theta(\sqrt{k}/\gamma)$.

(c) The keyless cost. Let A be any PPT attack mapping the released table to perturbations $(\mathbf{v}_1, \dots, \mathbf{v}_m)$, and define its on-axis fraction $\Phi = \mathbb{E}\big[\sum_{i,j}\langle \mathbf{v}_i, \mathbf{w}_j\rangle^2\big]/\mathbb{E}\big[\sum_i \|\mathbf{v}_i\|^2\big]$, the share of squared budget landing in the secret subspace. To place c_0/γ of noise on each of k directions, a budget Δ per row must satisfy $\Delta^2\Phi \geq kc_0^2/\gamma^2$. Two facts pin Φ down. First, in the *unmarked* world the generation process never touches $\mathbf{W}$, so A's output is independent of the Haar frame and $\Phi = k/n$ *exactly*, by exchangeability of coordinates under rotation. Second, in the marked world $\Phi \leq k/n\,(1 + o(1))$: an attack achieving $\Phi \geq k/n + 1/\operatorname{poly}(n)$ constitutes an estimator of the planted subspace with noticeable advantage over the isotropic prior (concretely, the top-k eigenspace of $\sum_i \mathbf{v}_i\mathbf{v}_i^\top$ carries noticeable alignment with $\mathbf{W}$), and by the search-to-decision reductions for CLWE (Bruna et al., 2021; Gupte et al., 2022) such an estimator yields a distinguisher contradicting Assumption 2.2. For the strongly aligned regime the reduction is self-contained: given candidate directions $\hat{\mathbf{u}}$ from A, compute the Rayleigh statistic $R(\hat{\mathbf{u}}) = \big|\frac{1}{m}\sum_i e^{2\pi i f\langle \mathbf{y}_i, \hat{\mathbf{u}}\rangle}\big|$ on the oracle samples; alignment $\rho = |\langle \hat{\mathbf{u}}, \mathbf{w}_j\rangle|$ leaves signal $\mu\, e^{-2\pi^2 f^2(1-\rho^2)}$, detectable at $1 - \rho^2 = O(1/f^2)$, with the threshold calibrated on the Gaussian oracle. The general (weak-alignment) case is the step that invokes the cited search-to-decision reductions.

Consequently the keyless budget obeys $\Delta^2 \geq kc_0^2/(\gamma^2\Phi) = nc_0^2/\gamma^2\,(1 - o(1))$, i.e. $\Delta = \Theta(\sqrt{n}/\gamma)$: a $\sqrt{n/k}$ inflation over (b). Figure 2 shows the measured frontier and the $\sqrt{n/k}$ law across k. □

Remark A.3 (Scope). The theorem concerns perturbations below the layer scale. Attacks that operate *at or above* the layer scale $\gtrsim 1/\gamma$—noise-and-denoise, regeneration by a retrained model—erase any bounded-perturbation watermark and are treated separately in Section 4.2, where their utility cost is measured rather than lower-bounded.

A.4 Proofs of Proposition 4.3 and Proposition 4.5

Proof of Proposition 4.3. Independent on-axis attack noise a_{ij} adds to the phase jitter ξ_{ij} of Theorem 4.2(i); for Gaussian a_{ij} with standard deviation $\sigma_{a,\parallel}$, the characteristic functions multiply, giving $\mu' = \mu\, e^{-2\pi^2\gamma^2\sigma_{a,\parallel}^2}$ exactly (general noise enters through $|\mathbb{E}e^{2\pi i f a}|$, with the Gaussian expression as the canonical case). Edits made in table space map to latent perturbations through the encoder; the local Lipschitz factor of that map is measured empirically and converts table-space budgets to σ_a. For a perturbation direction $\mathbf{v}$ chosen without knowledge of $\mathbf{W}$: under the Haar frame, $\mathbb{E}\langle \mathbf{v}, \mathbf{w}_j\rangle^2 = \|\mathbf{v}\|^2/n$ by exchangeability, so with per-coordinate noise level $\sigma_a^2 = \|\mathbf{v}\|^2/n$ we get $\mathbb{E}[\sigma_{a,\parallel}^2] = \sigma_a^2$ per direction, and μ' stays within a constant factor of μ iff $\sigma_a \lesssim c/\gamma$. □

