PATHSPEED: ACCELERATING STRUCTURAL RECOVERY IN FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning a generative distribution requires recovering its modes, connected components, and loops. We study how training-time sampling can accelerate this recovery in flow matching. PathSpeed measures the motion of path marginals before velocity training and uses it to construct a fixed training-time distribution. We estimate squared path speed from adjacent marginal samples, then smooth and standardize the estimates. We show that this finite-difference measurement splits exactly into changes in projected means and scales and changes in standardized shape. In the continuous limit, squared path speed measures the energy of the conditional-average motion that transports the projected marginals. Integrated speed bounds changes in persistence diagrams at a fixed observation scale. On controlled distributions, PathSpeed reaches the final structural error of uniform time sampling (Uniform) with 33–49% fewer updates and achieves lower final errors. Ablations show that shape information and the placement of sampling probabilities along time both contribute. On CIFAR-10 and class-conditional latent ImageNet-256, PathSpeed obtains the lowest mean FID among the compared schedules and reaches Uniform's final FID with 42% and 52% fewer updates, respectively. The cached distribution adds no velocity-network forward or backward pass during training.

1 INTRODUCTION

A generative model can place samples in the correct region yet join separate components, misallocate mass across modes, or fill holes that should remain empty. The choice of training times changes how frequently the model learns each part of the probability path. We show that sampling according to marginal motion accelerates component and loop recovery with the path, velocity model, and optimizer held fixed.

For the linear paths used in flow matching, intermediate marginals can be sampled directly from endpoint pairs before training the velocity model (Lipman et al., 2023; Liu et al., 2023; Albergo et al., 2025). Each pair follows a constant-velocity interpolation. The marginal distribution evolves unevenly along the path, with changes in both the location of probability mass and its shape. Its conditional-average velocity describes the motion of this mass, and variation around that mean reflects uncertainty in endpoint pairing. The marginal motion to be learned is therefore already determined by the data and the path.

We introduce **PathSpeed**, an offline time sampler that allocates updates according to how rapidly the path marginals move. Fixed schedules specify the time distribution (Esser et al., 2024; Hang et al., 2024), and online methods such as InfoNoise adapt it using training statistics (Raya et al., 2026). The velocity-regression loss contains both model error and variation inherent in endpoint pairing. PathSpeed bases its allocation on the energy of projected marginal motion. We estimate this quantity from adjacent marginal samples using squared sliced-Wasserstein distances (Bonneel et al., 2015). Random projections and sorting make the estimate directly computable from samples. Smoothing, standardization, and softmax then produce a fixed training-time distribution (Figure 1). This construction preserves the fixed-time regression target, architecture, and inference procedure.

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

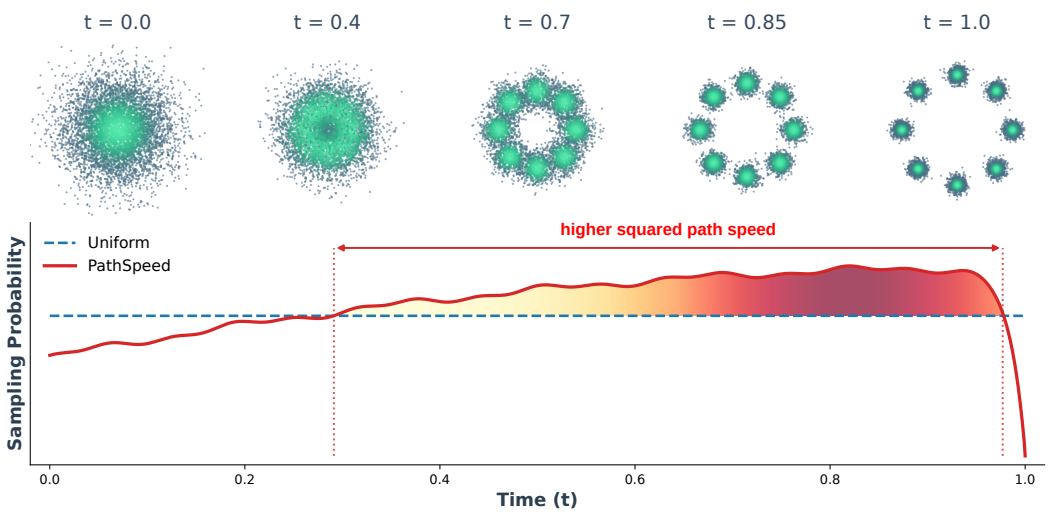


Figure 1: **Marginal motion guides the training-time distribution.** Top: marginals at five times on a multimodal FM path. Bottom: background color represents estimated squared path speed; the curves show the resulting PathSpeed time distribution and Uniform. The terminal drop reflects forward-difference padding (Section 4.1).

We decompose estimated squared path speed into moment and shape components using the location–scale–shape identity (Irpino & Verde, 2015). The former measures changes in projected means and scales; the latter captures evolution after centering and scaling each marginal. In the continuous limit, squared speed equals projected conditional-average velocity energy, and the shape component is the residual energy after its best affine approximation. We also bound persistence-diagram changes by integrated path speed at a fixed observation scale (Cohen-Steiner et al., 2007).

PathSpeed reaches Uniform's final structural error on Checkerboard, Gaussian25, and 2rings with 33–49% fewer updates and lower final errors (Table 1; Figures 2 and 3). On CIFAR-10, it lowers FID from InfoNoise's 5.92 to 4.61, a 22% reduction. On class-conditional latent ImageNet-256, PathSpeed-trained SiT-B/2 (Ma et al., 2024) obtains the lowest mean FID, 27.40, at 400k updates without classifier-free guidance. It reaches Uniform's final FID with an estimated 42% fewer updates on CIFAR-10 and 52% fewer on ImageNet-256 (Table 2). The component comparisons identify information that improves the allocation: adding the shape component lowers CIFAR-10 FID from 4.85 to 4.61. Permuting the same sampling probabilities across time removes the gain (Table 3).

Contributions. **(i)** We show that sampling training times according to marginal motion accelerates structural recovery and lowers final structural errors. **(ii)** We explain squared path speed through its moment and shape components, conditional-average motion, and persistence stability. **(iii)** We construct a fixed time distribution from this measurement and demonstrate its benefits on controlled distributions and natural images.

2 Related Work and Positioning

Structure formation along generative paths. Symmetry breaking and speciation describe how diffusion trajectories separate toward data regions (Raya & Ambrogioni, 2023; Biroli et al., 2024). Critical-window theory bounds feature emergence (Li & Chen, 2024). Hierarchical models exhibit sharp high-level and smoother low-level reconstruction transitions (Sclocchi et al., 2025). These results concern generation time; our question is how allocation over path time affects the training updates required for structural recovery.

Time sampling and loss weighting. Sampling frequency and loss weights jointly allocate training attention (Kingma & Gao, 2023). Min-SNR balances timestep objectives through clipped signal-to-noise ratios (Hang et al., 2023). Logit-normal sampling (Esser et al., 2024) and Laplace sampling

in log-SNR coordinates (Hang et al., 2024) emphasize intermediate levels; RF++ uses a U-shaped distribution motivated by endpoint losses in Reflow (Lee et al., 2024a).

Data- and model-dependent schedules. ANT selects schedules offline using autocorrelation statistics (Lee et al., 2024b). Improved DDPM uses online loss-second-moment sampling (Nichol & Dhariwal, 2021). CDCD fits a time-warping function to loss profiles (Dieleman et al., 2022). Our Loss-CDF comparator is a CDCD-inspired loss-profile sampler adapted to FM. Kim et al. (2025) instead estimate the improvement in the global training objective after updates. InfoNoise allocates samples from estimated conditional entropy rates (Raya et al., 2026). cDOL combines an irreducible-loss reference, calibrated weights, and remaining-gap sampling (Xu et al., 2026). PathSpeed constructs its time distribution from the motion of target-path marginals.

Transport geometry and structural evaluation. Sliced-Wasserstein distances compare projected distributions and have been used for generative modeling (Bonneel et al., 2015; Wu et al., 2019). We apply the one-dimensional location–scale–shape decomposition (Irpino & Verde, 2015) to the squared speed of the probability path. Persistent homology tracks components and loops across filtration levels (Edelsbrunner & Harer, 2010); stability bounds their response to perturbations (Cohen-Steiner et al., 2007). Geometry Score compares persistent summaries of generated and reference samples (Khrulkov & Oseledets, 2018). We use persistence stability to bound changes in the smoothed path marginals.

3 Path Motion and Structural Recovery

Problem setup. Let $(X_1, C) \sim p_{\mathrm{data}}$ and $X_0 \sim p_{\mathrm{base}}$ be independent, where C denotes an optional condition. For $t \in [0,1]$ and a time density q, the linear path and velocity-regression objective are

$$X_t = (1-t)X_0 + tX_1, \qquad U_t = X_1 - X_0, \tag{1}$$

$$\mathcal{L}_q(\theta) = \mathbb{E}_{t\sim q}\,\mathbb{E}\big[\|v_\theta(X_t, t, C) - U_t\|_2^2\big]. \tag{2}$$

At each time, the conditional-mean target is $v_t^*(x,c) = \mathbb{E}[U_t \mid X_t = x, C = c]$ (Lipman et al., 2023; Liu et al., 2023). The choice of q controls the frequency of these regression tasks. Path time t indexes a fixed probability path; training progress measures the updates needed to recover the structure of the data distribution. We examine which information about this structure is contained in squared path speed.

3.1 Moment and Shape Components

Write $\mu_t = \mathrm{Law}(X_t)$ and $\rho_t = \phi_{\#}\mu_t$ for the marginal in a fixed representation $\phi : \mathbb{R}^d \to \mathbb{R}^m$. For planar and pixel-space tasks, ϕ is the identity; ImageNet uses frozen features of decoded latent-path states (Appendix C). Assume that ρ_t has finite second moments. Let $P_{\omega,t}$ be the law of $Y_{\omega,t} = \omega^\top \phi(X_t)$, with ω uniform on $\mathbb{S}^{m-1}$. For finite-difference interval Δ, define a finite-difference estimate of squared path speed

$$S_{\phi,\Delta}(t) := \frac{\mathrm{SW}_2^2(\rho_{t+\Delta}, \rho_t)}{\Delta^2} = \mathbb{E}_\omega \frac{W_2^2(P_{\omega,t+\Delta}, P_{\omega,t})}{\Delta^2}, \tag{3}$$

where $\Delta > 0$ and both times lie in $[0,1]$ (Bonneel et al., 2015). We call this quantity squared path speed and its variation over time the speed profile. Its continuous limit is the squared metric speed in (5). We decompose it using the location–scale–shape identity (Irpino & Verde, 2015).

