YOUR FID CANNOT SEE THE TAIL: AUDITING AND REPAIRING THE RARE-EVENT CALIBRATION OF LARGE GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are increasingly used as data: synthetic corpora, simulators, and world models supply samples that downstream systems treat as draws from the real world. For these uses, the question is not only whether typical samples look good, but whether events of a chosen severity occur at the right rate. We study this question as tail calibration. Given a black-box sampler Q, a reference corpus P, and a fixed severity functional R, we audit the exceedance-rate ratio $\rho(\tau) = Q\{R \geq \tau\}/P\{R \geq \tau\}$ at named tail depths. We first show why this rate must be measured directly: FID-style finite-moment metrics can agree while tail mass changes, and any sample-based test pays a rare-event cost that scales with the inverse tail probability. The audit spends its sample budget on the tail indicator, reports intervals that combine exceedance-count error with threshold-estimation error, and measures its own error on real-versus-real splits before judging models. We also give a black-box wrapper that remaps generated samples to a calibrated severity law without retraining or weight access. The wrapper fixes exactly the severity law that the audit certifies, up to the audit's own measured error, and along shape coordinates that its transport preserves, no severity-keyed post-processor can produce conditional tail shapes beyond mixtures of the sampler's own. Across public image-generation samplers, two reference corpora, and multiple severity functionals, the certified ratios range from nearly sixfold under-production to more than fivefold over-production. Bulk-quality scores alone do not explain these errors, and the verdict can change when the severity functional changes. Even when exceedance rates are calibrated, conditional tail shapes remain compressed. These results suggest that samplers used as data should be reported with per-(configuration, functional) tail-calibration certificates, together with the audit's own measured error.

1 INTRODUCTION

Generated samples are increasingly used as data. Synthetic corpora augment or replace training sets (Gu et al., 2025), learned simulators supply the scenarios against which autonomous systems are validated (O'Kelly et al., 2018; Corso et al., 2021), and world models roll out the futures that planners score. In each case the consumer treats samples as draws from the world, and much of the risk lies in rare events. A safety case built on a learned simulator covers only the accidents that the simulator produces at the right rate, and simulation-based safety validation regards rare-event fidelity as its main obstacle (Liu & Feng, 2024). For these uses the relevant question is whether events of a given *severity* occur at the right *frequency*.

Standard evaluation does not answer this question. FID (Heusel et al., 2017) and the metrics proposed to replace it (Stein et al., 2023; Kynkäänniemi et al., 2022; Jayasumana et al., 2024) summarize where the bulk of the sample distribution lies. Section 2 identifies two obstacles that better features would not remove. First, a metric computed from finitely many feature moments can take the same value on two samplers whose tail-calibration ratios differ by any prescribed amount up to $1/p$ (Proposition 1). Second, there are pairs of samplers for which any test based on n draws, whatever statistic it uses, needs $n \gtrsim p^{-1}(\sqrt{c}-1)^{-2}$ draws to detect a factor-c error at tail probability

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.

p (Proposition 2). At the customary budget of 50k samples the 95% interval for a tail-rate ratio at $p = 10^{-4}$ is 0.9 to 2.4 units wide (§5), the same order as the factor-two errors we find.

We therefore measure the tail rate directly. The audited object is a *deployed sampler* Q, meaning a model together with its sampling configuration, accessible only through draws. It is compared with a reference corpus from the data law P along a severity functional R fixed in advance, and the audited quantity is the tail-calibration ratio $\rho(\tau) = Q\{R \geq \tau\}/P\{R \geq \tau\}$ at depths $p \in \{10^{-3}, 10^{-4}\}$. We borrow the term from Allen et al. (2025), who study forecasts whose law is known exactly; a sampler's exceedance rates must be estimated from draws (§6). The audit (§3) spends its budget on the tail indicator, reports an interval that combines exceedance-count error with threshold-estimation error, and first measures its own error by auditing held-out real data in the sampler's role.

When a sampler fails the audit, a black-box *wrapper* (§4) remaps its samples so that their severity law equals a calibration law fitted to real data. The wrapper needs no weights, gradients or retraining. Its guarantee has the same scope as the audit's estimand. The severity law is corrected up to the calibration law's measured error and conditional tail shapes are left as the sampler produced them. Along shape coordinates that the transport preserves, a post-processor that acts on a sample only through its severity can at most mix the sampler's own conditional shapes. The image-level implementation moves only the top 1% of draws, so its effect on Fréchet-type scores is bounded by the feature displacement of that 1% (Proposition 4).

Contributions. *(i) Theory.* We formalize tail calibration along a severity functional for black-box samplers and prove two limits of standard evaluation: finite-moment metrics can agree while tail mass differs, and sample-based tests can need a number of draws of order $1/p$ (Props. 1–2). *(ii) Audit and recalibration.* We give an audit whose intervals combine count and threshold error under a stated local assumption, checked on real-versus-real splits before any model is judged and on planted ratios. We give a wrapper whose guarantee matches the audit's estimand, transports for a general R, a proof that severity-keyed post-processing cannot leave the sampler's family of conditional tail shapes along transport-invariant coordinates (Prop. 3, Cor. 1), and a bound on the change in Fréchet distance caused by wrapping (Prop. 4). *(iii) Evidence.* We audit eight public ImageNet sampler configurations under three severity functionals and replicate the protocol on CelebA-HQ. The unit of reporting is a (configuration, functional) pair, because verdicts change along both.

Under the pixel-space severity R_1, $\hat{\rho}(10^{-3})$ ranges from 0.17 to 1.99 across the eight configurations and five intervals exclude 1. Similar bulk quality does not imply a similar tail: iDDPM and EDM have published FIDs of 2.90 and 2.60 and ratios of 0.17 and 1.81 (Table 1, Fig. 1). iDDPM over-produces under a representation-space severity R_2 (1.55), and disabling classifier-free guidance moves DiT-XL/2 from 0.98 to 0.64. Among the seven configurations wrapped at the image level, the four whose intervals exclude 1 move to 0.98–0.99 while the measured Inception FID changes by at most 0.0024. Conditional tail dispersion along the first principal component, which the wrapper cannot change, is 0.40–0.71 of the real value in all eight configurations. On CelebA-HQ, DDPM over-produces by factors of 3.6 to 5.6.

2 Tail Calibration and the Limits of Standard Evaluation

Definition 1 (Tail calibration)**.** Let $R : \mathcal{X} \to \mathbb{R}$ be a fixed measurable severity functional, P the data law and Q the law of a deployed sampler. For a level τ with $P\{R \geq \tau\} > 0$ the *tail-calibration ratio* is $\rho(\tau) = Q\{R \geq \tau\}/P\{R \geq \tau\}$. Let τ_p be the upper-p quantile of R under P; we assume the law of R under P has no atom at τ_p, so that $P\{R \geq \tau_p\} = p$. The sampler is *tail-calibrated along R at depth p* if $\rho(\tau_p) = 1$; it under-produces if $\rho(\tau_p) < 1$ and over-produces if $\rho(\tau_p) > 1$. We write $\rho(p)$ for $\rho(\tau_p)$.

The ratio ρ is a rate at a fixed depth, and the two results below show that distance-type scores say little about it. The first is structural and concerns finite-moment metrics; the second is statistical and concerns every test. We write $T_\tau = \{R \geq \tau\}$ for the tail region and B_τ for its complement.

Assumption 1 (Moment mixability)**.** For the moment map $\psi : \mathcal{X} \to \mathbb{R}^d$ that defines the metric (for FID, $\psi = (\varphi, \varphi\varphi^\top)$ with φ the Inception features, Heusel et al., 2017), there exist a law Q_T supported in T_τ and a law Q_B supported in B_τ with $\mathbb{E}_{Q_T}\psi = \mathbb{E}_{Q_B}\psi$.

Remark 1 (Reading Assumption 1). A metric of the form $m(\mathbb{E}_Q\psi)$ receives a sampler only through the vector $\mathbb{E}_Q\psi \in \mathbb{R}^d$. The assumption says that tail membership cannot be read from that vector:

some population of bulk images has the same average features as some population of tail images. By Carathéodory's theorem it is enough that one point of $\psi(T_\tau)$ be a convex combination of $d{+}1$ points of $\psi(B_\tau)$ (Fig. 2, left). If the assumption fails, the convex hulls of $\psi(T_\tau)$ and $\psi(B_\tau)$ are disjoint, so a hyperplane in $\mathbb{R}^d$ separates them; this happens when the squared severity is itself linear in ψ (App. A). It is natural when severity and metric use different representations, the case we test below.

Proposition 1 (Structural blindness of finite-moment metrics)**.** *Let $\mathcal{M}(P,Q) = m(\mathbb{E}_Q\psi)$ for an arbitrary function m and a moment map $\psi : \mathcal{X} \to \mathbb{R}^d$ satisfying Assumption 1 at level τ_p. Then for every c with $1 < c \le 1/p$ and every base sampler Q_0 there exist samplers Q_1, Q_2 with $\mathcal{M}(P,Q_1) = \mathcal{M}(P,Q_2)$ and $\rho_2(\tau_p) - \rho_1(\tau_p) \ge c$.*

The proof (App. A) perturbs Q_0 by mixture weight $\varepsilon = cp$, once toward Q_B and once toward Q_T. The two perturbed samplers share a moment vector by Assumption 1, so every m assigns them the same score, and their tail masses differ by exactly ε. The condition $c \le 1/p$ ensures that ε is a valid mixture weight.

Fig. 2 (App. A) illustrates the construction on real features, with the moments matched approximately. The tail is defined by R_1 at $p = 10^{-3}$ and ψ is the MAE ViT-B/16 mean embedding ($d = 768$). Q_T is uniform on the 536 held-out real images above $\tau_{10^{-3}}$. Q_B is a 529-atom law on held-out real images below it, within the Carathéodory bound $d{+}1 = 769$, and its mean embedding matches that of Q_T to a relative error of 9.4×10^{-4}. With Q_0 = EDM at $M = 10^5$, mixing in either law at $\varepsilon = cp$ changes $\rho(10^{-3})$ by exactly c and changes the Fréchet distance in that feature space by 3.2×10^{-5} (c=2) and 1.1×10^{-4} (c=5). The larger change is 1.8% of the smallest gap between two audited configurations and 6% of the metric's Monte-Carlo noise. An unmatched bulk law at the same ε moves the distance 193 times further, so the small change is due to the moment matching. The second-moment block is matched to 7.8–8.8% in Frobenius norm (App. A).

Remark 2 (Scope of Proposition 1). The proposition covers Fréchet-style scores from a fixed finite-dimensional feature map (Heusel et al., 2017; Stein et al., 2023) and functions of them. It does not cover characteristic-kernel scores such as CMMD (Jayasumana et al., 2024), whose mean embedding determines the law so that Assumption 1 cannot hold, or likelihood-based scores such as FLD (Jiralerspong et al., 2023), which are not moment functionals. Both families are subject to Proposition 2. On the pair constructed above, an RBF-kernel MMD separates Q_B from Q_T (permutation $p \le 0.04$ at 5,000 per side) but not the mixtures Q_1, Q_2, for which detection would need about 4×10^7 draws (App. A).

Proposition 2 (Sample-size blindness of any test)**.** *Let Q_1, Q_2 agree conditionally on the tail indicator $\mathbf{1}\{R \ge \tau_p\}$ and satisfy $Q_1(T_{\tau_p}) = p$, $Q_2(T_{\tau_p}) = cp$ with $cp \le \frac{1}{2}$. Any test (in particular any evaluation metric with any decision rule) based on n i.i.d. sampler draws that distinguishes Q_1 from Q_2 with error probability below $1/3$ requires $n \ge C \,/\, \big(p(\sqrt{c}-1)^2\big)$, C a universal constant.*

Remark 3 (What 50k evaluation samples resolve). The bound applies to moment, kernel and likelihood metrics alike because it limits the information in the draws. It gives the scaling in p and c (numerical values in App. A), and §5 measures the resolution of the 50k budget directly. The audit of §3 uses the tail indicator, the sufficient statistic of the two-point problem in the proof, and reports the interval width that the available budget gives.

3 The Audit

Severity functionals. A severity functional is a fixed measurable map $R : \mathcal{X} \to \mathbb{R}$, chosen before any model is seen and fitted on real data only. We use three and report each separately. R_1 is the whitened radius on the top K=50 eigenpairs of the pixel covariance and measures pixel-space atypicality. R_2 is the same radius on the top 50 components of an MAE-pretrained ViT-B/16 patch-token mean (He et al., 2022); it measures atypicality in a learned representation, not human semantics. R_3 is the maximum of 16 per-position whitened radii on a 4×4 patch grid. Definitions and Algorithm 1 are in App. B.

Data splits and audited quantities. The reference corpus, ImageNet-32 (Chrabaszcz et al., 2017), is split once under a fixed seed. The calibration split $\mathcal{D}_{\text{cal}}$ (n_f = 200,000) is used to fit the functional, the calibration law of §4 and the shape basis. The held-out split $\mathcal{D}_{\text{hold}}$ (n_h = 1,081,167) supplies the reference thresholds: τ_p is estimated by the empirical upper-p quantile of R on $\mathcal{D}_{\text{hold}}$.

The data law P is the distribution from which the corpus is drawn, so all claims concern that population and not images in general, and the threshold is an estimate of the population quantile. The sampler is drawn M times (15,104 to 100,000 in our experiments) and the audit reads only the severities $r_i = R(X_i)$. Beyond the depth that $\mathcal{D}_{\text{hold}}$ can resolve (the corpus floor), thresholds come from the generalized-Pareto tail of App. C and every such number is labeled as extrapolation. A separate *shape audit* projects the sampler's rank-matched tail directions and the held-out real exceedance directions onto a principal-component basis fitted on the calibration tail, and tests equality of dispersion per component.

Estimator and count interval. Let $k = \#\{i : r_i \geq \tau_p\}$ and $\hat{q} = k/M$, so that $\hat{\rho}(p) = \hat{q}/p$. By the central limit theorem $(\hat{q} - q)/\sqrt{q(1-q)/M}$ is approximately standard normal. The textbook (Wald) interval substitutes $\hat{q}$ for q in the standard error and gives $\hat{q} \pm z\sqrt{\hat{q}(1-\hat{q})/M}$. It is unreliable for rare events. At $p = 10^{-4}$ and $M = 10^5$ the expected count is $Mp = 10$. The Wald interval is symmetric, has zero width at $k = 0$, can have a negative lower limit (for iDDPM, $k = 1$ gives $[-0.10, 0.30]$ on the ρ scale), and covers well below the nominal level when Mq is small (Brown et al., 2001). The Wilson score interval (Wilson, 1927) keeps q in the standard error and solves $|\hat{q} - q| \leq z\sqrt{q(1-q)/M}$ for q. Its endpoints stay in $[0, 1]$, it is asymmetric at small counts, and its coverage stays close to nominal for counts of this size (Brown et al., 2001). We use it at $z = 1.96$ and write $[L_W, U_W]$ for its endpoints divided by p (App. B.5).

Threshold error and the certificate. The threshold τ_p is itself estimated from n_h held-out points, and a threshold that is too high lowers $\hat{\rho}$ whatever the sampler does. Let G_P, G_Q be the survival functions of R under P and Q, with hazard rates $h = -\mathrm{d}\log G/\mathrm{d}\tau$. To first order (App. B.5) the threshold adds a relative error $\eta_p\,\sigma_\tau$ to $\hat{\rho}$, where $\sigma_\tau = \sqrt{(1-p)/(n_h p)}$ and $\eta_p = h_Q(\tau_p)/h_P(\tau_p)$.

Assumption 2 (Locally constant ratio)**.** At τ_p, $\mathrm{d}\log\rho/\mathrm{d}\log p = 0$, which is equivalent to $\eta_p = 1$.

In general $\eta_p = 1 + \mathrm{d}\log\rho/\mathrm{d}\log p$ (App. B.5). A sufficient condition for the assumption is that G_Q be proportional to G_P near τ_p, as for a sampler calibrated throughout that neighborhood. Under the assumption the threshold term depends on neither density and equals σ_τ, which is 3% at $p = 10^{-3}$ and 10% at 10^{-4} for $n_h = 1.08$M. The count term dominates at our budgets, so the interval is insensitive to moderate departures from $\eta_p = 1$. The reported certificate combines the two sources in quadrature, widening each half-width separately:

$$\mathrm{CI}_{\text{quad}} = \Big[\hat{\rho} - \sqrt{(\hat{\rho} - L_W)^2 + (z\,\hat{\rho}\,\sigma_\tau)^2},\ \ \hat{\rho} + \sqrt{(U_W - \hat{\rho})^2 + (z\,\hat{\rho}\,\sigma_\tau)^2}\Big], \quad \sigma_\tau = \sqrt{\frac{1-p}{n_h\,p}}. \tag{1}$$

Every verdict in this paper is read from (1): a configuration is labeled under-producing or over-producing at depth p when the interval lies below or above 1, and *calibrated* when the interval contains 1. The last label means that calibration is not rejected; it does not certify $\rho = 1$. The interval is conditional on the fitted functional R: it accounts for the sampler's draws and for the held-out threshold, not for refitting R on another calibration split, whose effect we report separately (§5). Its 95% level is nominal, since both components are approximations, so we check coverage empirically. Against an independently planted ratio, adding the threshold term raises coverage for planted $\rho \geq 1$ from 0.82–0.89 to 0.92–0.94 (all planted ratios in App. F; the treatment of cells with $k \leq 1$ is in App. B.5).

Self-validation. Because ρ is a ratio of two estimated tail rates, the audit can err when no model does. Before any model is judged, SELFVALIDATE audits real data in the sampler's role, where the answer is 1 (§5). Beyond the corpus floor we report instead the calibration-law ratio $e(p)$ of §4 as a distribution over split seeds.

Lemma 1 (Downstream inheritance)**.** *Let synthetic data be drawn i.i.d. from a sampler Q with tail-calibration ratio $\rho(\tau)$, and let a downstream analyst estimate $P\{R \geq \tau\}$ by the plug-in frequency $\hat{p}_n = \frac{1}{n}\#\{i \leq n : R(X_i) \geq \tau\}$. Then $\mathbb{E}\,\hat{p}_n = \rho(\tau)\,P\{R \geq \tau\}$: the estimate is biased by exactly the factor $\rho(\tau)$, at every corpus size n.*

A larger synthetic corpus reduces the variance of $\hat{p}_n$ and leaves this multiplicative bias unchanged (proof in App. A).

The reporting unit. An audit result is a certificate for a (sampler configuration, severity functional) pair, and verdicts change along both indices (§5). We therefore write "configuration X is calibrated along R at depth p" and never "model X is tail-calibrated".

4 The Recalibration Wrapper

The wrapper maps the sampler's severity law F_Q to a *calibration law* $\widehat{F}$ fitted on real data: the empirical severity distribution of $\mathcal{D}_{\mathrm{cal}}$ below an adaptive threshold u, spliced to a generalized-Pareto tail above it (App. C). It applies the monotone map $T = \widehat{F}^{-1} \circ F_Q$. Each of the M draws is rank-transformed, its rank is pushed through $\widehat{F}^{-1}$, and the sample is moved to the new severity by a *transport* $\phi_{r \to r'}$ satisfying $R(\phi_{r \to r'}(x)) = r'$ (Alg. 2, App. B). The sampler is used only as a source of draws.

Proposition 3 (Wrapper guarantee)**.** *Assume F_Q is continuous and let $\tilde{Q}$ be the law of the wrapped sampler run with the population F_Q in place of the rank transform. Then:* (i) Exactness. *The severity law of $\tilde{Q}$ is exactly $\widehat{F}$. Hence $\tilde{\rho}(\tau_p) = (1 - \widehat{F}(\tau_p))/p$ for every sampler, and $|\tilde{\rho}(\tau_p) - 1|$ is the estimation error of $\widehat{F}$ against the data law, with no contribution from Q.* (ii) Shape passthrough. *Suppose T is strictly increasing on the range of severities considered. For any transport-invariant shape coordinate S ($S \circ \phi_{r \to r'} = S$), the wrapped conditional shape law at severity s equals the sampler's at severity $T^{-1}(s)$. Conditional tail shapes are inherited up to this monotone relabeling.* (iii) Plug-in error. *With the M-draw rank transform in place of F_Q, the wrapped severity law deviates from $\widehat{F}$ by the error induced by $\sup_v |\widehat{F}_Q - F_Q|$, bounded uniformly by $\sqrt{\log(2/\delta)/2M}$ with probability $1 - \delta$ (Dvoretzky–Kiefer–Wolfowitz, Massart, 1990).*

The monotonicity in (ii) holds above the splice point u, which lies below every audited threshold, provided F_Q is strictly increasing there: $\widehat{F}$ is a generalized-Pareto law above u, whereas its empirical part below u has steps. Part (ii) extends to a whole class. Let Φ be any post-processor that reads a sample only through its severity, $\Phi(X) = \phi_{R(X) \to g(R(X),U)}(X)$ with measurable g and $U \perp X$. For every shape coordinate S that the transport preserves, each conditional law of $S(\Phi(X))$ is a mixture of the sampler's own, so no such Φ produces a shape law outside the closed convex hull of the sampler's family (Cor. 1, App. A). A transport that alters S falls outside this statement, and the shape audit measures S along the coordinates that our transport preserves.

Constructing the transport. R_1 and R_3 have explicit transports (App. B.1). For a general differentiable R, including representation-space functionals such as R_2, App. B.4 gives three training-free constructions. *(a)* If a chart $x \leftrightarrow (r, s)$ with $r = R(x)$ is available, set $\phi_{r \to r'}(x) = x(r', s)$; for R_1 the chart is $(\|w\|, w/\|w\|)$ in whitened coordinates w. *(b)* Integrate the gradient flow $\dot{x} = G^{-1}\nabla R/(\nabla R^\top G^{-1} \nabla R)$ for time $r' - r$, with G a positive-definite metric. Along it $\frac{\mathrm{d}}{\mathrm{d}s} R(x(s)) = 1$, so the target severity is reached exactly whenever the flow exists up to time $r' - r$, and the flow-line label is a preserved shape coordinate. *(c)* Take the minimal displacement $\arg\min_y \|y - x\|_G^2$ subject to $R(y) = r'$, which satisfies parts (i) and (iii) of Proposition 3 but in general preserves no shape coordinate. For R_1 with G the whitening metric, (b) and (c) both reduce to the radial rescaling $w \mapsto (r'/r)w$ used in our experiments (Proposition 5); other metrics reach the same severity along a different path. A transport could also be learned, at the price of a second trained model.

The calibration-law ratio. By part (i) the sorted wrapped severities are $r'_{(i)} = \widehat{F}^{-1}\big(i/(M+1)\big)$, a deterministic function of $\widehat{F}$ and M. Two configurations wrapped at the same M therefore have identical wrapped severities whatever their raw ratios (verified in §5), and the residual miscalibration after wrapping is that of $\widehat{F}$. We call

$$e(p) \;=\; \frac{1 - \widehat{F}(\tau_p)}{p}, \qquad \text{so that} \qquad |\tilde{\rho}(\tau_p) - 1| = |e(p) - 1| \text{ identically,} \tag{2}$$

the *calibration-law ratio* and $|e(p) - 1|$ the *self-error* of the audit. On the reference split $e(10^{-3}) = 0.990$ and $e(10^{-4}) = 0.688$. With $M = 10^5$ draws the rank grid has resolution $1/(Mp) = 0.1$ at $p = 10^{-4}$ and the wrapped ratio there is 0.600. Over four further split seeds

$e(10^{-4}) \in \{0.87, 0.78, 1.04, 1.04\}$, and we report $e(p)$ as this distribution. A wrapped ratio of 0.60 at 10^{-4} thus reflects the accuracy of $\widehat{F}$ at that depth and carries no information about the sampler.

Remark 4 (The support boundary). A transported image can leave the pixel box. We restore feasibility by damped alternating projection onto the box and the severity level set, and report the residuals of this step. Only draws with rank $v_i > 0.99$ are transported, a region that contains every draw whose raw or target severity reaches $\tau_{10^{-3}}$, and 99.5–100% of them land within 0.05 of target. In floating point the image-realized rates equal the remapped rates; after rounding to 8-bit images they differ by at most four exceedance counts (App. E). Transports to much deeper severities have larger residuals (App. H), and a support-adapted chart, where one exists, removes the problem (App. C).

The cost of wrapping in Fréchet distance. Let φ be the feature map of a Fréchet-type score and let $d_F(P, Q)$ be the Fréchet distance between the Gaussian fits to $\varphi_{\#}P$ and $\varphi_{\#}Q$, so that $\mathrm{FID} = d_F^2$.

Proposition 4 (Fréchet cost of post-processing). *Let $X \sim Q$ and $X' = \Phi(X)$ for any post-processor Φ, with law $\tilde{Q}$. Then $\left|d_F(P, \tilde{Q}) - d_F(P, Q)\right| \le d_F(Q, \tilde{Q}) \le \left(\mathbb{E}\|\varphi(X') - \varphi(X)\|^2\right)^{1/2}$. If Φ changes only a set of draws of Q-probability π, the right-hand side equals $\sqrt{\pi}\, D$, where D is the root-mean-square feature displacement of the changed draws.*

The proof (App. A) combines the triangle inequality for d_F with Gelbrich's bound (Gelbrich, 1990) and also holds for the empirical laws of the M draws. The wrapper must transport every draw whose raw or target severity exceeds the shallowest audited threshold, so a fraction $\pi \approx p \max\{1, \rho(p)\}$ suffices and repairs at smaller tail probabilities cost less. The feature mean and covariance move by terms of order π, so to first order the change in FID is linear in π (App. A.6). Our implementation uses $\pi = 0.01$, and §5 measures the change.

5 Experiments

Setup. The reference corpus is ImageNet-32 with the split of §3. R_1 is the default functional, and numbers without a subscript refer to it. Its thresholds $\tau_{10^{-3}} = 13.93$ and $\tau_{10^{-4}} = 16.13$ are both empirical quantiles of $\mathcal{D}_{\mathrm{hold}}$; the second rests on about 108 held-out exceedances, and the generalized-Pareto tail is used only beyond that depth. We audit eight deployed sampler configurations from three model families released over five years; two pairs share a network and differ only in how it is sampled. $M = 100{,}000$ except for the DiT configurations (50,000) and EDM-churn, stopped by a wall-time limit at $M = 15{,}104$ (details in Apps. H and J).

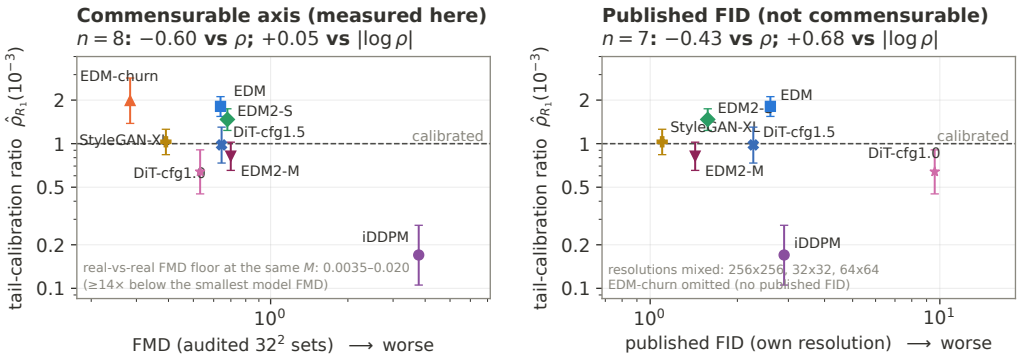

Figure 1: Audited $\hat{\rho}(10^{-3})$ under R_1 (log scales, 95% quadrature intervals, dashed line = calibrated) against FMD measured on the audited samples (**left**) and against published FID (**right**).