Proof of Proposition 4.5. Mixtures and the detection threshold. If a fraction ρ of rows is marked (mean μ) and the rest genuine (mean 0 up to the bias of Theorem 4.2(ii)), then $\mathbb{E}S = \rho\mu$, and the row certificate detects at level α once $\rho\mu \gtrsim \sqrt{2\ln(1/\alpha)/m}$, i.e. $\rho^* \asymp \sqrt{\log(1/\alpha)/m}\,/\mu$. If instead an adversary *arbitrarily modifies* a ρ fraction of a marked table, trimming the top and bottom ρm row scores before averaging removes the modified rows' influence up to the order statistics of the intact scores; the trimmed mean then concentrates around μ at the rate of $(1-\rho)m$ independent summands, up to constants—the stated effective sample size.

Conformal localization. Let $s_1, \ldots, s_{n_{\text{cal}}}$ be the scores of a disjoint claimed-genuine sample under the *realized* key, and for suspect row i set $p_i = \big(1 + \#\{\ell : s_\ell \geq s_i\}\big)/(n_{\text{cal}} + 1)$. Conditional on the key, a genuine suspect row is exchangeable with the calibration rows, so p_i is super-uniform *conditionally on the key*. Conformal p-values sharing one calibration set satisfy the PRDS property on the set of true nulls, so Benjamini–Hochberg at level q controls $\text{FDR} \leq q\, m_0/m \leq q(1-\rho)$ (Benjamini & Yekutieli, 2001; Bates et al., 2023)—the $q(1-\rho)$ decline visible in Figure 3. By contrast, the pooled fresh-key null is valid only *marginally* over the key draw: a realized key can align with a lattice direction of the genuine data and lift many genuine rows together, inflating the per-key FDR (measured on Default: per-seed FDR up to 0.16 at $q = 0.05$); this is why the reference null is the deployed choice.

Fraction estimation. $\mathbb{E}\bar{S} = \rho\mu_1 + (1-\rho)\mu_0$, so with verifier-calibrated endpoints $\hat{\mu}_0$ (calibration mean) and $\hat{\mu}_1$ (marked reference mean), computed from samples independent of the suspect, $\hat{\rho} = (\bar{S} - \hat{\mu}_0)/(\hat{\mu}_1 - \hat{\mu}_0)$ has $\mathbb{E}[\hat{\rho} \mid \hat{\mu}_0, \hat{\mu}_1] = (\rho\mu_1 + (1-\rho)\mu_0 - \hat{\mu}_0)/(\hat{\mu}_1 - \hat{\mu}_0)$: unbiased at the true endpoints, first-order unbiased in general, with delta-method variance $\big[\text{Var}(\bar{S}) + (1-\rho)^2\text{Var}(\hat{\mu}_0) + \rho^2\text{Var}(\hat{\mu}_1)\big]/(\mu_1 - \mu_0)^2$. Ignoring the null endpoint ($\hat{\mu}_0 \equiv 0$) biases $\hat{\rho}$ upward by $(1-\rho)\mu_0/(\mu_1 - \mu_0)$—the $+0.16$ overshoot measured on Shoppers, whose lattice coherence makes μ_0 macroscopic. *Cell-level re-imputation* perturbs the recovered latent through the encoder and therefore enters as additional encoding noise inside σ_{inv}, degrading μ per Proposition 4.3 rather than breaking validity. □

