TAIL-INDEX ANNEALING FOR LONG-TAILED DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical denoising diffusion models rely on Gaussian noise throughout the forward process. Recent stable-noise diffusion models, such as the Denoising Lévy Probabilistic Model (DLPM), replace it with stable noise that has a single fixed tail index, forcing a fixed compromise between terminal initialization and reverse estimation. Direct substitution of a time-varying tail index into DLPM fails since increments with different tail indices do not form a single stable law. We introduce *Annealed DLPM* (A-DLPM), a consistent framework that increases the tail index along a half-cosine forward schedule and traverses the reverse schedule during generation. We construct the corresponding conditional Gaussian reverse kernel, which reduces to the original DLPM when the tail index is fixed. The resulting annealed residual admits an exact, interpretable spectral-curvature characterization. On two-dimensional data, A-DLPM attains the lowest mean tail-sensitive error among evaluated methods; on long-tailed image benchmarks, it attains the highest mean Recall among all evaluated methods and improves both FID and Recall over the evaluated DLPM, DLIM, and Gaussian DDPM controls.

1 INTRODUCTION

Diffusion models adopt one noise family for the entire noising path. Classical denoising diffusion uses Gaussian noise as a perturbation (Sohl-Dickstein et al., 2015; Ho et al., 2020), and its score-based and implicit variants retain the same light-tailed endpoint (Song et al., 2021b;a). Recent Lévy models replace Gaussian noise by an α-stable law with a single constant tail index (Yoon et al., 2023; Shariatian et al., 2025). As shown by LIM, heavier-tailed perturbations can improve coverage of underrepresented modes on imbalanced data (Yoon et al., 2023). LIM attributes this effect to large Lévy jumps, which allow reverse trajectories to traverse separated regions of the sample space and reach multiple modes more readily. The choice of noise is therefore consequential: the generator must preserve common structure without abandoning weakly represented regions (Zhang et al., 2024), while heavier-tailed perturbations make reverse estimation harder. Recent error bounds for fixed tail indices formalize this tension as a trade-off between terminal initialization and accumulated training error (Cherkaoui et al., 2026). A single stable tail index imposes one global compromise throughout the path.

Varying the tail index with time is not a straightforward modification of a model designed for a constant tail index. Stable laws with a common tail index are closed under addition; changing that tail index across nondegenerate steps instead produces a mixed-power characteristic exponent, not a simply rescaled stable law. The marginal and posterior formulas for a fixed tail index α used by Denoising Lévy Probabilistic Models (DLPMs) therefore no longer apply (Shariatian et al., 2025). Ideally, a valid construction must define the joint law of increments with different tail indices, ensure that training and sampling operate on the same law, provide a tractable conditional distribution for reverse updates, and reduce exactly to DLPM when the path of tail indices is constant.

We introduce *Annealed DLPM* (A-DLPM), an extension of DLPM with a time-varying tail index. The stable tail index α rises smoothly from 1.5 to 2.0 along the forward noising path, and decreases from 2.0 back to 1.5 during generation; we call this *standard half-cosine reverse annealing*. Building on the scale-mixture representation of DLPM, we draw independent positive stable variables A_t with tail index $\alpha_t/2$ and update the scalar conditional variance Σ_t according to the forward recursion. We define the residual as $Y_t := X_t - \bar{\gamma}_t X_0$, the accumulated noise remaining after subtracting the

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.

prescribed signal component $\bar{\gamma}_t X_0$ from the noisy state X_t. Conditioning on $A_{0:t}$ makes the residual path Gaussian, so reverse sampling is analytically tractable and the original epsilon-prediction objective is preserved. The forward process is generally non-Markovian, because its transition depends on X_0; this is analogous to how implicit diffusion paths can share marginals without sharing a Markov forward chain (Song et al., 2021a). When sampling with a coarser time grid than training, we derive a corrected backward update that propagates the residual using the product of the original one-step forward coefficients.

Beyond the reverse update, the construction reveals the distribution family generated by annealing. For a characteristic-function argument u, set $r := \|u\| > 0$. The residual Y_t has characteristic exponent $\Psi_t(r) := \sum_{j=1}^{t} c_{j,t} r^{\alpha_j}$ with $c_{j,t} \geq 0$. When nonzero coefficients occur at distinct tail indices, this sum cannot be written as a single power Cr^{α} with $C > 0$. When at least one coefficient is nonzero, we show in Theorem 2 that $\mathrm{d}\log\Psi_t(r)/\mathrm{d}\log r$ is the average of the scheduled tail indices under the weights $c_{j,t}r^{\alpha_j}/\Psi_t(r)$, while $\mathrm{d}^2\log\Psi_t(r)/\mathrm{d}(\log r)^2$ is their variance under the same weights. Thus annealing can retain the heavy tails associated with lower values of the tail index, while incorporating components with higher values, providing a concrete way to redistribute the initialization–training trade-off identified by Cherkaoui et al. (2026) for heavy-tailed diffusion with a fixed tail index.

Beyond this spectral characterization, the discrete construction also admits a continuous-time limit. On every compact subinterval before the terminal time, the discrete residual also converges in the J_1 Skorokhod topology to an Ornstein–Uhlenbeck integral driven by an additive process. In the continuous limit, a time-varying tail index introduces a signal-correction drift that disappears for every path with a fixed tail index α.

Under matched configurations and budgets, annealing obtains the lowest mean upper-tail MSLE among all evaluated methods on the two-dimensional task (Shariatian et al., 2025). On CIFAR-10-LT and CIFAR-100-LT, averaging five generation seeds within each of five training runs, it attains the highest mean PRDC Recall among all evaluated methods and lower mean FID than DLPM and DLIM at every evaluated tail index and than Gaussian DDPM (Heusel et al., 2017; Naeem et al., 2020). On both benchmarks, its target-normalized prediction errors resemble the DLPM errors at $\alpha = 2.0$ early and $\alpha = 1.5$ near the data, consistent with redistributing the initialization–training compromise across reverse time.

Our contributions are as follows.

- **Joint construction with a variable tail index.** A time-varying path of stable tail indices whose accumulated positive stable scales keep the forward marginal, epsilon target, and reverse update mutually consistent.
- **Consistent reverse inference.** Conditional Gaussian reverse updates between arbitrary retained diffusion times, with exact reduction to DLPM whenever the tail index is constant.
- **Distributional characterization.** A mixed-power residual with an exact spectral-curvature identity, together with the corresponding continuous-time limit driven by an additive process and the signal correction induced by a changing tail index.
- **Improved fidelity–diversity trade-off.** Improved long-tailed generation with heavier-tailed increments near the data and near-Gaussian increments near the terminal noise. This temporal allocation corresponds to a redistribution of the compromise between terminal initialization and reverse-time training across time, instead of imposing it at every step.

2 Related Work

Non-Gaussian and heavy-tailed diffusion. Classical diffusion models use Gaussian noise throughout the forward path (Ho et al., 2020; Song et al., 2021b). Non-Gaussian diffusion has also been explored through alternative noise families, including gamma (Nachmani et al., 2021), generalized normal (Deasy et al., 2022), Laplace (Jolicoeur-Martineau et al., 2023), and Cauchy (Lian et al., 2025) perturbations, as well as finite-activity jump processes (Baule, 2025). Stable laws are another natural choice, as they directly model heavy tails and have a long history in finance and insurance (Embrechts et al., 1997).

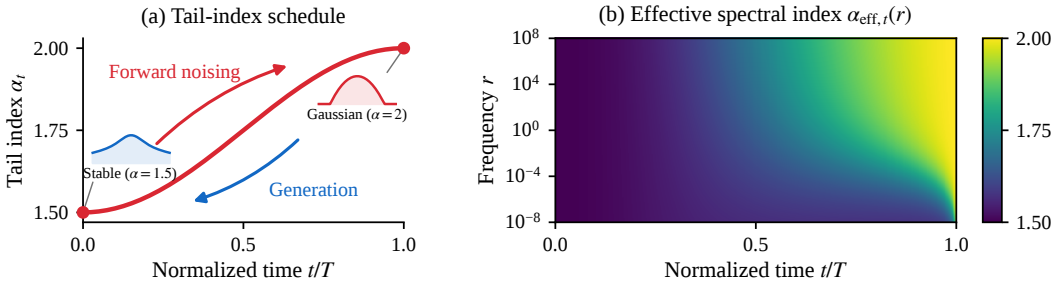


Figure 1: **Annealed schedule and accumulated spectrum.** (a) Tail-index schedule and increment densities (common logarithmic vertical scale); generation reverses the path. (b) Effective spectral index of the untruncated accumulated noise, computed from Theorem 2 on the standard 4000-step chain for $t > 0$.

The Lévy–Itô model (LIM) (Yoon et al., 2023) replaces Brownian motion in a continuous-time score model with an isotropic stable Lévy process of fixed tail index. Its exact reverse-time SDE is written in terms of a fractional score, which is learned by fractional denoising score matching. Subsequent work develops improved stochastic and deterministic samplers for LIM (Popov et al., 2025). DLPM (Shariatian et al., 2025) takes a different route: it represents each stable perturbation with a fixed tail index as $\sqrt{A}Z$, where A is a positive stable scale and Z is Gaussian. Conditional on the scales, its discrete forward transitions and reverse distributions (given X_0) are Gaussian, so training retains an epsilon-style objective and the DDPM network parameterization.

Student-t diffusion and flow models use fixed degrees of freedom (Pandey et al., 2025), while normalizing flows learn the source tail parameter or apply a terminal tail transformation (Jaini et al., 2020; Hickling & Prangle, 2025). Random-clock flow matching recovers stable or Student-t marginals by mixing clock-conditioned Gaussian sources, with a fixed tail index for each model (Yang et al., 2026). Pachebat (2026) applies coordinatewise soft-log transformations to the data to induce tail annealing; this annealing operates on the data distribution rather than on the noise tail index. A related line of work keeps a fixed tail index for heavy-tailed noise but adapts the effective noise scale during generation according to the denoising residual (Wakatsuki & Shimazaki, 2026).

Cherkaoui et al. (2026) analyze the trade-off inherent in models with a fixed tail index. They decompose the error in the final distribution into terminal initialization, approximation, and finite-sample estimation terms, with the latter two accumulating along generation. A heavier-tailed prior can reduce terminal mismatch while making the learned reverse problem harder, which explains why quality need not improve monotonically as the tail index decreases. Their analysis, however, applies to a Markov model with a fixed tail index and does not define a joint law with a variable tail index, nor does it cover a mixed-power residual or a scale-augmented epsilon target. A-DLPM fills this gap: it allocates the tail index across diffusion steps, while preserving the conditional Gaussian training formulation of DLPM.

Learning from long-tailed data. For long-tailed data, supervised methods use class-balanced weighting, tail-aware margins, decoupled classifiers, or logit adjustment (Cui et al., 2019; Cao et al., 2019; Kang et al., 2020; Menon et al., 2021). Oriented calibration brings frequency-aware correction to diffusion generation (Zhang et al., 2024). Our image experiments use the same benchmarks but deliberately omit labels, reweighting, resampling, and auxiliary experts, isolating the effect of the time-varying noise law.

3 Annealing with a Variable Tail Index

A-DLPM coordinates a forward process with a time-varying tail index, the accumulated conditional variance that defines the epsilon target, and a conditional Gaussian reverse distribution. As shown in Figure 1(a), forward noising follows the red path toward thinner-tailed increments, while generation traverses the same path in reverse. Note that a time-varying tail index cannot be inserted directly

into formulas for a fixed tail index: the sum of two independent stable increments with distinct tail indices has characteristic function $\exp\{-c_1\|u\|^{\alpha_1} - c_2\|u\|^{\alpha_2}\}$, which no rescaling turns into the single-power form $\exp\{-C\|u\|^{\alpha}\}$ unless $\alpha_1 = \alpha_2$.

3.1 Stable scale augmentation

A law is stable when sums of independent copies remain in the same location–scale family. The parameter $\alpha \in (0, 2]$ is the tail index: smaller α yields heavier polynomial tails, while $\alpha = 2$ gives the Gaussian distribution. Let $\mathcal{S}_\alpha$ denote the centered, symmetric, and isotropic stable law on $\mathbb{R}^d$ defined by

$$S \sim \mathcal{S}_\alpha \quad \Longleftrightarrow \quad \mathbb{E}e^{i\langle u, S\rangle} = e^{-\|u\|^\alpha}, \qquad u \in \mathbb{R}^d, \quad 0 < \alpha \le 2. \tag{1}$$

For $\alpha < 2$, this characteristic function equivalently defines a symmetric stable law with a uniform spectral measure (Samorodnitsky & Taqqu, 1994, Proposition 2.5.5). Note that under this normalization, $\mathcal{S}_2 = \mathcal{N}(0, 2I_d)$ rather than $\mathcal{N}(0, I_d)$. For $0 < \beta \le 1$, let $\mathcal{A}_\beta$ denote the positive stable law under the scale convention

$$A \sim \mathcal{A}_\beta \quad \Longleftrightarrow \quad \mathbb{E}e^{-sA} = e^{-(2s)^\beta}, \qquad s \ge 0. \tag{2}$$

This is a rescaled one-sided stable law, and Eq. (2) follows from Nolan (2020, Proposition 3.2). If $A \sim \mathcal{A}_{\alpha/2}$ and $Z \sim \mathcal{N}(0, I_d)$ are independent, then $\sqrt{A}Z \sim \mathcal{S}_\alpha$. Lemma 3 in Appendix A provides the proof.

3.2 Standard half-cosine reverse annealing

DLPM uses the same stable law at every noise level. In contrast, we let the stable tail index change smoothly along the diffusion horizon. For a chain indexed by $t = 0, 1, \ldots, N-1$, we choose the half-cosine path

$$\alpha_t := 1.5 + \frac{1 - \cos(\pi t/(N-1))}{4}, \quad t = 0, 1, \ldots, N-1 \tag{3}$$

Thus forward noising moves from the heavy-tailed endpoint $\alpha_0 = 1.5$ near the data to the Gaussian endpoint $\alpha_{N-1} = 2.0$. The half-cosine changes most rapidly near the midpoint and has zero slope at both ends. Generation traverses the same path backward, from 2.0 to 1.5; we call this *standard half-cosine reverse annealing*.

The forward schedule places heavier-tailed increments near the data and lighter-tailed ones toward the high-noise end. Note that $\alpha_{N-1} = 2.0$ makes only the final increment Gaussian; the accumulated noise is generally not a single stable law, as characterized in Section 4. The exact form of Eq. (3) is not essential: Appendix H shows that nearby monotone schedules yield comparable results.

3.3 Forward process and signal correction

We now construct the forward process associated with this tail-index path. We retain the variance-preserving cosine noise schedule $\{\beta_t\}$ used by DLPM (Nichol & Dhariwal, 2021; Shariatian et al., 2025). Let q_t denote its cumulative signal power:[1] $q_0 := 1$, $q_t := \prod_{j=1}^{t}(1 - \beta_j)$. We use q_t and the current tail index to define two cumulative coefficients:

$$\bar{\gamma}_t := q_t^{1/\alpha_t}, \qquad \bar{\sigma}_t := (1 - q_t)^{1/\alpha_t}. \tag{4}$$

When $\alpha_t \equiv \alpha$ is constant, these reduce to the usual DLPM marginal coefficients. For each step $t \ge 1$, define

$$\gamma_t := (1 - \beta_t)^{1/\alpha_t}, \qquad \sigma_t := \beta_t^{1/\alpha_t}, \tag{5}$$

with $\gamma_0 := 1$ and $\sigma_0 := 0$.

Draw positive stable scales $A_t \sim \mathcal{A}_{\alpha_t/2}$ and standard Gaussian innovations $Z_t \sim \mathcal{N}(0, I_d)$, with the collections $\{A_t\}$ and $\{Z_t\}$ mutually independent and independent of X_0. Write $A_{0:t} := (A_0, \ldots, A_t)$ and accumulate their conditional variance through

$$\Sigma_0 := 0, \qquad \Sigma_t := \gamma_t^2 \Sigma_{t-1} + \sigma_t^2 A_t, \quad t \ge 1. \tag{6}$$

[1] This quantity is commonly denoted $\bar{\alpha}_t$ in Gaussian diffusion models (Ho et al., 2020). We use q_t because α denotes the stable tail index throughout this paper.

The resulting forward transition is

$$X_t := \gamma_t X_{t-1} + (\bar{\gamma}_t - \gamma_t \bar{\gamma}_{t-1}) X_0 + \sigma_t \sqrt{A_t} Z_t. \tag{7}$$

The middle term is a signal correction that forces the coefficient of X_0 to remain $\bar{\gamma}_t$ in the conditional mean. Conditional on X_0 and $A_{0:t}$, Proposition 5 in Appendix B shows that the resulting marginal is

$$X_t = \bar{\gamma}_t X_0 + \sqrt{\Sigma_t} \widetilde{Z}_t = \bar{\gamma}_t X_0 + Y_t, \tag{8}$$

where $\widetilde{Z}_t \mid (X_0, A_{0:t}) \sim \mathcal{N}(0, I_d)$ and $Y_t := X_t - \bar{\gamma}_t X_0$ is the residual.

3.4 Epsilon target and conditional reverse distribution

To retain the epsilon-prediction parameterization of DLPM, we normalize the residual by $\bar{\sigma}_t$ for $t \geq 1$ and define

$$\epsilon_t := \frac{Y_t}{\bar{\sigma}_t} = \frac{\sqrt{\Sigma_t}}{\bar{\sigma}_t} \widetilde{Z}_t, \qquad X_t = \bar{\gamma}_t X_0 + \bar{\sigma}_t \epsilon_t. \tag{9}$$

Conditional on $A_{0:t}$, the target ϵ_t is Gaussian with covariance $(\Sigma_t / \bar{\sigma}_t^2) I_d$, and is generally not standard Gaussian. The transition in Eq. (7) depends explicitly on X_0, so the forward process is generally non-Markovian after marginalizing X_0; conditional on X_0 and the scale path, the residual process is Gaussian.

The network receives (X_t, t) and predicts ϵ_t. We use the same per-sample epsilon root-mean-square-error (epsilon-RMSE) objective as in DLPM,

$$\mathcal{L}_\epsilon(\theta) := \mathbb{E}\left[\frac{1}{\sqrt{d}} \|\epsilon_\theta(X_t, t) - \epsilon_t\|_2 \right], \tag{10}$$

where d is the data dimension and t is sampled uniformly from $\{1, \ldots, N-1\}$. The corresponding estimate for X_0 is

$$\widehat{X}_{0,\theta} := \frac{X_t - \bar{\sigma}_t \epsilon_\theta(X_t, t)}{\bar{\gamma}_t}. \tag{11}$$

When $\Sigma_t > 0$, the conditional Gaussian structure of the forward chain makes the joint distribution of (X_s, X_t) given X_0 and the scale path $A_{0:t}$ exactly Gaussian for any retained indices $s < t$. By Proposition 7 in Appendix B, standard Gaussian conditioning then gives the reverse conditional distribution

$$X_s \mid (X_t, X_0, A_{0:t}) \sim \mathcal{N}\big(a_{s,t} X_t + b_{s,t} X_0,\ \nu_{s,t} I_d\big), \tag{12}$$

with the following coefficients:

$$G_{s,t} := \prod_{k=s+1}^{t} \gamma_k, \tag{13}$$

$$a_{s,t} := G_{s,t} \frac{\Sigma_s}{\Sigma_t}, \qquad b_{s,t} := \bar{\gamma}_s - a_{s,t} \bar{\gamma}_t, \tag{14}$$

$$\nu_{s,t} := \Sigma_s - G_{s,t}^2 \frac{\Sigma_s^2}{\Sigma_t}. \tag{15}$$

For generation, we substitute the endpoint estimate $\widehat{X}_{0,\theta}$ from Eq. (11) for the unknown X_0.

When the schedule of tail indices is constant ($\alpha_t \equiv \alpha$), positive stable closure reduces Σ_t to the marginal scale used by DLPM; Proposition 8 in Appendix B proves the exact reduction. For the full training and sampling algorithms, see Appendix E.

4 What Distribution Does Annealing Create?

Recall that $Y_t = X_t - \bar{\gamma}_t X_0$ is the residual, namely the accumulated noise before normalization by $\bar{\sigma}_t$. Subtracting $\bar{\gamma}_t X_0$ from Eq. (7) gives

$$Y_t = \gamma_t Y_{t-1} + \sigma_t \sqrt{A_t} Z_t \tag{16}$$

for $t \geq 1$, with $Y_0 = 0$. Iterating this recursion gives

$$Y_t = \sum_{j=1}^{t} w_{j,t} \sqrt{A_j} Z_j, \qquad \text{where } w_{j,t} := \sigma_j \prod_{k=j+1}^{t} \gamma_k = \sigma_j G_{j,t}. \tag{17}$$

The empty product for $j = t$ in $w_{j,t}$ is defined by 1. The following result characterizes the marginal distribution of this residual Y_t, and shows that it generally falls outside the family of stable laws with a fixed tail index.