Bulk quality does not determine tail calibration. iDDPM and EDM have nearly the same published FID (2.90 and 2.60) and err in opposite directions, at 0.17 and 1.81 times the reference rate. Over all eight configurations $\hat{\rho}(10^{-3})$ ranges from 0.17 to 1.99 and five intervals exclude 1. At 10^{-4} only iDDPM still excludes 1, and the intervals are wide, up to $[0.62, 5.86]$ for EDM-churn at $M = 15{,}104$. At the customary budget of $M' = 50$k the same estimator has mean 95% widths

Table 1: Tail-calibration results under R_1, with R_2 and R_3 point estimates (full intervals in Table 21). Brackets are 95% quadrature intervals; k is the exceedance count at $\tau_{10^{-3}}$. $\downarrow$ under-produces, $\uparrow$ over-produces, unmarked = interval contains 1 (marks for R_2, R_3 use Wilson-only intervals; recomputing all 24 cells under Eq. (1) changes no verdict). FMD is measured on the audited 32^2 samples against one fixed real reference in MAE ViT-B/16 feature space. FID is the value published with each checkpoint, at its own resolution and protocol (64^2 for iDDPM, EDM and EDM2, 32^2 for StyleGAN-XL, 256^2 for DiT-XL/2); Inception FID measured on the audited samples is in Table 12.

		quality				R_1 (pixel PCA-50)		R_2	R_3
Configuration	Family	FID	FMD	M	k	$\hat{\rho}(10^{-3})$	$\hat{\rho}(10^{-4})$	$\hat{\rho}(10^{-3})$	$\hat{\rho}(10^{-3})$
iDDPM (Nichol & Dhariwal, 2021)	diff.	2.90	3.78	100k	17	$0.17^{\downarrow}$ [0.11, 0.27]	$0.10^{\downarrow}$ [0.02, 0.57]	$1.55^{\uparrow}$	$0.32^{\downarrow}$
EDM (Karras et al., 2022)	diff.	2.60	0.64	100k	181	$1.81^{\uparrow}$ [1.54, 2.11]	1.70 [0.99, 2.77]	0.93	$1.47^{\uparrow}$
EDM-churn†	diff.	–	0.28	15.1k	30	$1.99^{\uparrow}$ [1.38, 2.84]	1.99 [0.62, 5.86]	0.53	0.86
EDM2-S (Karras et al., 2024)	diff.	1.58	0.68	100k	147	$1.47^{\uparrow}$ [1.23, 1.74]	0.80 [0.38, 1.59]	0.84	0.89
EDM2-M (Karras et al., 2024)	diff.	1.43	0.70	100k	82	0.82 [0.65, 1.02]	0.50 [0.20, 1.18]	1.17	$0.39^{\downarrow}$
StyleGAN-XL (Sauer et al., 2022)	GAN	1.10	0.39	100k	103	1.03 [0.84, 1.26]	0.60 [0.26, 1.32]	1.03	$1.86^{\uparrow}$
DiT-cfg1.5 (Peebles & Xie, 2023)	LDM	2.27	0.64	50k	49	0.98 [0.73, 1.30]	0.60 [0.19, 1.77]	$0.72^{\downarrow}$	0.74
DiT-cfg1.0‡	LDM	9.60	0.53	50k	32	$0.64^{\downarrow}$ [0.45, 0.91]	0.20 [0.03, 1.13]	$0.64^{\downarrow}$	$0.40^{\downarrow}$

† EDM's network under a stochastic-churn sampler, $M = 15{,}104$. ‡ DiT-XL/2's weights with classifier-free guidance off.

of 0.24–0.75 at 10^{-3} and 0.93–2.42 at 10^{-4} (App. G). Published FID values come from different resolutions and protocols, so we also measure the Fréchet MAE distance (FMD) between MAE ViT-B/16 features of the audited 32^2 samples and one fixed real reference, whose real-versus-real floor is at least 14 times below the smallest model FMD. The two quality axes rank the configurations differently (Spearman $+0.11$). On both, configurations of similar bulk quality differ widely in tail rate: EDM and DiT-cfg1.5 have the same FMD (0.64) and ratios 1.81 and 0.98 (Fig. 1; rank statistics in App. H.3).

Verdicts depend on the functional and on the sampling configuration. The verdicts of the three functionals disagree (Fig. 10). Pairwise verdict agreement is $3/8$, $4/8$ and $3/8$, and Kendall's τ_b between R_1 and R_2 is -0.36 (descriptive at $n = 8$; all pairs in App. H). The clearest case is iDDPM, which under-produces under R_1 (0.17 $[0.11, 0.27]$) and over-produces under R_2 (1.55 $[1.31, 1.83]$). StyleGAN-XL and EDM2-M are calibrated under R_1 and R_2 but produce R_3-extreme images at 1.86 and 0.39 times the reference rate. Disabling classifier-free guidance on the same DiT-XL/2 weights moves $\hat{\rho}_{R_1}(10^{-3})$ from 0.98 to 0.64 $[0.45, 0.91]$ and also lowers the R_2 and R_3 estimates, whereas replacing EDM's Heun sampler by stochastic churn leaves it over-producing (1.81 and 1.99).

Recalibration. Within each sample-budget group the sorted wrapped severities are identical across configurations (maximum absolute difference 0.00), as §4 predicts. The wrapped ratio is 0.990 at 10^{-3} and 0.600 at 10^{-4} for all five configurations with $M = 100$k, 0.980 and 0.600 for both DiT configurations ($M = 50$k), and 0.927 and 0.662 for EDM-churn (Fig. 3), the value of $e(p)$ on the rank grid of each M. Downstream, plug-in rare-event estimates from raw synthetic corpora stay biased by the factor ρ at every corpus size $n \in \{10^3, \ldots, 10^5\}$ (Lemma 1; iDDPM 0.16–0.19 and EDM 1.73–1.86 at 10^{-3}, real-data control 0.97–1.01), and are 0.96–1.01 after wrapping (App. G).

Effect of wrapping on FID. We regenerated the samples of seven configurations (all but EDM-churn) from their checkpoints with the same seeds, which reproduces every count of Table 1, wrapped the images and rounded them to 8 bits. The four configurations whose intervals excluded 1 under R_1 then have image-realized $\hat{\rho}(10^{-3})$ of 0.98–0.99 and are labeled calibrated. Inception FID, measured on a fixed 50k subsample against 540,583 held-out real images, changes by at most 0.0024 (EDM, 2.357 $\rightarrow$ 2.359), about 7% of the resampling standard deviation of the FID itself (0.033–0.047). KID and FMD change by at most 2.7×10^{-6} and 1.6×10^{-3}. Transporting all M draws gives the same audited rates and raises FID by 0.05–0.19. In the three configurations where the tail-only change is resolved from zero (iDDPM, EDM, EDM2-S) the full wrap costs 48 to 94 times more, of the order of $1/\pi = 100$ that the discussion after Proposition 4 suggests. An FID that barely moves while $\hat{\rho}$ goes from 0.17 to 0.98 shows on real samplers the insensitivity that Proposition 1 describes, and for the same reason it is no evidence that the transported images are intact.

App. E examines them directly: the median SSIM of the transported images is at least 0.98 in every configuration, and their improved precision changes by at most 1.8 points.

Conditional tail dispersion remains reduced. Proposition 3(ii) leaves conditional tail shape unchanged, so a defect of shape would survive recalibration. We find one. On the principal-component basis of the calibration tail, at rank-matched depths, the PC1 dispersion ratio $\hat{\sigma}_{\mathrm{gen}}/\hat{\sigma}_{\mathrm{real}}$ lies in 0.40–0.71 for all eight configurations at both $p = 10^{-3}$ and $p = 10^{-2}$. Of the 32 cells, 25 reject equal dispersion under the Brown–Forsythe test at Holm-adjusted level 0.05, including 15 of the 16 PC1 cells. A permutation test of equal distributions rejects in 28 (App. H.7). The second component is less affected (0.68–1.00). The reduction is unrelated to rate calibration. StyleGAN-XL is calibrated under R_1 and has the smallest PC1 ratio, 0.40 $[0.34, 0.46]$. A sampler can therefore produce rare events at the right rate with too little variety among them.

Self-validation. We audit real data in the sampler's role, where $\rho = 1$. With the functional and $\widehat{F}$ held fixed and only the split redrawn ($B = 400$), which is the setting Eq. (1) addresses, the quadrature interval covers 1 at rates $0.950/0.953/0.958$ at the three depths, against $0.933/0.918/0.928$ for Wilson alone. When the functional and $\widehat{F}$ are also refitted for each of five split seeds, the interval covers 1 in 12 of 15 cells (11 for Wilson alone). Refitting spreads $\hat{\rho}_{\mathrm{self}}(10^{-3})$ over 0.89–1.18 against a split-only design effect of 1.05–1.10, a source of variation outside the interval that we report with it (App. D).

Validation and sensitivity studies. Further studies are in Apps. F and G. In the corpus-relative planted-ratio design, which isolates the counting error, with planted $\rho \in \{0.3, 1, 2, 5\}$, the estimator is unbiased to 0.5% and coverage is 0.94–0.96. Under four variants of the basis dimension and the data split no verdict changes between under- and over-production, and 26 of 32 verdicts (81%) are retained, every change being to an adjacent label. In the component ablations the generalized-Pareto splice gives no gain in point accuracy over the empirical quantile at checkable depths, so we use it only for profile-likelihood intervals and labeled extrapolation beyond the corpus floor (Apps. C and G).

Replication on CelebA-HQ. We reran the protocol unchanged on CelebA-HQ-256 (Karras et al., 2017), with $10{,}000$ calibration and $20{,}000$ held-out images, R_1 and $\widehat{F}$ refitted, StyleGAN2-ADA (Karras et al., 2020) drawn at $M = 100{,}000$ and DDPM (Ho et al., 2020) under DDIM-100 (Song et al., 2020) at $M = 30{,}000$. The held-out split certifies $p \in \{10^{-2}, 3 \times 10^{-3}, 10^{-3}\}$. StyleGAN2-ADA is calibrated at all three depths ($\hat{\rho} = 1.02$, 0.90, 0.95) and DDPM over-produces at all three (3.60, 4.10, 5.60, lower interval limits at least 2.99), the largest over-production in this paper. At $p = 10^{-4}$ the corpus is too small for an empirical threshold and we assign no verdict. Wrapped ratios follow that split's calibration-law ratio $e(p) = 1.050/0.803/0.604$ to within 0.007. With $20{,}000$ held-out images the fitted $\widehat{F}$ under-produces the held-out tail by about 40% at 10^{-3} ($e = 0.604$), and Proposition 3(i) guarantees no more than that (App. I).

6 Related Work

Evaluation of generative models. FID fits Gaussians in Inception space (Heusel et al., 2017), and FD-DINOv2 changes the backbone while remaining a moment metric (Stein et al., 2023). Precision–recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019) and density–coverage (Naeem et al., 2020) separate fidelity from diversity, and criticism of these metrics (Kynkäänniemi et al., 2022; Stein et al., 2023) has led to kernel (Jayasumana et al., 2024) and feature-likelihood (Jiralerspong et al., 2023) alternatives. All are distribution-level summaries, and none is a function of the exceedance rate at a fixed depth. Proposition 1 applies to the Fréchet-style scores among them and Proposition 2 to all of them. The rarity score (Han et al., 2022) ranks individual samples by k-NN uncommonness, whereas we estimate a rate; neither determines the other.

Tail calibration. Allen et al. (2025) define tail calibration for probabilistic forecasts: a forecast CDF is tail-calibrated when its conditional exceedance behavior matches that of the outcomes as the threshold approaches the upper endpoint. We adopt their decomposition into occurrence and severity; our ρ is an occurrence ratio at a finite depth. The audited object differs. Their forecast law

is known, so $1-F(t)$ can be evaluated directly and guarantees hold in the limit of high thresholds. A sampler is reachable only through draws, which requires finite-sample intervals at fixed depths and a check of the audit's own error. Their Proposition 13(b), that PIT calibration neither implies nor is implied by tail calibration, is the forecasting counterpart of our §2. Wessel et al. (2026) enforce tail calibration during training with threshold-weighted scoring rules, which needs the loss, gradients and retraining. Gneiting & Resin (2023) connect PIT-based and tail notions of calibration.

Recalibration and transport. Post-hoc recalibration of scalar predictors is classical (Platt et al., 1999; Zadrozny & Elkan, 2002; Guo et al., 2017). Our calibration/held-out design follows split-based conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023), although we have no paired ground truth per sample and certify a ratio of rates between distributions. The closest precedent for the wrapper is quantile-delta mapping with parametric tails in climate downscaling (Cannon et al., 2015). We apply the idea to samplers, where the correction has to be realized on the generated samples, and add the exactness result, the DKW control of estimation error (Massart, 1990), the ceiling result and the Fréchet-cost bound (Gelbrich, 1990). The generalized-Pareto tail relies on the peaks-over-threshold limit (Balkema & De Haan, 1974; Pickands III, 1975; Coles et al., 2001). Lipschitz pushforwards of light-tailed latents cannot produce heavier tails (Jaini et al., 2020), one reason to repair the tail outside the model.

Rare events in applications. Related questions arise for climate emulators (Stamatelopoulos & Sapsis, 2025; Manshausen et al., 2026), language models (Parulekar & Liu, 2026) and rare-event synthetic data (Gu et al., 2025), without black-box intervals, a self-check, or a repair. Two recent papers address different objects: Zhao et al. (2026) stabilize tail-risk estimates across training checkpoints, and Smith et al. (2025) fine-tune weights so that bulk sampling statistics satisfy distributional constraints. Simulation-based safety validation uses generative scenario models at the depths we audit (O'Kelly et al., 2018; Corso et al., 2021; Liu & Feng, 2024); Lemma 1 quantifies the consequence for such users.

7 Discussion

What to report. A tail-calibration result should state the severity functional, the sampling configuration, the reference corpus and the depth, since the verdict changed along each of these, together with the audit's own error: the self-validation coverage at checkable depths and the distribution of $e(p)$ beyond them. Exceedance rates can be repaired from outside the model, up to that error and at a cost in FID that Proposition 4 bounds and that we measured to be below the metric's resampling noise. Tail diversity is a different matter. The reduced dispersion appears in calibrated and miscalibrated configurations alike, Wrap leaves it unchanged, and by Corollary 1 a severity-keyed post-processor can only mix the sampler's own conditional shapes along the coordinates its transport preserves.

Choosing a severity functional. Since the verdict depends on R, the functional should be fixed before the audit, from the application that consumes the samples, and reported with the result as a loss is reported with a risk. An audit requires a reference corpus large enough for the depth of interest; recalibration requires in addition a transport, for which §4 gives three constructions. Without an application-specified R, we recommend reporting a panel of functionals and treating their disagreement as a result.

Limitations. We audit one functional at a time, and each of ours is a low-level notion of severity. All statements concern the population from which the reference corpus is drawn, and smaller corpora limit the auditable depths (10^{-3} on CelebA-HQ, where $e(p) = 0.60$). Beyond the corpus floor our numbers are generalized-Pareto extrapolations with propagated intervals, and the splice gave no gain in point accuracy at checkable depths. The threshold term of the interval assumes a locally constant ratio (Assumption 2) and is understated by the factor η_p otherwise, coverage falls to 85–87% for strongly under-producing samplers, and the interval does not include the variation from refitting the functional, under which self-validation covers $\rho = 1$ in 12 of 15 cells. Proposition 4 bounds the change in a Fréchet score, not the perceptual quality of individual transported images, which we assessed by proxies and without a human study. We report no audit outside images. Learned severity functionals and corpora that certify $p = 10^{-5}$ are the next steps.

AI USE STATEMENT

Large language models (Claude, Anthropic; ChatGPT, OpenAI) were used in this work solely to aid and polish the writing. Working from the authors' draft, the models suggested clearer and more concise wording, helped reorganize paragraphs and redistribute material between the main text and the appendices, standardized terminology, notation, and spelling, refined formatting, shortened figure and table captions, and flagged textual inconsistencies between the prose and the tables or LaTeX source. The authors formulated the research questions, conducted the theoretical analysis, designed the experiments, and analyzed the results. Every formal statement, proof, and reported numerical value was verified by the authors. The authors reviewed each edit suggested by the models, confirmed that it left the technical claims and reported numerical values unchanged, and take full responsibility for the content of this paper.

ETHICS STATEMENT

The audit informs the use of synthetic data by identifying discrepancies in the frequency of rare events, and the same information shows where a sampler could be pushed to produce more of them. The wrapper can raise the rate at which severe samples appear. The severity functionals used here measure statistical atypicality and should not be read as measures of harm. No human subjects are involved, and all corpora and checkpoints we use are publicly released. CelebA-HQ contains images of real people; we use the public release for research and show no face images from it.

REPRODUCIBILITY STATEMENT

Appendix J documents the datasets, checkpoints, preprocessing, sampler settings, seeds, software environment and computational cost, and gives the full constants of the calibration law. All corpora and checkpoints are public, and sampling is seeded per configuration. Proofs of all formal statements are in Appendices A and B. Appendix D reports self-validation across data splits. Code and result files are provided as anonymized supplementary material.

REFERENCES

Sam Allen, Jonathan Koh, Johan Segers, and Johanna Ziegel. Tail calibration of probabilistic forecasts. *Journal of the American Statistical Association*, 120(552):2796–2808, 2025.

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

August A Balkema and Laurens De Haan. Residual life time at great age. *The Annals of probability*, pp. 792–804, 1974.

Lawrence D Brown, T Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical science*, pp. 101–117, 2001.

Alex J Cannon, Stephen R Sobie, and Trevor Q Murdock. Bias correction of gcm precipitation by quantile mapping: how well do methods preserve changes in quantiles and extremes? *Journal of Climate*, 28(17):6938–6959, 2015.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Stuart Coles, Joanna Bawa, Lesley Trenner, and Pat Dorazio. *An introduction to statistical modeling of extreme values*, volume 208. Springer, 2001.

Anthony Corso, Robert Moss, Mark Koren, Ritchie Lee, and Mykel Kochenderfer. A survey of algorithms for black-box safety validation of cyber-physical systems. *Journal of Artificial Intelligence Research*, 72:377–428, 2021.

David C Dowson and BV Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

Tilmann Gneiting and Johannes Resin. Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination. *Electronic Journal of Statistics*, 17(2):3226–3286, 2023.

Jingyi Gu, Xuan Zhang, and Guiling Wang. Beyond the norm: A survey of synthetic data generation for rare events. *arXiv preprint arXiv:2506.06380*, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of lipschitz triangular flows. In *International Conference on Machine Learning*, pp. 4673–4681. PMLR, 2020.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

Marco Jiralerspong, Joey Bose, Ian Gemp, Chongli Qin, Yoram Bachrach, and Gauthier Gidel. Feature likelihood divergence: evaluating the generalization of generative models using samples. *Advances in Neural Information Processing Systems*, 36:33095–33119, 2023.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184. IEEE, 2024.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of imagenet classes in fr\'echet inception distance. *arXiv preprint arXiv:2203.06026*, 2022.

Henry X Liu and Shuo Feng. Curse of rarity for autonomous vehicles. *nature communications*, 15 (1):4808, 2024.

Peter Manshausen, Noah Brenowitz, Julius Berner, Karthik Kashinath, and Mike Pritchard. Towards accurate extreme event likelihoods from diffusion model climate emulators. *arXiv preprint arXiv:2605.03802*, 2026.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pp. 1269–1283, 1990.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Matthew O'Kelly, Aman Sinha, Hongseok Namkoong, Russ Tedrake, and John C Duchi. Scalable end-to-end autonomous vehicle testing via rare-event simulation. *Advances in neural information processing systems*, 31, 2018.

Nikita Y Parulekar and Anqi Liu. Estimating rare events in language models with proper evaluation. *arXiv preprint arXiv:2607.18454*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

James Pickands III. Statistical inference using extreme order statistics. *the Annals of Statistics*, pp. 119–131, 1975.

John Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 2018.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022.

Henry D Smith, Nathaniel L Diamant, and Brian L Trippe. Calibrating generative models to distributional constraints. *arXiv preprint arXiv:2510.10020*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Stamatis Stamatelopoulos and Themistoklis P Sapsis. Can diffusion models capture extreme event statistics? *Computer Methods in Applied Mechanics and Engineering*, 435:117589, 2025.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Jakob Benjamin Wessel, Maybritt Schillinger, Frank Kwasniok, and Sam Allen. Enforcing tail calibration when training probabilistic forecast models. *International Journal of Forecasting*, 2026.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Ming-Ju Wu, Jyh-Shing R Jang, and Jui-Long Chen. Wafer map failure pattern recognition and similarity ranking for large-scale data sets. *IEEE Transactions on Semiconductor Manufacturing*, 28(1):1–12, 2014.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 694–699, 2002.

Shuning Zhao, Patrick Wong, Leran Zhang, and Xiaolin Hu. Diachronic sample integration: Robust tail-risk estimation with generative models. *arXiv preprint arXiv:2607.10810*, 2026.

A PROOFS

A.1 PROPOSITION 1 (STRUCTURAL BLINDNESS OF FINITE-MOMENT METRICS)

Proof. Fix the depth p, write $\tau = \tau_p$, and let Q_T (supported in T_τ) and Q_B (supported in B_τ) be the laws supplied by Assumption 1, so that $\mathbb{E}_{Q_T}\psi = \mathbb{E}_{Q_B}\psi$. Let c satisfy $1 < c \le 1/p$ and set

$$\varepsilon = c\,p \ (\le 1 \text{ by the quantifier}), \qquad Q_1 = (1-\varepsilon)Q_0 + \varepsilon Q_B, \qquad Q_2 = (1-\varepsilon)Q_0 + \varepsilon Q_T.$$

Both are probability laws because $\varepsilon \in (0, 1]$, which is where the restriction $c \le 1/p$ is used. Their moment vectors coincide,

$$\mathbb{E}_{Q_1}\psi = (1-\varepsilon)\mathbb{E}_{Q_0}\psi + \varepsilon\mathbb{E}_{Q_B}\psi = (1-\varepsilon)\mathbb{E}_{Q_0}\psi + \varepsilon\mathbb{E}_{Q_T}\psi = \mathbb{E}_{Q_2}\psi,$$

so $\mathcal{M}(P, Q_1) = m(\mathbb{E}_{Q_1}\psi) = m(\mathbb{E}_{Q_2}\psi) = \mathcal{M}(P, Q_2)$ for every function m. Their tail masses differ by exactly ε, since $Q_B(T_\tau) = 0$ and $Q_T(T_\tau) = 1$:

$$Q_2(T_\tau) - Q_1(T_\tau) = \varepsilon\big(Q_T(T_\tau) - Q_B(T_\tau)\big) = \varepsilon.$$

Dividing by $P(T_{\tau_p}) = p$ (Definition 1) gives $\rho_2(\tau_p) - \rho_1(\tau_p) = \varepsilon/p = c$. □

The gap is attained with equality, and $c \le 1/p$ is the largest gap a mixture of weight $\varepsilon \le 1$ can produce. The planted-ratio simulation of App. F enforces the same constraint when it plants a ratio ρ by drawing exceedances at rate $\rho\,p$. The proof uses no property of ψ beyond Assumption 1 and no property of m. It therefore applies to the Fréchet distance, to any reweighting of the feature moments, and to any monotone transform of the resulting score.

Remark 5 (Checking Assumption 1). Carathéodory's theorem reduces mixability to a finite check. If some point of $\psi(T_\tau)$ lies in the convex hull of $d{+}1$ points of $\psi(B_\tau)$, then a $(d{+}1)$-atom bulk mixture Q_B matches the moment vector of a tail atom and the assumption holds. If it fails, the convex hulls of $\psi(T_\tau)$ and $\psi(B_\tau)$ are disjoint, and two disjoint convex sets in $\mathbb{R}^d$ are separated by a hyperplane: some affine functional of the feature moments is at least a constant on the tail and at most that constant on the bulk. An example is a severity that is a Mahalanobis radius in the metric's own feature space when ψ contains the second moments: R^2 is then linear in ψ and $\mathbb{E}_{Q_T}R^2 \ge \tau^2 > \mathbb{E}_{Q_B}R^2$, so no mixable pair exists. Mixability is thus a property of the pair (ψ, R). Among the pairs in this paper it excludes (FMD, R_2), which are computed from the same MAE features. Section A.7 tests the assumption numerically for a pixel-space severity and a deep-feature metric.

A.2 PROPOSITION 2 (SAMPLE-SIZE BLINDNESS OF ANY TEST)

Proof. Since Q_1 and Q_2 agree conditionally on the tail indicator $\mathbf{1}\{R \ge \tau_p\}$, the indicator is a sufficient statistic for the pair, and any test based on n i.i.d. draws can be taken to depend on the draws only through the n indicators. The problem reduces to distinguishing n Bernoulli trials with success probability p from n with probability cp. The squared Hellinger distance of the Bernoulli pair is

$$H^2 = \big(\sqrt{cp} - \sqrt{p}\big)^2 + \big(\sqrt{1-cp} - \sqrt{1-p}\big)^2.$$

The first term is $p(\sqrt{c} - 1)^2$. For the second, rationalize:

$$\big(\sqrt{1-cp} - \sqrt{1-p}\big)^2 = \frac{\big((1-p) - (1-cp)\big)^2}{\big(\sqrt{1-p} + \sqrt{1-cp}\big)^2} = \frac{(c-1)^2p^2}{\big(\sqrt{1-p} + \sqrt{1-cp}\big)^2} \ \le \ \frac{(c-1)^2p^2}{2},$$

where the last step uses $cp \le \frac{1}{2}$ and $c > 1$, hence $p \le cp \le \frac{1}{2}$, hence $\sqrt{1-p}, \sqrt{1-cp} \ge 1/\sqrt{2}$ and the denominator is at least 2. Writing $(c-1)^2p^2 = p(\sqrt{c}-1)^2 \cdot p(\sqrt{c}+1)^2$ and noting

$$p(\sqrt{c}+1)^2 = cp\Big(1+\tfrac{1}{\sqrt{c}}\Big)^2 \le \tfrac{1}{2}\cdot 4 = 2 \qquad (c>1,\ cp \le \tfrac{1}{2}),$$

the second term is at most $p(\sqrt{c}-1)^2$, so $H^2 \le 2p(\sqrt{c}-1)^2$. Tensorizing, $H^2\big(Q_1^{\otimes n}, Q_2^{\otimes n}\big) = 2\big(1-(1-H^2/2)^n\big) \le nH^2$, and since total variation is dominated by the Hellinger distance, $\mathrm{TV} \le H\big(Q_1^{\otimes n}, Q_2^{\otimes n}\big) \le \sqrt{nH^2}$. Le Cam's two-point bound says the error probability of any test is at least $\frac{1}{2}(1-\mathrm{TV})$, so an error probability below $1/3$ forces $\mathrm{TV} \ge 1/3$ and hence $n\,H^2 \ge 1/9$. Combining with $H^2 \le 2p(\sqrt{c}-1)^2$ gives $n \ge C/\big(p(\sqrt{c}-1)^2\big)$ with $C = 1/18$. □

With $C = 1/18$ the bound reads $n \ge 3{,}238$ at $c = 2$, $p = 10^{-4}$ and $n \ge 324$ at $p = 10^{-3}$, and it exceeds 5×10^4 only for $p \lesssim 6.5 \times 10^{-6}$. These values illustrate the lower bound. Because it is a lower bound on n, meeting it does not show that a given error probability is achievable at a given budget. The statements about the 50k budget in the main text rest on the interval widths measured in App. G; from this proposition we use only the scaling $n \asymp p^{-1}(\sqrt{c}-1)^{-2}$ and the fact that it holds for every statistic.

A.3 Lemma 1 (downstream inheritance)

Proof. The plug-in estimator is $\hat{p}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{R(X_i) \ge \tau\}$ with $X_i \sim Q$ i.i.d., so by linearity of expectation $\mathbb{E}\,\hat{p}_n = Q\{R \ge \tau\} = \rho(\tau)\,P\{R \ge \tau\}$ for every n. As an estimator of $P\{R \ge \tau\}$, $\hat{p}_n$ therefore has multiplicative bias factor $\rho(\tau)$ (relative bias $\rho(\tau) - 1$) for every corpus size n. □

The estimator is unbiased for the sampler's exceedance probability and biased by the constant factor $\rho(\tau)$ for the data's. The factor does not depend on n, so collecting more synthetic data reduces the variance of the downstream estimate and leaves its bias unchanged.

A.4 Proposition 3 (wrapper guarantee)

Proof. (i) Continuity of F_Q gives $V = F_Q(R(X)) \sim \mathrm{Unif}(0,1)$, so $\widehat{F}^{-1}(V)$ has law $\widehat{F}$ by the standard inverse-transform identity for generalized inverses. The transport realizes the target severity exactly, $R(X') = R(\phi_{r\to r'}(X)) = r' = \widehat{F}^{-1}(V)$, so the severity law of $\tilde{Q}$ is $\widehat{F}$. Consequently, for the depth-p threshold τ_p of the data law (which satisfies $P\{R \ge \tau_p\} = p$),

$$\tilde{\rho}(\tau_p) = \frac{\tilde{Q}\{R \ge \tau_p\}}{P\{R \ge \tau_p\}} = \frac{1-\widehat{F}(\tau_p)}{p} = e(p),$$

an expression in which the sampler does not appear.

(ii) Condition on $R(X) = r$ and apply $\phi_{r\to T(r)}$ with $T = \widehat{F}^{-1} \circ F_Q$. Transport invariance gives $S(X') = S(X)$, while $R(X') = T(r)$, so the joint law of the wrapped pair is $(S, T(R))$. Under the assumption of part (ii) that T is strictly increasing on the range of R considered, the event $\{T(R) = s\}$ equals $\{R = T^{-1}(s)\}$, whence $\mathcal{L}\big(S \mid R(X') = s\big) = \mathcal{L}\big(S \mid R(X) = T^{-1}(s)\big)$. The argument does not depend on the particular $\widehat{F}$. The wrapper relabels the severity axis and leaves the conditional shape family unchanged.

