Steering Dynamical Regimes of Diffusion Models by Breaking Detailed Balance

Haiqi Lu[1,2] **Ying Tang**[1,3,4,5]

[1]Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Yingcai Honor College, University of Electronic Science and Technology of China, Chengdu 611731, China
[3]School of Physics, University of Electronic Science and Technology of China, Chengdu 611731, China
[4]Key Laboratory of Quantum Physics and Photonic Quantum Information, Ministry of Education, University of Electronic Science and Technology of China, Chengdu 611731, China
[5]Non-classical Information Science Basic Discipline Research Center of Sichuan Province, University of Electronic Science and Technology of China, Chengdu 611731, China
Corresponding authors: Haiqi Lu <makarovjim@gmail.com>; Ying Tang <jamestang23@gmail.com>

Abstract

We show that deliberately breaking detailed balance in generative diffusion processes can reshape the reverse-time dynamical timeline of diffusion models. By leveraging the decomposition of the Ornstein–Uhlenbeck process into a symmetric component and a non-reversible anti-symmetric component that generates rotational probability currents, we construct a geometry-aware perturbation that is exponentially optimal for long-time relaxation. We analyze how such non-reversible control reshapes the macroscopic dynamical regimes of the phase transitions recently identified in generative diffusion models. We derive a general matrix criterion for speciation and show that suitable non-reversible perturbations can substantially advance speciation in absolute time. For collapse, we separate its entropy-based definition from a tractable high-dimensional approximation, which depends on covariance-normalized separation between training samples and can therefore acquire perturbation-dependent shifts. Gaussian-mixture and matched drift-specific DDPM experiments on MNIST and ImageNet16 support these findings; a limited MNIST quality-preservation check finds no evident degradation.

1. Introduction

Generative diffusion models are often described as reverse-time stochastic processes driven by a learned score (Song et al., 2021). Recent analyses resolve distinct time windows along this reverse trajectory associated with semantic mode selection, generalization, and memorization (Biroli et al., 2024; Biroli & Mézard, 2023; Ambrogioni, 2025). Following that literature, we use *speciation* for the time at which trajectories commit to a data mode, such as a mixture component or image class. We use *collapse* for the later onset of a memorization-dominated regime in which individual training samples become the relevant attractors (Achilli et al., 2025; Pham et al., 2025; Ye et al., 2026). Their dependence on the forward generator connects the noising dynamics to both generalization and memorization.

The usual linear forward process is reversible and isotropic. This convenient choice fixes a particular route toward the noise distribution. Linear nonequilibrium theory allows a broader family of routes: rotational probability currents can break detailed balance while preserving the asymptotic noise covariance (Ao, 2004; Kwon et al., 2005; Qian, 2013; Godrèche & Luck, 2019). Such currents can improve asymptotic relaxation for Gaussian diffusions (Hwang et al., 1993; Ichiki & Ohzeki, 2013; Lelièvre et al., 2013). We ask how they affect diffusion-model regime timing.

We study how non-reversible forward dynamics reshape speciation and collapse. We formulate a non-reversible OU process that preserves the asymptotic noise covariance, derive a matrix eigenvalue criterion for speciation, and extend

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

a high-dimensional collapse approximation to general linear drifts while retaining the entropic definition of collapse. Gaussian-mixture and trained-DDPM experiments then reveal a strong dependence on geometry. Suitable Lelièvre-style currents move speciation earlier, whereas the weak dense Simple $\mathbf{Q}$ perturbations tested here produce little change in real-data speciation and collapse. These results connect spectral-gap acceleration to finite-time regime displacement.

2. Related Work and Positioning

Dynamical regimes and memorization. Prior work identifies speciation and collapse as distinct diffusion-generation events (Biroli et al., 2024; Biroli & Mézard, 2023), and relates them to associative memory, manifold-driven memorization, entropy signatures, and training-time generalization (Ambrogioni, 2024; Pham et al., 2025; Achilli et al., 2025; 2024; Handke et al., 2026; Bonnaire et al., 2025; Ye et al., 2026). Our question is complementary: how reverse-time regimes move when the forward noising dynamics changes while its asymptotic noise covariance is held fixed.

Inductive bias through the forward generator. Diffusion models encode data geometry through noising and denoising (Stanczuk et al., 2024; Kadkhodaie et al., 2024). Our setting isolates a controlled dynamical bias: for a fixed symmetric drift geometry, adding an anti-symmetric component keeps the asymptotic noise covariance fixed while changing the probability current, transient marginals, and reverse linear drift.

Non-reversible dynamics. Nonequilibrium decompositions separate potential forces, probability currents, and noise (Ao, 2004; Kwon et al., 2005; Qian, 2013). For linear OU processes, anti-symmetric currents can preserve the asymptotic noise covariance (Kwon et al., 2011; Godrèche & Luck, 2019) and improve asymptotic Gaussian relaxation (Hwang et al., 1993; Ichiki & Ohzeki, 2013; Lelièvre et al., 2013). We use this machinery as a forward-process design and evaluate its effects on finite-time speciation and collapse.

Coupled OU dynamics and synchronization gaps. The coupled-OU work of Albrychiewicz et al. (2026a) identifies synchronization gaps from inter-modality coupling. Our intervention adds an anti-symmetric generator component while preserving the asymptotic noise covariance. Their speciation condition is a fixed-point bifurcation; ours is a curvature instability of the evolving density that also covers non-normal drifts. For collapse, forward-noise-normalized separation between training samples controls our high-dimensional approximation, allowing covariance geometry to affect the transition even under equal trace-level volume contraction.

3. Non-Reversible Forward Dynamics

Let the data distribution be supported on samples $\mathbf{a}_\mu \in \mathbb{R}^d$, where d is the data dimension. Let $\mathbf{I}$ be the $d \times d$ identity and $\mathbf{W}_t$ a standard d-dimensional Brownian motion. We use the drift $\mathbf{A} = (\mathbf{I} + \mathbf{Q})\mathbf{U}$, where $\mathbf{U} = \mathbf{U}^\top > 0$ is symmetric positive definite and $\mathbf{Q} = -\mathbf{Q}^\top$ is anti-symmetric. The resulting OU noising process is

$$d\mathbf{x}_t = -\mathbf{A}\mathbf{x}_t\, dt + \sqrt{2}\, d\mathbf{W}_t, \tag{1}$$

The symmetric matrix $\mathbf{U}$ defines the asymptotic noise covariance, while $\mathbf{Q}$ adds a circulating probability current and preserves this covariance. Let $\mathbf{\Sigma}_\infty$ denote this covariance, reached as $t \to \infty$ after continued forward noising. It solves

$$\mathbf{A}\mathbf{\Sigma}_\infty + \mathbf{\Sigma}_\infty \mathbf{A}^\top = 2\mathbf{I}, \qquad \mathbf{\Sigma}_\infty = \mathbf{U}^{-1}. \tag{2}$$

Thus the asymptotic noise covariance is invariant to $\mathbf{Q}$. The limiting distribution remains the same zero-mean Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\infty)$.

The forward solution from an initial state $\mathbf{a}$ is

$$\mathbf{x}_t = e^{-\mathbf{A}t}\mathbf{a} + \sqrt{2}\int_0^t e^{-\mathbf{A}(t-s)} d\mathbf{W}_s, \tag{3}$$

with stochastic covariance

$$\mathbf{\Sigma}_{\mathrm{sto}}(t) = 2\int_0^t e^{-\mathbf{A}s} e^{-\mathbf{A}^\top s}\, ds = \mathbf{\Sigma}_\infty - e^{-\mathbf{A}t}\mathbf{\Sigma}_\infty e^{-\mathbf{A}^\top t}. \tag{4}$$

Under time reversal, the learned reverse SDE has linear term $\mathbf{A}\mathbf{y}$ plus the usual score correction. Hence $\mathbf{Q}$ changes the reverse-time drift field while preserving the asymptotic noise covariance; the score correction is still the score of the forward marginal generated by the chosen drift.

More explicitly, let t_f be the final noising time, let $\tau = t_f - t$ denote reverse time, and let P_t denote the forward marginal density at time t. The reverse state $\mathbf{y}_\tau$ follows

$$d\mathbf{y}_\tau = \left[\mathbf{A}\mathbf{y}_\tau + 2\nabla_{\mathbf{y}} \log P_{t_f-\tau}(\mathbf{y}_\tau)\right] d\tau + \sqrt{2}\, d\mathbf{W}_\tau. \tag{5}$$

The extra non-reversible component is therefore the linear term $\mathbf{Q}\mathbf{U}\mathbf{y}_\tau$. In learned DDPMs we use the same network architecture and training protocol across drifts; the forward linear operator changes, and the reverse dynamics uses the corresponding linear term and learned score. Matching the architecture and training protocol isolates changes in dynamical regime timing from changes in model capacity.