Proposition 1 (Finite-difference decomposition)**.** Let $a_{\omega,t}$ and $r_{\omega,t}$ be the mean and standard deviation of $P_{\omega,t}$. Assume that the projected variances are positive at both compared times, and let $P^\circ_{\omega,t}$ denote the centered, unit-variance distribution. Then

$$\begin{aligned} S_{\phi,\Delta}(t) &= S_{G,\phi,\Delta}(t) + N_{\phi,\Delta}(t), \\ S_{G,\phi,\Delta}(t) &= \frac{\mathbb{E}_\omega\big[(a_{\omega,t+\Delta} - a_{\omega,t})^2 + (r_{\omega,t+\Delta} - r_{\omega,t})^2\big]}{\Delta^2}, \\ N_{\phi,\Delta}(t) &= \frac{\mathbb{E}_\omega\big[r_{\omega,t+\Delta} r_{\omega,t} W_2^2\big(P^\circ_{\omega,t+\Delta}, P^\circ_{\omega,t}\big)\big]}{\Delta^2}. \end{aligned} \tag{4}$$

Both terms are nonnegative. The identity also holds for empirical marginals and finite averages over projection directions.

The moment component measures changes in projected means and scales. The shape component measures the evolution that remains after centering and scaling each marginal. Gaussian-only retains the moment component using the same representation, interval, and calibration as PathSpeed. With these choices held fixed, adding the shape component improves loop recovery and CIFAR-10 FID (Table 3).

Marginal motion and pairing variability. For the population interpretation, suppose that the projected processes are mean-square differentiable, have smooth positive densities with vanishing boundary flux, and their quantiles are differentiable in L^2 over probability levels and projection directions. Define $b_{\omega,t}(y) = \mathbb{E}[\dot{Y}_{\omega,t} \mid Y_{\omega,t} = y]$ and its best affine approximation in $L^2(P_{\omega,t})$, $\ell_{\omega,t}(y) = \dot{a}_{\omega,t} + (\dot{r}_{\omega,t}/r_{\omega,t})(y - a_{\omega,t})$. Writing $S_\phi = \lim_{\Delta\to 0} S_{\phi,\Delta}$ and $N_\phi = \lim_{\Delta\to 0} N_{\phi,\Delta}$, we obtain

$$\begin{aligned} S_\phi(t) &= |\dot{\rho}_t|^2_{\mathrm{SW}_2} = \mathbb{E}_\omega \mathbb{E}\big[b_{\omega,t}(Y_{\omega,t})^2\big], \\ N_\phi(t) &= \mathbb{E}_\omega \mathbb{E}\Big[\big(b_{\omega,t}(Y_{\omega,t}) - \ell_{\omega,t}(Y_{\omega,t})\big)^2\Big] = \mathbb{E}_\omega\Big[r^2_{\omega,t} |\dot{P}^\circ_{\omega,t}|^2_{W_2}\Big]. \end{aligned} \tag{5}$$

Here $|\dot{P}|_{W_2}$ denotes metric speed. We obtain these identities from quantile transport and the continuity equation (Ambrosio et al., 2008). For a projected process, suppressing ω in the notation, conditional expectation separates its velocity energy as

$$\mathbb{E}[\dot{Y}_t^2] = \mathbb{E}[b_t(Y_t)^2] + \mathbb{E}[\mathrm{Var}(\dot{Y}_t \mid Y_t)].$$

The first term is the squared speed of the projected marginal in (5). The second measures velocity variation among endpoint pairs at the same projected position. For the identity representation in our planar and pixel-space experiments, the linear path gives $\dot{Y}_{\omega,t} = \omega^\top(X_1 - X_0)$, so the total velocity energy is constant in time. Appendix A gives the proof and the connection to the full FM regression target. Squared path speed measures how much paired velocity is expressed as conditional-average marginal motion at each time.

At a fixed time, the velocity-regression loss is the sum of approximation error and a pairing-variance term that remains at the optimal predictor. Sampling according to loss magnitude responds to both contributions. We allocate updates according to the conditional-average motion required to transport the marginals. Its shape component measures the evolution beyond projected shifts and changes in scale. This gives a distributional basis for choosing training times before fitting the velocity model.

3.2 Persistent Structure at an Observation Scale

We bound changes in persistent structure by integrated path speed at a fixed observation scale. Let $f_{\rho,\eta} = G_\eta * \rho$, where G_η is the $\mathcal{N}(0, \eta^2 I_m)$ density. On a fixed finite triangulation, let $\mathcal{D}_j^\eta(\rho)$ be the superlevel-set persistence diagram of the piecewise-linear interpolant of $f_{\rho,\eta}$. The observation scale η specifies spatial resolution; $j = 0$ and $j = 1$ track components and loops. We use d_B for the bottleneck distance between diagrams.

Proposition 2 (Path motion bounds changes in persistence diagrams)**.** For $\rho, \nu \in \mathcal{P}_2(\mathbb{R}^m)$, there is a constant $C_{m,\eta}$ depending only on m and η such that

$$d_B\big(\mathcal{D}_j^\eta(\rho), \mathcal{D}_j^\eta(\nu)\big) \le C_{m,\eta}\, \mathrm{SW}_2(\rho, \nu). \tag{6}$$

Consequently, along an absolutely continuous path with squared SW_2 speed $S_\phi(t)$, the change in the persistence diagram over $[a, b]$ is at most $C_{m,\eta} \int_a^b \sqrt{S_\phi(t)}\, dt$.

We combine the Fourier representation of Gaussian smoothing with persistence stability (Cohen-Steiner et al., 2007) to obtain (6); Appendix A gives the constant and proof. At a fixed observation scale, a substantial change in the persistence diagram requires corresponding marginal motion.

4 Method

We estimate squared path speed, smooth and calibrate the estimates, and train with the resulting fixed time distribution. Algorithm 1 summarizes the construction.

4.1 Estimating Squared Path Speed

Partition $[0,1]$ into K equal bins I_k with centers $t_k = (k - \frac{1}{2})/K$. For an offset $1 \le \delta \le \lfloor K/2 \rfloor$, set $\Delta = \delta/K$ and choose

$$j(k) = k + \delta, \qquad k = 1, \dots, K - \delta. \tag{7}$$

For each comparison, draw n endpoint pairs and construct the two marginal sample clouds from the same pairs. Sharing endpoint pairs makes the clouds coincide when the compared times coincide, suppressing sampling differences unrelated to temporal change. Project both clouds along the same L random unit directions. Let $y_{k,\ell,(i)}$ and $\widetilde{y}_{k,\ell,(i)}$ denote the separately sorted projections at t_k and $t_{j(k)}$. The estimated squared path speed is

$$\widehat{s}_k = \frac{1}{nL\Delta^2} \sum_{\ell=1}^{L} \sum_{i=1}^{n} \big(y_{k,\ell,(i)} - \widetilde{y}_{k,\ell,(i)}\big)^2. \tag{8}$$

The final δ bins have no valid forward neighbor and are assigned a fixed small floor $\widehat{s}_k = \epsilon_{\text{pad}}$ before smoothing. On valid forward comparisons, the same projections yield $\widehat{s}_G$ and $\widehat{N}$ through Proposition 1, with empirical variances computed using denominator n. For conditional generation, endpoint samples follow the training label distribution; the resulting time distribution is shared across conditions.

4.2 Smoothing and Calibrated Allocation

Let A be a fixed local averaging operator with nonnegative entries and $A\mathbf{1} = \mathbf{1}$. We smooth the estimated profile, $h = A\widehat{s}$, and, for a fixed $\epsilon_{\text{std}} > 0$, compute

$$\bar{h} = \frac{1}{K} \sum_k h_k, \qquad \sigma_h^2 = \frac{1}{K} \sum_k (h_k - \bar{h})^2, \qquad T_h = \sigma_h + \epsilon_{\text{std}}. \tag{9}$$

The allocation rewards higher smoothed squared speed and penalizes concentration relative to uniform time sampling.

Proposition 3 (Scale-calibrated allocation)**.** Let $u_k = 1/K$ and $T_h > 0$. The unique maximizer over the probability simplex Δ_K is

$$q_{\text{path}} = \arg\max_{q \in \Delta_K} \big\{ \langle q, h \rangle - T_h D_{\text{KL}}(q \| u) \big\}, \qquad q_{\text{path},k} = \frac{\exp((h_k - \bar{h})/T_h)}{\sum_{r=1}^{K} \exp((h_r - \bar{h})/T_h)}. \tag{10}$$

Calibration sets the allocation strength according to relative speed variation: the temperature tracks dispersion across time bins. A constant speed profile gives the uniform distribution. When $\epsilon_{\text{std}} = 0$ and $\sigma_h > 0$, replacing $\widehat{s}$ by $a\widehat{s} + b\mathbf{1}$, with $a > 0$, leaves the time distribution unchanged. Appendix B proves these statements.

Smoothing uses neighboring times to reduce fluctuations from finite samples and projections. Calibration expresses their relative variation on a shared scale, and softmax produces a directly sampleable distribution. Table 3(a) compares these constructions using the same speed estimates. Appendix B analyzes smoothing error and allocation stability; Appendix C describes the construction.

Training and cost. For each training example, sample $k \sim \text{Cat}(q_{\text{path}})$ and $t \sim \text{Unif}(I_k)$, then construct (X_t, U_t) using (1). The induced density is $q(t) = K q_{\text{path},k}$ on I_k. We optimize the unweighted objective (2), with the time distribution controlling the frequency of regression tasks. We construct the time distribution once and cache $O(K)$ probabilities. Sampling adds no extra velocity-network forward or backward pass during training.

