SMALL LOSS, POOR GENERATOR: WHERE CONSISTENCY TUNING SHOULD TRAIN FOR ONE-STEP GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training a one-step image generator is expensive. Easy Consistency Tuning (ECT) turns a pretrained diffusion model into a one-step consistency model by fine-tuning, and its smallest setting on CIFAR-10 uses about one million images. We study which distribution of training noise levels (which we call the training path) gives the best one-step generator for a fixed budget, and why. We introduce a novel theoretical framework with the following results. (i) The averaged consistency loss can tend to zero whilst the generator stays wrong. (ii) Adapting a known telescoping bound to the level T' at which generation starts, we bound the generator's error by quantities that involve only noise levels $t \leq T'$. (iii) In a simple chain model of training, the least-trained pairs of neighbouring noise levels below T' limit how fast the generator improves. (iv) A uniform draw of the noise levels gives every such pair nearly the same weight in the averaged loss. From this framework we derive the resistance of a training path, which is computed before any training. We apply these theoretical results to propose Uniform Consistency Tuning (UCT). Experiments on CIFAR-10 and ImageNet-64 show that UCT improves the one-step generation of ECT at small training budgets.

1 INTRODUCTION

Diffusion models generate an image by removing noise over many steps (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). Consistency models learn a flow map that directly connects a noisy image with a clean image, which lets a single pass generate an image from noise (Song et al., 2023). Generating in a single pass is cheaper than generating over multiple steps, but it is prone to higher error, because no later step can correct it.

Training a one-step image generator is expensive (Liu et al., 2024; Zhou et al., 2024). Recent one-step generators use training budgets of 330 to 820 million images even on CIFAR-10 (Lee et al.,

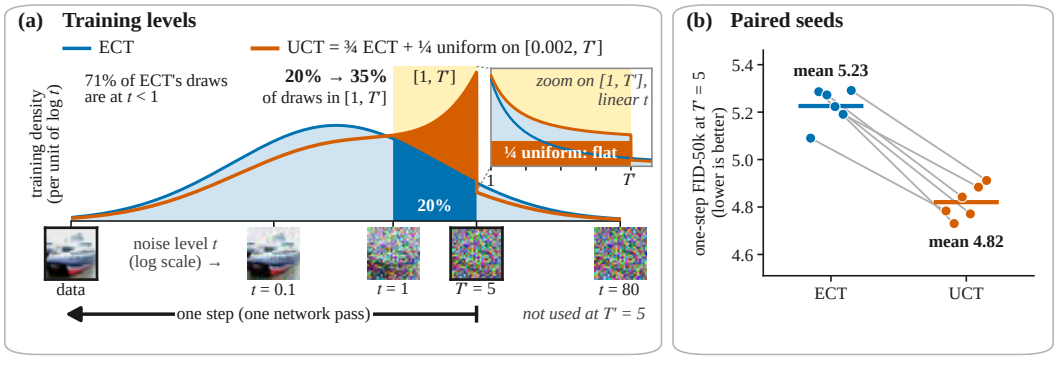


Figure 1: **ECT rarely trains on noisier levels.** (a) Training-level density for ECT and for UCT. The zoomed-in panel shows $[1, T']$ on a linear axis, where the uniform part of UCT's density is flat. (b) One-step FID-50k on CIFAR-10 after 8000 updates.

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.

2025; Song & Dhariwal, 2024; Geng et al., 2025a; Table 5), and cutting this cost is an active goal (Hu et al., 2026). Easy Consistency Tuning (ECT; Geng et al., 2025b) turns a pretrained diffusion model into a consistency model by fine-tuning, and its smallest setting on CIFAR-10 tunes on about one million images, 325 to 800 times fewer.

Because every update counts at such a small budget, we study which distribution of training noise levels (which we call the training path) gives the best one-step generator for a fixed budget, and why. ECT trains the pretrained diffusion model to give the same estimate of the clean image from two noisy versions of the image that share the noise but differ in how noisy they are (the noise level t). ECT draws noise levels from a lognormal distribution, and 71% of them lie below the noise level $t = 1$ (Figure 1a). A one-step generator starts from T', a much higher level than $t = 1$. Its output image is the sum of the changes in the network's estimates between neighbouring levels, from T' to $t = 0$ (the clean image), and an inconsistency between any two neighbouring levels therefore ends up in the output image. ECT trains mostly between the noise levels $t = 0$ and $t = 1$, and seldom on the noisier levels between $t = 1$ and $T' = 5$ (20%). Because the less noisy levels that ECT trains on most carry most of the weight in the averaged loss, this loss can be small whilst the connection between the starting noise and the output image is poor.

We introduce a novel theoretical framework that relates the training path to the averaged loss and to the error of the one-step generator (Section 4). We show with a counterexample that minimising the averaged loss need not lead to a better one-step generator (Theorem 1). Instead, adapting the telescoping bound of ECT and Kong et al. (2024) to the start level T', we bound the error of the generator by quantities that involve only levels $t \leq T'$ (Theorem 2). We call a pair of neighbouring noise levels below T' a link. In a simple chain model of training, each link pulls the output at its upper level towards a target set by its lower level, at a rate set by the training path, and the least-trained links limit how fast the chain becomes consistent (Theorem 3). In this model the training path sets the speed of training rather than its end point, which Kingma et al. (2021) and Kingma & Gao (2023) observed for the noise schedule of diffusion losses and which we quantify for consistency tuning; it therefore matters most when the budget is small. Under ECT's lognormal noise sampling, these least-trained links are the links just below T'. In contrast, a uniform draw gives every link below T' nearly the same weight in the averaged loss (Proposition 4).

Applying these results, we propose Uniform Consistency Tuning (UCT): we draw a quarter of the training levels uniformly from a range between the lowest level and T', and the rest from ECT's lognormal noise sampling (Algorithm 1). This gives more training to the links just below T', which ECT trains least. UCT changes nothing else, and both methods start generation from the same level.

At a budget of about one million images on CIFAR-10 (Krizhevsky, 2009), UCT lowers the one-step Fréchet Inception Distance (FID) (Heusel et al., 2017) in each of the six paired seeds (Figure 1b). From the data-matched start at $T' = 5$, UCT reaches ECT's final FID-10k with 19% fewer updates. After ECT's full schedule of $100{,}000$ updates, UCT also lowers one-step FID from both starts. On class-conditional ImageNet-64 (Deng et al., 2009), UCT lowers one-step FID from the data-matched start after both 8000 and $48{,}000$ updates, and after $100{,}000$ updates the two one-step FIDs are within 1%. In an ablation on CIFAR-10 with $T' = 5$, the chain model separates the training paths that improve on ECT from those that do not, except when every level is drawn from the range. Of the three ranges we tried, the one ending at the start level is the best, as the chain model predicts.

Our contributions are as follows. (i) Theory: we introduce a novel theoretical framework for the training path of consistency tuning, which is the main contribution of this paper. We prove that a small averaged loss does not certify the generator (Theorem 1), adapt the known telescoping bound of ECT and Kong et al. (2024) to the start level (Theorem 2), and show in a chain model how the training path sets the speed of training (Theorem 3, Proposition 4); from these we derive the resistance of a training path, computed before any training. (ii) Algorithm: we propose UCT, which changes only ECT's training path, drawing a quarter of its training levels uniformly up to T', and adds no extra cost. (iii) Empirical: after 8000 updates, UCT lowers one-step FID in each of the six paired seeds on CIFAR-10 and from the data-matched start in both ImageNet-64 seeds, and on CIFAR-10 the resistance ranks tuned generators consistently with FID, best of the measures we compared, whilst the averaged loss does not.

2 Preliminaries and Related Work

This section gives the background that the theory of Section 4 uses, including the probability-flow ordinary differential equation (ODE), and then turns to related work.

Noise levels and the probability-flow ODE. Let x_0 be a training image, drawn from the data distribution ρ with mean m and covariance Σ, and let z be standard Gaussian noise. The noisy image at noise level t is $x_0 + tz$, which at $t = 80$ is essentially pure noise, and we write ρ_t for its distribution. A diffusion model generates an image by following the data's probability-flow ODE (Song et al., 2021; Karras et al., 2022) from a high noise level ($t = 80$) down to $t = 0$, with a network in place of the conditional mean:

$$\dot{x} = \big(x - \mathbb{E}[x_0 \mid x_0 + tz = x]\big)/t. \tag{1}$$

Its solutions carry the distribution ρ_t of the noisy data at level t to the distribution ρ_r at every lower level r (Song et al., 2021). Following the ODE takes many network evaluations: the sampler of EDM, the diffusion model of Karras et al. (2022) that ECT fine-tunes, uses 35 on CIFAR-10.

Consistency models and one-step generation. A consistency model is a network $f_\theta(x, t)$ trained to map a noisy image at level t to the end of its probability-flow path, with $f_\theta(x, 0) = x$ (Song et al., 2023). Later work improved how consistency models are trained (Song & Dhariwal, 2024; Geng et al., 2025b; Lu & Song, 2025), and related methods learn flow maps, shortcuts, average velocities or moment-matched two-time samplers (Kim et al., 2024; Frans et al., 2025; Zhou et al., 2025; Boffi et al., 2025; Geng et al., 2025a; Hu et al., 2026). The standard start is pure noise, $\mathcal{N}(0, 80^2 I)$ at level 80. At a lower start level T' the noisy data have the same mean and covariance as the Gaussian $\gamma_{T'} = \mathcal{N}(m, \Sigma + T'^2 I)$, and their distance to it does not grow with T' (Section 4). The generator draws from this Gaussian and applies the network once:

$$G_{T'}(\xi) = f_\theta\big(m + (\Sigma + T'^2 I)^{1/2}\xi,\ T'\big), \qquad \xi \sim \mathcal{N}(0, I). \tag{2}$$

We take this start from Raya & Ambrogioni (2023, Section 5) and Wang & Vastola (2024), estimate m and Σ from the training images and call it the **data-matched start**.