Theorem 1 (Mixed-power spectrum and non-representability as a single stable law)**.** *For the mutually independent positive-stable scales A_j and standard Gaussian innovations Z_j appearing in Eq. (17), the characteristic function of Y_t is*

$$\mathbb{E}e^{\mathrm{i}\langle u, Y_t\rangle} = \exp\{-\Psi_t(\|u\|)\}, \qquad \textit{where } \Psi_t(r) := \sum_{j=1}^{t} w_{j,t}^{\alpha_j} r^{\alpha_j}, \qquad r \geq 0. \tag{18}$$

For a fixed t, call the jth term active when $w_{j,t} > 0$. If at least two active terms have distinct tail indices, then there do not exist $C > 0$ and $\alpha \in (0, 2]$ such that $\Psi_t(r) = Cr^\alpha$ for all $r > 0$. Hence the residual is not a rescaled member of any single stable law.

For any t with at least one active term and any $r > 0$, define the effective spectral tail index and the normalized weights by

$$\alpha_{\mathrm{eff},t}(r) := \frac{\mathrm{d}\log \Psi_t(r)}{\mathrm{d}\log r}, \qquad \pi_{j,t}(r) := \frac{w_{j,t}^{\alpha_j} r^{\alpha_j}}{\Psi_t(r)}.$$

Theorem 2 (Exact spectral-curvature identity)**.** *For every $r > 0$,*

$$\alpha_{\mathrm{eff},t}(r) = \sum_{j=1}^{t} \pi_{j,t}(r)\alpha_j, \qquad \frac{\mathrm{d}\alpha_{\mathrm{eff},t}(r)}{\mathrm{d}\log r} = \frac{\mathrm{d}^2 \log \Psi_t(r)}{\mathrm{d}(\log r)^2} = \sum_{j=1}^{t} \pi_{j,t}(r)\big(\alpha_j - \alpha_{\mathrm{eff},t}(r)\big)^2 \geq 0.$$

For every $r > 0$, the curvature is strictly positive whenever at least two active tail indices differ. It may tend to zero as $r \downarrow 0$ or $r \uparrow \infty$ when the normalized weights concentrate on one extreme tail index.

This identity makes the frequency dependence precise: as $r \downarrow 0$, $\alpha_{\mathrm{eff},t}(r)$ tends to $\alpha_{\min} := \min_{j:w_{j,t}>0} \alpha_j$; as $r \uparrow \infty$, it tends to $\alpha_{\max} := \max_{j:w_{j,t}>0} \alpha_j$. When the smallest active tail index is below 2, the heavy-tailed behavior of the residual Y_t is governed by this smallest tail index: for any nonzero vector v, the one-dimensional projection $\langle v, Y_t\rangle$ satisfies $\Pr(|\langle v, Y_t\rangle| > s) \sim C_v s^{-\alpha_{\min}}$ as $s \to \infty$ for some $C_v > 0$; see Proposition 9 in Appendix C. Consequently, A-DLPM retains the heavy-tailed behavior associated with low tail indices, whereas the high-frequency part of the characteristic exponent is controlled by the larger active tail indices. Figure 1(b) illustrates this separation: higher frequencies have larger effective indices at intermediate times. The nearly uniform early region reflects the small weighted variance of the early active indices, while the later frequency contrast reveals the coexistence of distinct contributions in the accumulated noise.

For DLPM, Cherkaoui et al. (2026) decompose the sampling error into two components: terminal mismatch and the errors accumulated along the reverse process. Although that decomposition does not directly apply to A-DLPM, it suggests a narrower empirical prediction: early in generation, where the forward increments have tail indices near 2.0, the prediction errors of A-DLPM should resemble those of DLPM with $\alpha = 2.0$; near the data, where the forward increments have heavier tails, they should resemble those of DLPM with $\alpha = 1.5$. Experiments in Appendix F confirm this predicted shift.

A continuous-time limit gives a complementary process-level description. Let $T > 0$ denote the limiting horizon, and let $q(t)$ and $\alpha(t)$ be smooth limits of the discrete schedules on $[0, T]$, with $q(t) > 0$ for $t < T$. Define $\beta(t) := -\partial_t \log q(t)$ and $\kappa(t) := \beta(t)/\alpha(t)$. When the tail index varies, the limit is driven by an additive process L, and an additional drift makes the coefficient of X_0 equal to $q(t)^{1/\alpha(t)}$, the continuous-time counterpart of $\bar{\gamma}_t$ in Eq. (8). For $0 \leq s \leq t < T$, set $K(s, t) := \exp\{-\int_s^t \kappa(r)\mathrm{d}r\}$, the continuous-time limit of $G_{j,t}$ in Eq. (13). Then for every $\tau < T$, the piecewise-constant interpolation of the discrete residual converges in the J_1 Skorokhod topology

to

$$Y_t = \int_0^t K(s,t)\mathrm{d}L_s, \tag{19}$$

where L is an additive process whose increments satisfy

$$\mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, L_t - L_s\rangle} = \exp\left\{-\int_s^t \beta(r)\|u\|^{\alpha(r)}\mathrm{d}r\right\}, \qquad 0 \le s \le t < T.$$

Define the continuous state by $X_t := q(t)^{1/\alpha(t)}X_0 + Y_t$; it obeys

$$\mathrm{d}X_t = -\kappa(t)X_{t-}\mathrm{d}t - q(t)^{1/\alpha(t)}\frac{\alpha'(t)}{\alpha(t)^2}\log q(t)X_0\mathrm{d}t + \mathrm{d}L_t. \tag{20}$$

If the tail index is constant and κ is also constant, then $K(s,t)$ reduces to $\mathrm{e}^{-\kappa(t-s)}$, the finite-horizon counterpart of the classical symmetric α-stable Ornstein–Uhlenbeck kernel (Samorodnitsky & Taqqu, 1994, Example 3.6.3). For a varying tail index, L generally has nonstationary increments, so the process Y_t defined by Eq. (19) is generally neither stationary nor a single stable process with one fixed tail index.

The notation X_{t-} in Eq. (20) denotes the left limit of the càdlàg state, because the driving additive process L may have jumps when $\alpha(t) < 2$ and $\beta(t) > 0$. Since $X_{t-} = X_t$ for almost every t, this distinction does not affect the $\mathrm{d}t$ drift integral. The second drift in Eq. (20) is the continuous-time counterpart of the discrete signal correction in Eq. (7). It follows from

$$\frac{\mathrm{d}}{\mathrm{d}t}q(t)^{1/\alpha(t)} = -\kappa(t)q(t)^{1/\alpha(t)} - q(t)^{1/\alpha(t)}\frac{\alpha'(t)}{\alpha(t)^2}\log q(t).$$

The term involving $\alpha'(t)$ produces the second drift. It vanishes for a constant tail index and, more generally, vanishes pointwise whenever $\alpha'(t)\log q(t) = 0$.

Appendix B derives the discrete forward marginal (Proposition 5), the signal-correction identity (Lemma 6), the reverse conditional (Proposition 7), and exact recovery at a fixed tail index (Proposition 8). Appendix C proves the mixed-power spectrum and spectral-curvature identity (Theorems 1–2), together with the asymptotic result for the spatial tail index (Proposition 9). The continuous-time limit is established in Appendix D (Theorem 13).

5 Experiments

We evaluate the annealed path by its upper-tail MSLE on the two-dimensional stable target, and by FID and Recall on CIFAR-10-LT and CIFAR-100-LT. We also report Precision, Density, Coverage, grouped long-tail coverage, and annealing schedule sensitivity.

5.1 Two-dimensional stable-tail task

We use the centered isotropic 1.7-stable task with scale matrix $0.05I_2$, denoted $\mathcal{S}_{1.7}(0, 0.05I_2)$, from the original DLPM evaluation (Shariatian et al., 2025). All trained configurations share 32,000 training samples, 100 diffusion steps, batch size 1024, 10,016 updates, and 20 paired seeds; each model generates 25,000 samples once with a seed matched to training. We compare A-DLPM with Gaussian DDPM, DLPM and LIM at each $\alpha \in \{1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$, and Denoising Lévy Implicit Models (DLIM) sampled deterministically from the corresponding DLPM checkpoints.

The primary metric is normalized mean squared logarithmic error (MSLE) between real and generated positive-tail quantiles of the first coordinate over $[0.95, 1 - 1/25000]$; lower is better. Figure 2 shows that A-DLPM achieves the lowest mean Tail MSLE among all evaluated configurations. Appendix G.1 defines the metric and also reports 250-bin Wasserstein distance; the Wasserstein score of A-DLPM remains close to the best result of competitors.

5.2 Long-tailed image experiments

Our image models are trained on the exponentially imbalanced CIFAR-10-LT and CIFAR-100-LT training splits (Krizhevsky, 2009; Cui et al., 2019); the class-resolved evaluation uses the correspond-

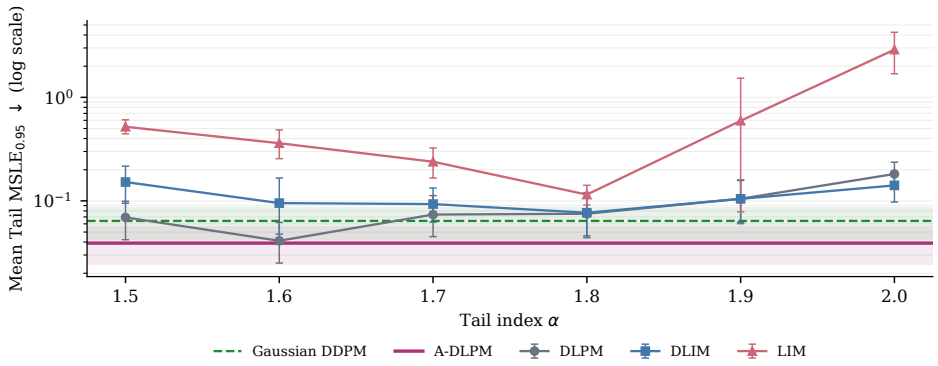


Figure 2: **Tail MSLE on the two-dimensional $\mathcal{S}_{1.7}$ task.** Curves show means and 95% bootstrap intervals over 20 paired training seeds for methods evaluated across α; horizontal lines and shaded regions show the corresponding means and intervals for A-DLPM and Gaussian DDPM.

ing original balanced CIFAR test splits. Every model uses the same improved-DDPM U-Net (Nichol & Dhariwal, 2021), and each method makes 999 network evaluations. We evaluate DLPM, DLIM, and LIM at $\alpha \in \{1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$, together with Gaussian DDPM and A-DLPM. DLIM applies deterministic sampling to the corresponding DLPM checkpoints and requires no additional training. For numerical stability, stable-noise image runs clip A_t at 20, following LIM and DLPM; Appendix H.6 examines this implementation choice. Training uses no labels, class balancing, auxiliary loss, or additional network. Appendix G gives the experiment details.

The main image comparisons use five paired training seeds and a shared training window of approximately 75k updates on each dataset. The training trajectories in Appendix H.1 instead follow one training run at three training checkpoints (50k, 75k, and 100k updates) and average FID and Recall over five generation runs per checkpoint. For the main comparisons, we evaluate each checkpoint with five nested generation seeds, average them within each training run, and report means over the five training seeds. FID and PRDC use a 5000-image reference drawn from the long-tailed training pool while preserving class proportions.

Table 1: **Main image results.** Means across five training runs after averaging five generation seeds within each run. DLIM uses the corresponding DLPM checkpoints. Boldface marks the best mean without implying significance.

	CIFAR-10-LT, 75k updates					*CIFAR-100-LT, 74,883 updates*				
Method	FID ↓	Prec. ↑	Recall ↑	Density ↑	Coverage ↑	FID ↓	Prec. ↑	Recall ↑	Density ↑	Coverage ↑
DLPM $\alpha = 1.5$	23.4528	0.7471	0.6270	0.9617	0.8493	30.5234	0.7618	0.6124	0.9472	0.7927
DLPM $\alpha = 1.6$	21.1991	0.7593	0.6317	0.9860	0.8585	30.1980	0.7756	0.6137	0.9648	0.7978
DLPM $\alpha = 1.7$	20.0888	0.7661	0.6338	1.0021	0.8652	28.7264	0.7878	0.6091	0.9944	0.8028
DLPM $\alpha = 1.8$	19.1624	0.7698	0.6374	1.0072	0.8660	28.6762	0.7987	0.6038	1.0160	0.8052
DLPM $\alpha = 1.9$	22.2891	0.7680	0.6370	0.9968	0.8721	30.3516	0.8041	0.5992	**1.0262**	0.7977
DLPM $\alpha = 2.0$	18.9972	0.7753	0.6369	1.0184	0.8789	31.6663	**0.8045**	0.6003	1.0138	0.7932
DLIM $\alpha = 1.5$	32.5045	0.7051	0.6443	0.8101	0.7928	37.2626	0.7339	0.6123	0.8402	0.7482
DLIM $\alpha = 1.6$	29.3500	0.7189	0.6434	0.8273	0.8005	35.8401	0.7500	0.6041	0.8657	0.7520
DLIM $\alpha = 1.7$	26.4444	0.7313	0.6424	0.8516	0.8103	33.7027	0.7703	0.5947	0.8983	0.7539
DLIM $\alpha = 1.8$	24.3380	0.7427	0.6395	0.8647	0.8131	32.4739	0.7824	0.5903	0.9233	0.7581
DLIM $\alpha = 1.9$	23.2319	0.7496	0.6399	0.8763	0.8242	33.6604	0.7943	0.5787	0.9511	0.7522
DLIM $\alpha = 2.0$	20.7598	0.7575	0.6447	0.9155	0.8439	33.2714	0.7981	0.5873	0.9628	0.7632
LIM $\alpha = 1.5$	23.2581	0.7276	0.6434	0.8301	0.7712	28.7539	0.7462	0.6166	0.8090	0.7297
LIM $\alpha = 1.6$	21.9072	0.7415	0.6369	0.8723	0.7885	27.2693	0.7556	0.6174	0.8576	0.7537
LIM $\alpha = 1.7$	21.1846	0.7427	0.6442	0.8906	0.8042	26.2500	0.7634	0.6246	0.8814	0.7780
LIM $\alpha = 1.8$	19.9418	0.7455	0.6498	0.8921	0.8235	**24.8241**	0.7736	0.6268	0.9285	0.8002
LIM $\alpha = 1.9$	21.2464	0.7385	0.6369	0.8598	0.8191	27.4396	0.7774	0.6158	0.9350	0.8017
LIM $\alpha = 2.0$	20.1665	0.7668	0.6472	0.9554	0.8731	34.7944	0.8012	0.5923	1.0060	0.7898
Gaussian DDPM	18.5799	0.7768	0.6456	1.0224	0.8866	31.5007	0.8010	0.6027	1.0107	0.8020
A-DLPM	**17.3010**	**0.7788**	**0.6537**	**1.0509**	**0.8891**	26.5218	0.7976	**0.6333**	1.0228	**0.8248**

Table 1 reports all five metrics, while Figure 3 shows the joint FID–Recall comparison. Across both datasets, A-DLPM has lower mean FID and higher mean Recall than every DLPM, DLIM, and Gaussian DDPM row; on CIFAR-10-LT this also holds for every LIM. On CIFAR-100-LT, A-DLPM has the highest Recall and Coverage. LIM at $\alpha = 1.7$ and 1.8 trades lower FID for lower Recall. Appendix H.2 reports intervals across paired training runs for all FID–Recall comparisons in Table 1.

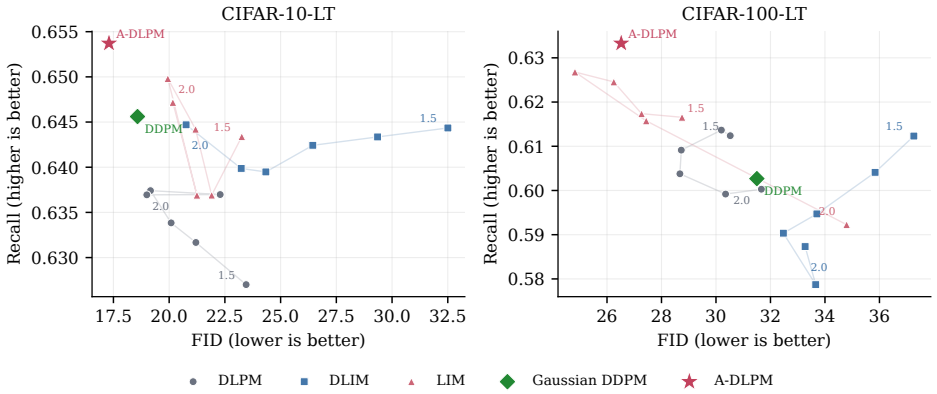


Figure 3: **FID–Recall.** Points are five-run means from Table 1. Curves connect $\alpha = 1.5$ through 2.0 within each family; labels mark the endpoints. A-DLPM has the highest mean Recall on both datasets.

The same Coverage pattern holds across class frequencies: in Figure 4, A-DLPM has the highest mean among the DLPM baselines for head, medium, and tail classes on both datasets. Appendix H.7 further resolves Recall and Coverage by individual class.

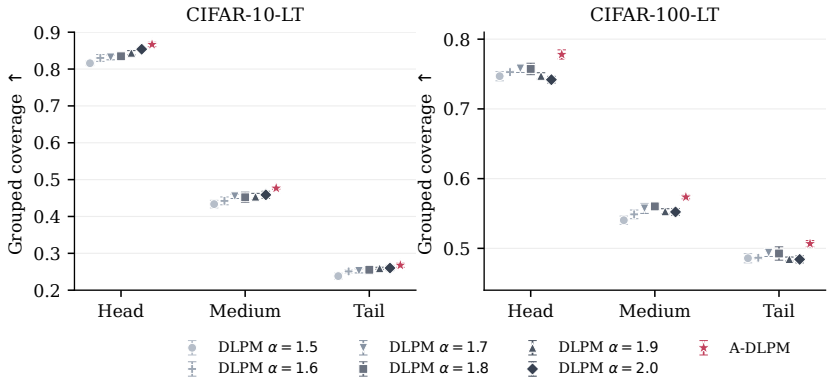


Figure 4: **Grouped coverage by class frequency.** Means and 95% Student-t intervals across five training runs after averaging generation seeds within each run. The original balanced CIFAR test splits provide the real queries.

We also examine the robustness of the annealed schedule from Eq. (3) under nearby design choices, such as varying the half-cosine curvature or moderately changing the lower endpoint 1.5; these variations lead to only modest changes in the reported metrics on both datasets. See Appendices H.3 and H.5. A whole-path control that mirrors the annealing direction instead produces a large deterioration on both datasets; Appendix H.4 gives the complete comparison.

6 Conclusion

A-DLPM couples a time-varying stable tail index to denoising diffusion, exactly recovers DLPM for a constant tail index, and admits mixed-power and continuous-time characterizations. On the two-dimensional task, A-DLPM achieves the lowest mean Tail MSLE. On long-tailed image benchmarks, A-DLPM improves the FID–Recall trade-off over baselines with a fixed tail index. Together with the prediction-error analysis, these results suggest that temporal allocation contributes to the improved fidelity–diversity trade-off.

Reproducibility Statement

We have made every effort to ensure that the results presented in this paper are reproducible. The paper specifies the model construction and learning objective, while the appendices provide complete algorithms and experimental settings. The experiments use procedurally generated synthetic data, and the publicly available CIFAR-10-LT and CIFAR-100-LT datasets (derived from CIFAR-10 and CIFAR-100). We provide an anonymous repository containing the implementation, configurations, and scripts for reproducing the reported experiments at `https://anonymous.4open.science/r/A-DLPM-DC0A/`. The repository records the exact training, generation, and data-loading settings, enabling reproduction of individual runs.

AI Use Statement

Large language models (LLMs) were used to improve the readability and language of the manuscript and to assist with code implementation. They were not involved in designing the research, conducting experiments, or performing the analysis. The authors reviewed and verified all LLM-assisted text and code, and take full responsibility for the scientific content.

References

Antoine Ayache. Sharp estimates on the tail behavior of a multistable distribution. *Statistics & Probability Letters*, 83(3):680–688, 2013.

Adrian Baule. Generative modelling with jump-diffusions. *arXiv preprint arXiv:2503.06558*, 2025. URL https://arxiv.org/abs/2503.06558.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pp. 1565–1576, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html.