(iii) Write $\varepsilon_M = \sup_v |\widehat{F}_Q(v) - F_Q(v)|$. For any continuity point t, using the Galois property of the generalized inverse and $F_Q(R) \sim \mathrm{Unif}(0,1)$,

$$\Pr\big(\widehat{F}^{-1}(\widehat{F}_Q(R)) \le t\big) = \Pr\big(\widehat{F}_Q(R) \le \widehat{F}(t)\big) \in \big[\widehat{F}(t) - \varepsilon_M,\ \widehat{F}(t) + \varepsilon_M\big],$$

so the wrapped severity law deviates from $\widehat{F}$ by at most ε_M uniformly. Dvoretzky–Kiefer–Wolfowitz with Massart's constant (Massart, 1990) gives $\Pr\big(\varepsilon_M > \sqrt{\log(2/\delta)/2M}\big) \le \delta$. This bound concerns the transform estimated from M draws and applied to a new draw. Applied to the same M draws, the rank remap produces the deterministic grid $\widehat{F}^{-1}\big(i/(M+1)\big)$, whose discretization is discussed in the next remark. □

Remark 6 (The rank convention). The implementation replaces $\widehat{F}_Q$ by $v_i = \mathrm{rank}(r_i)/(M{+}1)$. Since $\mathrm{rank}(r_i)/(M{+}1) = \frac{M}{M+1}\widehat{F}_Q(r_i)$, the two differ by at most $1/(M{+}1)$ uniformly, so part (iii) holds with ε_M replaced by $\varepsilon_M + 1/(M{+}1)$. On the ρ scale at depth p the grid has step $1/(Mp)$, which is 0.1 at $M = 10^5$ and $p = 10^{-4}$ and 0.66 for EDM-churn ($M = 15{,}104$). This discretization is visible at the deepest audited depth, and we therefore report the rank-grid value e_{grid} separately from the continuous-law value of $e(p)$ (App. D). The convention avoids $v = 1 \mapsto +\infty$ when the splice has a heavy tail, and every wrapped number we report uses it.

A.5 Corollary 1 (ceiling for severity-keyed post-processing)

Corollary 1 (Shape ceiling for severity-keyed post-processing)**.** *Let Φ read a sample only through its severity: $\Phi(X) = \phi_{R(X)\to g(R(X),U)}(X)$ for measurable g and $U \perp X$. Then the severity law of $\Phi(X)$ depends on Q only through F_Q, and for every transport-invariant S and almost every s, the conditional shape law of $\Phi(X)$ at severity s is a mixture of the sampler's own conditional shape laws. No such Φ produces a conditional shape law outside the closed convex hull of the sampler's family. If g is deterministic and strictly increasing, as for* Wrap*, each output conditional shape law equals one of the sampler's own. Within the class,* Wrap *attains the target severity law exactly by Proposition 3(i).*

Proof. Let $\Phi(X) = \phi_{R(X)\to g(R(X),U)}(X)$ with g measurable and $U \perp X$. By the defining property of the transport, $R(\Phi(X)) = g(R(X), U)$, whose law is the pushforward of $\mathcal{L}(R(X))\otimes\mathcal{L}(U)$ under g: it depends on the sampler only through F_Q, which is the first claim. For the second, transport invariance gives $S(\Phi(X)) = S(X)$, so for any s

$$\mathcal{L}\big(S(\Phi(X)) \mid R(\Phi(X)) = s\big) = \int \mathcal{L}\big(S(X) \mid R(X) = r\big)\, \nu_s(\mathrm{d}r),$$

where ν_s is a regular conditional distribution of $R(X)$ given $g(R(X), U) = s$, defined for almost every s. The identity uses that U is independent of X, so that given $R(X) = r$ the shape $S(X)$ is independent of $g(r, U)$. The output conditional shape law is therefore a mixture of the sampler's own conditional shape laws and lies in their closed convex hull. If $g(r, u) = T(r)$ with T strictly increasing, ν_s is the point mass at $T^{-1}(s)$ and the mixture has a single component, which is Proposition 3(ii). Finally, the deterministic choice $g(r, \cdot) = \widehat{F}^{-1}(F_Q(r))$ used by Wrap attains the target law $\widehat{F}$ exactly by Proposition 3(i). □

The class in the corollary is the class of post-processors that can be built from an audit's output, since the audit reports only severities. A randomized member can reweight the sampler's conditional shape laws across severities, but it cannot produce a shape law that lies outside their closed convex hull.

A.6 Proposition 4 (Fréchet cost of post-processing)

For a law μ on $\mathbb{R}^D$ with mean m_μ and covariance Σ_μ write $N_\mu = \mathcal{N}(m_\mu, \Sigma_\mu)$. The Fréchet distance used by FID is the 2-Wasserstein distance between Gaussian fits (Dowson & Landau, 1982),

$$d_F(P, Q) = W_2\big(N_{\varphi_\# P}, N_{\varphi_\# Q}\big), \qquad d_F^2 = \|m_P - m_Q\|^2 + \mathrm{tr}\Big(\Sigma_P + \Sigma_Q - 2\big(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}\big)^{1/2}\Big).$$

Proof. W_2 is a metric on laws with finite second moments, so its restriction to Gaussian laws satisfies the triangle inequality, and $|d_F(P, \tilde{Q}) - d_F(P, Q)| \le d_F(Q, \tilde{Q})$. By Gelbrich's bound (Gelbrich, 1990), for any two laws μ, ν with finite second moments $W_2(N_\mu, N_\nu) \le W_2(\mu, \nu)$. Applying it to $\mu = \varphi_\# Q$ and $\nu = \varphi_\# \tilde{Q}$ gives $d_F(Q, \tilde{Q}) \le W_2(\varphi_\# Q, \varphi_\# \tilde{Q})$. The pair $(\varphi(X), \varphi(X'))$ with $X \sim Q$ and $X' = \Phi(X)$ is a coupling of these two laws, and W_2 is the infimum of the root-mean-square cost over couplings, so $W_2(\varphi_\# Q, \varphi_\# \tilde{Q}) \le (\mathbb{E}\|\varphi(X') - \varphi(X)\|^2)^{1/2}$. If $\Phi(X) = X$ outside an event A with $Q(A) = \pi$, then $\mathbb{E}\|\varphi(X') - \varphi(X)\|^2 = \pi\, \mathbb{E}[\|\varphi(X') - \varphi(X)\|^2 \mid A] = \pi D^2$. □

The argument uses no property of Q beyond finite second moments of the features, so it applies to the empirical law of M draws and their images under Φ, with the expectation replaced by the sample

mean (and covariances normalized by M). In terms of FID itself, with $\delta = \sqrt{\pi} D$,

$$\big|\mathrm{FID}(P, \tilde{Q}) - \mathrm{FID}(P, Q)\big| \le 2\, d_F(P, Q)\, \delta + \delta^2.$$

Remark 7 (Order of the moment shift). The bound above is of order $\sqrt{\pi}$ and is a worst case over all laws with the given displacement. The moments themselves move by order π. Writing $\Delta = \varphi(X') - \varphi(X)$, the means satisfy $m_{\tilde{Q}} - m_Q = \pi\, \mathbb{E}[\Delta \mid A]$, hence $\|m_{\tilde{Q}} - m_Q\| \le \pi D$, and

$$\Sigma_{\tilde{Q}} - \Sigma_Q = \pi\, \mathbb{E}\big[\varphi(X')\varphi(X')^\top - \varphi(X)\varphi(X)^\top \mid A\big] - \big(m_{\tilde{Q}} m_{\tilde{Q}}^\top - m_Q m_Q^\top\big),$$

each term of which is of order π when the features have bounded second moments on A. Where d_F^2 is differentiable in (m, Σ), which holds when $\Sigma_P^{1/2} \Sigma_Q \Sigma_P^{1/2}$ is positive definite, the change in FID is therefore of order π. This is the scaling compared with measurements in §5 and App. E.

Remark 8 (Severity in the metric's own feature space). Suppose R is the whitened radius on the top K principal components (λ_k, v_k) of the features φ used by the metric, and the transport is the radial rescaling of the whitened coordinates. Then $\|\varphi(X') - \varphi(X)\| \le \sqrt{\lambda_1}\, |r' - r|$, and because $T = \widehat{F}^{-1} \circ F_Q$ is the monotone rearrangement,

$$\mathbb{E}\|\varphi(X') - \varphi(X)\|^2 \le \lambda_1 \int_{v_0}^1 \big(\widehat{F}^{-1}(v) - F_Q^{-1}(v)\big)^2 \,\mathrm{d}v$$

when only ranks above v_0 are transported. For $v_0 = 0$ the integral is the squared one-dimensional 2-Wasserstein distance between the sampler's severity law and the calibration law. In this setting the Fréchet cost of the wrapper is controlled by the miscalibration it removes. This is the setting of R_2 with FMD; we did not run the image-level wrapper under R_2.

A.7 Empirical illustration of moment matching

This subsection reports the numerical construction behind Fig. 2 and the scope statement of Remark 2. All quantities are measured in MAE ViT-B/16 feature space (the features of R_2), and the tail is defined by R_1 at $p = 10^{-3}$.

Construction. Q_T is uniform on the 536 held-out real images above $\tau_{10^{-3}}$. We minimize $\|\sum_i w_i \psi(x_i) - \mathbb{E}_{Q_T}\psi\|$ over the simplex on a pool of N bulk candidates. The minimum falls from 1.59×10^{-1} at $N = 5{,}000$ to 9.9×10^{-3} at $N = 540{,}048$, which is every eligible held-out real image below τ. It decays as $N^{-0.59}$ with no visible floor and is attained on 482 atoms, within the Carathéodory bound $d{+}1 = 769$. This is consistent with $\mathbb{E}_{Q_T}\psi$ lying in the closure of the hull; it does not prove that it lies in the hull. The law Q_B used for Fig. 2 and for the metric comparisons below comes from a separate sparse solve and has 529 atoms, with mean residual 9.7×10^{-3} (9.4×10^{-4} relative to $\|\mathbb{E}_{Q_T}\psi\|$).

Residuals. The mean block of the Fréchet moment map is matched to within 0.27% of the tail law's displacement from the real mean. The second-moment block $\psi\psi^\top$ is matched to 8.8% in Frobenius norm (7.8% for a variant that targets the second moments). We attribute this to our first-order solver on a 540,048-variable problem; infeasibility is not demonstrated, since 15.4% of bulk candidates exceed the tail law's mean whitened energy. Fig. 2 is accordingly described as sharing a moment vector up to a reported residual, and its numbers are measured in MAE feature space and not in Inception space.

Metric comparisons. With $Q_0 =$ EDM at $M = 10^5$ the two perturbed samplers differ in $\rho(10^{-3})$ by exactly $c \in \{2, 5\}$, while their Fréchet distances differ by 3.2×10^{-5} and 1.1×10^{-4} (5.4×10^{-5} and 1.6×10^{-4} with $Q_0 =$ StyleGAN-XL). An unmatched control at the same ε moves the metric 193 times further. For the mean-embedding distance $\|\mathbb{E}_Q\varphi - \mathbb{E}_P\varphi\|^2$, whose moment map is the one we match, the two samplers agree to a relative 4.4×10^{-7} and 1.2×10^{-6}.

Kernel contrast. On the same pair Q_B, Q_T, an RBF-kernel MMD with median-heuristic bandwidth has population $\mathrm{MMD}^2(Q_B, Q_T) = 4.15 \times 10^{-4}$, non-zero as the characteristic property requires. A permutation test separates the two laws at $n = 5{,}000$ draws per side ($p \le 0.04$ for

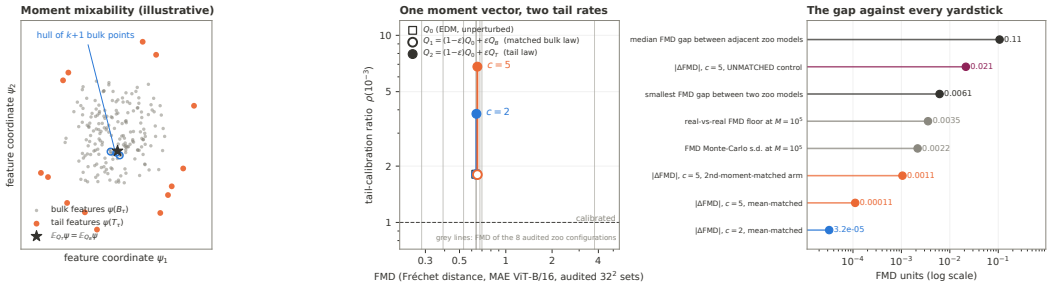

Figure 2: The construction of Proposition 1. **Left (schematic):** the moment vector of a tail law inside the convex hull of $d{+}1$ bulk feature vectors. **Center (computed):** the construction on real features (§A.7). Mixing either law into Q_0 at $\varepsilon = cp$ changes the tail-calibration ratio by exactly c; the change in Fréchet distance is not visible at this scale. Gray lines mark the FMD of the eight audited configurations. **Right:** the change in Fréchet distance compared with the metric's sampling noise, the real-versus-real floor, the smallest gap between two audited configurations, and an unmatched control at the same ε. The second-moment block is matched to 7.8–8.8%; the variant that trades mean accuracy for a closer second-moment match also stays below all four reference scales.

every bulk law we built), which the finite moment map cannot do. Matching the mean embedding lowers the kernel discrepancy by a factor of 168 and does not remove it. At the mixture level the kernel is limited by sample size: the population value $\mathrm{MMD}^2(Q_1, Q_2) = \varepsilon^2\,\mathrm{MMD}^2(Q_B, Q_T)$ is 1.7×10^{-9}–1.9×10^{-8}, and a permutation test at $n = 2 \times 10^4$ per side does not reject in three of four cells ($p = 0.51$–1.00). The fourth cell has $p = 0.047$ with a population discrepancy 4,700 times below the 95th percentile of the null, and we regard it as a chance rejection. Detection would need $n \approx 4 \times 10^7$–4×10^8. For kernel scores the applicable statement is thus Proposition 2, as Remark 2 says.

B Audit and Wrapper: Functionals, Algorithms, Transports, and Interval Construction

B.1 The three severity functionals and their transports

All three functionals are fitted on the 200,000-image calibration split $\mathcal{D}_{\mathrm{cal}}$ of ImageNet-32 (Chrabaszcz et al., 2017) and are reported separately.

R_1 **(pixel atypicality).** Let m be the pixel mean of $\mathcal{D}_{\mathrm{cal}}$ and (λ_k, v_k), $k \le K = 50$, the top eigenpairs of the pixel covariance. With $W = [v_1/\sqrt{\lambda_1}, \ldots, v_K/\sqrt{\lambda_K}]$ the whitened coordinates are $w = W^\top(x - m)$ and $R_1(x) = \|w\|_2$, the whitened radius on the retained subspace. The transport rescales the whitened coordinates, $w \mapsto (r'/r)w$, which in pixel space reads $x' = x + \sum_k \sqrt{\lambda_k}\,(r'/r - 1)\,w_k\,v_k$. The component of x orthogonal to the retained subspace is unchanged.

R_2 **(representation atypicality).** The same whitened radius, $R_2(x) = \|\Lambda_2^{-1/2} V_2^\top (e(x) - m_2)\|_2$, on the top $K_2{=}50$ components of $e(x) \in \mathbb{R}^{768}$, the mean of the 196 final-layer patch tokens of an MAE-pretrained ViT-B/16 (He et al., 2022) applied to the upsampled image. R_2 measures atypicality in this representation and makes no claim about human semantics. Because e is a nonlinear network, R_2 has no closed-form image-space transport; the constructions of §B.4 apply to it, and we do not wrap images under R_2.

R_3 **(worst-patch extremes).** On the 4×4 grid of $8 \times 8 \times 3$ patches, each position is whitened separately on $K_3{=}32$ components to a radius $r_j(x)$, and $R_3(x) = \max_{j \le 16} r_j(x)$. An explicit transport acts on the whitened coordinates of each patch, as for R_1. To raise the severity it scales the whitened coordinates of the argmax patch by r'/r_{j^*}. To lower it, it rescales every patch whose radius exceeds r' to radius exactly r'. We audit under R_3 but do not wrap images under it.

The transports of R_1 and R_3 reach their target severity exactly before the state-space constraint is imposed (Remark 4).

B.2 Algorithm 1: Audit

Algorithm 1 AUDIT. Lines 1–3 use real data only.

Require: sampler Q (black box); severity R; splits $\mathcal{D}_{\mathrm{cal}}, \mathcal{D}_{\mathrm{hold}}$; depths $p_1 > \cdots > p_L$

1: fit the severity functional on $\mathcal{D}_{\mathrm{cal}}$; then the adaptive generalized-Pareto tail $(u, \hat{\xi}, \hat{\sigma}, \zeta)$ of the calibration law on its excesses ▷ App. C
2: $\tau_\ell \leftarrow$ empirical upper-p_ℓ quantile of R on $\mathcal{D}_{\mathrm{hold}}$, or its generalized-Pareto extrapolation beyond the corpus floor
3: SELFVALIDATE($\mathcal{D}_{\mathrm{cal}}$) ▷ run *before* any model is judged
4: draw $X_1, \ldots, X_M \sim Q$; $r_i \leftarrow R(X_i)$
5: $\hat{\rho}(\tau_\ell) \leftarrow \#\{i : r_i \geq \tau_\ell\}/(Mp_\ell)$ with certificate (1); verdict = under / calibrated / over as that interval lies below, contains, or lies above 1
6: *shape audit*: project the sampler's rank-matched tail directions and the held-out real exceedance directions onto the calibration-tail principal-component basis; test dispersion equality per component

routine SELFVALIDATE($\mathcal{D}$): rerun lines 4–5 with $\mathcal{D}$ in the *model* role against the same τ_ℓ, reporting $\hat{\rho}_{\mathrm{self}}(p_\ell)$ with certificate at every checkable depth and the calibration-law ratio $e(p) = (1 - \widehat{F}(\tau_p))/p$

All fitted components use $\mathcal{D}_{\mathrm{cal}}$, and empirical audit thresholds use $\mathcal{D}_{\mathrm{hold}}$. Thresholds beyond the corpus floor come from the generalized-Pareto tail of App. C and are labeled as extrapolation wherever they appear.

B.3 Algorithm 2: Wrap

The sampler's severity law F_Q is unknown and is replaced by the rank transform of its own M draws. The implementation uses $v_i = \mathrm{rank}(r_i)/(M+1)$ in place of the raw empirical CDF $\widehat{F}_Q$. This avoids $v = 1 \mapsto +\infty$ under a heavy-tailed splice and shifts the uniform bound of Proposition 3(iii) by at most $1/(M+1)$. Every wrapped number we report uses this convention.

Algorithm 2 WRAP: black-box tail recalibration. It requires no weights, gradients or retraining; the sampler is used only as a source of draws.

Require: sampler Q; severity R with transport $\phi_{r \to r'}$ satisfying $R(\phi_{r \to r'}(x)) = r'$; calibration law $\widehat{F}$; draws $X_1, \ldots, X_M \sim Q$ with $r_i = R(X_i)$

1: $v_i \leftarrow \mathrm{rank}(r_i)/(M+1)$ ▷ plug-in for $F_Q(r_i)$
2: $r'_i \leftarrow \widehat{F}^{-1}(v_i)$ ▷ empirical quantile of $\mathcal{D}_{\mathrm{cal}}$ below u; $u + \frac{\hat{\sigma}}{\hat{\xi}}\big((\frac{1-v}{\zeta})^{-\hat{\xi}} - 1\big)$ above
3: $X'_i \leftarrow \phi_{r_i \to r'_i}(X_i)$ ▷ support-aware chart; project onto the state space if needed
4: **return** $X'_1, \ldots, X'_M$ ▷ rates read off $\{r'_i\}$; images returned when samples, not only rates, are needed

Which samples are transported. Line 3 of Algorithm 2 is stated for every draw. Every image-level result in this paper transports only the draws with $v_i > v_0 = 0.99$, the sampler's top 1% by severity rank, and returns the rest unchanged. For every level t above both $F_Q^{-1}(v_0)$ and $\widehat{F}^{-1}(v_0)$ the wrapped exceedance probability is then $1 - \widehat{F}(t)$, as under the full map, so Proposition 3(i) holds on that range. The region contains every draw whose raw or target severity reaches $\tau_{10^{-3}}$, and the audited rates are identical to those of the full map. App. E compares the two implementations: transporting all M draws changes no audited rate and raises FID by 0.05–0.19.

B.4 TRANSPORTS FOR A GENERAL SEVERITY FUNCTIONAL

Proposition 3(i) and (iii) require of the transport only that $R(\phi_{r\to r'}(x)) = r'$. Part (ii) concerns shape coordinates S with $S \circ \phi_{r\to r'} = S$. Of the three functionals above, R_1 and R_3 have explicit transports and R_2 does not. For a general R, including R_2, we describe three constructions that use only R and, for (b) and (c), its gradient.

(a) Chart. Suppose there is a bijection $x \leftrightarrow (r, s)$ with $r = R(x)$ and s ranging over a fixed set. Define $\phi_{r\to r'}(x) = x(r', s)$. The target severity is reached by construction and $S = s$ is transport-invariant. For R_1 the chart is $(r, u, x_\perp)$ with $r = \|w\|$, $u = w/\|w\|$ and $x_\perp$ the component of x outside the retained subspace. This is the preferred construction when a chart is known.

(b) Gradient flow. Let R be continuously differentiable with $\nabla R \neq 0$ on the region of interest and let G be a symmetric positive-definite matrix (a metric on $\mathcal{X}$). Define $\phi_{r\to r'}(x)$ as the solution at time $s = r' - r$ of

$$\dot{x}(s) = \frac{G^{-1}\nabla R(x)}{\nabla R(x)^\top G^{-1}\nabla R(x)}, \qquad x(0) = x. \tag{3}$$

Wherever the solution exists, $\frac{\mathrm{d}}{\mathrm{d}s}R(x(s)) = \nabla R^\top \dot{x} = 1$, so $R(x(s)) = r + s$, and the target is reached exactly at $s = r' - r$ provided the solution exists on the whole interval between 0 and $r' - r$ (negative times lower the severity). Local Lipschitz continuity of the vector field gives existence only locally; reaching a given r' additionally requires that the trajectory not leave the region where $\nabla R \neq 0$. With a locally Lipschitz field, flow lines do not cross, so the label of the flow line through x is a shape coordinate S with $S \circ \phi_{r\to r'} = S$, and Proposition 3(ii) applies with this S. The transports also compose, $\phi_{r'\to r''} \circ \phi_{r\to r'} = \phi_{r\to r''}$. In practice (3) is integrated numerically; the two-dimensional examples of App. H with a curved severity use a fourth-order Runge–Kutta scheme with 30 steps.

(c) Minimal displacement. Define $\phi_{r\to r'}(x) \in \arg\min_y\{\|y - x\|_G^2 : R(y) = r'\}$ with $\|z\|_G^2 = z^\top G z$. A minimizer exists when the level set $\{R = r'\}$ is nonempty and closed, which holds for continuous R and r' in its range, since $\|y - x\|_G^2$ is coercive. The constraint gives the target severity, so parts (i) and (iii) of Proposition 3 hold. Such projections do not compose in general and need not preserve any coordinate, so part (ii) has no counterpart. When a box constraint is added to the feasible set, this is the problem that the alternating projection of Remark 4 solves approximately.

Proposition 5 (Consistency with the closed form for R_1)**.** *Let $R = R_1$, write $x = m + V\Lambda^{1/2}w + x_\perp$ with $V = [v_1, \ldots, v_K]$, $\Lambda = \mathrm{diag}(\lambda_k)$ and $V^\top x_\perp = 0$, and let $G = V\Lambda^{-1}V^\top + (I - VV^\top)$ be the whitening metric, for which $\|\mathrm{d}x\|_G^2 = \|\mathrm{d}w\|^2 + \|\mathrm{d}x_\perp\|^2$. For every x with $w \neq 0$ and every $r' > 0$, constructions (b) and (c) with this G both give the radial rescaling $w \mapsto (r'/r)w$ with $x_\perp$ unchanged, which is the closed-form transport of §B.1.*

Proof. Since $R(x) = \|w\|$ and $w = \Lambda^{-1/2}V^\top(x - m)$, we have $\nabla R(x) = V\Lambda^{-1/2}u$ with $u = w/\|w\|$. Because $G^{-1} = V\Lambda V^\top + (I - VV^\top)$, it follows that $G^{-1}\nabla R = V\Lambda^{1/2}u$ and $\nabla R^\top G^{-1}\nabla R = u^\top u = 1$. The flow (3) is therefore $\dot{x} = V\Lambda^{1/2}u$, that is $\dot{w} = u$ and $\dot{x}_\perp = 0$. Then $\frac{\mathrm{d}}{\mathrm{d}s}\|w\| = u^\top \dot{w} = 1$ and $\dot{u} = \dot{w}/\|w\| - w\,(u^\top \dot{w})/\|w\|^2 = 0$. The direction u is constant and $w(s) = (r + s)u$, which at $s = r' - r$ equals $r'u = (r'/r)w$. For (c), $\|y - x\|_G^2 = \|w_y - w\|^2 + \|y_\perp - x_\perp\|^2$ and the constraint is $\|w_y\| = r'$. The minimum is attained at $y_\perp = x_\perp$ and at the point of the sphere of radius r' closest to w, which is $w_y = r'u$. □

The metric matters. With the Euclidean pixel metric $G = I$ the flow becomes $\dot{w} = \Lambda^{-1}u/(u^\top\Lambda^{-1}u)$. The severity still grows at unit rate, but the direction u rotates toward the low-variance components unless Λ is a multiple of the identity or u is an eigenvector. Both metrics give a valid transport for parts (i) and (iii). They preserve different shape coordinates, and the shape audit of §3 is defined for the direction u, which the whitening metric preserves. A transport can also be learned, for instance by training a conditional editing network to satisfy $R(\phi_{r\to r'}(x)) = r'$ under a fidelity penalty. We list only training-free constructions because the wrapper is meant to work without a second trained model.

B.5 THE TWO INTERVAL COMPONENTS

Count error: from the Wald interval to the Wilson interval. With $k = \#\{i : r_i \geq \tau_p\}$ the count is binomial, $k \sim \mathrm{Bin}(M, q)$ with $q = Q\{R \geq \tau_p\}$, and $\hat{q} = k/M$ has mean q and variance $q(1-q)/M$. By the central limit theorem

$$\frac{\hat{q} - q}{\sqrt{q(1-q)/M}} \;\approx\; \mathcal{N}(0, 1). \tag{4}$$

The Wald interval replaces the unknown q in the denominator by $\hat{q}$ and inverts, which gives $\hat{q} \pm z\sqrt{\hat{q}(1-\hat{q})/M}$. The substitution is harmless when Mq is large. In the rare-event regime it is not, for three reasons. The standard error is estimated from the same few exceedances that determine $\hat{q}$, so it is too small whenever $\hat{q}$ falls below q; at $k = 0$ it is zero and the interval is the single point $\{0\}$. The interval is symmetric although the binomial law is strongly right-skewed at small Mq, so its lower limit can be negative. With $k = 1$, $M = 10^5$ and $p = 10^{-4}$ (iDDPM) it is $[-0.10, 0.30]$ on the ρ scale. Finally, its actual coverage is well below the nominal level for small Mq and oscillates with M and q (Brown et al., 2001). At our deepest audited depth the expected count is $Mp = 10$ for $M = 10^5$, 5 for the DiT pair and 1.5 for EDM-churn, and the observed counts range from 1 to 17.

The Wilson score interval (Wilson, 1927) keeps q in the denominator of (4) and collects all q with $(\hat{q} - q)^2 \leq z^2 q(1-q)/M$. This is a quadratic inequality in q whose solution set, divided by p, is

$$[L_W, U_W] \;=\; \frac{1}{p} \cdot \frac{\hat{q} + \frac{z^2}{2M} \;\pm\; z\sqrt{\frac{\hat{q}(1-\hat{q})}{M} + \frac{z^2}{4M^2}}}{1 + z^2/M}. \tag{5}$$

On the q scale the interval is centered at a value shrunk toward $1/2$; on the ρ scale it lies inside $[0, 1/p]$, has positive width at $k = 0$, and is asymmetric around $\hat{q}$ in the direction of the binomial skew. For the iDDPM cell above it gives $[0.02, 0.57]$. Brown et al. (2001) recommend it for small n on the basis of its coverage, and for rare events the relevant notion of small is the expected count Mq. We use $z = 1.96$. At $k \leq 1$ we also compute the exact Clopper–Pearson interval and assign a verdict only if it agrees with the reported interval. Two cells have $k = 1$, both at 10^{-4}: iDDPM ($M = 10^5$), whose Clopper–Pearson upper limit 0.56 agrees with its under-production verdict, and DiT-cfg1.0 ($M = 5 \times 10^4$), whose Clopper–Pearson interval $[0.005, 1.11]$ agrees with its interval containing 1.