Easy Consistency Tuning. ECT (Geng et al., 2025b) fine-tunes a pretrained diffusion model into a consistency model. It draws a level t from a distribution p, pairs it with a slightly lower level $\tau(t)$, with gap $\Delta(t) = t - \tau(t) > 0$, and minimises the averaged loss

$$\mathcal{L}_p(\theta) = \mathbb{E}_{x_0, z,\, t\sim p}\left[\frac{\big\|f_\theta(x_0 + tz, t) - \operatorname{sg} f_\theta(x_0 + \tau(t)z, \tau(t))\big\|_2}{\Delta(t)}\right], \tag{3}$$

where sg stops the gradient, which makes the output at the lower level the target and removes the need for a teacher. On CIFAR-10, ECT draws t from a lognormal p_{ECT}, under which $\log t$ is normal with mean -1.1 and standard deviation 2, the continuous form of the discrete distribution of levels in improved consistency training (iCT; Song & Dhariwal, 2024). Appendix B.1 gives τ and the training settings. ECT generates with an exponential moving average (EMA) of its weights. ECT's training schedule sets how the gap between paired levels shrinks over the stages of training (Geng et al., 2025b, Section 3.2), whilst p stays fixed. We keep this schedule and call p the training path (Section 4). The two levels of a pair share the noise but need not lie on one probability-flow path, and even the ideal map has a positive loss on a pair (Appendix A.2).

Which noise levels to train. Reweighting and resampling noise levels are well studied for diffusion models (Choi et al., 2022; Hang et al., 2023; Kingma & Gao, 2023; Hang et al., 2025; Ambrogioni et al., 2026), and reshaping the time distribution helps one-step rectified flows (Lee et al., 2024). The original consistency model trains every link equally (Song et al., 2023). iCT replaced this equal training with a lognormal that puts much less weight on high levels, because errors at lower levels carry over to higher ones (Song & Dhariwal, 2024, Sections 3.1 and 3.5), and ECT keeps it. Gokmen et al. (2024) argue that this lognormal trains high noise levels too rarely, and Xu et al. (2025) draw most levels near the start of one-step generation. Truncated consistency models (TCM) train a second stage only at and above a dividing level (Lee et al., 2025), and SANA-Sprint puts a fixed share of its adversarial training at the level where generation starts (Chen et al., 2025a). Adaptive Discretization for Consistency Models (ADCM) adapts the gap between paired levels, and because its loss is not divided by the gap, this changes each link's weight as a change of the training

Algorithm 1 Uniform Consistency Tuning (UCT), in the notation of ECT (Geng et al., 2025b, Algorithm 1). The highlighted line is the only change to ECT. On CIFAR-10, $w(t) = 1/(t - r)$ and $d(x, y) = \|x - y\|_2$, which gives the loss (3), and $r = \tau(t)$ in the first stage of ECT's schedule. On ImageNet-64, w and d follow ECT's ImageNet-64 code (Appendix B.6).

Input: dataset $\mathcal{D}$, pretrained weights θ_0, noise distribution $p_{\mathrm{ECT}}(t)$, mapping function $p(r \mid t, \mathrm{Iters})$, weighting $w(t)$, metric d, learning rate η, range $[a, T']$, probability π, EMA decay β, updates S.
Init: $\theta \leftarrow \theta_0$, $\theta_{\mathrm{ema}} \leftarrow \theta_0$, $\mathrm{Iters} = 0$.
repeat
  Sample $x_0 \sim \mathcal{D}$, $z \sim \mathcal{N}(0, I)$, $t \sim p_{\mathrm{ECT}}(t)$
  With probability π, resample $t \sim \mathcal{U}[a, T']$
  Sample $r \sim p(r \mid t, \mathrm{Iters})$
  Compute $x_t = x_0 + t \cdot z$, $x_r = x_0 + r \cdot z$
  $\mathcal{L}(\theta) = w(t) \cdot d\big(f_\theta(x_t, t),\ f_{\mathrm{sg}(\theta)}(x_r, r)\big)$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}(\theta)$
  $\theta_{\mathrm{ema}} \leftarrow \beta\, \theta_{\mathrm{ema}} + (1 - \beta)\, \theta$
  $\mathrm{Iters} = \mathrm{Iters} + 1$
until $\mathrm{Iters} = S$
return θ_{ema}

path does (Bai et al., 2025). Like Xu et al. (2025) and SANA-Sprint, UCT trains more near the start, but it keeps ECT's single stage and pairing and spreads a quarter of its draws uniformly between the lowest level and T', which the chain model favours over concentrating them at the start (Section 5.2).

Error bounds. Theorem 2 adapts the telescoping argument of ECT (Geng et al., 2025b, Section 3.1) and the bound of Kong et al. (2024, Theorem 3.1), which already starts from an intermediate level and sums the consistency losses below it, but uses the model's Lipschitz constant; we use the ideal map's constant and the consistency gaps, which makes the start term independent of training. The guarantees of Song et al. (2023, Theorem 1), Lyu et al. (2024), Li et al. (2025), Chen et al. (2025b) and Yang et al. (2025), which start from pure noise, have the same structure. Stable consistency tuning views consistency training as temporal-difference learning (Wang et al., 2025), and our chain model is a simple case of this view; in asynchronous Q-learning, likewise, the least-visited state-action pairs limit convergence (Li et al., 2022). Our work differs from these in how the training path sets the rate of each link, and Appendix A relates each of our results to these earlier ones. Wang & Vastola (2024, Section 8) also observe that, in principle, a denoiser need not be trained at high noise levels when sampling begins at the data-matched start.

3 Uniform Consistency Tuning

We apply our framework to ECT and change only its training path, which gives Uniform Consistency Tuning (UCT, Algorithm 1). With probability π, UCT draws the level uniformly from the **range** $[a, T']$, where a is the smallest nonzero noise level of EDM's sampler, and otherwise from ECT's lognormal noise sampling:

$$p_\pi = (1 - \pi)\, p_{\mathrm{ECT}} + \pi\, \mathcal{U}[a, T'], \qquad \pi = 0.25, \qquad a = 0.002. \tag{4}$$

The range ends at T' because only levels at or below T' enter the error bound of Section 4. It is uniform because the loss divides by the gap, which gives each link of the chain below T' nearly the same weight (Proposition 4), as the original consistency model does (Song et al., 2023). We keep three quarters of ECT's draws, because drawing every level from the range is worse than ECT at $T' = 5$ and fails from pure noise (Section 5.2). Everything else follows ECT, including all hyperparameters and, on CIFAR-10, the pretrained EDM model and the Rectified Adam (RAdam) optimiser (Liu et al., 2020) (Appendices B.1 and B.6). An update takes about the same time in both methods, and generation adds one precomputed matrix product.

4 THEORY OF TRAINING PATHS

We call the distribution p of training noise levels the **training path**, because it decides how much training each part of the path from T' down to the clean image receives. We ask whether a small averaged loss guarantees a good one-step generator, which levels control its error, and how p sets the speed at which these levels become consistent. We write f for the network, $\gamma_t = \mathcal{N}(m, \Sigma + t^2 I)$ for the Gaussian at level t, $F_{\#}\mu$ for the distribution of $F(X)$ when $X \sim \mu$, and W_2 for the 2-Wasserstein distance (Villani, 2009, Chapter 6), $W_2(\mu,\nu) = \inf\left(\mathbb{E}\|X - X'\|_2^2\right)^{1/2}$ over all pairs of random vectors (X, X') with distributions μ and ν. The chain below the start is $t_0 = T' > t_1 > \cdots > t_N$ with $t_{j+1} = \tau(t_j)$, where t_N is the last level not below a, and **link** j is the pair (t_j, t_{j+1}). The ideal consistency map $f^*(\cdot, t)$ moves a point at level t along the probability-flow ODE (1) down to a level $0 < \epsilon \le t_N$, and it therefore maps $\rho_{T'}$ to ρ_ϵ.

Assumptions. The results use three assumptions. (A1) The data distribution ρ has a finite second moment, and the noise z is independent of x_0. (A2) The pairing $\tau : (0,\infty) \to (0,\infty)$ is continuous and strictly increasing, with $0 < \tau(t) < t$ and $\tau(t) \to \infty$ as $t \to \infty$, which ECT's pairing satisfies (Appendix A). (A3) The probability-flow ODE (1) has a unique solution from every point, as it does for bounded data such as images. Appendix A gives the proofs.

Theorem 1 (A small loss does not certify the generator)**.** *Assume (A2). Let the data ρ be a single point y, and let p be a training path on $(0,\infty)$ without a point mass at T' or at $\tau^{-1}(T')$, such as ECT's lognormal or the training path of UCT. For every unit vector v and every $\alpha \neq 0$ there are smooth maps f_n with $f_n(x,0) = x$ and $\|f_n(x,t) - f_n(x',t)\|_2 \le \|x - x'\|_2$ such that $\mathcal{L}_p(f_n) \to 0$ as $n \to \infty$, also with the squared norm in (3), whilst $W_2\big(f_n(\cdot,T')_{\#}\gamma_{T'},\, \rho\big) = |\alpha|$ for every n.*

The proof builds maps that send every start to $y + \alpha v$. They change with the level only in two narrow bands, near 0 and near T', whose contribution to the loss tends to zero as the bands shrink with n (Appendix A.1). Such maps are ruled out by exact consistency on the chain below T', which the error bound of Song et al. (2023, Theorem 1) assumes, but not by a small averaged loss. TCM adds a point mass at its dividing level because the same mechanism would otherwise collapse its objective to a constant (Lee et al., 2025, Section 3); Theorem 1 proves the same weakness at the start level. Hence a training path cannot be judged by how small it makes the averaged loss. Theorem 2 looks at the chain below T' instead; we obtain it by adapting the known telescoping bound of ECT (Geng et al., 2025b, Section 3.1) and Kong et al. (2024, Theorem 3.1) to the data-matched start.