John M. Chambers, Colin L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976. doi: 10.1080/01621459.1976.10480344.

Hamza Cherkaoui, Hélène Halconruy, and Antonio Ocello. Do heavy tails help diffusion? on the subtle trade-off between initialization and training. *arXiv preprint arXiv:2605.13175*, 2026. URL https://arxiv.org/abs/2605.13175.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019. doi: 10.1109/CVPR.2019.00949.

Jacob Deasy, Nikola Simidjievski, and Pietro Liò. Heavy-tailed denoising score matching. *arXiv preprint arXiv:2112.09788*, 2022.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*, volume 33 of *Applications of Mathematics: Stochastic Modelling and Applied Probability*. Springer, 1997. doi: 10.1007/978-3-642-33483-2.

Kenneth J. Falconer and Jacques Lévy Véhel. Multifractional, multistable, and other processes with prescribed local form. *Journal of Theoretical Probability*, 22(2):375–401, 2009. doi: 10.1007/s10959-008-0147-9.

Kenneth J. Falconer and Lining Liu. Multistable processes and localizability. *Stochastic Models*, 28 (3):503–526, 2012. doi: 10.1080/15326349.2012.699766.

Xiequan Fan and Jacques Lévy Véhel. Multistable Lévy motions and their continuous approximations. *arXiv preprint arXiv:1503.06623*, 2015.

Dongyoon Han, Jiwhan Kim, and Junmo Kim. Deep pyramidal residual networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5927–5935, 2017.

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. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tennessee Hickling and Dennis Prangle. Flexible tails for normalizing flows. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 23155–23178. PMLR, 2025. URL https://proceedings.mlr.press/v267/hickling25a.html.

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

Jean Jacod and Albert N. Shiryaev. *Limit Theorems for Stochastic Processes*, volume 288 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2 edition, 2003.

Priyank Jaini, Ivan Kobyzev, Yaoliang Yu, and Marcus Brubaker. Tails of Lipschitz triangular flows. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4673–4681. PMLR, 2020. URL https://proceedings.mlr.press/v119/jaini20a.html.

Yuchen Jiao, Yuchen Zhou, and Gen Li. Optimal convergence analysis of DDPM for general distributions. *arXiv preprint arXiv:2510.27562*, 2025. URL https://arxiv.org/abs/2510.27562.

Alexia Jolicoeur-Martineau, Kilian Fatras, Ke Li, and Tal Kachman. Diffusion models with location-scale noise. *arXiv preprint arXiv:2304.05907*, 2023.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Gen Li and Yuling Yan. $O(d/T)$ convergence theory for diffusion probabilistic models under minimal assumptions. In *International Conference on Learning Representations*, 2025.

Qi Lian, Yu Qi, and Yueming Wang. Cauchy diffusion: A heavy-tailed denoising diffusion probabilistic model for speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24549–24557, 2025. doi: 10.1609/aaai.v39i23.34634.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. arXiv:1711.05101.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Eliya Nachmani, Robin San Roman, and Lior Wolf. Denoising diffusion gamma models. *arXiv preprint arXiv:2110.05948*, 2021.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

John P. Nolan. *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer Series in Operations Research and Financial Engineering. Springer, 2020. doi: 10.1007/978-3-030-52915-4.

Jean Pachebat. Tail annealing for heavy-tailed flow matching. *arXiv preprint arXiv:2605.20068*, 2026. URL https://arxiv.org/abs/2605.20068.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.14171. arXiv:2410.14171.

Vadim Popov, Assel Yermekova, Tasnima Sadekova, Artem Khrapov, and Mikhail Kudinov. Improved sampling algorithms for Lévy–Itô diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XxCgeWSTNp.

Philip E. Protter. *Stochastic Integration and Differential Equations*, volume 21 of *Stochastic Modelling and Applied Probability*. Springer-Verlag, Berlin, 2 edition, 2005. doi: 10.1007/978-3-662-10061-5.

Balram S. Rajput and Jan Rosiński. Spectral representations of infinitely divisible processes. *Probability Theory and Related Fields*, 82(3):451–487, 1989. doi: 10.1007/BF00339998.

Gennady Samorodnitsky and Murad S. Taqqu. *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, 1994. ISBN 978-0-412-05171-5.

Ken-Iti Sato. *Lévy Processes and Infinitely Divisible Distributions*, volume 68 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, 1999.

Ken-iti Sato. Additive processes and stochastic integrals. *Illinois Journal of Mathematics*, 50(1-4): 825–851, 2006. doi: 10.1215/ijm/1258059494.

Dario Shariatian, Umut Simsekli, and Alain Durmus. Denoising Lévy probabilistic models. In *International Conference on Learning Representations*, 2025. arXiv:2407.18609.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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

Keito Wakatsuki and Hideaki Shimazaki. Self-regulating annealing in heavy-tailed diffusion models. *arXiv preprint arXiv:2606.01645*, 2026. URL https://arxiv.org/abs/2606.01645.

Ward Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.

Zhouhao Yang, Yezhen Wang, Kenji Kawaguchi, Vladimir Braverman, and Haoyang Cao. Heavy-tailed flow matching via random clocks. *arXiv preprint arXiv:2607.13841*, 2026. URL https://arxiv.org/abs/2607.13841.

Eun Bi Yoon, Keehun Park, Sungwoong Kim, and Sungbin Lim. Score-based generative models with Lévy processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40694–40707, 2023.

Tianjiao Zhang, Huangjie Zheng, Jiangchao Yao, Xiangfeng Wang, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed diffusion models with oriented calibration. In *International Conference on Learning Representations*, 2024.

Shuai Zhao, Liguang Zhou, Wenxiao Wang, Deng Cai, Tin Lun Lam, and Yangsheng Xu. Toward better accuracy-efficiency trade-offs: Divide and co-training. *IEEE Transactions on Image Processing*, 31:5869–5880, 2022. doi: 10.1109/TIP.2022.3201602.

APPENDIX

The appendix is organized as follows.

- Appendix A collects basic facts about positive stable scales and the training objective. We prove the positive-scale Gaussian representation, weighted closure at a constant tail index, and integrability of the training target.
- Appendix B provides proofs for the discrete construction: we derive the conditional forward marginal and the signal-correction identity in Appendix B.1, the conditional Gaussian reverse distribution for arbitrary retained indices in Appendix B.2, and exact recovery of DLPM when the tail index is constant in Appendix B.3.
- Appendix C provides proofs for the mixed-power spectrum: we show that the accumulated residual has a mixed-power characteristic exponent, identify its spatial tail index, and establish the spectral-curvature identity.
- Appendix D establishes the continuous-time limit: it constructs the additive driving process, verifies the deterministic stochastic integrals, proves J_1 Skorokhod convergence of the discrete residual, and derives the additional signal drift induced by a varying tail index.
- Appendix E details the training and sampling algorithms: it gives the discrete schedule construction and the complete pseudocode for training and respaced reverse sampling.
- Appendix F analyzes reverse-time prediction-error allocation: it defines the affinity statistic and hierarchical bootstrap, and reports the shift in error behavior across reverse time.
- Appendix G records experimental details: Appendix G.1 for the two-dimensional task and Appendix G.2 for the image task.
- Appendix H provides additional experimental results: Appendix H.1 reports checkpoint trajectories; Appendix H.2 reports paired differences and confidence intervals for the main image comparison; Appendices H.3, H.4, and H.5 examine sensitivity to curvature and path shape, annealing direction, and endpoints, respectively; Appendix H.6 examines sensitivity to scale clipping; Appendix H.7 reports per-class Recall and Coverage; and Appendix H.8 presents qualitative samples.

A NOTATION AND STABLE SCALE FACTS

We write $\|\cdot\|$ for the Euclidean norm, $\langle\cdot,\cdot\rangle$ for the Euclidean inner product, I_d for the $d \times d$ identity matrix, $\mathbb{E}$ for expectation, $\overset{\mathrm{d}}{=}$ for equality in distribution, $\Rightarrow$ for weak convergence (equivalently, convergence in distribution for random variables). The symmetric α-stable distribution $\mathcal{S}_\alpha$ and the positive stable distribution $\mathcal{A}_\beta$ are defined by Eqs. (1) and (2).

Lemma 3 (Positive-scale representation)**.** *Let $A \sim \mathcal{A}_{\alpha/2}$ and $Z \sim \mathcal{N}(0, I_d)$ be independent. Then $\sqrt{A}Z \sim \mathcal{S}_\alpha$.*

Proof. For every $u \in \mathbb{R}^d$, conditioning on A, using the Gaussian characteristic function and applying the Laplace transform of A gives

$$\begin{aligned}\mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, \sqrt{A}Z\rangle} &= \mathbb{E}\exp\left\{-\tfrac{1}{2}A\|u\|^2\right\} \\ &= \exp\left\{-\left(2\cdot\tfrac{1}{2}\|u\|^2\right)^{\alpha/2}\right\} = \mathrm{e}^{-\|u\|^\alpha}.\end{aligned}$$

Uniqueness of characteristic functions completes the proof. See also Samorodnitsky & Taqqu (1994, Equation 2.5.3). □

For $\alpha < 2$, our scale convention $\mathbb{E}\mathrm{e}^{-sA} = \mathrm{e}^{-(2s)^{\alpha/2}}$ follows by multiplying by 2 a positive $\alpha/2$-stable variable with Laplace transform $\exp(-s^{\alpha/2})$ (Nolan, 2020, Proposition 3.2; Samorodnitsky & Taqqu, 1994, Proposition 1.2.12); it matches the DLPM implementation, while at $\alpha = 2$ the same Laplace transform gives that $A = 2$ a.s. for $A \sim \mathcal{A}_1$, and $\sqrt{A}Z \sim \mathcal{S}_2 = \mathcal{N}(0, 2I_d)$.

Lemma 4 (Weighted closure for positive stable scales)**.** *Let $A, A_1, \dots, A_m \sim \mathcal{A}_{\alpha/2}$ be independent and identically distributed (i.i.d.), and let $b_1, \dots, b_m \geq 0$. Then*

$$\sum_{j=1}^{m} b_j^2 A_j \overset{\mathrm{d}}{=} \left(\sum_{j=1}^{m} b_j^\alpha \right)^{2/\alpha} A.$$

Proof. For $s \geq 0$, by the mutual independence of $A_1, \dots, A_m$ and Eq. (2),

$$\begin{aligned} \mathbb{E}\left[\exp\left\{ -s \sum_{j=1}^{m} b_j^2 A_j \right\} \right] &= \prod_{j=1}^{m} \exp\{-(2sb_j^2)^{\alpha/2}\} \\ &= \exp\left\{ -(2s)^{\alpha/2} \sum_{j=1}^{m} b_j^\alpha \right\}. \end{aligned}$$

Uniqueness of the Laplace transform (Nolan, 2020, Theorem A.3(a)) proves the claim. □

In the proof above, the common value of α is essential: with unequal tail indices, the product contains unequal powers $s^{\alpha_j/2}$ and cannot be reduced to one member of the same scale family. This failure of closure is precisely what gives rise to the mixed-power form in Theorem 1.

For $S \sim \mathcal{S}_\alpha$ with $0 < \alpha < 2$, each coordinate is a one-dimensional symmetric α-stable random variable. Samorodnitsky & Taqqu (1994, Property 1.2.16) gives

$$\mathbb{E}|S_i|^p < \infty \quad \text{for } 0 < p < \alpha, \qquad \mathbb{E}|S_i|^p = \infty \quad \text{for } p \geq \alpha.$$

Consequently, $\mathbb{E}\|S\|^p < \infty$ for $0 < p < \alpha$ and $\mathbb{E}\|S\|^p = \infty$ for $p \geq \alpha$. In particular, when $1 < \alpha < 2$, each coordinate of S has a finite mean but infinite variance, whereas the Gaussian endpoint $\alpha = 2$ has finite moments of all orders. Since all schedules considered here have tail indices greater than 1, every stable innovation $\sqrt{A_j} Z_j \sim \mathcal{S}_{\alpha_j}$ has a finite first moment. For $t \geq 1$, the training target $\epsilon_t = \bar{\sigma}_t^{-1} \sum_{j=1}^{t} w_{j,t} \sqrt{A_j} Z_j$ is therefore integrable by the triangle inequality. Consequently, the epsilon-RMSE population risk in Eq. (10) is well-defined.

B Forward Law and Conditional Reverse Distribution

B.1 Forward marginal

Proposition 5 (Conditional forward marginal)**.** *Under Eqs. (6) and (7),*

$$X_t \mid X_0, A_{0:t} \sim \mathcal{N}(\bar{\gamma}_t X_0, \Sigma_t I_d).$$

Proof. At $t = 0$, $\bar{\gamma}_0 = 1$ and $\Sigma_0 = 0$, so the stated law is the point mass at X_0. Now assume the claim holds at $t-1$, so that for some $\widetilde{Z}_{t-1} \mid (X_0, A_{0:t-1}) \sim \mathcal{N}(0, I_d)$, $X_{t-1} = \bar{\gamma}_{t-1} X_0 + \sqrt{\Sigma_{t-1}} \widetilde{Z}_{t-1}$ almost surely.

Substituting into Eq. (7) gives

$$\begin{aligned} X_t &= \gamma_t \bar{\gamma}_{t-1} X_0 + (\bar{\gamma}_t - \gamma_t \bar{\gamma}_{t-1}) X_0 + \gamma_t \sqrt{\Sigma_{t-1}} \widetilde{Z}_{t-1} + \sigma_t \sqrt{A_t} Z_t \\ &= \bar{\gamma}_t X_0 + \gamma_t \sqrt{\Sigma_{t-1}} \widetilde{Z}_{t-1} + \sigma_t \sqrt{A_t} Z_t. \end{aligned}$$

Conditional on X_0 and the scales $A_{0:t}$, the final two terms are independent centered Gaussians because Z_t is independent of the preceding innovations. Their conditional covariance matrices sum to $(\gamma_t^2 \Sigma_{t-1} + \sigma_t^2 A_t) I_d = \Sigma_t I_d$ by Eq. (6), proving the marginal by induction. □

Unrolling the same recursion in Eq. (6) gives the pathwise variance identity

$$\Sigma_t = \sum_{j=1}^{t} \left(\sigma_j \prod_{k=j+1}^{t} \gamma_k \right)^2 A_j = \sum_{j=1}^{t} w_{j,t}^2 A_j. \tag{21}$$

Whenever $\Sigma_t > 0$, the aggregate Gaussian in Eq. (8) can therefore be constructed explicitly as

$$\widetilde{Z}_t := \frac{1}{\sqrt{\Sigma_t}} \sum_{j=1}^{t} w_{j,t} \sqrt{A_j} Z_j.$$

Conditional on $A_{0:t}$, its covariance is I_d. When $\Sigma_t = 0$, take $\widetilde{Z}_t$ to be a standard Gaussian vector independent of X_0, the scale path, and the other Gaussian innovations. This preserves the conditional distribution $\widetilde{Z}_t \mid (X_0, A_{0:t}) \sim \mathcal{N}(0, I_d)$ stated after Eq. (8), without changing X_t because $\sqrt{\Sigma_t}\widetilde{Z}_t = 0$.

Lemma 6 (Discrete signal correction)**.** *Let $\delta_t := \bar{\gamma}_t - \gamma_t \bar{\gamma}_{t-1}$. For $t \geq 1$, if $\alpha_t \geq \alpha_{t-1}$ and $0 < q_{t-1} \leq 1$, we have $\delta_t \geq 0$, and if $\alpha_t \equiv \alpha$ then $\delta_t = 0$ for all t. Moreover,*

$$0 \leq \delta_t \leq \bar{\gamma}_t(-\log q_{t-1}) \left(\frac{1}{\alpha_{t-1}} - \frac{1}{\alpha_t} \right).$$

Proof. Using Eqs. (4) and (5), first note that

$$\frac{\gamma_t \bar{\gamma}_{t-1}}{\bar{\gamma}_t} = q_{t-1}^{1/\alpha_{t-1} - 1/\alpha_t}.$$

Therefore,

$$\delta_t = \bar{\gamma}_t \big(1 - q_{t-1}^{1/\alpha_{t-1} - 1/\alpha_t}\big). \tag{22}$$

The exponent is nonnegative because $\alpha_t \geq \alpha_{t-1}$; since $0 < q_{t-1} \leq 1$, this proves $\delta_t \geq 0$. When $\alpha_t \equiv \alpha$, the exponent is zero, hence $\delta_t = 0$. Finally set $x := -(1/\alpha_{t-1} - 1/\alpha_t) \log q_{t-1} \geq 0$. Then $\delta_t = \bar{\gamma}_t(1 - \mathrm{e}^{-x}) \leq \bar{\gamma}_t x$, which gives the bound.

□

Since the transition in Eq. (7) depends explicitly on X_0, the conditional law of X_t cannot be specified from X_{t-1} alone. Once $(X_0, A_{0:t})$ are fixed, however, the residual recursion is Markov and Gaussian. We therefore derive the reverse conditional distribution under the same conditioning.

B.2 Arbitrary-jump conditioning and the corrected coefficient

Proposition 7 (Conditional Gaussian reverse distribution)**.** *For indices $s < t$ from the original chain with $\Sigma_t > 0$, conditional on $(X_0, A_{0:t})$, the distribution of X_s given X_t is specified by Eqs. (12)–(15).*

Proof. Define $G_{s,t}$ by Eq. (13). Then Eq. (16) gives

$$Y_t = G_{s,t} Y_s + R_{s,t}, \qquad \text{where } R_{s,t} := \sum_{j=s+1}^{t} \sigma_j G_{j,t} \sqrt{A_j} Z_j.$$

Conditional on the scale path $A_{0:t}$, $R_{s,t}$ depends only on $Z_{s+1}, \ldots, Z_t$, hence it is centered Gaussian and independent of Y_s. Meanwhile, Eq. (6) gives

$$\Sigma_t = G_{s,t}^2 \Sigma_s + \sum_{j=s+1}^{t} \sigma_j^2 G_{j,t}^2 A_j.$$

We have

$$\mathrm{Var}(R_{s,t} \mid A_{0:t}) = \left(\sum_{j=s+1}^{t} \sigma_j^2 G_{j,t}^2 A_j \right) I_d = \big(\Sigma_t - G_{s,t}^2 \Sigma_s\big)\, I_d.$$

This identity also shows that $\Sigma_t - G_{s,t}^2 \Sigma_s \geq 0$.

Using $X_t - \bar{\gamma}_t X_0 = Y_t$ and $Y_t = G_{s,t} Y_s + R_{s,t}$,

$$\begin{aligned}\mathrm{Cov}(X_s, X_t \mid X_0, A_{0:t}) &= \mathbb{E}[Y_s Y_t^\top \mid X_0, A_{0:t}] \\ &= G_{s,t} \underbrace{\mathbb{E}[Y_s Y_s^\top \mid X_0, A_{0:t}]}_{\Sigma_s I_d} + \underbrace{\mathbb{E}[Y_s R_{s,t}^\top \mid X_0, A_{0:t}]}_{0} \\ &= G_{s,t} \Sigma_s I_d.\end{aligned}$$

Here Proposition 5 gives $\mathrm{Var}(Y_s \mid X_0, A_{0:t}) = \Sigma_s I_d$ and $\mathbb{E}[Y_s \mid X_0, A_{0:t}] = 0$.

Applying the Gaussian conditioning formula,

$$\begin{aligned}\mathbb{E}[X_s \mid X_t, X_0, A_{0:t}] &= \bar{\gamma}_s X_0 + G_{s,t} \frac{\Sigma_s}{\Sigma_t} (X_t - \bar{\gamma}_t X_0) \\ &= a_{s,t} X_t + b_{s,t} X_0,\end{aligned}$$

and

$$\mathrm{Var}(X_s \mid X_t, X_0, A_{0:t}) = \Sigma_s I_d - G_{s,t}^2 \frac{\Sigma_s^2}{\Sigma_t} I_d.$$

These are exactly Eqs. (14)–(15).

□

Proposition 7 assumes $\Sigma_t > 0$; if $\Sigma_t = 0$, all noise contributions up to time t vanish, and the conditional law is a point mass. For adjacent steps ($s = t - 1$), the coefficient reduces to $G_{s,t} = \gamma_t$. More generally, for respaced sampling, s and t must be retained indices from the original training chain, so $G_{s,t}$ accumulates all omitted one-step coefficients; constructing a fresh shorter schedule would define a different joint law outside the scope of the proposition.

Finally, note that the conditioning on $(X_0, A_{0:t})$ is essential: marginalizing the scale path turns the Gaussian pair into a scale mixture, and marginalizing X_0 introduces the unknown data posterior. Therefore the reverse update is exact only when both X_0 and the realized scales are known.

