THE TRUNCATION TRAP: SILENT FAILURE OF FIXED-WINDOW DIFFUSION LIKELIHOOD AND INFORMATION ESTIMATORS UNDER REPRESENTATION DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based estimators of likelihood and mutual information evaluate an integral of denoising error over signal-to-noise ratio; in practice the integral runs over a fixed, truncated log-SNR window whose placement and data-scale assumptions are frozen at configuration time. We document a latent failure mode of this family: when the representation's scale drifts, most commonly because a jointly learned embedding grows under cross-entropy training, integrand mass silently migrates out of the window and the estimator retains an arbitrarily small fraction of the full-line integral, with no warning. The mechanism is predictive: the integrand peak sits at $\log \mathrm{SNR} \approx \mathrm{anchor} - \ln s$ for representation scale s. We demonstrate it end to end on a 306M-parameter diffusion language model whose truncated readout reports 0.78 BPC against a converged dense-grid value of 4.27. Auditing the released code of four published systems, we find that each carries an undocumented guard: per-forward embedding normalization (Plaid), default input standardization (MINDE, MMG), and data-adaptive windows (ITD). Disabling MINDE's guard shifts a scale-invariant MI estimate by 4.2 nats; ablating Plaid's guard in a from-scratch retrain leaves the model intact but sends the fixed-window readout to 278 BPC, whereas a guarded positive control reproduces Plaid's reported 1.12 BPC in three independent ways. We distill a defense kit, consisting of an entropy-floor assertion, convergence-verified windows with measured tail bounds, and a rescale stress test, that detects or precludes this failure mode before a number is reported.

1 INTRODUCTION

Information-theoretic diffusion (ITD) estimators (Kong et al., 2023) express the negative log-likelihood of data as an integral of minimum mean-square denoising error over the log-SNR axis, and the same identity underlies a growing family of mutual-information estimators (Franzese et al., 2024). The identities motivate finite-window numerical estimators. We study reuse settings in which a window and its tail corrections are fixed from an initial data-scale snapshot and are not refreshed as the representation changes.

This paper is about what happens when that snapshot assumption breaks. Our motivating observation is stark: a 306M-parameter character-level diffusion language model, evaluated with the standard truncated importance-sampling readout inherited from released ITD code, reported 0.78 bits per character on `text8`, apparently far surpassing autoregressive baselines. The number is an artifact. A deterministic dense-grid evaluation of the *same* integral converges to 4.27 BPC. The discrepancy is not sampling noise, nor a bug in the usual sense: every component behaves as documented; the failure is architectural. Cross-entropy training inflated the learned embedding matrix's per-dimension second moment by $\sim 150\times$; the integrand's peak location tracks $-\ln s$ for this second moment s, and the evaluation window never moved. By the end of training, more than half of the integral's mass lay outside the window, and the closed-form tail bound that was supposed to certify the truncation (computed, as is standard, at initialization scale) understated the escaped mass by two orders of magnitude (§3).

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.

We argue this is not one project's mishap but a *latent failure mode of the estimator family*, and we support the claim with four lines of evidence:

(1) Mechanism with predictive power (§3). The integrand-peak location obeys $\lambda_{\text{peak}} \approx \text{anchor} - \ln s$; within a training run this predicts the truncation-onset step from norm telemetry alone (residuals ≤ 0.66 log-SNR units over a $62\times$ scale range).

(2) Controlled demonstration (§4). Within a single run, the truncated readout dives from 8.04 to 2.10 BPC while the honest dense value stays flat at 4.17; the honest value is stable across 43M/172M/306M model sizes, while the truncated readout is governed solely by escaped mass.

(3) In-the-wild audit (§5). In released code of four published systems, each carries a guard that the corresponding paper does not document: Plaid (Gulrajani & Hashimoto, 2023) L2-normalizes embedding rows each forward pass (its raw norms drift $24\times$ under the same optimizer pressure we measured); MINDE and MMG rest on an unmentioned default `StandardScaler`, without which the official code misreads a scale-invariant MI by up to 4.97 nats; the original ITD image estimator recomputes its window from the eval data at eval time, the guard our LM setup unknowingly discarded; and MMG's released code additionally omits the adaptive window scheme its paper describes.

(4) Positive control and defense (§6). With guards in place the framework delivers: our convergence-verified honest protocol yields 1.08–1.10 BPC through three readouts, close to Plaid's reported 1.12, and a corrected recipe with normalization, adaptive windows, and endpoint terms yields a stable honest 4.12 BPC where the artifact is structurally impossible.

Along the way we isolate a second, opposing trap: floating-point error floors amplified by e^λ inflate over-extended integrals (§3.4; three codebases), so wider windows are not automatically safer.

No published headline number is invalidated by our audit (§7); the risk is prospective: a reimplementation from the papers alone can silently omit an implementation-level guard, and an extension to learned, drifting representations can discard a protection whose role was never stated. (An independently discovered chain-rule omission in a diffusion-LM likelihood pipeline (Yang et al., 2026) deflates through an orthogonal mechanism: silent deflation in this family is a recurring hazard, §7.)

2 Background: diffusion likelihoods as SNR integrals

The identity. Let $x \in \mathbb{R}^D$ be a data representation and consider the Gaussian channel $z_\lambda = \sqrt{\alpha_\lambda}\, x + \sqrt{1-\alpha_\lambda}\,\epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$ and log-SNR $\lambda = \log \frac{\alpha_\lambda}{1-\alpha_\lambda}$. The I-MMSE relation (Kong et al., 2023) ties mutual information to denoising error: $\frac{d}{d\lambda} I(x; z_\lambda) = \frac{1}{2} e^\lambda \,\mathrm{mmse}(\lambda)$, where $\mathrm{mmse}(\lambda) = \mathbb{E}\,\|x - \mathbb{E}[x \mid z_\lambda]\|^2$. For a diffusion model, replacing the optimal denoiser by the learned $\hat{x}_\theta$ gives the same integral of the model's own (suboptimal) denoising error,

$$\mathcal{L}(\theta) \;=\; \frac{1}{2}\int_{-\infty}^{\infty} e^\lambda\, \mathbb{E}\left\|x - \hat{x}_\theta(z_\lambda, \lambda)\right\|^2 d\lambda \;+\; \text{const}, \tag{1}$$

which upper-bounds the data entropy. When the data are *discrete* tokens s embedded as $x = E(s)$, the constant vanishes: $I(s; z_\lambda) \to H(s)$ as $\lambda \to \infty$, so the integral of the denoising-error integrand $g(\lambda) = \frac{1}{2} e^\lambda\, \mathbb{E}\|x - \hat{x}\|^2$ over the whole real line is a valid log-loss for the token sequence in the one-sided sense made precise in App. A: it never falls below the sequence entropy, and a truncated evaluation forfeits exactly this guarantee. Two properties matter for everything that follows: the identity holds only for the integral over *all* of $\mathbb{R}$, and the integrand is a bump whose *location* on the λ axis depends on the scale of x.

The practice. Numerical implementations use finite integration ranges. The released ITD code draws λ from a truncated logistic (window $[\mu - c\sigma,\ \mu + c\sigma]$, default $c{=}4$) and importance-weights the integrand; mass outside the window is covered by closed-form tail bounds computed from data-scale statistics. In our LM reuse, both the window and the tail-bound constants were frozen at initialization; the released image pipeline refreshes its data-dependent window (§5.2). The MINDE family (Franzese et al., 2024) reaches the same structure by a different route: fixed VP-SDE constants define an implicit static log-SNR window of roughly $[-10, +9]$ (App. B), with the truncation error dismissed by a "sufficiently large T" argument. For a frozen-window reuse, three pieces of

bookkeeping must remain consistent: (1) **the window** (where the estimator looks); (2) **the data scale** (where the integrand actually is); and (3) **the tail terms** (the certificate that what it does not look at is negligible). The bookkeeping is sound when the three agree. The trap arms when leg (2) moves after legs (1) and (3) were frozen. By contrast, VLB-style estimators (Kingma et al., 2021; Gulrajani & Hashimoto, 2023) *measure* their endpoint terms (reconstruction likelihood, prior KL) on the live model, so escaped mass inflates a measured term instead of vanishing, and the bound stays valid. The direct I-MMSE integral has no such self-accounting term; nothing in it can notice that mass has left the window.

3 THE TRUNCATION TRAP

3.1 REPRESENTATION DRIFT IS THE DEFAULT, NOT A CORNER CASE

When the embedding E is trained jointly with the denoiser under a cross-entropy objective (the standard recipe for continuous diffusion language models), E appears on both sides of the channel: it is the signal being corrupted *and* a parameter of the loss. For a fixed noise level, scaling E by k multiplies the effective SNR of the token-identity signal by k^2, and the CE loss at every sampled λ improves monotonically with class separation. With no weight decay and no normalization on E (our base recipe, inherited defaults), nothing opposes this incentive (the implicit bias toward growing logit norms (Soudry et al., 2018) pushes the same way), so growth is slow, persistent, and unbounded. Empirically, every unconstrained run we trained (43M and 306M parameters, embedding-init σ_E from 0.02 to 1.0) grew its embedding per-dimension second moment from 1.0 at initialization to 145–186 by 100k steps, with the growth accumulating smoothly over training (Fig. 1b). The phenomenon is documented in the text-diffusion literature as a training pathology (CDCD (Dieleman et al., 2022) normalizes embeddings to prevent it; Difformer (Gao et al., 2024) reports collapse without constraints); undocumented is its consequence for likelihood estimators.