Theorem 2 (What bounds the one-step error)**.** *Assume (A1)–(A3), and that $f^*(\cdot, T')$ is $L_{T'}$-Lipschitz. Let f be measurable, let $V_0 \sim \gamma_{T'}$, let $V_1, \ldots, V_N$ be random vectors on the same probability space such that all $f(V_j, t_j)$ and $f^*(V_j, t_j)$ are square-integrable, and put $e_j = f(V_j, t_j) - f^*(V_j, t_j)$ and $d_j = e_j - e_{j+1}$. Then*

$$\underbrace{W_2\big(f(\cdot,T')_{\#}\gamma_{T'},\, \rho_\epsilon\big)}_{\substack{\text{one-step error}\\ \text{(generated vs. data)}}} \le \underbrace{L_{T'}\,\delta_{T'}}_{\substack{\text{start mismatch}\\ \text{(Gaussian vs. noisy data)}}} + \underbrace{\textstyle\sum_{j<N}\|d_j\|}_{\substack{\text{consistency gaps}\\ \text{on the links below } T'}} + \underbrace{\|e_N\|}_{\substack{\text{error at the}\\ \text{last level}}}, \tag{5}$$

where $\delta_{T'} = W_2(\rho_{T'}, \gamma_{T'})$ and $\|X\| = \left(\mathbb{E}\|X\|_2^2\right)^{1/2}$.

The consistency gap d_j is the model's change across link j minus the ideal map's change, which is the model's inconsistency on link j when the V_j lie on one path of the data's ODE. The proof couples V_0 with a sample of $\rho_{T'}$ and writes $e_0 = e_N + \sum_{j<N} d_j$, with the ideal map's Lipschitz constant in place of the model's. The bound therefore tells us where training matters: only levels at or below T' enter it (draws above T' do not reduce it directly), and every link below T' enters with the same weight, however rarely the training path visits it. The start term does not depend on training, and its mismatch $\delta_{T'}$ cannot grow with T', because adding the same independent noise to ρ_t and γ_t does not increase their distance, but a higher start lengthens the chain and can raise $L_{T'}$.

To see how the training path affects the consistency gaps d_j, consider a **chain model** of training. Number the levels upward, from the lowest one ($i = 0$) to the start ($i = N$). Let y_i be the output at level i, y_i^* its ideal value, $q_i = y_i - y_i^*$ its error and $c_i = y_i^* - y_{i-1}^*$. The model assumes that (M1) every level has its own output and no weights are shared, (M2) the i-th link from the bottom carries the loss $\frac{\lambda_i}{2}\|y_i - \mathrm{sg}(y_{i-1}) - c_i\|_2^2$, (M3) training is gradient flow on the sum of these losses, and (M4) the lowest level is exact, $q_0 = 0$. Because the output y_i enters the loss of the link above

only through sg, gradient flow gives $\dot{y}_i = -\lambda_i(y_i - y_{i-1} - c_i)$, and the errors therefore follow $\dot{q}_i = -\lambda_i(q_i - q_{i-1})$ in each coordinate. The i-th link from the bottom is link $N - i$ of the chain, and we take its rate λ_i to be the loss weight that p puts on it,

$$\lambda_i = \int_{t_{N-i+1}}^{t_{N-i}} \frac{p(t)}{\Delta(t)}\, dt. \tag{6}$$

Theorem 3 (The least-trained links limit the chain)**.** *Let $\lambda_1, \dots, \lambda_N > 0$, and let q solve $\dot{q}_i = -\lambda_i(q_i - q_{i-1})$ for $i = 1, \dots, N$, with $q_0 \equiv 0$ and ordered initial errors $0 \le q_1(0) \le \cdots \le q_N(0)$. Let $E_1, \dots, E_N$ be independent* Exp(1) *variables and $H_k = \sum_{i=k+1}^{N} E_i/\lambda_i$. Then, for all $s \ge 0$,*

(i) $q_N(s) = \sum_{k=1}^{N} \{q_k(0) - q_{k-1}(0)\}\, \mathbb{P}(H_{k-1} > s)$*;*

(ii) if $\tilde{\lambda}_i \ge \lambda_i$ for every i, the solution $\tilde{q}$ with rates $\tilde{\lambda}_i$ and the same initial errors satisfies $0 \le \tilde{q}_N(s) \le q_N(s)$;

(iii) if $s > 0$,

$$q_N(0)\, e^{-\lambda_N s} \;\le\; q_N(s) \;\le\; \frac{q_N(0)}{s} \sum_{i=1}^{N} \frac{1}{\lambda_i}. \tag{7}$$

For the proof, consider a walker that starts at level N, waits for the time E_i/λ_i at each level $i \ge 1$, then steps down one level, and stops at 0. It reaches level $k - 1$ at time H_{k-1}. Conditioning on the first wait shows that $q_N(s)$ is the mean initial error at the walker's level at time s, which gives (i). Higher rates shorten every wait, which gives (ii) because the increments $q_k(0) - q_{k-1}(0)$ are nonnegative, and Markov's inequality for H_{k-1} and $H_{k-1} \ge E_N/\lambda_N$ give (iii) (Appendix A.3).

A walker from the top crosses every link before it reaches level 0, and this takes the mean time $\mathbb{E}H_0 = R = \sum_i 1/\lambda_i$, which we call the chain's **resistance**. The upper bound in (7) is proportional to R, and the lower bound shows that the top link alone caps the speed. The resistance compares training paths from p alone, before any training (Sections 5.2 and 5.3). From the same initial error, a path with a smaller resistance has a smaller upper bound on the error at every time, and a path that starves the top link is slow however much it trains the other links. Under ECT's lognormal on CIFAR-10, the top link below $T' = 5$ has a rate 279 times smaller than the fastest links, and the links that reach above $t = 1$ carry 94.9% of R. The training path of UCT lowers R to 0.46 times ECT's, mainly by raising the rates of these links. It also lowers the rates of the lower links, by at most a quarter, and part (ii) therefore does not apply; numerically, the model gives UCT an error no larger than ECT's at every time we checked, for every ordered initial error (Appendix A.3).

Proposition 4 (A uniform range weights every link nearly equally)**.** *Assume (A2). Let p be uniform on $[a, T']$, and let link j lie in $[a, T']$. If Δ is differentiable with $\kappa_- \le \Delta' \le \kappa_+ < 1$ on the link, then its weight $w_j = \int_{t_{j+1}}^{t_j} p(t)/\Delta(t)\, dt$ satisfies*

$$\varphi(\kappa_-) \;\le\; (T' - a)\, w_j \;\le\; \varphi(\kappa_+), \qquad \varphi(\kappa) = \frac{-\log(1-\kappa)}{\kappa}, \quad \varphi(0) = 1.$$

In particular, a gap proportional to the noise level gives every link the same weight.

The width of a link cancels, because a wide link is hit more often by a uniform draw but each draw is divided by a larger gap. For ECT's pairing τ the slope of the gap stays below 0.02, and all 585 links below $T' = 5$ receive the same weight to within 1% (Appendix A.4). A uniform range therefore adds nearly the same rate to every link and, relative to their rates, helps the slowest links most without our needing to know where they are. In the chain model, a density that favours low levels, where ECT's lognormal trains fast, helps less (Table 2).

5 Experiments

Setup. On unconditional CIFAR-10 we follow ECT's training settings for 8000 updates with six training seeds. Within a seed, ECT and UCT share the data order and ECT's level draws, except that UCT replaces each draw by a uniform one with probability π. Seeds 0–4 share the random stream

Table 1: **One-step FID-50k after tuning** (mean ± standard deviation over paired seeds; lower is better). Data-matched start: the Gaussian at $T' = 5$. Pure noise: ECT's standard start at level 80. Rows with 48k and 100k updates follow ECT's full schedule with one paired seed.

Dataset	Start	Updates	Seeds	ECT	UCT
CIFAR-10	data-matched, $T' = 5$	8k	6	5.23 ± 0.08	4.82 ± 0.07
CIFAR-10	pure noise	8k	6	7.35 ± 0.12	6.34 ± 0.21
CIFAR-10	data-matched, $T' = 5$	100k	1	4.13	4.00
CIFAR-10	pure noise	100k	1	4.50	4.30
ImageNet-64	data-matched, $T' = 5$	8k	2	49.75 ± 0.25	48.39 ± 0.74
ImageNet-64	data-matched, $T' = 5$	48k	1	19.66	19.17
ImageNet-64	data-matched, $T' = 5$	100k	1	8.50	8.53

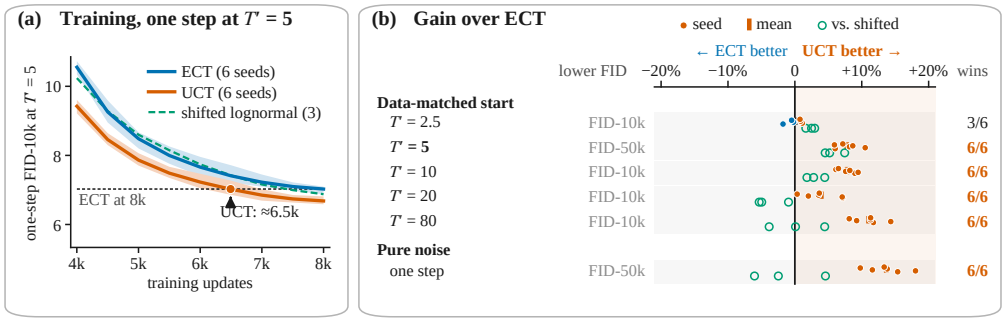


Figure 2: **UCT improves on ECT at every start level shown except $T' = 2.5$, where the two tie.** Unconditional CIFAR-10, 8000 updates, one step (lower is better). (a) FID-10k at $T' = 5$, updates 4000–8000 (mean and min–max over six seeds; dashed: shifted lognormal, seeds 0–2). (b) FID reduction relative to ECT per paired seed (dots) and on average (bars); the right column counts seeds where UCT is better, and hollow circles compare UCT with the shifted lognormal (seeds 0–2).

of these replacements. We evaluate one-step generation from the data-matched start at $T' = 5$, the start level with the lowest FID-10k for ECT, and from pure noise at level 80. We score both starts with FID-50k, which compares the Inception features of 50,000 generated images with those of the training images (Heusel et al., 2017). Other start levels and the training curves use FID-10k, and all methods use the same evaluation noise. On class-conditional ImageNet-64 we run ECT's code from the EDM2-S model (Karras et al., 2024) with two paired seeds. To test longer training, we also train one pair of runs for 100,000 updates on each dataset. Appendix B gives the details.