B.3 Exact recovery at a fixed tail index

Proposition 8 (Reduction to a fixed tail index)**.** *Suppose $\alpha_t \equiv \alpha$ and $q_t > 0$ throughout the original chain, and let $A \sim \mathcal{A}_{\alpha/2}$. Then: (i) the correction in Eq. (7) is zero; (ii) $G_{s,t} = \bar{\gamma}_t / \bar{\gamma}_s$; and (iii) $\Sigma_t \overset{\mathrm{d}}{=} \bar{\sigma}_t^2 A$.*

Proof. For constant α,

$$\gamma_t \bar{\gamma}_{t-1} = (q_t / q_{t-1})^{1/\alpha} q_{t-1}^{1/\alpha} = q_t^{1/\alpha} = \bar{\gamma}_t,$$

which proves (i). For part (ii), write the product definition of $G_{s,t}$ in Eq. (13):

$$G_{s,t} = \prod_{k=s+1}^{t} \gamma_k = \prod_{k=s+1}^{t} (q_k / q_{k-1})^{1/\alpha} = (q_t / q_s)^{1/\alpha} = \frac{\bar{\gamma}_t}{\bar{\gamma}_s}.$$

Eq. (21) and Lemma 4 give

$$\Sigma_t = \sum_{j=1}^{t} w_{j,t}^2 A_j \overset{\mathrm{d}}{=} \left(\sum_{j=1}^{t} w_{j,t}^\alpha \right)^{2/\alpha} A.$$

Define $C_t := \sum_{j=1}^{t} w_{j,t}^\alpha$. Since $w_{t,t} = \sigma_t$ and $w_{j,t} = \gamma_t w_{j,t-1}$ for $j < t$ (by the definition in Eq. (17)), we have

$$C_t = \gamma_t^\alpha C_{t-1} + \sigma_t^\alpha.$$

At a fixed tail index α, $\gamma_t^\alpha = q_t / q_{t-1}$, $\sigma_t^\alpha = 1 - q_t / q_{t-1}$, and $\bar{\sigma}_t^\alpha = 1 - q_t$ (by Eqs. (4) and (5)). Starting from $C_0 := 0 = \bar{\sigma}_0^\alpha$, induction gives $C_t = 1 - q_t = \bar{\sigma}_t^\alpha$. The mutual independence of

$A_1, \dots, A_t$ and Eq. (2) imply

$$\begin{aligned}\mathbb{E}\mathrm{e}^{-s\Sigma_t} &= \prod_{j=1}^{t} \exp\{-(2sw_{j,t}^2)^{\alpha/2}\} \\ &= \exp\left\{-(2s)^{\alpha/2}\sum_{j=1}^{t} w_{j,t}^{\alpha}\right\} = \exp\left\{-(2s)^{\alpha/2}C_t\right\} \\ &= \exp\{-(2s\bar{\sigma}_t^2)^{\alpha/2}\}.\end{aligned}$$

This is the Laplace transform of $\bar{\sigma}_t^2 A$ in Eq. (2), which proves (iii). □

Proposition 8 also establishes the case used by the DLPM baselines. Part (i) reduces the forward transition to

$$X_t = \gamma_t X_{t-1} + \sigma_t \sqrt{A_t} Z_t.$$

This recursion matches the augmented DLPM construction in initial state, coefficients, scales, and innovations. Hence the full joint law of $(X_0, \dots, X_{N-1}, A_{0:N-1})$ coincides with that of DLPM. Part (ii) reduces their residual propagation coefficient to the marginal ratio for a fixed tail index. Part (iii) shows that the accumulated scale Σ_t collapses to the single-scale DLPM marginal when the tail index is constant.

C Proofs for the Mixed-Power Spectrum

Proof of Theorem 1. Each term in Eq. (17) is an independent isotropic stable vector, with characteristic function (by Eq. (1))

$$\mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, w_{j,t}\sqrt{A_j}Z_j\rangle} = \exp\{-w_{j,t}^{\alpha_j}\|u\|^{\alpha_j}\}.$$

Multiplying the characteristic functions proves Eq. (18).

All terms with $w_{j,t} = 0$ can be removed. Let $J_t := \{j : w_{j,t} > 0\}$ be the set of active terms, $\alpha_{\min} := \min_{j\in J_t} \alpha_j$ and $\alpha_{\max} := \max_{j\in J_t} \alpha_j$. Suppose $\alpha_{\min} < \alpha_{\max}$ and $\Psi_t(r) = Cr^{\alpha}$ for some $C > 0$ and all $r > 0$. Factoring out the smallest power gives $\Psi_t(r)/r^{\alpha_{\min}} \to \sum_{j\in J_t:\alpha_j=\alpha_{\min}} w_{j,t}^{\alpha_j} > 0$ as $r \downarrow 0$, forcing $\alpha = \alpha_{\min}$. Similarly, factoring out the largest power and letting $r \uparrow \infty$ yields $\alpha = \alpha_{\max}$, a contradiction. □

The proof covers the possible degeneracies. If there is no active term, Y_t is deterministic zero. If all active terms share one tail index, their coefficients add and the residual is stable at that common tail index. The conclusion that the residual is not a rescaled stable law therefore holds only when at least two active tail indices differ.

The smallest active tail index, which controls the low-frequency limit of $\alpha_{\mathrm{eff},t}(r)$, also determines the tail of the residual Y_t. However, this connection is not immediate from the characteristic function in Eq. (18), and is established separately in the following proposition.

Proposition 9 (Tail index of the mixed-power residual)**.** *Let $J_t := \{j : w_{j,t} > 0\}$ be nonempty and set $\alpha_{\min} := \min_{j\in J_t} \alpha_j < 2$. For every nonzero $v \in \mathbb{R}^d$, there exists a constant $C_v > 0$ such that*

$$\Pr(|\langle v, Y_t\rangle| > x) \sim C_v x^{-\alpha_{\min}} \qquad \text{as } x \to \infty.$$

Proof. Group all terms in Eq. (17) with tail index $\alpha_{\min}$ into U and all remaining terms into V, that is,

$$\begin{aligned}U &:= \sum_{j:\, w_{j,t}>0,\ \alpha_j=\alpha_{\min}} w_{j,t}\langle v, \sqrt{A_j}Z_j\rangle, \\ V &:= \sum_{j:\, w_{j,t}>0,\ \alpha_j>\alpha_{\min}} w_{j,t}\langle v, \sqrt{A_j}Z_j\rangle.\end{aligned}$$

Then $\langle v, Y_t\rangle = U + V$, and U is independent of V.

All summands in U are independent symmetric $\alpha_{\min}$-stable variables with a common tail index, so U is also symmetric $\alpha_{\min}$-stable. By Nolan (2020, Theorem 1.2), there exists $C_v > 0$ such that $\Pr(|U| > x) = C_v x^{-\alpha_{\min}}(1 + o(1))$. If $V \neq 0$, choose

$$\alpha_{\min} < p < \min_{j \in J_t:\, \alpha_j > \alpha_{\min}} \alpha_j.$$

Every summand of V has a finite pth absolute moment, hence so does V. Markov's inequality then gives $\Pr(|V| > \delta x) = O((\delta x)^{-p}) = o((\delta x)^{-\alpha_{\min}})$ for every fixed $\delta \in (0, 1)$. The triangle inequality yields

$$\begin{aligned}\Pr(|U + V| > x) &\le \Pr(|U| > (1 - \delta)x) + \Pr(|V| > \delta x) \\ &= C_v\big((1 - \delta)x\big)^{-\alpha_{\min}}(1 + o(1)) + o\big((\delta x)^{-\alpha_{\min}}\big).\end{aligned}$$

Likewise,

$$\begin{aligned}\Pr(|U + V| > x) &\ge \Pr(|U| > (1 + \delta)x) - \Pr(|V| > \delta x) \\ &= C_v\big((1 + \delta)x\big)^{-\alpha_{\min}}(1 + o(1)) - o\big((\delta x)^{-\alpha_{\min}}\big).\end{aligned}$$

Dividing these bounds by $x^{-\alpha_{\min}}$ and letting $x \to \infty$ gives

$$\begin{aligned}C_v(1 + \delta)^{-\alpha_{\min}} &\le \liminf_{x\to\infty} x^{\alpha_{\min}} \Pr(|U + V| > x) \\ &\le \limsup_{x\to\infty} x^{\alpha_{\min}} \Pr(|U + V| > x) \\ &\le C_v(1 - \delta)^{-\alpha_{\min}}.\end{aligned}$$

Letting $\delta \downarrow 0$ proves the stated asymptotic. If $V = 0$, the conclusion follows directly from the same stable-tail asymptotic (Nolan, 2020, Theorem 1.2). □

Proposition 9 shows that, although components with higher tail indices alter the distribution of the residual, they do not remove the asymptotic power-law tail inherited from the smallest active tail index. Related tail asymptotics for multistable stochastic integrals are also studied by Ayache (2013).

Proof of Theorem 2. Write $s := \log r$, $c_j := w_{j,t}^{\alpha_j}$, and $M_t(s) := \Psi_t(\mathrm{e}^s) = \sum_{j=1}^t c_j \mathrm{e}^{\alpha_j s}$. Then

$$\frac{\mathrm{d}M_t(s)}{\mathrm{d}s} = \sum_{j=1}^t c_j \alpha_j \mathrm{e}^{\alpha_j s}, \qquad \frac{\mathrm{d}^2 M_t(s)}{\mathrm{d}s^2} = \sum_{j=1}^t c_j \alpha_j^2 \mathrm{e}^{\alpha_j s}.$$

Differentiating the log-sum-exp gives

$$\alpha_{\mathrm{eff},t}(\mathrm{e}^s) := \frac{\mathrm{d}}{\mathrm{d}s} \log \Psi_t(\mathrm{e}^s) = \frac{\sum_{j=1}^t c_j \alpha_j \mathrm{e}^{\alpha_j s}}{\sum_{j=1}^t c_j \mathrm{e}^{\alpha_j s}} = \sum_{j=1}^t \pi_{j,t}(\mathrm{e}^s)\alpha_j.$$

Furthermore,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\pi_{j,t}(\mathrm{e}^s) = \frac{\mathrm{d}}{\mathrm{d}s}\frac{c_j \mathrm{e}^{\alpha_j s}}{M_t(s)} &= \frac{c_j \alpha_j \mathrm{e}^{\alpha_j s} M_t(s) - c_j \mathrm{e}^{\alpha_j s} \frac{\mathrm{d}M_t(s)}{\mathrm{d}s}}{M_t(s)^2} \\ &= \frac{c_j \mathrm{e}^{\alpha_j s}}{M_t(s)}\left(\alpha_j - \frac{1}{M_t(s)}\frac{\mathrm{d}M_t(s)}{\mathrm{d}s}\right) \\ &= \pi_{j,t}(\mathrm{e}^s)\big(\alpha_j - \alpha_{\mathrm{eff},t}(\mathrm{e}^s)\big).\end{aligned}$$

Then differentiating $\alpha_{\mathrm{eff},t}(\mathrm{e}^s)$ yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\alpha_{\mathrm{eff},t}(\mathrm{e}^s) = \sum_{j=1}^t \alpha_j \frac{\mathrm{d}}{\mathrm{d}s}\pi_{j,t}(\mathrm{e}^s) &= \sum_{j=1}^t \pi_{j,t}(\mathrm{e}^s)\alpha_j\big(\alpha_j - \alpha_{\mathrm{eff},t}(\mathrm{e}^s)\big) \\ &= \sum_{j=1}^t \pi_{j,t}(\mathrm{e}^s)\alpha_j^2 - \left(\sum_{j=1}^t \pi_{j,t}(\mathrm{e}^s)\alpha_j\right)^2,\end{aligned}$$

which is the stated weighted variance. Its equivalent pairwise form is

$$\sum_{j=1}^{t} \pi_{j,t}(\mathrm{e}^s)\big(\alpha_j - \alpha_{\mathrm{eff},t}(\mathrm{e}^s)\big)^2 = \frac{1}{2}\sum_{j=1}^{t}\sum_{k=1}^{t} \pi_{j,t}(\mathrm{e}^s)\pi_{k,t}(\mathrm{e}^s)(\alpha_j - \alpha_k)^2.$$

Since $\pi_{j,t}(r) > 0$ for every active term at each finite $r > 0$, the pairwise form is strictly positive exactly when two active tail indices differ, and zero exactly when all active tail indices agree. □

Theorem 2 shows that $\alpha_{\mathrm{eff},t}(r)$ is a weighted average of the active tail indices, so it lies between $\alpha_{\min}$ and $\alpha_{\max}$. These bounds are attained in the limit: since $\Psi_t(r)/r^{\alpha_{\min}} \to C_{\min} > 0$ as $r \downarrow 0$ and $\Psi_t(r)/r^{\alpha_{\max}} \to C_{\max} > 0$ as $r \uparrow \infty$, $\alpha_{\mathrm{eff},t}(r) \to \alpha_{\min}$ and $\alpha_{\mathrm{eff},t}(r) \to \alpha_{\max}$ at the respective endpoints. Together with nonnegative curvature, the effective tail index $\alpha_{\mathrm{eff},t}(r)$ moves monotonically between these endpoints.

D Continuous-Time Limit

Processes with a time-varying stable tail index have been studied as multistable processes in the probability literature, including local-form and stochastic-integral constructions (Falconer & Lévy Véhel, 2009; Falconer & Liu, 2012), sharp tail estimates (Ayache, 2013), and functional limit theorems with continuous approximations (Fan & Lévy Véhel, 2015). These constructions establish well-defined stochastic processes but do not specify a denoising target or a learned reverse sampler. In our setting, the intensity $\beta(t)$ and propagation kernel $K(s,t)$ are determined by the discrete A-DLPM coefficients. We therefore establish the functional limit of this specific recursion directly.

Fix a finite time horizon $T > 0$.

Assumption 10 (Regular schedules)**.** Assume $q : [0,T] \to [0,1]$ is continuously differentiable, $q(0) = 1$, $q(t) > 0$ for $t < T$, and, for $t \in [0,T)$, $\beta(t) := -\partial_t \log q(t)$ is bounded and nonnegative on compact subsets of $[0,T)$. The tail index function $\alpha : [0,T] \to (1,2]$ is continuously differentiable.

Since the range of α is compact in $(1,2]$, fix constants α_- and α_+ such that $1 < \alpha_- \le \alpha(t) \le \alpha_+ \le 2$ on $[0,T]$.

Since $\alpha(t)$ and $\beta(t)$ vary with time, the driving noise must be an additive process: a stochastically continuous process with independent but generally nonstationary increments. To construct such a process, we specify the increment law on each interval through its Lévy–Khintchine triplet, denoted by (b, Q, ν): b is the drift vector, Q the Gaussian covariance matrix, and ν the Lévy measure, with

$$b \in \mathbb{R}^d, \quad Q = Q^\top \succeq 0, \quad \nu(\{0\}) = 0, \quad \int_{\mathbb{R}^d} (1 \wedge \|x\|^2)\, \nu(\mathrm{d}x) < \infty.$$

For the symmetric laws used below, $b = 0$ and ν is symmetric, so the negative log characteristic function reduces to

$$\psi(u) := \frac{1}{2}\langle u, Qu\rangle + \int_{\mathbb{R}^d} (1 - \cos\langle u, x\rangle)\, \nu(\mathrm{d}x).$$

Define the reference measure $\mathfrak{m}(\mathrm{d}r) := \beta(r)\mathrm{d}r$ on $[0,T)$, which intuitively measures the noise intensity along the time axis. It is finite on every compact subinterval because β is locally bounded.

The next lemma constructs an additive process by requiring this ψ to match the desired interval exponent.

Lemma 11 (Existence of the stable-type additive process)**.** *Under Assumption 10, there exists a càdlàg additive process L on $[0,T)$, with $L_0 = 0$, whose increments satisfy, for $0 \le s \le t < T$ and $u \in \mathbb{R}^d$,*

$$\mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, L_t - L_s\rangle} = \exp\left\{-\int_s^t \beta(r)\|u\|^{\alpha(r)}\mathrm{d}r\right\}. \tag{23}$$

Proof. For clarity, we divide the proof into four steps.

Step 1. We first identify the Lévy–Khintchine characteristics whose characteristic exponent is $\|u\|^{\alpha(r)}$. The case $\alpha(r) < 2$ is represented by a symmetric Lévy measure, whereas $\alpha(r) = 2$ is represented by a Gaussian covariance.

For $a \in (1, 2)$, define

$$c_{d,a} := \frac{a2^{a-1}\Gamma((d+a)/2)}{\pi^{d/2}\Gamma(1-a/2)}.$$

Whenever $\alpha(r) < 2$, let

$$\rho_r(\mathrm{d}x) := c_{d,\alpha(r)}\|x\|^{-d-\alpha(r)}\mathrm{d}x.$$

A polar-coordinate calculation gives

$$\int_{\mathbb{R}^d}(1 \wedge \|x\|^2)\rho_r(\mathrm{d}x) = |\mathbb{S}^{d-1}|c_{d,\alpha(r)}\left(\frac{1}{2-\alpha(r)} + \frac{1}{\alpha(r)}\right).$$

Here $|\mathbb{S}^{d-1}|$ is the surface area of the unit sphere in $\mathbb{R}^d$. This quantity is uniformly bounded for $\alpha(r) \in [\alpha_-, 2)$: on compact subsets of $(1, 2)$ it follows from continuity; while $1/\Gamma(1-a/2) = O(2-a)$ cancels $(2-a)^{-1}$ as $a \uparrow 2$. Hence ρ_r is a symmetric Lévy measure, and $\sup_r \int_{\mathbb{R}^d}(1 \wedge \|x\|^2)\rho_r(\mathrm{d}x) < \infty$.

For $u \neq 0$, we have

$$\begin{aligned}
\int_{\mathbb{R}^d}(1-\cos\langle u, x\rangle)\rho_r(\mathrm{d}x) &= c_{d,\alpha(r)}\int_{\mathbb{R}^d}\frac{1-\cos\langle u, x\rangle}{\|x\|^{d+\alpha(r)}}\mathrm{d}x \\
&= c_{d,\alpha(r)}\|u\|^{\alpha(r)}\int_{\mathbb{R}^d}\frac{1-\cos y_1}{\|y\|^{d+\alpha(r)}}\mathrm{d}y \\
&= c_{d,\alpha(r)}\|u\|^{\alpha(r)}\frac{\pi^{d/2}\Gamma(1-\alpha(r)/2)}{\alpha(r)2^{\alpha(r)-1}\Gamma((d+\alpha(r))/2)} \\
&= \|u\|^{\alpha(r)},
\end{aligned}$$

where y_1 is the first coordinate of y, and the second equality follows from rotational invariance and the change of variables $y = \|u\|x$. For $u = 0$, the identity above is immediate.

Since α is continuous, the mapping $(r, x) \mapsto c_{d,\alpha(r)}\|x\|^{-d-\alpha(r)}$ is jointly measurable on $\{\alpha < 2\} \times (\mathbb{R}^d \setminus \{0\})$, so $r \mapsto \rho_r$ is a measurable kernel. For $\alpha(r) = 2$, the Gaussian covariance is $2I_d$ per unit $\mathfrak{m}$-mass, yielding the exponent $\|u\|^2$. The drift is zero for all r.

Step 2. We next integrate these characteristics over $[s, t]$ to obtain a valid infinitely divisible law. The Lévy measure of the increment law over $[s, t]$ is

$$\nu_{s,t}(B) := \int_{\{r\in[s,t]:\alpha(r)<2\}}\rho_r(B)\,\mathfrak{m}(\mathrm{d}r), \qquad B \subseteq \mathbb{R}^d \setminus \{0\} \text{ Borel}.$$

This is a symmetric Lévy measure because Tonelli's theorem and the uniform bound above give

$$\int_{\mathbb{R}^d}(1 \wedge \|x\|^2)\,\nu_{s,t}(\mathrm{d}x) = \int_{\{r\in[s,t]:\alpha(r)<2\}}\left(\int_{\mathbb{R}^d}(1 \wedge \|x\|^2)\rho_r(\mathrm{d}x)\right)\mathfrak{m}(\mathrm{d}r) < \infty.$$

The Gaussian component of the integrated law is $2I_d\,\mathfrak{m}(\{r \in [s, t] : \alpha(r) = 2\})$, and the drift is zero. Hence the Lévy–Khintchine exponent of this triplet is

$$\begin{aligned}
\psi_{s,t}(u) &:= \mathfrak{m}(\{r \in [s, t] : \alpha(r) = 2\})\|u\|^2 + \int_{\{r\in[s,t]:\alpha(r)<2\}}\underbrace{\left(\int_{\mathbb{R}^d}(1-\cos\langle u, x\rangle)\rho_r(\mathrm{d}x)\right)}_{\|u\|^{\alpha(r)}}\mathfrak{m}(\mathrm{d}r) \\
&= \int_s^t \beta(r)\|u\|^{\alpha(r)}\mathrm{d}r,
\end{aligned}$$

where the second equality uses the identity from Step 1 on $\{\alpha < 2\}$ and $\|u\|^2 = \|u\|^{\alpha(r)}$ on $\{\alpha = 2\}$. Define

$$\widehat{\mu}_{s,t}(u) := \exp\{-\psi_{s,t}(u)\} = \exp\left\{-\int_s^t \beta(r)\|u\|^{\alpha(r)}\mathrm{d}r\right\}.$$

The Lévy–Khintchine theorem (Sato, 1999, Theorem 8.1) shows that this is the characteristic function of an infinitely divisible probability law $\mu_{s,t}$.