3.2 THE PEAK-LOCATION LAW

Where does the integrand $g(\lambda)$ sit? For embedded tokens, the denoising error transitions from prior variance (token unresolvable) to $\sim$0 (token resolved) around the log-SNR at which typical pairwise embedding separations become distinguishable through the noise: $e^{\lambda}\,\delta^2 \gtrsim O(1)$, where δ^2 scales with the per-dimension second moment s of the embedding (App. A). Holding the embedding geometry fixed, the transition, and with it the integrand peak, obeys

$$\lambda_{\text{peak}}(s) \;=\; \lambda_{\text{peak}}(s_0) \;-\; \ln(s/s_0). \tag{2}$$

Growth of s translates the bump *left*, toward low SNR, at a predictable rate. We verify Eq. 2 two ways. Across five independently trained 43M models spanning a $178\times$ range of final embedding scale, measured peaks match the prediction with max residual 0.12 log-SNR units (grid resolution 0.25). Along a single training run (architecture, geometry, and denoiser co-evolving; 19 checkpoints, $62\times$ scale range), the residual never exceeds 0.66 (App. A, Table 3). The law has predictive force: given only norm telemetry, one can forecast the training step at which the integrand escapes any given window; for the 43M telemetry run, forecast and measured gap-opening coincide at step 45k (Fig. 1a).

Scale convention and transfer. The scale s in Eq. 2 is a second moment, not an embedding amplitude: multiplying every row by k multiplies s by k^2, hence shifts the predicted peak by $-2\ln k$. The anchor is model-dependent. Predicting 306M peaks from a 43M anchor leaves an approximately -1.1 offset, confounded by width and denoiser quality; the within-architecture fits, not universal anchor transfer, support the mechanism (App. A, Table 3).

3.3 WHY THE CERTIFICATE FAILS TOO

The truncation would be harmless if the tail bounds caught the escaped mass. They cannot, because they are scale snapshots. Our inherited protocol bounds the omitted low-SNR mass by $\text{tail}_{\text{low}} \le \frac{2R^2}{\ln 2}\, e^{\mu - c\sigma}$ BPC with $R^2 = \max_v \|E_v\|^2$. This is correct as a bound, but it was evaluated at the initialization scale ($\sigma_E = 1 \Rightarrow R^2/d \approx 1$), where it gives a reassuring 0.134 BPC. Recomputed

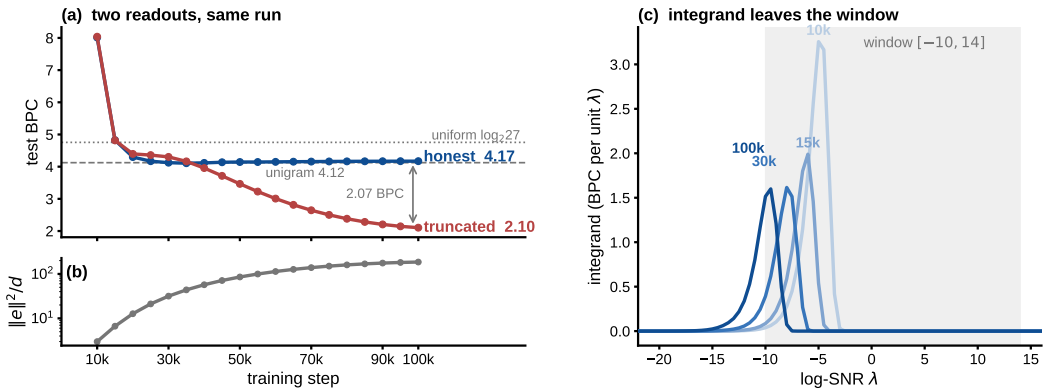


Figure 1: **The artifact opens within a single training run** (43M, original recipe, `text8`). **(a)** The truncated readout (window $[-10, 14]$) falls to 2.10 BPC while the honest dense-grid value stays flat near 4.17; the gap opens at step 45k, forecast from norm telemetry by Eq. 2. **(b)** Embedding norm growth. **(c)** The integrand translates left by $-\ln s$, out of the fixed window (shaded); by 100k steps only 49.6% of its mass remains inside.

at the trained scale ($R^2/d = 181$ at 100k steps) it is a vacuous 24.3 BPC, and the actually omitted mass, measured by dense integration, is 3.404 BPC of a 4.27 BPC total (App. A.3). The reported 0.78 was thereby *certified* by a bound whose constant was stale by two orders of magnitude. The general lesson: any error bound whose constants depend on learned quantities must be recomputed from the trained model at evaluation time, and a bound that certifies a surprising result deserves a direct check: dense integration of the omitted region cost $\sim$30 GPU-minutes and falsified the certificate immediately.

3.4 The opposing trap: precision floors amplified by e^{λ}

The obvious fix, integrating a much wider window empirically, meets a second, opposing artifact. At high log-SNR the true denoising error decays like $e^{-\lambda}$, but any floating-point floor on the measured MSE (fp32 rounding of near-zero errors) is *multiplied by* e^{λ} in the integrand, so over-extending the empirical integral into the far right tail inflates the estimate without bound. We observed this mechanism in our LM pipeline (evaluating in bf16 alone shifts the converged dense value by $+0.11$ BPC; final integration should run in fp32), reproduced it in the released ITD image code, and it powers the ablation blow-up of §6; App. G details all three instances. Wider is not safer: the correct treatment is an empirical integral over a *convergence-verified* window plus analytic tails, which is what the original ITD configuration does, and what our defense kit enforces.

4 Controlled demonstration

Setup. We train character-level diffusion LMs on `text8` (27-token vocabulary, 90/5/5M split) with a DiT backbone over a jointly learned embedding, at three scales (43M/172M/306M parameters; App. F). The training and evaluation noise window is the truncated logistic inherited from released ITD code with hard-coded $(\mu, \sigma, c) = (2, 3, 4)$, i.e. $\lambda \in [-10, 14]$: the configuration a from-the-paper reimplementation would naturally adopt (§5 traces this lineage). "Honest" values are deterministic trapezoid integrals of $g(\lambda)$ on a dense fp32 grid over $[-22, 22]$, extended until the edge integrand vanishes (convergence checked; App. F).

Dense-grid convergence protocol. Deterministic trapezoid on a 0.25-spaced log-SNR grid, fp32 accumulation, EMA weights, fixed noise seed per grid point; the grid is extended until the edge integrand is negligible relative to the total (headline model: left-edge integrand 1.3×10^{-7} BPC/unit on the $[-28, 22]$ grid, contributing 0.0024 BPC beyond $[-22, 22]$). A readout is reported only with its convergence evidence (edge integrands and in-window mass fraction).

Table 1: Left: the honest value is stable across model scales; the truncated readout is governed by escaped mass. Right: 2×2 factorization at 43M/100k steps; the trap requires *both* free norms and a truncated window.

	43M	172M	306M
dense BPC	4.30	4.38	4.27
truncated BPC	4.40	—	0.78
mass retained	96%	96%	20%

dense BPC	truncated	full
free norm	4.158	4.095
fixed norm	4.104	4.099

The artifact, end to end. The 306M model trained 100k steps reports 0.78 BPC under the truncated IS readout: below every autoregressive baseline and, more to the point, below the `text8` unigram reference of 4.12 BPC. The honest integral of the same model is 4.27 BPC: the window retains only 0.861 BPC of mass (0.78's scale), while 3.404 BPC sits below $\lambda = -10$, exactly where norm growth pushed the token-resolution transition. Figure 1 shows the within-run anatomy on the matched 43M telemetry run: truncated and honest readouts agree early ($\sim$98% in-window at 20k steps), then diverge precisely when the norm proxy crosses the predicted threshold; the readout decays smoothly toward zero with no visible symptom of failure.

The silence deserves emphasis, because the terminal 0.78 is the *least* dangerous form of the artifact: a number below every known compressor invites suspicion on its own. The trajectory that produces it does not. The same mechanism leaves the 43M run's truncated readout at 2.10 BPC, a value indistinguishable from legitimate progress (our masked-diffusion baseline reads 2.03), reached while the honest value sits at 4.17, and flagged by no plausibility heuristic. Every intermediate deflation between "correct" and "absurd" is a publishable-looking number. The MI side is worse still: drifts of -0.8 to -4.2 nats (§5.1) have no best-known-compressor reference to trip at any magnitude.

Stability of the honest value. Across the three model scales under the original recipe, the honest integral is flat: 4.30 / 4.38 / 4.27 BPC (Table 1, left; checkpoint horizons differ across scales, a $\sim$0.13-BPC slack bounded by the within-run trace, §8). Extending across seven runs varying model size, σ_E init, sampler center, and training clip, the honest total stays within 4.10–4.38 (App. F); truncated readouts of the same checkpoints meanwhile range from 8.04 down to 0.78, tracking only the escaped fraction. A 2×2 factorization at 100k steps (Table 1, right) isolates the conjunction: with norms fixed *or* the window widened to cover the integrand, honest and windowed values agree to within 0.06 BPC; the 0.78-style collapse requires free norms *and* a fixed truncated window together.

The 2×2 factorization (100k steps, 43M). Free vs fixed norm $\times$ truncated $[-10, 14]$ vs full $[-22, 22]$ window, honest dense fp32 values: free/truncated 4.158 (49.8% of mass in its own window), fixed/truncated 4.104 (99.7%), fixed/full 4.099 (100%), free/full 4.095 (100% in its window; 46.5% in the original one, final $\|e\|^2/d = 214$). The honest value is flat across all four cells; only the free-norm cells have escaped mass for a truncated *readout* to lose. Notably, widening only the *training* window (free/full cell) does not stop norm growth; it merely keeps the eval window ahead of the migrating integrand.

What the honest number means. For context: matched-compute autoregressive and masked-diffusion (Sahoo et al., 2024) baselines on our setup reach 1.34 and 2.03 BPC respectively, and the unigram reference is 4.12. The honest full-line value, 4.27 BPC, is worse than even the unigram reference, a modeling outcome we do not defend and do not need to: the paper's subject is that the readout reported 0.78 for a checkpoint whose full-line value does not even beat a unigram, with no warning. The positive control (§6) shows the estimator family reads out $\sim$1.1 BPC faithfully on a model that achieves it: model quality and estimator honesty are fully disentangled.