Algorithm 1 PathSpeed training

Input: Endpoint sampler; ϕ, A; $K, \delta, n, L, \epsilon_{\text{pad}}, \epsilon_{\text{std}}$.
Output: Time distribution q_{path}; model v_θ.

```
Initialize ŝ_k = ε_pad for all k.
for k = 1, ..., K − δ do
    Sample paired clouds at t_k, t_{j(k)} via (7).
    Project, sort; estimate ŝ_k by (8).
end for
Smooth h = Aŝ; cache q_path by (10).
for each minibatch do
    Draw endpoints and times per example.
    Form (X_t, U_t) by (1); update on (2).
end for
```

5 EXPERIMENTS

We test structural recovery on controlled distributions and image quality, coverage, and connectivity, then ablate the time-distribution construction.

Experimental setup. Within each dataset, methods share the path, architecture, optimizer, batch size, training budget, and generation protocol. We compare uniform time sampling (Uniform), Logit-normal (Esser et al., 2024), Laplace-λ with fixed $b = 1$ (Hang et al., 2024), Loss-CDF, and InfoNoise (Raya et al., 2026). Loss-CDF is a CDCD-inspired loss-profile sampler adapted to FM (Dieleman et al., 2022). Checkerboard uses 5k updates; Gaussian25 and 2rings use 10k. CIFAR-10 (Krizhevsky, 2009) uses a pixel-space U-Net at 100k updates. Class-conditional latent ImageNet-256 (Deng et al., 2009) uses SiT-B/2 (Ma et al., 2024) at 400k without classifier-free guidance. COIL-100 (Nene et al., 1996) tests recovery of image-feature connectivity across object views in Figure 3. Squared path speed uses coordinates for the planar tasks, pixels for CIFAR-10 and COIL-100, and frozen Inception features of decoded latent-path states for ImageNet-256. Appendix D defines the comparison settings and metrics.

Metrics and recovery thresholds. CellDisc is the Jensen–Shannon divergence between grid occupancies on Checkerboard and Gaussian25; D_{β_1} compares density-superlevel Betti curves on 2rings. Final image metrics are FID-50K (Heusel et al., 2017), feature-space Recall (Kynkäänniemi et al., 2019), Geometry Score (GS) (Khrulkov & Oseledets, 2018), and MST-based H_0 connectivity discrepancy. We report mean $\pm$ sample standard deviation over three training seeds. Thresholds are Uniform's final structural error or image FID.

Let $\overline{M}_m(b)$ be the mean lower-is-better metric for sampler m after b updates and B the final budget. For controlled tasks, $\mathcal{B}$ contains evaluated update counts; for image FID, $\overline{M}_m$ is interpolated between checkpoints and $\mathcal{B}$ spans their update range. With U denoting Uniform, define

$$b^\star_m = \min\{b \in \mathcal{B} : \overline{M}_m(b) \le \overline{M}_{\mathrm{U}}(B)\}, \qquad R_{\mathrm{steps}}(m) = \frac{b^\star_m}{b^\star_{\mathrm{U}}}. \tag{11}$$

Table 1 reports the first evaluated crossing; Table 2 reports R_{steps} from interpolated FID crossings, normalized by Uniform's crossing.

5.1 FASTER RECOVERY OF DISTRIBUTIONAL STRUCTURE

Figure 2 shows when Checkerboard components separate: at matched updates, PathSpeed weakens the spurious bridges that remain under Uniform.

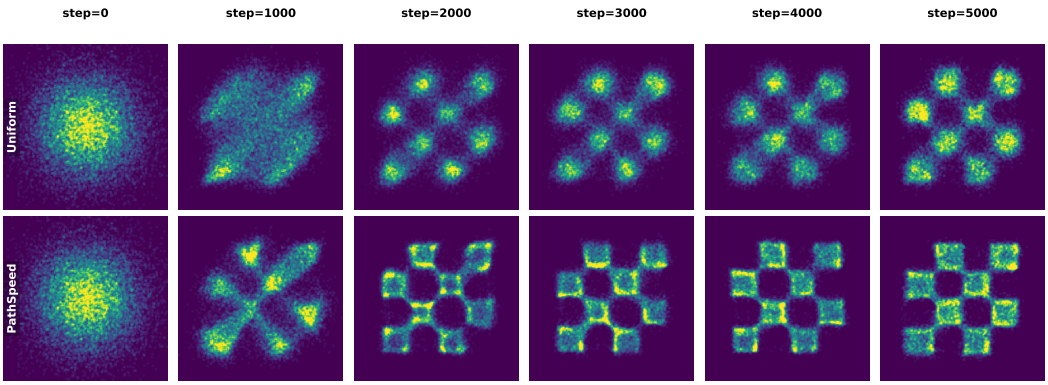


Figure 2: **PathSpeed separates Checkerboard components earlier.** Uniform and PathSpeed snapshots at matched training checkpoints. Components become distinct and bridges weaken earlier under PathSpeed.

PathSpeed reaches all three structural error thresholds with 33–49% fewer updates than Uniform and attains the lowest final errors (Table 1). The lower final errors show that the gains persist beyond first crossing the threshold. On 2rings, final error falls from 0.015 ± 0.008 to 0.001 ± 0.001.

Table 1: **Earlier structural recovery and lower final error.** Step is the first evaluated update reaching the final Uniform error; $>$ indicates no attainment within budget. Final is mean $\pm$ standard deviation at the stated budget.

	Checkerboard CellDisc; 5k updates		**Gaussian25** CellDisc; 10k updates		**2rings** D_{β_1}; 10k updates	
Method	**Step** $\downarrow$	**Final** $\downarrow$	**Step** $\downarrow$	**Final** $\downarrow$	**Step** $\downarrow$	**Final** $\downarrow$
Fixed schedules						
Uniform	4900	0.323 ± 0.002	6000	0.131 ± 0.005	1500	0.015 ± 0.008
Logit-normal	4200	0.325 ± 0.003	6500	0.125 ± 0.004	1400	0.012 ± 0.006
Laplace-λ ($b = 1$)	4700	0.326 ± 0.003	7800	0.133 ± 0.005	1800	0.016 ± 0.007
Online schedules						
Loss-CDF	4800	0.325 ± 0.004	> 10000	0.240 ± 0.015	1900	0.018 ± 0.005
InfoNoise	4500	0.324 ± 0.004	9200	0.130 ± 0.006	1700	0.016 ± 0.006
Offline path-aware						
PathSpeed	**2500**	$\mathbf{0.276} \pm 0.010$	**3500**	$\mathbf{0.098} \pm 0.007$	**1000**	$\mathbf{0.001} \pm 0.001$

Figure 3 follows recovery throughout training, including image-feature connectivity on COIL-100. PathSpeed crosses the final Uniform error earlier in each panel and continues to reduce structural error after the crossing.

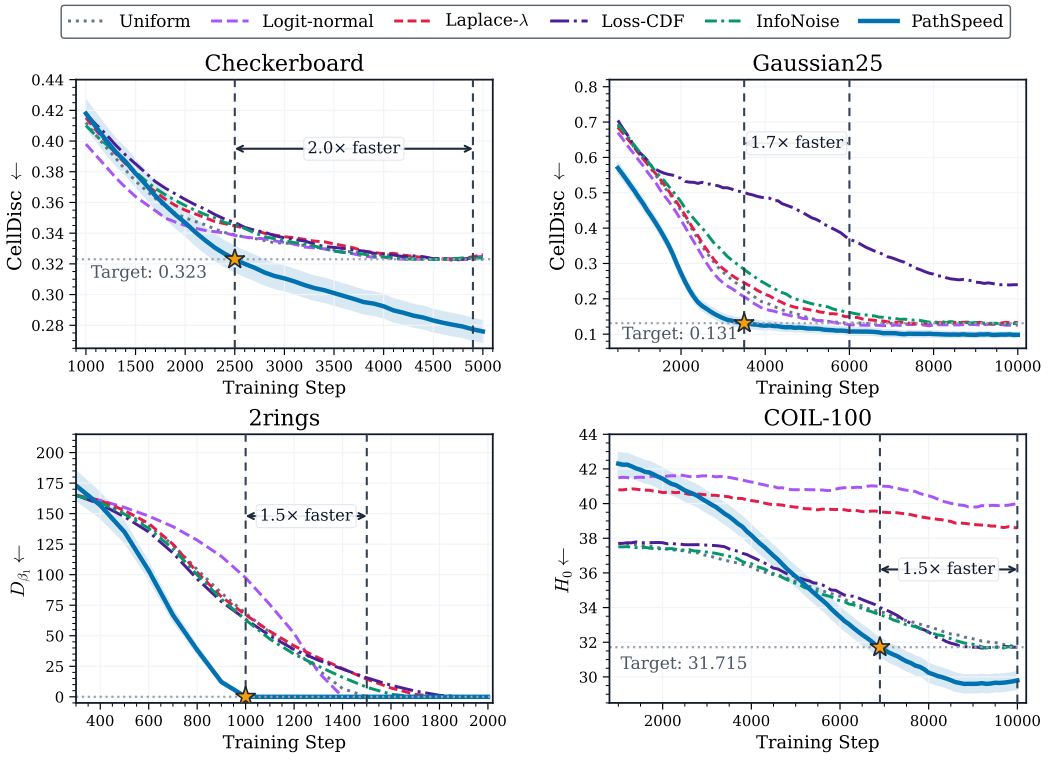


Figure 3: **PathSpeed reaches Uniform's final structural errors earlier.** Metrics are CellDisc on Checkerboard and Gaussian25, D_{β_1} on 2rings, and feature-connectivity H_0 on COIL-100. Curves compare Uniform, Logit-normal, Laplace-λ, Loss-CDF, InfoNoise, and PathSpeed; each horizontal threshold is the final Uniform mean for that metric.

5.2 Image Quality, Coverage, and Training Speed

Quality improves with coverage and connectivity. On CIFAR-10, PathSpeed obtains FID 4.61 ± 0.12 versus InfoNoise's 5.92 ± 0.15, a 22% reduction, together with the highest mean Recall and lowest mean H_0 (Table 2). On ImageNet-256, PathSpeed-trained SiT-B/2 obtains the lowest mean FID, 27.40, at 400k updates without classifier-free guidance, together with the highest mean Recall and lowest mean H_0. The gains in FID are accompanied by broader feature-space coverage and closer connectivity statistics.