Step 3. We verify that the interval laws are consistent and stochastically continuous. Since the exponent is additive, for $0 \le s \le t \le v < T$,

$$\widehat{\mu}_{s,t}(u)\widehat{\mu}_{t,v}(u) = \widehat{\mu}_{s,v}(u),$$

so $\mu_{s,t} * \mu_{t,v} = \mu_{s,v}$, while $\mu_{s,s} = \delta_0$, the point mass at zero. Fix $\tau < T$. For $0 \le s \le t \le \tau$,

$$\int_s^t \beta(r)\|u\|^{\alpha(r)}\mathrm{d}r \le \|\beta\|_{L^\infty(0,\tau)}|t-s|\max\{\|u\|^{\alpha_-}, \|u\|^{\alpha_+}\} \longrightarrow 0$$

whenever $t - s \to 0$ within that compact interval. Lévy's continuity theorem therefore gives both $\mu_{s,t} \Rightarrow \delta_0$ as $s \uparrow t$ and $t \downarrow s$. This establishes the consistency and stochastic continuity needed for the existence theorem in the next step.

Step 4. By Sato (1999, Theorem 9.7(ii)), the increment laws verified in Step 3 determine an additive process in law; by Sato (1999, Theorem 11.5), it admits a càdlàg modification. Although both results are stated on $[0, \infty)$, the same conclusion holds on $[0, T)$ through the increasing homeomorphism $h(a) := Ta/(1 + a)$. Indeed, apply them to $\widetilde{\mu}_{a,b} := \mu_{h(a),h(b)}$ and pull the resulting process $\widetilde{L}$ back as $L_r := \widetilde{L}_{h^{-1}(r)}$. Then L has increment laws $\mu_{s,t}$, proving the lemma.

□

The increments of L are independent but generally nonstationary. If $\alpha(r) \equiv \alpha$, the characteristic exponent in Eq. (23) reduces to $\|u\|^\alpha \int_s^t \beta(r)\mathrm{d}r$, so the increment law is symmetric α-stable with scale $\left(\int_s^t \beta(r)\mathrm{d}r\right)^{1/\alpha}$. Equivalently, L is a symmetric α-stable Lévy process under the deterministic time change $t \mapsto \int_0^t \beta(r)\mathrm{d}r$. When α is nonconstant on a set of positive measure, the integral $\int_s^t \beta(r)\|u\|^{\alpha(r)}\mathrm{d}r$ generally cannot be written as $c\|u\|^\alpha$ for any fixed α, so L is not a single stable Lévy process.

To formulate the continuous limit of the discrete process, for every $\tau < T$ let $0 = t_{0,n} < \cdots < t_{m_n,n} = \tau$ be partitions whose maximal subinterval width tends to zero. Write $\Delta t_{k,n} := t_{k,n} - t_{k-1,n}$ and mesh $|\pi_n| := \max_k \Delta t_{k,n}$, which satisfies $|\pi_n| \to 0$ as $n \to \infty$, and set $q_{k,n} := q(t_{k,n})$ and $\alpha_{k,n} := \alpha(t_{k,n})$. On each partition, define

$$\gamma_{k,n} := \left(\frac{q_{k,n}}{q_{k-1,n}}\right)^{1/\alpha_{k,n}}, \qquad \sigma_{k,n} := \left(1 - \frac{q_{k,n}}{q_{k-1,n}}\right)^{1/\alpha_{k,n}}, \qquad k = 1, \ldots, m_n.$$

These definitions are consistent with the discrete coefficients in Eq. (5). Nonnegativity of β makes q nonincreasing, so these coefficients are real and nonnegative. Let $\xi_{k,n}$ be independent isotropic stable innovations such that, for $u \in \mathbb{R}^d$,

$$\mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, \xi_{k,n}\rangle} := \exp\{-\sigma_{k,n}^{\alpha_{k,n}}\|u\|^{\alpha_{k,n}}\}. \tag{24}$$

Define the discrete residual by $Y^{(n)}_{t_{0,n}} := 0$ and $Y^{(n)}_{t_{k,n}} := \gamma_{k,n} Y^{(n)}_{t_{k-1,n}} + \xi_{k,n}$, which is consistent with Eq. (16). For $s \in [0, \tau]$, extend it piecewise constantly by $Y^{(n)}_s := Y^{(n)}_{\rho_n(s)}$, where $\rho_n(s) := \max\{t_{j,n} : t_{j,n} \le s\}$. Hence $Y^{(n)}_s$ is càdlàg.

With L supplied by Lemma 11, the remainder of this section uses the same $\mathfrak{m}$ and ρ_r as its proof. Denote $\kappa(t) := \beta(t)/\alpha(t)$, which is bounded on every compact subinterval of $[0, T)$. Throughout, $\mathrm{d}L_s$ denotes integration with respect to the independently scattered increment measure induced by L, whose value on $(s, t]$ is $L_t - L_s$.

Lemma 12 (Bounded deterministic integrands)**.** *Fix $t < T$. If $g : [0, t] \to \mathbb{R}^d$ is bounded and Borel measurable, then*

$$I_g := \int_0^t \langle g(s), \mathrm{d}L_s\rangle$$

is well-defined and, for every $\theta \in \mathbb{R}$,

$$\mathbb{E}\mathrm{e}^{\mathrm{i}\theta I_g} = \exp\left\{-\int_0^t \beta(s)\|\theta g(s)\|^{\alpha(s)}\mathrm{d}s\right\}.$$

Proof. Let $M := \sup_{s\in[0,t]}\|g(s)\| < \infty$. The jump condition satisfies

$$\int_{\{s\in[0,t]:\alpha(s)<2\}}\int_{\mathbb{R}^d}\big(1\wedge|\langle g(s),x\rangle|^2\big)\rho_s(\mathrm{d}x)\mathfrak{m}(\mathrm{d}s)$$
$$\le \max\{1,M^2\}\int_{\{s\in[0,t]:\alpha(s)<2\}}\int_{\mathbb{R}^d}(1\wedge\|x\|^2)\rho_s(\mathrm{d}x)\mathfrak{m}(\mathrm{d}s) < \infty,$$

and the Gaussian condition satisfies

$$\int_{\{s\in[0,t]:\alpha(s)=2\}} 2\|g(s)\|^2\mathfrak{m}(\mathrm{d}s) \le 2M^2\mathfrak{m}([0,t]) < \infty.$$

The drift condition vanishes by symmetry. So I_g is well-defined by Rajput & Rosiński (1989, Theorem 2.7). Identities

$$\begin{cases}\displaystyle\int_{\mathbb{R}^d}\big(1-\cos\langle\theta g(s),x\rangle\big)\rho_s(\mathrm{d}x) = \|\theta g(s)\|^{\alpha(s)}, & \alpha(s)<2,\\ \frac{1}{2}\langle\theta g(s), 2I_d\,\theta g(s)\rangle = \|\theta g(s)\|^2, & \alpha(s)=2,\end{cases}$$

together with $\mathfrak{m}(\mathrm{d}s) = \beta(s)\mathrm{d}s$ give the stated formula. See also Sato (2006) for stochastic integration with respect to additive processes. □

The deterministic kernel $K(s,t)$ in Eq. (19) satisfies $0 < K(s,t) \le 1$ for $0 \le s \le t < T$. The vector integral $Y_t = \int_0^t K(s,t)\mathrm{d}L_s$ is defined coordinatewise by Lemma 12. Applying the same lemma to $g(s) = K(s,t)u$ and using linearity gives, for every $u \in \mathbb{R}^d$,

$$\mathbb{E}\mathrm{e}^{\mathrm{i}\langle u,Y_t\rangle} = \exp\left\{-\int_0^t \beta(s)K(s,t)^{\alpha(s)}\|u\|^{\alpha(s)}\mathrm{d}s\right\}.$$

The Lévy measure and Gaussian covariance components arise from the disjoint measurable sets $\{\alpha<2\}$ and $\{\alpha=2\}$, so their exponents add without taking the limit of a stable Lévy measure as $\alpha\uparrow 2$.

When $\alpha(s)\equiv\alpha\in(1,2)$, the integrability condition for the stochastic integral $\int_0^t K(s,t)\,\mathrm{d}L_s$ reduces to

$$\int_0^t |K(s,t)|^\alpha\mathfrak{m}(\mathrm{d}s) < \infty,$$

which is the standard condition for a symmetric α-stable stochastic integral (Samorodnitsky & Taqqu, 1994, Eq. (3.4.1) and Proposition 3.4.1). For a varying tail index, this condition is no longer applicable, which is why we instead verify the general conditions of Rajput & Rosiński (1989, Theorem 2.7).

Theorem 13 (Functional limit)**.** *Under Assumption 10, for every $\tau<T$, let $Y^{(n)}$ be the piecewise-constant càdlàg interpolation of the discrete residual on the partitions of $[0,\tau]$ defined above. Then*

$$Y^{(n)} \;\Rightarrow\; Y \qquad \text{in } \big(D([0,\tau];\mathbb{R}^d), J_1\big), \tag{25}$$

where $Y_t = \int_0^t K(s,t)\mathrm{d}L_s$ is the continuous-time residual in Eq. (19), with L supplied by Lemma 11. In particular, the finite-dimensional convergence holds at every finite collection of times in $[0,T)$. If $q(T)>0$ and β is bounded on $[0,T]$, then the same conclusion holds with $\tau=T$.

At radius $\rho\ge 0$, the marginal characteristic exponent of Y_t is

$$\Psi_t^{\mathrm{ct}}(\rho) := \int_0^t \beta(s)K(s,t)^{\alpha(s)}\rho^{\alpha(s)}\mathrm{d}s. \tag{26}$$

For $X_t := q(t)^{1/\alpha(t)}X_0 + Y_t$, the full state satisfies Eq. (20).

Proof. We first construct an auxiliary additive process $\widetilde{L}^{(n)}$ and an auxiliary residual $\widetilde{Y}^{(n)}$, then identify $\widetilde{Y}^{(n)}$ with the discrete residual through a grid-matching argument.

Step 1. For $k = 1, \dots, m_n$, set $\Delta t_{k,n} := t_{k,n} - t_{k-1,n}$ and define, for $s \in (t_{k-1,n}, t_{k,n}]$,

$$\alpha_n(s) := \alpha(t_{k,n}), \qquad \beta_n(s) := \frac{1}{\Delta t_{k,n}} \int_{t_{k-1,n}}^{t_{k,n}} \beta(r)\mathrm{d}r, \qquad \kappa_n(s) := \frac{\beta_n(s)}{\alpha_n(s)}.$$

At $s = 0$, set $\alpha_n(0) := \alpha(0)$, $\beta_n(0) := \beta(0)$, and $\kappa_n(0) := \kappa(0)$; these endpoint values do not affect any integral. Under this definition, we have

$$\gamma_{k,n} = \exp\left\{\frac{\log q_{k,n} - \log q_{k-1,n}}{\alpha_{k,n}}\right\} = \exp\left\{-\frac{1}{\alpha_{k,n}} \int_{t_{k-1,n}}^{t_{k,n}} \beta(r)\mathrm{d}r\right\} = \mathrm{e}^{-\kappa_n \Delta t_{k,n}},$$

$$\sigma_{k,n} = \left(1 - \gamma_{k,n}^{\alpha_{k,n}}\right)^{1/\alpha_{k,n}} = \left(1 - \mathrm{e}^{-\beta_n \Delta t_{k,n}}\right)^{1/\alpha_{k,n}}.$$

Here and below, when the kth grid interval is fixed, β_n and κ_n denote their constant values on that interval.

Let $\widetilde{L}^{(n)}$ be the càdlàg additive process obtained by concatenating independent isotropic $\alpha(t_{k,n})$-stable Lévy processes on the successive grid intervals, each with characteristic exponent $\beta_n \|u\|^{\alpha_{k,n}}$; its increment has characteristic function

$$\mathbb{E}\exp\{\mathrm{i}\langle u, \widetilde{L}_t^{(n)} - \widetilde{L}_s^{(n)}\rangle\} = \exp\left\{-\int_s^t \beta_n(r)\|u\|^{\alpha_n(r)}\mathrm{d}r\right\}, \qquad 0 \le s \le t \le \tau. \tag{27}$$

On $[0, \tau]$, both L and $\widetilde{L}^{(n)}$ have locally finite, absolutely continuous, zero-drift characteristics, by Lemma 11 and the construction above. The Lévy–Itô decomposition for additive processes (Sato, 1999, Theorem 19.2) therefore shows that both processes are semimartingales. Since α and β are continuous on $[0, \tau]$, the piecewise-constant approximations satisfy $\alpha_n \to \alpha$, $\beta_n \to \beta$, and hence $\kappa_n \to \kappa$ uniformly on $[0, \tau]$. For grid points $t_{k,n} \le t_{j,n}$, the corresponding accumulated coefficient satisfies

$$\prod_{\ell=k+1}^{j} \gamma_{\ell,n} = \exp\left\{-\int_{t_{k,n}}^{t_{j,n}} \kappa_n(r)\mathrm{d}r\right\}.$$

It therefore converges to $K(s,t) = \exp\{-\int_s^t \kappa(r)\mathrm{d}r\}$ whenever $t_{k,n} \to s$ and $t_{j,n} \to t$. Thus $K(s,t)$ is the continuous limit of the discrete coefficient $G_{k,j}$. Consequently, for every $R < \infty$,

$$\sup_{\substack{0 \le t \le \tau \\ \|u\| \le R}} \left| \int_0^t \beta_n(s)\|u\|^{\alpha_n(s)}\mathrm{d}s - \int_0^t \beta(s)\|u\|^{\alpha(s)}\mathrm{d}s \right| \longrightarrow 0. \tag{28}$$

The additive process L is stochastically continuous and therefore has no fixed time of discontinuity. Moreover, since $x \mapsto \mathrm{e}^{-x}$ is 1-Lipschitz on $[0, \infty)$, Eq. (28) implies, for every fixed $u \in \mathbb{R}^d$,

$$\sup_{0 \le t \le \tau} \left| \mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, \widetilde{L}_t^{(n)}\rangle} - \mathbb{E}\mathrm{e}^{\mathrm{i}\langle u, L_t\rangle} \right| \longrightarrow 0.$$

This is exactly the characteristic-function criterion for processes with independent increments in Jacod & Shiryaev (2003, Corollary VII.4.43). Hence

$$\widetilde{L}^{(n)} \;\Rightarrow\; L \qquad \text{in } \big(D([0,\tau];\mathbb{R}^d), J_1\big). \tag{29}$$

Step 2. Define $\widetilde{Y}^{(n)}$ as the solution of the linear SDE

$$\mathrm{d}\widetilde{Y}_t^{(n)} = -\kappa_n(t)\,\widetilde{Y}_{t-}^{(n)}\mathrm{d}t + \mathrm{d}\widetilde{L}_t^{(n)}, \qquad \widetilde{Y}_0^{(n)} = 0. \tag{30}$$

For a bounded measurable function c, set $A_c(t) := \int_0^t c(r)\mathrm{d}r$ and define the continuous map $\Phi_c : D([0,\tau];\mathbb{R}^d) \to D([0,\tau];\mathbb{R}^d)$ by

$$\Phi_c(x)(t) := x(t) - \mathrm{e}^{-A_c(t)} \int_0^t \mathrm{e}^{A_c(s)} c(s)x(s)\mathrm{d}s, \quad 0 \le t \le \tau. \tag{31}$$

Integration by parts shows that $\widetilde{Y}^{(n)} = \Phi_{\kappa_n}(\widetilde{L}^{(n)})$ and $Y = \Phi_\kappa(L)$.

We claim that the map $(c, x) \mapsto \Phi_c(x)$ is continuous, when c is equipped with the uniform convergence topology and x with the J_1 topology. Indeed, if $c_n \to c$ uniformly and $x_n \to x$ in J_1, then after applying the time changes implicit in J_1 convergence, the sequence x_n is uniformly bounded on $[0, \tau]$ as a standard consequence of J_1 convergence. Moreover, $x_n(s) \to x(s)$ at every continuity point of x, and the set of discontinuities of x is at most countable, so convergence holds for Lebesgue-a.e. s. Since c_n and x_n are uniformly bounded, the integrands $\mathrm{e}^{A_{c_n}(s)} c_n(s) x_n(s)$ are uniformly bounded. Dominated convergence thus gives, uniformly in t,

$$\int_0^t \mathrm{e}^{A_{c_n}(s)} c_n(s) x_n(s) \mathrm{d}s \to \int_0^t \mathrm{e}^{A_c(s)} c(s) x(s) \mathrm{d}s.$$

Since the second term in (31) converges uniformly to a continuous function of t, and the first term converges in J_1, their sum converges in J_1, that is, $\Phi_{c_n}(x_n) \to \Phi_c(x)$ in J_1. Combining this continuity with (29) and $\kappa_n \to \kappa$ via the continuous mapping theorem yields

$$\widetilde{Y}^{(n)} \Rightarrow Y \qquad \text{in } \big(D([0,\tau]; \mathbb{R}^d), J_1\big). \tag{32}$$

Step 3. By the solution of SDE (30) in Step 2, for $t \in (t_{k-1,n}, t_{k,n}]$ we have

$$\begin{aligned}
\widetilde{Y}^{(n)}_t &= \Phi_{\kappa_n}(\widetilde{L}^{(n)})_t = \widetilde{L}^{(n)}_t - \mathrm{e}^{-A_{\kappa_n}(t)} \int_0^t \mathrm{e}^{A_{\kappa_n}(s)} \kappa_n(s) \widetilde{L}^{(n)}_s \mathrm{d}s \\
&= \widetilde{L}^{(n)}_t - \mathrm{e}^{-A_{\kappa_n}(t)} \Big(\int_0^{t_{k-1,n}} \mathrm{e}^{A_{\kappa_n}(s)} \kappa_n(s) \widetilde{L}^{(n)}_s \mathrm{d}s + \kappa_n \int_{t_{k-1,n}}^t \mathrm{e}^{A_{\kappa_n}(s)} \widetilde{L}^{(n)}_s \mathrm{d}s \Big) \\
&= \widetilde{L}^{(n)}_t - \mathrm{e}^{-A_{\kappa_n}(t)} \Big(\mathrm{e}^{A_{\kappa_n}(t_{k-1,n})} \big(\widetilde{L}^{(n)}_{t_{k-1,n}} - \widetilde{Y}^{(n)}_{t_{k-1,n}} \big) + \kappa_n \int_{t_{k-1,n}}^t \mathrm{e}^{A_{\kappa_n}(s)} \widetilde{L}^{(n)}_s \mathrm{d}s \Big) \\
&= \mathrm{e}^{-\kappa_n (t - t_{k-1,n})} \widetilde{Y}^{(n)}_{t_{k-1,n}} + \widetilde{L}^{(n)}_t - \mathrm{e}^{-\kappa_n (t - t_{k-1,n})} \widetilde{L}^{(n)}_{t_{k-1,n}} - \kappa_n \mathrm{e}^{-\kappa_n t} \int_{t_{k-1,n}}^t \mathrm{e}^{\kappa_n s} \widetilde{L}^{(n)}_s \mathrm{d}s \\
&= \mathrm{e}^{-\kappa_n (t - t_{k-1,n})} \widetilde{Y}^{(n)}_{t_{k-1,n}} + \int_{t_{k-1,n}}^t \mathrm{e}^{-\kappa_n (t-s)} \mathrm{d}\widetilde{L}^{(n)}_s.
\end{aligned}$$

The last step applies the semimartingale product formula (Protter, 2005, Corollary 2 to Theorem II.22); the deterministic exponential factor has continuous finite variation, so its quadratic covariation with $\widetilde{L}^{(n)}$ vanishes. At $t = t_{k,n}$ we obtain $\widetilde{Y}^{(n)}_{t_{k,n}} = \gamma_{k,n} \widetilde{Y}^{(n)}_{t_{k-1,n}} + \eta_{k,n}$, where

$$\eta_{k,n} := \int_{t_{k-1,n}}^{t_{k,n}} \mathrm{e}^{-\kappa_n (t_{k,n} - s)} \mathrm{d}\widetilde{L}^{(n)}_s.$$

The stochastic integral $\eta_{k,n}$ is an isotropic stable random vector whose characteristic function satisfies

$$\begin{aligned}
-\log \mathbb{E} \mathrm{e}^{\mathrm{i}\langle u, \eta_{k,n} \rangle} &= \int_{t_{k-1,n}}^{t_{k,n}} \beta_n \mathrm{e}^{-\alpha(t_{k,n}) \kappa_n (t_{k,n} - s)} \|u\|^{\alpha(t_{k,n})} \mathrm{d}s \\
&= \beta_n \|u\|^{\alpha(t_{k,n})} \int_{t_{k-1,n}}^{t_{k,n}} \mathrm{e}^{-\alpha(t_{k,n}) \kappa_n (t_{k,n} - s)} \mathrm{d}s \\
&= \big(1 - \mathrm{e}^{-\beta_n \Delta t_{k,n}}\big) \|u\|^{\alpha(t_{k,n})} \\
&= \sigma_{k,n}^{\alpha_{k,n}} \|u\|^{\alpha_{k,n}},
\end{aligned}$$

where the first equality uses (27) and Lemma 12 for the corresponding local characteristics of $\widetilde{L}^{(n)}$, with $g(s) = \mathrm{e}^{-\kappa_n (t_{k,n} - s)} u$. The identity remains valid when $\beta_n = 0$, in which case the integral is zero.