We use two kinds of anti-symmetric perturbations. The first is a simple dense perturbation with $Q_{ij} = q$ for $i < j$ and $Q_{ji} = -q$. The second is the exponentially optimal construction of Lelièvre et al. (Lelièvre et al., 2013), which can equalize the real parts of the spectrum. Writing $\sigma(\mathbf{B})$ for the spectrum of a matrix $\mathbf{B}$ and $\mathrm{Tr}(\mathbf{B})$ for its trace, the construction satisfies

$$\max_{\mathbf{Q}^\top = -\mathbf{Q}} \min \Re\sigma((\mathbf{I} + \mathbf{Q})\mathbf{U}) = \frac{\mathrm{Tr}(\mathbf{U})}{d}. \tag{6}$$

Equation (6) characterizes asymptotic relaxation. The transition criteria below identify the additional objects that control speciation and collapse.

A dense Simple $\mathbf{Q}$ rotates coordinates independently of the data signal and the slow eigenspaces of $\mathbf{U}$, whereas the Lelièvre perturbation redistributes decay rates across the spectrum. For non-normal drift matrices, non-orthogonal eigenvectors can stretch some directions before their eventual decay, causing finite-time behavior to depart from the asymptotic spectral rate. We use the spectral-gap optimum as a reference and evaluate each transition through its corresponding finite-time criterion.

4. Dynamical-Regime Criteria

4.1. Speciation

Speciation is the time at which the evolving distribution becomes locally unstable to separation between data modes. In a two-class problem this is the moment when the density near the midpoint changes from a single class-ambiguous basin to two class-specific basins. Let t_S denote the speciation time. Let $\mathbf{\Sigma}_{\mathrm{B}}$ be the between-mode covariance, which records how component means differ, and let $\mathbf{\Sigma}_{\mathrm{W}}$ be the within-mode covariance. We write $\mathbf{\Sigma}_{\mathrm{loc}}(t)$ for the covariance of one propagated component after adding forward noise. A second-order expansion of $-\log P_t(\mathbf{x})$ around the symmetric point gives the following criterion, where $\lambda_{\min}$ denotes the smallest eigenvalue:

$$\begin{aligned}
&\widetilde{\mathbf{M}}(t) = \mathbf{\Sigma}_{\mathrm{loc}}(t) - e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{B}}e^{-\mathbf{A}^\top t},\\
&\mathbf{\Sigma}_{\mathrm{loc}}(t) = \mathbf{\Sigma}_{\mathrm{sto}}(t) + e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{W}}e^{-\mathbf{A}^\top t}.\\
&\lambda_{\min}\left(\widetilde{\mathbf{M}}(t_S)\right) = 0.
\end{aligned} \tag{7}$$

For a symmetric two-component Gaussian mixture with means $\pm\mathbf{m}$ and within-component covariance $\sigma^2\mathbf{I}$, $\mathbf{\Sigma}_{\mathrm{B}} = \mathbf{m}\mathbf{m}^\top$ and $\mathbf{\Sigma}_{\mathrm{W}} = \sigma^2\mathbf{I}$. Equation (7) compares the local blur of one component against the class-separation vector after it has been propagated by the drift. When $\mathbf{A}$, $\mathbf{\Sigma}_{\mathrm{B}}$, and $\mathbf{\Sigma}_{\mathrm{W}}$ are simultaneously diagonalizable, let Λ denote the leading between-mode variance, w_Λ the within-mode variance in that direction, and d_Λ the corresponding drift eigenvalue. The criterion then reduces to

$$t_S = \frac{\log(1 + d_\Lambda(\Lambda - w_\Lambda))}{2d_\Lambda}, \tag{8}$$

The point-mass limit has $w_\Lambda = 0$ and recovers the previous form. The full matrix form is needed for non-normal non-reversible drifts. The derivation is given in Section C.

The speciation criterion compares two evolving quantities. The local covariance $\mathbf{\Sigma}_{\mathrm{loc}}(t)$ grows as noising accumulates uncertainty and propagates within-class variability, while the signal covariance $e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{B}}e^{-\mathbf{A}^\top t}$ measures the remaining class information. Changing $\mathbf{Q}$ changes both terms through $e^{-\mathbf{A}t}$ and the Lyapunov covariance in (4). With the invariant density fixed, rotating or redistributing the class signal relative to the noise geometry shifts the instability time.

In the isotropic reversible point-mass limit $\mathbf{A} = \mathbf{I}$ and $w_\Lambda = 0$, (8) reduces to $t_S = \frac{1}{2}\log(1+\Lambda)$. With nonzero within-mode variance, the isotropic reversible expression is $t_S = \frac{1}{2}\log(1+\Lambda-w_\Lambda)$, and for large signal strength this matches the familiar $\frac{1}{2}\log\Lambda$ scaling. In the anisotropic reversible case, the decay rate along the leading signal direction determines the crossing. In the non-reversible case, the signal generally evolves across eigendirections, so the crossing must be computed from the full matrix in (7).

To estimate this time experimentally, the speciation figures use the cloning probability $\phi(t)$: two reverse trajectories are continued from the same noisy state at time t, and $\phi(t)$ is the probability that their final class labels agree. Values near $1/2$ indicate no class commitment, while values near 1 indicate that the shared noisy state has already selected a class.

4.2. Collapse

Collapse is a different transition: reverse samples become tied to individual training examples beyond their association with a broader semantic mode. Let t_C denote the collapse time, let $s(t)$ be the entropy density of the true forward distribution, let $s_G(t)$ be the entropy density of one Gaussian neighborhood around a propagated training example, and let $\alpha = \log n/d$ be the entropy density contributed by choosing one of n examples. We write $s^{\mathrm{sep}}(t) = \alpha + s_G(t)$ for the entropy density of the corresponding separated neighborhoods. The entropic criterion says that collapse occurs when the effective volume of the true forward distribution becomes comparable to the volume of these n separated neighborhoods,

$$s(t_C) = s^{\mathrm{sep}}(t_C) = \alpha + s_G(t_C), \qquad \alpha = \frac{\log n}{d}. \tag{9}$$

Equation (9) defines collapse through entropy balance. The REM approximation below yields a prediction in a tractable high-dimensional model, and the experiments track the transition using empirical excess entropy density and a nearest-neighbour switching proxy.

Following high-dimensional analyses of diffusion memorization (Biroli et al., 2024; Achilli et al., 2025), we approximate the many squared distances from a reference sample to the other training samples with a Gaussian Random Energy Model (REM). Define

$$\mathbf{M}_t = e^{-\mathbf{A}t}, \qquad \mathbf{\Gamma}_t = \mathbf{\Sigma}_{\mathrm{sto}}(t), \qquad \mathbf{K}_t = \mathbf{M}_t^\top \mathbf{\Gamma}_t^{-1}\mathbf{M}_t. \tag{10}$$

The matrix $\mathbf{K}_t$ converts a difference between two training samples into a squared distance measured in units of the forward noise covariance $\mathbf{\Gamma}_t$. Thus its eigenvalues $\kappa_j(t)$ are direction-wise signal-to-noise ratios for separating training examples. Under the REM approximation, let σ_0^2 denote the per-coordinate sample-variance scale, so that sample differences follow $\mathbf{a}_1 - \mathbf{a}_\mu \sim \mathcal{N}(0, 2\sigma_0^2\mathbf{I})$. Collapse is then determined implicitly by

$$2\alpha + 1 = \frac{1}{d}\sum_{j=1}^{d}\left[\log(1+2\sigma_0^2\kappa_j(t_C)) \; + \frac{1}{1+2\sigma_0^2\kappa_j(t_C)}\right]. \tag{11}$$

Equation (11) is an implicit REM prediction under the high-dimensional Gaussian assumptions above. It shows how $\mathbf{Q}$ enters collapse: the trace identity $\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{U})$ fixes Euclidean volume contraction, while the REM criterion depends on the full spectrum of $\mathbf{K}_t$. The derivation is given in Section E.