Threshold error. Write $G_P = 1 - F_R$ for the survival function of R under the data law and G_Q for the sampler's, so that $\tau_p = G_P^{-1}(p)$. The empirical quantile $\hat{\tau}_p$ of n_h held-out points satisfies the first-order expansion $\hat{\tau}_p - \tau_p \approx -(\hat{G}_P(\tau_p) - p)/G_P'(\tau_p)$ with $\mathrm{Var}\,\hat{G}_P(\tau_p) = p(1-p)/n_h$. The audit evaluates the sampler's exceedance rate at $\hat{\tau}_p$ instead of τ_p, and

$$\frac{G_Q(\hat{\tau}_p) - G_Q(\tau_p)}{G_Q(\tau_p)} \approx \frac{G_Q'(\tau_p)}{G_Q(\tau_p)}\,(\hat{\tau}_p - \tau_p) \approx -\frac{h_Q(\tau_p)}{h_P(\tau_p)} \cdot \frac{\hat{G}_P(\tau_p) - p}{p}, \tag{6}$$

where $h = -G'/G$ is the hazard rate and we used $G_P(\tau_p) = p$. The relative error that the threshold contributes to $\hat{\rho}$ therefore has standard deviation

$$\eta_p\,\sigma_\tau, \qquad \sigma_\tau = \sqrt{\frac{1-p}{n_h\,p}}, \qquad \eta_p = \frac{h_Q(\tau_p)}{h_P(\tau_p)}.$$

Since $\log\rho = \log G_Q - \log G_P$, we have $\mathrm{d}\log\rho/\mathrm{d}\tau = h_P - h_Q$ and $\mathrm{d}\log p/\mathrm{d}\tau = -h_P$, hence $\eta_p = 1 + \mathrm{d}\log\rho/\mathrm{d}\log p$. Under Assumption 2, $\mathrm{d}\log\rho/\mathrm{d}\log p = 0$ at τ_p, so $\eta_p = 1$, and the term equals σ_τ whatever the densities: the local slope that converts a threshold shift into a rate shift under P is converted back by the same slope under Q. This is the form used in (1) and in our code, which sets $\eta_p = 1$ throughout. The value is 3% at $p = 10^{-3}$ and 10% at 10^{-4} for $n_h = 1.08$M.

Assumption 2 holds in particular when G_Q is proportional to G_P in a neighborhood of τ_p, for instance for a sampler that is calibrated throughout that neighborhood. It is a condition on the slope of ρ at τ_p and is not implied by $\rho(\tau_p) = 1$ alone. It fails when ρ varies with depth. A ratio that falls as p decreases gives $\eta_p > 1$ and an understated threshold term. The effect on the reported interval is limited at our budgets because the count term dominates: with $k = 100$ and $p = 10^{-3}$ the relative half-width is $(0.196^2 + 0.059^2)^{1/2} = 0.205$ for $\eta_p = 1$ and 0.229 for $\eta_p = 2$, a widening

of about 12%. The depth scans of App. H show how $\hat{\rho}$ varies with p for each configuration. We did not estimate η_p, because the difference of two noisy log-ratios at adjacent depths is too variable at our counts, and we list the assumption among the limitations. Where a threshold comes from the generalized-Pareto tail, the profile interval of App. C replaces this term.

Status of the interval components. The Wilson interval is an asymptotic approximation with good small-count behavior. The threshold term is a first-order delta-method approximation under Assumption 2. Extrapolated depths add a model-based component, the generalized-Pareto fit of App. C, whose profile-likelihood interval propagates into $\hat{\rho}$ and whose validity is checked empirically and is not guaranteed. Among the uncertainty statements, the only exact finite-sample bound is the Dvoretzky–Kiefer–Wolfowitz step of Proposition 3(iii) (Massart, 1990), apart from the Clopper–Pearson check above. Empirical coverage of the combined interval is reported in Apps. D and F.

B.6 How the theory, the audit and the wrapper fit together

Propositions 1–2 identify the severity law of R as a quantity that finite-moment metrics can miss and that no statistic resolves, for some pairs of samplers, without a budget of order $p^{-1}(\sqrt{c}-1)^{-2}$. The audit estimates that law's tail. By Proposition 3(i) the wrapper restores every functional of the severity law, and hence $\rho(\tau)$ at every depth, to the calibration law. By part (ii) the wrapper leaves conditional tail shapes unchanged up to the monotone relabeling, and by Corollary 1 any severity-keyed post-processor can only mix them along transport-invariant coordinates. The shape statistics of the audit measure exactly this remaining component (§5). After wrapping, the tail probability of the sampler is determined by $\widehat{F}$ alone, and the generative model contributes the content of the samples.

Same-M comparisons. Two configurations wrapped at the same M have identical sorted wrapped severities, because $r'_{(i)} = \widehat{F}^{-1}(i/(M+1))$. The collection contains two groups with more than one member, the five configurations at $M = 10^5$ and the DiT pair at $M = 5 \times 10^4$. We compare each member with the first member of its group, which gives four comparisons in the first group and one in the second, covering seven of the eight configurations. The maximum absolute discrepancy in the sorted wrapped severities is exactly 0 within each group. EDM-churn ($M = 15{,}104$) is alone in its group. The wrapped ratios of all groups are given in §5.

C The Spliced Calibration Law: Threshold Selection, Profile Intervals, and the Support-Aware Chart

This appendix specifies the spliced calibration law $\widehat{F}$, the rule that selects its threshold, the intervals attached to its tail probabilities, the support-aware transport chart, and the basis used for shape comparisons, and it reports the validation of each. These are the components that the audit and the wrapper use at depths beyond what the held-out corpus can certify.

C.1 The spliced calibration law $\widehat{F}$

Let $r_1, \ldots, r_{n_f}$ be the severities of the calibration split $\mathcal{D}_{\mathrm{cal}}$ and let u be a threshold selected as in §C.2. Write N_u for the number of exceedances, $\zeta = N_u/n_f$ for the exceedance fraction, and $y = r - u$ for the excesses. We fit a generalized Pareto law to the excesses by maximum likelihood with the location pinned at zero, which is the classical peaks-over-threshold model (Balkema & De Haan, 1974; Pickands III, 1975; Coles et al., 2001). For $\xi \neq 0$,

$$1 - \widehat{F}(t) \;=\; \zeta\Big(1 + \hat{\xi}\,\frac{t-u}{\hat{\sigma}}\Big)^{-1/\hat{\xi}}, \qquad t > u, \tag{7}$$

and for $t \leq u$ the empirical survival function of the calibration split is used unchanged. The $\xi = 0$ branch replaces the bracket by $\exp\{-(t-u)/\hat{\sigma}\}$. The corresponding quantile function, which is

what WRAP evaluates, is

$$\widehat{F}^{-1}(v) \;=\; \begin{cases} \text{empirical } v\text{-quantile of } \mathcal{D}_{\mathrm{cal}}, & v \le 1-\zeta, \\ u + \dfrac{\hat\sigma}{\hat\xi}\Big(\Big(\dfrac{1-v}{\zeta}\Big)^{-\hat\xi} - 1\Big), & v > 1-\zeta, \end{cases} \tag{8}$$

again with the $\hat\xi = 0$ branch $u + \hat\sigma \log\big(\zeta/(1-v)\big)$.

Two caveats apply. First, for $\hat\xi < 0$ the fitted law has a finite upper endpoint $u - \hat\sigma/\hat\xi$, and probabilities requested beyond it return zero. The "zero" columns of the tables below report how often this happens. Second, the splice is a model above u. Every number that depends on Eq. (7) beyond the held-out floor is labeled as extrapolation in the main text and carries the interval of §C.3.

The deep-tail transport demonstration needs severities strictly beyond the deepest certified threshold τ. We sample them from the threshold-stability form of Eq. (7). With $\sigma_\tau^{\mathrm{gpd}} = \hat\sigma + \hat\xi(\tau - u)$ and $E \sim \mathrm{Exp}(1)$, $r_g = \tau + \sigma_\tau^{\mathrm{gpd}}\,(e^{\hat\xi E} - 1)/\hat\xi$ is an exact draw from the fitted excess law above τ.

C.2 THE ADAPTIVE THRESHOLD RULE

A peaks-over-threshold fit depends on its threshold. A threshold that is too low biases the generalized-Pareto approximation, and one that is too high leaves too few excesses to fit. We choose the threshold automatically, per corpus, by goodness of fit. Over the grid of candidate exceedance quantiles

$$q \in \mathcal{Q} = \{0.98,\ 0.985,\ 0.99,\ 0.9925,\ 0.995\},$$

fit the generalized Pareto by maximum likelihood to the excesses over the q-quantile u_q of the calibration split, compute the Kolmogorov–Smirnov p-value of those excesses against the fitted law, and select

$$q^* \;=\; \arg\max_{q \in \mathcal{Q}} \; \mathrm{KS}p\big(\mathrm{GPD}(\hat\xi_q, \hat\sigma_q)\big).$$

Because the parameters are fitted to the same excesses, this p-value is not a calibrated goodness-of-fit test; we use it only as a score for ranking the candidate thresholds, and the performance of the resulting rule is assessed on synthetic families below. Ties are broken toward the deeper threshold, which the implementation does by adding $10^{-12} j$ to the j-th grid score. All intervals then condition on the selected quantile. We do not average over the grid, and we report q^* alongside every fit.

Validation design. We validated the rule on four synthetic families whose tail probabilities at extrapolated depths are known analytically. Following the row labels of Table 3, example 1 (ex1) is the linear Gaussian family, example 2 (ex2) the radial t_3 family, example 3 (ex3) the curved family, and example 4 (ex4) the mean family in dimension $d = 10$. Their targets lie 4.1, 4.0, 3.0 and 3.0 decades beyond the empirical floor. In every table of this subsection, "ratio" is the estimated tail probability divided by the true one at the extrapolated depth, "zero" is the fraction of replicates whose fitted endpoint falls short of the target, and intervals are at 90% nominal level. Two studies of different size were run, and their numbers are reported separately.

Pilot study (30 replicates per cell). The pilot compares the adaptive rule with every fixed-q choice on the same replicates (Table 2). In the pilot the rule's "zero" column is 0 in all four families. In the example-1 grid, where every fixed choice is shown, the rule's coverage equals that of the best fixed choice for both interval types, and the rule reaches it without knowing which fixed choice is best.

Main validation study (100 replicates per family). The larger study applies the adaptive rule independently on each replicate and gives the reference numbers in Table 3. The four-family coverage figures quoted in §C.3 come from this study. We also ran the same 100-replicate design at a fixed $q = 0.98$. Relative to that run, the adaptive rule moves the example-1 ratio median from 0.28 to 0.63 and its profile coverage from 0.81 to 0.94. The ratio median of example 3 moves from 0.60 to 0.74, and that of example 4 from 0.62 to 0.72. A fixed threshold has to be chosen without knowledge of the true tail, and these differences are the cost of fixing $q = 0.98$ in advance.

Table 2: Pilot study, 30 replicates per cell: adaptive threshold rule against fixed thresholds q. Column heads: "ratio med [IQR]" is the median and interquartile range of the estimated over true tail probability at the extrapolated depth; "zero" is the fraction of replicates whose fitted endpoint falls short of the target; "prof. cov / w" and "boot. cov / w" give, for the profile-likelihood and bootstrap intervals, the coverage at 90% nominal and the interval width in decades. The full grid is shown for example 1 (ex1); for examples 2–4 (ex2–ex4) the fixed-0.98 baseline is shown against the rule. The rule's median selected quantile is .99 on ex1–ex3 and .985 on ex4.

example	q	ratio med [IQR]	zero	prof. cov / w	boot. cov / w
ex1	0.98	0.43 [0.05,0.88]	.03	.80 / 2.8	.77 / 3.6
ex1	0.985	0.61 [0.06,1.09]	.00	.87 / 2.8	.87 / 3.4
ex1	0.99	0.57 [0.18,1.48]	.03	.87 / 3.0	.80 / 3.4
ex1	0.9925	0.52 [0.10,1.57]	.00	.90 / 3.4	.90 / 3.2
ex1	0.995	0.45 [0.05,1.07]	.03	.97 / 2.5	.80 / 4.2
ex1	**adaptive**	**0.63 [0.14,2.43]**	.00	**.97 / 2.8**	.90 / 3.2
ex2	0.98	0.77 [0.40,1.05]	.00	.87 / 1.3	.87 / 1.3
ex2	adaptive	0.73 [0.38,1.06]	.00	.87 / 1.6	.83 / 1.7
ex3	0.98	0.72 [0.35,1.08]	.00	.87 / 1.2	.80 / 1.3
ex3	adaptive	0.84 [0.44,1.33]	.00	.87 / 1.2	.87 / 1.5
ex4	0.98	0.74 [0.37,1.44]	.00	.90 / 1.3	.77 / 1.3
ex4	adaptive	0.75 [0.33,1.48]	.00	.83 / 1.3	.77 / 1.3

Table 3: Main validation study, 100 replicates per family: adaptive rule, 90% nominal intervals. Column heads: "ratio med [IQR]" is the median and interquartile range of the estimated over true tail probability; "zero" is the fraction of replicates whose fitted endpoint falls short of the target; "cov. boot" and "cov. prof" are the coverage of the bootstrap and profile-likelihood intervals; "width b / p (dec)" is the width of the bootstrap / profile interval in decades; "med. q" is the median selected quantile. The extrapolation distance beyond the empirical floor, in decades, is given in parentheses after each family.

family	ratio med [IQR]	zero	cov. boot	cov. prof	width b / p (dec)	med. q
ex1 linear, Gauss. (4.1 dec)	0.63 [0.06,1.52]	.02	.86	.94	3.8 / 2.9	.99
ex2 radial, t_3 (4.0 dec)	0.82 [0.37,1.39]	.00	.84	.86	1.5 / 1.6	.99
ex3 curved (3.0 dec)	0.74 [0.38,1.32]	.01	.85	.88	1.6 / 1.3	.99
ex4 d=10 mean (3.0 dec)	0.72 [0.21,1.56]	.00	.71	.77	1.6 / 1.3	.99

C.3 Profile-likelihood intervals for tail probabilities

Let $\ell(\xi, \sigma)$ be the generalized-Pareto log-likelihood of the N_u excesses of u in a corpus of n points, and let τ be the level whose exceedance probability is wanted. Reparameterize $(\xi, \sigma) \mapsto (\xi, p)$ through the model's own tail probability,

$$p \;=\; \frac{N_u}{n}\Big(1 + \xi\,\frac{\tau - u}{\sigma}\Big)^{-1/\xi},$$

invert for

$$\sigma(\xi, p) \;=\; \frac{\xi\,(\tau - u)}{(p/\zeta)^{-\xi} - 1}, \qquad \zeta = \frac{N_u}{n}, \tag{9}$$

(the $\xi \to 0$ limit being $\sigma = (\tau - u)/\log(\zeta/p)$), and profile out the shape parameter, $\ell^*(p) = \max_\xi \ell\big(\xi, \sigma(\xi, p)\big)$ over $\xi \in [-0.95, 2]$. The $1 - \alpha$ interval is the likelihood-ratio set

$$\mathrm{CI}_{1-\alpha}(p) \;=\; \Big\{p :\; 2\big[\ell^*(p^\dagger) - \ell^*(p)\big] \;\le\; \chi^2_{1,1-\alpha}\Big\}, \qquad p^\dagger = \arg\max_{p \in G}\; \ell^*(p), \tag{10}$$

where G is a grid of 25 log-spaced points spanning two decades either side of the nominal depth and $p^\dagger$ is the profile maximizer over that grid. The two endpoints are obtained by bracketing and bisecting $2[\ell^*(p^\dagger) - \ell^*(p)] - \chi^2_{1,1-\alpha}$ in $\log_{10} p$. The upper endpoint is capped at ζ, because the peaks-over-threshold model does not apply above ζ, and an endpoint is left at its cap if no crossing is bracketed. For the CelebA rows of App. I we use $1 - \alpha = 0.90$.

Table 4: Calibration law $\widehat{F}$ on ImageNet-32 under R_1: 20 splits, 200k calibration points, 1.08M held out, adaptive threshold reselected per split. Column heads: "ratio med [IQR]" is the median and interquartile range of the predicted over held-out empirical exceedance probability; "zero" is the fraction of splits whose fitted endpoint falls short of the target; "$\hat{\xi}$ med" is the median shape estimate; "cov. boot / prof" and "width boot / prof (dec)" are the coverage and the width in decades of the bootstrap / profile-likelihood intervals at 90% nominal.

depth p	ratio med [IQR]	zero	$\hat{\xi}$ med	cov. boot / prof	width boot / prof (dec)
10^{-3}	1.04 [0.97, 1.07]	.00	-0.03	.80 / .75	0.08 / 0.07
10^{-4}	0.88 [0.75, 1.03]	.00	-0.03	.75 / .80	0.26 / 0.24
3×10^{-5}	0.79 [0.61, 0.93]	.00	-0.03	.70 / .75	0.45 / 0.41

We use Eq. (10) in place of the percentile bootstrap because of its measured coverage. In an earlier run in the example-1 setting, a percentile bootstrap that refits the threshold inside each resample covered 68% at 90% nominal, because it cannot represent the asymmetry of the likelihood surface in ξ. In the 100-replicate study (Table 3), the profile interval covers $0.94/0.86/0.88/0.77$ across the four validation families against $0.86/0.84/0.85/0.71$ for the bootstrap, at 0.90 nominal and at comparable or smaller width.

Two limitations remain. In the hardest family, example 4, coverage is below nominal for both interval types (0.77 for the profile interval and 0.71 for the bootstrap). On clustered corpora, both interval types must take the cluster, and not the single observation, as the resampling unit. Otherwise their coverage falls to a small fraction of nominal.

C.4 The calibration law fitted for the image audit

For the pixel-space severity R_1 on ImageNet-32, $\widehat{F}$ is fitted once on the 200,000-image calibration split and then held fixed for every configuration in the paper. The fit selects the quantile $q^* = 0.9925$, with threshold $u = 12.06$, $\hat{\xi} = -0.050$, $\hat{\sigma} = 0.974$ and exceedance fraction $\zeta = 0.0075$. The audited level $\tau_{10^{-4}} = 16.13$ does not come from this fit. It is the empirical upper-10^{-4} quantile of $\mathcal{D}_{\mathrm{hold}}$ (≈ 108 exceedances). The GPD tail of $\widehat{F}$ places the same quantile at 15.84, within 2% of the held-out value. This comparison belongs to the component ablation summarized in §5 and reported in full in App. G (Table 19), and 15.84 is not used as an audit threshold. The shape estimate is close to the exponential boundary $\xi = 0$, where profile intervals are widest. For that reason we report profile intervals and do not use a normal approximation.

Table 4 reports the accuracy of $\widehat{F}$ on the image corpus under a 20-split replication. Each split refits the chart, reselects the threshold, and predicts held-out exceedance probabilities at three depths. The median ratio of predicted to held-out probability is 1.04 at 10^{-3}, 0.88 at 10^{-4} and 0.79 at 3×10^{-5}. The law is accurate at 10^{-3} and underestimates the held-out tail by about 12% at 10^{-4} and 21% at 3×10^{-5}, which is why wrapped ratios fall below 1 at 10^{-4}. Both interval types cover 0.70–0.80 at 90% nominal here. This is the accuracy available from a calibration law trained on 2×10^5 points, which hold 200 exceedances at 10^{-3} and 6 at 3×10^{-5}, and it is the reason the audit's headline depths are 10^{-3} and 10^{-4}.

C.5 The support-aware chart

Remark 4 reports the failure mode of a raw-coordinate transport at depth. The chart update raises energy that a bounded state space cannot absorb, so coordinates leave the box and must be projected back, which costs severity accuracy. Where the state space is a box, the alternative to projection is to transport in a chart in which the box is the whole space.

For data supported in $(0,1)^d$ (dequantized so that the map is finite), take the componentwise logit chart $\psi(x) = \mathrm{logit}(x)$ and transport by a *common shift along the all-ones direction*:

$$x \longmapsto \varsigma\big(\mathrm{logit}(x) + t\mathbf{1}\big), \qquad t \text{ the unique root of } R\big(\varsigma(\mathrm{logit}(x) + t\mathbf{1})\big) = r',$$

with ς the componentwise logistic function. If the severity is continuous and strictly increasing along this path and r' lies in its range along the path, the root exists, is unique, and is found by bi-

section. Every coordinate stays inside $(0, 1)$ by construction, and no projection step is needed. The shape state at a reference severity r_0 is $Z = \mathrm{logit}(x) + t_0(x)\mathbf{1}$, the point where the sample's own severity clock reaches r_0. Per-coordinate drift and a common scale are then fitted nonparametrically in r by severity binning and linear interpolation, so the shape model is smoothed and never extrapolated. Parametric extrapolation is reserved for the GPD tail of the calibration law above, where its uncertainty is quantified.

Illustration on a bounded-support corpus. As a separate illustration of support-preserving transport, we applied the chart to the WM-811K wafer-map corpus (Wu et al., 2014): $811{,}390$ wafer maps in $46{,}292$ production lots, with severity the mean cell failure fraction, which takes values in $[0, 1]$. This corpus is distinct from the image corpora audited in the main text. The transport is exact to machine precision: $\max |R(Z) - r_0| = 2.2 \times 10^{-16}$ on the reference level set and $\max |R - \text{target}| = 4.4 \times 10^{-16}$ for generated records. The fraction of coordinates falling outside the physical box is 0.000%, against 40.175% for the same construction performed in raw coordinates.

The chart leaves the severity law unchanged. The predicted-over-true tail probability at 10^{-3} is 1.05, and the minimum generated severity equals the target threshold to four decimals. The shape statistics improve. Leading principal components of the generated exceedances are 2–4 held-out standard deviations off in raw coordinates and within a third of one in the chart. As mean / standard deviation, PC1 is $+0.77/2.70$ generated against $-0.00/2.51$ held out, and PC2 is $+0.59/1.89$ against $-0.00/1.83$. The per-cell mean absolute gaps have median 0.012 and maximum 0.044, against a held-out cell standard deviation of 0.089.

The same corpus gives one methodological caution. With clustered records, an i.i.d. independence diagnostic returns $p = 0.00$ while the valid cluster-level permutation returns $p = 0.15$, so the resampling unit must be the cluster.

The image corpus of the main text has no such chart, because a pixel box is not a product of open intervals once quantization and clipping are taken into account. There we project and report the residuals (Remark 4).

C.6 Shape basis fitted on the calibration-split tail

Every shape comparison in this paper projects generated and held-out exceedance directions onto a principal-component basis fitted on the tail of the calibration split, never on the held-out exceedances being tested. A basis fitted on the held-out exceedances themselves inflates the spread along its own leading component by selection. We measured this on ImageNet-32 under R_1, with a tail generated from the calibration law during development of the protocol. With the basis fitted on the held-out exceedances, the held-out PC1 standard deviation was 2.00 against 1.17 for the generated tail, a spurious factor-of-two dispersion mismatch. With the basis fitted on the calibration tail, the same comparison gives 1.68 against 1.35 on PC1 and 1.29 against 1.28 on PC2. The factor-of-two mismatch is therefore an artifact of the basis. We fit the shape basis on the calibration split and evaluate dispersion on held-out exceedances. The shape collapse reported in §5 is measured against a basis that has never seen the held-out exceedances, and the PC1 under-dispersion that remains with this basis is much smaller than the artifact.

D Real-versus-Real Self-Validation

Before any generator is audited, the protocol audits real data against real data. The calibration split $\mathcal{D}_{\mathrm{cal}}$ takes the role of the model, and the thresholds τ_p come from the held-out split $\mathcal{D}_{\mathrm{hold}}$. Both splits are drawn from the same population, so the population target is $\rho_{\mathrm{self}} = 1$, and a departure of the estimate from 1 measures the error of the audit itself. This appendix reports every cell of that test, then separates the two sources of its variability, and then defines the calibration-law ratio $e(p)$ that the wrapper inherits.

D.1 COVERAGE AGAINST THE SELF-VALIDATION TARGET

Before any model was audited, we set a target for the audit's own error: the 95% interval should contain 1 in at least 14 of the 15 seed $\times$ depth cells formed by five split seeds (8000–8004) and three depths. The target measures how far the audit can be trusted; the model audits do not depend on meeting it, and a shortfall is reported and carried into the interpretation of every verdict. This check refits the functional and $\widehat{F}$ for each seed, whereas the interval of Eq. (1) is conditional on the fitted functional; the next subsection separates the two sources of variation. Table 5 lists all cells. With the quadrature interval of §3, 12 of the 15 cells contain 1, below this target. With Wilson-only intervals the count is $11/15$. Seed 8004 misses at all three depths under both intervals, and seed 8003 misses at 3×10^{-3} under the Wilson-only interval. The next subsection shows where the shortfall comes from: with the functional and calibration law held fixed and only the split redrawn, the quadrature interval reaches nominal coverage, so the excess variability comes from refitting the functional and $\widehat{F}$ for each seed.

Table 5: Self-validation, all 15 cells. The calibration split ($M = 200{,}000$ real images) is audited in the model role against the held-out thresholds τ_p, so the target value is $\rho_{\text{self}} = 1$. In the last column, ✓ and × mark whether the 95% Wilson-only interval (W) and the 95% quadrature interval (Q) contain 1. The totals are $12/15$ for quadrature and $11/15$ for Wilson-only.

Seed	p	τ_p	k	$\hat{\rho}_{\text{self}}$	95% Wilson	95% quadrature	W / Q
8000	10^{-2}	11.753	2014	1.007	[0.964, 1.052]	[0.960, 1.056]	✓ ✓
8000	3×10^{-3}	12.921	597	0.995	[0.918, 1.078]	[0.911, 1.085]	✓ ✓
8000	10^{-3}	13.935	204	1.020	[0.889, 1.170]	[0.876, 1.182]	✓ ✓
8001	10^{-2}	11.762	1955	0.978	[0.935, 1.022]	[0.932, 1.025]	✓ ✓
8001	3×10^{-3}	12.939	575	0.958	[0.883, 1.040]	[0.876, 1.046]	✓ ✓
8001	10^{-3}	13.964	178	0.890	[0.769, 1.031]	[0.757, 1.040]	✓ ✓
8002	10^{-2}	11.764	2037	1.019	[0.975, 1.063]	[0.971, 1.067]	✓ ✓
8002	3×10^{-3}	12.932	642	1.070	[0.990, 1.156]	[0.982, 1.163]	✓ ✓
8002	10^{-3}	13.944	212	1.060	[0.927, 1.213]	[0.912, 1.225]	✓ ✓
8003	10^{-2}	11.746	2073	1.037	[0.993, 1.082]	[0.989, 1.086]	✓ ✓
8003	3×10^{-3}	12.915	652	1.087	[1.007, 1.173]	[0.998, 1.181]	× ✓
8003	10^{-3}	13.935	206	1.030	[0.899, 1.180]	[0.885, 1.193]	✓ ✓
8004	10^{-2}	11.754	2118	1.059	[1.015, 1.105]	[1.011, 1.109]	× ×
8004	3×10^{-3}	12.922	656	1.093	[1.013, 1.180]	[1.005, 1.188]	× ×
8004	10^{-3}	13.922	236	1.180	[1.039, 1.340]	[1.022, 1.355]	× ×

D.2 SEPARATING SPLIT VARIABILITY FROM REFITTING VARIABILITY

The shortfall has two possible causes. The intervals may be too narrow, or the protocol may contain a source of variability that the intervals do not model. A *fixed-functional re-split null* separates the two. The severity functional and the calibration law $\widehat{F}$ are frozen at their seed-8000 fit, and only the assignment of real images to $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{hold}}$ is redrawn, $B = 400$ times. The spread of $\hat{\rho}_{\text{self}}$ over these draws is due to split sampling alone (Table 6).

Table 6: Fixed-functional re-split null ($B = 400$ redrawn splits, severity functional and calibration law frozen at the seed-8000 fit). The design effect is the ratio of the realized s.d. of $\hat{\rho}_{\text{self}}$ to the Wilson-implied s.d. The last two columns give the fraction of draws in which the 95% Wilson-only and quadrature intervals contain 1.

p	mean $\hat{\rho}_{\text{self}}$	s.d.	Wilson s.d.	design eff.	cov. (Wilson)	cov. (quad.)
10^{-2}	0.998	0.0239	0.0222	$\times1.077$	0.933	0.950
3×10^{-3}	0.999	0.0447	0.0408	$\times1.097$	0.918	0.953
10^{-3}	1.002	0.0744	0.0708	$\times1.051$	0.928	0.958

With the functional frozen, the Wilson-only interval is too narrow by a measurable amount. The realized s.d. exceeds the Wilson-implied s.d. by a factor of 1.05 to 1.10, and Wilson-only coverage

is 0.933, 0.918 and 0.928 at the three depths. Adding the delta-method threshold term in quadrature raises coverage to 0.950, 0.953 and 0.958, which is nominal. For this reason the quadrature interval is the certificate used throughout the paper.

The wider interval does not account for the spread across seeds. With the functional frozen, the realized s.d. at 10^{-3} is 0.074. In Table 5 the severity functional and $\widehat{F}$ are refitted for every seed, and there $\hat{\rho}_{\text{self}}(10^{-3})$ ranges over 0.89–1.18 across the five seeds, with s.d. ≈ 0.10. Each seed therefore audits a slightly different axis, with its own PCA basis, adaptive threshold and generalized-Pareto fit. This refitting variability is a protocol-level uncertainty that no count-based interval can absorb, and verdicts at smaller tail probabilities should be read with it in mind. It is also why the main claims of the paper are stated at 10^{-3}, where this component is smallest relative to the reported effects: ratios of 0.17 and 1.99 lie far outside a band of ± 0.10.

D.3 The calibration-law ratio $e(p)$, its rank-grid counterpart, and the forward diagnostic

Three related ratios appear in this paper. They evaluate different threshold constructions, and Table 7 reports them separately for every seed.