The shape component contributes to the CIFAR-10 improvement. With the representation, interval, and calibration held fixed, the full speed profile gives lower FID than its moment-only component (Table 3). This comparison connects image training to the shape-sensitive allocation used for structural recovery.

Table 2: **Lower mean FID and fewer updates to Uniform's final FID.** CIFAR-10: 100k updates; ImageNet-256: SiT-B/2, 400k, no CFG. Final metrics are mean ± standard deviation; bold marks the best mean in each dataset block. Relative updates uses interpolated FID crossings in (11); $>$ denotes a lower bound.

Method	FID-50K ↓	Recall ↑	GS ↓	H_0 ↓	Relative updates ↓
(a) CIFAR-10					
Fixed schedules					
Uniform	7.95 ± 0.22	0.615 ± 0.012	0.019 ± 0.002	1.38 ± 0.09	1.00
Logit-normal	6.45 ± 0.18	0.645 ± 0.010	0.016 ± 0.002	1.24 ± 0.08	0.80
Laplace-λ ($b = 1$)	6.55 ± 0.18	0.642 ± 0.011	0.016 ± 0.002	1.25 ± 0.08	0.82
Online schedules					
Loss-CDF	8.15 ± 0.28	0.608 ± 0.014	$\mathbf{0.013} \pm 0.002$	1.42 ± 0.10	> 1.00
InfoNoise	5.92 ± 0.15	0.655 ± 0.010	0.015 ± 0.002	1.16 ± 0.07	0.72
Offline path-aware allocation					
Gaussian-only	4.85 ± 0.16	0.648 ± 0.009	0.015 ± 0.002	1.22 ± 0.07	0.79
PathSpeed	$\mathbf{4.61} \pm 0.12$	$\mathbf{0.672} \pm 0.008$	0.014 ± 0.002	$\mathbf{1.08} \pm 0.06$	**0.58**
(b) ImageNet-256					
Fixed schedules					
Uniform	33.20 ± 0.65	0.565 ± 0.014	0.035 ± 0.003	17.20 ± 0.65	1.00
Logit-normal	28.40 ± 0.72	0.612 ± 0.011	0.028 ± 0.002	14.10 ± 0.52	0.58
Laplace-λ ($b = 1$)	28.65 ± 0.70	0.609 ± 0.012	0.029 ± 0.002	14.25 ± 0.55	0.60
Online schedules					
Loss-CDF	33.10 ± 0.78	0.568 ± 0.015	$\mathbf{0.027} \pm 0.002$	17.10 ± 0.70	0.96
InfoNoise	27.95 ± 0.68	0.616 ± 0.011	$\mathbf{0.027} \pm 0.002$	13.85 ± 0.48	0.52
Offline path-aware allocation					
PathSpeed	$\mathbf{27.40} \pm 0.65$	$\mathbf{0.622} \pm 0.010$	$\mathbf{0.027} \pm 0.002$	$\mathbf{13.45} \pm 0.45$	**0.48**

Fewer updates to Uniform's final FID. Curve interpolation estimates 42% fewer updates on CIFAR-10 and 52% fewer on ImageNet-256 to reach this threshold. PathSpeed also improves final-budget FID.

5.3 Path Information and Time-Distribution Construction

Calibration controls the time distribution. With the estimated squared path speed fixed, raw linear normalization retains little of the gain. Removing standardization raises CIFAR-10 FID to 14.50, compared with 4.61 for PathSpeed (Table 3(a)). Proposition 3 shows how standardization makes allocation strength follow relative variation across time. Smoothing further improves the calibrated distribution. A log transform performs worse than the final construction. Checkerboard shows the same ordering.

Table 3: **Calibration, shape, and sampling at the right times.** (a) Time distributions from the same estimated squared path speed; (b) comparisons of the moment component, time locations, and Beta fit. Here $h = A\widehat{s}$ and $\mathcal{Q} = \mathrm{softmax} \circ Z$, where $Z(x) = (x - \bar{x}\mathbf{1})/(\sigma_x + \epsilon_{\mathrm{std}})$. Beta fit is fitted to q_{path}. Final metrics are mean $\pm$ standard deviation; Step follows (11).

Variant	**Time distribution**	**Checkerboard**	**2rings**		**CIFAR-10**
		CellDisc ↓	Step ↓	D_{β_1} ↓	FID ↓
Uniform	$q_k = 1/K$	0.323 ± 0.002	1500	0.015 ± 0.008	7.95 ± 0.22
(a) Constructing the time distribution					
Raw normalization	$q_k \propto \widehat{s}_k + \epsilon$	0.320 ± 0.005	—	—	7.85 ± 0.25
No smoothing	$q = \mathcal{Q}(\widehat{s})$	0.288 ± 0.012	—	—	5.12 ± 0.16
No standardization	$q = \mathrm{softmax}(h)$	0.462 ± 0.018	—	—	14.50 ± 0.85
With log	$q = \mathcal{Q}(\log(h + \epsilon))$	0.298 ± 0.008	—	—	5.95 ± 0.18
(b) Shape component and allocation across time					
Gaussian-only	$q = \mathcal{Q}(A\widehat{s}_G)$	—	1400	0.010 ± 0.005	4.85 ± 0.16
Permuted	$q_k = q_{\mathrm{path},\pi(k)}$	—	> 10000	0.018 ± 0.012	8.45 ± 0.55
Negated	$q_k \propto 1/q_{\mathrm{path},k}$	—	> 10000	0.024 ± 0.009	10.20 ± 0.38
Beta fit	fitted to q_{path}	—	1300	0.007 ± 0.004	6.40 ± 0.14
PathSpeed	$q = \mathcal{Q}(h)$	$\mathbf{0.276} \pm 0.010$	**1000**	$\mathbf{0.001} \pm 0.001$	$\mathbf{4.61} \pm 0.12$

Shape information improves training. Gaussian-only and PathSpeed use the same representation, interval, and calibration. Adding the shape component of Proposition 1 lowers CIFAR-10 FID from 4.85 to 4.61 and 2rings final error from 0.010 to 0.001 (Table 3(b)). These gains show that changes beyond projected means and scales provide useful information for choosing training times.

Sampling at the right times matters. Permuting the time distribution preserves its probability values and entropy but reallocates them to different times. CIFAR-10 FID rises to 8.45, and 2rings fails to reach the threshold within 10k updates. Negating the speed preference also removes the benefit. Thus the location of higher sampling probabilities along the path contributes to structural recovery.

A Beta fit partially retains the gain. A two-parameter Beta distribution fitted to PathSpeed lowers CIFAR-10 FID from Uniform's 7.95 to 6.40. Gaussian-only and PathSpeed obtain 4.85 and 4.61, respectively. On 2rings, the full distribution recovers loops more accurately, with final error 0.001 compared with 0.007 for the Beta fit.

6 Discussion and Conclusion

The marginal motion of a target path provides information for allocating training updates before the velocity model is trained. We use squared path speed to measure this motion and separate changes in projected location and scale from changes in shape. PathSpeed converts the measurement into a fixed time distribution through smoothing and scale calibration. The resulting allocation recovers structures earlier and more accurately, with accompanying gains in image quality, coverage, and connectivity.

The comparisons identify how this information contributes. Adding the shape component improves training beyond moment-only sampling. Permuting the same probabilities removes the benefit of sampling at the right times. A two-parameter Beta fit improves over Uniform, while the full distribution gives lower CIFAR-10 FID and more accurate loop recovery on 2rings.

Our experiments use linear paths, fixed representations, and a global time distribution. For molecular conformational ensembles, we will study whether symmetry- and periodicity-aware representations can reduce the updates needed to recover conformational basins and relative occupancies while preserving physical validity.

DISCLOSURE OF AI USE

We used large language models to assist with literature discovery, research brainstorming, drafting and refining theoretical arguments, experimental planning, and code writing. The authors directed the research, made the final methodological and editorial decisions, and take responsibility for the accuracy and integrity of the manuscript, including its theoretical claims, experimental results, references, and accompanying code.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL https://www.jmlr.org/papers/v26/23-1605.html.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2 edition, 2008.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024. doi: 10.1038/s41467-024-54281-3. URL https://www.nature.com/articles/s41467-024-54281-3.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi: 10.1007/s10851-014-0506-3.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5. URL https://doi.org/10.1007/s00454-006-1276-5.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

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

Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010. ISBN 9780821849255.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7441–7451, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Hang_Efficient_Diffusion_Training_via_Min-SNR_Weighting_Strategy_ICCV_2023_paper.html.

Tiankai Hang, Shuyang Gu, Xin Geng, and Baining Guo. Improved noise schedule for diffusion training, 2024. URL https://arxiv.org/abs/2407.03297.

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. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Antonio Irpino and Rosanna Verde. Basic statistics for distributional symbolic variables: A new metric-based approach. *Advances in Data Analysis and Classification*, 9(2):143–175, 2015. doi: 10.1007/s11634-014-0176-4. URL https://doi.org/10.1007/s11634-014-0176-4.

Valentin Khrulkov and Ivan Oseledets. Geometry score: A method for comparing generative adversarial networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2621–2629. PMLR, 2018. URL https://proceedings.mlr.press/v80/khrulkov18a.html.

Myunsoo Kim, Donghyeon Ki, Seong-Woong Shim, and Byung-Jun Lee. Adaptive non-uniform timestep sampling for accelerating diffusion model training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2513–2522, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Kim_Adaptive_Non-Uniform_Timestep_Sampling_for_Accelerating_Diffusion_Model_Training_CVPR_2025_paper.html.

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. URL https://openreview.net/forum?id=NnMEadcdyD.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/7343a5c976f8399880b695267f1f9e9f-Abstract-Conference.html.

Seunghan Lee, Kibok Lee, and Taeyoung Park. ANT: Adaptive noise schedule for time series diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/db5ca61dbc08cf5143c05ad2d1b0b2ca-Abstract-Conference.html.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27474–27498. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24g.html.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *Computer Vision – ECCV 2024*, volume 15135 of *Lecture Notes in Computer Science*, pp. 23–40. Springer, 2024. doi: 10.1007/978-3-031-72980-5_2. URL https://doi.org/10.1007/978-3-031-72980-5_2.