Under $\mathbf{Q} = 0$ and $\mathbf{A} = \mathbf{U} = \mathrm{diag}(u_1, \ldots, u_d)$, all directions decay independently and the familiar coordinate-wise reversible formula is recovered. In that case

$$\kappa_j(t) = \frac{u_j e^{-2u_j t}}{1 - e^{-2u_j t}}, \tag{12}$$

and (11) becomes the standard diagonal implicit REM equation. For general linear drifts, the REM approximation instead makes collapse depend on covariance-normalized separation between training samples, a quantity distinct from the object controlling speciation.

5. Experiments

5.1. Gaussian Mixtures

We first use a two-component Gaussian mixture in $d = 1024$ with means $\pm\mathbf{m}$, within-component covariance $\sigma^2\mathbf{I}$, $\mathbf{U} = \mathrm{diag}(u_1, \ldots, u_d)$, and $u_j = 1 + 8(j-1)/(d-1)$. Speciation is measured by the trajectory-cloning probability

$\phi(t)$. At forward time t, two independent reverse continuations are initialized from the same noisy state $\mathbf{y}_t$; if $c(\mathbf{x}_0)$ denotes the final mixture component, then

$$\phi(t) = \mathbb{P}\left[c(\mathbf{x}_0^{(1)}) = c(\mathbf{x}_0^{(2)}) \,\middle|\, \mathbf{x}_t^{(1)} = \mathbf{x}_t^{(2)} = \mathbf{y}_t\right]. \tag{13}$$

Thus $\phi(t) \approx 1/2$ indicates no class commitment, while $\phi(t) \approx 1$ indicates that the shared noisy state already fixes the final component. For image data below, the same definition is used with $c(\mathbf{x}_0)$ given by a classifier label.

In the Gaussian mixture, the class-separation vector is spread across eigendirections of $\mathbf{U}$, allowing rotations to redistribute signal weight between slow and fast directions. This setup exposes the mechanism predicted by (7); the analytic cloning form is derived in Section D.

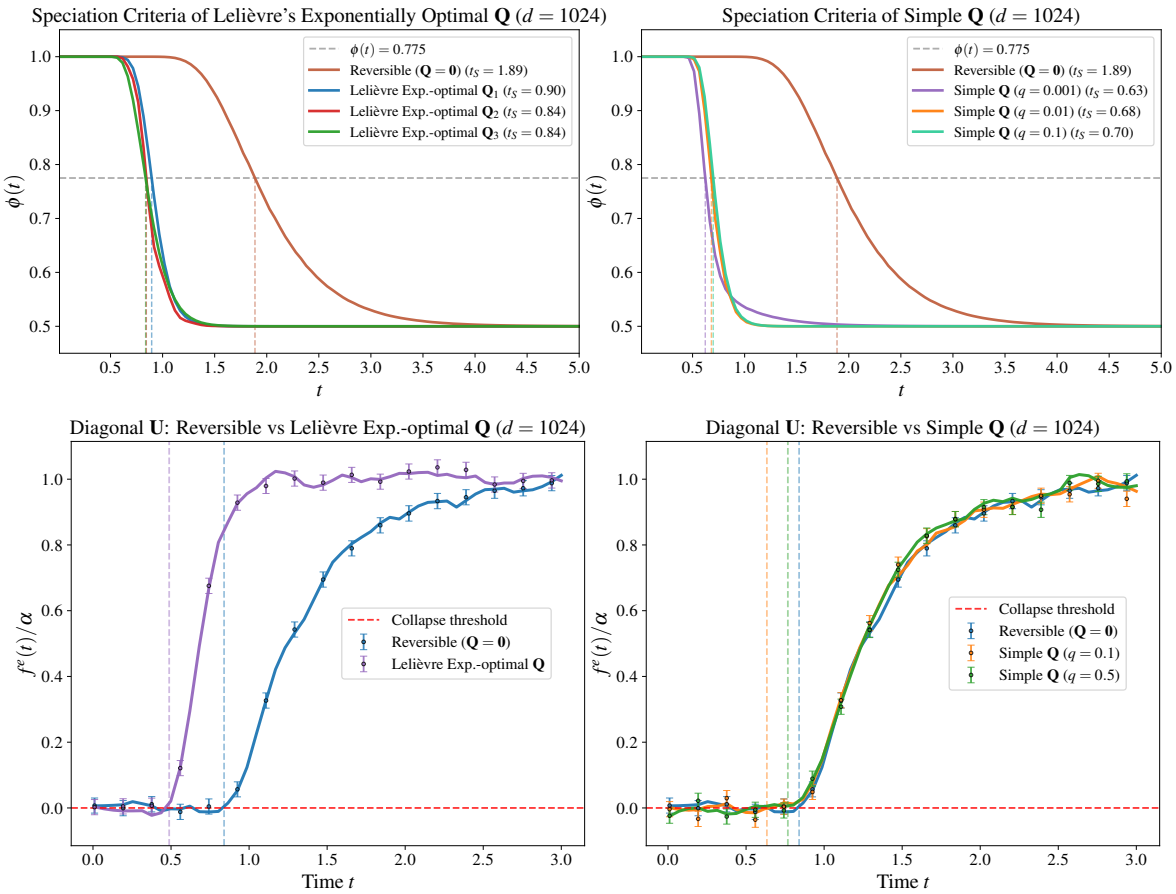

Figure 1. Gaussian mixtures show that non-reversibility can advance speciation, while collapse shifts remain perturbation- and diagnostic-dependent in these tests. Top: Lelièvre Exp.-optimal perturbations use the spectral-gap equalization principle in (6), while Simple $\mathbf{Q}$ uses $Q_{ij} = q$ for $i < j$ and $Q_{ji} = -q$; both can shift the cloning transition when they couple to the mixture signal. The normalized curves of Gaussian mixture in Figure 5 cross at $(t/t_S, \phi) = (1, 0.775)$, motivating the plotted speciation threshold. Bottom: collapse responds through the covariance-normalized spectrum in (10), with smaller shifts whose magnitude depends on the diagnostic.

Figure 1 shows that non-reversible drift can substantially advance speciation when $\mathbf{Q}$ changes the propagated signal and local covariance in (7). Collapse displacement is smaller and diagnostic-dependent: weak simple perturbations induce modest shifts, while Lelièvre $\mathbf{Q}$ can move the zero-crossing by reshaping $\mathbf{K}_t$. After rescaling time by the matrix-predicted t_S, the normalized curves in Figure 5 align near the common cloning threshold; the raw curves show the displacement in absolute time.

5.2. Trained DDPMs on Image Data

We train matched drift-specific DDPMs with four forward drifts: standard DDPM ($\mathbf{A} = \mathbf{I}$), reversible anisotropic control ($\mathbf{Q} = 0, \mathbf{A} = \mathbf{U}$), Lelièvre Exp.-optimal $\mathbf{Q}$, and Simple $\mathbf{Q}$. They share the same U-Net architecture, objective, and $350{,}000$-step budget while using drift-specific forward noising and corresponding reverse sampling. Anisotropic operators are defined in the full PCA basis and rotated back to pixel space with all dimensions retained.

We test whether the Gaussian-mixture effect persists on MNIST 0 vs 7 and a binary ImageNet16 task. Speciation uses the same trajectory-cloning protocol, with a ResNet-18 classifier to identify the final class. For collapse, let $s^e(t)$ be the empirical forward entropy density and $s_G(t)$ the entropy density of the corresponding single-Gaussian reference. We measure the empirical excess entropy density

$$f^e(t) = s^{\mathrm{sep}}(t) - s^e(t), \qquad s^{\mathrm{sep}}(t) = \frac{\log n}{d} + s_G(t). \tag{14}$$

We also use a nearest-neighbour switching proxy $\hat{t}_C$, defined by the last reverse-time point at which the nearest training index changes. Together, these two measurements track collapse in finite-dimensional trained models.