B HCLWE HARDNESS: PRECISE STATEMENTS AND ATTACK LANDSCAPE

The hardness behind Assumption 2.2 is as follows. For $\gamma \geq 2\sqrt{n}$ and $1/\text{poly}(n) \leq \beta < 1$, distinguishing $\text{poly}(n)$ samples of $H_{\mathbf{w},\beta,\gamma,\delta}$ (over uniformly random $\mathbf{w}$) from $\mathcal{N}(0, I_n)$ with non-negligible advantage is as hard as the worst-case quantum hardness of $\tilde{O}(n/\beta)$-approximate GapSVP/SIVP (Bruna et al., 2021), and as hard as classical LWE (Gupte et al., 2022). The assumption extends to $k = O(n)$ orthonormal secret directions and to arbitrary phases δ (Bruna et al., 2021; Goldwasser et al., 2022), and, under subexponential LWE, holds already for γ polylogarithmic in n (Gupte et al., 2022). The known attacks delimit the parameter range: LLL-type lattice reduction solves the noiseless and exponentially-small-noise regimes (Song et al., 2021b), which is why $\beta \geq 1/\text{poly}(n)$ is required, and a covariance attack succeeds in time $\exp(O(\gamma^2))$ (Bruna et al., 2021). For small γ—the robustness-favoring end of the dial—asymptotic hardness therefore gives way to the sample-constrained concrete-security analysis of Shehata et al. (2024), which bounds the adversary's advantage as a function of the number of released rows; Section 4.3 instantiates it for our parameter choices, with per-key release caps as the operational consequence, and Section F reports the measured attack landscape.

C EXACT HCLWE SAMPLING

We use the projection sampler of Shehata et al. (2024), in the phase-shifted form. Along a secret direction with phase δ, the target one-dimensional marginal is

$$p(t) \;\propto\; e^{-t^2/2} \sum_{m \in \mathbb{Z}} \exp\Big(-\frac{(m + \delta - \gamma t)^2}{2\beta^2} \Big).$$

Completing the square in t—exactly, with no approximation—gives, with $\gamma' = \sqrt{\gamma^2 + \beta^2}$,

$$t^2 + \frac{(\gamma t - (m+\delta))^2}{\beta^2} = \frac{\gamma'^2}{\beta^2}\Big(t - \frac{(m+\delta)\gamma}{\gamma'^2}\Big)^2 + \frac{(m+\delta)^2}{\gamma'^2},$$

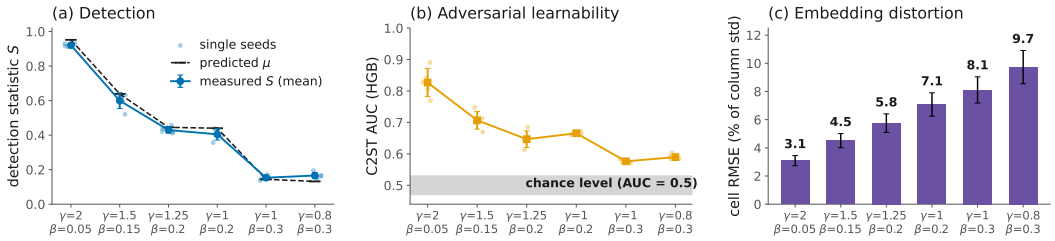


Figure 4: **The (γ, β) dial on real Adult** (5 seeds; operating points ordered strong to weak). **(a)** The measured statistic S follows the predicted μ. **(b)** Learnability by the strongest audit classifier (HGB) falls toward chance as the mark weakens; certified verdicts hold at every point (Table 4).