Thus $\eta_{k,n} \overset{\mathrm{d}}{=} \xi_{k,n}$ by Eq. (24), and the $\eta_{k,n}$ are independent because they depend on increments of $\widetilde{L}^{(n)}$ over disjoint intervals. Hence the grid values of $\widetilde{Y}^{(n)}$ have the same joint distribution as the discrete residual $Y^{(n)}$.

Step 4. Recall $\rho_n(s) = \max\{t_{j,n} : t_{j,n} \le s\}$ from the definition of the piecewise-constant interpolation of $Y^{(n)}$. Each ρ_n is nondecreasing, and $Y^{(n)} \overset{\mathrm{d}}{=} \widetilde{Y}^{(n)} \circ \rho_n$. Moreover, $\|\rho_n - \mathrm{id}\|_\infty \le |\pi_n| \to 0$, which implies $\rho_n \to \mathrm{id}$ uniformly, and hence also in J_1 (Jacod & Shiryaev, 2003, Proposition VI.1.17(a)). Since $\rho_n \to \mathrm{id}$ deterministically, we have $(\widetilde{Y}^{(n)}, \rho_n) \Rightarrow (Y, \mathrm{id})$ in J_1. As id is continuous and strictly increasing, the composition mapping is J_1-continuous at (Y, id) (Whitt, 2002, Theorem 13.2.2). Hence, by the continuous mapping theorem, $Y^{(n)} \overset{\mathrm{d}}{=} \widetilde{Y}^{(n)} \circ \rho_n \Rightarrow Y$ in J_1.

If $q(T) > 0$, then β is continuous and bounded on the entire interval $[0, T]$, so all preceding arguments apply with $\tau = T$.

Step 5. We convert the limiting residual integral into its differential form. The kernel satisfies $K(s,t) = K(s,r)K(r,t)$; hence, for $r < t$,

$$Y_t = K(r,t)Y_r + \int_r^t K(s,t)\mathrm{d}L_s.$$

To derive the differential form directly, set

$$b(t) := \exp\left\{\int_0^t \kappa(r)\mathrm{d}r\right\}.$$

Then $K(s,t) = b(s)/b(t)$, $Y_t = b(t)^{-1}\int_0^t b(s)\mathrm{d}L_s$. As established in Step 1, L is a semimartingale, and hence so is $\int_0^{\cdot} b(s)\mathrm{d}L_s$. Integration by parts with the continuously differentiable finite-variation factor b^{-1} therefore gives

$$\begin{aligned}
\mathrm{d}Y_t &= \mathrm{d}\left(b(t)^{-1}\int_0^t b(s)\,\mathrm{d}L_s\right) \\
&= (b^{-1})'(t)\left(\int_0^{t-} b(s)\,\mathrm{d}L_s\right)\mathrm{d}t + b(t)^{-1}\,\mathrm{d}\left(\int_0^t b(s)\,\mathrm{d}L_s\right) \\
&= -\kappa(t)b(t)^{-1}\big(b(t)Y_{t-}\big)\mathrm{d}t + b(t)^{-1}b(t)\,\mathrm{d}L_t \\
&= -\kappa(t)Y_{t-}\,\mathrm{d}t + \mathrm{d}L_t.
\end{aligned}$$

Equivalently, $Y_t = L_t - \int_0^t \kappa(r)Y_{r-}\mathrm{d}r$. The integral on the right is continuous in time, so Y has no fixed time of discontinuity because L has none. Thus Eq. (25) also implies the finite-dimensional convergence asserted in the theorem at every deterministic collection of times.

Step 6. In passing from the residual Y_t to the full state $X_t = a(t)X_0 + Y_t$, the deterministic factor $a(t) := q(t)^{1/\alpha(t)}$ generates an additional drift term. Using $\mathrm{d}Y_t = -\kappa(t)Y_{t-}\mathrm{d}t + \mathrm{d}L_t$ from Step 5 gives

$$\begin{aligned}
\mathrm{d}X_t &= a'(t)X_0\mathrm{d}t + \mathrm{d}Y_t \\
&= a'(t)X_0\mathrm{d}t - \kappa(t)Y_{t-}\mathrm{d}t + \mathrm{d}L_t \\
&= a'(t)X_0\mathrm{d}t - \kappa(t)(X_{t-} - a(t)X_0)\mathrm{d}t + \mathrm{d}L_t \\
&= -\kappa(t)X_{t-}\mathrm{d}t + a(t)\left(\kappa(t) + \frac{a'(t)}{a(t)}\right)X_0\mathrm{d}t + \mathrm{d}L_t.
\end{aligned}$$

The term in parentheses can be rewritten by direct differentiation,

$$\frac{a'(t)}{a(t)} = \frac{1}{\alpha(t)}\frac{q'(t)}{q(t)} - \frac{\alpha'(t)}{\alpha(t)^2}\log q(t) = -\kappa(t) - \frac{\alpha'(t)}{\alpha(t)^2}\log q(t).$$

This gives Eq. (20). This extra drift is necessary: since $a'(t)/a(t) = -\kappa(t) - (\alpha'(t)/\alpha(t)^2)\log q(t)$, discarding the $\log q$-term would be equivalent to assuming $a'(t)/a(t) = -\kappa(t)$, which contradicts the prescribed $a(t) = q(t)^{1/\alpha(t)}$ unless $\alpha'(t)\log q(t) \equiv 0$. The algebraic identity in Lemma 6 is its exact discrete counterpart.

□

The following corollary is the continuous-time analogue of Theorem 2, replacing the finite sum over active increments by an integral over the continuous schedule.

Corollary 14. *Under the conditions of Theorem 13, fix $t < T$ and $\rho > 0$ such that $\Psi_t^{\mathrm{ct}}(\rho) > 0$, and define*

$$\omega_t(s) := \beta(s)K(s,t)^{\alpha(s)}, \qquad p_{t,\rho}(s) := \frac{\omega_t(s)\rho^{\alpha(s)}}{\Psi_t^{\mathrm{ct}}(\rho)}, \qquad 0 \le s \le t.$$

Then

$$\frac{\mathrm{d}\log\Psi_t^{\mathrm{ct}}(\rho)}{\mathrm{d}\log\rho} = \int_0^t \alpha(s)p_{t,\rho}(s)\mathrm{d}s,$$
$$\frac{\mathrm{d}^2\log\Psi_t^{\mathrm{ct}}(\rho)}{\mathrm{d}(\log\rho)^2} = \int_0^t \left(\alpha(s) - \int_0^t \alpha(v)p_{t,\rho}(v)\mathrm{d}v\right)^2 p_{t,\rho}(s)\mathrm{d}s \ge 0.$$

The inequality is strict exactly when α is not constant $\omega_t(s)\mathrm{d}s$-almost everywhere.

Proof. For $\rho > 0$, whenever $\Psi_t^{\mathrm{ct}}(\rho) > 0$, $p_{t,\rho}$ is a probability density on $[0,t]$ (by definition of Ψ_t^{ct} in (26)). Write $\ell := \log\rho$. On every compact ℓ-interval, the integrand $\omega_t(s)\mathrm{e}^{\alpha(s)\ell}$ and its first two ℓ-derivatives are bounded by a fixed constant times $\omega_t(s)$, because $\alpha(s) \in [\alpha_-, \alpha_+]$. Since $\int_0^t \omega_t(s)\mathrm{d}s \le \int_0^t \beta(s)\mathrm{d}s < \infty$, dominated convergence permits two differentiations under the integral. Set

$$M_j(\ell) := \int_0^t \alpha(s)^j \omega_t(s)\mathrm{e}^{\alpha(s)\ell}\mathrm{d}s, \qquad j \in \{0,1,2\}.$$

Then $M_0(\ell) = \Psi_t^{\mathrm{ct}}(\mathrm{e}^\ell)$, $M_0'(\ell) = M_1(\ell)$, and $M_1'(\ell) = M_2(\ell)$. Therefore

$$\frac{\mathrm{d}\log\Psi_t^{\mathrm{ct}}(\mathrm{e}^\ell)}{\mathrm{d}\ell} = \frac{M_1(\ell)}{M_0(\ell)}, \qquad \frac{\mathrm{d}^2\log\Psi_t^{\mathrm{ct}}(\mathrm{e}^\ell)}{\mathrm{d}\ell^2} = \frac{M_2(\ell)}{M_0(\ell)} - \left(\frac{M_1(\ell)}{M_0(\ell)}\right)^2.$$

This is exactly the formula in the corollary. Because $\mathrm{e}^{\alpha(s)\ell} > 0$, the measures $p_{t,\mathrm{e}^\ell}(s)\mathrm{d}s$ and $\omega_t(s)\mathrm{d}s$ have the same null sets. The variance is therefore strictly positive precisely when $\alpha(s)$ is not constant $\omega_t(s)\mathrm{d}s$-almost everywhere. □

Figure 5 illustrates a finite-dimensional consequence of Theorem 13. The deterministic exponent calculations show the effect of grid refinement, while the simulated projection visualizes the corresponding limit of the joint distribution. Here the maximum relative discrepancy decays approximately linearly with mesh width, while the deviations in (a) shrink toward zero and the empirical curves in (c) closely track the limit for $m_n = 8$ and 32.

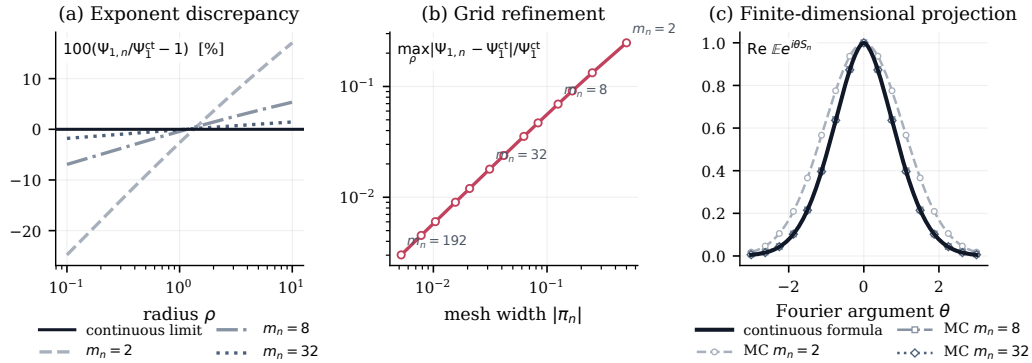


Figure 5: Numerical illustration of Theorem 13. We use $\alpha(t) = 1.5 + 0.25(1 - \cos(\pi t))$ and $q(t) = \mathrm{e}^{-2t}$, and inspect equal partitions of $[0,1]$; m_n is the number of subintervals. (a) shows the signed percentage discrepancy between the discrete and limiting marginal exponents. (b) reports the maximum relative discrepancy over 240 logarithmically spaced radii in $[10^{-1}, 10]$. (c) compares the limiting characteristic function with empirical estimates for the projection $S_n = Y_{0.25}^{(n)} - 0.7Y_{0.5}^{(n)} + 0.8Y_1^{(n)}$, using 100,000 independent draws for each n.

Scope of the theoretical results Several qualifications on the scope of the theory merit explicit notice. After marginalizing X_0, the forward chain is generally non-Markovian, whereas conditioning on X_0 gives the one-step recursion in Eq. (7). In the plug-in reverse update, the learned endpoint estimate replaces the unknown X_0, so the conditional mean is approximate while the conditional covariance remains exact. The continuous-time limit is established in the J_1 Skorokhod topology on every $[0, \tau]$ with $\tau < T$; for the idealized continuous cosine signal schedule, $q(T) = 0$ and $\beta(t)$ becomes singular as $t \uparrow T$, so the endpoint extension stated in Theorem 13 does not apply. For numerical stability, the image implementation clips positive stable scales. This does not affect the conditional Gaussian construction: Propositions 5 and 7 remain valid, while the distribution-level results in Lemmas 4 and 3, Theorem 1, Proposition 9, and Theorem 13 apply only to the untruncated scale A_t. Appendix H.6 examines sensitivity to the clip threshold. We do not provide a formal convergence bound to the data distribution; such guarantees remain an active research area even for Gaussian diffusion (Li & Yan, 2025; Jiao et al., 2025), and their extension to mixed-power settings is beyond the present scope. Finally, the half-cosine schedule is not shown to be optimal among all possible tail-index paths; it is pre-specified based on smoothness and endpoint conditions, with robustness examined in Appendices H.3 and H.5.

E Training and Sampling Algorithms

For a diffusion chain of length N, the implementation retains the cosine $\{\beta_t\}$ sequence that generates the variance-preserving stepwise coefficients of DLPM (Nichol & Dhariwal, 2021). Its normalized cumulative power is

$$q_t := \frac{\cos^2\left(\frac{t/N+\eta}{1+\eta}\frac{\pi}{2}\right)}{\cos^2\left(\frac{\eta}{1+\eta}\frac{\pi}{2}\right)}, \qquad \eta := 0.008, \qquad t = 0, \dots, N-1.$$

Set $\beta_0 := 0$ and $\beta_t := 1 - q_t/q_{t-1}$ for $t \geq 1$. Then $q_t = \prod_{j=1}^{t}(1-\beta_j)$, and it satisfies $q_0 = 1$ and $q_{N-1} > 0$.

The 1.5–2.0 endpoints in Eq. (3) are a pre-specified, dataset-independent design choice, not a requirement of the construction. The half-cosine interpolation has zero slope at both endpoints and therefore changes smoothly near the boundaries. Other choices of the annealing schedule for the tail index and its endpoints are evaluated as sensitivity analyses in Appendices H.3 and H.5.

Algorithm 1 A-DLPM epsilon training

Require: data distribution p_{data}; original chain length N; batch size B; data dimension D.
Schedules $\{\alpha_t, \bar{\gamma}_t, \bar{\sigma}_t, \gamma_t, \sigma_t\}_{t=0}^{N-1}$; network ϵ_θ.
repeat
Sample a mini-batch $\{X_0^{(i)}\}_{i=1}^{B} \sim p_{\text{data}}$ and $t_i \sim \text{Unif}\{1, \dots, N-1\}$.
for $i = 1, \dots, B$ **do**
Sample independent $A_j^{(i)} \sim \mathcal{A}_{\alpha_j/2}$ for $j = 0, \dots, t_i$ and $Z^{(i)} \sim \mathcal{N}(0, I_D)$.
Set $\Sigma_0^{(i)} := \sigma_0^2 A_0^{(i)}$.
for $j = 1, \dots, t_i$ **do**
$\Sigma_j^{(i)} := \gamma_j^2 \Sigma_{j-1}^{(i)} + \sigma_j^2 A_j^{(i)}$. (by Eq. (6))
end for
$X_{t_i}^{(i)} := \bar{\gamma}_{t_i} X_0^{(i)} + \sqrt{\Sigma_{t_i}^{(i)}} Z^{(i)}$. (by Eq. (8))
$\epsilon_{t_i}^{(i)} := (\sqrt{\Sigma_{t_i}^{(i)}}/\bar{\sigma}_{t_i}) Z^{(i)}$. (by Eq. (9))
end for
$\mathcal{L} := B^{-1} \sum_{i=1}^{B} \left(D^{-1} \|\epsilon_\theta(X_{t_i}^{(i)}, t_i) - \epsilon_{t_i}^{(i)}\|_2^2 \right)^{1/2}$. (by Eq. (10))
Update θ to minimize $\mathcal{L}$.
until the training budget is exhausted
Return: θ.

Algorithm 1 samples only the scales $A_0, \dots, A_t$ up to the chosen time step t. Given these scales, a single aggregate Gaussian draw is sufficient to sample the exact marginal in Eq. (8). The network

never observes the scales $A_{0:t}$, the labels, or the class frequencies. Feeding the scale A_t to the network as an additional input leads to a clear degradation in performance, which is also reported by Shariatian et al. (2025, Appendix G).

In practice, for $\alpha < 2$, a draw from $\mathcal{A}_{\alpha/2}$ is generated using the Chambers–Mallows–Stuck (CMS) algorithm (Chambers et al., 1976). Sampling one such scale costs $O(1)$. Drawing all $A_0, \dots, A_t$ and computing Σ_t costs $O(N)$ per training sample for a chain of length N and requires no additional network evaluations.

Algorithm 2 Respaced reverse sampling for A-DLPM

Require: trained ϵ_θ; original chain length N; data dimension D.
Schedules $\{\alpha_t, \bar{\gamma}_t, \bar{\sigma}_t, \gamma_t, \sigma_t\}_{t=0}^{N-1}$.
Retained indices from the original chain $0 = \tau_0 < \tau_1 < \cdots < \tau_K = N-1$.
Sample independent $A_t \sim \mathcal{A}_{\alpha_t/2}$ for $t = 0, \dots, N-1$.
Set $\Sigma_0 := \sigma_0^2 A_0$ and, for $t = 1, \dots, N-1$, set $\Sigma_t := \gamma_t^2 \Sigma_{t-1} + \sigma_t^2 A_t$.
Sample $Z \sim \mathcal{N}(0, I_D)$ and initialize $X_{\tau_K} := \sqrt{\Sigma_{\tau_K}} Z$.
for $k = K, K-1, \dots, 1$ **do**
  Set $t := \tau_k$ and $s := \tau_{k-1}$.
  $\widehat{\epsilon} := \epsilon_\theta(X_t, t), \quad \widehat{X}_0 := (X_t - \bar{\sigma}_t \widehat{\epsilon})/\bar{\gamma}_t.$ (by Eq. (11))
  if $s = 0$ **then**
    $X_s := \widehat{X}_0$.
  else
    $G_{s,t} := \prod_{j=s+1}^{t} \gamma_j$. (by Eq. (13))
    $a_{s,t} := G_{s,t}\Sigma_s/\Sigma_t, \quad b_{s,t} := \bar{\gamma}_s - a_{s,t}\bar{\gamma}_t$. (by Eq. (14))
    $\nu_{s,t} := \Sigma_s - G_{s,t}^2 \Sigma_s^2/\Sigma_t$. (by Eq. (15))
    Sample $\xi \sim \mathcal{N}(0, I_D)$ and set $X_s := a_{s,t} X_t + b_{s,t}\widehat{X}_0 + \sqrt{\nu_{s,t}}\xi$. (by Eq. (12))
  end if
end for
Return: X_0.

Algorithm 2 makes the respacing rule explicit. Retained indices $\{\tau_k\}_{k=0}^{K}$ are a subset of the original training chain, thus the conditional Gaussian law in Eq. (12) remains exact for retained pairs, conditional on X_0 and $A_{0:t}$. In our image experiments, we retain $K + 1 = 1000$ indices from the original $N = 4000$ chain by setting $\tau_k = \mathrm{round}(k(N-1)/K)$ for $k = 0, \dots, K$, where round denotes rounding to the nearest integer. The loop therefore makes $K = 999$ network evaluations.

The zero-mean initialization in Algorithm 2 omits the small forward signal $\bar{\gamma}_{N-1} X_0$ in Proposition 5, as in standard diffusion prior initialization.