Sameer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (COIL-100). Technical Report CUCS-006-96, Department of Computer Science, Columbia University, February 1996. URL https://cave.cs.columbia.edu/repository/COIL-100.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL https://proceedings.mlr.press/v139/nichol21a.html.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d0da30e312b75a3fffd9e9191f8bc1b0-Abstract-Conference.html.

Gabriel Raya, Bac Nguyen, Georgios Batzolis, Yuhta Takida, Dejan Stancevic, Naoki Murata, Chieh-Hsin Lai, Yuki Mitsufuji, and Luca Ambrogioni. Noise scheduling as information-guided allocation in diffusion training. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026. URL https://openreview.net/forum?id=VnnK7CYAx2.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1): e2408799121, 2025. doi: 10.1073/pnas.2408799121. URL https://doi.org/10.1073/pnas.2408799121.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. Sliced wasserstein generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3713–3722, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wu_Sliced_Wasserstein_Generative_Models_CVPR_2019_paper.html.

Yixian Xu, Shengjie Luo, Liwei Wang, Di He, and Chang Liu. Diagnosing and improving diffusion models by estimating the optimal loss value. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=X7JfjLKKLQ.

A Squared Path Speed and Structural Control

This section proves the identities and bound used in Section 3. The finite-difference decomposition is an application of the one-dimensional location–scale–shape identity of Irpino & Verde (2015). The persistence argument uses stability of persistence diagrams (Cohen-Steiner et al., 2007).

A.1 Moment and shape components of squared speed

Let $Q_0, Q_1 \in L^2(0,1)$ be quantile functions of two one-dimensional probability laws with means a_0, a_1 and positive standard deviations r_0, r_1. Write $Q_i = a_i + r_i Z_i$, where $\int_0^1 Z_i(u)\,du = 0$ and $\int_0^1 Z_i(u)^2\,du = 1$. The quantile formula for quadratic transport gives

$$W_2^2(P_0, P_1) = \int_0^1 (Q_1 - Q_0)^2\,du \tag{A.1}$$

$$= (a_1 - a_0)^2 + (r_1 - r_0)^2 + r_0 r_1 \int_0^1 (Z_1 - Z_0)^2\,du. \tag{A.2}$$

Indeed, the cross term with $a_1 - a_0$ vanishes, and the remaining term is $r_1^2 + r_0^2 - 2r_0 r_1 \int Z_0 Z_1$. The last integral in (A.2) is $W_2^2(P_0^\circ, P_1^\circ)$. Apply this identity at t and $t + \Delta$, divide by Δ^2, and average over uniform projection directions to obtain Proposition 1.

The same proof applies to empirical quantile functions, which are piecewise constant. For equal-size samples, the integrals become averages over separately sorted projections, with means and variances computed using denominator n. Replacing the directional expectation by a finite average preserves the identity exactly. Positive variances are needed to define the standardized laws. For two Gaussian projected laws their standardized quantiles coincide, so their shape component is zero. The moment component equals the quadratic transport between moment-matched one-dimensional Gaussians. The decomposition applies to the valid finite-difference comparisons; boundary padding is specified separately in Appendix C.

A.2 Continuous marginal speed and affine motion

Fix a direction ω and suppress it in the notation. Under the regularity conditions in Section 3.1, let p_t be the projected density, F_t its distribution function, and Q_t its quantile function. The projected process obeys the continuity equation

$$\partial_t p_t(y) + \partial_y\{p_t(y) b_t(y)\} = 0, \qquad b_t(y) = \mathbb{E}[\dot{Y}_t \mid Y_t = y]. \tag{A.3}$$

Vanishing boundary flux implies $\partial_t F_t(y) = -p_t(y) b_t(y)$. Differentiating $F_t(Q_t(u)) = u$ therefore yields $\dot{Q}_t(u) = b_t(Q_t(u))$. Differentiability in L^2 gives

$$\lim_{\Delta \downarrow 0} \frac{W_2^2(P_{t+\Delta}, P_t)}{\Delta^2} = \int_0^1 \dot{Q}_t(u)^2\,du = \mathbb{E}[b_t(Y_t)^2]. \tag{A.4}$$

The assumed L^2 differentiability over both probability levels and directions justifies the directional average and identifies the squared sliced-Wasserstein metric speed (Ambrosio et al., 2008).

Writing $Q_t = a_t + r_t Z_t$, differentiation of $\int Z_t = 0$ and $\int Z_t^2 = 1$ gives $\langle 1, \dot{Z}_t \rangle = \langle Z_t, \dot{Z}_t \rangle = 0$. Consequently

$$\int \dot{Q}_t^2 = \dot{a}_t^2 + \dot{r}_t^2 + r_t^2 \int \dot{Z}_t^2. \tag{A.5}$$

The affine function $\ell_t(y) = \dot{a}_t + (\dot{r}_t / r_t)(y - a_t)$ is the $L^2(P_t)$ projection of b_t onto affine functions: $\mathbb{E} b_t(Y_t) = \dot{a}_t$ and $\mathbb{E}[(Y_t - a_t) b_t(Y_t)] = r_t \dot{r}_t$. It follows that

$$\mathbb{E}[(b_t(Y_t) - \ell_t(Y_t))^2] = r_t^2 \int \dot{Z}_t^2 = r_t^2 |\dot{P}_t^\circ|_{W_2}^2. \tag{A.6}$$

Averaging over directions proves (5). The limit of the finite-difference shape component can also be obtained by subtracting the moment limit from the total limit in (4); this uses the stated joint L^2 regularity.

For differentiable representations the chain rule gives $\dot{Y}_{\omega,t} = \omega^\top D\phi(X_t)U_t$. For $\phi = \mathrm{id}$, conditioning first on (X_t, C) and then on $Y_{\omega,t}$ gives

$$b_{\omega,t}(Y_{\omega,t}) = \mathbb{E}[\omega^\top U_t \mid Y_{\omega,t}] = \mathbb{E}[\omega^\top v_t^*(X_t, C) \mid Y_{\omega,t}]. \tag{A.7}$$

This conditions on a single projected position, whereas the full FM regression conditions on (X_t, C).

A.3 Marginal motion and pairing variability

For any square-integrable projected velocity, write $b_{\omega,t}(Y_{\omega,t}) = \mathbb{E}[\dot{Y}_{\omega,t} \mid Y_{\omega,t}]$. Orthogonality of conditional expectation gives

$$\mathbb{E}[(\dot{Y}_{\omega,t})^2] = \mathbb{E}[b_{\omega,t}(Y_{\omega,t})^2] + \mathbb{E}\,\mathrm{Var}(\dot{Y}_{\omega,t} \mid Y_{\omega,t}). \tag{A.8}$$

Indeed, the residual $\dot{Y}_{\omega,t} - b_{\omega,t}(Y_{\omega,t})$ has zero conditional mean, so its cross term with $b_{\omega,t}(Y_{\omega,t})$ vanishes. Under the regularity conditions above, the first term is the squared speed of the projected marginal. The second is the minimum mean-squared error for predicting projected velocity from projected position and measures variability of the pairs passing through that position.

For the identity representation and linear coupling, $\dot{Y}_{\omega,t} = V_\omega = \omega^\top(X_1 - X_0)$, whose law is independent of t. Thus conditional-average motion and pairing variability sum to the constant second moment $\mathbb{E}[V_\omega^2]$. For differentiable nonlinear ϕ, the total projected velocity energy is $\mathbb{E}[(\omega^\top D\phi(X_t)U_t)^2]$ and can vary with t. The scalar conditioning used here is $Y_{\omega,t}$; the full FM regression conditions on (X_t, C), as shown in the preceding subsection.

A.4 Changes in persistence diagrams at a fixed observation scale

Let ω be uniform on $\mathbb{S}^{m-1}$ and use this normalized average in the definition of SW_2. For characteristic functions χ_ρ, χ_ν and $r \geq 0$, any coupling of the projected laws gives

$$|\chi_\rho(r\omega) - \chi_\nu(r\omega)| \leq rW_1(P_{\omega,\rho}, P_{\omega,\nu}) \leq rW_2(P_{\omega,\rho}, P_{\omega,\nu}), \tag{A.9}$$

using $|e^{irx} - e^{iry}| \leq r|x - y|$. Let G_η be the $\mathcal{N}(0, \eta^2 I_m)$ density. Fourier inversion of its convolution is absolutely integrable, since characteristic functions are bounded and the Gaussian Fourier transform is integrable. Polar coordinates and Cauchy–Schwarz over directions then give

$$\|G_\eta * (\rho - \nu)\|_\infty \leq (2\pi)^{-m} \int_{\mathbb{R}^m} e^{-\eta^2\|\xi\|^2/2} |\chi_\rho(\xi) - \chi_\nu(\xi)|\, d\xi \tag{A.10}$$

$$\leq C_{m,\eta}\, \mathrm{SW}_2(\rho, \nu), \tag{A.11}$$

where a valid explicit constant is

$$C_{m,\eta} = \frac{|\mathbb{S}^{m-1}|}{(2\pi)^m} \int_0^\infty r^m e^{-\eta^2 r^2/2}\, dr = \frac{\sqrt{2}\,\Gamma((m+1)/2)}{(2\pi)^{m/2}\Gamma(m/2)\eta^{m+1}}. \tag{A.12}$$

For a fixed finite triangulation, the sup norm of the difference of the piecewise-linear interpolants is bounded by the maximum difference at vertices, hence by this density sup norm. Persistence stability for tame functions on the same complex, applied to the negatives of the interpolants for superlevel filtrations, gives Proposition 2 (Cohen-Steiner et al., 2007). For an absolutely continuous sliced-Wasserstein path, $\mathrm{SW}_2(\rho_a, \rho_b) \leq \int_a^b \sqrt{S_\phi(t)}\, dt$, which yields the integrated bound.

The constant depends on observation scale and dimension. The bound uses full directional transport and the persistence diagrams of the specified smoothed densities on a fixed complex.