so p is exactly the mixture

$$\Pr[M = m] \propto e^{-(m+\delta)^2/(2\gamma'^2)}, \qquad t \mid M = m \;\sim\; \mathcal{N}\Big(\frac{(m+\delta)\gamma}{\gamma'^2},\; \frac{\beta^2}{\gamma'^2}\Big) :$$

a shifted discrete-Gaussian layer index and Gaussian noise within the layer. Equivalently, the marked on-axis coordinate is $t_{ij} = (m_{ij} + \delta_j)\gamma/\gamma'^2 + e_{ij}/\gamma'$ with $e_{ij} \sim \mathcal{N}(0, \beta^2)$, the form used in Equation (1); layer spacing is γ/γ'^2 and layer width β/γ'. We sample M by enumeration over $|m + \delta| \le 8\gamma'$, exact to double precision. Two identities drive the paper. *Moment exactness*: $\mathbb{E}[t] = 0$ and $\mathrm{Var}(t) = \gamma^2/\gamma'^4 \cdot \mathrm{Var}(M + \delta) + \beta^2/\gamma'^2 = (\gamma^2 + \beta^2)/\gamma'^2 = 1$, using $\mathrm{Var}(M + \delta) = \gamma'^2$, which holds up to a theta-function correction $O(e^{-2\pi^2\gamma'^2})$—below double precision for $\gamma' \ge 1.5$. The mixture of layer centers and within-layer noise thus reproduces the standard-normal moments, in latent space and in the data-space variant of Section 3.4 alike. *Detection identity*: at the score frequency $f = \gamma'^2/\gamma$,

$$f\, t_{ij} \;=\; m_{ij} + \delta_j + \sigma_0\, \varepsilon_{ij}, \qquad \sigma_0 = \frac{\beta\gamma'}{\gamma}, \;\; \varepsilon_{ij} \sim \mathcal{N}(0, 1),$$

so $ft \bmod 1$ concentrates at δ_j with exactly Gaussian jitter—the premise of Theorem 4.2(i).

D Additional Experimental Results

Controlled-host validation. The σ_{inv} protocol recovers a planted encoder noise of 0.15 as 0.1500 ± 0.0003; μ predicted from the measured residual tracks the round trip within 0.01 on all hosts; rounding to one decimal costs μ: $0.952 \to 0.935 \pm 0.002$, and integer rounding leaves $\mu = 0.17 \pm 0.04$. Under cell noise the empirical μ matches prediction to $\sim$0.005 across 1–20% of column std; row replacement gives $0.762/0.571/0.380/0.190$ at $\rho = .2/.4/.6/.8$ (stds ≤ 0.003); 50 adversarial cell edits of 2 std leave $p < 10^{-6}$; 5% winsorization gives $\mu = 0.161 \pm 0.010$ and heavy MICE re-imputation (15% of cells) $\mu = 0.006 \pm 0.002$. Flow-host localization: AUC ≥ 0.999 for $\rho \in [0.02, 0.8]$, FDR ≤ 0.046, power 0.73 at $\rho = 0.02$; in an imputed-arm case study the treated arm flags at $p_{\mathrm{bonf}} = 1.6 \times 10^{-94}$ with clean arms at $p \approx 1$. Keyless removal to $\mu < 0.05$ costs $\approx 20\%$ per-cell RMSE.

False-accusation audit. Datasets: Breast Cancer, Wine, Diabetes, plus heavy-tailed and correlated stress variants; 100 fresh keys per cell; minimum observed $p = 0.34$. C2ST at m=20,000: marked-vs-unmarked 0.502 ± 0.002 against same-vs-same 0.500 ± 0.002.

The (γ, β) dial on real data. Figure 4 and Table 4 report the full dial sweep on TabSyn–Adult (data-space variant, $m = 6000$, 5 seeds): detection weakens and adversarial learnability falls monotonically together, the measured S tracks the predicted μ throughout, and every point retains a certified verdict ($p_{\mathrm{MC}} \le 2.4 \times 10^{-3}$).

E Measured Limitation Details

(i) *Latent variant on diffusion–VAE hosts*: the recovered statistic is indistinguishable from zero across all 36 dataset–parameter cells (Section 3.3); the data-space variant's undetectability is audited by

Table 4: Dial sweep on TabSyn–Adult (data-space, m=6000, 5 seeds; p_{MC} by key rerandomization; RMSE in units of column std).