5.1 Main Results

CIFAR-10. Our run of ECT reproduces the result that Geng et al. (2025b) report for this budget (Appendix B.2). From the data-matched start at $T' = 5$, UCT lowers FID-50k in each of the six paired seeds (Table 1, and Table 6 in Appendix B.7). The improvement is systematic rather than noise between seeds: the 95% interval for the mean reduction is $[0.31, 0.51]$, and a one-sided sign test gives a p-value of $1/64$. The gain is not tied to the data-matched start. Starting instead from noised training images, which match the noisy data exactly, UCT is again lower in each of seeds 0–2 (mean FID-10k 5.96 against 6.34), and it is also lower from pure noise and at every other start level we tried except $T' = 2.5$, where the two methods tie (Figure 2b). The gain also appears early in training: at $T' = 5$, UCT is below ECT at every checkpoint we evaluated and reaches ECT's final FID-10k with 19% fewer updates (Figure 2a). Finally, the gain is not specific to the Inception features of FID, because the Fréchet distance in DINOv2 features (Stein et al., 2023) is also lower for UCT from both starts and in each of seeds 0–2 (Table 7 in Appendix B.7). With ECT's full schedule of 100,000 updates, UCT is also lower from both starts (Table 1, one paired seed).

Other one-step methods. At our budget, $1/200$ to $1/800$ of the training images in their published settings, every run of TCM, consistency distillation, continuous-time consistency training and MeanFlow is worse than each of seeds 0–2 of UCT from the same start (Table 4 in Appendix B.4). The

Table 2: **Training paths on unconditional CIFAR-10** (Appendix B.3; 8000 updates; one-step FID-50k, lower is better). Unless stated otherwise, a quarter of the draws comes from the listed distribution and the rest from ECT's lognormal, with $a = 0.002$. Resistance: the chain model's $\sum_i 1/\lambda_i$ below $T' = 5$, relative to ECT's. Seeds 0–2: mean ± standard deviation.

		Start at $T' = 5$		Pure noise
Training path	Resistance	Seed 0	Seeds 0–2	Seed 0
ECT ($\pi = 0$)	1.00	5.09	5.22 ± 0.11	7.54
Range, $\pi = 0.1$	0.65	4.82		6.96
Range, $\pi = 0.25$ (**UCT**)	0.46	4.78	4.79 ± 0.06	6.38
Range, $\pi = 0.5$	0.34	4.82	4.86 ± 0.11	6.33
Range alone ($\pi = 1$)	0.40	5.96		318.2
Log-uniform on $[0.3, 5]$	0.63	4.97		7.17
Density $\propto \Delta(t)/t$ on $[a, 5]$	0.58	4.89		6.78
Density $\propto \Delta(t)^{-1/2}$ on $[a, 5]$	0.51	5.01		7.01
Half on $[a, 5]$, half on $[a, 0.821]$	0.65	5.01		7.10
Uniform on $[a, 80]$	1.17	5.25	5.42 ± 0.15	7.33
Log-uniform on $[a, 80]$	1.02	5.83		7.54
Log-Student-t on $[a, 80]$	1.07	5.21		7.59
Point mass at 5	1.32	5.42		6.09
Point masses at 80 and 0.821	1.33	5.29		9.42
Shifted lognormal, no range	0.65	5.01	5.08 ± 0.06	6.01
Matched lognormal, no range	2.42	7.66	6.97 ± 0.84	10.98
Shifted lognormal + range	0.42	5.01	5.00 ± 0.06	6.25
ECT's draws above 5 redrawn, + range	0.45	4.93		330.9

closest are TCM at $T' = 5$ (7.92 against UCT's 4.79) and consistency distillation from pure noise (9.41 against 6.25), both with their settings compressed to our budget (mean FID-50k over seeds 0–2).

ImageNet-64. On ImageNet-64, the data-matched start at level T' uses the statistics of the class being generated. For a label c drawn uniformly, the start is $\mathcal{N}(m_c, \Sigma_w + T'^2 I)$, where m_c is the mean of the training images of class c and Σ_w is the within-class covariance, pooled over classes. This start follows the condition-dependent Gaussian starts of Lee et al. (2022) and Issachar et al. (2025). We pool because about 1300 images per class are too few to estimate a covariance in 12,288 dimensions. As on CIFAR-10, $T' = 5$ gives the lowest FID-10k of ECT among the start levels we tried (seed 0). Both methods use the EMA profile of relative width 0.050 (Appendix B.6). After 8000 updates, UCT lowers one-step FID-50k from the data-matched start by 1.70 and 1.01 in the two paired seeds. When the seed-0 pair continues to 100,000 updates, UCT is lower at 48,000 updates (19.17 against 19.66), and at 100,000 updates they are within 1% (8.53 against 8.50).

5.2 Which training paths help

Resistance. In the chain model, the resistance of a training path measures how slowly the error at the start level is removed (Section 4). We compute and test the resistance for the chain below the data-matched start, $T' = 5$; we do not use the chain model to predict FID from pure noise. At seed 0, every path in Table 2 with a lower resistance than ECT improves FID at $T' = 5$, except the range alone ($\pi = 1$), and every path with a higher resistance is worse than ECT. The range alone and the variant that redraws ECT's draws above 5 train no level above T', and neither can generate from pure noise. Of the range shapes we tried, the uniform one gives the lowest FID, and each of $\pi = 0.1$, 0.25 and 0.5 improves on ECT.

Alternatives to the range. A simpler way to train noisier levels more is to shift ECT's lognormal towards higher noise. At $T' = 5$ this closes about a third of the gap to UCT (seeds 0–2), UCT is better in each of these seeds, and adding the range to it does not reach UCT. Spreading the extra draws up to level 80, or putting them all at T', makes FID at $T' = 5$ worse than ECT's and raises the resistance above ECT's. A lognormal that matches UCT's shares of draws below 1, between 1

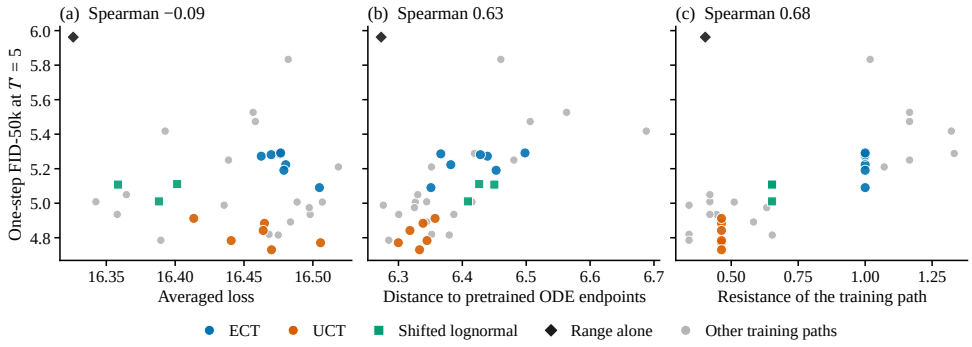


Figure 3: **What ranks the generators.** Each point is one of the 36 tuned models of Section 5.3 (8000 updates; Appendix B.5), against its one-step FID-50k at $T' = 5$ (lower is better). (a) The averaged loss (3). (b) The distance between the one-step outputs and the endpoints of the pretrained model's ODE paths from the same 2000 starts. (c) The resistance of its training path.

and 5 and above 5 moves weight away from the lowest links, has 2.42 times ECT's resistance and is much worse than ECT: matching these coarse shares is not enough.

Where the range ends. The chain model also predicts where the range should end. For each start level $T' \in \{2.5, 5, 10\}$, it gives the smallest resistance to the range that ends at T', and among ranges ending at 2.5, 5 and 10, trained with three seeds each, this range also gives the lowest mean FID-10k at that start, although at $T' = 10$ the margin is small (Table 3 in Appendix B.3).

5.3 What ranks the generators

To see whether Theorem 1 matters in practice, we took 36 tuned models from CIFAR-10 runs with different training paths and seeds, each scored with FID-50k at $T' = 5$ (Appendix B.5), and compared three ways of ranking them. The averaged loss (3), evaluated on the 10,000 test images with the same draws of noise and level for every model, does not rank them (Spearman correlation -0.09, Figure 3a). The quantities of Section 4 do. In Theorem 2 with the pretrained model's ODE in place of the data's, the term that depends on the tuned model is the distance between its one-step outputs and the endpoints of the pretrained model's ODE paths (Appendix A.2). Measured from 2000 starts at $T' = 5$, this distance ranks the models with Spearman correlation 0.63 (bootstrap 95% interval $[0.28, 0.85]$, Figure 3b). The resistance of each model's training path ranks them best, with Spearman correlation 0.68 ($[0.37, 0.88]$, Figure 3c). UCT is closer to the pretrained model's endpoints than ECT in all six paired seeds, and the range alone is the closest of all but has the worst FID. Vouitsis et al. (2024) likewise find that solving the ODE more accurately can give worse samples.

6 Conclusion

We introduced a novel theoretical framework that relates the training path of consistency tuning to its averaged loss and to the error of the resulting one-step generator. We showed that a small averaged loss does not guarantee a good one-step generator, whose error, by a known telescoping bound that we adapted, depends on every link below its start level. In a simple chain model of training, the least-trained links limit how fast the generator improves. On CIFAR-10, the resistance of a training path, computed before any training, ranks the tuned generators consistently with FID, best of the measures we compared, whilst the averaged loss does not. We applied our theoretical framework to ECT and proposed Uniform Consistency Tuning (UCT), which changes only its training path: it draws a quarter of the training levels uniformly up to the start level, adding nearly the same weight to every link. At a budget of about one million images, UCT lowers the one-step FID of ECT from the data-matched start on CIFAR-10 and ImageNet-64 and from pure noise on CIFAR-10, where the improvement remains after 100,000 updates.