5 AUDITING RELEASED ESTIMATORS

We audited the released code and checkpoints of four published systems spanning the estimator family's design space (Table 2), asking of each: where is the window, what pins the data scale, and what happens when the pin is removed? Combined with a 30-paper structural survey (App. E), the

Table 2: The audit at a glance. Each system carries a guard that is live in code but stated nowhere in its paper; the last column is the measured effect of removing or bypassing it (§5.1–§5.4, §6).

system	window	guard in code	exposed result
MINDE	VP-SDE, implicit	default input standardization	-4.2 nats (true MI 10)
ITD (images)	logistic, clip 4	adaptive (μ, σ); pixel scale	$+0.79$ bpd (rescale test)
Plaid	VLB, learned endpoints	per-forward L2 row norm	278 BPC (ablation + transplant)
MMG	logistic $(2, 3)$, clip 4	default input standardization	-4.97 nats (true MI 9.9)

answer is uniform: *no published headline number moves, and every system carries a guard its paper does not document* (Table 2).

5.1 MINDE: EXPOSED ONCE ITS UNDOCUMENTED GUARD IS DISABLED

Mutual information is invariant to rescaling ($\mathrm{MI}(sX; sY) = \mathrm{MI}(X, Y)$), so any drift of the estimate under input rescaling is pure estimator error, an ideal cleanroom for the trap. MINDE's identity is integrated over the fixed VP-SDE window above. Its official code, however, applies a `StandardScaler` to both variables by default (stated nowhere in the paper), silently undoing any global rescale before the fixed window sees it. We ran the official code on benchmark Gaussian tasks with known ground truth (Czyż et al., 2023) at input scales $s \in \{0.01, 0.1, 1, 10, 100\}$, three seeds, in two modes. **Guarded** (default): estimates are flat at ground truth for every s (max drift 0.18 nats). **Exposed** (`-preprocessing none`, a supported code path): on the d=20/MI=10 task the estimate reads 5.80$\pm$0.01 nats at s=0.01 (drift -4.20) and 7.04$\pm$0.04 at s=100 (drift -2.96). The two failure edges are mechanistically distinct and both predicted by code audit before measurement: at $s \ll 1$ the integrand escapes past the high-SNR cutoff t_ϵ, where the model itself is fine (a dense t-grid on the same trained score network recovers $\sim$10.2) and the official importance sampler simply cannot see the mass; at $s \gg 1$ the mass escapes past T=1 and is genuinely unrecoverable. An exact integral of the true Gaussian integrand restricted to the official window predicts the measured values within $\sim$1 nat at every scale (App. B), confirming window truncation as the dominant bias. The paper's MINDE-σ variant is more fragile still: its reference score hard-codes unit scale, and at s=100 it reads 36–43 nats for a true MI of 10. The flag is our cleanroom knob, but the exposed configuration is not hypothetical: it is what any reimplementation from the paper would build (the scaler appears in no equation, algorithm, or experimental description), and the MINDE-σ variant is exposed exactly as described. Even the guard is partial: a single global standardization pins only the *average* scale, so mixtures of lengthscales or heavy tails can keep mass near the window edges in guarded mode (App. B).

5.2 ITD ON IMAGES: SAFE BECAUSE PINNED, AND THE LINEAGE OF OUR TRAP

The original ITD image estimator (Kong et al., 2023) is the ancestor of the truncated-logistic readout; we reproduced its shipped evaluation artifacts to four decimal places on CIFAR-10, then swept the integration window with a dense deterministic grid (App. C). The published numbers are essentially window-safe: the continuous estimate moves -0.0084 bpd from clip 4 to 5 and is then flat to clip 8; 99.54% of the converged gap mass lies inside the default window; the dequantized and discrete estimates are converged at the default clip (the $+0.08$ bpd rise at clips 7–8 is §3.4's fp32-floor artifact in an independent codebase). Crucially, this safety is *conditional*, and the conditions are exactly the guards our LM recipe lost: the data scale is pinned by $[-1, 1]$ pixel preprocessing, and the window is *data-adaptive*: (μ, σ) are recomputed from the evaluation data's covariance spectrum at eval time, so the integrand cannot silently walk away from the window. Our LM setup inherited this code's truncated logistic and clip=4 but hard-coded $(\mu, \sigma) = (2, 3)$ at init-time values over a representation that then grew $150\times$: two guards dropped in one reuse step. The step is the default one, not an idiosyncratic blunder: the adaptive (μ, σ) computation is an undocumented internal of the image pipeline, while what the interface exports is the sampler shape and its constants; Info-

Decomp (Kong et al., 2024) (manual constants, clip 3) and MMG (Yu et al., 2025) (repository defaults, verbatim) made the same move and survive only because their representations are frozen and standardized (§5.4); ours was the same reuse in the one setting, a learned representation, where the dropped guard was load-bearing. A rescale-without-refresh stress test on the original pipeline confirms the bookkeeping (not the model) is what fails when the pin is detached: scaling inputs by $s{=}2$ while freezing the snapshot statistics collapses the estimator's data-dependent term to zero (deviation $+0.794$ bpd), a silent, window-independent bias no widening can recover (App. C).

5.3 PLAID: THE DRIFT IS REAL, THE GUARD IS UNDOCUMENTED

Plaid (Gulrajani & Hashimoto, 2023) trains learned embeddings jointly with a VLB objective: on paper, the exact configuration that armed our trap; its text states the embeddings are optimized "without additional constraints." The released code tells a different story: `EmbeddingMatrix.forward()` L2-normalizes every embedding row on every forward pass. Auditing the released 1B checkpoint, we find the *raw* (stored) embedding norms drifted $24\times$ during training (the same optimizer pressure we measured, alive inside a published system), while the per-forward projection neutralizes it before the channel ever sees the drift. Plaid's two remaining legs are also closed: its schedule endpoints are learned parameters, and its VLB measures reconstruction and prior-KL endpoint terms on the live model. All three legs of the trap are broken; the scale-pinning one by a guard whose existence the paper's own text denies.

5.4 MMG: THE DESCRIBED GUARD IS ABSENT FROM THE CODE

MMG (Yu et al., 2025) is Plaid's mirror image. Its paper describes a two-stage adaptive scheme (fit the window's (μ, σ) from a preliminary model's *measured* conditional-MMSE curve, then re-train), which is exactly the measured-scale calibration that neutralizes the trap, though motivated there purely as efficiency. The released repositories contain no implementation of it: the only window in the experiment code is the inherited fixed `logistic_integrate` with defaults $(2, 3)$, clip 4 (a constructor flag for updating the location is stored but never read; App. H), and the live protection is instead a default input standardization stated nowhere in the paper: the third undocumented guard in as many codebases. With standardization disabled, the as-released estimator drifts under pure input rescaling by up to -4.97 nats on a true MI of 9.90 ($s{=}0.01$), matching the analytic window-restricted integral (4.79). A faithful reimplementation of the adaptive rule tracks the predicted $-2 \ln s$ window migration at moderate scales, but its width rule under-covers the high-MI task even at native scale (5.84 vs 9.90 at $s{=}1$, also predicted analytically) and its location fit degenerates at extreme scales (fitted $\mu \approx -34$, estimate zero). The two paper–code gaps face opposite directions: Plaid's code contains a guard its paper denies; MMG's paper describes a guard its code lacks; each defeats a faithful reader of the other artifact.

5.5 THE FIELD-LEVEL PICTURE

A structural survey of 30 published noise-integral likelihood and MI estimators (App. E) completes the pattern: **0** are exposed as published; **7** in the ITD/MINDE lineage carry a structurally exposed *scheme* (fixed window and/or snapshot-scale tail argument) that binds only under reuse with a drifting representation; every learned-representation system that reports a noise-integral likelihood pins its scale in code or self-accounts via measured VLB endpoint terms. No surveyed paper documents the failure mode. The protection is real, but its protective role is nowhere stated: implementation-level guards can be silently omitted, and even documented mechanisms can be dropped in reuse with no one noticing what was lost. That is how our recipe, a faithful implementation of the *described* method, walked into it.

6 POSITIVE CONTROL AND THE GUARDED PROTOCOL

The framework delivers when guarded. A skeptic could read §4 as evidence that honest evaluation of continuous-embedding diffusion LMs is impossible, or that our honest protocol is miscalibrated; the released Plaid 1B model refutes both. Applying our convergence-verified honest readout (dense-grid diffusion integral plus measured endpoint terms, no component reused from Plaid's own evaluator), we obtain three `text8` readouts close to its published result: Plaid's own evaluator gives

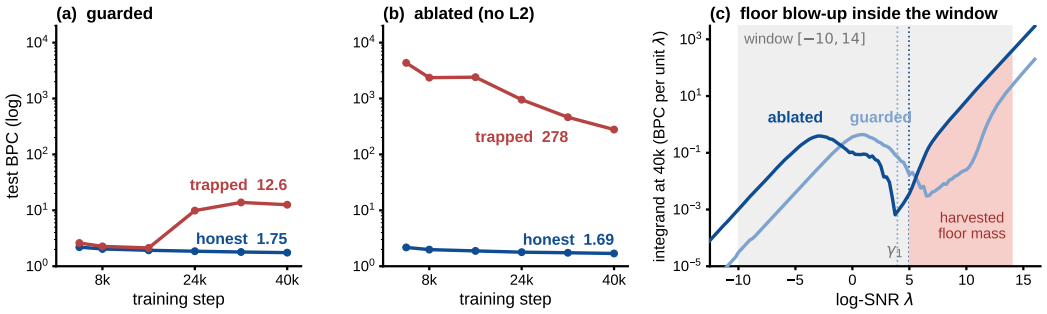


Figure 2: **Ablating Plaid's undocumented guard: the model stays good, the fixed-window readout does not** (28.4M Plaid-small recipe, character-level `text8`). **(a)** Guarded: the transplanted readout drifts to 12.6 BPC late in training as the error floor rises. **(b)** Ablated: the model trains slightly *better* (honest 1.69 vs 1.75 BPC), but the transplanted readout is wrong by two orders of magnitude throughout. **(c)** Integrands at 40k steps: both learned windows end at $\gamma_1 \approx 4$–5 (dotted); the transplanted window extends to $+14$ and harvests $e^{\lambda}\times$error-floor mass (red fill, §3.4).