The first is the calibration-law ratio, the quantity a wrapped sampler inherits. It is the exceedance rate of the calibration law $\widehat{F}$ at the held-out threshold, divided by p:

$$e(p) \;=\; \frac{1-\widehat{F}(\tau_p)}{p}, \qquad \tau_p = \text{hold-split upper-}p\text{ quantile}, \tag{11}$$

which equals 1 exactly when $\widehat{F}$ reproduces the held-out tail at depth p. Evaluated on the continuous spliced law, it is written $e_{\text{cont}}(p)$ in the table.

The second quantity, $e_{\text{grid}}(p)$, is the same ratio realized on the rank grid of a wrapped sample of size M. The wrapped severities lie on a grid of M ranks, so the ratio is resolved only in steps of $1/(Mp)$, and e_{grid} is the value that a wrapped sampler of that size inherits. The table uses $M = 100$k, for which the step at 10^{-4} is 0.1.

The third quantity is the forward diagnostic $\kappa(p) = P_{\mathcal{D}_{\text{hold}}}(R \geq \widehat{F}^{-1}(1-p))/p$. It takes the threshold from the calibration law and counts exceedances on the held-out split, which is the opposite order from $e(p)$. The two are not reciprocals of each other, because they depend on the local tail slope at different thresholds. On the reference split (seed 8000) at 10^{-4}, $e_{\text{cont}} = 0.688$, $e_{\text{grid}} = 0.600$ and $\kappa = 1.286$, so $e(p)$ and $\kappa(p)$ differ by a factor of nearly two. Only $e(p)$ enters Proposition 3(i).

Table 7: Calibration-law ratio by seed. e_{cont} evaluates Eq. (11) on the continuous spliced law. e_{grid} is the same quantity on the rank grid of an $M = 100$k wrapped sample, which is the value a wrapped sampler of that size inherits. κ is the forward diagnostic, listed for comparison. The last column is the calibration-law threshold $\widehat{F}^{-1}(1-p)$ used by κ.

Seed	p	$e_{\text{cont}}(p)$	$e_{\text{grid}}(p)$	$\kappa(p)$ (forward)	$\widehat{F}^{-1}(1-p)$
8000	10^{-3}	0.990	0.99	1.008	13.927
8000	10^{-4}	0.688	0.60	1.286	15.839
8001	10^{-3}	0.906	0.90	1.105	13.873
8001	10^{-4}	0.870	0.80	1.082	15.996
8002	10^{-3}	1.075	1.07	0.920	14.009
8002	10^{-4}	0.775	0.70	1.212	15.956
8003	10^{-3}	1.064	1.06	0.940	13.993
8003	10^{-4}	1.037	1.00	0.971	16.074
8004	10^{-3}	1.143	1.14	0.883	14.046
8004	10^{-4}	1.041	1.00	0.943	16.071

The ratio varies with the split. Across the five seeds $e(10^{-4})$ spans 0.69–1.04 and $e(10^{-3})$ spans 0.91–1.14, so we report $e(p)$ as a range over seeds. The calibration law is accurate to roughly $\pm 10\%$ at 10^{-3} and to tens of percent at 10^{-4}. By Proposition 3(i), a wrapped sampler has exactly this accuracy and no better. Every wrapped ratio in §5 is therefore compared with $e(p)$ instead of with

1. On the reference split the wrapped ratios at $M = 100$k are 0.990 and 0.600, equal to e_{grid} at both depths (Fig. 3). On the CelebA-HQ replication (App. I) the wrapped ratios are within 0.007 of $e(p)$ at all three certified depths, although they are as low as 0.60.

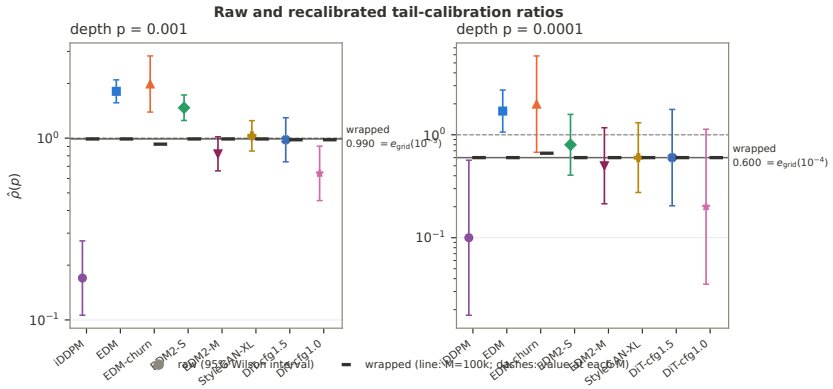

Figure 3: Raw and wrapped tail ratios for the eight configurations on R_1, at $p = 10^{-3}$ (left) and $p = 10^{-4}$ (right), on a logarithmic axis. Colored markers are the raw $\hat{\rho}_{R_1}(p)$ of each configuration with 95% Wilson intervals (verdicts in the paper use the quadrature interval of Eq. (1); at 10^{-4} the two differ for EDM). Short black dashes are the wrapped ratios, and the thin dashed line marks 1. Wrapped severities are bitwise identical across configurations at equal M, so the dashes take one value per budget group. The five $M = 100$k configurations share the solid horizontal line at 0.990 and 0.600, the two DiT configurations ($M = 50$k) have 0.980 and 0.600, and EDM-churn ($M = 15{,}104$) has 0.927 and 0.662. The values labeled e_{grid} in the panels are the rank-grid quantities at $M = 100$k. The continuous-law value at 10^{-4} is $e_{\text{cont}}(10^{-4}) = 0.688$ (Table 7).

E The Effect of Wrapping on FID and on the Transported Images

Proposition 3 describes what the wrapper does to the severity law, and Proposition 4 bounds what any post-processor can do to a Fréchet score. Neither says what the transport and the projection onto the pixel box do to individual images. This appendix measures both effects. It also reports Inception FID measured on the audited samples (Table 12).

Setup. Seven of the eight configurations of Table 1 (all but EDM-churn) were regenerated from the public checkpoints with the same sampler settings and seeds as in App. J: $600{,}000$ images in 20.6 GPU-hours. This is a rerun of the same draws, not an independent sample, and it reproduces the exceedance counts, $\hat{\rho}$ and intervals of Table 1 at both depths in all seven configurations, and FMD agrees to within 6×10^{-5}. The wrapper is applied to images as in Remark 4: the top 1% of draws by severity rank are transported with 120 steps of damped alternating projection, and the result is rounded to 8-bit images, the form in which samples are consumed. FID follows the standard recipe (Inception-v3 pool features, 32^2 images upsampled to 299^2 inside the extractor) against one fixed reference of $540{,}583$ held-out real images. KID uses the same features, and FMD is the axis of Table 1. FID is biased in the sample size, so all arms of a configuration are evaluated on the same fixed $50{,}000$-image subsample and their differences are paired. Uncertainty comes from a paired bootstrap ($B = 200$, the same resampled indices in every arm). As a control, $100{,}000$ held-out real images disjoint from the reference are passed through the same pipeline in the sampler's role. These FID values are computed at 32^2 against our reference split and are not comparable with published values.

Tail-only wrapping. Table 8 and Fig. 4 give the result. The four configurations whose intervals excluded 1 under R_1 (iDDPM and DiT-cfg1.0 under-producing, EDM and EDM2-S over-producing) are labeled calibrated after wrapping, with image-realized $\hat{\rho}(10^{-3})$ of 0.98 to 0.99 from raw values of 0.17 to 1.81. At 10^{-4} every configuration has 0.60, the rank-grid value of the calibration-law ratio (Eq. (2)). Over the same seven configurations FID changes by at most 0.0024 (EDM). The

Table 8: Inception FID ($n = 50$k, fixed paired subsample) and $\hat{\rho}_{R_1}$ before and after wrapping. ΔFID is the paired difference on that subsample, with the 95% paired-bootstrap interval ($B = 200$) in brackets; "sd" is the bootstrap standard deviation of the raw FID. $\hat{\rho}$ is computed from the realized 8-bit images with the quadrature interval of Eq. (1). * 100k held-out real images in the sampler's role.

		FID (Inception, 32^2)				$\hat{\rho}_{R_1}$ at depth		
Configuration	M	raw	wrapped	ΔFID [95%]	sd	raw, 10^{-3}	wrapped, 10^{-3}	wr., 10^{-4}
iDDPM	100k	3.067	3.065	-0.0016 [-0.0036, -0.0004]	0.035	$0.17^{\downarrow}$ [0.11, 0.27]	0.98 [0.79, 1.20]	0.60
EDM	100k	2.357	2.359	$+0.0024$ [$+0.0017$, $+0.0028$]	0.033	$1.81^{\uparrow}$ [1.54, 2.11]	0.98 [0.79, 1.20]	0.60
EDM2-S	100k	3.024	3.025	$+0.0017$ [$+0.0010$, $+0.0018$]	0.043	$1.47^{\uparrow}$ [1.23, 1.74]	0.99 [0.80, 1.21]	0.60
EDM2-M	100k	3.057	3.057	$+0.0001$ [-0.0001, $+0.0003$]	0.046	0.82 [0.65, 1.02]	0.99 [0.80, 1.21]	0.60
StyleGAN-XL	100k	3.539	3.539	$+0.0000$ [-0.0001, $+0.0002$]	0.039	1.03 [0.84, 1.26]	1.03 [0.84, 1.26]	0.60
DiT-cfg1.5	50k	3.107	3.107	-0.0002 [-0.0005, $+0.0001$]	0.047	0.98 [0.73, 1.30]	0.98 [0.73, 1.30]	0.60
DiT-cfg1.0	50k	3.137	3.137	$+0.0005$ [-0.0001, $+0.0014$]	0.040	$0.64^{\downarrow}$ [0.45, 0.91]	0.98 [0.73, 1.30]	0.60
Real images*	100k	0.716	0.716	-0.0000 [-0.0001, $+0.0001$]	0.007	1.01 [0.82, 1.24]	1.01 [0.82, 1.24]	0.60

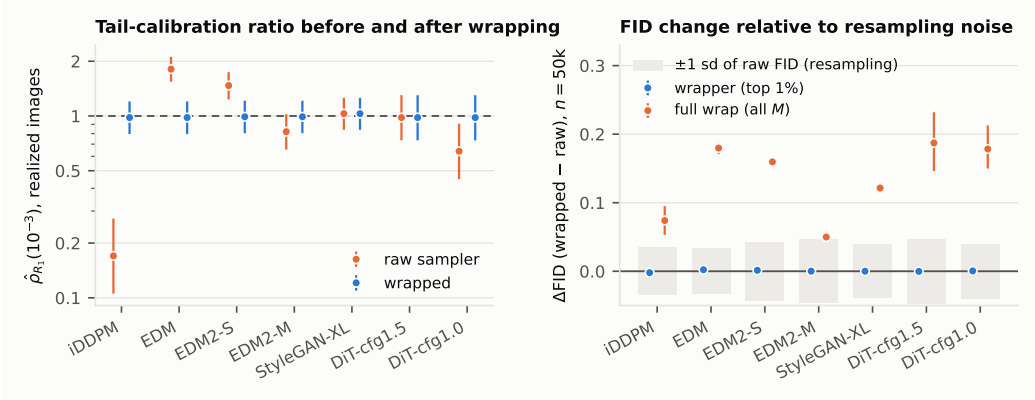

Figure 4: Left: $\hat{\rho}_{R_1}(10^{-3})$ computed from the realized 8-bit images before and after wrapping (95% quadrature intervals, dashed line = calibrated). Right: paired-bootstrap change in Inception FID ($B = 200$, 95% intervals) for the tail-only wrapper (top 1% of draws) and for a full wrap of all M draws. Gray bars span ± 1 bootstrap standard deviation of the raw FID.

bootstrap interval excludes zero in three configurations, in one of them (iDDPM) as a decrease. In every configuration the change is at most 7% of the bootstrap standard deviation of the FID itself (0.033 to 0.047 at $n = 50$k). KID changes by at most 2.7×10^{-6} and FMD by at most 1.6×10^{-3}. For scale, the FID of the real-image control against the same reference is 0.72.

Relation to Propositions 1 and 4. The raw and wrapped versions of iDDPM are two samplers whose FID differs by 0.002 and whose $\hat{\rho}(10^{-3})$ differs by 0.81. This shows on real samplers the insensitivity that Proposition 1 describes; the FIDs are close, not equal, and the comparison does not test Assumption 1. The same insensitivity means that the small FID change cannot serve as evidence that the transported images are intact, which is why they are examined one by one below. Proposition 4 bounds the change in d_F by $\sqrt{\pi}D$ with $\pi = 0.01$. We did not store per-image Inception features, so D is not evaluated. The measured changes are consistent with the order-π scaling of the moment shift: where the tail-only change is resolved from zero, the change under the full wrap ($\pi \approx 1$) is 48 to 94 times larger in magnitude (point estimates on the fixed subsample).

Two variants of the implementation. *Full wrap.* Applying the transport to all M draws, as Algorithm 2 is written, gives the same audited rates and raises FID by 0.05 to 0.19, which is one to six bootstrap standard deviations and positive in all 200 resamples. On real images it costs 0.05. Its effect on FMD is mixed, with two configurations improving, so we do not conclude that it lowers quality. It changes the bulk measurably without changing any audited quantity, and the implementation therefore transports the tail region only. *Hard clip.* With one radial step and a single clip, FID is unchanged but under-producing samplers are not repaired. iDDPM reaches 0.59 and remains

Table 9: Two variants of the implementation. *Full wrap* applies Algorithm 2 to all M samples; *hard clip* replaces the damped alternating projection by one radial step and one clip. ΔFID values are paired-bootstrap means, with 95% intervals for the full wrap. "within" is the share of transported samples whose realized severity is within 0.05 of its target; the last column gives this share for the damped scheme of Table 8.

	full wrap			hard clip			damped
Configuration	ΔFID [95%]	modified	$\hat{\rho}(10^{-3})$	ΔFID	$\hat{\rho}(10^{-3})$	within	within
iDDPM	$+0.074$ [$+0.053$, $+0.095$]	100%	0.98 [0.79, 1.20]	-0.0013	$0.59^{\downarrow}$ [0.45, 0.76]	21%	99.5%
EDM	$+0.180$ [$+0.171$, $+0.187$]	100%	0.98 [0.79, 1.20]	$+0.0023$	0.98 [0.79, 1.20]	100%	100%
EDM2-S	$+0.160$ [$+0.152$, $+0.166$]	100%	0.99 [0.80, 1.21]	$+0.0014$	0.99 [0.80, 1.21]	100%	100%
EDM2-M	$+0.050$ [$+0.046$, $+0.054$]	98%	0.99 [0.80, 1.21]	$+0.0002$	0.94 [0.76, 1.16]	95%	100%
StyleGAN-XL	$+0.121$ [$+0.118$, $+0.125$]	99%	1.03 [0.84, 1.26]	$+0.0001$	1.03 [0.84, 1.26]	99%	100%
DiT-cfg1.5	$+0.187$ [$+0.146$, $+0.232$]	99%	0.98 [0.73, 1.30]	-0.0001	0.98 [0.73, 1.30]	84%	100%
DiT-cfg1.0	$+0.178$ [$+0.150$, $+0.213$]	99%	0.98 [0.73, 1.30]	$+0.0005$	0.82 [0.60, 1.12]	42%	100%
Real images*	$+0.047$ [$+0.033$, $+0.068$]	10%	1.01 [0.82, 1.24]	-0.0000	1.01 [0.82, 1.24]	95%	100%

Table 10: Per-image change of the transported samples (medians, with the 5th or 95th percentile where shown), share of pixel values on the box boundary before and after, improved precision of the same samples before and after against a real-image manifold ("bulk" is the precision of the same sampler's untransported samples), and KID of the transported set to the real tail. The held-out real tail has precision 0.658.

Configuration	changed	$\lvert\Delta\text{pix}\rvert$	PSNR	SSIM med / p05	LPIPS med / p95	sat. %	precision	bulk	tail KID
iDDPM	1000/1000	6.23	30.6	0.982 / 0.904	0.0095 / 0.0396	3.8 $\rightarrow$ 16.8	0.632 $\rightarrow$ 0.615	0.700	0.0050 $\rightarrow$ 0.0050
EDM	1000/1000	3.05	37.0	0.997 / 0.947	0.0014 / 0.0024	14.1 $\rightarrow$ 0.0	0.673 $\rightarrow$ 0.669	0.660	0.0023 $\rightarrow$ 0.0029
EDM2-S	999/1000	1.82	41.4	0.999 / 0.972	0.0006 / 0.0011	13.1 $\rightarrow$ 0.1	0.709 $\rightarrow$ 0.707	0.682	0.0024 $\rightarrow$ 0.0028
EDM2-M	747/1000	0.39	52.1	1.000 / 0.998	0.0001 / 0.0004	12.7 $\rightarrow$ 8.2	0.760 $\rightarrow$ 0.758	0.701	0.0056 $\rightarrow$ 0.0056
StyleGAN-XL	307/1000	0.07	59.4	1.000 / 1.000	0.0000 / 0.0001	14.0 $\rightarrow$ 12.7	0.675 $\rightarrow$ 0.678	0.682	0.0035 $\rightarrow$ 0.0035
DiT-cfg1.5	454/500	0.64	49.6	1.000 / 0.994	0.0002 / 0.0007	10.8 $\rightarrow$ 12.9	0.700 $\rightarrow$ 0.692	0.726	0.0034 $\rightarrow$ 0.0034
DiT-cfg1.0	500/500	2.57	38.0	0.995 / 0.955	0.0018 / 0.0093	9.4 $\rightarrow$ 18.2	0.680 $\rightarrow$ 0.662	0.636	0.0023 $\rightarrow$ 0.0024
Real images*	852/1000	0.35	52.6	1.000 / 0.997	0.0001 / 0.0003	13.7 $\rightarrow$ 15.1	0.658 $\rightarrow$ 0.660	0.656	0.0000 $\rightarrow$ 0.0000

under-producing, with 21% of its transported samples within 0.05 of target against 99.5% for the damped scheme. Over-producing samplers are unaffected, because shrinking a radius cannot leave the box. The alternating projection is thus needed for upward severity moves, and clipping by itself has no effect on FID.

Quality of the transported images. Table 10 considers only the draws that the wrapper changes. In six of the seven configurations the change is small (median SSIM at least 0.995, median LPIPS at most 0.002). The largest repair is iDDPM, which has to move from 0.17 to calibration: the median change is 6.2 gray levels per pixel and the median SSIM 0.982. To assess whether the results still resemble real images we use improved precision (Kynkäänniemi et al., 2019). A sample counts as realistic if it lies inside the 3-nearest-neighbor ball of some real image in Inception feature space, with a reference manifold of $50{,}000$ random real images plus the $5{,}406$ real images of highest severity. For the same samples before and after wrapping, precision falls by at most 1.8 points (DiT-cfg1.0 $0.680 \rightarrow 0.662$, iDDPM $0.632 \rightarrow 0.615$) and by less than one point in the other five configurations. Held-out real tail images have precision 0.658 against this manifold, so the wrapped tails are about as realistic by this measure as the real tail. KID between the transported set and the real tail changes by at most 0.0006.

Artifacts of the projection. The projection has two side effects. First, it changes the share of saturated pixels in both directions. Real tail images have 13.7% of their pixel values on the box boundary. Under-producing samplers are under-saturated in the tail (iDDPM 3.8%) and the wrapper moves them toward the real level (16.8%). Over-producing samplers are contracted and their saturated pixels nearly vanish (EDM $14.1\% \rightarrow 0.05\%$), so pure white becomes slightly gray. The perceptual change is small (median SSIM 0.997). Second, the projection fails on a few samples. For iDDPM, 5 of $1{,}000$ transported draws do not converge and 8 more converge with a visible change (SSIM below 0.80). The real-image control has 2 and 0.

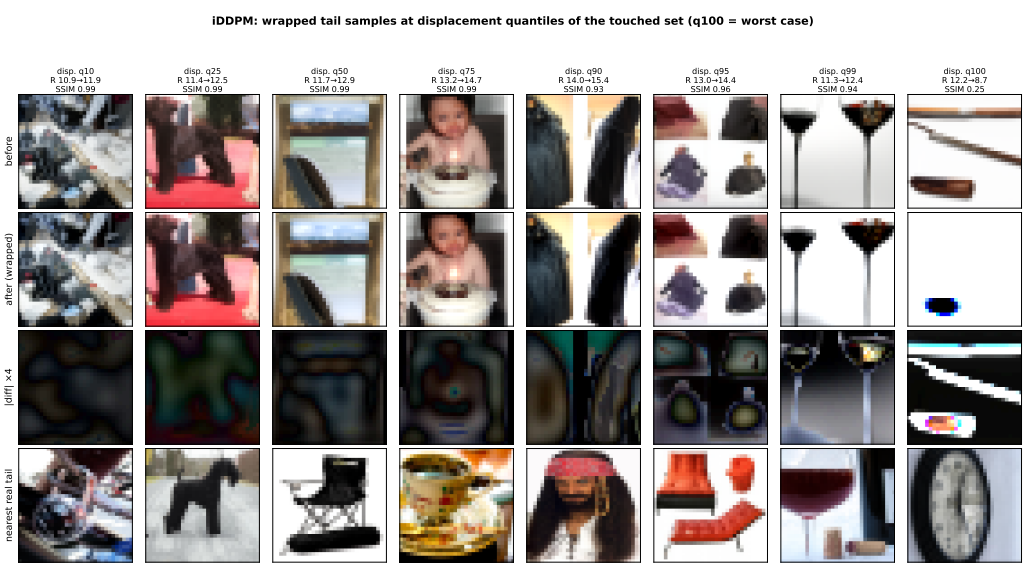

Figure 5: iDDPM, the largest repair among the seven configurations. Columns are fixed quantiles of pixel displacement among the transported samples, so the rightmost column is the worst case. Rows: before, after, absolute difference ($\times 4$), nearest real tail image in Inception space. In the worst case a nearly white image is driven into saturation and its severity falls.

Table 11: An optional per-sample guard. A transported image is replaced by the original sample when its severity residual exceeds 0.05 ("residual") or, in the second variant, when in addition SSIM against the original is below 0.80. ΔFID values are paired-bootstrap means. Configurations not listed revert no sample under either guard.

Configuration	guard	reverted	min SSIM	#SSIM<0.8	ΔFID	$\hat{\rho}(10^{-3})$	$\hat{\rho}(10^{-4})$	precision
iDDPM	none	0	0.25	11	-0.0020	0.98 [0.79, 1.20]	0.60	0.615
iDDPM	residual	5	0.45	8	-0.0017	0.98 [0.79, 1.20]	0.60	0.617
iDDPM	residual or SSIM	13	0.80	0	-0.0018	0.96 [0.78, 1.18]	0.50	0.616
EDM	none	0	0.67	5	$+0.0023$	0.98 [0.79, 1.20]	0.60	0.669
EDM	residual	0	0.67	5	$+0.0023$	0.98 [0.79, 1.20]	0.60	0.669
EDM	residual or SSIM	5	0.83	0	$+0.0022$	0.98 [0.79, 1.20]	0.60	0.667
DiT-cfg1.0	none	0	0.77	1	$+0.0005$	0.98 [0.73, 1.30]	0.60	0.662
DiT-cfg1.0	residual	0	0.77	1	$+0.0005$	0.98 [0.73, 1.30]	0.60	0.662
DiT-cfg1.0	residual or SSIM	1	0.81	0	$+0.0005$	0.98 [0.73, 1.30]	0.60	0.664
Real images*	none	0	0.30	1	-0.0000	1.01 [0.82, 1.24]	0.60	0.660
Real images*	residual	2	0.96	0	-0.0000	1.01 [0.82, 1.24]	0.60	0.660
Real images*	residual or SSIM	2	0.96	0	-0.0000	1.01 [0.82, 1.24]	0.60	0.660

An optional guard. Both failures can be detected per sample, so the original draw can be returned when the realized severity misses its target by more than 0.05 or when SSIM against the original falls below 0.80 (Table 11). The residual test alone is insufficient, because it does not detect an image that converged to a silhouette. Together the two tests revert 19 of the $6{,}000$ transported generated images (13 for iDDPM, 5 for EDM, 1 for DiT-cfg1.0) and 2 of the $1{,}000$ real ones. They raise the worst-case SSIM to at least 0.80 in every configuration, leave FID unchanged and change no verdict; iDDPM moves from 0.98 to 0.96 at 10^{-3} and loses one count at 10^{-4}. With the guard the wrapped severity law is no longer exactly $\widehat{F}$, since reverted draws keep their raw severity. Proposition 3(i) then holds up to the reverted fraction of each sampler's transported draws, at most 1.3% (iDDPM).

Scope. These results are for R_1 on ImageNet-32 and seven configurations. The image-realized rates can differ from the remapped rates by up to four counts, because a severity move smaller than half a gray level rounds back to the original image. StyleGAN-XL, which was already calibrated, keeps 103 exceedances against a remap target of 99, and only 307 of its $1{,}000$ transported images change at all. Precision, LPIPS and KID are computed in feature spaces pretrained at higher reso-

Table 12: Inception FID measured on the audited 32^2 sample sets (Inception-v3 pool features, one fixed reference of 540,583 held-out images), with the published values and the MAE-space FMD of Table 1. KID is $\times 10^3$. * Held-out real images in the sampler's role.

Configuration	FID publ.	FID (M)	FID (50k)	KID	FMD	$\hat{\rho}_{R_1}(10^{-3})$
iDDPM	2.90	2.74	3.07	1.34	3.776	$0.17^{\downarrow}$ [0.11, 0.27]
EDM	2.60	2.05	2.36	0.91	0.638	$1.81^{\uparrow}$ [1.54, 2.11]
EDM2-S	1.58	2.71	3.02	1.69	0.678	$1.47^{\uparrow}$ [1.23, 1.74]
EDM2-M	1.43	2.73	3.06	1.91	0.699	0.82 [0.65, 1.02]
StyleGAN-XL	1.10	3.18	3.54	2.25	0.391	1.03 [0.84, 1.26]
DiT-cfg1.5	2.27	3.11	3.11	2.04	0.644	0.98 [0.73, 1.30]
DiT-cfg1.0	9.60	3.14	3.14	1.39	0.531	$0.64^{\downarrow}$ [0.45, 0.91]
Real images*	–	0.39	0.72	0.01	0.004	1.01 [0.82, 1.24]

lution and are proxies for the usability of 32^2 images. We ran no human study, no sweep over the transported fraction π, and no image-level wrap under R_2 or R_3.

Measured FID of the audited configurations. Table 12 reports the Inception FID of the audited sample sets. It orders the configurations differently from both other quality axes. EDM has the lowest measured FID and StyleGAN-XL the highest, whereas StyleGAN-XL has the lowest FMD, and iDDPM, which has by far the highest FMD, is in the middle of the measured-FID ranking.

F Planted-Truth Recovery Under a Known Tail-Calibration Ratio

Every ratio in §5 estimates a quantity that cannot be observed for a real generative model. To test the estimator we construct samplers whose tail-calibration ratio is known by construction and audit them with the unmodified protocol.

Construction and the two designs. We use two designs, which differ in the split on which the planted ratio is defined. In the *corpus-relative design* (Variant A) the planted ratio is defined at the same held-out threshold that the audit uses, so the planted truth and the audited threshold refer to one fixed corpus. This design isolates the counting error of the estimator. In the *independent-threshold design* (Variant B) the planted ratio is defined on a split that is independent of the one from which the audit takes its threshold, so threshold-estimation error enters the comparison. Variant B matches the estimand of §3, where τ_p is an estimate of the population quantile, and it is the design that tests the interval of Eq. (1).

In both designs we fix a depth p and its threshold τ_p on the held-out split $\mathcal{D}_{\text{hold}}$. Starting from a base cache Q_0 (StyleGAN-XL, truncated below τ_p), we mix in held-out real tail images at weight ε,

$$Q_\varepsilon \;=\; (1-\varepsilon)\,Q_0|\{R<\tau_p\} \;+\; \varepsilon\,P_{\text{tail}}|\{R\geq\tau_p\}, \qquad \text{so} \qquad Q_\varepsilon(R\geq\tau_p)=\varepsilon \tag{12}$$

exactly, and the planted ratio is $\rho(p) = \varepsilon / P_{\mathcal{D}_{\text{hold}}}(R \geq \tau_p)$. This is the ε-mixing construction of Proposition 1. The mixture is built in sample space, so the audit receives only images. We plant $\rho \in \{0.3, 1, 2, 5\}$ at $p \in \{10^{-2}, 10^{-3}\}$ with $B = 500$ independent replicates per cell at $M = 100{,}000$.

The independent-threshold design was run twice, with different splits of $\mathcal{D}_{\text{hold}}$. The first run (Table 13, Variant B) reports Wilson-only intervals. The second run (Table 14) halves $\mathcal{D}_{\text{hold}}$, so that the audit's threshold and the planted truth are estimated on disjoint halves ($n_{\text{threshold}} = 540{,}583$), and computes the Wilson-only and the quadrature interval on the same replicates.