Reproducibility Statement

Appendix A gives complete proofs, and Appendix B gives the training and evaluation protocols. The supplementary material contains our change to ECT, as a patch against the public ECT code, and the commands for the runs in the paper. It also contains the scripts that compute the data-matched start, FID, the distance to the pretrained model's ODE endpoints and the chain-model numbers. Finally, it contains Lean 4 proofs, checked against Mathlib, of Theorems 1–3 and Proposition 4, where the Lean proof of Theorem 2 takes as given that the probability-flow ODE carries $\rho_{T'}$ to ρ_ϵ.

Ethics Statement

This work tunes pretrained image generators on the public research datasets CIFAR-10 and ImageNet, used under their licences. The tuned models inherit the risks of the pretrained models they start from.

AI Use Statement

Large language model assistants were used in three ways. *Writing:* to edit and polish the text, captions and tables. *Retrieval and discovery:* to search for and summarise related work; every cited work was read and checked against its source before citation. *Research ideation and execution:* to brainstorm and check derivations, to write and review training, evaluation and plotting code, and to manage cluster runs; every reported number comes from logged runs. All claims, proofs, numbers and code were verified by the authors with independent checks before inclusion, and the authors take full responsibility for the final content.

References

Luca Ambrogioni, Giulio Franzese, Alberto Foresti, Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Naoki Murata, Chieh-Hsin Lai, and Yuki Mitsufuji. From atoms to entropy: Optimal noise allocation for diffusion training in the convex regime. *arXiv preprint arXiv:2607.20540*, 2026.

Jiayu Bai, Zhanbo Feng, Zhijie Deng, Tianqi Hou, Robert C. Qiu, and Zenan Ling. Adaptive discretization for consistency models. In *Advances in Neural Information Processing Systems 38*, 2025.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025.

Junsong Chen, Shuchen Xue, Yuyang Zhao, Jincheng Yu, Sayak Paul, Junyu Chen, Han Cai, Song Han, and Enze Xie. SANA-Sprint: One-step diffusion with continuous-time consistency distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16185–16195, 2025a.

Yiding Chen, Yiyi Zhang, Owen Oertell, and Wen Sun. Convergence of consistency model with multistep sampling under general data assumptions. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9764–9782, 2025b.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2204.00227.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems 38*, 2025a. URL https://arxiv.org/abs/2505.13447.

Zhengyang Geng, Ashwini Pokle, Weijian Luo, Justin Lin, and J. Zico Kolter. Consistency models made easy. In *International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=xQVxo9dSID.

Mahmut S. Gokmen, Cody Bumgardner, Jie Zhang, Ge Wang, and Jin Chen. High noise scheduling is a must. *arXiv preprint arXiv:2404.06353*, 2024.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via Min-SNR weighting strategy. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Tiankai Hang, Shuyang Gu, Jianmin Bao, Fangyun Wei, Dong Chen, Xin Geng, and Baining Guo. Improved noise schedule for diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4796–4806, 2025.

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 30*, 2017.

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

Zheyuan Hu, Chieh-Hsin Lai, Yuki Mitsufuji, and Stefano Ermon. CMT: Mid-training for efficient learning of consistency, mean flow, and flow-map models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2B8GkGTgmY.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a conditional prior distribution for flow-based generative models. *Transactions on Machine Learning Research*, 2025.

Thibaut Issenhuth, Sangchul Lee, Ludovic Dos Santos, Jean-Yves Franceschi, Chansoo Kim, and Alain Rakotomamonjy. Improving consistency models with generator-augmented flows. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 26586–26610, 2025. URL https://proceedings.mlr.press/v267/issenhuth25a.html.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35*, 2022.

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

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *International Conference on Learning Representations*, 2024.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems 34*, 2021.

Fei Kong, Jinhao Duan, Lichao Sun, Hao Cheng, Renjing Xu, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. ACT-Diffusion: Efficient adversarial consistency training for one-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8890–8899, 2024.

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

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *International Conference on Learning Representations*, 2022.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, volume 37, pp. 63082–63109, 2024.

Sangyun Lee, Yilun Xu, Tomas Geffner, Giulia Fanti, Karsten Kreis, Arash Vahdat, and Weili Nie. Truncated consistency models. In *International Conference on Learning Representations*, 2025.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2022.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1621–1629, 2025.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*, 2020.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. InstaFlow: One step is enough for high-quality diffusion-based text-to-image generation. In *International Conference on Learning Representations*, 2024. URL https://iclr.cc/virtual/2024/poster/19575.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025.

Junlong Lyu, Zhitang Chen, and Shoubo Feng. Sampling is as easy as keeping the consistency: convergence guarantee for consistency models. In *Proceedings of the 41st International Conference on Machine Learning, PMLR 235*, pp. 33664–33685, 2024.

J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems 36*, 2023.

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

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *International Conference on Learning Representations*, 2024.

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*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

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

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Noël Vouitsis, Rasa Hosseinzadeh, Brendan Leigh Ross, Valentin Villecroze, Satya Krishna Gorti, Jesse C. Cresswell, and Gabriel Loaiza-Ganem. Inconsistencies in consistency models: Better ODE solving does not imply better samples. In *NeurIPS 2024 Workshop on Attributing Model Behavior at Scale (ATTRIB)*, 2024. Workshop paper.

Binxu Wang and John J. Vastola. The unreasonable effectiveness of Gaussian score approximation for diffusion models and its applications. *Transactions on Machine Learning Research*, 2024.

Fu-Yun Wang, Zhengyang Geng, and Hongsheng Li. Stable consistency tuning: Understanding and improving consistency models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy (DeLTa)*, 2025. Workshop paper.

Jian Xu, Wei Chen, Shigui Li, Delu Zeng, John Paisley, and Qibin Zhao. Consist-Retinex: One-step noise-emphasized consistency training accelerates high-quality Retinex enhancement. *arXiv preprint arXiv:2512.08982*, 2025.

Ruofeng Yang, Bo Jiang, Cheng Chen, and Shuai Li. Improved discretization complexity analysis of consistency models: Variance exploding forward process and decay discretization scheme. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70756–70775, 2025.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. Simple and fast distillation of diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 40831–40860, 2024.

A PROOFS

Relation to earlier results. We obtain Theorem 2 by adapting the known telescoping argument of Geng et al. (2025b, Section 3.1) and the bound of Kong et al. (2024, Theorem 3.1) to the data-matched start, and we prove it in the form we use. The adaptation replaces the model's Lipschitz constant by that of the ideal map and the consistency losses by the consistency gaps on the links below T', which makes the start term independent of training and leaves only levels at or below T' in every other term. Recursions of the same kind, unrolled from pure noise, underlie the other guarantees cited in Section 2. Theorem 1 proves, for the start level, the failure that Lee et al. (2025, Section 3) prevent with a point mass at their dividing level: a training path that gives a level no mass cannot enforce consistency there. Theorem 3 casts the chain model as the pure-death process of the walker in Appendix A.3, whose standard properties (Norris, 1997, Chapter 2) give parts (i)–(iii); the new element is the rate (6), through which the training path sets the speed of each link, and the resistance that follows from it. The chain model keeps the premise of iCT that errors at lower levels carry over to higher ones (Song & Dhariwal, 2024, Sections 3.1 and 3.5), and it is a simple case of the temporal-difference view of Wang et al. (2025). Because the lower links are trained fastest, the least-trained upper links set the speed, as the least-visited state-action pairs do in asynchronous Q-learning (Li et al., 2022). Proposition 4 carries the equal training of every link, which the original consistency model obtains by drawing the link index uniformly (Song et al., 2023), over to ECT's pairing with the loss divided by the gap as in iCT and ECT, and it bounds how far each link's weight can deviate from equal. That the training path sets the speed of training rather than its end point was observed by Kingma et al. (2021) and Kingma & Gao (2023) for the noise schedule of diffusion losses; the chain model makes this quantitative for consistency tuning through the resistance, and it holds there because, under (M1), every link can reach zero loss. The data-matched start is taken from Raya & Ambrogioni (2023, Section 5) and Wang & Vastola (2024), and its class-conditional form from Lee et al. (2022) and Issachar et al. (2025).

Assumption (A2) asks that $\tau : (0,\infty) \to (0,\infty)$ is continuous and strictly increasing, with $0 < \tau(t) < t$ and $\tau(t) \to \infty$ as $t \to \infty$. Let $\Delta(t) = t - \tau(t)$. ECT's pairing $\tau(t) = t\{1 - (1 + 8\sigma(-t))/256\}$, with $\sigma(u) = 1/(1+e^{-u})$, satisfies (A2), because $251t/256 < \tau(t) < t$ and $\tau'(t) = 1 - \Delta'(t) \geq 1 - 5/256$ for $t > 0$. By (A1), the data $x_0 \sim \rho$ have a finite second moment, mean m and covariance Σ, and the noise $z \sim \mathcal{N}(0, I_d)$ is independent of x_0. As in the main text, ρ_t is the distribution of $x_0 + tz$, $\gamma_t = \mathcal{N}(m, \Sigma + t^2 I)$, $\|X\| = (\mathbb{E}\|X\|_2^2)^{1/2}$ for a random vector X, and $F_{\#}\nu$ is the distribution of $F(X)$ for $X \sim \nu$.

A.1 PROOF OF THEOREM 1

The stop-gradient in (3) does not change the value of the loss, and we omit it. Let

$$\mathcal{L}_1(f) = \int \frac{\mathbb{E}\|f(x_0 + tz, t) - f(x_0 + \tau(t)z, \tau(t))\|_2}{\Delta(t)}\, p(dt),$$

and let $\mathcal{L}_2$ be the same expression with the squared norm in the numerator.