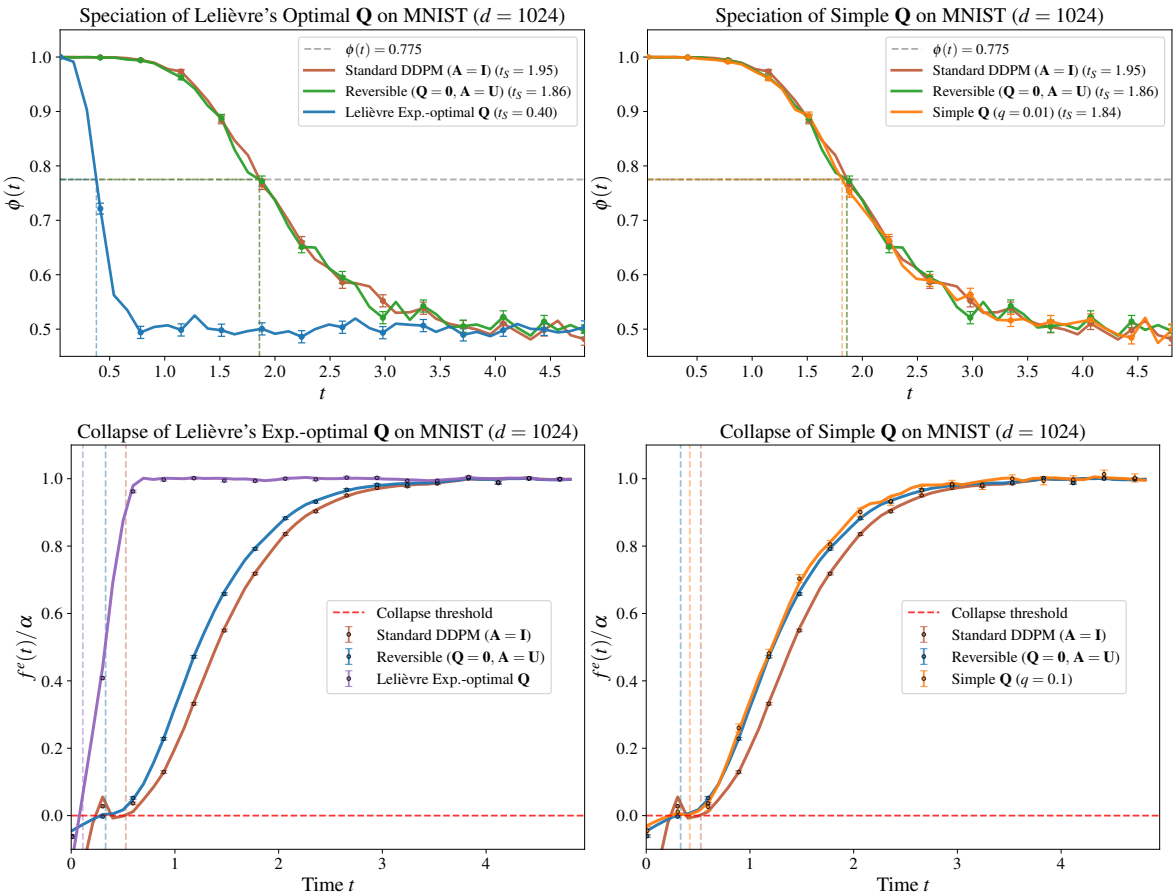

Figure 2. MNIST shows perturbation-dependent displacement of the two regimes. Top: Lelièvre $\mathbf{Q}$ advances t_S from 1.86 to 0.40, while Simple $\mathbf{Q}$ stays near the reversible reference with $t_S^{\mathrm{SQ}} = 1.84$. Bottom: collapse displacement is modest for Simple $\mathbf{Q}$ and stronger for Lelièvre $\mathbf{Q}$.

The standard DDPM supplies the isotropic baseline, and the reversible anisotropic model isolates the effect of changing $\mathbf{U}$ at $\mathbf{Q} = 0$. At the same fixed $\mathbf{U}$, Simple $\mathbf{Q}$ introduces a generic probability current, while the Lelièvre perturbation uses the drift spectrum to target slow directions. With architecture and training budget matched, this comparison separates the effects of anisotropic decay, detailed-balance breaking, and geometry-aware current design.

Figure 2 shows that regime displacement in the tested MNIST DDPMs depends strongly on perturbation geometry, and that the two regimes respond differently. Lelièvre $\mathbf{Q}$ moves the cloning transition strongly, whereas Simple $\mathbf{Q}$

remains close to the reversible reference and produces only modest collapse displacement. In MNIST, about 98.6% of the class-separation energy lies in the leading principal component, so a weak generic rotation has little signal mass to redistribute.

The near-overlap between the Simple $\mathbf{Q}$ and reversible cloning curves follows from the signal term in (7). Let $\mathbf{m}$ denote the empirical class-separation vector, and let $\mathbf{v}_k$ be drift modes satisfying $\mathbf{A}\mathbf{v}_k = \lambda_k \mathbf{v}_k$, with $\mathbf{m} = \sum_k \alpha_k \mathbf{v}_k$. Then, up to finite-time prefactors,

$$\|e^{-\mathbf{A}t}\mathbf{m}\|^2 \approx \sum_{k=1}^{d} |\alpha_k|^2 e^{-2\mathrm{Re}(\lambda_k)t}. \tag{15}$$

Simple $\mathbf{Q}$ shifts speciation by moving appreciable signal weight between directions with different decay rates, a mechanism that is more effective in the synthetic mixture than in the low-dimensional real-data signal sectors.

On ImageNet16, Figure 3 shows the same hierarchy. Lelièvre $\mathbf{Q}$ advances speciation, whereas Simple $\mathbf{Q}$ nearly overlaps with the reversible reference because about 97.1% of the class-separation energy lies in the first 20 principal components.

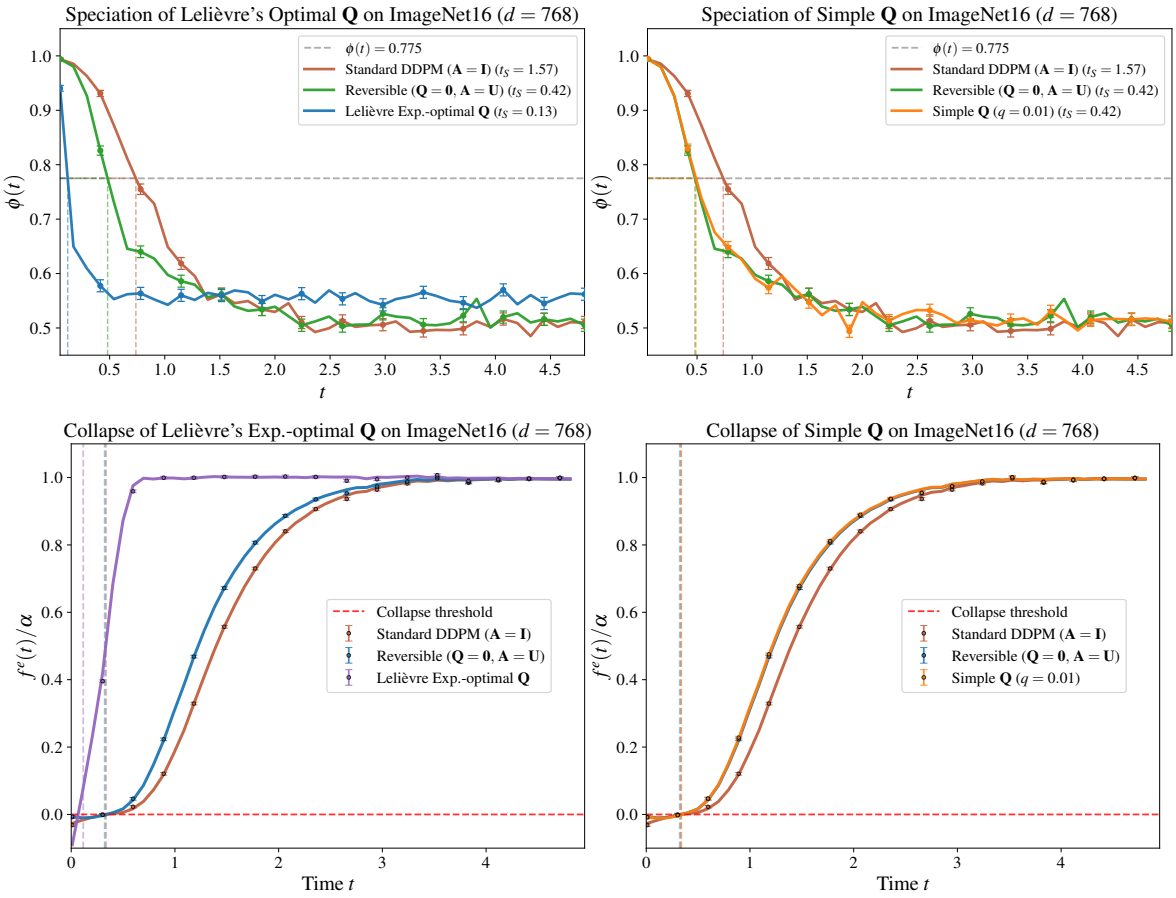

Figure 3. ImageNet16 reproduces the perturbation-dependent hierarchy observed on MNIST. Top: Lelièvre $\mathbf{Q}$ advances t_S from 0.424 to 0.131, while Simple $\mathbf{Q}$ with $q = 0.01$ remains at $t_S^{\mathrm{SQ}} = 0.424$. Bottom: weak Simple $\mathbf{Q}$ leaves collapse nearly unchanged, while Lelièvre $\mathbf{Q}$ moves the finite-dimensional marker.