1.1016 BPC, our deterministic reassembly (reconstruction CE + window integral + prior KL) gives 1.1044, and a floor-aware pure integral gives 1.0830, against a reported 1.12 (`enwik8`: 1.156 vs 1.18). The honest protocol neither inflates nor deflates a genuinely good model; the 0.78→4.27 correction in §4 was recipe failure, not framework failure. These readouts share one checkpoint.

Ablating the guard on the competitive recipe. If Plaid's undocumented L2 projection were load-bearing for *modeling*, the trap would remain a corner case; if it protects only the *readout*, then a reimplementation that drops it trains a good model and still reports a wrong number. We tested this directly: two 28.4M-parameter models with Plaid's default small recipe, trained from scratch on character-level `text8`, identical except that the ablated variant removes the per-forward row normalization (Fig. 2). The guard is not what makes the model good: honest readouts (dense grid over the learned window plus measured endpoint terms) give 1.75 (guarded) and 1.69 BPC (ablated) at 40k steps. But transplanting our LM recipe's readout verbatim (fixed $[-10, 14]$ window, no endpoint terms) misreads the ablated model by two orders of magnitude, 278 BPC, this time *upward*: with the projection gone the representation scale floats $\sim 15\times$ higher, its error floor scales with it, and the window's upper edge sits ~ 9 log-SNR units above the learned endpoint, so the readout harvests $e^{\lambda}\times$floor mass (Fig. 2c; §3.4) with the edge integrand still rising; the convergence check fails loudly (measured floors and per-checkpoint table in App. D). Even the guarded model is over-read (12.6 vs 1.75) once its own floor rises late in training. Together with §4, this brackets the failure mode: the same frozen-window readout *deflates* our LM ($150\times$ norm growth, mass escaping below the window) and *explodes* on the Plaid-small recipe (floor mass harvested above it). It is wrong in whichever direction the geometry dictates, and silent only in the first case, while the honest protocol's convergence and floor checks pass on every model here.

Ablation attribution. At $\lambda \geq 14$, error floors of 4.5×10^{-4} (ablated) versus 3.1×10^{-5} (guarded) track the 15-fold second-moment ratio. Less than 10^{-3} nats escapes below the window, confirming high-edge inflation (App. D).

Defense kit. Each component below is cheap, and each would have caught the artifact independently. (i) *Entropy-floor assertion:* feed i.i.d. uniform tokens through the full pipeline; any estimate below $\log_2 |V|$ bits/token proves mass loss, with zero reference implementations needed. Our pipeline reads 2.43 BPC on uniform data against a 4.755 floor, an instant failure; the dense readout (6.06) and Plaid (App. D) pass. (ii) *Convergence-verified adaptive window:* place the window from the *measured* integrand of the trained model, widen until edge integrands vanish, run the final integral in fp32, and add analytic tails evaluated at measured constants (§3.3). (iii) *Rescale stress test:* multiply inputs by s and assert the estimate shifts by exactly the analytic amount (NLL) or not at all (MI); §5.1 and §5.2 are this test at work. (iv) *Norm telemetry:* log representation scale per eval; Eq. 2 converts the trace into an escape forecast (full practitioner checklist: App. F).

Validation. Retraining the 43M model with the guarded recipe (fixed-norm embedding, adaptive full-coverage window, measured tails, fp32 integration) yields honest 4.12 BPC with 100% of integrand mass in-window, guardrails passing at all 50 evals, and importance-sampling and dense readouts in agreement (4.228 ± 0.130 vs. 4.235 in matched precision). The scale-drift failure mode is structurally precluded, at zero cost in CE (App. F). The residual gap to the AR baseline (1.34) is no longer attributable to the estimator: a modeling problem, visible as such.

7 Related work

Diffusion likelihood and MI estimation. The I-MMSE identity enters likelihood estimation through ITD (Kong et al., 2023) and extends to MI and information decompositions, Poisson channels, and O-information (Franzese et al., 2024; Kong et al., 2024; Yu et al., 2025; Bhattacharya et al., 2025; Bounoua et al., 2024). Warm-start ITD (Shen et al., 2025) argues truncation error is negligible in its fixed-scale image setting, consistent with our audit: the danger is not truncation per se but its conjunction with drift. VLB-based estimators (Kingma et al., 2021; Gulrajani & Hashimoto, 2023; Yang et al., 2026; Li et al., 2022) form the self-accounting contrast class (§2).

Silent evaluation failures in diffusion LMs. RePlaid (Yang et al., 2026) reports an independent silent-deflation incident (a dropped chain-rule term in a Hutchinson-trace NELBO) and states Plaid's unit-norm embeddings as fact without flagging the paper-vs-code discrepancy we document in §5.3. On the sampling side, fp32 Gumbel truncation silently deflates generative perplexity (Zheng et al., 2025); our fp32-floor artifact (§3.4) is the likelihood-side sibling, and both echo broader critiques of likelihood-adjacent evaluation (Theis et al., 2016; Schaeffer et al., 2023).

MI estimator reliability. Benchmark-driven critiques (Czyż et al., 2023) probe MI estimators with distributional structure; concurrent work (Foresti et al., 2026) studies ϵ-floor "perturbed-MI" bias; ours is orthogonal to both: a scale-invariance stress test tied to a code-level mechanism, with the bias predicted analytically before measurement.

Norm growth as a training pathology. Embedding-scale instability in text diffusion is documented as a *training* problem, with normalization as the fix (Dieleman et al., 2022; Gao et al., 2024; Ye et al., 2024); the implicit-bias literature explains the pressure (Soudry et al., 2018). Our contribution is the quantified consequence for *evaluation*: the $-\ln s$ peak-migration law, its interaction with frozen windows and snapshot certificates, and the resulting silent collapse of the reported value.

8 Discussion and limitations

Scope of the claim. We do not invalidate any published number; the audit actively confirms three of them (§5). The claim is prospective: the guards, or their protective role, are undocumented, so the two channels by which methods spread (reimplementation from papers, extension to learned representations) can silently shed them. Our own headline model is the existence proof: the failure occurs in practice, at scale, under a faithful implementation of the described method.

Limitations. (i) The rescale-without-refresh stress test on the ITD image pipeline puts the model out of distribution; we isolate the bookkeeping-only effects and do not interpret the OOD component (App. C). (ii) The honest BPC of our unguarded models mixes checkpoint horizons across scales (§4; slack ≤ 0.13 BPC by the within-run trace). (iii) Our LM experiments are character-level and `text8`-scale; the mechanism depends only on (learned scale) $\times$ (frozen window), neither of which is dataset-specific, but we have not demonstrated the artifact in a subword-vocabulary system. (iv) The survey classifies structural exposure from papers and released code; UNCLEAR entries are marked, not resolved (App. E).

Outlook. The estimator family is worth defending: the identity is exact, and its estimators are variance-controlled and, as the positive control shows, accurate when their bookkeeping is honest. The defense kit costs minutes per evaluation; we release it, with code and claims ledger, to make the guard layer explicit and inherited by default.

REPRODUCIBILITY STATEMENT

All experiments run on a single 8-GPU node; every number in this paper is backed by a committed artifact (JSON/Markdown) in the released repository, and the mapping from claims to artifacts is machine-checkable: a claims ledger (`CLAIMS_LEDGER.json`) maps 35 headline values to their source artifacts. All 35 entries pass the source-file and value-presence checks in `evidence_check.py`; these checks establish traceability, not independent verification of the experiments or conclusions. Training configs, dense-grid evaluation scripts, the entropy-floor module, audit scripts for the four external systems, and raw sweep outputs (52 MINDE configs, 193-point ITD grids, Plaid readout logs, the 10-config MMG scale sweep, and the from-scratch Plaid-recipe ablation) are included. External audits use only publicly released code and checkpoints.

ETHICS STATEMENT

This work documents a latent failure mode in published estimator code without invalidating published results, and we have taken care to state exposure claims structurally ("would bind under reuse") rather than as criticism of prior work. We are notifying the authors of the audited systems with our measurement code before any public naming of the undocumented-guard findings. The models trained here are small-scale character LMs on public data; no human subjects, personal data, or dual-use concerns apply.

AI USE STATEMENT

Large language model assistants were used extensively in this project, in every case under author direction and with author verification of the outputs. Uses in required-disclosure categories: LLM assistants helped refine the mechanism hypotheses and design the measurement protocols (including the dense-grid convergence protocol and the audit and ablation experiment designs); implemented the training, evaluation, and audit code and orchestrated the corresponding runs; drafted the mathematical derivations in App. A, which the authors subsequently checked; and assisted in interpreting experimental results. Additional uses: literature search and the initial pass of the structural survey (App. E), figure scripts, and drafting and revising the manuscript text. All survey classifications and file:line audit claims were verified by the authors against the cited papers and repositories; the research questions and final claims are the authors' own; the headline values are mapped to measurement artifacts by the ledger described in the Reproducibility statement; and the authors take full responsibility for all content of this paper.

REFERENCES

Sagnik Bhattacharya, Abhiram Gorle, Ahsan Bilal, Connor Ding, Amit Kumar Singh Yadav, and Tsachy Weissman. Itdpdm: Information-theoretic discrete poisson diffusion model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.05082.

Mustapha Bounoua, Giulio Franzese, and Pietro Michiardi. Sωi: Score-based o-information estimation. In *International Conference on Machine Learning (ICML)*, 2024. arXiv:2402.05667.

Paweł Czyż, Frederic Grabowski, Julia E. Vogt, Niko Beerenwinkel, and Alexander Marx. Beyond normal: On the evaluation of mutual information estimators. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.11078.

Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

Alberto Foresti, Ivan Butakov, Alexander Tolmachev, Giulio Franzese, Alexey Frolov, and Pietro Michiardi. Towards diverse and comprehensive benchmarks for mutual information estimation. *arXiv preprint arXiv:2607.03487*, 2026.

Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. Minde: Mutual information neural diffusion estimation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.09031.

Zhujin Gao, Junliang Guo, Xu Tan, Yongxin Zhu, Fang Zhang, Jiang Bian, and Linli Xu. Empowering diffusion models on the embedding space for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 4664–4683, 2024. arXiv:2212.09412.

Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2305.18619.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2107.00630.

Xianghao Kong, Rob Brekelmans, and Greg Ver Steeg. Information-theoretic diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2302.03792.

Xianghao Kong, Ollie Liu, Han Li, Dani Yogatama, and Greg Ver Steeg. Interpretable diffusion via information decomposition. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.07972.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-lm improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2205.14217.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.07524.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yirong Shen, Lu Gan, and Cong Ling. Information theoretic learning for diffusion models with warm start. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2510.20903.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70), 2018.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026.

Jiasheng Ye, Zaixiang Zheng, Yu Bao, Lihua Qian, and Mingxuan Wang. Dinoiser: Diffused conditional sequence learning by manipulating noises. *Transactions of the Association for Computational Linguistics*, 12, 2024. arXiv:2302.10025.

Longxuan Yu, Xing Shi, Xianghao Kong, Tong Jia, and Greg Ver Steeg. Mmg: Mutual information estimation via the mmse gap in diffusion. In *NeurIPS Workshop on Structured Probabilistic Inference and Generative Modeling (SPIGM)*, 2025. arXiv:2509.20609.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.02908.

A Derivations

A.1 The discrete-data identity has no differential-entropy constant

Write the channel as $z_\lambda = \sqrt{\alpha_\lambda}\, x + \sqrt{1-\alpha_\lambda}\,\epsilon$ with SNR $\gamma = e^\lambda = \alpha_\lambda/(1-\alpha_\lambda)$. The I-MMSE relation of Guo, Shamai, and Verdú, in the form used by Kong et al. (2023), states $\frac{d}{d\gamma} I(x; z) = \frac{1}{2}\,\mathrm{mmse}(\gamma)$; the change of variables $\gamma = e^\lambda$ gives $\frac{d}{d\lambda} I(x; z_\lambda) = \frac{1}{2}\, e^\lambda \mathrm{mmse}(\lambda)$. Now let s be a sequence of discrete tokens and $x = E(s)$ its embedding. Because E is injective on the vocabulary, $I(s; z_\lambda) = I(x; z_\lambda)$, and the two limits are pinned: $I \to 0$ as $\lambda \to -\infty$ (pure noise) and $I \to H(s)$ as $\lambda \to +\infty$ (tokens exactly recoverable: s is discrete, so mutual information saturates at the entropy rather than diverging like a differential entropy). Integrating the derivative between the two limits:

$$H(s) \;=\; \frac{1}{2}\int_{-\infty}^{\infty} e^\lambda\, \mathrm{mmse}(\lambda)\, d\lambda, \tag{3}$$

with no additive constant. Replacing the MMSE-optimal denoiser $\mathbb{E}[x \mid z_\lambda]$ by a model $\hat{x}_\theta$ can only increase the integrand pointwise ($\mathbb{E}\|x - \hat{x}_\theta\|^2 \geq \mathrm{mmse}$), so

$$\mathcal{L}(\theta) \;=\; \frac{1}{2}\int_{-\infty}^{\infty} e^\lambda\, \mathbb{E}\big\|x - \hat{x}_\theta(z_\lambda, \lambda)\big\|^2\, d\lambda \;\geq\; H(s), \tag{4}$$

which we use as a *one-sided* log-loss for the token sequence, and nothing stronger: Eq. 4 does not by itself exhibit a normalized model attaining $\mathcal{L}(\theta)$ (Kong et al., 2023 likewise derive the suboptimal-denoiser integral as an upper bound, not an identity), and no claim in this paper requires one. What the full-line integral does carry is the entropy guarantee: it can only err *upward*, toward longer code lengths, never below the entropy of the data. Autoregressive cross-entropy and the masked-diffusion NELBO are quoted alongside it in §4 as context, each of its own type (exact loss and variational bound, respectively). The trap is precisely the loss of the one-sidedness: a *truncated* evaluation of Eq. 4 carries no bound in either direction and can fall below $H(s)$. Dividing by $T \ln 2$ for a T-token sequence converts nats to bits per character. For our character-level ITD LMs, "honest" denotes the converged dense-grid evaluation of Eq. 4. Plaid readouts instead use the endpoint-corrected VLB or the separately specified floor-aware integral (§6, App. D). Truncation loses the full-line entropy guarantee when integral mass is omitted without valid correction terms.

The entropy-floor test (§6) is the contrapositive of Eq. 4: for i.i.d. uniform tokens $H(s) = T \log |V|$ exactly, so any pipeline that reports less (beyond MC error) cannot be computing Eq. 4: it is losing mass somewhere, no reference implementation required.

A.2 The peak-location law

The integrand $g(\lambda) = \frac{1}{2} e^\lambda \mathbb{E}\|x - \hat{x}\|^2$ transitions from its low-SNR plateau (denoising error = prior variance; g grows like e^λ) to its high-SNR decay (tokens resolved; error $\to 0$ faster than $e^{-\lambda}$) around the log-SNR at which the observation begins to distinguish embedding rows. At noise level λ, the posterior over token v at a position depends on likelihood ratios $\exp\{-\|z - \sqrt{\alpha}\, e_v\|^2/(2(1-\alpha))\}$, i.e. on the separations $e^\lambda\, \|e_v - e_w\|^2$ (for $\lambda \ll 0$, $\alpha \approx e^\lambda$). A token becomes resolvable from z alone when $e^\lambda\, \delta^2 \gtrsim O(1)$, with δ^2 a typical squared row separation. We assume:

1. *scale-invariant geometry:* during norm growth the direction structure of E is approximately preserved, so $\delta^2 \;\propto\; s\, d$ where s is the per-dimension second moment (empirically the growth is near-uniform across rows: min/max row norm$^2/d$ spread $\sim 4\times$ against a $150\times$ mean growth);
2. *context absorbs into the anchor:* the denoiser's use of context shifts the transition left by a model-quality-dependent constant, which the anchor point absorbs;
3. the low-SNR plateau tracks the prior variance $\propto s$, so the transition dominates the location of the bump's mass.

Then the transition satisfies $e^{\lambda^*} s = \text{const}$, i.e. $\lambda^*(s) = \lambda^*(s_0) - \ln(s/s_0)$, which is Eq. 2.

Table 3: Empirical test of the peak-location law. Top: five independently trained 43M models (embedding-init and sampler variants), anchored on the fixed-norm run. Bottom: 19 checkpoints of a single training run (`r2a1_orig43m`), anchored at step 10k; geometry, denoiser, and scale co-evolve, yet the prediction from norm telemetry alone stays within 0.66 log-SNR units (grid resolution 0.25). Cross-architecture transfer (306M predicted from a 43M anchor) shows a consistent ≈ -1.1 offset, confounded by width and denoiser quality; the within-architecture fits carry the mechanism.

point	scale s	peak	prediction	residual
cross-run 43M				
σ_E=0.5 (20k)	5.5	-6.00	-5.96	-0.04
σ_E=1.0 (20k)	7.6	-6.25	-6.28	$+0.03$
sampler loc= -6 (20k)	8.3	-6.25	-6.37	$+0.12$
σ_E=0.02 (100k)	178.0	-9.50	-9.43	-0.07
within-run 43M				
step 20k	12.7	-7.00	-6.45	-0.55
step 40k	57.1	-8.50	-7.95	-0.55
step 60k	113.5	-9.00	-8.64	-0.36
step 70k	138.9	-9.50	-8.84	-0.66
step 85k	168.6	-9.50	-9.03	-0.47
step 100k	185.4	-9.50	-9.13	-0.37
cross-arch 306M				
smoke5k (100k)	168.6	-10.50	-9.38	-1.12
train-c=6 (100k)	186.0	-10.50	-9.48	-1.02

Escape forecasting. Combining Eq. 2 with norm telemetry s_t yields a running forecast $\lambda^*(t)$ of the integrand peak. For the telemetry run, the measured bump extends ≈ 2 log-SNR units below its peak before decaying; the forecast therefore predicts truncation losses becoming material once $\lambda^*(t)$ comes within ≈ 2 units of the window edge at -10, i.e. $\lambda^* \approx -8$, which the anchored forecast places at $s \approx 60$–70 (steps 40–45k). The measured truncated-vs-honest gap first exceeds 0.3 BPC at step 45k ($s = 71$). Full per-checkpoint data: 19 checkpoints, max residual 0.66 over the $62\times$ range $s \in [3.0, 185.4]$ (Table 3 shows a representative subset; the complete table is in the released `a2_scaling_check` artifact).

A.3 Tail-bound arithmetic

The low-SNR tail of Eq. 4 below a window edge $\lambda_{\min} = \mu - c\sigma$ is bounded using $\mathbb{E}\|x - \hat{x}\|^2 \le 4R^2T$ with $R = \max_v \|E_v\|$. The constant 4 is not an assumption about the network: each position of x is an embedding row, and the denoiser output at each position is by construction a convex combination of embedding rows, $\hat{x} = \mathrm{softmax}(\text{logits})^\top E$, so both lie in the embedding ball of radius R and $\|x - \hat{x}\| \le 2R$ per position. (For an unconstrained regression head the same bound holds with R^2 replaced by $\max(R^2, \sup \|\hat{x}\|^2)$, a measured quantity.) Then: $\frac{1}{2}\int_{-\infty}^{\lambda_{\min}} e^{\lambda}\, \mathbb{E}\|x - \hat{x}\|^2\, d\lambda \le 2R^2T\, e^{\lambda_{\min}}$, i.e. $\text{tail}_{\text{low}} \le \frac{2R^2}{\ln 2}\, e^{\mu - c\sigma}$ BPC. With $(\mu, \sigma, c) = (2, 3, 4)$, $e^{-10} = 4.54 \times 10^{-5}$. At initialization ($\sigma_E = 1$, $R^2 \approx d = 1024$) the bound is $\frac{2 \cdot 1024}{\ln 2} \cdot 4.54 \times 10^{-5} = 0.134$ BPC. At the trained scale the measured max row norm gives $R^2/d = 181$, scaling the same bound to $0.134 \times 181 = 24.3$ BPC. The bound was never wrong; its *constant* was a learned quantity quoted at configuration time. The measured escaped mass, 3.404 BPC, sits comfortably under the honest bound and $25\times$ over the stale one.