Theorem (Theorem 1 in full)**.** *Assume (A2). Let p be a training path on $(0,\infty)$ with $p(\{T'\}) = p(\{\tau^{-1}(T')\}) = 0$, and let $\rho = \delta_y$ for a point y. Then $\gamma_{T'} = \rho_{T'} = \mathcal{N}(y, T'^2 I)$. For every unit vector v and every $\alpha \neq 0$ there are smooth maps f_n with $f_n(x,0) = x$ and spatial Lipschitz constant at most 1 such that $\mathcal{L}_1(f_n) \to 0$ and $\mathcal{L}_2(f_n) \to 0$ as $n \to \infty$, whilst $W_2\big(f_n(\cdot, T')_{\#}\gamma_{T'}, \rho\big) = |\alpha|$ for every n.*

Proof. The maps send every start at T' to $y + \alpha v$ and change with the level only in a band near 0 and a band near T', whose contribution to the loss vanishes as the bands shrink. Let $\chi(u) = \exp\{-u^2/(1-u^2)\}$ for $|u| < 1$ and $\chi(u) = 0$ otherwise. This is a smooth bump with $0 \leq \chi \leq 1$ and $\chi(0) = 1$. For $0 < \zeta_0, \zeta_1 < T'$ put

$$f(x,t) = y + \chi(t/\zeta_0)\,(x - y) + \alpha\,\chi\big((t - T')/\zeta_1\big)\,v.$$

Because $\chi(0) = 1$ and $\chi(-T'/\zeta_1) = 0$, we have $f(x,0) = x$, and $x \mapsto f(x,t)$ is Lipschitz with constant $\chi(t/\zeta_0) \leq 1$. At $t = T'$ the second term vanishes because $T'/\zeta_0 > 1$, and the last term equals αv. The generator therefore sends every start to $y + \alpha v$, and $W_2(\delta_{y+\alpha v}, \delta_y) = |\alpha|$.

At the input $y + tz$, the second term of f equals $\beta(t)z$ with $\beta(t) = t\,\chi(t/\zeta_0)$. For $0 < t < \zeta_0$ put $u = t/\zeta_0$ and $r = u^2/(1-u^2)$. Then $\beta'(t) = \chi(u) + u\chi'(u)$ and $u\chi'(u) = -2u^2(1-u^2)^{-2}\chi(u) = -2r(1+r)e^{-r}$, which gives $\beta'(t) = e^{-r}(1 - 2r - 2r^2)$. Because $re^{-r} \le 1/e$ and $r^2e^{-r} \le 4/e^2$, we get $|\beta'| \le 1 + 2/e + 8/e^2 < 3$, and $\beta = 0$ for $t \ge \zeta_0$. The second term therefore differs between the two levels of a pair only if $\tau(t) < \zeta_0$, and then, by the mean value theorem, by at most $3\Delta(t)\|z\|_2$. Its contribution to $\mathcal{L}_1$ is at most $3\,\mathbb{E}\|z\|_2\; p\{t : \tau(t) < \zeta_0\}$. In $\mathcal{L}_2$ one factor $\Delta(t)$ remains, and $\Delta(t) < t < \tau^{-1}(\zeta_0)$ when $\tau(t) < \zeta_0$; with $\mathbb{E}\|z\|_2^2 = d$, the contribution to $\mathcal{L}_2$ is at most $9d\,\tau^{-1}(\zeta_0)\; p\{t : \tau(t) < \zeta_0\}$. As $\zeta_0 \to 0$, $\tau^{-1}(\zeta_0)$ decreases to 0 and the sets $\{t : \tau(t) < \zeta_0\} = (0, \tau^{-1}(\zeta_0))$ decrease to the empty set. Both bounds therefore tend to 0.

The last term differs between the two levels of a pair only if t or $\tau(t)$ lies in $S = (T' - \zeta_1, T' + \zeta_1)$, and then by at most $|\alpha|$. With $\mu(dt) = p(dt)/\Delta(t)$, its contributions to $\mathcal{L}_1$ and $\mathcal{L}_2$ are at most $|\alpha|\{\mu(S) + \mu(\tau^{-1}(S))\}$ and $\alpha^2\{\mu(S) + \mu(\tau^{-1}(S))\}$. Because Δ is continuous and positive, it is bounded below by a positive constant near each of the points T' and $\tau^{-1}(T')$, and μ is finite there. As $\zeta_1 \to 0$, S and $\tau^{-1}(S)$ decrease to $\{T'\}$ and $\{\tau^{-1}(T')\}$, which have zero p-mass. Hence $\mu(S)$ and $\mu(\tau^{-1}(S))$ tend to 0.

By the triangle inequality, $\mathcal{L}_1(f)$ is at most the sum of the two contributions to $\mathcal{L}_1$, and because the squared norm of a sum of two vectors is at most twice the sum of their squared norms, $\mathcal{L}_2(f)$ is at most twice the sum of the two contributions to $\mathcal{L}_2$. Taking $f_n = f$ with $\zeta_0 = \zeta_1 = \min\{1/n, T'/2\}$ completes the proof. □

A.2 Proof of Theorem 2

Fix a level $0 < \epsilon \le t_N$. Assumption (A3) asks that the probability-flow ODE (1) of the data (Song et al., 2021; Karras et al., 2022) has a unique solution on $[\epsilon, T']$ from every point. This holds when ρ has bounded support, because the velocity is then continuous in t and Lipschitz in x, uniformly for $t \in [\epsilon, T']$. Let $f^*(\cdot, t)$ send a point at level t to the end of its path at level ϵ. Because the probability-flow ODE preserves the marginals ρ_t (Song et al., 2021), $f^*(\cdot, T')_{\#}\rho_{T'} = \rho_\epsilon$. Let $T' = t_0 > t_1 > \cdots > t_N \ge \epsilon$ be the chain of Section 4, let $V_0 \sim \gamma_{T'}$, and let $V_1, \ldots, V_N$ be random vectors on the same probability space such that all $f(V_j, t_j)$ and $f^*(V_j, t_j)$ are square-integrable.

Theorem (Theorem 2 in full)**.** *Assume (A1)–(A3). Let $D_{T'} = W_2\big(f^*(\cdot, T')_{\#}\gamma_{T'}, \rho_\epsilon\big)$, the error of the ideal map from the data-matched start. With $e_j = f(V_j, t_j) - f^*(V_j, t_j)$ and $d_j = e_j - e_{j+1}$,*

$$W_2\big(f(\cdot, T')_{\#}\gamma_{T'}, \rho_\epsilon\big) \;\le\; D_{T'} + \|e_0\| \;\le\; D_{T'} + \|e_N\| + \sum_{j<N} \|d_j\|. \tag{8}$$

If $f^*(\cdot, T')$ is $L_{T'}$-Lipschitz, then $D_{T'} \le L_{T'}\, W_2(\rho_{T'}, \gamma_{T'})$, which gives (5).

Proof. By the triangle inequality for W_2,

$$W_2\big(f(\cdot, T')_{\#}\gamma_{T'}, \rho_\epsilon\big) \le W_2\big(f(\cdot, T')_{\#}\gamma_{T'}, f^*(\cdot, T')_{\#}\gamma_{T'}\big) + D_{T'}.$$

The pair $(f(V_0, T'), f^*(V_0, T'))$ couples the two distributions in the first term on the right, which is therefore at most $\|e_0\|$. This proves the first inequality in (8). Summing $d_j = e_j - e_{j+1}$ over $j < N$ gives $e_0 = e_N + \sum_{j<N} d_j$, and Minkowski's inequality in L^2 proves the second. This is the L^2 form of ECT's telescoping bound (Geng et al., 2025b, Section 3.1, Eq. 13). If $f^*(\cdot, T')$ is $L_{T'}$-Lipschitz, it maps each coupling (V, X) of $\gamma_{T'}$ and $\rho_{T'}$ to a coupling of $f^*(\cdot, T')_{\#}\gamma_{T'}$ and ρ_ϵ with $\|f^*(V, T') - f^*(X, T')\| \le L_{T'}\|V - X\|$. Taking the infimum over couplings gives $D_{T'} \le L_{T'} W_2(\rho_{T'}, \gamma_{T'})$. Because $W_2(\rho_\epsilon, \rho) \le \epsilon\sqrt{d}$, the same bound compares the generator with the data up to the term $\epsilon\sqrt{d}$. □

The start mismatch does not grow with T'. Let $\delta_t = W_2(\rho_t, \gamma_t)$ and $t \ge r > 0$. Add the same independent $\mathcal{N}(0, (t^2 - r^2)I)$ to both members of an optimal coupling of ρ_r and γ_r. Because independent Gaussian variances add, the sums have distributions ρ_t and γ_t, and their difference is unchanged. Hence $\delta_t \le \delta_r$. This says nothing about $L_{T'}$, which can grow with T'.

The pretrained model in place of the data. Let Φ send a point at level T' to the end of its path at level ϵ, under the pretrained model's probability-flow ODE, which we also assume has a unique

solution from every point, and assume that $\Phi(V_0)$ is square-integrable. The first inequality in (8) holds with Φ in place of $f^*(\cdot, T')$, by the same argument:

$$W_2\big(f(\cdot,T')_{\#}\gamma_{T'},\rho_\epsilon\big) \le W_2\big(\Phi_{\#}\gamma_{T'},\rho_\epsilon\big) + \|f(V_0,T') - \Phi(V_0)\|.$$

The first term is the pretrained model's own error from the same start, which tuning does not change. The second is the distance measured in Section 5.3. Let V_j be the point at level t_j on the pretrained model's path from V_0. Telescoping gives $f(V_0,T') - \Phi(V_0) = f(V_N,t_N) - \Phi(V_0) + \sum_{j<N}\{f(V_j,t_j) - f(V_{j+1},t_{j+1})\}$. All V_j have the same endpoint $\Phi(V_0)$, and $f(V_N,t_N) - \Phi(V_0)$ is the tuned model's deviation at level t_N from the pretrained model's map. By Minkowski's inequality, the distance is at most this deviation plus the sizes of the model's changes across the links.