Recovery and Wilson-only coverage. Table 13 gives the mean estimate, the relative bias and the coverage of the Wilson-only interval in each cell. In the corpus-relative design the audit is unbiased: the largest relative bias over the eight cells is 0.5%, at planted ratios spanning a factor of 17. Coverage is 0.940–0.964 against a nominal 0.95. In the independent-threshold design the threshold draw adds variance that the Wilson-only interval does not include, and coverage falls to 0.836–0.906.

Table 13: Planted-truth recovery with Wilson-only intervals, $B = 500$ replicates per cell, $M = 100{,}000$. Variant A is the corpus-relative design (planted ratio defined at the audit's own held-out threshold). Variant B is the independent-threshold design, first run (planted ratio defined on an independent split of $\mathcal{D}_{\text{hold}}$).

Variant	p	planted ρ	mean $\hat{\rho}$	rel. bias	coverage
A (corpus-relative truth)	10^{-2}	0.3	0.299	-0.2%	0.940
	10^{-2}	1.0	1.001	$+0.1\%$	0.950
	10^{-2}	2.0	2.000	$+0.0\%$	0.962
	10^{-2}	5.0	5.005	$+0.1\%$	0.946
	10^{-3}	0.3	0.300	$+0.1\%$	0.950
	10^{-3}	1.0	1.004	$+0.4\%$	0.964
	10^{-3}	2.0	1.990	-0.5%	0.942
	10^{-3}	5.0	5.002	$+0.0\%$	0.948
B (independent truth)	10^{-2}	0.3	0.308	$+2.7\%$	0.842
	10^{-2}	1.0	1.002	$+0.2\%$	0.900
	10^{-2}	2.0	1.995	-0.3%	0.892
	10^{-2}	5.0	4.972	-0.6%	0.852
	10^{-3}	0.3	0.323	$+7.6\%$	0.836
	10^{-3}	1.0	1.012	$+1.2\%$	0.906
	10^{-3}	2.0	1.980	-1.0%	0.870
	10^{-3}	5.0	4.933	-1.3%	0.866

Coverage under independent thresholds. The quadrature certificate of Eq. (1) was designed for the independent-threshold design. There the threshold contributes a relative error $\sigma_\tau = \sqrt{(1-p)/(n\,p)}$, which is 0.0135 at 10^{-2} and 0.0430 at 10^{-3} for the $n = 540{,}583$ images used here. Table 14 reports the coverage of both interval constructions on the same replicates. For every planted $\rho \geq 1$ the quadrature certificate raises coverage from 0.816–0.892 to 0.918–0.944. This range of ρ contains every over-production verdict in the paper.

At planted $\rho = 0.3$ the quadrature certificate does not reach nominal coverage. It covers at 0.850 for $p = 10^{-2}$ and at 0.866 for $p = 10^{-3}$. The cause is the construction, which truncates Q_0 exactly at τ_p. A large amount of probability mass then lies immediately below the threshold, and a small perturbation of the threshold moves an unusually large number of samples across it. Real samplers do not truncate exactly at the audited threshold, so the effect is specific to this construction at that one point. Even so, a user who reads a strong under-production verdict (such as iDDPM's 0.17 under R_1) against an independent corpus should treat the nominal 95% as roughly 85–87%.

Table 14: Coverage of the planted truth in the independent-threshold design, second run, under Wilson-only and quadrature intervals computed on the same replicates ($B = 500$ per cell). The audit's threshold and the planted truth are estimated on disjoint halves of $\mathcal{D}_{\text{hold}}$ ($n_{\text{threshold}} = 540{,}583$).

p	planted ρ	σ_τ	coverage (Wilson)	coverage (quadrature)
10^{-2}	0.3	0.0135	0.836	0.850
10^{-2}	1.0	0.0135	0.890	0.928
10^{-2}	2.0	0.0135	0.892	0.944
10^{-2}	5.0	0.0135	0.850	0.938
10^{-3}	0.3	0.0430	0.832	0.866
10^{-3}	1.0	0.0430	0.878	0.918
10^{-3}	2.0	0.0430	0.858	0.930
10^{-3}	5.0	0.0430	0.816	0.922

Wilson-only coverage in the two runs. The two Wilson-coverage summaries (Variant B of Table 13 and the Wilson column of Table 14) use different held-out splits, and they differ by up to 0.05 (at $p = 10^{-3}$, $\rho = 5$: 0.866 against 0.816). Comparisons between interval constructions use the common replicates of Table 14, and these are the coverage values quoted in the main text.

Image-level equivalence. The mixtures are built in sample space but audited through the severity map. Recomputing severities from the physically mixed images reproduces the severity-space exceedance count exactly, and both routes give $\hat{\rho} = 2.05$.

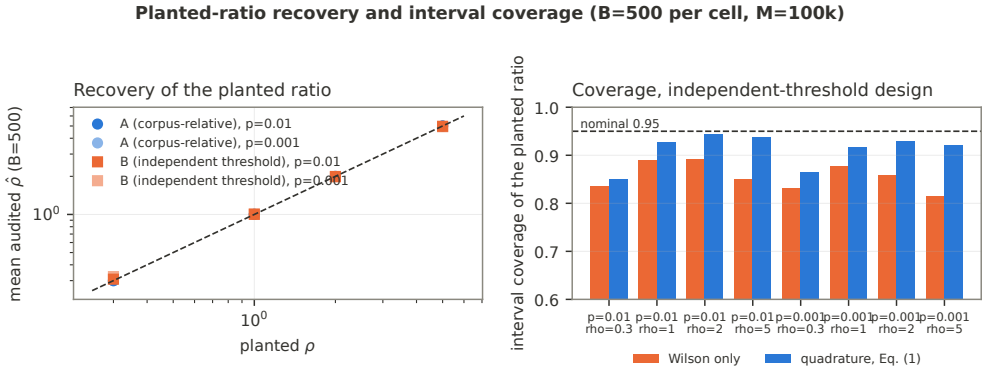

Figure 6: Planted-truth recovery and interval coverage ($B = 500$ replicates per cell, $M = 100{,}000$). Left: mean $\hat{\rho}$ against the planted ρ for both designs and both depths, on logarithmic axes; the dashed line marks exact recovery. Right: coverage of the planted truth in the independent-threshold design for each depth and planted ρ (values of Table 14); the dashed line marks the nominal 0.95. Legend: "Wilson only" is the counting interval alone and "quadrature" is the interval of Eq. (1).

G Downstream Estimates, Sensitivity, Sampling Budget, and Component Ablations

G.1 Downstream estimates by corpus size

Lemma 1 states that a plug-in rare-event estimate computed on synthetic data is biased by exactly ρ at every corpus size. To measure this, we draw bootstrap synthetic corpora of size n from an audited cache, compute $\hat{p}_n = \frac{1}{n}\#\{R(X_i) \geq \tau_p\}$, and report $\hat{p}_n/p$. Each point uses $B = 500$ corpora. The sizes n are logarithmically spaced from 10^3 to 10^5, and the $n = 10^5$ endpoint is the full cache. There are three arms: the raw sampler, the wrapped sampler (severity remap), and a real-data control drawn from held-out images. We use iDDPM ($\rho = 0.17$) and EDM ($\rho = 1.81$), the most miscalibrated configuration in each direction among those with $M = 10^5$ (EDM-churn, at 1.99, has only $M = 15{,}104$), so that the effect is visible against the bootstrap noise.

Table 15: Mean $\hat{p}_n/p$ over $B = 500$ bootstrap corpora of size n, with the s.d. in parentheses, for severity R_1 at two depths p. Arms are the raw sampler, the wrapped sampler, and a real-data control drawn from held-out images.

p	arm	n=10^3	n=3162	n=10^4	n=31623	n=10^5
10^{-3}	iDDPM raw	0.186 (0.42)	0.185 (0.24)	0.161 (0.13)	0.170 (0.07)	0.170 (0.00)
	iDDPM wrapped	1.012 (0.97)	0.970 (0.54)	0.969 (0.32)	0.984 (0.18)	0.990 (0.00)
	EDM raw	1.726 (1.36)	1.863 (0.75)	1.834 (0.43)	1.809 (0.23)	1.810 (0.00)
	EDM wrapped	0.958 (0.99)	0.977 (0.54)	0.979 (0.32)	1.007 (0.18)	0.990 (0.00)
	real control	0.992 (0.97)	0.975 (0.55)	0.971 (0.31)	1.000 (0.17)	1.008 (0.10)
10^{-4}	iDDPM raw	0.180 (1.33)	0.108 (0.61)	0.082 (0.27)	0.102 (0.18)	0.100 (0.00)
	iDDPM wrapped	0.600 (2.54)	0.607 (1.41)	0.644 (0.82)	0.579 (0.43)	0.600 (0.00)
	EDM raw	1.680 (4.15)	1.790 (2.36)	1.614 (1.22)	1.725 (0.76)	1.700 (0.00)
	EDM wrapped	0.600 (2.54)	0.683 (1.53)	0.618 (0.82)	0.583 (0.44)	0.600 (0.00)
	real control	1.220 (3.45)	1.037 (1.86)	1.026 (1.01)	1.022 (0.56)	0.994 (0.33)

Table 15 and Figure 7 show the same pattern in every arm. The mean of $\hat{p}_n/p$ does not depend on n, and only its dispersion shrinks as n grows. The raw arms stay at the audited ρ. At $n = 10^5$

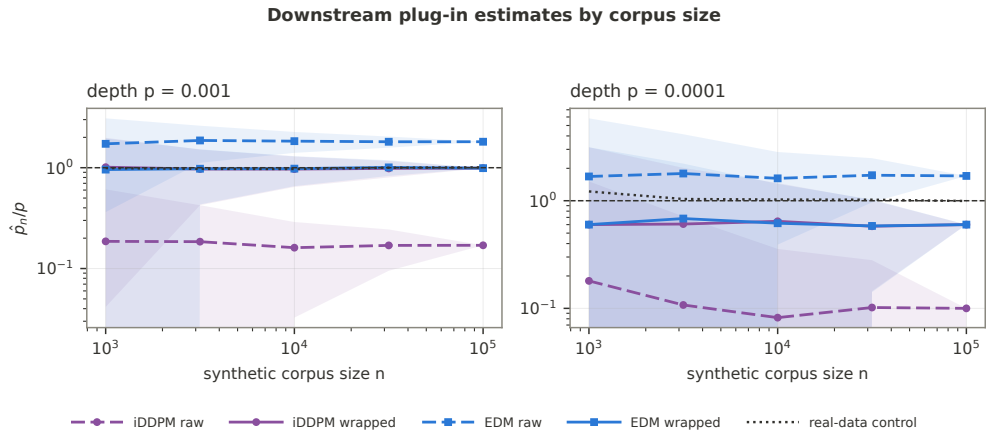

Figure 7: Mean $\hat{p}_n/p$ against corpus size n for the arms of Table 15: raw sampler (dashed), wrapped sampler (solid), real-data control (dotted). Shaded bands are ± 1 s.d. over $B = 500$ bootstrap corpora.

the iDDPM plug-in estimate is still 0.17 of the reference probability, which is exactly its audited ρ. More synthetic data therefore does not remove the bias, and a rare-event risk estimate computed on generated scenarios is off by the factor ρ. The real-data control stays at 1.

The wrapped arms stay at the calibration-law ratio $e(p)$, so the wrapper corrects the estimand up to the accuracy of the calibration law $\widehat{F}$. At 10^{-3} the wrapped means are 0.96–1.01, inside the $[0.9, 1.1]$ criterion for $n \geq 10^4$. At 10^{-4} they are 0.58–0.68, because $e(10^{-4}) = 0.688$ on this split and the rank grid of each n quantizes it (0.60 at $n = 10^5$; App. D). The shortfall at this depth is that of $\widehat{F}$, which the wrapper inherits.

G.2 Sensitivity to basis dimension, data split, and GPD threshold selection

The audit has three free analysis settings: the dimension K of the whitened basis that defines R_1, the seed of the split into $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{hold}}$, and the rule that selects the generalized-Pareto threshold. We re-ran all eight configurations under six variants and compared each variant with the default setting (Table 16, Figure 8). Three criteria were prespecified. The first is that no verdict changes from under-production to over-production or the reverse. The second is that at least 80% of verdicts are retained and that any change is adjacent, that is, to or from "calibrated". The third is that the per-configuration median $|\log(\hat{\rho}_v/\hat{\rho}_{\text{default}})|$ is at most $\log 1.25 = 0.223$. All three are met. There are zero sign flips and the maximum median drift is 0.090. The retention criterion was specified over all 48 comparisons, where it gives $42/48 = 87.5\%$ with every change adjacent. The two fixed-q variants cannot change a verdict at $p = 10^{-3}$ (see below), so we also report the stricter count over the other four variants, $26/32 = 81\%$, which meets the same threshold.

The reference for the first criterion is the default verdict. No variant changes an under-production default verdict to over-production, or the reverse. A configuration whose default interval contains 1 can still fall on either side of 1 under different variants. StyleGAN-XL, with default 1.03, is read as under-producing at $K{=}30$ (0.750) and as over-producing at $K{=}100$ (1.330), and each of these is an adjacent change from "calibrated". The six verdict changes fall on EDM2-S, EDM2-M and StyleGAN-XL, that is, on 3 of the 8 configurations, and occur only under the basis-dimension and split-seed variants. They are not confined to configurations whose default interval contains 1. The default interval of EDM2-S excludes 1 and its verdict changes, while the default interval of DiT-cfg1.5 contains 1 and its verdict never changes.

Threshold-selection variants at $p = 10^{-3}$. The two fixed-q columns of Table 16 reproduce the default column exactly. This agreement holds by construction and is no evidence of robustness. At $p = 10^{-3}$ the audited threshold τ_p is an empirical quantile of the held-out split $\mathcal{D}_{\text{hold}}$, so changing q does not change it. The threshold rule affects the GPD tail of the calibration law, and Table 17

Table 16: $\hat{\rho}_{R_1}(10^{-3})$ under the default setting and six analysis variants. Marks give the verdict read from the 95% Wilson interval, the interval prespecified for this stability check: $\downarrow$ under-producing, $\uparrow$ over-producing, unmarked calibrated. The last column, "drift", is the median over the six variants of $|\log(\hat{\rho}_v/\hat{\rho}_{\text{default}})|$, which the third stability criterion requires to be at most $\log 1.25 = 0.223$.

Configuration	default	K=30	K=100	seed 8001	seed 8002	q=0.98	q=0.995	drift
iDDPM	$0.17^{\downarrow}$	$0.14^{\downarrow}$	$0.15^{\downarrow}$	$0.17^{\downarrow}$	$0.17^{\downarrow}$	$0.17^{\downarrow}$	$0.17^{\downarrow}$	0.000
EDM	$1.81^{\uparrow}$	$1.53^{\uparrow}$	$1.61^{\uparrow}$	$1.75^{\uparrow}$	$1.82^{\uparrow}$	$1.81^{\uparrow}$	$1.81^{\uparrow}$	0.020
EDM-churn	$1.99^{\uparrow}$	$1.59^{\uparrow}$	$1.99^{\uparrow}$	$1.99^{\uparrow}$	$1.99^{\uparrow}$	$1.99^{\uparrow}$	$1.99^{\uparrow}$	0.000
EDM2-S	$1.47^{\uparrow}$	1.17	$1.20^{\uparrow}$	$1.40^{\uparrow}$	$1.48^{\uparrow}$	$1.47^{\uparrow}$	$1.47^{\uparrow}$	0.028
EDM2-M	0.82	1.04	$0.77^{\downarrow}$	$0.79^{\downarrow}$	$0.80^{\downarrow}$	0.82	0.82	0.031
StyleGAN-XL	1.03	$0.75^{\downarrow}$	$1.33^{\uparrow}$	0.98	1.04	1.03	1.03	0.030
DiT-cfg1.5*	0.99	0.94	0.63	0.99	1.04	0.99	0.99	0.026
DiT-cfg1.0*	$0.67^{\downarrow}$	$0.44^{\downarrow}$	$0.42^{\downarrow}$	$0.60^{\downarrow}$	$0.63^{\downarrow}$	$0.67^{\downarrow}$	$0.67^{\downarrow}$	0.090

* The DiT rows of this table use $M = 19{,}200$ (cfg1.5) and $M = 43{,}200$ (cfg1.0). The corresponding rows of Table 1 use $M = 50{,}000$, so the default entries read 0.99 and 0.67 here and 0.98 and 0.64 there. The sensitivity criteria concern the movement of a row relative to its own default entry.

reports that effect separately. Across the six variants the fitted shape $\hat{\xi}$ ranges from -0.057 to $+0.028$. Among the variants that share the default basis and split, the extrapolated $\widehat{F}^{-1}(1-10^{-4})$ moves by at most 0.5% (15.836 to 15.915). The K variants change the severity axis itself, so their thresholds are not comparable in absolute units. Their verdicts are comparable, and Table 16 compares those.

Table 17: Generalized-Pareto fit under the default setting and the same six variants: selected quantile level q^*, threshold u, shape $\hat{\xi}$, scale $\hat{\sigma}$, KS score (App. C.2), and extrapolated threshold. $\kappa(10^{-4}) = P_{\mathcal{D}_{\text{hold}}}(R \geq \widehat{F}^{-1}(1-10^{-4}))/10^{-4}$ is the forward diagnostic of App. D. It is a different quantity from the calibration-law ratio e that the wrapper inherits.

Variant	q^*	u	$\hat{\xi}$	$\hat{\sigma}$	KS score	$\widehat{F}^{-1}(1-10^{-4})$	$\kappa(10^{-4})$
default	0.9925	12.06	-0.050	0.974	0.970	15.839	1.286
K=30	0.99	9.25	-0.043	0.829	0.982	12.714	1.128
K=100	0.995	17.27	$+0.028$	1.238	0.989	22.389	0.934
seed 8001	0.9925	12.02	$+0.001$	0.918	0.844	15.996	1.082
seed 8002	0.9925	12.08	-0.057	1.012	0.995	15.956	1.212
q=0.98 (fixed)	0.98	11.07	-0.035	1.003	0.186	15.915	1.239
q=0.995 (fixed)	0.995	12.45	-0.053	0.958	0.888	15.836	1.286

The fixed $q = 0.98$ rule gives a KS score of 0.186, against 0.970 for the adaptively chosen threshold; both exceed the minimum score of 0.05 that we require of a fit. Since the parameters are estimated from the same excesses, these values rank candidate thresholds and are not calibrated p-values.

G.3 Interval width by sampling budget

We subsample each audit cache to $M' \in \{5\text{k}, 10\text{k}, 20\text{k}, 50\text{k}, 100\text{k}\}$ draws, with $B = 200$ resamples per point, and record the point estimate and the width of the 95% Wilson interval (Table 18, Figure 9). The point estimates stay centered at every budget. At $M' = 50$k and $p = 10^{-3}$, the mean estimate is within 0.65% of the full-cache value. The interval width scales as $1/\sqrt{M'p}$.

The $M' = 50$k row is the plug-in baseline that §6 refers to. At this community-standard evaluation budget, the interval of a plug-in tail count is 0.24–0.75 ρ-units wide at 10^{-3} and 0.93–2.42 ρ-units wide at 10^{-4}. At that depth the uncertainty is of the same order as the factor-two miscalibrations found among the audited configurations, so a 50k evaluation does not reliably resolve them: whether an interval excludes 1 depends on the realized count. For example, a sampler with $\rho = 2$ has 10 expected exceedances at $M' = 50$k and $p = 10^{-4}$, and at exactly 10 the Wilson interval is $[1.09, 3.68]$, while at 6 it includes 1. These are measured widths; Proposition 2 gives the scaling of the required sample size, not a threshold for this budget. This is why the audit spends its whole sampling budget on the tail indicator.

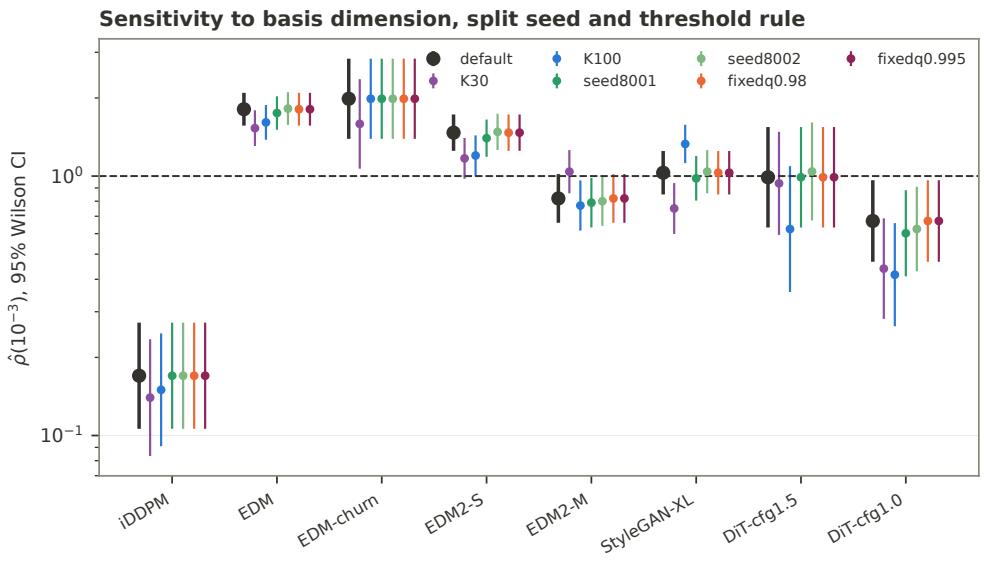

Figure 8: $\hat{\rho}_{R_1}(10^{-3})$ with 95% Wilson intervals under the seven analysis settings of Table 16 (the default and six variants), shown per configuration. The DiT sample sizes are those given below Table 16.

Table 18: Interval width by sampling budget M', as ranges across configurations ($B = 200$ subsamples per cell). "Width" is the mean 95% Wilson width in ρ units. "Covers full" is the fraction of subsample intervals that contain the full-cache estimate. A cache of M draws cannot supply a row with $M' > M$, so fewer configurations enter at large M'. EDM-churn ($M = 15{,}104$) is absent above $M' = 10{,}000$. The two DiT caches ($M = 50{,}000$) supply the $M' = 50{,}000$ row and are absent above it.

	$p = 10^{-3}$		$p = 10^{-4}$		
Budget M'	width	covers full	width	covers full	#configs
5,000	1.02–2.57	0.93–0.99	7.89–10.86	0.89–0.99	8
10,000	0.61–1.79	0.95–1.00	4.05–6.67	0.87–1.00	8
20,000	0.41–1.19	0.96–1.00	2.09–4.05	0.96–1.00	7
50,000	0.24–0.75	0.99–1.00	0.93–2.42	1.00	7
100,000	0.17–0.53	1.00	0.55–1.66	1.00	5

G.4 Component ablations

Table 19 compares threshold estimates, wrapped rates, and transport residuals after removing individual components.

Transport chart. Removing the chart degrades the transport. We replace the damped alternating-projection realization with a one-shot rescale, or drop the chart and work in raw pixels. Either change leaves transported samples about two severity units from their targets, and no sample lands within 0.05 of its target, against 71–76% for the full construction.

Generalized-Pareto splice. The splice gives no gain in point accuracy at depths the calibration split $\mathcal{D}_{\text{cal}}$ reaches. At 10^{-4} its threshold estimate is within 0.02 severity units of the empirical quantile of $\mathcal{D}_{\text{cal}}$, and both are about 0.3 units below the held-out reference. At 10^{-5} the empirical quantile is the closer of the two. In the wrapper, the rates with and without the splice agree to 0.02 in ρ at 10^{-3} and exactly at 10^{-4}. We therefore do not claim that the extreme-value layer improves the audit's point estimates. Its use is in interval construction and labeled extrapolation: it supplies profile-likelihood intervals and explicitly labeled extrapolation beyond the corpus floor (App. C).

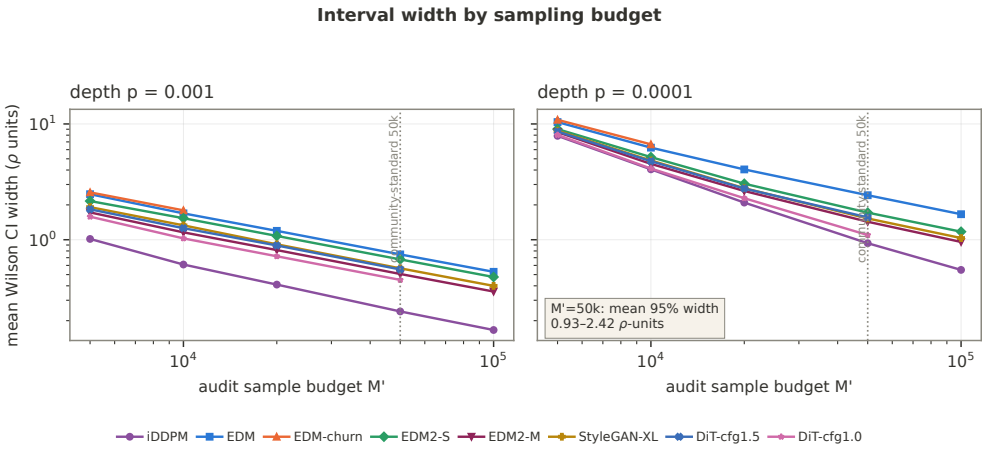

Figure 9: Mean 95% Wilson interval width, in ρ units, against sampling budget M', per configuration, at $p = 10^{-3}$ and $p = 10^{-4}$ ($B = 200$ subsamples per point). The dotted line marks the community-standard 50k budget.

Table 19: Component ablations. Thresholds and transport residuals are in severity units, and wrapped rates are in ρ units. "Truth" is the held-out reference threshold. "Overshoot" and "truncated" denote over-production and under-production. The transport rows come from a separate run of the full construction, so their residuals differ slightly from the per-configuration values of Table 22.

Component	Condition	Variant	Result
GPD audit threshold	$p = 10^{-4}$ (truth 16.13)	spliced GPD	15.84
		fit-split empirical	15.82
	$p = 10^{-5}$ (truth 18.11)	spliced GPD	17.54
		fit-split empirical	17.74
GPD in the wrapper	all three samplers, 10^{-3}	on / off	0.99 / 1.01
	all three samplers, 10^{-4}	on / off	0.60 / 0.60
Transport chart	EDM (overshoot)	full (damped alt. proj.)	0.0002 / 71.4%
		one-shot; pixel-space	1.88, 1.98 / 0%
	iDDPM (truncated)	full (damped alt. proj.)	0.0000 / 76.1%
		one-shot; pixel-space	2.19, 2.49 / 0%

Transport rows report median $|R(\phi(x)) - r'|$ / fraction of transported samples within 0.05 of their severity target.

With these, a depth the corpus cannot reach is reported as a flagged interval, as in the CelebA-HQ 10^{-4} rows of App. I, and not as an unlabeled point estimate.

H Per-Configuration Gallery and Shape Evidence

H.1 Shared calibration-law constants and per-configuration statistics

All eight configurations are audited with one calibration law $\widehat{F}$ and the same two thresholds (constants in Table 28). On $\mathcal{D}_{\text{cal}}$ the severity quantiles of the real images are $q_{50} = 6.72$, $q_{99} = 11.76$ and $q_{99.9} = 13.94$. The held-out split contains 109 images beyond $\tau_{10^{-4}}$ and 1,082 beyond $\tau_{10^{-3}}$. Conditional on exceeding $\tau_{10^{-4}}$, the held-out severity has $q_{50} = 16.81$ and $q_{90} = 18.11$. On the principal-component basis fitted to the tail of $\mathcal{D}_{\text{cal}}$ (the calibration-tail basis), the held-out dispersions are $\hat{\sigma}_{\text{PC1}} = 1.68$ and $\hat{\sigma}_{\text{PC2}} = 1.29$. Since $\widehat{F}$ and the thresholds are common to all configurations, every cross-model comparison is made on the same axis.

Table 20 compares bulk and tail summaries of severity across the eight configurations. The median severity is nearly the same everywhere (q_{50} between 6.13 and 7.12, against 6.72 for real images),

Table 20: Per-configuration severity statistics on the shared R_1 axis. $k(\tau_3)$ and $k(\tau_4)$ are the exceedance counts at $\tau_{10^{-3}}$ and $\tau_{10^{-4}}$. The columns q are severity quantiles of the configuration's own samples, and "tail range" is the minimum and maximum severity among the kept top-K samples. The last four columns describe the matched-severity panels of §H.8, in which these top-K samples are transported to target severities drawn from the generalized-Pareto tail of $\widehat{F}$ above $\tau_{10^{-4}}$: PC1 and PC2 are standard deviations of the transported samples on the fixed calibration-tail basis (held-out real exceedances of $\tau_{10^{-3}}$: 1.68 and 1.29), and "cond." are the median and 90th percentile of their realized severities (held-out real images above $\tau_{10^{-4}}$: 16.81 and 18.11). These columns therefore reflect the target law and the transport, not the sampler's own tail.