(γ, β)	predicted μ	S (marked)	$\log_{10} p_{\mathrm{Hoeff}}$	$p_{\mathrm{MC}} \leq$	C2ST HGB	RMSE
$(2, .05)$	0.952	$0.921 \pm .008$	≤ -300	2.8×10^{-4}	$.827 \pm .045$	3.1%
$(1.5, .15)$	0.639	$0.600 \pm .046$	≤ -300	5.8×10^{-4}	$.707 \pm .028$	4.5%
$(1.25, .2)$	0.445	$0.429 \pm .018$	-241 ± 21	7.6×10^{-4}	$.647 \pm .026$	5.8%
$(1, .2)$	0.440	$0.405 \pm .033$	-214 ± 34	9.6×10^{-4}	$.666 \pm .005$	7.1%
$(1, .3)$	0.144	$0.153 \pm .012$	-31 ± 5	1.2×10^{-3}	$.576 \pm .005$	8.1%
$(0.8, .3)$	0.132	$0.166 \pm .018$	-36 ± 8	2.4×10^{-3}	$.590 \pm .008$	9.7%

Table 5: Covariance-attack AUC against samples released per key (n=32, 30 trials per cell; 0.5 is chance). Bold marks decisive attacks (AUC ≥ 0.9).

	samples per key			
γ	200	500	1000	3000
0.25	**0.92**	**1.00**	**1.00**	**1.00**
0.35	0.50	**1.00**	**1.00**	**1.00**
0.50	0.44	0.54	0.39	0.62
0.75	0.55	0.55	0.57	0.58
1.0	0.64	0.38	0.55	0.53
2.0	0.47	0.54	0.45	0.42

C2ST and inherits only the hardness of recovering $\mathbf{W}$ from whitened rows. (ii) *Narrow numeric blocks*: on Adult (d=2), HGB C2ST reaches 0.962 at full strength while logistic regression remains at chance; the exact MC certificate saturates near 10^{-3}–10^{-4} because a rerandomized key direction occasionally aligns with the true one—the honest ceiling on the evidence a two-column mark can carry—and per-row flagging power is zero. At the shallow point $(\gamma, \beta) = (1, 0.3)$, certification passes on Magic ($p_{\mathrm{MC}} = 10^{-4}$) but fails on Shoppers, where the unmarked lattice coherence ($S_0 = 0.26$) matches the mark's lift ($S = 0.27$). (iii) *Laundering (D3)*: a Gaussian-copula student trained on fully marked output shows no residual signal ($S = 0.0004 \pm 0.0039$, $p \geq 0.87$ every seed), while the null-side control never false-fires. Tail clipping and heavy re-imputation gradually degrade detection (Section 5.1); temporal and MNAR benchmarks remain future work.

F Concrete-security parameter study

Table 5 reports the best known keyless attack—the covariance test of Bruna et al. (2021), implemented after Shehata et al. (2024)—against the dial at $n = 32$, as a function of samples released under one key. The attack is decisive at $\gamma = 0.25$ from 200 samples and at $\gamma = 0.35$ from 500; at $\gamma = 0.5$ it is marginal even at 3,000 samples (succeeding only near 10^4 in extended runs), and $\gamma \geq 0.75$ resists every budget tested. This is the sample-constrained regime of Section 4.3: an adversary observes one release, and per-key exposure caps of 10^3–10^4 rows keep even mid-dial operating points outside the attack's reach, while the asymptotic regime $\gamma \geq 2\sqrt{n}$ needs no cap. Tolerated perturbation follows Proposition 4.3: at $(\gamma, \beta) = (2, 0.05)$ the detector's predictions hold across cell noise of 1–20% of column std (Section D), with the noise budget scaling as c/γ along the dial. Finite precision cuts both ways: rounding of released tables adds attack-like noise (the measured μ costs in Section D), but it also destroys the exponentially-small-noise regime that LLL-type lattice reduction requires (Song et al., 2021b), so the per-sample attack does not apply to released tables at any realistic quantization.