F Reverse-Time Prediction-Error Allocation

This section examines whether the learned prediction errors follow the scheduled tail index. The analysis concerns their allocation over time rather than a uniform reduction in the local errors of A-DLPM. The analysis of diffusion with a fixed tail index by Cherkaoui et al. (2026, Theorem 2.2) bounds the final sampling error by a terminal initialization term and approximation and estimation terms accumulated along the reverse trajectory. Although that theorem does not cover our construction with a variable tail index after marginalizing X_0, it suggests a focused empirical question: does annealing behave as if the tail index were fixed throughout generation, or does the error behavior shift with the time-varying tail index?

Let $m \in \{1.5, 1.8, 2.0, \mathrm{ann}\}$, where $1.5, 1.8, 2.0$ index DLPM baselines and $m = \mathrm{ann}$ denotes A-DLPM. Let $u \in \{0, \dots, 998\}$ index the retained noise levels in reverse order, from the terminal state $\tau_K = N - 1$ down to the noise level τ_1 closest to the data; the data endpoint $\tau_0 = 0$ is excluded. Let c index pairs of a held-out image and a sampled scale path. The empirical mean $\mathbb{E}_c$ averages over the four paths per image and then over images. Let $t_m(u)$ denote the time index in the original chain of method m that corresponds to reverse ordinal u. Let d denote the image dimension, $\epsilon_{\theta_m}(X_{m,c}(u), t_m(u))$ denote the model prediction, and $\epsilon_{m,c}(u)$ denote its matched forward-noise

target. Their per-coordinate prediction-error and target energies are

$$E_{m,c}(u) := \frac{\|\epsilon_{\theta_m}(X_{m,c}(u), t_m(u)) - \epsilon_{m,c}(u)\|_2^2}{d}, \qquad T_{m,c}(u) := \frac{\|\epsilon_{m,c}(u)\|_2^2}{d}.$$

The raw error curve is computed using the same aggregation as the implemented epsilon-RMSE loss,

$$L_m(u) := \mathbb{E}_c \sqrt{E_{m,c}(u)}.$$

Figure 6 displays this quantity in the left column. Since $T_{m,c}(u)$ varies with the stable tail index, these raw curves reflect absolute prediction-error magnitude but are not directly comparable across different tail indices. Formal comparisons therefore use the log energy ratio $R_{m,c}(u)$ and the affinity statistic $D_c(u)$ defined below.

$$R_{m,c}(u) := \frac{1}{2} \log \frac{E_{m,c}(u) + 10^{-12}}{T_{m,c}(u) + 10^{-12}},$$
$$D_c(u) := |R_{\mathrm{ann},c}(u) - R_{1.5,c}(u)| - |R_{\mathrm{ann},c}(u) - R_{2.0,c}(u)|.$$

Thus $D_c(u) > 0$ means that A-DLPM is closer to DLPM with $\alpha = 2.0$, while $D_c(u) < 0$ means that it is closer to DLPM with $\alpha = 1.5$. DLPM with $\alpha = 1.8$ is shown as an intermediate baseline but is not used in the affinity statistic.

For each dataset, we use 256 held-out images (positions 64–319 of the balanced test loader) and four path seeds. All methods use a single checkpoint from one training run, at 75,000 updates for CIFAR-10-LT and 74,883 updates for CIFAR-100-LT (EMA 0.9999). DLPM is evaluated on its inherited 1000-step rescaled path, while A-DLPM uses 1000 retained states from its original 4000-step path. The 999 noisy states are partitioned into equal early, middle, and late thirds before evaluation. For each pair c, the averages of the affinity statistic over the early and late thirds are

$$\overline{D}_{\mathrm{E},c} := \frac{1}{333} \sum_{u=0}^{332} D_c(u), \qquad \overline{D}_{\mathrm{L},c} := \frac{1}{333} \sum_{u=666}^{998} D_c(u).$$

We test the predictions that early in generation A-DLPM is closer to DLPM with $\alpha = 2.0$ (i.e., $\mathbb{E}_c \overline{D}_{\mathrm{E},c} > 0$), that late in generation it is closer to DLPM with $\alpha = 1.5$ (i.e., $\mathbb{E}_c \overline{D}_{\mathrm{L},c} < 0$), and that the affinity shifts from 2.0-like to 1.5-like over reverse time (i.e., $\mathbb{E}_c(\overline{D}_{\mathrm{L},c} - \overline{D}_{\mathrm{E},c}) < 0$). Accordingly, we report the following summary statistics:

$$S_{\mathrm{E}} := \mathbb{E}_c \overline{D}_{\mathrm{E},c}, \quad S_{\mathrm{L}} := \mathbb{E}_c \overline{D}_{\mathrm{L},c}, \quad S_{\mathrm{shift}} := \mathbb{E}_c\big(\overline{D}_{\mathrm{L},c} - \overline{D}_{\mathrm{E},c}\big).$$

Table 2 reports these statistics. Every confidence interval is based on 20,000 hierarchical bootstrap replicates. Each replicate samples 256 images with replacement and one of the four paths uniformly for each selected image, then averages the resulting values.

Table 2: Reverse-time affinity tests. The predicted pattern is $S_{\mathrm{E}} > 0$, $S_{\mathrm{L}} < 0$, and $S_{\mathrm{shift}} < 0$.

Dataset	Criterion	Estimate	95% bootstrap interval
CIFAR-10-LT	Early DLPM-2.0 affinity (S_{E})	0.364736	$[0.357055, 0.372293]$
	Late DLPM-1.5 affinity (S_{L})	-0.129921	$[-0.133089, -0.126846]$
	Temporal shift (S_{shift})	-0.494657	$[-0.502234, -0.486943]$
CIFAR-100-LT	Early DLPM-2.0 affinity (S_{E})	0.361721	$[0.354192, 0.369174]$
	Late DLPM-1.5 affinity (S_{L})	-0.098797	$[-0.103756, -0.093607]$
	Temporal shift (S_{shift})	-0.460518	$[-0.470040, -0.450777]$

The left column of Figure 6 reports the raw absolute prediction-error magnitude. The raw error of A-DLPM lies below all three DLPM baselines at every retained middle state, and at 97.6% of retained late states on both datasets. The right panels show that A-DLPM is closer to DLPM with $\alpha = 2.0$ in the high-noise third (early in reverse time) and closer to DLPM with $\alpha = 1.5$ near the data end (late in reverse time). This shift is confirmed by the fact that all six bootstrap intervals in Table 2

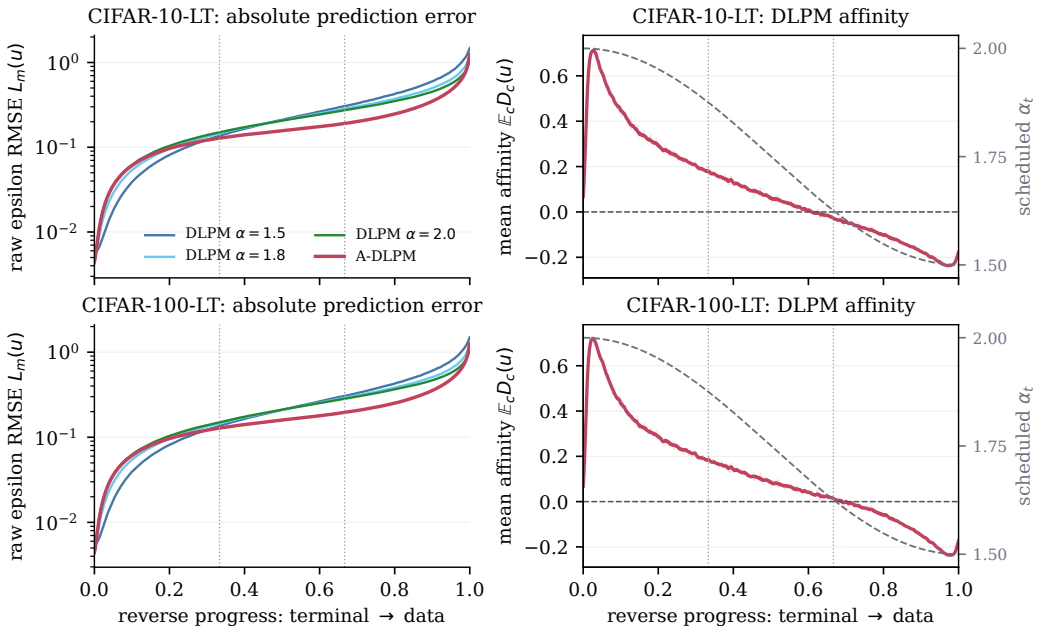


Figure 6: **Absolute prediction error and temporal allocation.** Left: raw held-out epsilon RMSE $L_m(u)$ for DLPM with $\alpha \in \{1.5, 1.8, 2.0\}$ and A-DLPM. Right: mean affinity $\mathbb{E}_c D_c(u)$; positive values indicate similarity to DLPM with $\alpha = 2.0$, negative to DLPM with $\alpha = 1.5$. Gray dashed: scheduled α. Vertical dotted lines delimit the regions used to compute S_{E}, S_{L}, and S_{shift} in Table 2.

exclude zero. Moreover, the two affinity curves have Pearson correlation 0.99785, indicating that the same temporal transition occurs consistently across the two datasets. Thus the learned prediction-error behavior follows the scheduled change in tail index rather than remaining tied to one fixed tail index.

These results support a temporal reallocation rather than a uniform reduction of local errors. A fixed tail index ties the noise family to a single reverse estimation regime, whereas the target-normalized prediction errors of A-DLPM resemble DLPM with $\alpha = 2.0$ early in reverse time and DLPM with $\alpha = 1.5$ near the data. Together with the mixed-power spectrum in Theorem 1 and the FID–Recall comparison in Table 1, this suggests temporal reallocation as an explanation for the improved fidelity–diversity trade-off.

G Experimental Details

G.1 Two-dimensional setup and paired analysis

All trained configurations use 20 paired training seeds, and each run uses batch size 1024 for 313 epochs (10,016 optimizer updates) and is evaluated once with the generation seed equal to the training seed. Gaussian DDPM and LIM use the same data, network, and update budget as A-DLPM and DLPM. DLIM uses deterministic sampling with $\eta = 0$ in the notation of Shariatian et al. (2025) from each corresponding DLPM checkpoint, and therefore requires no additional training. Both metrics are lower-is-better and are computed as follows.

Tail MSLE. Let $n = 25000$ be the number of generated samples. Let $\widehat{Q}_R$ and $\widehat{Q}_G$ denote the empirical quantile functions of the first coordinates of the real and generated samples, respectively, and define the evaluation grid

$$u_j := 0.95 + \frac{j}{511}\left(1 - \frac{1}{n} - 0.95\right), \qquad j = 0, \ldots, 511.$$

For every grid point at which both quantiles are positive, set $e_j := \left(\log \widehat{Q}_R(u_j) - \log \widehat{Q}_G(u_j)\right)^2$. If $j_1 < \cdots < j_M$ enumerate those valid grid points, the Tail MSLE is the normalized trapezoidal integral

$$\mathrm{TailMSLE}_{0.95} := \frac{1}{u_{j_M} - u_{j_1}} \sum_{r=1}^{M-1} \frac{e_{j_r} + e_{j_{r+1}}}{2} \left(u_{j_{r+1}} - u_{j_r}\right).$$

Wasserstein. For each method, we flatten all scalar coordinates of the generated and real samples and accumulate them into normalized histograms $(a_i)_{i=1}^B$ and $(b_i)_{i=1}^B$ on the same $B := 250$ bins. Both histograms are normalized so that $\sum_{i=1}^B a_i = \sum_{i=1}^B b_i = 1$. If c_i is the center of bin i, the empirical Wasserstein-1 distance is

$$\widehat{W}_1 := \min_{P_{ij} \geq 0} \sum_{i=1}^B \sum_{j=1}^B P_{ij} |c_i - c_j| \quad \text{subject to} \quad \sum_{j=1}^B P_{ij} = a_i, \quad \sum_{i=1}^B P_{ij} = b_j.$$

Figure 2 displays the sample means with 95% percentile bootstrap intervals obtained from 20,000 resamples of the 20 paired seed indices. Resampling the common indices preserves the paired experimental design. Table 3 provides the numerical values underlying Figure 2. A-DLPM attains the lowest mean Tail $\mathrm{MSLE}_{0.95}$ among all evaluated configurations. The DLPM values form a U-shaped curve with a minimum near $\alpha = 1.6$, while no LIM or DLIM baseline has a lower mean than A-DLPM. The Wasserstein score of A-DLPM remains close to the best mean among all evaluated configurations.

Table 3: **Two-dimensional stable-tail task.** Means $\pm$ sample standard deviations over 20 paired training seeds. Columns give the tail index for DLPM, DLIM, and LIM; Gaussian DDPM and A-DLPM do not use a tail-index column. Boldface marks the best mean in each metric.

Method	$\alpha = 1.5$	$\alpha = 1.6$	$\alpha = 1.7$	$\alpha = 1.8$	$\alpha = 1.9$	$\alpha = 2.0$
Tail MSLE$_{0.95}$ ↓						
DLPM	0.0692 ± 0.0671	0.0412 ± 0.0437	0.0738 ± 0.0798	0.0751 ± 0.0927	0.1048 ± 0.1159	0.1819 ± 0.1207
DLIM	0.1521 ± 0.1419	0.0952 ± 0.1431	0.0932 ± 0.0833	0.0768 ± 0.0863	0.1046 ± 0.1124	0.1411 ± 0.1096
LIM	0.5198 ± 0.1893	0.3607 ± 0.2733	0.2389 ± 0.1867	0.1150 ± 0.0586	0.5939 ± 1.9574	2.8881 ± 3.0335
DDPM			0.0641 ± 0.0517			
A-DLPM			$\mathbf{0.0391} \pm 0.0382$			
Wasserstein ↓						
DLPM	0.0245 ± 0.0243	$\mathbf{0.0180} \pm 0.0242$	0.0191 ± 0.0239	0.0206 ± 0.0228	0.0222 ± 0.0224	0.0261 ± 0.0217
DLIM	0.0320 ± 0.0237	0.0245 ± 0.0237	0.0203 ± 0.0235	0.0215 ± 0.0233	0.0214 ± 0.0226	0.0246 ± 0.0224
LIM	0.0665 ± 0.0331	0.0563 ± 0.0286	0.0467 ± 0.0300	0.0335 ± 0.0250	0.0443 ± 0.0557	0.2065 ± 0.1490
DDPM			0.0193 ± 0.0230			
A-DLPM			0.0190 ± 0.0250			

G.2 Long-tailed image experiments

We construct the standard CIFAR-10-LT and CIFAR-100-LT training sets by exponentially subsampling the original CIFAR-10 and CIFAR-100 training splits (Krizhevsky, 2009; Cui et al., 2019). The original balanced test splits are left unchanged.

The CIFAR-10-LT training set uses the class counts $[5000, 2997, 1796, 1077, 645, 387, 232, 139, 83, 50]$, which yield 125 optimizer updates per epoch, placing epoch 600 exactly at 75,000 updates. The CIFAR-100-LT training set uses a monotone exponential class-count schedule from 500 images in the most frequent class to 5 images in the rarest, totaling 10,847 training images. With batch size 100, CIFAR-100-LT has 109 updates per epoch, so epochs 458, 687, and 916 correspond to 49,922, 74,883, and 99,844 updates, respectively. Both datasets use 32×32 images, random horizontal flips during training, and the improved-DDPM U-Net (Nichol & Dhariwal, 2021). The main CIFAR-10-LT checkpoint is taken at 75,000 updates. The main CIFAR-100-LT checkpoint is the 687-epoch checkpoint at 74,883 updates.

The optimizer is AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01. The learning rate is 2×10^{-4}, and StepLR decays the learning rate by 0.99 every 1000 optimizer updates. We use EMA 0.9999 and global gradient clipping at 5. Every model is trained

unconditionally, without class labels, class rebalancing, an auxiliary loss, or an additional network. DLPM, Gaussian DDPM, and A-DLPM use 4000-step training chains. LIM is trained in continuous time, with training times sampled from a continuous interval rather than from a fixed discrete chain.

DLPM and LIM are trained separately at each $\alpha \in \{1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$ with the same five paired training seeds and update budget as A-DLPM and Gaussian DDPM. DLIM applies its deterministic update to each corresponding DLPM checkpoint and therefore requires no additional training. Each checkpoint is evaluated with the same five generation seeds. Metrics are first averaged over those five generation seeds within a training run; Table 1 reports the mean over the five training runs.

Every stable-noise image configuration clips the positive mixing scale at 20 during both training and sampling for numerical stability; Appendix H.6 examines this implementation choice.

During evaluation, every method performs 999 network evaluations over 1000 states. A-DLPM retains 1000 indices from its original 4000-step chain and uses the corresponding conditional reverse updates for arbitrary pairs of retained indices. DLPM and DLIM use the inherited rescaling to a 1000-step chain, while LIM uses its stochastic reverse update on the same number of states.

The CIFAR-10-LT reference set contains $[2015, 1208, 724, 434, 260, 156, 94, 56, 33, 20]$ images drawn from the long-tailed training pool. For CIFAR-100-LT, the reference is constructed by the same deterministic proportional allocation from its training pool, so that the class proportions match the training distribution.

FID uses 2048-dimensional features from an ImageNet-pretrained Inception-v3 network (Heusel et al., 2017). Precision, Recall, Density, and Coverage are computed using PRDC with $k = 5$ (Naeem et al., 2020). Each overall FID and PRDC comparison uses 5000 real and 5000 generated images.

The complete balanced 10,000-image test set for each dataset is reserved for grouped Recall and Coverage: it supplies the real queries and their class labels, while each comparison uses the same 5000 generated images. The $k = 5$ neighborhoods are computed globally before the real queries are partitioned by decreasing training frequency into head (first third of classes), medium (second third), and tail (remaining classes).

H Additional Experimental Results

H.1 Training trajectories

Figure 7 traces mean FID and Recall for one training run per dataset across the three archived checkpoints, evaluating each checkpoint with five generation runs. Table 4 gives the corresponding numerical values. Together they show why the shared middle checkpoint is treated as one operating point rather than as a per-dataset optimum.

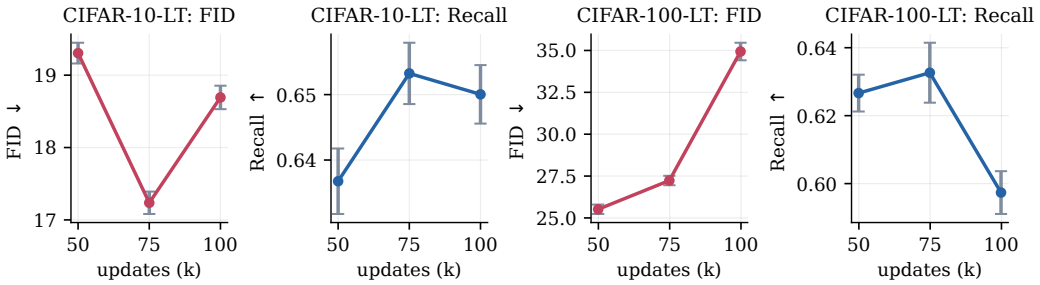


Figure 7: FID and Recall trajectories for A-DLPM with the standard half-cosine schedule, using one training run per dataset. Points and error bars show means and sample standard deviations over five generation runs, with 5,000 images per run. The horizontal axis gives training updates in thousands. The shared main checkpoint is the middle point, not the per-metric optimum.

We retain the pre-specified common window of approximately 75k updates (75,000 and 74,883, respectively), rather than selecting a different checkpoint for each dataset. Mean Recall is highest at this middle checkpoint on both datasets; FID is lowest there on CIFAR-10-LT and at the earlier checkpoint on CIFAR-100-LT.

Table 4: Trajectory metrics from the training runs used in Figure 7. Entries are means ± sample standard deviations over five generation runs.

Dataset	Updates	FID↓	Recall↑
CIFAR-10-LT	50k	19.3049 ± 0.1432	0.6368 ± 0.0050
	75k	17.2365 ± 0.1546	0.6532 ± 0.0047
	100k	18.6925 ± 0.1620	0.6500 ± 0.0045
CIFAR-100-LT	49,922	25.5169 ± 0.2809	0.6266 ± 0.0054
	74,883	27.2338 ± 0.2802	0.6326 ± 0.0088
	99,844	34.9371 ± 0.5236	0.5974 ± 0.0063

H.2 Paired uncertainty for the main image comparison

Table 5 reports the paired differences for the main image results presented in Table 1 and Figure 3. Within each training run, metrics are first averaged over the five generation seeds. For each of the five paired training runs, we define $\Delta\mathrm{FID} := \mathrm{FID}_{\mathrm{baseline}} - \mathrm{FID}_{\text{A-DLPM}}$ and $\Delta\mathrm{Recall} := \mathrm{Recall}_{\text{A-DLPM}} - \mathrm{Recall}_{\mathrm{baseline}}$, so positive values favor A-DLPM in both columns. The intervals are pointwise 95% Student-t intervals over the five paired differences. They are descriptive and not adjusted for multiple comparisons.