B Calibration and Estimation Error

B.1 Gibbs allocation and scale calibration

For a fixed vector $h \in \mathbb{R}^K$, temperature $T > 0$, and uniform $u_k = 1/K$, define $F_{h,T}(q) = \langle q, h\rangle - TD_{\mathrm{KL}}(q\|u)$. Set $Z = \sum_k u_k e^{h_k/T}$ and $q_k^* = u_k e^{h_k/T}/Z$. Direct substitution gives

$$F_{h,T}(q) = T\log Z - TD_{\mathrm{KL}}(q\|q^*). \tag{B.1}$$

Nonnegativity of KL divergence proves uniqueness, including on the boundary of the simplex. Subtracting h from the logits leaves the softmax unchanged, giving (10) with $T = \sigma_h + \epsilon_{\mathrm{std}}$. A constant vector yields $q^* = u$.

Because $A\mathbf{1} = \mathbf{1}$, transforming $\widehat{s}$ to $a\widehat{s} + b\mathbf{1}$ transforms h to $ah + b\mathbf{1}$ and σ_h to $a\sigma_h$ for $a > 0$. If $\epsilon_{\mathrm{std}} = 0$ and $\sigma_h > 0$, the centered normalized logits, and hence the time distribution, are identical. For positive ϵ_{std}, the denominator becomes $a\sigma_h + \epsilon_{\mathrm{std}}$, so exact scale invariance need not hold.

B.2 Smoothing bias and variance

Let s be the population finite-difference estimate of squared path speed at the valid comparison times, extended to the remaining bins by the fixed boundary floor. Suppose $\widehat{s}$ has finite second moments, with $\mathbb{E}\widehat{s} = s + \beta$ and covariance Σ; A is deterministic. Then

$$\mathbb{E}\|A\widehat{s} - s\|_2^2 = \|(A - I)s + A\beta\|_2^2 + \mathrm{tr}(A\Sigma A^\top). \tag{B.2}$$

To see this, write $A\widehat{s} - s = (A - I)s + A\beta + A(\widehat{s} - \mathbb{E}\widehat{s})$. The last summand has zero mean; the cross term vanishes in expectation, and its expected squared norm is the stated trace. The identity allows correlations across bins, including those induced by shared endpoint samples or projection directions. Smoothing can reduce fluctuations while introducing distortion through $(A - I)s$; it need not reduce total error for every profile.

B.3 Stability of the calibrated allocation

A deterministic bound makes the nondegeneracy requirement explicit. Let $P = I - \mathbf{1}\mathbf{1}^\top/K$, $\sigma_x = \|Px\|_2/\sqrt{K}$, and $z(x) = Px/(\sigma_x + \epsilon)$ for fixed $\epsilon \geq 0$. For vectors h, g satisfying $\sigma_h, \sigma_g \geq a > 0$,

$$\|z(h) - z(g)\|_2 \leq \frac{2}{a + \epsilon}\|h - g\|_2. \tag{B.3}$$

Indeed, $|\sigma_h - \sigma_g| \leq \|h - g\|_2/\sqrt{K}$. Subtracting the two normalized vectors, bounding the numerator difference by $\|h - g\|_2$, and using $\|Pg\|_2/(\sigma_g + \epsilon) \leq \sqrt{K}$ bounds each of the two terms by $\|h - g\|_2/(a + \epsilon)$. The softmax Jacobian $\mathrm{diag}(q) - qq^\top$ is positive semidefinite with trace at most one, hence spectral norm at most one. Thus the time distributions satisfy the same Lipschitz bound.

For $T_h = \sigma_h + \epsilon$, let $q(h) = \mathrm{softmax}(z(h))$ and $q(g) = \mathrm{softmax}(z(g))$. The log-sum-exp Hessian has norm at most one, so its Bregman divergence, which equals the corresponding KL divergence, is at most half the squared logit distance. Therefore

$$0 \leq F_{h,T_h}(q(h)) - F_{h,T_h}(q(g)) = T_h D_{\mathrm{KL}}(q(g)\|q(h)) \leq \frac{2T_h}{(a + \epsilon)^2}\|h - g\|_2^2. \tag{B.4}$$

This establishes stability of the speed-based allocation utility under the stated dispersion condition.

C Time-Distribution Construction and Training

Representations and estimation. We estimate squared sliced-Wasserstein speed using the mean squared difference of separately sorted projections in (8). The compared clouds share endpoint pairs and projection directions. The planar experiments use coordinates directly, with $n = 12{,}000$, $L = 64$, and $\delta = 2$; $K = 50$ for Checkerboard and $K = 64$ for Gaussian25 and 2rings. CIFAR-10 and COIL-100 use $(K, \delta, n, L) = (64, 2, 512, 64)$ and estimate squared path speed directly in pixel space, with ϕ the identity on the image tensors used for velocity regression. ImageNet-256 uses the same (K, δ, n, L) as CIFAR-10, with a path constructed in scaled VAE latents and features extracted after decoding:

$$z_t = (1 - t)z_0 + tz_1, \qquad \phi(z_t) = \mathrm{Inception}(\mathrm{Decode}(z_t/0.18215)).$$

We interpolate latent endpoints, decode each intermediate state, and extract features with frozen VAE and Inception networks. The time distribution is computed offline and cached.

Smoothing and sampling. We use forward differences and fill the final δ bins before smoothing, as in Section 4.1. For the planar and image configurations, the smoothing kernel is $(0.25, 0.5, 0.25)$ on interior bins, with the first and last entries unchanged; $\epsilon_{\text{pad}} = \epsilon_{\text{std}} = 10^{-12}$. Standardization uses the population standard deviation and unit logit multiplier. The resulting distribution is cached before training. The class-conditional ImageNet model uses one global time distribution with endpoints drawn from the training label distribution. Sampling a bin and then uniformly within it induces $q(t) = Kq_{\text{path},k}$ on I_k. We use unweighted velocity MSE.

Construction cost and regression target. After representation extraction, projected comparisons cost $O(KLn(m + \log n))$ operations. The cached distribution stores $O(K)$ probabilities and adds no velocity-network evaluation during training. For a fixed time, the square-loss decomposition gives

$$\mathbb{E}[\|v(X_t, t, C) - U_t\|^2 \mid X_t, C] = \|v - v_t^*\|^2 + \mathbb{E}[\|U_t - v_t^*\|^2 \mid X_t, C]. \tag{C.1}$$

The conditional minimizer is unchanged by q; the time-weighted risk and finite-budget optimization depend on q. The implementation averages squared errors over coordinates, a fixed normalization of (2) within each dataset.

D Experimental Settings and Metric Definitions

Controlled distributions. Gaussian25 is an equally weighted 5×5 Gaussian mixture with center spacing 2 and component standard deviation 0.1. 2rings consists of two equally weighted concentric circles of radii 2 and 4, with coordinate noise of standard deviation 0.05. The planar models are MLPs with three hidden layers of width 512, SiLU activations, and a 64-dimensional sinusoidal time embedding processed by a time MLP. We use AdamW with zero weight decay, batch size 1024, and gradient norm clipping at 1. Checkerboard uses learning rate 10^{-3} for 5k updates; Gaussian25 and 2rings use 3×10^{-3} for 10k. Generation uses raw model weights and explicit Euler integration: 20 steps and 10,000 real/generated samples for Checkerboard; 50 steps and 4,096 real/generated samples for Gaussian25 and 2rings.

CIFAR-10 and COIL-100. We use TorchCFM U-Nets with base width 128, two residual blocks per level, channel multipliers $(1, 2, 2, 2)$, and dropout 0.1. CIFAR-10 uses 32×32 RGB inputs and COIL-100 uses 64×64 RGB inputs, normalized to $[-1, 1]$. Both use Adam with learning rate 2×10^{-4}, batch size 128, zero weight decay, and gradient clipping at 1. The budgets are 100k and 60k updates, with linear warmup for 5k and 1k updates, respectively. CIFAR-10 uses random horizontal flips and EMA decay 0.9999. COIL-100 uses no augmentation and EMA decay $\min(0.9999, (b+1)/(b+10))$ at update b. Generation uses EMA weights and adaptive `dopri5` integration with absolute and relative tolerances 10^{-5}. CIFAR-10 final evaluation uses 50,000 generated images; COIL-100 uses 4,096, with its 7,200 training images as the real reference set.

Class-conditional latent ImageNet-256. We train SiT-B/2 with 12 layers, width 768, and 12 attention heads on ImageNet-1K. A frozen VAE encodes 256×256 images into $4 \times 32 \times 32$ latents, scaled by 0.18215. Patches of size 2×2 give 256 tokens. Training uses AdamW with constant learning rate 10^{-4}, zero weight decay, global batch size 256, and 400k updates. We use center cropping, random horizontal flips, class dropout 0.1, and EMA decay 0.9999. The target is unweighted velocity MSE along the linear latent path. Final evaluation uses EMA weights, 250-step Euler–Maruyama sampling, 50,000 images, and no classifier-free guidance. Uniform is the uniform-time comparison; the other samplers replace its time distribution under the same model and evaluation settings.

Comparison samplers. Logit-normal draws $z \sim \mathcal{N}(0, 1)$ and sets $t = \text{sigmoid}(z)$, clipped to $[10^{-6}, 1 - 10^{-6}]$. Laplace-λ uses $b = 1$. Loss-CDF tracks unweighted velocity losses in K bins with EMA decay 0.99. After 1,000 uniform warmup updates, it refreshes the distribution every 100 updates using three-point smoothing and a 2% uniform mixture. For the CIFAR-10 InfoNoise FM adaptation, we form $\widehat{x}_1 = x_t + (1 - t)v_\theta(x_t, t)$ and sample with density

$$\pi(t) \propto \frac{t}{1-t}\widehat{\text{MSE}}_x(t)\frac{\sigma(t)^3}{\sigma(t)^3 + 0.15^3}, \qquad \sigma(t) = \frac{1-t}{t}.$$

The estimator uses 128 log-noise bins with endpoint coverage, a minimum 5k-update warmup, and refreshes every 100 updates with smoothing and EMA new-value weight 0.1 (Raya et al., 2026). On

Checkerboard, InfoNoise uses 500 sampler-warmup updates; the learning-rate schedule follows the common Checkerboard training protocol. All samplers train with the same unweighted velocity-regression target.