The ImageNet16 collapse diagnostics show a weaker and less uniform response. The entropy zero-crossings are approximately 0.323, 0.337, and 0.116 for reversible, Simple $\mathbf{Q}$, and Lelièvre curves; nearest-neighbour estimates are $\hat{t}_C = 0.214 \pm 0.141$, 0.214 ± 0.186, and 0.011 ± 0.013. Together with Figure 2, these results agree with the REM interpretation: substantial changes in the covariance-normalized separation spectrum move the collapse diagnostics, with direction and magnitude varying across the binary-image experiments.

For the reversible anisotropic, Lelièvre, and Simple $\mathbf{Q}$ models, the legend values of t_S are computed independently from the matrix criterion (7). These models share the same $\mathbf{U}$, and their rescaled cloning curves satisfy $\phi(t) \approx 0.775$ near $t/t_S = 1$ in Figure 5. The Standard DDPM uses $\mathbf{U} = \mathbf{I}$ and serves separately as the isotropic baseline.

5.3. MNIST Generation-Quality Preservation Check

We supplement the regime-timing measurements with a sample-quality check on the MNIST 0-versus-7 task. For each drift, we draw 1000 samples from its $350{,}000$-step final EMA checkpoint using the native full $T = 1000$ ancestral schedule, the same seed, and the same batch boundaries. Each checkpoint is trained separately under the matched architecture, data, optimization, and training-budget protocol described above, so the comparison includes the learned score associated with each drift.

All methods are evaluated by one common binary ResNet-18 evaluator. It obtains 99.90% accuracy on the held-out official MNIST 0-versus-7 test split and exposes a 512-dimensional penultimate feature. Table 1 reports evaluator-predicted class balance, average maximum-softmax confidence, and nearest-neighbour distances after ℓ_2-normalizing this feature. Class balance measures whether a drift favors one terminal class, and confidence tracks evaluator-level recognizability. The generated-to-test distance measures proximity to held-out examples in the shared feature space, while the generated-to-train distance tracks proximity to the training set.

Table 1. MNIST generation-quality preservation check with 1000 samples per method. NN-test and NN-train are mean nearest-neighbour distances in the common evaluator feature space. The real test-to-train NN mean is 0.0509. All four methods produce similar values on these task-specific metrics; perceptual quality and diversity require broader metrics and datasets.

Method	Pred. 0 (%)	Pred. 7 (%)	Confidence	NN-test	NN-train
Standard DDPM	50.6	49.4	0.9997	0.0643	0.0363
Reversible	50.1	49.9	0.9998	0.0635	0.0407
Lelièvre $\mathbf{Q}$	48.9	51.1	0.9987	0.0647	0.0471
Simple $\mathbf{Q}$	50.4	49.6	0.9997	0.0644	0.0408

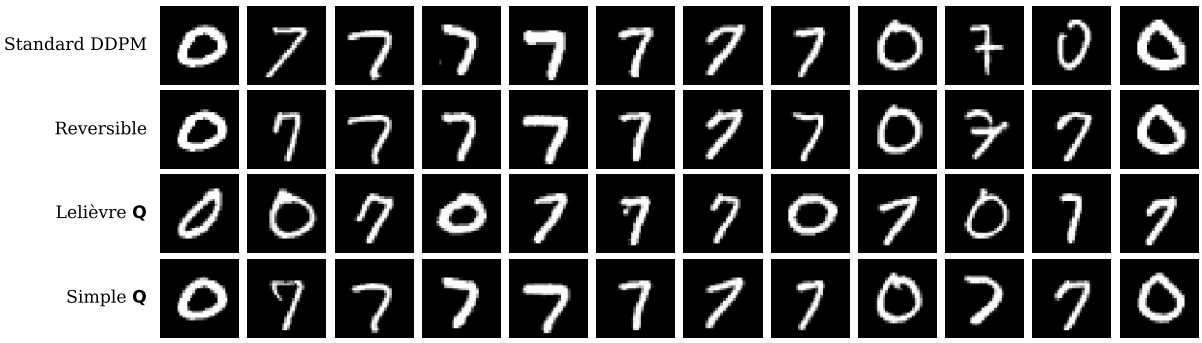

Figure 4. Fixed-index qualitative comparison for the MNIST quality-preservation check. Each row shows the first twelve samples in stored order from the corresponding 1000-sample set. All rows use the same seed policy, final EMA policy, and full $T = 1000$ sampling schedule. Digit recognizability appears comparable across the modified dynamics and Standard DDPM.

The predicted class proportions remain close to the balanced target for all four methods, and all average confidences exceed 0.998. Their generated-to-test means lie in the narrow range 0.0635–0.0647. The fixed-index examples in Figure 4 show a similar level of digit recognizability. Across class balance, confidence, and feature distance, the modified dynamics remain close to the Standard DDPM baseline on this binary task. The largest class-balance deviation from 50% is 1.1 percentage points, and the spread in average confidence is 0.0011. These small differences agree with the visual comparison, in which every row contains recognizable examples from both classes. The small spread in generated-to-test distance indicates that the evaluator places samples from all four models at similar distances from held-out digits. Generated-to-train distance varies more across drifts, from 0.0363 to 0.0471, motivating a separate study of training-set proximity. Broader evaluations are required to resolve perceptual-quality differences and memorization behavior.

6. Discussion

Breaking detailed balance changes the finite-time reverse trajectory even when the asymptotic noise distribution is unchanged. Speciation occurs when $\lambda_{\min}(\widetilde{\mathbf{M}}(t_S)) = 0$, so a geometry-aware $\mathbf{Q}$ can advance semantic commitment by changing the propagated class signal and local covariance. Collapse follows an entropic criterion, and its REM approximation depends on the spectrum of $\mathbf{K}_t = e^{-\mathbf{A}^\top t}\mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}$. Consequently, equal trace-level volume contraction can coexist with different collapse behavior. The propagated between-class signal and the covariance-normalized sample-separation spectrum thus offer separate targets for controlling the two transitions.

Synchronization gaps in coupled-OU systems and Diffusion Transformers arise from coupling-induced relaxation channels (Albrychiewicz et al., 2026a;b). Here, the asymptotic noise covariance remains fixed while the probability current changes. Both interventions can alter mode-selection and memorization times; for our linear dynamics, the REM spectrum $\{\kappa_j(t)\}$ encodes the collapse-side response.

The forward generator acts as an inductive bias beyond its role as a variance schedule. Forward processes with the same asymptotic noise covariance can expose different amounts of class evidence at intermediate noise levels. A geometry-aware current rotates or redistributes this evidence before noising erases it, causing the learned reverse trajectory to commit to a semantic mode at a different time. Simple $\mathbf{Q}$ uses generic dense rotations independent of the class-separation directions, and its cloning curves remain close to the reversible reference on the image tasks. The magnitude of the probability current alone therefore says little about the timing shift; its orientation relative to the data signal and drift eigenspaces is equally important.

Spectral relaxation and regime timing capture different meanings of "acceleration." Lelièvre $\mathbf{Q}$ accelerates forward OU relaxation in the spectral sense, while the experiments measure shifts in reverse-time speciation and collapse. Earlier speciation moves semantic commitment along the fixed schedule. Earlier collapse moves the onset of the regime associated with individual training examples, with consequences that depend on the intended generation behavior. A useful forward generator may favor one timing shift over another according to the desired balance between semantic commitment and sample specificity.

The MNIST check in Section 5.3 finds comparable digit recognizability and evaluator-space fidelity across the observed timing shifts. Precision/recall-style diversity metrics, FID, and multiclass evaluations can test how these shifts affect generation beyond the binary MNIST and ImageNet16 settings considered here.