We point out that, for CIFAR-10-LT against DLPM with $\alpha = 1.9$, all five paired FID differences are positive, but one high-FID baseline training run yields a much larger positive difference, inflating the sample variance and widening the interval across zero. More broadly, 46 of the 48 pointwise FID and Recall intervals against DLPM and DLIM lie entirely above zero across the two datasets, supporting A-DLPM's FID–Recall advantage over these baselines.

Table 5: Paired differences for all FID–Recall comparisons in Table 1. Entries are mean differences with pointwise 95% Student-t intervals across five paired training runs.

	CIFAR-10-LT		CIFAR-100-LT	
Baseline	ΔFID	ΔRecall	ΔFID	ΔRecall
DLPM $\alpha = 1.5$	6.152 [5.780, 6.523]	0.0267 [0.0209, 0.0325]	4.002 [2.877, 5.126]	0.0209 [0.0172, 0.0246]
DLPM $\alpha = 1.6$	3.898 [3.465, 4.331]	0.0220 [0.0187, 0.0254]	3.676 [1.542, 5.811]	0.0196 [0.0111, 0.0282]
DLPM $\alpha = 1.7$	2.788 [2.211, 3.364]	0.0199 [0.0156, 0.0241]	2.205 [1.482, 2.928]	0.0242 [0.0152, 0.0331]
DLPM $\alpha = 1.8$	1.861 [1.614, 2.109]	0.0163 [0.0123, 0.0204]	2.154 [1.254, 3.055]	0.0295 [0.0221, 0.0370]
DLPM $\alpha = 1.9$	4.988 [−4.625, 14.601]	0.0167 [0.0135, 0.0200]	3.830 [2.860, 4.800]	0.0341 [0.0240, 0.0442]
DLPM $\alpha = 2.0$	1.696 [1.404, 1.989]	0.0168 [0.0113, 0.0222]	5.144 [3.809, 6.480]	0.0330 [0.0256, 0.0404]
DLIM $\alpha = 1.5$	15.203 [14.461, 15.945]	0.0094 [−0.0016, 0.0203]	10.741 [9.880, 11.602]	0.0210 [0.0122, 0.0297]
DLIM $\alpha = 1.6$	12.049 [11.453, 12.645]	0.0104 [0.0069, 0.0138]	9.318 [8.286, 10.350]	0.0292 [0.0207, 0.0377]
DLIM $\alpha = 1.7$	9.143 [8.266, 10.021]	0.0113 [0.0063, 0.0162]	7.181 [6.143, 8.219]	0.0386 [0.0233, 0.0538]
DLIM $\alpha = 1.8$	7.037 [6.444, 7.630]	0.0142 [0.0095, 0.0190]	5.952 [4.373, 7.531]	0.0429 [0.0355, 0.0504]
DLIM $\alpha = 1.9$	5.931 [4.886, 6.975]	0.0138 [0.0048, 0.0229]	7.139 [5.420, 8.857]	0.0546 [0.0490, 0.0602]
DLIM $\alpha = 2.0$	3.459 [3.085, 3.833]	0.0090 [0.0016, 0.0164]	6.750 [5.558, 7.942]	0.0460 [0.0401, 0.0518]
LIM $\alpha = 1.5$	5.957 [5.206, 6.708]	0.0103 [0.0060, 0.0146]	2.232 [0.128, 4.337]	0.0167 [0.0105, 0.0229]
LIM $\alpha = 1.6$	4.606 [3.587, 5.626]	0.0168 [0.0081, 0.0255]	0.747 [−0.650, 2.145]	0.0159 [0.0076, 0.0242]
LIM $\alpha = 1.7$	3.884 [3.238, 4.529]	0.0095 [0.0081, 0.0109]	−0.272 [−1.725, 1.181]	0.0087 [0.0020, 0.0154]
LIM $\alpha = 1.8$	2.641 [2.344, 2.937]	0.0039 [−0.0025, 0.0103]	−1.698 [−2.421, −0.975]	0.0065 [−0.0011, 0.0141]
LIM $\alpha = 1.9$	3.945 [1.230, 6.661]	0.0168 [0.0025, 0.0311]	0.918 [0.470, 1.365]	0.0175 [0.0071, 0.0279]
LIM $\alpha = 2.0$	2.866 [1.954, 3.777]	0.0065 [−0.0014, 0.0145]	8.273 [5.372, 11.173]	0.0410 [0.0307, 0.0513]
Gaussian DDPM	1.279 [0.812, 1.745]	0.0081 [0.0030, 0.0132]	4.979 [2.562, 7.396]	0.0306 [0.0269, 0.0343]

H.3 Robustness to curvature and path shape

Table 6 compares the standard half-cosine path in Eq. (3) with nearby curvature perturbations and a linear path under the same training budget on both image benchmarks. The purpose is to test whether the reported operating point depends on one fragile interpolation choice. The curvature

family is defined by $\alpha_t(p) := 1.5 + 0.5\left(\frac{1-\cos(\pi t/(N-1))}{2}\right)^p$ for $p > 0$, and the original half-cosine path of tail indices has $p = 1$. The linear path is $\alpha_t = 1.5 + 0.5t/(N-1)$.

Table 6: Curvature sensitivity at 75,000 updates for CIFAR-10-LT and 74,883 updates for CIFAR-100-LT. Each path uses one training run and five generation runs; entries are the mean ± sample standard deviation across generation runs. The core method is the standard half-cosine row in each dataset block.

Path	FID↓	Precision↑	Recall↑	Density↑	Coverage↑
CIFAR-10-LT					
Half-cosine $p = 0.5$	17.4833 ± 0.1966	0.7767 ± 0.0057	0.6514 ± 0.0051	1.0434 ± 0.0197	0.8902 ± 0.0045
Half-cosine $p = 1$	17.2365 ± 0.1546	0.7780 ± 0.0062	0.6535 ± 0.0045	1.0564 ± 0.0213	0.8938 ± 0.0049
Half-cosine $p = 2$	17.2827 ± 0.1525	0.7790 ± 0.0024	0.6542 ± 0.0052	1.0627 ± 0.0341	0.8944 ± 0.0086
Half-cosine $p = 4$	17.5070 ± 0.2260	0.7763 ± 0.0051	0.6617 ± 0.0122	1.0421 ± 0.0216	0.8956 ± 0.0045
Linear	17.6331 ± 0.1311	0.7720 ± 0.0090	0.6521 ± 0.0067	1.0276 ± 0.0307	0.8846 ± 0.0033
CIFAR-100-LT					
Half-cosine $p = 0.5$	27.5526 ± 0.2349	0.8002 ± 0.0063	0.6312 ± 0.0066	1.0181 ± 0.0304	0.8175 ± 0.0063
Half-cosine $p = 1$	27.2338 ± 0.2802	0.8007 ± 0.0099	0.6324 ± 0.0090	1.0315 ± 0.0451	0.8246 ± 0.0069
Half-cosine $p = 2$	27.1042 ± 0.4260	0.7960 ± 0.0076	0.6328 ± 0.0068	1.0177 ± 0.0344	0.8206 ± 0.0076
Half-cosine $p = 4$	29.1249 ± 0.4737	0.8028 ± 0.0045	0.6201 ± 0.0051	1.0466 ± 0.0232	0.8212 ± 0.0060
Linear	26.6535 ± 0.3281	0.7968 ± 0.0051	0.6322 ± 0.0063	1.0253 ± 0.0246	0.8304 ± 0.0050

On CIFAR-10-LT, the four half-cosine curvatures produce similar metric values, and replacing the half-cosine with a linear path does not lead to a qualitative change in performance. On CIFAR-100-LT, results for $p \in \{0.5, 1, 2\}$ remain closely grouped, while $p = 4$ has higher FID and lower Recall but remains competitive on the other metrics. The linear path has the lowest mean FID and highest mean Coverage on this dataset, with Recall close to the standard path. No single path dominates all metrics, and several gaps are comparable to the variation across generation runs. These results support robustness to moderate curvature changes, but not that any single curvature is best across all metrics or universally optimal. The variants are reported for sensitivity analysis only and are not used as the main method.

H.4 Mirrored annealing path

We compare the standard forward path $1.5 \to 2.0$ with the mirrored path $2.0 \to 1.5$ on both datasets. During generation, this mirrored path is traversed in reverse, so the scheduled tail index moves from 1.5 to 2.0. For each dataset, the standard row reuses five generation archives from one checkpoint; the mirrored checkpoint is trained with the same training seed, and evaluated with the same five seeds, all else unchanged.

Table 7: Mirrored path control. Paths are stated in the forward direction; entries are means ± sample standard deviations over five generation seeds, from one training run per row.

Forward path	FID↓	Precision↑	Recall↑	Density↑	Coverage↑
CIFAR-10-LT					
Standard $1.5 \to 2.0$	17.2377 ± 0.1563	0.7780 ± 0.0064	0.6532 ± 0.0045	1.0562 ± 0.0213	0.8937 ± 0.0047
Mirrored $2.0 \to 1.5$	40.9148 ± 1.2173	0.6920 ± 0.0064	0.6313 ± 0.0076	0.9051 ± 0.0188	0.8618 ± 0.0093
CIFAR-100-LT					
Standard $1.5 \to 2.0$	27.2348 ± 0.2806	0.8005 ± 0.0098	0.6326 ± 0.0088	1.0314 ± 0.0449	0.8245 ± 0.0068
Mirrored $2.0 \to 1.5$	50.9620 ± 1.1897	0.7300 ± 0.0072	0.5872 ± 0.0079	0.9049 ± 0.0201	0.7510 ± 0.0088

The standard path has lower FID and higher means for every other overall metric on both datasets. This deterioration is consistent with the temporal allocation described in Section 4 and Appendix F. The standard path places near-Gaussian increments at the high-noise end and heavier-tailed increments near the data; the mirrored path reverses this placement, opposite to the prediction-error shift observed for the standard path. This comparison supports the chosen orientation.

H.5 SENSITIVITY TO THE ANNEALING ENDPOINTS

To isolate endpoint sensitivity, we replace the standard lower endpoint 1.5 with 1.4 or 1.6, or the standard upper endpoint 2.0 with 1.9. We retain the standard half-cosine path, training budget, and training seed. This sensitivity study uses one training run per path, and each trained checkpoint is evaluated with the same five generation seeds. The standard $1.5 \to 2.0$ path serves as the reference.

Table 8: Endpoint sensitivity at the matched training budget. Entries are means over the same five generation seeds, with one trained checkpoint per row.

Path	FID↓	Precision↑	Recall↑	Density↑	Coverage↑
CIFAR-10-LT					
Standard $1.5 \to 2.0$	17.2377	0.7780	0.6532	1.0562	0.8937
$1.4 \to 2.0$	17.0571	0.7785	0.6553	1.0315	0.8926
$1.6 \to 2.0$	17.5589	0.7836	0.6515	1.0560	0.8905
$1.5 \to 1.9$	19.8089	0.7676	0.6541	1.0346	0.8872
CIFAR-100-LT					
Standard $1.5 \to 2.0$	27.2348	0.8005	0.6326	1.0314	0.8245
$1.4 \to 2.0$	26.1814	0.7910	0.6386	0.9944	0.8183
$1.6 \to 2.0$	27.0712	0.8046	0.6283	1.0305	0.8231
$1.5 \to 1.9$	28.9925	0.7877	0.6354	1.0068	0.8148

These results show that the lower endpoint is robust and that the main result does not depend on tuning it precisely to 1.5: both directions preserve the low-FID, high-Recall region relative to the DLPM baselines on both datasets. The more heavily tailed lower endpoint can improve FID and Recall, but does not uniformly improve the other coordinates.

Moving only the upper endpoint produces a qualitatively different response. Ending at 1.9 consistently worsens FID and no longer retains the simultaneous FID–Recall advantage over all evaluated DLPM baselines on either dataset. This FID response is consistent with the intended placement of near-Gaussian increments at the high-noise end and heavier-tailed increments closer to the data.

H.6 SCALE CLIPPING SENSITIVITY

For numerical stability, the image implementation truncates large positive stable scale draws during both training and generation, following DLPM and LIM:

$$A_t^{(C)} := \min\{A_t, C\}. \tag{33}$$

The truncation is an implementation detail rather than part of the ideal process analyzed in Sections 3 and 4. The main experiments use $C = 20$. To measure the effect of clipping, we compare $C \in \{10, 20, 50, \infty\}$ on both image datasets, where $C = \infty$ applies no clipping. We reuse the main $C = 20$ checkpoint and its generated archives, and train one matched checkpoint for each of the other clipping thresholds $C \in \{10, 50, \infty\}$. All other settings remain fixed: the standard $1.5 \to 2.0$ half-cosine path, training seed, training budget, and corrected reverse update. We use one training run for each clipping threshold and evaluate each trained checkpoint with five generation seeds, each producing 5,000 images.

The results in Table 9 show that removing clipping degrades every overall metric on both datasets, and differences among finite thresholds are minor compared with the degradation caused by removing clipping.

Table 9: Positive stable scale clipping at the main checkpoint. Entries are means ± sample standard deviations over five generation seeds, from one training run per row. The $C = 20$ rows reuse the corresponding checkpoints from the main configuration.

Threshold C	FID↓	Precision↑	Recall↑	Density↑	Coverage↑
CIFAR-10-LT					
10	17.2798 ± 0.1767	0.7794 ± 0.0075	0.6562 ± 0.0059	1.0419 ± 0.0181	0.8890 ± 0.0061
20 (main)	17.2377 ± 0.1563	0.7780 ± 0.0064	0.6532 ± 0.0045	1.0562 ± 0.0213	0.8937 ± 0.0047
50	17.6362 ± 0.0922	0.7770 ± 0.0031	0.6484 ± 0.0040	1.0560 ± 0.0146	0.8815 ± 0.0060
∞	24.7582 ± 0.3803	0.7361 ± 0.0103	0.6048 ± 0.0091	0.8541 ± 0.0146	0.7939 ± 0.0107
CIFAR-100-LT					
10	25.6871 ± 0.2669	0.7966 ± 0.0076	0.6444 ± 0.0099	1.0240 ± 0.0195	0.8323 ± 0.0102
20 (main)	27.2348 ± 0.2806	0.8005 ± 0.0098	0.6326 ± 0.0088	1.0314 ± 0.0449	0.8245 ± 0.0068
50	27.9788 ± 0.4528	0.8015 ± 0.0056	0.6306 ± 0.0056	1.0157 ± 0.0299	0.8220 ± 0.0085
∞	32.4134 ± 0.3772	0.7610 ± 0.0064	0.5729 ± 0.0096	0.8616 ± 0.0118	0.7316 ± 0.0080

H.7 Per-class Recall and Coverage

We further separate the grouped analysis of Figure 4 into individual classes. For each generated archive, we use the complete original balanced CIFAR test split as real queries with their ground truth labels and compute the global $k = 5$ neighborhoods. The Recall and Coverage differences in Figure 8 are computed from means over five training runs, averaging five generation seeds within each run.

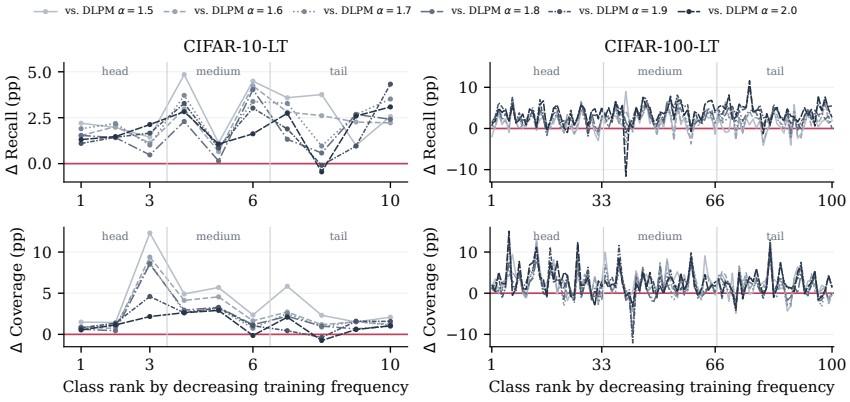


Figure 8: **Per-class differences against DLPM.** Curves show A-DLPM minus each DLPM baseline, in percentage points, so positive values favor A-DLPM. Classes follow decreasing training frequency.

On CIFAR-10-LT, A-DLPM exceeds all six fixed DLPM means in Recall for 9 of 10 classes and in Coverage for 8 of 10 classes, showing broad gains. The corresponding counts on CIFAR-100-LT are 68 of 100 and 51 of 100; most negative values are moderate in size, and the overall tendency remains positive.

H.8 Qualitative Samples

Figure 9 shows samples from A-DLPM stratified by predicted class. For each dataset, candidates come from one trained checkpoint and five generation runs. For classification, we use one public pretrained PyramidNet-272 checkpoint per dataset, released by Zhao et al. (2022); PyramidNet is the classifier architecture introduced by Han et al. (2017). On the original balanced test splits, we reproduce accuracies of 98.67% for CIFAR-10 and 88.98% for CIFAR-100.

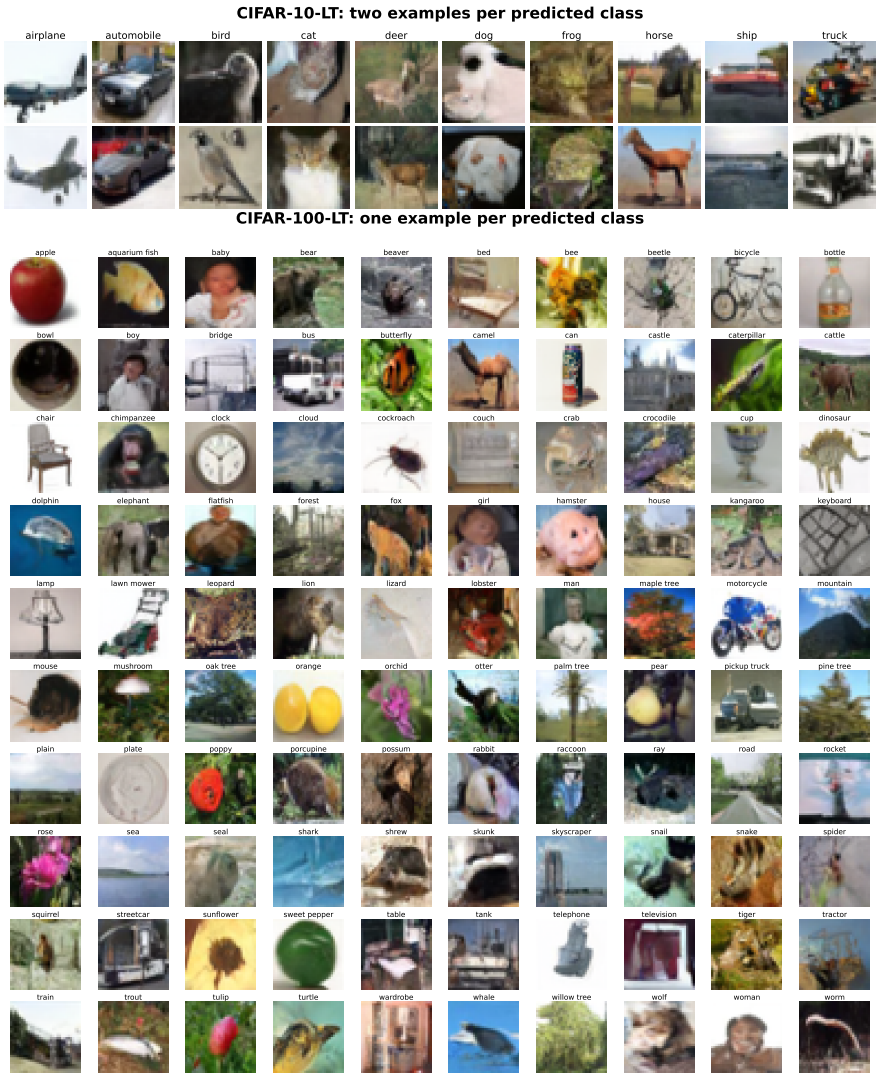


Figure 9: **Samples from A-DLPM organized by predicted class.** The upper panel contains two samples for each predicted CIFAR-10 class. The lower panel contains one sample for each of the 100 predicted CIFAR-100 classes.