Relation to velocity-regression error. Let $\ell_\theta(t)$ denote the population velocity MSE, $e_\theta(t)$ the approximation error relative to v_t^*, $r^*(t)$ the irreducible velocity variance, and $M_{x,\theta}(t)$ the population MSE of $\widehat{x}_1$, all averaged over data coordinates. Let $g(t) = \sigma(t)^3/(\sigma(t)^3 + 0.15^3)$ be the gate above. Then

$$\ell_\theta(t) = e_\theta(t) + r^*(t), \qquad M_{x,\theta}(t) = (1-t)^2 \ell_\theta(t).$$

At a fixed time, the denoising estimate used above satisfies $\widehat{x}_1 - X_1 = (1-t)(v_\theta - U_t)$. Its population MSE is therefore $(1-t)^2$ times the velocity-regression MSE. Combining this identity with (C.1) gives the population form of the sampling density below.

$$\pi_{\mathrm{pop},\theta}(t) \propto t(1-t)g(t)\big[e_\theta(t) + r^*(t)\big].$$

The allocation contains the current approximation error and the irreducible velocity variance, modulated by the path factor and gate. The implemented sampler estimates the denoising profile using the binning, EMA, and smoothing procedure specified above.

Statistics and recovery thresholds. We report the mean and sample standard deviation across three training seeds. The controlled distributions are evaluated every 100 updates. Image models are evaluated at saved training checkpoints, with final FID-50K reported in Table 2. Recovery thresholds are the final Uniform mean values. For controlled tasks, we report the first evaluated threshold crossing. For image FID, we interpolate the mean learning curves between evaluated checkpoints and estimate their first crossing of Uniform's final FID. Relative updates is the ratio of this estimated crossing to Uniform's crossing of the same threshold, as in (11).

Components, time locations, and compression. Gaussian-only uses the moment component with the same representation, interval, and calibration. Permuted reassigns the probabilities to time bins, preserving their values and entropy. Negated normalizes the reciprocal distribution, reversing the logit preference. The two-parameter Beta fit is fitted to q_{path}.

Cell discrepancy and loop error. Let p and r be the generated and real occupancy distributions on the evaluation grid, and $m = (p+r)/2$. We define

$$\mathrm{CellDisc} = \mathrm{JS}(p, r) = \tfrac{1}{2} D_{\mathrm{KL}}(p\|m) + \tfrac{1}{2} D_{\mathrm{KL}}(r\|m). \tag{D.1}$$

Checkerboard uses a 120×120 grid on $[-6, 6]^2$. Gaussian25 uses a 120×120 occupancy grid on $[-6, 6)^2$, with cell width 0.1 and 4,096 real and 4,096 generated samples per evaluation. Its saved Betti-1 diagnostic uses a 64×64 grid on the same domain, four-neighbor adjacency, and top-mass level 0.90, reporting $|\beta_1^{\mathrm{fake}} - \beta_1^{\mathrm{real}}|$. Gaussian25 Recall uses the project's `precision_recall_knn` implementation on raw two-dimensional coordinates with $k = 5$. For 2rings, we compare density-superlevel Betti curves parameterized by top-mass levels m_j:

$$D_{\beta_1} = \left\{ \frac{1}{M} \sum_{j=1}^{M} [\beta_1^{\mathrm{gen}}(m_j) - \beta_1^{\mathrm{real}}(m_j)]^2 \right\}^{1/2}. \tag{D.2}$$

Image diagnostics. FID measures feature-distribution discrepancy (Heusel et al., 2017). Recall uses $k = 5$ with 1,024 real and generated feature points in the COIL-100 evaluation (Kynkäänniemi et al., 2019). GS is the L^2 norm of the difference between mean relative living-time vectors (Khrulkov & Oseledets, 2018). For equal-size point clouds, MST-H_0 compares sorted minimum-spanning-tree edge lengths:

$$H_0 = \frac{1}{n-1} \sum_{i=1}^{n-1} |e_i(Z_{\mathrm{gen}}) - e_i(Z_{\mathrm{real}})|. \tag{D.3}$$

For COIL-100, GS uses 512 feature points, 64 landmarks, and five internal repetitions; H_0 uses 512 points per cloud. Both metrics use three evaluation repetitions per model.

E PATH SPEED AND TRAINING-TIME DISTRIBUTIONS

We examine how marginal motion determines the distribution of training times. Figure 4 compares the resulting time distributions across datasets. Figure 5 separates the moment and shape components that enter the PathSpeed construction.

E.1 WHERE TRAINING SAMPLES ARE ALLOCATED

We compare PathSpeed with Gaussian-only, the Beta distribution fitted to PathSpeed, Logit-normal, and Laplace-λ. Uniform time sampling provides a constant-density reference. The InfoNoise curves show the sampling distributions at the end of training on each dataset.

All curves are expressed as densities with respect to the flow-matching time t, increasing from noise to data. For a common plotting partition $\{J_j\}$, we display

$$\overline{q}_j = \frac{\Pr(t \in J_j)}{|J_j|}. \tag{E.1}$$

This preserves probability mass when the samplers use different native discretizations. For the equal-width PathSpeed bins, the density is $Kq_{\mathrm{path},k}$ on I_k.

Gaussian-only uses the same representation, finite-difference interval, and calibration as PathSpeed. Their difference therefore shows where shape information changes the allocation. The fitted Beta curve shows which features of the resulting time distribution survive two-parameter compression.

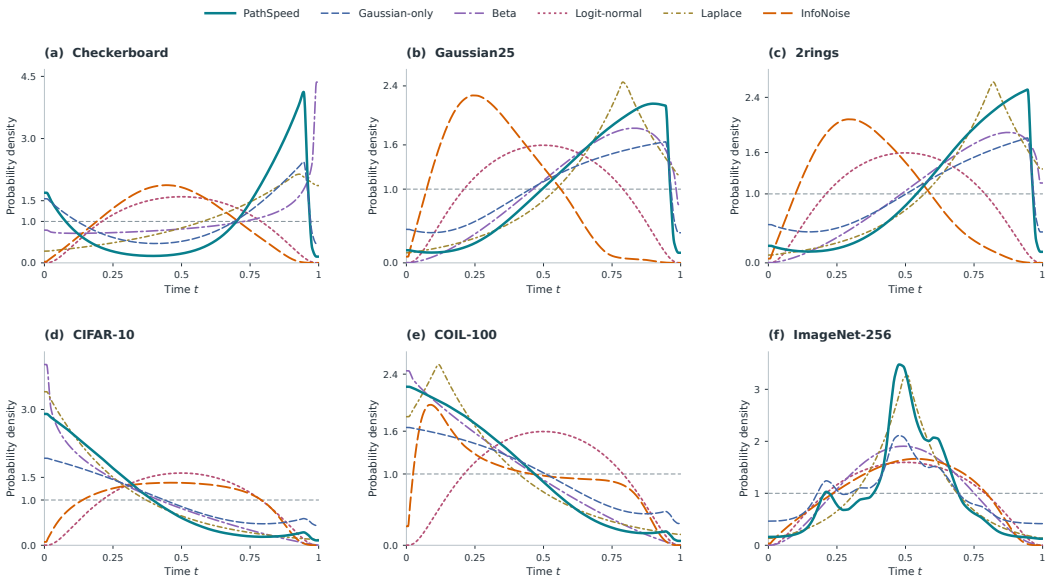


Figure 4: **Training allocation across the probability path.** Curves are densities in the common FM time coordinate. Gaussian-only isolates allocation from moment changes, and the fitted Beta compresses the PathSpeed distribution. InfoNoise is shown at each dataset's final training checkpoint. The horizontal dashed line marks the Uniform density of one. Each panel uses the representation and path of its dataset.

The allocation varies across datasets. CIFAR-10 and COIL-100 place substantial mass near the data endpoint, whereas ImageNet-256 concentrates more mass in the interior of the path.

E.2 MOMENT AND SHAPE COMPONENTS ALONG THE PATH

For Checkerboard, Gaussian25, and 2rings, we compute the two components from the same projected samples used to estimate squared path speed. On every valid forward comparison,

$$\widehat{s}_k = \widehat{s}_{G,k} + \widehat{N}_k. \tag{E.2}$$

The moment component measures changes in projected means and standard deviations. The shape component measures the change remaining after standardization of each projected marginal.

Figure 5 displays the components before temporal smoothing. Within each dataset, all three curves use the same normalization constant

$$c_d = \max_{k \in \mathcal{V}} \widehat{s}_k, \qquad \mathcal{V} = \{1, \dots, K - \delta\}. \tag{E.3}$$

The displayed components therefore retain their relative sizes and sum to the displayed total. The final δ bins are marked separately as padded bins.

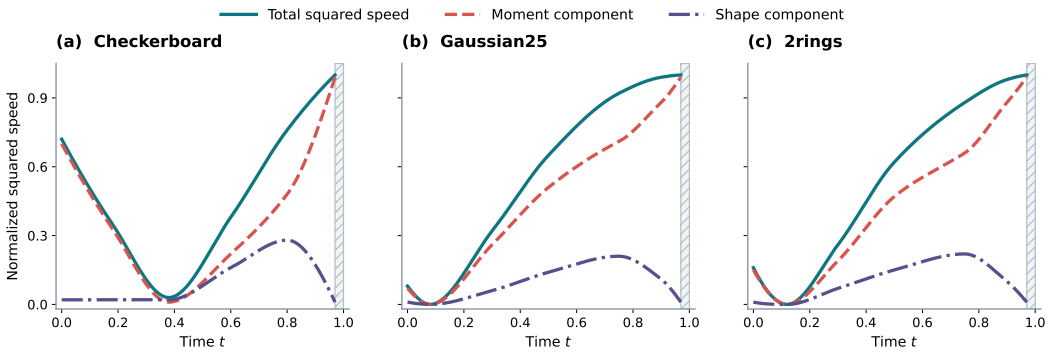


Figure 5: **The information in squared path speed.** The total is decomposed using the same samples and projections as the estimator. A common scale within each panel preserves the additive decomposition. The hatched terminal region marks the final δ bins, filled by forward-difference padding.

Across these planar examples, the moment component accounts for most of the total squared speed, while the smaller shape component has a nonconstant temporal profile.

F Estimating Squared Path Speed

We evaluate the sample-based estimator against a population reference on Gaussian25, then examine how the resulting time distribution changes with the estimation settings.