Data geometry explains the contrast between the synthetic and image experiments. In the Gaussian mixtures, the signal is spread across drift eigendirections by construction, allowing rotational currents to transfer substantial signal weight between slow and fast directions. In MNIST and ImageNet16, much of the measured class-separation energy lies in a small principal subspace. A weak dense $\mathbf{Q}$ can then create probability current while barely changing the signal-to-noise geometry that controls the observed transitions.

Useful non-reversible generators should therefore account for the signal directions and noise levels of interest. Current concentrated in directions orthogonal to the semantic signal may have little effect on class commitment, even if it strongly violates detailed balance. A perturbation matched to slow signal directions can produce a much larger timing shift. The Lelièvre construction supplies a geometry-aware reference; learned or data-adaptive currents could target task-relevant subspaces directly and vary their mixing strength across the noising trajectory.

The analysis covers linear forward processes, a high-dimensional approximation for collapse, and binary trained-image tasks. Extending the construction to multiclass data, nonlinear drifts, or learned state-dependent generators requires controlling both the invariant measure and the induced score. Discretization is another open issue: numerical-stability and compute constraints may shift the optimum for a finite-step sampler away from its continuous-time counterpart. Moreover, finite-time behavior of non-normal drifts depends on prefactors as well as spectral exponents. Evaluating these effects on larger datasets and practical samplers will determine how far spectral design principles transfer beyond the present setting.

References

Achilli, B., Ventura, E., Silvestri, G., Pham, B., Raya, G., Krotov, D., Lucibello, C., and Ambrogioni, L. Losing dimensions: Geometric memorization in generative diffusion. *arXiv*, 2024. URL https://arxiv.org/abs/2410.08727.

Achilli, B., Ambrogioni, L., Lucibello, C., Mézard, M., and Ventura, E. Memorization and generalization in generative diffusion under the manifold hypothesis. *J. Stat. Mech.: Theory Exp.*, 2025(7):073401, 2025. doi: 10.1088/1742-5468/ade136. URL https://iopscience.iop.org/article/10.1088/1742-5468/ade136/meta.

Albrychiewicz, E., Franco Valiente, A., and Chen, L.-C. Dynamical regimes of multimodal diffusion models. *arXiv*, 2026a. URL https://arxiv.org/abs/2602.04780.

Albrychiewicz, E., Franco Valiente, A., Chen, L.-C., and Zhao, V. Z. Interpreting the synchronization gap: The hidden mechanism inside diffusion transformers. *arXiv preprint arXiv:2603.20987*, 2026b. doi: 10.48550/arXiv.2603.20987. URL https://arxiv.org/abs/2603.20987.

Ambrogioni, L. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024. doi: 10.3390/e26050381. URL https://doi.org/10.3390/e26050381.

Ambrogioni, L. The statistical thermodynamics of generative diffusion models: Phase transitions, symmetry breaking and critical instability. *Entropy*, 27(3):291, 2025. doi: 10.3390/e27030291. URL https://doi.org/10.3390/e27030291.

Ao, P. Potential in stochastic differential equations: novel construction. *J. Phys. A*, 37:L25, 2004. doi: 10.1088/0305-4470/37/3/L01. URL http://iopscience.iop.org/0305-4470/37/3/L01.

Biroli, G. and Mézard, M. Generative diffusion in very large dimensions. *J. Stat. Mech.: Theory Exp.*, 2023 (9):093402, 2023. doi: 10.1088/1742-5468/acf8ba. URL https://iopscience.iop.org/article/10.1088/1742-5468/acf8ba/meta.

Biroli, G., Bonnaire, T., De Bortoli, V., and Mézard, M. 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.

Bonnaire, T., Urfin, R., Biroli, G., and Mezard, M. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BSZqpqgqM0.

Godrèche, C. and Luck, J.-M. Characterising the nonequilibrium stationary states of ornstein-uhlenbeck processes. *J. Phys. A: Math. Theor.*, 52(3):035002, 2019. doi: 10.1088/1751-8121/aaf190. URL https://doi.org/10.1088/1751-8121/aaf190.

Handke, F., Stančević, D., Koulischer, F., Demeester, T., and Ambrogioni, L. The entropic signature of class speciation in diffusion models. *arXiv preprint arXiv:2602.09651*, 2026. doi: 10.48550/arXiv.2602.09651. URL https://arxiv.org/abs/2602.09651.

Hwang, C.-R., Hwang-Ma, S.-Y., and Sheu, S.-J. Accelerating gaussian diffusions. *Ann. Appl. Probab.*, 3(3):897–913, 1993. doi: 10.1214/aoap/1177005371. URL https://doi.org/10.1214/aoap/1177005371.

Ichiki, A. and Ohzeki, M. Violation of detailed balance accelerates relaxation. *Phys. Rev. E*, 88(2):020101, 2013. doi: 10.1103/PhysRevE.88.020101. URL https://doi.org/10.1103/PhysRevE.88.020101.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *ICLR*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0.

Kwon, C., Ao, P., and Thouless, D. J. Structure of stochastic dynamics near fixed points. *Proc. Natl. Acad. Sci. U.S.A.*, 102(37):13029–13033, 2005. doi: 10.1073/pnas.0506347102. URL https://doi.org/10.1073/pnas.0506347102.

Kwon, C., Noh, J. D., and Park, H. Non-equilibrium fluctuations for linear diffusion dynamics. *Phys. Rev. E*, 83 (6):061145, 2011. doi: 10.1103/PhysRevE.83.061145. URL https://doi.org/10.1103/PhysRevE.83.061145.

Lelièvre, T., Nier, F., and Pavliotis, G. A. Optimal non-reversible linear drift for the convergence to equilibrium of a diffusion. *J. Stat. Phys.*, 152(2):237–274, 2013. doi: 10.1007/s10955-013-0769-x. URL https://link.springer.com/article/10.1007/s10955-013-0769-x.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory networks. In *New Frontiers in Associative Memories*, 2025. URL https://openreview.net/forum?id=IWZnhP3YgK.

Qian, H. A decomposition of irreversible diffusion processes without detailed balance. *J. Math. Phys.*, 54(5):053302, 2013. doi: 10.1063/1.4803847. URL https://doi.org/10.1063/1.4803847.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Stanczuk, J. P., Batzolis, G., Deveney, T., and Schönlieb, C.-B. Diffusion models encode the intrinsic dimension of data manifolds. In Salakhutdinov, R., Kolter, Z., Heller, K., Weller, A., Oliver, N., Scarlett, J., and Berkenkamp, F. (eds.), *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46412–46440. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/stanczuk24a.html.

Ye, Z., Zhu, Q., Tao, M., and Chen, M. Provable separations between memorization and generalization in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=42gfTZzyvV.

A. Forward OU Algebra and Time Reversal

For a dataset of n points $\mathbf{a}_\mu \in \mathbb{R}^d$, the empirical data law is

$$P_0^e(\mathbf{a}) = \frac{1}{n}\sum_{\mu=1}^{n} \delta(\mathbf{a} - \mathbf{a}_\mu). \tag{16}$$

The generalized forward noising process is

$$d\mathbf{x}_t = -\mathbf{A}\mathbf{x}_t\, dt + \sqrt{2}\, d\mathbf{W}_t, \qquad \mathbf{A} = (\mathbf{I} + \mathbf{Q})\mathbf{U}, \qquad \mathbf{U} = \mathbf{U}^\top > 0, \quad \mathbf{Q} = -\mathbf{Q}^\top. \tag{17}$$

The explicit solution from $\mathbf{x}_0 = \mathbf{a}$ is

$$\mathbf{x}_t = e^{-\mathbf{A}t}\mathbf{a} + \sqrt{2}\int_0^t e^{-\mathbf{A}(t-s)} d\mathbf{W}_s. \tag{18}$$

The asymptotic noise covariance $\mathbf{\Sigma}_\infty$ solves the Lyapunov equation

$$\mathbf{A}\mathbf{\Sigma}_\infty + \mathbf{\Sigma}_\infty \mathbf{A}^\top = 2\mathbf{I}. \tag{19}$$

For the chosen decomposition, $\mathbf{\Sigma}_\infty = \mathbf{U}^{-1}$ because

$$(\mathbf{I} + \mathbf{Q})\mathbf{U}\mathbf{U}^{-1} + \mathbf{U}^{-1}\mathbf{U}(\mathbf{I} + \mathbf{Q})^\top = \mathbf{I} + \mathbf{Q} + \mathbf{I} - \mathbf{Q} = 2\mathbf{I}. \tag{20}$$

Thus the anti-symmetric perturbation changes the probability current while preserving the asymptotic noise covariance.