B MINDE audit details

Where the window lives. `src/libs/SDE.py` (lines 10–24) fixes `VP_SDE(beta_min=0.1, beta_max=20, N=1000)` with T=1; `src/libs/importance.py` (lines 8–25) truncates importance sampling at `t_epsilon` $= 10^{-3}$ with a linear approximation below. Together these bound the log-SNR the estimator can

Table 4: MINDE input-scale sweep (mean ± std over 3 seeds, nats). MI is scale-invariant, so every drift from ground truth (GT) is estimator error. "Exposed" = official code with `-preprocessing none`; "guarded" = official default (undocumented `StandardScaler`). The analytic window prediction accounts for the bulk of the exposed bias.

mode	s=0.01	s=0.1	s=1	s=10	s=100
d=5, true MI = 2.00 nats					
exposed	1.20±.01	1.96±.01	2.00±.01	1.93±.02	1.45±.01
analytic window	0.76	1.97	2.00	1.99	1.55
guarded	2.00±.01	2.00	2.00	2.00	2.00±.01
d=20, true MI = 10.00 nats					
exposed	5.80±.01	9.90±.04	9.84±.02	9.65±.04	7.04±.04
analytic window	3.58	9.82	10.00	9.97	7.89
guarded	9.84±.02	9.82	9.82	9.82	9.84±.02

ever see to $\approx [-10.05, +9.12]$. The undocumented guard: `src/scripts/config.py:18` defaults `-preprocessing` to `rescale`, and `src/scripts/helper.py` (lines 38–40) applies `sklearn StandardScaler` to X and Y, stated nowhere in the paper. Any non-`rescale` value (a supported code path) disables it; nothing else in the pipeline pins the input scale. The guard is also weaker than it looks: a single global affine standardization pins the *average* scale, so data whose effective scale varies across the support (mixtures of lengthscales, heavy tails) can keep integrand mass near the window edges even in guarded mode.

Protocol. Benchmark Gaussian tasks with analytic ground truth (Czyż et al., 2023): d=5/MI=2 and d=20/MI=10 nats. Inputs multiplied by $s \in \{0.01, 0.1, 1, 10, 100\}$; MINDE-c retrained from scratch at each configuration (official training loop, 3 seeds, 100k/10k train/test samples); MI read out with the official importance-sampled estimator, a dense t-grid trapezoid on the same score network, and the paper's MINDE-σ variant. The *analytic window prediction* is the exact integral of the true Gaussian integrand restricted to $t \in [10^{-3}, 1]$ at the raw scale the SDE sees: what a perfect score model could at best report given the window.

Results. Table 4. Baseline validation at s=1 matches the paper's own accuracy (their bmi suite: `multinormal-dense-5-5` 0.584 ± 0.010 vs GT 0.593, paper reports 0.6). The two exposed failure edges are mechanistically distinct:

- $s \ll 1$ (*front edge*, high-SNR cutoff): the official importance sampler reads 5.80 on d=20 while a dense t-grid on the *same* trained network recovers 10.2–10.6: the information is inside the model but outside the sampler's $t \geq t_\epsilon$ view.
- $s \gg 1$ (*back edge*, T=1): sampler and dense grid agree (7.04 vs 6.95); the mass has escaped past T and is genuinely invisible to the trained model. No readout fix can recover it.

Both edges track the analytic prediction, confirming window truncation (not optimization failure) as the dominant mechanism. The MINDE-σ variant, whose reference score hard-codes unit scale, degrades non-monotonically and in the *opposite* direction at the back edge: $36.0/42.6/38.5$ nats across seeds at s=100 on d=20 (GT 10). All 52 configuration JSONs (task × s × seed × mode, with per-run MC replicates and dense-grid curves) are released.

C ITDIFFUSION AUDIT DETAILS

Reproduction anchor. We first re-derive the paper's numbers from the repo's shipped evaluation artifacts by exact re-computation of the estimator bookkeeping: all six stored values (continuous / dequantized / discrete bpd for tuned DDPM and IDDPM) reproduce to four decimals, and a verbatim GPU rerun of their `scripts/test.py` on the full 10k test set matches to ± 0.0002 bpd. The measurement machinery is therefore exact before any window is varied.

Table 5: ITD image estimator vs integration window: tuned IDDPM, CIFAR-10 test set, dense deterministic grid (193 points to clip 8), analytic discrete tails recomputed at each window edge. The continuous estimator is mildly under-integrated at the default window and converges by clip 5; the discrete column's clip 7–8 rise is the fp32-floor artifact (App. G), not signal.

clip	window (log-SNR)	cont. bpd	dequant. bpd	discrete bpd
2	$[0.07, 12.46]$	-4.0189	-3.8351	2.8875
3	$[-3.03, 15.55]$	-4.2258	-3.8624	2.8814
4 (default)	$[-6.13, 18.65]$	-4.2849	-3.8625	2.8814
5	$[-9.23, 21.75]$	-4.2933	-3.8625	2.8814
6	$[-12.32, 24.85]$	-4.2933	-3.8625	2.8816
8	$[-18.52, 31.04]$	-4.2933	-3.8625	2.9612

Window sweep. Table 5. The default window $(\mu, \sigma) = (6.26, 3.10)$ is *data-adaptive*: recomputed from the evaluation data's covariance spectrum at eval time (`diffusionmodel.py:246-251`); only c=4 is fixed. Fraction of the converged gap integral inside the default window: 99.540%; continuous shift clip $4 \to 5$: -0.0084 bpd, then flat to clip 8. Re-weighting their own shipped 1000-point importance sample to shrunken windows shows the continuous number still moving at the default edge (clip $3.5 \to 4$: -0.025 bpd) with a nonzero gap integrand at the right edge for the ITD-tuned model (35.5 of $d = 3072$): a mild, real, converging-by-clip-5 under-integration, in sharp contrast to our LM's unbounded one.

Rescale-without-refresh stress test. We scale inputs by s while freezing everything the estimator derived from its data snapshot (window, Gaussian reference entropy h_g, mmse_g curve, discrete x-interval), mimicking a scale drift the bookkeeping does not know about; the analytic truth shifts by exactly $\log_2 s$ bpd/dim. At s=0.5 the measured shift misses the analytic one by $+0.32$ bpd. At s=2.0 the estimator's clamped data-dependent term ($\mathrm{mmse}_g - \mathrm{mse}$, floored at 0) is *identically zero* at every grid point: the estimate collapses to the frozen Gaussian baseline ($+0.794$ bpd deviation), a window-independent silent bias that no widening can recover, while the discrete estimator simultaneously explodes ($> 10^6$ bpd; its rounding bookkeeping assumes $x \in [-1, 1]$). Caveat: the model itself is OOD on rescaled inputs, so part of the s=0.5 deviation is genuine model failure; the s=2.0 gap-collapse and the discrete blowup are, however, pure bookkeeping effects (the clamp and the frozen reference do not query the model at all when the gap is clamped to zero).

fp32 timestep saturation. Their UNet wrapper maps log-SNR to a continuous timestep via $\sqrt{\mathrm{sigmoid}(\lambda)}$ (`unet.py:577-579`); in fp32, $\mathrm{sigmoid}(\lambda)$ rounds to exactly 1.0 for $\lambda \gtrsim 17.3$, freezing the conditioning: $t(18.65) = t(25) = t(31)$. The default window's right edge already sits in this saturated regime, so wide-window evaluations beyond it measure extrapolation, not schedule-conditioned prediction: one more reason "just integrate wider" is not a fix (the discrete estimator is unaffected; its right tail is analytic).

D PLAID AUDIT DETAILS

Setup. Official released Plaid 1B checkpoint (1.275B params), zero-shot `text8` per their protocol (non-overlapping 1024-token BPE sequences, self-conditioning unrolled at eval, fp64 outer loop). Their pinned fused kernels were replaced by numerics-equivalent pure-torch shims (SDPA attention, tanh-GELU MLP, RMSNorm); the state dict loads with zero key mismatches.

Three readouts of one checkpoint (`text8`). (i) Their evaluator, run verbatim (3000 sequences, stochastic-t VLB): **1.1016** BPC. (ii) Our deterministic reassembly (reconstruction CE at γ_0 + dense-grid window integral + prior KL at γ_1; $0.2658 + 3.4004 + 0.0169$ nats/token): **1.1044** BPC. (iii) A floor-aware pure dense integral, cut at the integrand's clean trough at $\lambda = 8.75$: **1.0830** BPC. All within $\sim$3% of each other and of the reported 1.12 (`enwik8`: 1.1564 vs reported 1.18). Their learned integration window is $[-3.56, +6.80]$ in log-SNR (`gamma_bounds`, learned parameters), holding 94% of the floor-aware total.