Config	M	top-K	$k(\tau_3)$	$k(\tau_4)$	q_{50}	q_{99}	$q_{99.9}$	tail range	PC1	PC2	cond. q_{50}	cond. q_{90}
iDDPM	100k	1000	17	1	6.13	10.75	12.54	10.75–17.27	0.73	1.01	16.56	17.66
EDM	100k	1000	181	17	7.04	12.33	14.46	12.33–19.05	0.85	1.09	16.55	17.72
EDM-churn	15.1k	250	30	3	7.12	12.32	14.53	11.77–16.62	0.77	1.18	16.43	17.65
EDM2-S	100k	1000	147	8	6.98	12.14	14.24	12.14–18.11	0.93	1.12	16.53	17.67
EDM2-M	100k	1000	82	5	6.84	11.85	13.77	11.85–17.09	0.95	1.17	16.49	17.65
StyleGAN-XL	100k	1000	103	6	6.82	11.81	13.95	11.81–17.12	0.64	1.07	16.53	17.72
DiT-cfg1.5	50k	500	49	3	6.65	11.75	13.89	11.75–16.25	0.73	1.07	16.50	17.69
DiT-cfg1.0	50k	500	32	1	6.55	11.38	13.27	11.38–16.46	0.70	1.03	16.49	17.76

while the exceedance count at $\tau_{10^{-3}}$ ranges from 17 to 181. Bulk summaries of severity therefore do not separate configurations whose tail rates differ tenfold, which is consistent with the phenomenon behind Proposition 1. The dispersion test of §H.7 compares the sampler's own tail directions with those of real images.

H.2 Certificates under the three severity functionals

Table 1 reports point estimates only for R_2 and R_3. Table 21 adds the certificates, and Figure 10 plots them. The three functionals are built by the same recipe on different representations. Each must pass the same operability gate before it is used for a leaderboard: a goodness-of-fit test for the GPD tail of its calibration law, and a self-audit whose certificate contains 1 at 10^{-2}, 3×10^{-3} and 10^{-3}.

Table 21: Per-(configuration, functional) certificates at $p = 10^{-3}$: $\hat{\rho}$ with 95% quadrature intervals, exceedance count k, and the verdict under that functional. Verdict agreement is $3/8$ (R_1–R_2), $4/8$ (R_1–R_3) and $3/8$ (R_2–R_3); Kendall τ_b is -0.36, $+0.57$ and -0.07 respectively.

	R_1 (pixel PCA-50)		R_2 (MAE ViT-B/16)		R_3 (max-patch)	
Config	k	$\hat{\rho}$ [95%]	k	$\hat{\rho}$ [95%]	k	$\hat{\rho}$ [95%]
iDDPM	17	0.17 [0.11, 0.27] *under*	155	1.55 [1.31, 1.83] *over*	32	0.32 [0.22, 0.45] *under*
EDM	181	1.81 [1.54, 2.11] *over*	93	0.93 [0.75, 1.15]	147	1.47 [1.23, 1.74] *over*
EDM-churn	30	1.99 [1.38, 2.84] *over*	8	0.53 [0.27, 1.05]	13	0.86 [0.50, 1.47]
EDM2-S	147	1.47 [1.23, 1.74] *over*	84	0.84 [0.67, 1.05]	89	0.89 [0.72, 1.10]
EDM2-M	82	0.82 [0.65, 1.02]	117	1.17 [0.96, 1.41]	39	0.39 [0.28, 0.53] *under*
StyleGAN-XL	103	1.03 [0.84, 1.26]	103	1.03 [0.84, 1.26]	186	1.86 [1.59, 2.17] *over*
DiT-cfg1.5	49	0.98 [0.73, 1.30]	36	0.72 [0.52, 1.00] *under*	37	0.74 [0.53, 1.02]
DiT-cfg1.0	32	0.64 [0.45, 0.91] *under*	32	0.64 [0.45, 0.91] *under*	20	0.40 [0.26, 0.62] *under*

Rounding note. For DiT-cfg1.5 under R_2 the quadrature upper limit is 0.99988, which is printed above as 1.00. The interval excludes 1 by 1.2×10^{-4}, so the verdict is *under*. This is the only cell in the paper whose verdict cannot be read from the printed interval, and it should be treated as a boundary case.

Real-versus-real FMD floor. The FMD column of Table 1 is measured against one fixed real reference set. As a floor we compute the FMD of a disjoint real sample of the same size M against that reference. The floor is 0.003545 at $M = 100{,}000$, 0.008012 at $M = 50{,}000$ and 0.019878 at $M = 15{,}104$. Every configuration's FMD is at least 14 times the floor at its own M; the smallest ratio is that of EDM-churn (FMD 0.28). Figure 1 plots the FMD, FID and $\hat{\rho}$ columns of Table 1.

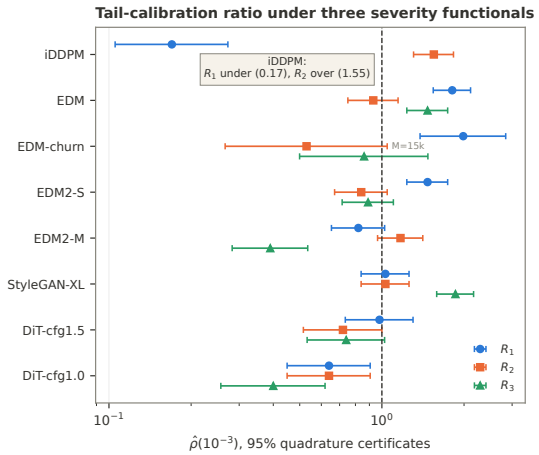

Figure 10: $\hat{\rho}(10^{-3})$ with 95% quadrature certificates under the three severity functionals, for the eight configurations. Verdict agreement is $3/8$ (R_1–R_2), $4/8$ (R_1–R_3) and $3/8$ (R_2–R_3). iDDPM is under-producing under R_1 and R_3 and over-producing under R_2. Values are in Table 21.

H.3 Rank correlations between quality axes and tail calibration

This subsection gives the rank correlations between the two quality axes and $|\log \hat{\rho}_{R_1}(10^{-3})|$, and their sensitivity. All of them are descriptive, since at most eight configurations enter. Published FID and measured FMD rank the configurations differently (Spearman $+0.11$, p-value 0.82). On the seven configurations that have both axes, the Spearman correlation with $|\log \hat{\rho}(10^{-3})|$ is $+0.68$ (p-value 0.09) for published FID and $+0.32$ for FMD. We quote this matched pair because the two axes are otherwise defined on different sets of configurations. On all eight configurations the FMD correlation is $+0.05$ (p-value 0.91), and $+0.14$ when every configuration is subsampled to the common $M = 15{,}104$. Under leave-one-out deletion the published-FID correlation stays between $+0.60$ and $+0.83$, while the FMD correlation ranges from -0.09 to $+0.49$. The FMD result is driven by iDDPM, whose FMD of 3.78 is 5.4 times the next-largest and which has the most under-producing tail, while its published 64^2 FID of 2.90 ranks sixth of seven. We draw no conclusion from the contrast between the two axes.

H.4 Transport residuals of wrapped samples

Wrapped rates are computed from the severity remap, which is exact. Wrapped images require an additional step. The transport is realized by damped alternating projection between the pixel box and the target severity level set, and this realization is approximate. Table 22 reports the residuals for each configuration. Before projection, the raw update sends 20.3–34.6% of pixel coordinates outside $[0, 255]$. After projection the median severity error is at most 0.0055 units. However, only 60.8–75.4% of samples land within 0.05 of their target, and the 95th-percentile residual is 1.8–2.9 severity units. The severity level set is not convex, so the alternating projection carries no convergence guarantee and we report its residuals instead. Every transported image shown in this appendix is therefore an approximate realization with the residual distribution of Table 22, while the wrapped rates are unaffected. These residuals belong to the panel transport, which carries a sampler's top draws to severities beyond the 10^{-4} level. The moves made by the wrapper itself are much smaller, and 99.5–100% of wrapped samples land within 0.05 of target (App. E).

H.5 Tail-calibration ratio over a ten-point depth grid

Table 1 audits two depths. Figure 11 shows $\hat{\rho}_{R_1}(p)$ at ten depths on a logarithmic grid from 10^{-2} to 10^{-4}. Over the nine depths above 10^{-4} some configurations are nearly flat: EDM stays between 1.77 and 1.92 and StyleGAN-XL between 0.86 and 1.05. Others drift: iDDPM moves between 0.15 and 0.29, EDM2-S falls from 1.52 to 1.11, and EDM2-M from 1.11 to 0.58, crossing 1. At 10^{-4} the estimates of EDM, StyleGAN-XL and iDDPM are 1.70, 0.60 and 0.10, based on counts of 17,

Table 22: Transport residuals per configuration, in severity units. The column oor_0 is the fraction of pixel coordinates outside $[0, 255]$ after the raw update, before projection.

Config	oor_0	median $\lvert R(\phi(x)) - r' \rvert$	p95	max	within 0.05
iDDPM	20.3%	0.0000	2.66	12.93	75.4%
EDM	31.8%	0.0002	1.82	6.21	71.0%
EDM-churn	34.6%	0.0055	2.86	12.13	60.8%
EDM2-S	29.9%	0.0004	2.04	12.53	69.4%
EDM2-M	29.1%	0.0008	2.78	13.25	64.5%
StyleGAN-XL	29.4%	0.0002	2.47	11.76	71.4%
DiT-cfg1.5	29.8%	0.0017	2.75	12.40	66.0%
DiT-cfg1.0	28.8%	0.0013	2.94	11.78	64.2%

6 and 1, so the change at the last depth is within the sampling variability of the estimator. A verdict therefore applies to the depth at which it is certified.

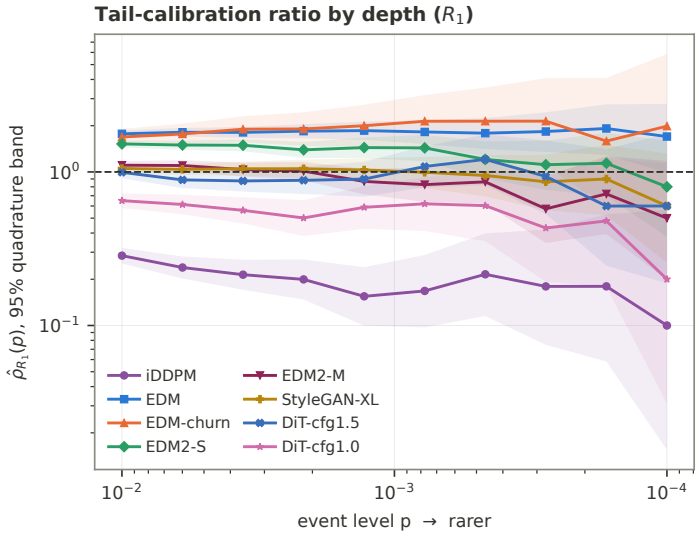

Figure 11: $\hat{\rho}_{R_1}(p)$ over a ten-point depth grid with 95% quadrature bands, one curve per configuration. Rarer events are to the right. The dashed line marks $\rho = 1$.

H.6 Most severe samples per configuration

Figure 12 shows the eight most severe samples produced by each configuration and the eight most severe held-out real images, on one R_1 scale. The exceedance counts of iDDPM and EDM at $\tau_{10^{-3}}$ differ by a factor of 10, and visual comparison of these rows does not reveal that difference. The audited quantity is a rate, and the rates are in Tables 1 and 20.

H.7 Tail dispersion test: all cells

The test compares the dispersion of generated and real tail samples along the first two principal components. The basis is fixed in advance from the top 2% of $\mathcal{D}_{\text{cal}}$ by severity. Depths are matched by rank, so that $n_g = \lceil qM \rceil$ generated samples are compared with the corresponding real tail. Equality of dispersion is tested with the Brown–Forsythe test (Levene's test with median centering), Holm-corrected over all 32 cells. It rejects in 25 of the 32 cells, including 15 of the 16 PC1 cells. As a secondary check we ran a permutation test whose statistic is the absolute log ratio of the two sample variances, with 10^4 random reassignments of the pooled projections to the two groups. Because it permutes raw projections, its null hypothesis is equality of the two distributions and not



Figure 12: The eight most severe samples of each configuration (rows) with their R_1 values. The top row shows the eight most severe held-out real images. The DiT rows use the final $M = 50{,}000$ sample sets.

of their dispersions alone. It rejects in 28 cells; the three additional cells are PC2 cells at $p = 10^{-3}$ (EDM, DiT-cfg1.5, DiT-cfg1.0). Table 23 lists every cell and Figure 13 plots them.

The PC1 ratio lies between 0.40 and 0.71 in every cell. PC2, which carries less of the calibration-tail variance, is compressed more mildly (0.68–1.00). Under-dispersion is therefore strongest along the leading direction of variation in the calibration-tail shape basis. By Proposition 3(ii) the wrapper leaves this direction unchanged.

Table 23: Tail dispersion ratios $\hat{\sigma}_{\text{gen}}/\hat{\sigma}_{\text{real}}$ with bootstrap 95% CIs, on the fixed calibration-tail PC basis. All 32 point estimates are below 1 (the largest, 0.999, prints as 1.00), and 25 reject equal dispersion under the Brown–Forsythe test at Holm-adjusted 0.05. Cells marked * are not rejected by either test; cells marked † are rejected only by the secondary permutation test.

Config	depth p	n_g	n_{real}	PC1 ratio [95% CI]	PC2 ratio [95% CI]
iDDPM	10^{-3}	100	1082	0.44 [0.34, 0.55]	0.68 [0.57, 0.78]
iDDPM	10^{-2}	1000	10812	0.55 [0.51, 0.59]	0.80 [0.76, 0.84]
EDM	10^{-3}	100	1082	0.49 [0.37, 0.61]	0.80 [0.68, 0.91]†
EDM	10^{-2}	1000	10812	0.63 [0.57, 0.68]	0.88 [0.84, 0.93]
EDM-churn	10^{-3}	16	1082	0.54 [0.32, 0.70]*	0.75 [0.51, 0.88]*
EDM-churn	10^{-2}	152	10812	0.63 [0.52, 0.74]	1.00 [0.88, 1.11]*
EDM2-S	10^{-3}	100	1082	0.64 [0.44, 0.81]	0.80 [0.66, 0.93]
EDM2-S	10^{-2}	1000	10812	0.69 [0.64, 0.75]	0.91 [0.87, 0.96]
EDM2-M	10^{-3}	100	1082	0.70 [0.49, 0.90]	0.86 [0.72, 0.98]*
EDM2-M	10^{-2}	1000	10812	0.71 [0.65, 0.77]	0.93 [0.88, 0.98]
StyleGAN-XL	10^{-3}	100	1082	0.40 [0.34, 0.46]	0.75 [0.63, 0.86]
StyleGAN-XL	10^{-2}	1000	10812	0.49 [0.46, 0.52]	0.84 [0.80, 0.88]
DiT-cfg1.5	10^{-3}	50	1082	0.44 [0.32, 0.56]	0.72 [0.58, 0.84]†
DiT-cfg1.5	10^{-2}	500	10812	0.56 [0.51, 0.62]	0.85 [0.80, 0.90]
DiT-cfg1.0	10^{-3}	50	1082	0.51 [0.37, 0.64]	0.69 [0.57, 0.79]†
DiT-cfg1.0	10^{-2}	500	10812	0.55 [0.49, 0.61]	0.83 [0.78, 0.88]

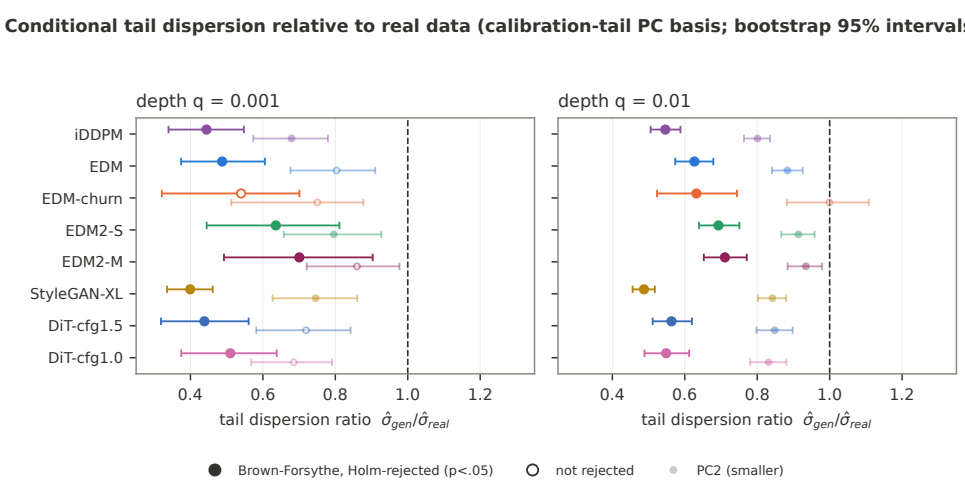

Figure 13: Tail dispersion ratios with bootstrap 95% intervals at both rank-matched depths, for PC1 and PC2. Filled markers are cells rejected by the Brown–Forsythe test after Holm correction. The PC1 ratio is below 1 for every configuration, including those whose R_1 certificate contains 1. Values are in Table 23.

H.8 Matched-severity audit panels

Figures 14–21 show one panel per configuration. The top row of each panel holds six held-out real images whose severity exceeds $\tau_{10^{-4}}$, the 1-in-10,000 level of R_1. The bottom row holds six samples of the configuration after the panel transport of Table 22, chosen so that their severities match those of the real images to two decimals. The bottom rows are approximate realizations, and the transport residuals of Table 22 apply to each of them. The panels are qualitative. The rates that determine the verdicts are in Table 1, and the sample size, kept top-K and exceedance counts of each configuration are in Table 20. EDM-churn has only $M = 15{,}104$ samples, with $k = 30$ and $k = 3$ exceedances at the two levels, so its certificates in Table 1 are wide and should be quoted with their intervals. The rates printed inside the two DiT panels equal those of Table 1.

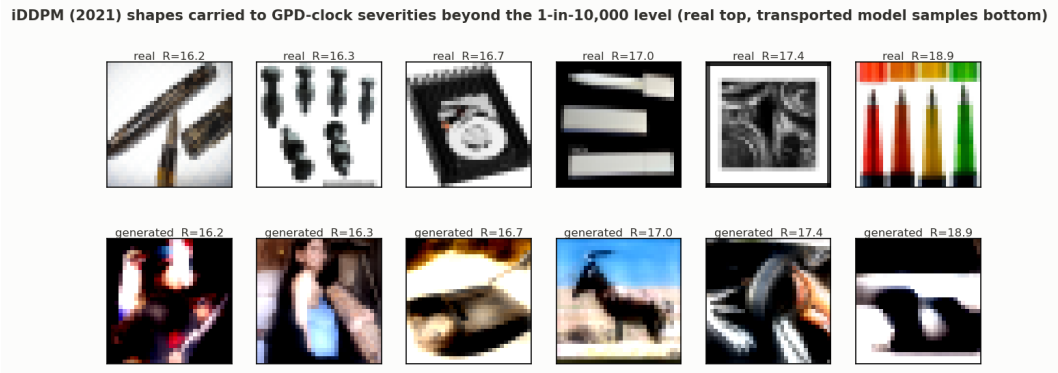

Figure 14: Matched-severity panel for iDDPM, $M = 100{,}000$, kept top-1000. Top row: six held-out real images beyond $\tau_{10^{-4}}$ under R_1. Bottom row: six transported iDDPM samples with severities matched to the real images to two decimals. The bottom row is an approximate realization with the transport residuals of Table 22. iDDPM has $k = 17$ exceedances at $\tau_{10^{-3}}$ and $k = 1$ at $\tau_{10^{-4}}$ (rates 1.70×10^{-4} and 1.00×10^{-5}), the lowest of the eight configurations under R_1.

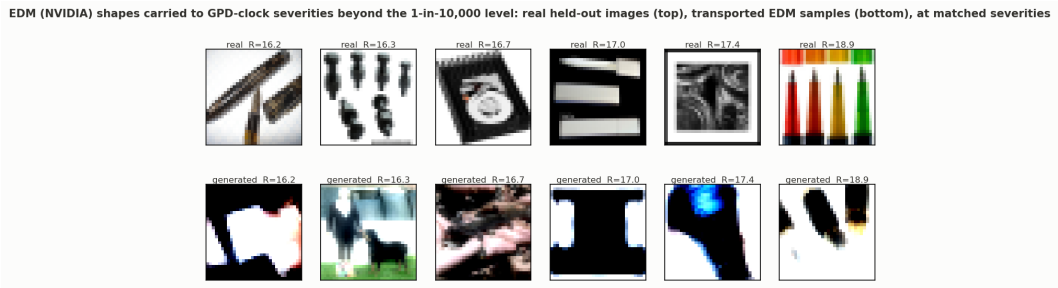

Figure 15: EDM (deterministic Heun sampler), $M = 100{,}000$, kept top-1000. Layout as in Figure 14.

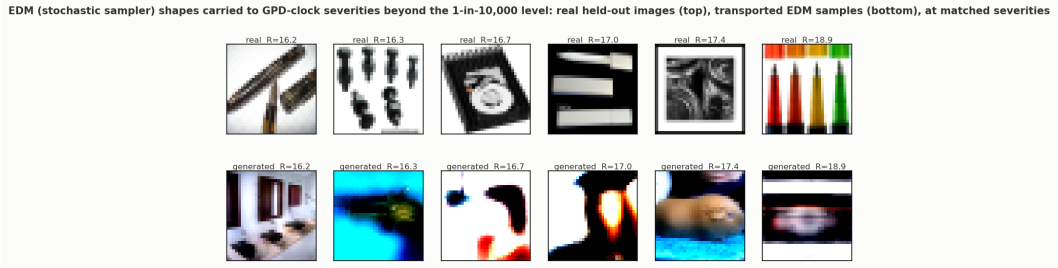

Figure 16: EDM-churn (same network as EDM, stochastic sampler), $M = 15{,}104$, kept top-250. Layout as in Figure 14.

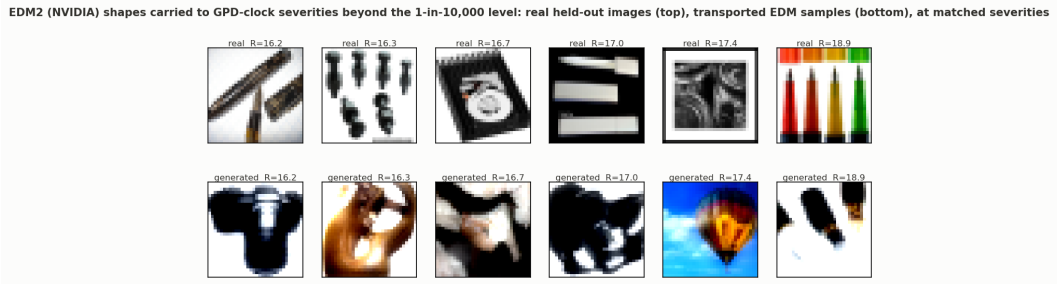

Figure 17: EDM2-S, $M = 100{,}000$, kept top-1000. Layout as in Figure 14.

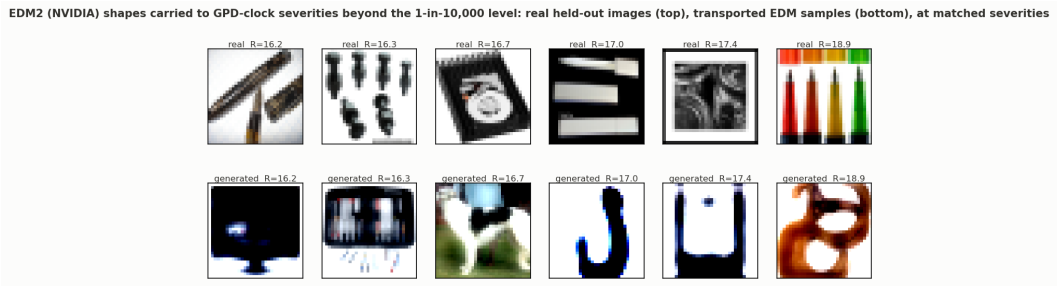

Figure 18: EDM2-M, $M = 100{,}000$, kept top-1000. Layout as in Figure 14.

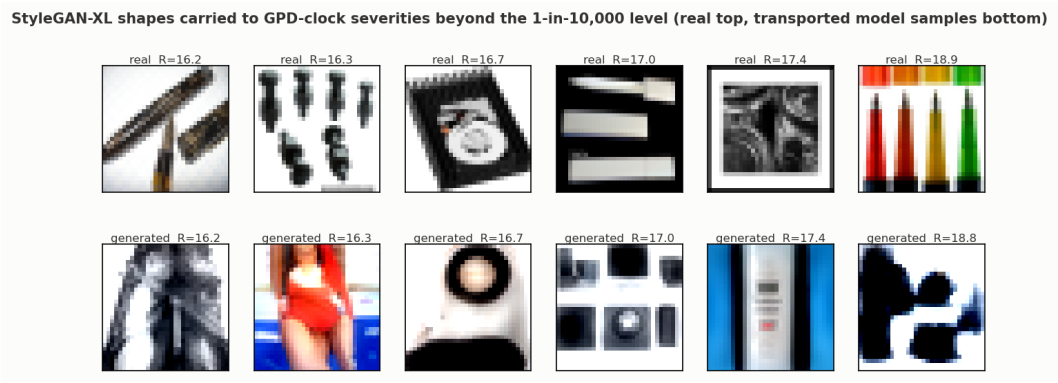

Figure 19: StyleGAN-XL, $M = 100{,}000$, kept top-1000. Layout as in Figure 14.

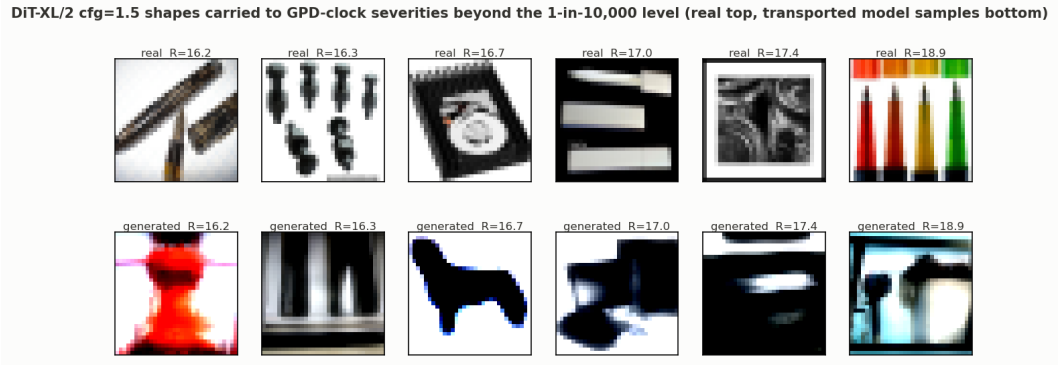

Figure 20: DiT-XL/2 at guidance 1.5, $M = 50{,}000$, kept top-500. Layout as in Figure 14. The printed rates (9.80×10^{-4}, $k = 49$; 6.00×10^{-5}, $k = 3$) equal those of Table 1.

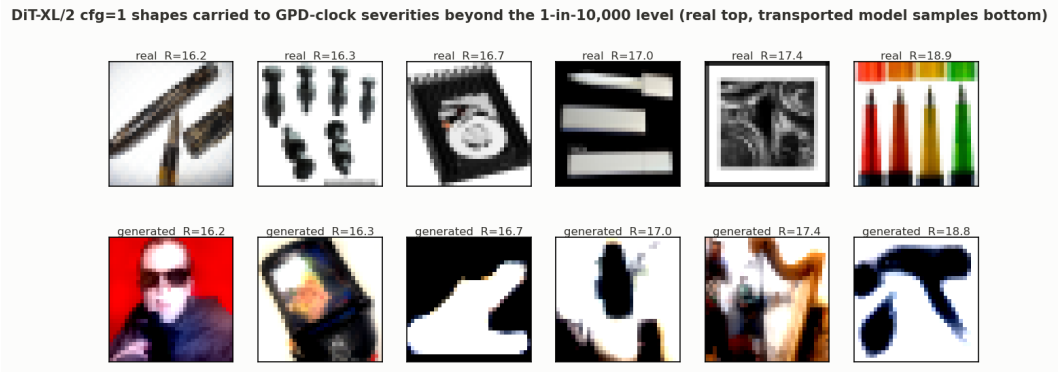

Figure 21: DiT-XL/2 without guidance, $M = 50{,}000$, kept top-500. Layout as in Figure 14. The printed rates, 6.40×10^{-4} ($k = 32$) and 2.00×10^{-5} ($k = 1$), equal those of Table 1.

H.9 ILLUSTRATIVE FIGURES ON TOY LAWS WITH CLOSED-FORM TAILS

The figures in this subsection are drawn on toy laws whose tail probabilities are known in closed form. They illustrate the construction and provide no evidence about any image model. Two caveats apply to all of them. First, they use fixed generalized-Pareto thresholds, because they were made before the adaptive maximum-KS rule used in the rest of the paper. They therefore illustrate the splice of bulk and GPD tail, and they do not illustrate threshold selection. Second, the ratio they report is $\hat{p}/p$, the ratio of the calibration law's tail probability to the exact one. This is a different quantity from the tail-calibration ratio ρ of a sampler, which has a different denominator.