F.1 Population Reference on Gaussian25

Let the Gaussian25 data distribution be

$$X_1 \sim \frac{1}{25} \sum_{j=1}^{25} \mathcal{N}(\mu_j, \sigma_c^2 I_2), \qquad \sigma_c = 0.1, \tag{F.1}$$

with the component means specified by the dataset. For $X_0 \sim \mathcal{N}(0, I_2)$ independent of X_1, the projected marginal along a unit direction ω is

$$P_{\omega,t} = \frac{1}{25} \sum_{j=1}^{25} \mathcal{N}\left(t\,\omega^\top \mu_j,\, \tau_t^2\right), \qquad \tau_t^2 = (1-t)^2 + t^2 \sigma_c^2. \tag{F.2}$$

Its cumulative distribution function is

$$F_{\omega,t}(y) = \frac{1}{25} \sum_{j=1}^{25} \Phi\left(\frac{y - t\,\omega^\top \mu_j}{\tau_t}\right), \tag{F.3}$$

where Φ is the standard normal CDF.

For a fixed set of projection directions, numerical inversion of these CDFs gives a population finite-difference reference:

$$s_k^{\mathrm{ref}} = \frac{1}{L\Delta^2} \sum_{\ell=1}^{L} \int_0^1 \left[F^{-1}_{\omega_\ell, t_k+\Delta}(u) - F^{-1}_{\omega_\ell, t_k}(u) \right]^2 du. \tag{F.4}$$

The reference uses the same time pairs and projection directions as the empirical estimate. We compute it in double precision with $L = 64$ projections. For each time and projection, the mixture CDF is evaluated on 16,385 equally spaced points spanning the smallest to largest projected component means, extended by $9\tau_t$ on each side. We invert this grid by linear interpolation and approximate the quantile integral using 4,096 equally spaced midpoints, $u_j = (j + 1/2)/4096$ for $j = 0, \ldots, 4095$. The squared quantile differences are averaged over these points and the projections, then divided by Δ^2.

For each independent endpoint-sampling repetition r, the Relative L^2 error on the valid bins is

$$e_r = \left[\frac{\sum_{k \in \mathcal{V}} (\widehat{s}_k^{(r)} - s_k^{\mathrm{ref}})^2}{\sum_{k \in \mathcal{V}} (s_k^{\mathrm{ref}})^2} \right]^{1/2}. \tag{F.5}$$

Each setting uses five independent endpoint-sampling repetitions. We report the mean and sample standard deviation of e_r.

F.2 Shared Endpoint Pairs

We compare two ways to construct the marginal sample clouds. The shared-pair estimator uses the same endpoint pairs at both times. The independent-cloud estimator draws a separate set of endpoint pairs for each time. Both use n samples per marginal, the same projection directions, and the same sorting-based transport calculation.

With shared endpoint pairs, the clouds coincide when the two times coincide. Independent clouds retain a finite-sample discrepancy even at the same time. We examine the effect of this sampling difference as the finite-difference interval and sample count change.

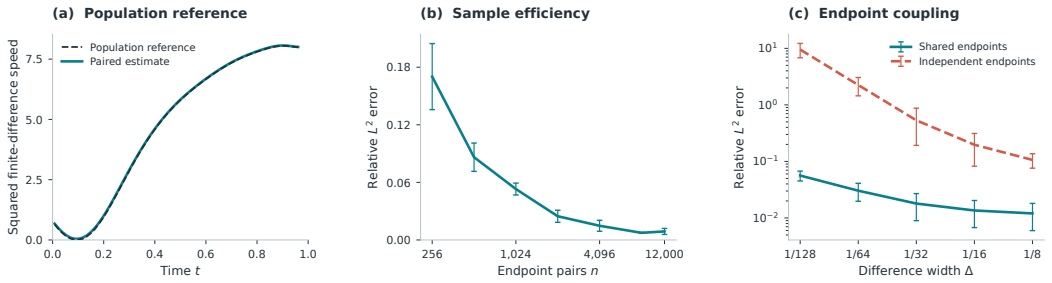


Figure 6: **Accuracy of the sample-based speed estimate.** Left: the estimated profile and the population reference at $n = 12000$ and $\Delta = 1/32$. Middle: Relative L^2 error for $n \in \{256, 512, 1024, 2048, 4096, 8192, 12000\}$. Right: Relative L^2 error for shared endpoint pairs and independent marginal clouds at $n = 4096$ and $\Delta \in \{1/128, 1/64, 1/32, 1/16, 1/8\}$. Shading and error bars show means $\pm$ sample standard deviations over five independent endpoint-sampling repetitions per setting and coupling method.

The paired estimate follows the population reference, and its error decreases as the sample count grows. Sharing endpoints is especially helpful at small difference widths, where independent marginal clouds produce larger errors.

F.3 Sensitivity of the Time Distribution

The default Gaussian25 construction uses $(K, \delta, n, L) = (64, 2, 12000, 64)$ and the binomial3 smoothing kernel $(0.25, 0.5, 0.25)$. Table 4 varies one construction choice around these defaults and reports differences between the resulting priors, together with their construction times.

For two constructed priors q and q', we compare their piecewise-constant time densities using

$$d_{\mathrm{TV}}(q, q') = \frac{1}{2} \int_0^1 |q(t) - q'(t)| \, dt. \tag{F.6}$$

This measures a difference in training-time allocation. For different grids, the integral is evaluated on their common refinement. The grid sweep fixes $\delta = 2$, so changing K also changes the finite-difference interval $\Delta = 2/K$. The offset sweep fixes $K = 64$.

Table 4: **Sensitivity of the Gaussian25 prior.** TV summarizes differences between constructed time-sampling priors. Each row lists the tested settings, including the default, and the reported TV and construction-time ranges.

Factor	Tested settings	Default	TV range	Build time range (s)
Time bins K	32, 64, 128	64	0.03633–0.07539	0.0078–0.0311
Offset δ	1, 2, 4	2	0.02809–0.06131	0.0153–0.0157
Endpoint pairs n	3000, 6000, 12000, 24000	12000	0.00026–0.00343	0.0095–0.0550
Projections L	16, 32, 64, 128	64	0.00069–0.00382	0.0132–0.0217
Smoothing	none, box3, binomial3, binomial5	binomial3	0.00381–0.01767	0.0155–0.0156

G One-time Prior Construction Cost

PathSpeed constructs its time distribution once and caches it throughout training. Table 5 reports construction times measured on an NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition GPU. We report the first measurement and the median and range over three constructions for each dataset. The timing benchmark uses $K = 64$ throughout; training-specific settings are given in Appendix C.

Timing starts with data in the velocity-regression representation and ends with cached time probabilities. It includes endpoint sampling, marginal construction, projections, sorting, smoothing, and calibration. For ImageNet-256, it also includes decoding interpolated latents and extracting frozen Inception features. Device operations are synchronized before and after construction.

Table 5: **One-time prior construction cost.** Times are in seconds. First is the initial measurement; median and range summarize three constructions. The listed settings apply to the timing benchmark.

Dataset	$K/\delta/n/L$	First	Median	Range
Checkerboard	$64/2/12000/64$	0.223	0.016	0.016–0.223
Gaussian25	$64/2/12000/64$	0.016	0.016	0.016–0.016
2rings	$64/2/12000/64$	0.016	0.016	0.016–0.016
CIFAR-10	$64/2/512/64$	10.307	10.261	10.216–10.307
COIL-100	$64/2/512/64$	10.304	10.351	10.304–10.392
ImageNet-256	$64/2/512/64$	239.654	239.081	239.043–239.654

H GENERATED SAMPLES

We compare Uniform and PathSpeed at matched training budgets using the architectures and generation settings in Appendix D. For each dataset, both methods use the same initial noise samples, shown in the same order.

H.1 COIL-100

The COIL-100 comparison complements the feature-connectivity curves in the main text.

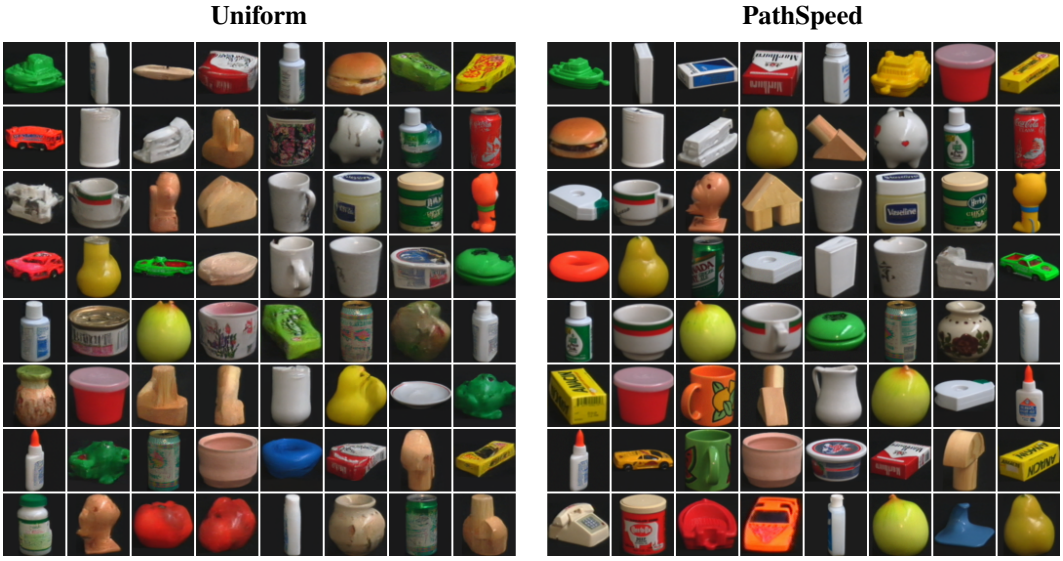


Figure 7: **COIL-100 generation after 60k training updates.** Left: Uniform. Right: PathSpeed. Each side shows 64 samples in an 8×8 grid.

H.2 CIFAR-10

Figure 8 compares the final 100k-update models from the CIFAR-10 setting in the main results, using EMA weights.



Figure 8: **CIFAR-10 generation after 100k training updates.** Left: Uniform. Right: PathSpeed. Each side shows 64 samples in an 8×8 grid.