What the loss measures. Consistency training compares two levels of the same image with the same noise, $x_0 + tz$ and $x_0 + rz$ (Song et al., 2023), and these two points need not be on one probability-flow path. For $x_0, z \sim \mathcal{N}(0,1)$ the ideal map to the data at level 0 is $f^*(x,t) = x/\sqrt{1+t^2}$, and for $r \neq t$

$$\mathbb{E}|f^*(x_0+tz,t) - f^*(x_0+rz,r)|^2 = 2 - \frac{2(1+tr)}{\sqrt{(1+t^2)(1+r^2)}} > 0,$$

by the Cauchy–Schwarz inequality for the vectors $(1,t)$ and $(1,r)$, which are not parallel. Even the ideal map has a positive loss on a training pair, and the loss therefore does not measure the gaps d_j directly. Issenhuth et al. (2025) show that the resulting discrepancy between consistency training and distillation remains in the continuous-time limit.

A.3 Proof of Theorem 3

Take the notation of Theorem 3 and put $u_k = q_k(0) - q_{k-1}(0) \ge 0$. By definition, $H_{k-1} = \sum_{i=k}^{N} E_i/\lambda_i$.

Proof. Consider a walker that starts at level N, waits at each level $i \ge 1$ for the time E_i/λ_i, then steps down to $i-1$, and stops at 0. Let J_s be its level at time s. For a walker started at level i, let $v_i(s) = \mathbb{E}^{(i)}[q_{J_s}(0)]$; then $v_0 \equiv 0$ and $v_i(0) = q_i(0)$. Conditioning on the first wait gives $v_i(s) = e^{-\lambda_i s}q_i(0) + \int_0^s \lambda_i e^{-\lambda_i r} v_{i-1}(s-r)\,dr$, and differentiating in s gives $\dot v_i = -\lambda_i(v_i - v_{i-1})$. The q_i solve the same linear system with the same initial values, and $q_i = v_i$ by uniqueness. In particular, $q_N(s) = \mathbb{E}^{(N)}[q_{J_s}(0)]$.

(i) Because $q_i(0) = \sum_{k\le i} u_k$, we have $q_{J_s}(0) = \sum_k u_k \mathbf{1}\{J_s \ge k\}$. The walker reaches level $k-1$ at time H_{k-1}, and $\{J_s \ge k\} = \{H_{k-1} > s\}$. Taking expectations gives (i).

(ii) Build both walkers from the same E_i. Then $\tilde H_{k-1} = \sum_{i\ge k} E_i/\tilde\lambda_i \le H_{k-1}$ for every k, and $\mathbb{P}(\tilde H_{k-1} > s) \le \mathbb{P}(H_{k-1} > s)$. Because $u_k \ge 0$, (i) gives $0 \le \tilde q_N(s) \le q_N(s)$.

(iii) By Markov's inequality, $\mathbb{P}(H_{k-1} > s) \le \mathbb{E}H_{k-1}/s = s^{-1}\sum_{i\ge k} 1/\lambda_i$, and (i) gives $q_N(s) \le s^{-1}\sum_k u_k \sum_{i\ge k} 1/\lambda_i$. The upper bound in (7) follows from $\sum_{i\ge k} 1/\lambda_i \le \sum_i 1/\lambda_i$ and $\sum_k u_k = q_N(0)$. For the lower bound, $H_{k-1} \ge E_N/\lambda_N$ for every k. Hence $\mathbb{P}(H_{k-1} > s) \ge \mathbb{P}(E_N > \lambda_N s) = e^{-\lambda_N s}$, and (i) gives $q_N(s) \ge q_N(0)e^{-\lambda_N s}$. □

The chain model for $T' = 5$. With ECT's pairing, the chain from $T' = 5$ has 585 links down to $t_N = 0.00203$, the last level not below $a = 0.002$. Under ECT's lognormal, the rate of the top link is 0.0160, and the largest rate is 4.47, near $t = 0.0061$. The 246 links that reach above $t = 1$ carry 94.9% of the resistance R. The training path of UCT raises the top rate to 0.0621 and gives $R/R_{\mathrm{ECT}} = 0.464$. It lowers the rates of the 330 links below $t = 0.887$, by at most 23.9%. For a point mass, the mass divided by the gap at its level is added to the rate of the link $(t_{j+1}, t_j]$ that contains it, and mass outside $[t_N, T']$ is not counted. The resistances in Tables 2 and 3 are computed in this way.

Numerical comparison with ECT. Because UCT lowers some rates, part (ii) does not compare it with ECT, and part (iii) compares only upper bounds. We computed the distribution of the walker's level for both training paths by integrating its forward equation with an implicit Runge–Kutta method (Radau, relative tolerance 10^{-10}, absolute tolerance 10^{-14}) at 3250 log-spaced training times from

$s = 1$ to $s = 10^6$. At each of these times and for every level k, $\mathbb{P}_{\mathrm{UCT}}(J_s \geq k)$ exceeds $\mathbb{P}_{\mathrm{ECT}}(J_s \geq k)$ by at most 2×10^{-15}, far below the solver's error of about 10^{-12}, measured against the matrix exponential. By (i), the error of UCT is then at most ECT's for every ordered initial error. The supplementary material includes the script that computes all numbers of this subsection.

A.4 Proof of Proposition 4

Proof. On link j write $t = t_j - u$ with $0 \leq u \leq \Delta_j = \Delta(t_j)$; the link is $[t_j - \Delta_j, t_j]$. By the mean value theorem, $\Delta_j - \kappa_+ u \leq \Delta(t_j - u) \leq \Delta_j - \kappa_- u$. Both bounds are positive because $u \leq \Delta_j$ and $\kappa_- \leq \kappa_+ < 1$. Substituting $u = \Delta_j v$,

$$\varphi(\kappa_-) = \int_0^1 \frac{dv}{1 - \kappa_- v} \;\leq\; \int_{t_{j+1}}^{t_j} \frac{dt}{\Delta(t)} \;\leq\; \int_0^1 \frac{dv}{1 - \kappa_+ v} = \varphi(\kappa_+),$$

and the link's width cancels. The link lies in $[a, T']$, where $p = 1/(T' - a)$, and the middle integral therefore equals $(T' - a)\, w_j$. For $\tau(t) = ct$ the slope Δ' equals $1 - c$ everywhere, and both bounds equal $\log(1/c)/(1 - c)$ on every link. □

For ECT's pairing, $\Delta(t) = t(1 + 8\omega)/256$ with $\omega = \sigma(-t)$, and $\frac{d}{dt}\sigma(-t) = -\omega(1 - \omega)$ gives $\Delta'(t) = \{1 + 8\omega - 8t\omega(1 - \omega)\}/256$. On $[a, 5]$ this slope has its minimum near $t = 2.40$ and its maximum at $t = a$, and it lies between $\kappa_- = 0.000786$ and $\kappa_+ = 0.019501$. With these values, the proposition gives $1.00039 < (T' - a)\, w_j < 1.00988$ on each of the 585 links below $T' = 5$.

B Experimental Details

B.1 Training

We use the public ECT code with its unconditional CIFAR-10 recipe. It starts from the EDM checkpoint (Karras et al., 2022) and uses RAdam with learning rate 10^{-4}, batch size 128, dropout 0.2 and no augmentation, and it evaluates an EMA of the weights with decay 0.9993. All 8000 updates lie in the first stage of ECT's schedule, where the lower level of a pair is

$$\tau(t) = t\Big(1 - \frac{1 + 8\sigma(-t)}{256}\Big), \qquad \sigma(u) = \frac{1}{1 + e^{-u}}. \tag{9}$$

ECT draws $\log t$ from $\mathcal{N}(-1.1, 2^2)$. UCT keeps this draw and, with probability $\pi = 0.25$, replaces it by a uniform draw on $[0.002, T']$. The coin flip and the uniform draw come from a separate random stream, and within a seed ECT and UCT therefore share the data order and ECT's draws. Seeds 0–4 share this separate stream, and seed 5 uses it shifted by 6000 draws. On RTX A6000 GPUs, the median update takes 0.883 s for ECT and 0.876 s for UCT. The 100,000-update pair uses ECT's published recipe for this length with seed 0, and both of its runs trained on an RTX A6000.

Most runs were trained on RTX A6000 GPUs, and the two methods of a seed always used the same GPU type. The exceptions are seed 4 of both methods and six variants of Table 2 (the log-uniform draws on $[0.3, 5]$ and on $[a, 80]$, the density $\propto \Delta(t)^{-1/2}$, the log-Student-t draws, the point mass at 5 and the variant that redraws ECT's draws above 5), trained on A100 GPUs (for seed 4, ECT on an 80 GB and UCT on a 40 GB model); the matched lognormal and the range-end runs, trained on V100 GPUs; and seed 2 of the shifted lognormal with the range, trained on an RTX A6000 up to update 6000 and on an A100 after that. ECT's level draws, the noise and the dropout masks come from the GPU's main random stream, whose values depend on the GPU type. Runs on different GPU types therefore share the data order and the separate stream's draws (checked on a V100), but not the main stream's draws. All CIFAR-10 evaluations use the same recorded evaluation noise, and the final models were scored on RTX A6000 GPUs. For seed 0, UCT trained on a V100 has one-step FID-50k 4.91 at $T' = 5$, against 4.78 when trained on an RTX A6000.

B.2 Start and Evaluation

The start uses the mean and the unbiased covariance estimate of the 50,000 training images, with pixel values scaled to $[-1, 1]$, and we compute the square root in (2) by an eigendecomposition.

Among the start levels 2.5, 5, 10, 20 and 80, $T' = 5$ gives the lowest FID-10k of ECT in each of seeds 0–4. FID uses ECT's evaluation code with reference statistics from the 50,000 training images. FID-50k scores the final models at $T' = 5$ and from pure noise. FID-10k scores the other start levels, the range-end runs and the training curves, which are evaluated at updates $1001, 1501, \ldots, 7501$ and 8000. At update 1001 the mean FID-10k at $T' = 5$ over the six seeds is 86.8 for ECT and 82.5 for UCT. UCT is lower in every seed at each of the 15 checkpoints. The Fréchet distance in DINOv2 features of Table 7 follows Stein et al. (2023) and uses the same 50,000 samples as FID-50k, with the 50,000 training images as reference. For the start from noised training images, we draw 10,000 training images with replacement, add noise of level T' and apply the network once. To check our run of ECT against the result reported for this budget, we use ECT's two-step sampler, which adds fresh noise to the one-step output at level 0.821 and applies the network once more. With this sampler, our run of ECT reproduces the published result: its FID-50k lies between 2.72 and 2.86 over the six seeds, against the two-step FID of 2.73 reported by Geng et al. (2025b).