Table 6: Guard ablation on the Plaid-small recipe (BPC). The trapped column tracks the error-floor dynamics rather than the model: it moves 4341→278 on the ablated run while the honest value moves 2.17→1.69. Mass below the transplanted window is negligible throughout ($< 10^{-3}$ nats); the entire misread is the high edge.

	guarded		ablated	
step	honest	trapped	honest	trapped
4k	2.194	2.60	2.170	4340.7
8k	2.038	2.27	1.983	2362.2
16k	1.928	2.12	1.879	2408.6
24k	1.856	9.89	1.784	954.1
32k	1.796	13.84	1.742	462.8
40k	1.750	12.60	1.692	278.5

Trap-leg audit (file:line). (1) *Free norms (closed):* `lib/models.py:96-98` L2-normalizes every embedding row on every forward pass. Checkpoint measurement: *raw* stored rows have mean norm 23.57 (max 38.99), a $24\times$ drift from unit init under the same optimizer pressure that grew our embeddings $150\times$, while the effective rows the likelihood sees measure 1.0 exactly. (2) *Fixed window (not present):* the limits are learned parameters (`lib/models.py:124-130`) and both tails are replaced by exact measured endpoint terms (reconstruction CE, `train.py:306-309`; prior KL, `train.py:313-318`). (3) *Snapshot tails (not present):* no tail extrapolation exists anywhere in the objective; the bound is a bona fide variational upper bound evaluated on the live model.

Symmetric checks. Where extension is *valid* (noise side), the extended tail integral over $[-20, -3.56]$ gives 0.0122 nats/token against their exact prior-KL endpoint term 0.0169: honest integration and exact endpoint terms agree to 0.005 nats. Where it is *invalid* (beyond the trained window, $\lambda \gtrsim 9$), the pure integral diverges loudly ($\hat{x}$-MSE floors at $\sim 10^{-5}$ while e^{λ} explodes; the convergence check flags it) rather than deviating silently: the inflationary mirror image of our silent deflation, caught by the same discipline. The entropy-floor test on i.i.d. uniform tokens passes in its endpoint-corrected form: 17.07 bits/token $\geq \log_2 32768 = 15$.

Caveats. Shim-vs-fused numerics differ at bf16 rounding level; the dense sweep used 64 sequences (window-integral MC error $\sim$0.01–0.02 nats/token); the CE endpoint was sampled at the nearest grid point ($\lambda = 7.0$ vs learned 6.80). None of these affect the $\sim$3% three-way agreement.

From-scratch guard ablation (setup). Plaid's default small configuration (dim 384, 16 blocks, 6 heads, embedding dim 16, sequence length 256; 28.4M parameters), trained from scratch on character-level `text8` (90/5/5M split, 27-token vocabulary) for 40k steps with identical hyperparameters and seeds; the ablated variant removes the L2 row normalization in `EmbeddingMatrix.forward()` (initialization unchanged; one-line diff shipped as a patch). Six checkpoints per variant, each read two ways: *honest* (reconstruction CE at γ_0 + dense deterministic grid over the learned window $[\gamma_0, \gamma_1]$ + closed-form prior KL at measured norms, convergence-checked) and *trapped* (verbatim transplant of the LM recipe: importance-sampled logistic$(2, 3)$, clip 4, i.e. window $[-10, 14]$, no endpoint terms). Table 6.

Reading the ablation. (1) *Mechanism attribution:* the learned windows sit at $[\gamma_0, \gamma_1] \approx [-3, 5]$, entirely inside the transplanted $[-10, 14]$, and the mass below -10 is $\sim 10^{-3}$ nats, so unlike our LM, nothing escapes *below*. The misread is the opposing trap of §3.4: the measured x-MSE floors at 4.5×10^{-4} (ablated) vs 3.1×10^{-5} (guarded) at $\lambda \geq 14$ (a $\sim 15\times$ ratio matching the $\sim 15\times$ ratio of effective scales, $\|\cdot\|^2/d$: 0.948 vs $1/16$), and $e^{\lambda}\times$floor over the nine units between γ_1 and the window edge dominates the integral. The edge integrand is still rising at the window boundary (the deterministic-grid transplant reads 393.5 BPC vs the importance sampler's variance-dominated 278.5; both meaningless), so any convergence check fires. (2) *Norm drift:* raw row norms drift in both variants (guarded raw mean $\|\cdot\|^2/d$ $0.28 \rightarrow 1.36$, unopposed because the projection makes raw scale a free direction; ablated $0.42 \rightarrow 0.95$ mean with rows up to 5.8), consistent with the

released 1B checkpoint's $24\times$ drift over a much longer schedule; drift accumulates with training duration, and this 40k-step run is short. (3) *Scope:* at this scale and duration the ablated run's misread direction is upward (floor harvesting), not the downward escape of our LM; the bracketing claim in §6 rests on both demonstrations jointly, not on this one reproducing deflation. Artifacts: `logs/route2_ext/plaid_ablation/`.

E STRUCTURAL SURVEY

Method and wording rule. Web-only survey of published noise-integral likelihood and MI estimators (arXiv/OpenReview PDFs plus GitHub file inspection; no external code executed at survey time), asking three questions of each system: is there a fixed/truncated integration window; can the representation scale drift after the window is set; and how are the tails handled (measured, bounded-from-snapshot, or assumed)? Nothing in this survey states or implies that any published *result* is wrong; the strongest classification is "structurally exposed *scheme*," which binds only under reuse with a drifting representation. The systematic sweep covers 27 estimator systems (Table 7) plus the sampling-side sibling (Zheng et al., 2025); two evaluation-reliability works were examined at full text after the sweep and are discussed below (the ϵ-floor MI benchmark (Foresti et al., 2026), and the LangFlow NELBO-sensitivity observation reported by Yang et al., 2026), for 30 works in total.

Table 7: Structural survey summary. **0 of 27 systems are exposed as published**; the 7 marked † (ITD/MINDE lineage) carry a structurally exposed scheme (fixed window and/or snapshot-scale tail) that would bind only under reuse with a drifting representation. Guards column: **P** scale pinned by fixed/normalized representation, **A** data-adaptive or learned window, **M** measured endpoint terms, **–** no NLL surface / no window.

system (arXiv)	estimator	window	guards
ITD† (2302.03792)	I-MMSE NLL	logistic IS, clip 4	P, A
Info-Decomp† (2310.07972)	MMSE-gap MI	manual loc/scale, clip 3	P
MMG† (2509.20609)	MMSE-gap MI	logistic (2,3); adaptive unrel.	P (undoc. scaler)
ItDPDM† (2505.05082)	Poisson NLL	fixed (α_0, α_1) + analytic tails	P
MINDE† (2310.09031)	score-diff KL	VP-SDE $[0.1, 20]$, $t_\epsilon{=}10^{-3}$	P (undoc. scaler)
SΩI† (2402.05667)	score-diff O-info	$[0, T]$, endpoints unstated	P
DIPHINE† (2606.18997)	score-diff Φ-ID	VP-SDE, 10-pt MC	P
IT discrete diff. (2510.24088)	discrete NLL	time-free (no window)	–
Warm-start ITD (2510.20903)	thermo. integration	$[\sigma_0^2, \sigma_1^2]$	P
VDM (2107.00630)	VLB	learnable endpoints	P, A, M
Plaid (2305.18619)	VLB	learned endpoints	P (undoc. L2), A, M
RePlaid (2605.18530)	VLB	learnable endpoints + shape	P, A, M
ScoreFlow (2101.09258)	ODE / SM bound	$[\epsilon, T]$ + Jensen corr.	P
i-DODE (2305.03935)	ODE MLE	$\gamma_\epsilon \approx -13.3$	P
LSGM (2106.05931)	latent NELBO	$[0, 1]$, small-t impl.-defined	(P), M; *unclear*
Diffusion-LM (2205.14217)	discrete-chain VLB	finite sum	M
SEDD / MDLM / D3PM / RADD	discrete NELBOs	schedule-const. tail or exact	–
BFN (2308.07037)	BFN loss	$[0, 1]$, no truncation	P

Continued on next page

Table 7: Structural survey summary (continued).

system (arXiv)	estimator	window	guards
CDCD / Differmer / DINOISER	no NLL reported	n/a	–
LD4LG / TEncDM (latent text)	no NLL reported	n/a	P
Analog Bits (2208.04202)	no BPD reported	n/a	P
sibling: Gumbel fp32 (2409.02908)	sampling-side	precision truncation	–

Cross-cutting observations. (1) *A single ancestral integrator:* the truncated-logistic sampler is defined once in the ITD repository and propagates by code reuse (Info-Decomp: clip 3; MMG: $(2, 3)$ defaults; our LM: $(2, 3, 4)$ hard-coded) and by design imitation (ItDPDM); the MINDE family reaches the same structure through fixed SDE constants. (2) *Tail certificates are scale snapshots* wherever they exist; the closest published mitigation (MMG's adaptive re-fit of the window from a measured MMSE curve) turns out to be described in that paper but absent from its released code, and its width rule under-covers at high MI (§5.4, App. H). (3) *The trap requires a conjunction no published paper completes:* fixed-window papers have fixed representations; learned-representation papers pin the scale in code or self-account via measured VLB endpoint terms. (4) *The drift leg is independently documented* as a training pathology (Dieleman et al., 2022; Gao et al., 2024; Ye et al., 2024). (5) *The protective pattern is consistent:* measured endpoint terms count escaped mass; fixed-window I-MMSE integrals have no term that can.

Note added in proof. Two works appeared after the survey was compiled. Foresti et al. (2026) document an ϵ-floor "perturbed-MI" bias in MI benchmarking, a distinct mechanism (no drift, no window escape, no code audit), complementary to our scale-invariance stress test. Yang et al. (2026), read at full text, is tabulated above; beyond its own (guarded) estimator it contributes an independent silent-deflation incident (a dropped chain-rule term in a Hutchinson-trace NELBO) and reports that a third system's ODE-based NELBO was "highly sensitive to design choices": convergent evidence that likelihood evaluation in this family fails quietly absent explicit discipline.

F Recipes, guardrails, and the checklist

Common setup. `text8`, 27-character vocabulary, 90/5/5M split, sequence length 256, DiT backbone over a learned embedding, AdamW (lr 3×10^{-4}, no weight decay), bf16 autocast training, EMA 0.9999 for all evaluations. Scales: 43M (d=384, L=16), 172M (d=768, L=16), 306M (d=1024, L=16/24). Original (trap) recipe: free-norm embedding, training and evaluation noise drawn from the truncated logistic with hard-coded $(\mu, \sigma, c) = (2, 3, 4)$.