The stochastic covariance in (18) is

$$\mathbf{\Sigma}_{\text{sto}}(t) = 2\int_0^t e^{-\mathbf{A}s} e^{-\mathbf{A}^\top s}\, ds. \tag{21}$$

It can also be written in closed Lyapunov form:

$$\mathbf{\Sigma}_{\text{sto}}(t) = \mathbf{\Sigma}_\infty - e^{-\mathbf{A}t}\mathbf{\Sigma}_\infty e^{-\mathbf{A}^\top t}. \tag{22}$$

Indeed, differentiating $e^{-\mathbf{A}t}\mathbf{\Sigma}_\infty e^{-\mathbf{A}^\top t}$ gives

$$\frac{d}{dt}\left(e^{-\mathbf{A}t}\mathbf{\Sigma}_\infty e^{-\mathbf{A}^\top t}\right) = -e^{-\mathbf{A}t}\left(\mathbf{A}\mathbf{\Sigma}_\infty + \mathbf{\Sigma}_\infty\mathbf{A}^\top\right)e^{-\mathbf{A}^\top t} = -2e^{-\mathbf{A}t}e^{-\mathbf{A}^\top t}, \tag{23}$$

and integration from 0 to t gives (22).

For time reversal, let t_f be the final noising time, let $\tau = t_f - t$, let $\mathbf{y}_\tau = \mathbf{x}_{t_f-\tau}$, and let P_t denote the forward marginal density. For a forward diffusion $d\mathbf{x}_t = \mathbf{f}(\mathbf{x}_t, t)\, dt + \sqrt{2}\, d\mathbf{W}_t$, the reverse-time process satisfies

$$d\mathbf{y}_\tau = [-\mathbf{f}(\mathbf{y}_\tau, t) + 2\nabla_{\mathbf{y}} \log P_t(\mathbf{y}_\tau)]\, d\tau + \sqrt{2}\, d\mathbf{W}_\tau, \qquad t = t_f - \tau. \tag{24}$$

Specializing to $\mathbf{f}(\mathbf{x}, t) = -\mathbf{A}\mathbf{x}$ gives

$$d\mathbf{y}_\tau = \left[\mathbf{A}\mathbf{y}_\tau + 2\nabla_{\mathbf{y}} \log p_{t_f-\tau}(\mathbf{y}_\tau)\right] d\tau + \sqrt{2}\, d\mathbf{W}_\tau. \tag{25}$$

Relative to the reversible linear model, the reverse drift gains the linear term $\mathbf{Q}\mathbf{U}\mathbf{y}_\tau$; the score term remains the score of the forward marginal generated by the chosen noising process.

B. Anti-Symmetric Perturbation Families

The optimized non-reversible perturbation used in the experiments follows the construction of Lelièvre et al. (Lelièvre et al., 2013). For a fixed positive definite $\mathbf{U}$, the spectral optimization problem is

$$\max_{\mathbf{Q}^\top = -\mathbf{Q}} \min \Re\sigma((\mathbf{I} + \mathbf{Q})\mathbf{U}). \tag{26}$$

The maximal achievable value is the mean curvature:

Proposition 1 (Maximal spectral gap (Lelièvre et al., 2013)). *For* $\mathbf{A} = (\mathbf{I} + \mathbf{Q})\mathbf{U}$ *with* $\mathbf{Q}^\top = -\mathbf{Q}$ *and* $\mathbf{U} = \mathbf{U}^\top > 0$,

$$\max_{\mathbf{Q}^\top = -\mathbf{Q}} \min \Re\sigma(\mathbf{A}) = \frac{\mathrm{Tr}(\mathbf{U})}{d}. \tag{27}$$

The constructive algorithm starts from an orthonormal basis and repeatedly identifies one vector whose quadratic level $(\psi, \mathbf{U}\psi)$ lies above $\mathrm{Tr}(\mathbf{U})/d$ and one whose level lies below it. A two-dimensional rotation within their span produces a new vector whose level is exactly $\mathrm{Tr}(\mathbf{U})/d$, followed by Gram–Schmidt orthonormalization of the remaining directions. Repeating these pair rotations yields an equilibrated basis from which an anti-symmetric perturbation attaining the maximal spectral gap can be constructed.

We use this construction as a geometry-aware reference for transient regime control. The theorem characterizes the asymptotic exponential relaxation rate of the forward OU process, while (7) and (11) describe the finite-time reverse-process regimes of speciation and collapse. Non-normal drift matrices can also stretch some directions before asymptotic decay, adding finite-time behavior beyond the spectral exponent.

For comparison, the Simple perturbation family uses a fixed dense pattern:

$$Q_{ij} = q \quad (i < j), \qquad Q_{ji} = -q, \qquad Q_{ii} = 0. \tag{28}$$

It breaks detailed balance independently of the slow eigenspaces of $\mathbf{U}$ and the empirical class-separation direction. We use it as a generic reference against perturbations designed from the drift geometry.

C. Detailed Speciation Derivation

C.1. Landau Expansion

To derive (7), write the forward marginal as

$$P_t(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma}_{\mathrm{sto}}(t))}} \exp\left(-\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)\mathbf{x} + g(\mathbf{x})\right), \tag{29}$$

where

$$g(\mathbf{x}) = \log \int d\mathbf{a}\, P_0(\mathbf{a}) \exp\Big[-\frac{1}{2}(e^{-\mathbf{A}t}\mathbf{a})^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)(e^{-\mathbf{A}t}\mathbf{a}) + \mathbf{x}^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\mathbf{a}\Big]. \tag{30}$$

Let

$$\Phi(\mathbf{a}; \mathbf{x}, t) = -\frac{1}{2}\mathbf{a}^\top e^{-\mathbf{A}^\top t}\mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\mathbf{a} + \mathbf{x}^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\mathbf{a}. \tag{31}$$

When the propagated signal is small enough for a local expansion, $\exp(\Phi) = 1 + \Phi + \frac{1}{2}\Phi^2 + \cdots$. Integrating term by term with respect to P_0 gives the $\mathbf{x}$-dependent contributions

$$\langle \mathrm{linear} \rangle = \mathbf{x}^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\langle \mathbf{a} \rangle, \tag{32}$$

$$\langle \mathrm{quadratic} \rangle = \frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\langle \mathbf{a}\mathbf{a}^\top \rangle e^{-\mathbf{A}^\top t}\mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)\mathbf{x}. \tag{33}$$

Using $\log(1+z) = z - \frac{1}{2}z^2 + \cdots$, the second-order term from the square of the linear contribution subtracts

$$\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\langle \mathbf{a} \rangle \langle \mathbf{a} \rangle^\top e^{-\mathbf{A}^\top t}\mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)\mathbf{x}. \tag{34}$$

Thus the total quadratic form in $-\log P_t(\mathbf{x})$ has curvature

$$\mathbf{M}(t) = \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t) - \mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t)e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{data}}e^{-\mathbf{A}^\top t}\mathbf{\Sigma}_{\mathrm{sto}}^{-1}(t), \tag{35}$$

where

$$\mathbf{\Sigma}_{\mathrm{data}} = \langle \mathbf{a}\mathbf{a}^\top \rangle - \langle \mathbf{a} \rangle \langle \mathbf{a} \rangle^\top. \tag{36}$$

For speciation with nonzero within-component covariance, one must distinguish the local width of one propagated component from the between-component signal. Let $\mathbf{\Sigma}_{\mathrm{W}}$ denote the covariance within one component. For a symmetric two-component Gaussian mixture with means $\pm\mathbf{m}$ and isotropic within-component covariance $\sigma^2\mathbf{I}$,

$$\mathbf{\Sigma}_{\mathrm{data}} = \sigma^2\mathbf{I} + \mathbf{m}\mathbf{m}^\top, \qquad \mathbf{\Sigma}_{\mathrm{W}} = \sigma^2\mathbf{I}, \qquad \mathbf{\Sigma}_{\mathrm{B}} = \mathbf{m}\mathbf{m}^\top. \tag{37}$$

The within-component part describes local component width; after the forward OU map it contributes to the covariance of one component,

$$\mathbf{\Sigma}_{\mathrm{loc}}(t) = \mathbf{\Sigma}_{\mathrm{sto}}(t) + e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{W}}e^{-\mathbf{A}^\top t}. \tag{38}$$

The between-component part $\mathbf{\Sigma}_{\mathrm{B}}$ drives the order-parameter instability. The stability loss is therefore equivalent to

$$\lambda_{\min}\Big(\mathbf{\Sigma}_{\mathrm{loc}}(t) - e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{B}}e^{-\mathbf{A}^\top t}\Big) = 0. \tag{39}$$

The equivalence follows from a congruence transformation by the positive-definite matrix $\mathbf{\Sigma}_{\mathrm{loc}}^{-1/2}(t)$.