B.3 Training paths

The log-Student-t variant draws $t = \exp(0.2S)$ conditional on $t \in [a, 80]$, where S has a Student-t distribution with 0.01 degrees of freedom. TCM uses the same draw, truncated to $[1, 80]$, in its second stage, together with a point mass at $t = 1$ (Lee et al., 2025, Section 4.4 and Eq. 7). The shifted lognormal replaces ECT's log-mean -1.1 by 0.02, the mean log-level of the variant with a uniform draw on $[a, 80]$. The matched lognormal has log-mean -0.255 and log-standard deviation 1.236. Like UCT, it puts 58.2% of its draws below 1, 35.3% between 1 and 5 and 6.6% above 5. The variant with half of the range on $[a, 0.821]$ draws its uniform level on $[a, 5]$ or on $[a, 0.821]$ with equal probability. The point masses at 80 and 0.821 share their quarter of the draws equally. The variant that redraws ECT's draws above 5 repeats each such draw until the level is at most 5, and then applies the range as UCT does.

Table 3: **Where the range ends.** Ranges $0.75\, p_{\mathrm{ECT}} + 0.25\, \mathcal{U}[a, T_r]$ with end T_r, evaluated from starts T'. Left: chain resistance relative to ECT's on the chain below T'. Right: one-step FID-10k, mean over seeds 0–2 (lower is better). Bold: the best value for each start. These runs were trained on V100 GPUs and are compared only with each other.

	Resistance at start T'			FID-10k at start T'		
Range end T_r	2.5	5	10	2.5	5	10
2.5	**0.60**	1.17	1.29	**9.07**	6.88	7.88
5 (UCT)	0.80	**0.46**	1.11	9.30	**6.66**	7.20
10	0.99	0.66	**0.36**	9.59	6.90	**7.12**

B.4 Other one-step methods at our budget

We trained TCM (Lee et al., 2025), consistency distillation (CD; Song et al., 2023), continuous-time consistency training (sCT; Lu & Song, 2025) and MeanFlow (Geng et al., 2025a) on unconditional CIFAR-10 for 8000 updates at batch size 128, with three seeds each. TCM and MeanFlow use their official code. CD uses a PyTorch implementation whose optimiser and EMA updates match the official JAX code in a 12-step test. sCT has no public code, and we implemented the algorithm in its paper. TCM's first stage, CD and sCT start from the pretrained EDM model that UCT starts from, and CD also uses it as its teacher. MeanFlow trains from scratch, as in its paper. The rows marked as published use each paper's settings with our budget and batch, with 4000 updates in each of TCM's two stages. These settings are written for much longer runs, and the compressed rows change at most two of them. TCM's schedule for the gap between paired levels is compressed into each stage, the warmups of sCT and MeanFlow take 800 updates, and the compressed rows of TCM, CD and MeanFlow use our EMA decay of 0.9993. Each model is scored from pure noise with its own one-step sampler. TCM and CD, which share ECT's noise parameterisation, are also scored from the data-matched start at $T' = 5$. Every model was scored on an RTX A6000. Table 4 gives the results. At their own budgets, the four papers use 200 to 800 times as many training images as our runs and report one-step FID between 2.46 and 3.55 from pure noise (Table 5). Our evaluation

code scores TCM's released CIFAR-10 checkpoint at FID-50k 2.51, against the 2.46 reported by its authors.

Table 4: Other one-step methods trained on unconditional CIFAR-10 at our budget (8000 updates, batch size 128): one-step FID-50k, mean over seeds 0–2 with the three seeds in parentheses (lower is better). A dash marks a start that does not apply.

Method	Settings	Pure noise	Data-matched start, $T' = 5$
TCM	As published	283.92 (283.95, 283.90, 283.91)	135.50 (135.63, 135.38, 135.49)
	Compressed	10.40 (8.18, 7.92, 15.10)	7.92 (6.92, 7.15, 9.70)
CD	As published	170.48 (204.34, 146.41, 160.68)	21.87 (22.22, 21.01, 22.39)
	Compressed	9.41 (9.41, 9.70, 9.13)	10.35 (10.30, 10.70, 10.06)
sCT	As published	60.26 (61.13, 58.94, 60.71)	–
	Compressed	31.27 (31.21, 30.89, 31.73)	–
MeanFlow	As published	372.43 (372.13, 372.43, 372.72)	–
	Compressed	60.12 (59.13, 62.00, 59.22)	–
UCT	–	**6.25** (6.38, 6.02, 6.35)	**4.79** (4.78, 4.73, 4.84)

Table 5: Published one-step results on unconditional CIFAR-10 at the methods' own budgets, as reported in each paper (FID from pure noise; lower is better). Training images: updates $\times$ batch size. The sCT paper gives neither the number of updates nor the batch size for this result, and its number of training images is taken from Bai et al. (2025). Its FID is from the final version of the paper, and an earlier version gives 2.97.

Method	Updates	Batch size	Training images	One-step FID
CD (Song et al., 2023)	800k	512	409.6M	3.55
iCT (Song & Dhariwal, 2024)	400k	1024	409.6M	2.83
sCT (Lu & Song, 2025)	–	–	204.8M	2.85
TCM (Lee et al., 2025)	250k + 200k	512, 1024	332.8M	2.46
MeanFlow (Geng et al., 2025a)	800k	1024	819.2M	2.92
ECT (Geng et al., 2025b)	100k	128	12.8M	4.54
	200k	128	25.6M	3.86
	400k	128	51.2M	3.60
ADCM (Bai et al., 2025)	100k	128	12.8M	3.16
	75k	1024	76.8M	2.80
UCT, our budget	8k	128	1.02M	6.34

B.5 What ranks the generators

We use the final EMA checkpoints (8000 updates) of the runs in Tables 6 and 2 that were trained on RTX A6000 or A100 GPUs, and of an earlier run of ECT, 36 in all. The runs on V100 GPUs were trained after these distances were computed. All checkpoints share 2000 starts at $T' = 5$. For each start, we solve the pretrained model's probability-flow ODE from T' to 0.002 with Heun's method (Karras et al., 2022), with one step between consecutive levels of the grid $\ell_0 = T'$, $\ell_{k+1} = \max(0.9\,\ell_k, 0.002)$. On the first 200 starts, halving the step size changes the endpoints by 0.23% of their root-mean-square norm. The distance is the root-mean-square Euclidean distance between a checkpoint's one-step outputs and these endpoints. The averaged loss is ECT's loss (3) under ECT's lognormal, evaluated as in Section 5.3 with the EMA weights and without dropout. The resistance of each run's training path is computed with the chain model of Appendix A.3. We report Spearman correlations (Spearman, 1904) with FID-50k at $T' = 5$ and percentile intervals from a bootstrap (Efron, 1979) over checkpoints with 5000 resamples.

B.6 ImageNet-64

Both methods use ECT's ImageNet-64 code and settings and start from the EDM2-S checkpoint (Karras et al., 2024). They train for 8000 updates at batch size 128 with seeds 0 and 1. ECT's lognormal on ImageNet-64 has log-mean -0.8 and log-standard deviation 1.6. UCT replaces ECT's

draw of the level by a uniform draw on $[0.002, T']$ with probability 0.25, using a separate random stream. With this probability set to 0, the code computes exactly ECT's loss. The code keeps three EMA profiles of the power-function form of Karras et al. (2024, Section 3), and we chose the one of relative width 0.050 before computing any FID. On the seed-0 model of ECT we tried the start levels 2.5, 5, 10 and 20. FID-50k uses $50{,}000$ samples with uniform labels and reference statistics from the full training set. With these statistics the pretrained model reaches FID-10k 4.49 with EDM2's deterministic sampler, the same value as with EDM2's statistics. Both methods of each comparison were scored on the same GPU type with the same recorded evaluation noise. For the longer run, both seed-0 models continue from 8000 updates under ECT's full $100{,}000$-update schedule, and we report the checkpoints at $48{,}000$ and $100{,}000$ updates. The continuation was stopped and resumed at the same updates for both methods, on several GPU types. At each restart the data order and, for UCT, the range draws continue exactly. ECT's level draws, the noise and the dropout masks restart from a new seed that is the same for both methods, but, as on CIFAR-10 (Appendix B.1), their values depend on the GPU type, and the two methods did not always run on the same GPU type. In both runs the stages of ECT's schedule changed after updates $25{,}050$, $50{,}050$ and $75{,}000$, instead of $25{,}001$, $50{,}001$ and $75{,}001$ in an uninterrupted run.

B.7 Per-seed results

Tables 6 and 7 give the per-seed values behind Table 1 and the DINOv2 comparison of Section 5.1.

Table 6: **One-step FID-50k on unconditional CIFAR-10 per training seed** after 8000 updates (lower is better).

	Data-matched start, $T' = 5$		Pure noise	
Seed	ECT	UCT	ECT	UCT
0	5.090	4.783	7.535	6.377
1	5.286	4.730	7.343	6.019
2	5.272	4.842	7.186	6.354
3	5.223	4.771	7.249	6.248
4	5.191	4.884	7.359	6.639
5	5.291	4.912	7.425	6.433

Table 7: **Fréchet distance in DINOv2 features** of one-step samples on unconditional CIFAR-10 after 8000 updates, per training seed (lower is better). Appendix B.3 defines the shifted lognormal.

	Data-matched start, $T' = 5$			Pure noise		
Training path	Seed 0	Seed 1	Seed 2	Seed 0	Seed 1	Seed 2
ECT	350.9	354.1	351.8	451.3	457.4	451.7
UCT	343.7	349.1	347.4	443.7	439.6	444.2
Shifted lognormal	356.3	357.3	358.2	408.7	419.6	418.8