The four-cell norm/window factorization and its quantitative interpretation are given in §4, Table 1.

The guarded recipe (`fix_v1`). Fixed-norm embedding (rows reparameterized to constant norm, the Plaid/CDCD guard), window placed from the measured integrand (μ=0, σ=4, c=5.5, covering $[-22, 22]$), and three runtime guardrails asserted at every evaluation: (i) the entropy-floor check on synthetic uniform batches; (ii) the measured-constant tail bound of App. A.3 logged next to every reported number (3×10^{-7} BPC at final eval); (iii) embedding-norm telemetry (mean/max row norm$^2/d$). Outcome over 100k steps: guardrails pass at all 50 evals, honest dense fp32 $=$ 4.1205 BPC with 100% in-window mass, importance-sampling and dense readouts agree in matched precision (4.228 ± 0.130 vs 4.235 bf16), and CE matches the unconstrained run (0.62 vs 0.65 nats): the norm growth was buying rescaling, not modeling power.

The dense-grid convergence protocol is specified in §4.

Practitioner checklist. For any likelihood/NLL/MI estimate built on a noise-level integral of a denoising statistic:

Table 8: Cross-run invariance of the honest value (dense fp32 grid, $[-22, 22]$). The honest total stays within 4.10–4.38 BPC across model size, embedding init, sampler center, training clip, and the fixed-norm reparameterization, while in-window mass collapses only in the long free-norm runs. Dense, window ($[-10, 14]$), and low-tail ($\lambda < -10$) values are in BPC; $s = \|e\|^2/d$.

run	dense	window	tail	peak λ	s
306M headline (100k)	4.265	0.861	3.404	-10.5	168.6
306M train-c=6 (100k)	4.133	0.731	3.402	-10.5	186.0
43M σ_E=0.02 (100k)	4.174	2.300	1.874	-9.5	178.0
43M σ_E=0.5 (20k)	4.324	4.255	0.069	-6.0	5.5
43M σ_E=1.0 (20k)	4.328	4.233	0.095	-6.25	7.6
43M sampler loc= -6 (20k)	4.316	4.214	0.102	-6.25	8.3
43M fixed-norm (100k)	4.120	4.108	0.013	-4.25	1.0
43M original recipe @20k	4.299	4.130	—	-7.0	12.7
172M original recipe @20k	4.380	4.191	0.189	-7.0	7.7×

1. **Entropy-floor test:** run i.i.d. uniform tokens through the full pipeline; below $\log_2 |V|$ bits/token ⇒ mass loss.
2. **Integrand-vs-window check at the final checkpoint**, not at init: plot the integrand on a grid wider than the window; report the in-window mass fraction next to the headline number (ours decayed from $\sim$98% at 20k steps to 20% at 100k).
3. **Recompute tail bounds from measured constants** at every evaluation; log them beside the numbers they certify.
4. **Train coverage ⊇ eval coverage ⊇ integrand support:** a model never trained at low SNR cannot be honestly evaluated there; widening only the eval window measures an OOD denoiser.
5. **Treat "better than the best known compression" as an alarm**, not a result, and check 1–4 before interpreting.
6. **Integrate in fp32:** bf16 alone shifted our converged dense integral by $+0.11$ BPC.
7. **Monitor representation scale during training** whenever the representation is learned; Eq. 2 turns the trace into an escape forecast.

Each item maps to a released artifact; item 1 ships as a standalone estimator-agnostic module.

G THE FP32-FLOOR ARTIFACT IN THREE CODEBASES

Mechanism. Let the measured denoising error at high log-SNR be $\widehat{\mathrm{mse}}(\lambda) = \mathrm{mse}(\lambda) + \delta$, where δ is a floating-point floor (rounding of near-zero squared errors, accumulation granularity, or saturated conditioning). The true term $e^{\lambda}\,\mathrm{mse}$ decays, but the artifact term $e^{\lambda}\,\delta$ *grows exponentially*, so the cumulative integral develops a spurious rise beyond some λ. Over-extending an empirical integral therefore *inflates* the estimate (the mirror image of truncation), and the correct treatment is an empirical integral inside a convergence-verified window plus analytic tails. Detection is cheap: a cumulative integral that develops a second rise after a clean trough is measuring its own precision floor.

Three independent instances. (1) *Our LM:* evaluating the dense integral in bf16 shifts the converged value $+0.11$ BPC (4.235 vs 4.120); the effect concentrates exactly where $e^{\lambda}\times$ (rounding floor) dominates the true integrand. (2) *ITD image code:* the discrete estimator rises $+0.0797$ bpd from clip 6 to 8 (Table 5): at $\lambda \gtrsim 25$ the fp32 floor on x-MSE is amplified by e^{λ}, inflating the empirical integral precisely where their analytic right-tail bound already covers the mass; their default configuration (empirical inside clip 4, analytic outside) is the correct treatment, and "wider" would have made their number *worse*. (3) *Plaid:* extending the pure integral beyond the trained window diverges ($\hat{x}$-MSE floors at $\sim 10^{-5}$/token while e^{λ} explodes; edge integrand 1.3×10^{5} nats/token/unit),

conflating the precision floor with OOD extrapolation; either way, mass manufactured by the readout, flagged loudly by the convergence check. The from-scratch ablation (App. D) weaponizes this same edge: a transplanted fixed window that integrates through it without a convergence check turns the floor into a 278-BPC headline. The sampling-side sibling is the fp32 Gumbel truncation of Zheng et al. (2025), where single-precision categorical sampling silently deflates generative perplexity $\sim 4\times$: same genre (a precision boundary quietly reshaping a headline metric), opposite side of the model (sampling vs likelihood readout).

H MMG AUDIT DETAILS

Where the window lives. `MMG_Unet.py` (lines 133–141) defines `logistic_integrate` with defaults $(\mu, \sigma) = (2, 3)$, clip 4 (the window $[-10, 14]$), fed by `-logsnr_loc` / `-logsnr_scale` (`main.py`, lines 39–42) and identical in the MOE variant (`MMG_Unet_MOE.py`, lines 191–199). The companion library carries an `update_logsnr_loc_flag` constructor argument that is stored but never read: a dead hook. A full-history search of both public repositories (all branches) finds no implementation of the paper's adaptive scheme; the experiment pipeline's default preprocessing standardizes both variables (`main.py`, lines 151–159), a step stated nowhere in the paper.

Protocol. Benchmark Gaussian tasks with analytic ground truth (Czyż et al., 2023): a dense d=5 task (MI 0.593 nats) and a sparse d=10 task (MI 9.904 nats), input scales $s \in \{0.01, 0.1, 1, 10, 100\}$, exposed protocol (`-preprocess none`), seed 0. Since MI is invariant to $X \to sX$, $Y \to sY$, any drift is estimator error. The *fixed* arm is the repository default. The *adaptive* arm reimplements the paper's two-stage scheme from its text: μ = the log-SNR where the measured conditional-MMSE curve crosses $d/2$, σ from the quarter-crossing, then a fresh model trained with the refit window (stage-1 preliminary model = the fixed run). For each cell we also compute the *analytic window prediction*: the exact integral of the true Gaussian integrand restricted to that run's window, i.e. what a perfectly trained denoiser could at best report given the window.

Table 9: MMG scale sweep, exposed protocol (nats; seed 0). "anl." = analytic window-restricted prediction. Fixed-window drift at extreme scales and the adaptive rule's under-coverage at high MI are both predicted by the analytic column; degenerate adaptive fits (fitted $\mu \approx -34$) collapse the estimate to zero.

s	fixed	anl.	drift	adaptive (μ, σ)	adaptive	anl.
Dense d=5, true MI = 0.593 nats						
0.01	0.568	0.590	−0.02	$(-33.5, 0.5)$	0.000	0.000
0.1	0.534	0.593	−0.06	$(5.2, 1.1)$	0.577	0.571
1	0.613	0.593	+0.02	$(0.5, 1.1)$	0.574	0.570
10	0.721	0.588	+0.13	$(-4.0, 1.1)$	0.619	0.572
100	0.398	0.338	−0.20	$(-8.7, 0.5)$	0.450	0.399
Sparse d=10, true MI = 9.904 nats						
0.01	4.930	4.791	−4.97	$(9.8, 1.3)$	5.416	5.742
0.1	8.926	8.924	−0.98	$(5.1, 1.3)$	5.575	5.765
1	9.774	9.888	−0.13	$(0.5, 1.3)$	5.835	5.782
10	9.735	9.900	−0.17	$(-4.2, 0.5)$	2.268	2.291
100	9.260	9.531	−0.64	$(-35.0, 0.5)$	−0.000	0.000

Results. Table 9. The fixed window is accurate at native scale (9.774 vs 9.904 at s=1) and drifts at the extremes (worst −4.97 nats at s=0.01 on the d=10 task; −0.20 at s=100 on d=5); the analytic window-restricted integral reproduces the measured values throughout, pinning the bias on the window rather than on training. The adaptive rule's location fit tracks the predicted $-2 \ln s$ migration where the crossing exists (fitted $\mu = 9.8, 5.1, 0.5, -4.2$ across s = 0.01–10 on d=10; prediction: −4.61 per decade), but the width rule under-covers the high-MI task at *every* scale including s=1 (5.84 vs 9.90; analytic prediction 5.78), a constant ∼4-nat blind spot, and when the measured curve fails to cross $d/2$ within the probed range the fit degenerates ($\mu \approx -34$) and the estimate collapses to zero. Both failure modes are properties of the described scheme, not of our reimplementation: the analytic column, which assumes a perfect denoiser, reproduces them.

Caveats. Single seed (the MC integration variance on these Gaussian tasks is small relative to the multi-nat effects, and the deterministic analytic column carries the attribution); the adaptive arm is necessarily a reimplementation, since no released implementation exists. Raw sweep outputs: `logs/route2_ext/mmg/`.