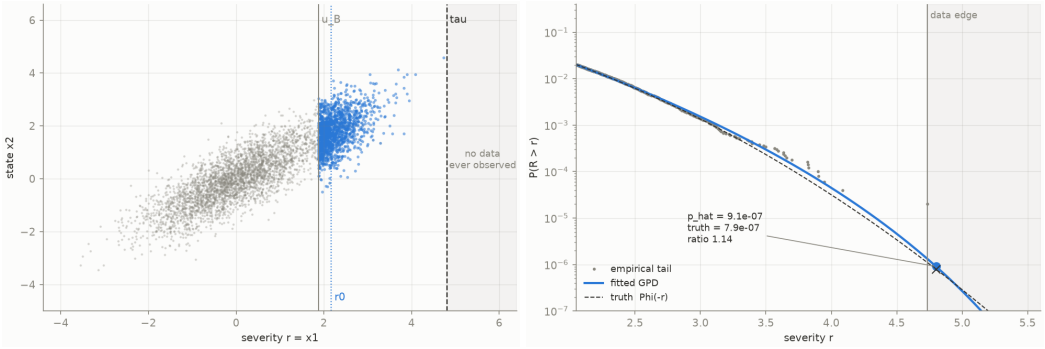

Figure 22: A bivariate Gaussian with $R(x) = x_1$ and a query level beyond every observation. Left: no sample lies at the queried depth, which is the setting of Proposition 2. Right: the parametric tail of the calibration law continues past the largest observation, and its extrapolation is compared with the closed-form tail probability.

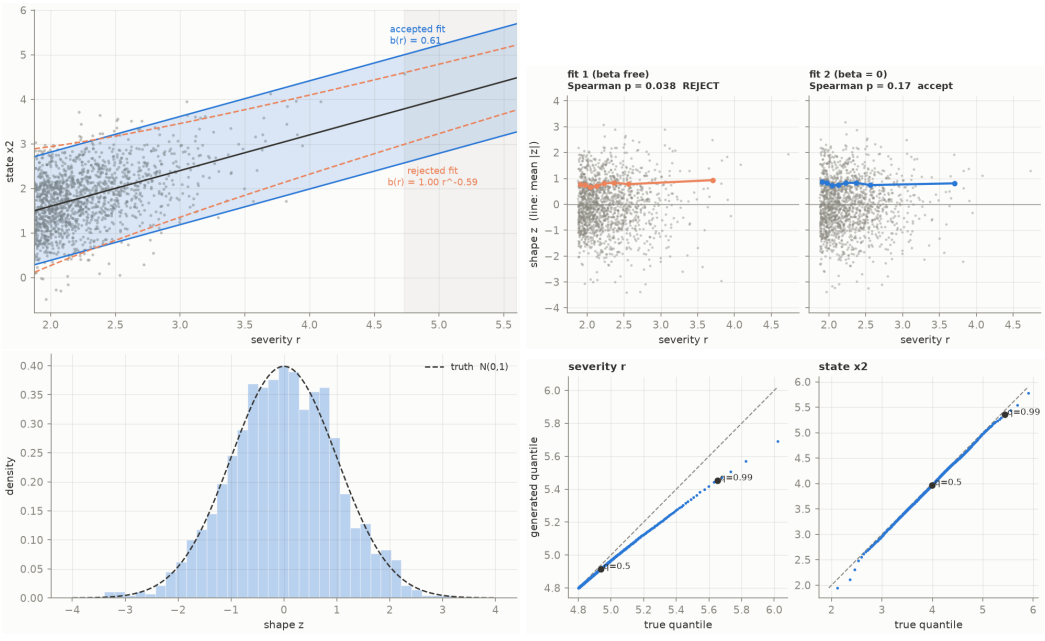

Figure 23: Operability diagnostics of the calibration law on the toy law. Top left: accepted and rejected center and spread fits. Top right: the independence diagnostic. Bottom left: the pooled shape law against the exact one. Bottom right: severity and state QQ plots against the exact law. The audit's operability gate automates these checks.

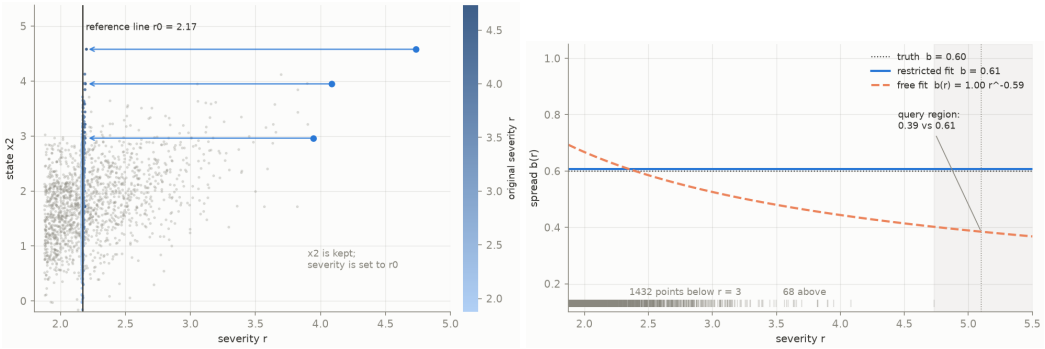

Figure 24: Left: the transport applied point by point on the toy law. The severity of each point is remapped and the transport-invariant shape coordinate is unchanged, as stated in Proposition 3(ii). Right: the spread curve and the query region into which it is extrapolated.

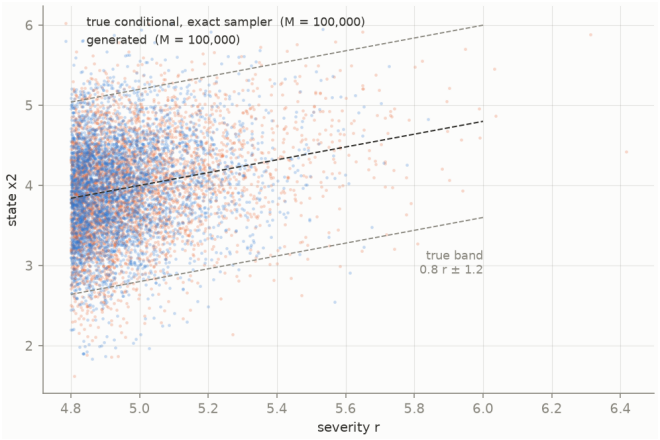

Figure 25: A generated rare set on the toy law, at a depth where direct sampling needs about a hundred million draws to observe a few hundred exceedances, far beyond a 50k evaluation budget.

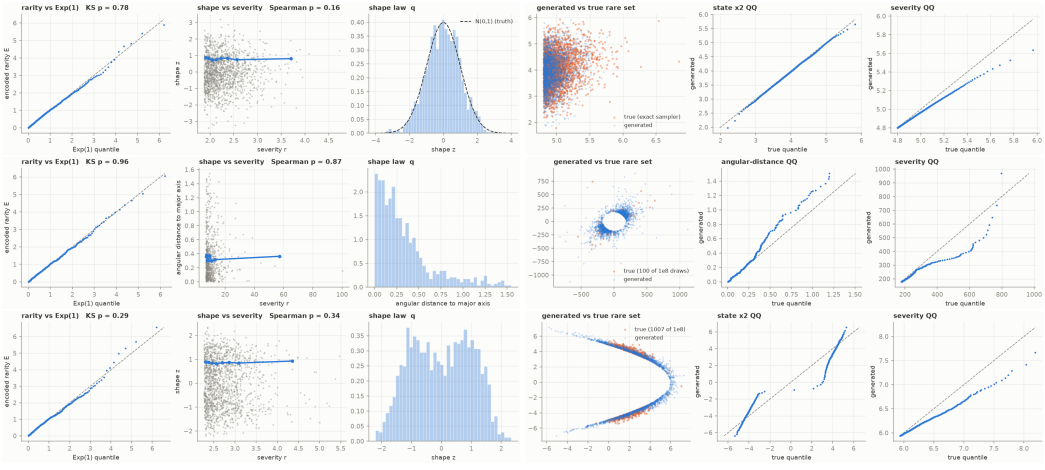

Figure 26: Three worked examples with closed-form tails (diagnostics left, validation right). Top: a linear severity. Middle: a radial severity on an elliptical t_3 law, the toy analogue of R_1. Here the transport is a radial rescaling and the invariant shape coordinate is the direction. The toy uses the raw norm, whereas the audit whitens first. Bottom: a curved severity whose transport follows a gradient flow, an example with a non-radial functional.

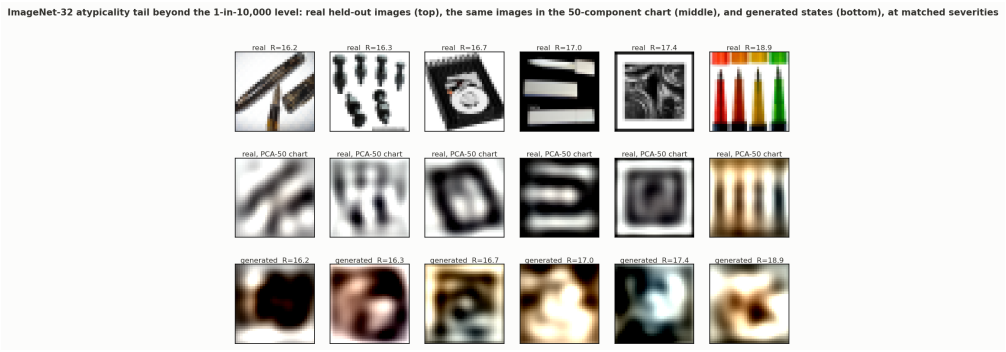

Figure 27: The radial ImageNet-32 construction from which the R_1 severity map is derived. The whitened radius is the severity coordinate. The tail region and its transport are drawn in the chart used by the wrapper.

I REPLICATION ON CELEBA-HQ-256

The ImageNet audit uses one corpus, one calibration law, and one set of thresholds. To check that the protocol does not depend on that corpus, we repeated all of it on CelebA-HQ-256 with two samplers, StyleGAN2-ADA and DDPM under a 100-step DDIM sampler. The severity construction, the adaptive threshold rule, the certificates, and the wrapper were held fixed. The corpus is much smaller than ImageNet, so this appendix also reports the depth beyond which it no longer supports a verdict. It gives the full intervals and wrapped rates that §5 summarizes by their ranges.

I.1 CORPUS, CALIBRATION LAW, AND CERTIFIABLE DEPTHS

The reference corpus is the 30,000-image CelebA-HQ-256 set (Karras et al., 2017). It passes through the same preprocessing as ImageNet (box downsampling to the 32×32 audit grid) and is divided into a calibration split $\mathcal{D}_{\mathrm{cal}}$ of 10,000 images and a held-out split $\mathcal{D}_{\mathrm{hold}}$ of 20,000 images under the same seed-8000 recipe. The calibration law $\widehat{F}$ for R_1 is refitted from scratch on $\mathcal{D}_{\mathrm{cal}}$. The adaptive rule selects the deepest threshold on its grid, $q^* = 0.995$ ($u = 11.429$, 50 calibration excesses), and returns $\hat{\xi} = -0.276$, $\hat{\sigma} = 0.764$, KS $p = 0.952$, and exceedance mass $\zeta = 0.0050$. Three audited thresholds are empirical quantiles of $\mathcal{D}_{\mathrm{hold}}$, namely 10.913 at $p = 10^{-2}$, 11.935 at $p = 3\times10^{-3}$, and 12.652 at $p = 10^{-3}$. We call these the three certified depths. A fourth threshold, 13.256 at $p = 10^{-4}$, is a generalized-Pareto extrapolation and is marked as such wherever it appears.

With $n_h = 20{,}000$ held-out images the threshold-estimation term is much larger than on ImageNet. The relative error $\sigma_\tau = \sqrt{(1-p)/(n_h p)}$ equals 0.070, 0.129 and 0.224 at the three certified depths, against 0.030 at 10^{-3} for the ImageNet corpus. The effect appears in the certificates below, where the quadrature interval is much wider than the Wilson interval. A 30k-image corpus therefore cannot certify as deeply as a 1.28M-image one.

I.2 SELF-VALIDATION ON CELEBA-HQ

For self-validation the 10,000 calibration images are audited in the model role against the thresholds estimated from the 20,000 held-out images (Table 24). At $p = 10^{-2}$, 3×10^{-3} and 10^{-3} the exceedance counts are 105, 24 and 5, and the point estimates of ρ_{self} are 1.050, 0.800 and 0.500. Both the Wilson and the quadrature interval contain $\rho_{\mathrm{self}} = 1$ at all three depths, as on the ImageNet reference split. The decline of the point estimate with depth is within counting noise ($k = 5$ at 10^{-3}). The calibration-law ratio $e(p)$ of $\widehat{F}$ at the same depths is listed in Table 26, where the wrapper is evaluated against it.

Table 24: CelebA-HQ self-validation. The calibration split $\mathcal{D}_{\mathrm{cal}}$ (10,000 real images) is audited in the model role against thresholds from the held-out split $\mathcal{D}_{\mathrm{hold}}$. Both 95% intervals contain 1 at all three depths.

p	τ_p	k	$\hat{\rho}_{\mathrm{self}}$	95% Wilson	95% quadrature
10^{-2}	10.913	105	1.050	[0.868, 1.269]	[0.818, 1.313]
3×10^{-3}	11.935	24	0.800	[0.538, 1.190]	[0.469, 1.239]
10^{-3}	12.652	5	0.500	[0.214, 1.170]	[0.139, 1.205]

At $p = 10^{-4}$ the calibration-law ratio is degenerate. Because $\hat{\xi} < 0$, the fitted severity law has a finite upper endpoint, and the empirical 10^{-4} threshold of $\mathcal{D}_{\mathrm{hold}}$, which is close to the sample maximum, lies above that endpoint. Against that threshold $e(10^{-4})$ evaluates to 0.000. Against the spliced $\tau_{10^{-4}}$ of $\widehat{F}$ itself it evaluates to 1.000 by construction. Neither value can serve as a target, and for this reason the 10^{-4} rows below are labeled extrapolation only.

I.3 AUDIT OF STYLEGAN2-ADA AND DDPM

We audit StyleGAN2-ADA (Karras et al., 2020) at $M = 100{,}000$ samples and DDPM (Ho et al., 2020) under a 100-step DDIM sampler (Song et al., 2020) at $M = 30{,}000$ (Table 25). Under R_1, the StyleGAN2-ADA certificate contains 1 at each of the three certified depths, with point estimates

1.020, 0.897 and 0.950 at $p = 10^{-2}$, 3×10^{-3} and 10^{-3}. DDPM-DDIM100 over-produces at each of them. Its point estimate grows from 3.6 to 5.6 times the reference rate as p decreases, and each of its three certificates lies entirely above 1. This is the largest over-production reported in this paper.

The certificate column of Table 25 has a mixed nominal level. At the three certified depths it is the 95% quadrature interval. The two entries marked ^{prof}, at $p = 10^{-4}$, are built differently. Each is a 95% Wilson bound on the model's own exceedance count, divided by a 90% profile-likelihood interval for the real tail probability at the extrapolated threshold (Eq. (10) with $1 - \alpha = 0.90$, App. C.3). Their overall nominal level is therefore neither 95% nor 90%, and no single coverage statement attaches to them. They bound the ratio under the fitted tail model and do not certify it.

Table 25: CelebA-HQ-256 audit under R_1. The certificate is the 95% quadrature interval, and the Wilson interval is shown for comparison. The 10^{-4} rows are beyond the corpus floor. Their thresholds come from the generalized-Pareto splice, their intervals (^{prof}) are propagated from the profile likelihood of the tail fit, and they carry no verdict. *The ^{prof} entries combine a 95% Wilson bound with a 90% profile interval and have no single nominal level (see text).

Model	p	k	$\hat{\rho}$	95% Wilson	certificate*	verdict
StyleGAN2-ADA	10^{-2}	1020	1.020	[0.960, 1.084]	[0.867, 1.175]	calibrated
($M = 100$k)	3×10^{-3}	269	0.897	[0.796, 1.010]	[0.649, 1.150]	calibrated
	10^{-3}	95	0.950	[0.777, 1.161]	[0.499, 1.417]	calibrated
	10^{-4}	38	3.800	[2.769, 5.215]	[0.924, 19.455]^{prof}	*extrapolation only*
DDPM-DDIM100	10^{-2}	1079	3.597	[3.392, 3.813]	[3.060, 4.138]	over
($M = 30$k)	3×10^{-3}	369	4.100	[3.705, 4.537]	[2.991, 5.224]	over
	10^{-3}	168	5.600	[4.817, 6.510]	[3.025, 8.216]	over
	10^{-4}	73	24.333	[19.359, 30.582]	[6.460, 114.087]^{prof}	*extrapolation only*

The 10^{-4} rows show how the audit reports a depth that the corpus cannot resolve. The point estimates there are 3.80 for StyleGAN2-ADA and 24.33 for DDPM-DDIM100, and they are not reported alone. The threshold is marked as extrapolated, and the profile-likelihood uncertainty of the tail fit is propagated into the interval. At $p = 10^{-4}$ this gives $[0.92, 19.46]$ for StyleGAN2-ADA and $[6.46, 114.09]$ for DDPM-DDIM100. The intervals are wide, and no verdict is assigned at this depth. The DDPM interval excludes 1 and the StyleGAN2-ADA interval does not.

I.4 Wrapped exceedance rates and the calibration-law ratio $e(p)$

Wrapping both samplers through the CelebA-HQ calibration law gives the same pattern as the ImageNet wrapper experiment of §5. The wrapped rate is determined by the calibration law and the sample count, and it does not depend on the model (Table 26). On this corpus the pattern tests Proposition 3(i) more strictly than on ImageNet, because the calibration-law ratio is far from 1. At $p = 10^{-2}$, 3×10^{-3} and 10^{-3} it is $e(p) = 1.050$, 0.803 and 0.604. The wrapped rates at the same three depths are 1.044 (StyleGAN2-ADA) and 1.043 (DDPM-DDIM100), then 0.800 and 0.600 for both samplers. Every cell is within 0.007 of $e(p)$, for two models whose raw ratios differ by a factor of 3.5 to 5.9.

Table 26: Wrapped exceedance rates on CelebA-HQ under R_1, compared with the calibration-law ratio $e(p)$ of the same split. "dev" $= \tilde{\rho} - e(p)$.

Model	p	raw $\hat{\rho}$	wrapped $\tilde{\rho}$	$e(p)$	dev
StyleGAN2-ADA	10^{-2}	1.020	1.044	1.050	−0.006
	3×10^{-3}	0.897	0.800	0.803	−0.003
	10^{-3}	0.950	0.600	0.604	−0.004
DDPM-DDIM100	10^{-2}	3.597	1.043	1.050	−0.007
	3×10^{-3}	4.100	0.800	0.803	−0.003
	10^{-3}	5.600	0.600	0.604	−0.004

A criterion that asks for the wrapped rate to lie in $[0.9, 1.1]$ is met at $p = 10^{-2}$ and missed at the other two depths. Proposition 3(i) makes a different statement. It says that the wrapped severity law equals $\widehat{F}$, so the wrapped exceedance rate equals $e(p)$, and on a 10,000-image calibration split

$e(10^{-3})$ is 0.604. The wrapper cannot supply calibration accuracy that the calibration law does not have. The ImageNet audit reports the same quantity as $e(10^{-3}) = 0.99$ and $e(10^{-4}) = 0.69$ (App. D). Since $e(10^{-3})$ is close to 1 there, the ImageNet experiment separates the target $e(p)$ from the target 1 less clearly than the CelebA-HQ replication does. The shortfall reflects the size of the reference corpus. On this corpus the wrapped rate is close to the reference rate at $p = 10^{-2}$, where $e = 1.05$, and falls increasingly below it at the two smaller probabilities. Calibration at smaller tail probabilities requires a larger reference corpus to estimate the target law.

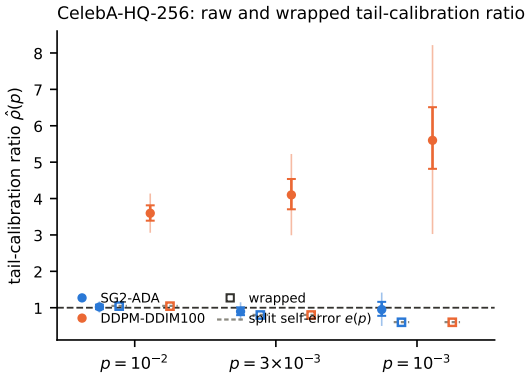

Figure 28: Replication on CelebA-HQ-256 under R_1, with $\widehat{F}$ refitted on the new corpus ($M = 100{,}000$ for StyleGAN2-ADA, $M = 30{,}000$ for DDPM-DDIM100). Circles: raw $\hat{\rho}$ with Wilson caps and the wider quadrature whisker. Squares: wrapped rate. Grey dashes: the calibration-law ratio $e(p)$ of the same split, which is the value the wrapped rate is compared with.

J Reproducibility Details

This appendix lists the hardware and software, the corpora, the checkpoints and sampler settings, the seeds, the sample counts and sampling cost, and the constants of the calibration law $\widehat{F}$.

J.1 Hardware and software

Sampling and embedding ran on single NVIDIA GH200 GPUs with 16 CPU cores and 64 GB of host memory; all later analysis is CPU-only. The environment uses Python 3.12.9, PyTorch 2.10.0 (CUDA 12.9) and SciPy 1.17.1, with `diffusers` 0.39.0, `timm` 0.6.13 and `numpy` 2.5.2 for the model code. The DiT-XL/2 code requires `timm` 0.6.13. The host is `aarch64`, where the custom CUDA extensions of the StyleGAN code bases do not compile; StyleGAN-XL and StyleGAN2-ADA then use the reference operators that the upstream repositories provide. A reproduction with compiled extensions should recover the same sample distribution, but not bit-identical samples.

J.2 Corpora

ImageNet-32. The reference corpus is the ImageNet-32 training set of 1,281,167 images (Chrabaszcz et al., 2017), stored as `uint8` arrays of size $32 \times 32 \times 3$. One permutation under split seed 8000 gives the calibration split $\mathcal{D}_{\text{cal}}$ with $n_f = 200{,}000$ images and the held-out split $\mathcal{D}_{\text{hold}}$ with $n_h = 1{,}081{,}167$ images. The PCA chart, the calibration law and the shape basis are fitted on $\mathcal{D}_{\text{cal}}$, and every audited threshold is read from $\mathcal{D}_{\text{hold}}$. All headline numbers use this split.

CelebA-HQ-256. The replication corpus is the 30,000 images of CelebA-HQ-256 (Karras et al., 2017), read in a fixed order, box-downsampled by a factor of 8 to 32^2 and quantized to `uint8`. The same seed-8000 permutation splits them into 10,000 calibration and 20,000 held-out images.

Reduction to the corpus grid. Every sampler output is quantized to `uint8` at the model's native resolution, box-averaged (by 2 from 64^2, by 8 from 256^2) and rounded before the severity map is

applied. The corpus itself is produced by the same path, so the calibration law involves no model-specific preprocessing.

J.3 Checkpoints and sampler settings

All checkpoints are public releases, used unchanged with the sampler settings published in their repositories. No model is fine-tuned, distilled or retrained. Two pairs share a network and differ only in the sampler (EDM and EDM-churn, DiT-cfg1.5 and DiT-cfg1.0).

- **iDDPM**: `iddpm-imagenet64_cond_270M_250K.pt`; DDIM-50 with `clip_denoised`, cosine schedule, 4000 diffusion steps, class-conditional, learned sigma.
- **EDM**: `edm-imagenet-64x64-cond-adm.pkl`; deterministic Heun, 18 steps (NFE 35).
- **EDM-churn**: the same checkpoint under the repository's ImageNet-64 stochastic setting, with 256 steps, $S_{\text{churn}} = 40$, $S_{\min} = 0.05$, $S_{\max} = 50$ and $S_{\text{noise}} = 1.003$ (NFE 511).
- **EDM2-S**: `edm2-img64-s-1073741-0.075.pkl`; deterministic Heun, 32 steps (NFE 63). **EDM2-M**: `edm2-img64-m-2147483-0.060.pkl`, same sampler.
- **StyleGAN-XL**: `sgxl-imagenet32.pkl` (G_{ema}); one forward pass at $\psi_{\text{trunc}} = 1$. No truncation is applied, because truncation changes the tail that is audited.
- **DiT-cfg**1.5: `DiT-XL-2-256x256.pt` with the `sd-vae-ft-ema` decoder; DDIM-50, classifier-free guidance 1.5. **DiT-cfg**1.0: same weights, guidance off.
- **StyleGAN2-ADA** (CelebA-HQ): the `celebahq-256` pickle (G_{ema}); one forward pass. **DDPM-DDIM**100: `ddpm-celebahq-256`, DDIM respaced to 100 steps.
- R_2 **backbone**: a frozen MAE-pretrained ViT-B/16 (He et al., 2022); the mean of its 196 final-layer patch tokens.

J.4 Seeds and replicates

Split seed 8000 fixes $\mathcal{D}_{\text{cal}}$, $\mathcal{D}_{\text{hold}}$, the PCA chart, the calibration law and the shape basis for both corpora. Seeds 8001–8004 are used in the self-validation study and seeds 8001 and 8002 also in the sensitivity study (App. G); no headline verdict is read from them. Each sampler has a base seed (5555 for iDDPM, 1234 for the EDM and EDM2 configurations, 4321 for StyleGAN-XL, 7777 for DiT, 71000 and 72000 for StyleGAN2-ADA and DDPM on CelebA-HQ), and sampling that resumes from a partial cache adds the number of samples already drawn, so the random stream continues and is not repeated. The planted-truth study uses $B = 500$ replicates per cell, the fixed-functional re-split null $B = 400$, the downstream-inheritance study $B = 500$ per cell, and the scale curves $B = 200$ subsamples per budget. The secondary permutation test of the shape audit uses 10^4 draws per cell and was run at two seeds with identical verdicts.

J.5 Sample counts and sampling cost

Sampling ran under a wall-time limit per job. EDM-churn, at 1.3 images per second, was stopped at $M = 15{,}104$ of a planned 25,000; both DiT configurations reached $M = 50{,}000$ over several jobs. Over the ten samplers the study generated 745,104 images in about 27 GPU-hours (Table 27). Computing the R_2 embeddings of all real images and model samples took a further 7 GPU-minutes. Every later number is computed from the cached severities of these draws.

J.6 Calibration-law constants

Every ImageNet-32 number in this paper is computed with one calibration law $\widehat{F}$, fitted once on $\mathcal{D}_{\text{cal}}$ and never refitted for a model. Table 28 lists its constants for R_1 next to those of the CelebA-HQ replication. The threshold quantile q^* is selected by the KS score of App. C.2 over the grid $\{0.98, 0.985, 0.99, 0.9925, 0.995\}$. Both ImageNet-32 thresholds are empirical quantiles of $\mathcal{D}_{\text{hold}}$, so the generalized-Pareto tail enters no audited threshold and is used only beyond the corpus floor. The shape basis is fitted on the $n = 4000$ images in the top 2% of $\mathcal{D}_{\text{cal}}$ by severity. The other two functionals are fitted in the same way: R_2 uses $K_2 = 50$ components (explained variance ratio 0.991, KS score 0.891) and R_3 uses $K_3 = 32$ components per patch position (KS score 0.784). The

Table 27: Sample sizes and sampling time per configuration.

Configuration	Corpus	M	sampling time (s)
iDDPM	ImageNet-32	100,000	7,275
EDM	ImageNet-32	100,000	5,497
EDM-churn	ImageNet-32	15,104	11,679
EDM2-S	ImageNet-32	100,000	5,438
EDM2-M	ImageNet-32	100,000	7,593
StyleGAN-XL	ImageNet-32	100,000	1,553
DiT-cfg1.5	ImageNet-32	50,000	31,711
DiT-cfg1.0	ImageNet-32	50,000	15,324
StyleGAN2-ADA	CelebA-HQ	100,000	113
DDPM-DDIM100	CelebA-HQ	30,000	11,062

CelebA-HQ law is built with the same procedure without retuning, and its $\tau_{10^{-4}}$ is an extrapolation that is never read as a verdict.

Table 28: Constants of the calibration law $\widehat{F}$ for R_1 on the two corpora (split seed 8000).

Quantity	ImageNet-32	CelebA-HQ-256
Threshold quantile	$q^* = 0.9925$	$q = 0.995$
GPD threshold u	12.06	11.429
Shape $\hat{\xi}$	-0.050	-0.276
Scale $\hat{\sigma}$	0.974	0.764
Exceedance mass ζ	0.0075	0.0050 (50 calibration excesses)
KS score of the GPD fit	0.970	0.952
Audited thresholds τ_p	$\tau_{10^{-3}} = 13.9350$, $\tau_{10^{-4}} = 16.1291$	$\tau_{10^{-2}} = 10.913$, $\tau_{3\times10^{-3}} = 11.935$, $\tau_{10^{-3}} = 12.652$, $\tau_{10^{-4}} = 13.256$ (extrapolation)