G Additional experimental details

Datasets and generators. The four benchmarks are the standard mixed-type tabular suites: Adult (census income; 6 numeric, 8 categorical columns; 2 eligible for marking), Magic (10 numeric; all eligible), Default (14 numeric, 9 categorical; all numeric eligible), and Shoppers (10 numeric, 8 categorical; 3 eligible), with standard train splits. TabSyn is trained per dataset through the TabWak

reference pipeline (Zhu et al., 2025) with upstream configurations; SMOTE, TabDDPM, STaSy, CoDi, and GReaT are trained on Adult through the upstream TabSyn benchmark suite with default configurations (GReaT fine-tunes a GPT-2–family model). Released tables use $m = 6{,}000$ rows ($8{,}000$ for the baseline comparison; $10{,}000$ per repetition for the fidelity audit).

Canonicalization and schema matching. Preprocessing is frozen and versioned per release and bound into the commitment (Section 3.4). The schema matcher aligns suspect columns to the generator schema by distributional fingerprints (type, support, quantile sketches); residual ambiguity is resolved by testing plausible permutations, with the multiplicity correction folded into the certified p-value. Per-column affine changes are absorbed by robust standardization refit on the suspect table, and this invariance is verified per seed to be exact (bit-identical S under rescaling).

Data-space codec. Eligibility is deterministic and key-independent: a column is markable if its interquartile range is positive, it takes at least 20 distinct values, and no single value carries more than 0.3 of the rows (excluding zero-inflated and spike-dominated columns, whose marking is trivially learnable and destabilizes standardization refits). Eligible columns are standardized by $(x - \text{median})/(\text{IQR}/1.349)$ and whitened by the shrunk covariance $(1-\lambda)S + \lambda I$ with $\lambda = 0.05$ and eigenvalue floor 10^{-4}. Format preservation re-rounds integer-valued columns and clips marked columns to their observed ranges. Because re-running the eligibility rule on a *marked* table can flip boundary columns (Default: 14 eligible pre-mark, 9 post-mark), the marked column list is carried as key material and the verifier refits standardization on exactly those columns.

Certification and localization protocols. Certified verdicts use exact Monte-Carlo key rerandomization: $9{,}999$ draws for the dataset matrix, $4{,}999$ for the generator matrix, 999 elsewhere; the Hoeffding bound is reported as an analytic scale only where its zero-mean premise is verified (Theorem 4.2(ii)). D2 localization uses conformal p-values against a disjoint claimed-genuine calibration sample of up to $4m$ rows scored under the realized key, Benjamini–Hochberg at $q = 0.05$, and the two-endpoint debiased $\hat{\rho}$ (Proposition 4.5). For the categorical channel, the verifier recomputes $u_{c_{\text{obs}}} = \text{PRF}_\kappa(\text{sync}_i, c_{\text{obs}})$ for each observed category and scores $2(u_{c_{\text{obs}}} - \frac{1}{2}) \in [-1, 1]$. On data generated independently of κ, $u_{c_{\text{obs}}}$ is uniform under the PRF assumption—irrespective of the unknown conditionals—so the scores are mean-zero and admit the same Hoeffding and key-rerandomization certificates as the numeric channel; under keyed sampling, Gumbel-max selects categories with large u_c, so the score is stochastically large, with more signal in higher-entropy cells (deterministic cells carry none).

Seeds, error bars, compute. All results are mean$\pm$std over 5 seeds (3 for the conditioning-wall grid), with per-seed values retained in the released result files. Experiments ran on an institutional SLURM cluster, one GPU per job; generator training dominates the budget (GReaT ≈ 6 h on Adult, the diffusion families 1–3 h each), and a full replication of every table and figure costs roughly 30 GPU-hours.