C.2. Simultaneously Diagonalizable Limit

Assume that $\mathbf{A}$, $\mathbf{\Sigma}_{\mathrm{B}}$, and $\mathbf{\Sigma}_{\mathrm{W}}$ are simultaneously orthogonally diagonalizable:

$$\mathbf{P}^\top\mathbf{A}\mathbf{P} = \mathrm{diag}(d_1,\dots,d_d), \qquad \mathbf{P}^\top\mathbf{\Sigma}_{\mathrm{B}}\mathbf{P} = \mathrm{diag}(c_1,\dots,c_d), \qquad \mathbf{P}^\top\mathbf{\Sigma}_{\mathrm{W}}\mathbf{P} = \mathrm{diag}(w_1,\dots,w_d), \qquad d_i > 0. \tag{40}$$

In this basis,

$$\mathbf{\Sigma}_{\mathrm{loc}}(t) = \mathbf{P}\,\mathrm{diag}\left(\frac{1-e^{-2d_1t}}{d_1} + w_1e^{-2d_1t}, \dots, \frac{1-e^{-2d_dt}}{d_d} + w_de^{-2d_dt}\right)\mathbf{P}^\top \tag{41}$$

and

$$e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{B}}e^{-\mathbf{A}^\top t} = \mathbf{P}\,\mathrm{diag}\big(c_1e^{-2d_1t}, \dots, c_de^{-2d_dt}\big)\,\mathbf{P}^\top. \tag{42}$$

The scalar stability entries are therefore

$$\widetilde{m}_i(t) = \frac{1-e^{-2d_it}}{d_i} + (w_i - c_i)e^{-2d_it}. \tag{43}$$

If $\Lambda = c_i$ is the leading between-component variance, $w_\Lambda = w_i$ is the corresponding within-component variance, and d_Λ is the drift eigenvalue in that direction, setting $\widetilde{m}_\Lambda(t_S) = 0$ gives

$$\frac{1-e^{-2d_\Lambda t_S}}{d_\Lambda} = (\Lambda - w_\Lambda)e^{-2d_\Lambda t_S}, \qquad t_S = \frac{\log(1 + d_\Lambda(\Lambda - w_\Lambda))}{2d_\Lambda}. \tag{44}$$

The point-mass expression follows by setting $w_\Lambda = 0$.

D. Gaussian-Mixture Cloning Observable

Consider two equally weighted Gaussian clusters with means $\pm\mathbf{m}$ and initial within-cluster covariance $\sigma^2\mathbf{I}$. Under the linear forward process, the marginal at time t can be written as

$$P_t(\mathbf{y}) = \frac{1}{2}G(\mathbf{y}, \boldsymbol{\mu}(t), \mathbf{\Sigma}_G(t)) + \frac{1}{2}G(\mathbf{y}, -\boldsymbol{\mu}(t), \mathbf{\Sigma}_G(t)), \tag{45}$$

where

$$\boldsymbol{\mu}(t) = e^{-\mathbf{A}t}\mathbf{m}, \qquad \mathbf{\Sigma}_G(t) = \mathbf{\Sigma}_{\mathrm{sto}}(t) + \sigma^2e^{-\mathbf{A}t}e^{-\mathbf{A}^\top t}. \tag{46}$$

Here $G(\mathbf{y}, \boldsymbol{\mu}, \mathbf{\Sigma})$ denotes the d-dimensional Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$.

Given a noisy state $\mathbf{y}$ at time t, the posterior probabilities of the two components determine whether two independent reverse clones are likely to terminate in the same component. Integrating over $\mathbf{y}$ gives

$$\begin{aligned}\phi(t) &= \frac{1}{2}\int d\mathbf{y}\,\frac{G(\mathbf{y},\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t))^2 + G(\mathbf{y},-\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t))^2}{G(\mathbf{y},\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t)) + G(\mathbf{y},-\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t))} \\ &= 1 - \int d\mathbf{y}\,\frac{G(\mathbf{y},\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t))\,G(\mathbf{y},-\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t))}{G(\mathbf{y},\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t)) + G(\mathbf{y},-\boldsymbol{\mu}(t),\boldsymbol{\Sigma}_G(t))}.\end{aligned} \tag{47}$$

Substituting the Gaussian density and collecting terms gives the equivalent form

$$\phi(t) = 1 - \frac{1}{2}\exp\!\left(-\frac{1}{2}\boldsymbol{\mu}(t)^\top\boldsymbol{\Sigma}_G(t)^{-1}\boldsymbol{\mu}(t)\right)\int \frac{d\mathbf{y}}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_G(t))}}\,\frac{\exp\!\left(-\frac{1}{2}\mathbf{y}^\top\boldsymbol{\Sigma}_G(t)^{-1}\mathbf{y}\right)}{\cosh(\mathbf{y}^\top\boldsymbol{\Sigma}_G(t)^{-1}\boldsymbol{\mu}(t))}. \tag{48}$$

The change of variables $\mathbf{z} = \boldsymbol{\Sigma}_G(t)^{-1/2}\mathbf{y}$ whitens the Gaussian measure in (48). In the isotropic reversible case, all directions orthogonal to $\mathbf{m}$ factorize and the expression reduces to the standard one-dimensional integral. For the non-reversible drifts used here, $\boldsymbol{\Sigma}_G(t)$ is generally non-diagonal, so we measure $\phi(t)$ by cloned reverse simulations and compare it with the matrix-predicted t_S.

E. Random Energy Model Analysis for Collapse

E.1. General Linear Drift

Let

$$\mathbf{M}_t = e^{-\mathbf{A}t}, \qquad \boldsymbol{\Gamma}_t = \boldsymbol{\Sigma}_{\mathrm{sto}}(t). \tag{49}$$

The empirical forward density is the Gaussian mixture

$$P_t^e(\mathbf{x}) \propto \sum_{\mu=1}^{n} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{M}_t\mathbf{a}_\mu)^\top\boldsymbol{\Gamma}_t^{-1}(\mathbf{x}-\mathbf{M}_t\mathbf{a}_\mu)\right]. \tag{50}$$

We probe the density near the forward image of a reference point:

$$\mathbf{x} = \mathbf{M}_t\mathbf{a}_1 + \mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0,\boldsymbol{\Gamma}_t). \tag{51}$$

For a background point $\mu \neq 1$, define $\boldsymbol{\delta}_\mu = \mathbf{a}_1 - \mathbf{a}_\mu$. Its energy is

$$E_\mu(t) = \frac{1}{2}\left(\mathbf{M}_t\boldsymbol{\delta}_\mu + \mathbf{z}\right)^\top\boldsymbol{\Gamma}_t^{-1}\left(\mathbf{M}_t\boldsymbol{\delta}_\mu + \mathbf{z}\right). \tag{52}$$

The REM approximation treats the background differences as independent Gaussian vectors, $\boldsymbol{\delta}_\mu \sim \mathcal{N}(0, 2\sigma_0^2\mathbf{I})$. Introduce

$$\mathbf{K}_t = \mathbf{M}_t^\top\boldsymbol{\Gamma}_t^{-1}\mathbf{M}_t, \qquad \mathbf{h}_t = \mathbf{M}_t^\top\boldsymbol{\Gamma}_t^{-1}\mathbf{z}. \tag{53}$$

Then

$$E_\mu(t) = \frac{1}{2}\boldsymbol{\delta}_\mu^\top\mathbf{K}_t\boldsymbol{\delta}_\mu + \boldsymbol{\delta}_\mu^\top\mathbf{h}_t + \frac{1}{2}\mathbf{z}^\top\boldsymbol{\Gamma}_t^{-1}\mathbf{z}. \tag{54}$$

The Gaussian integral over $\boldsymbol{\delta}_\mu$ gives

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\delta}}\exp[-\beta E_\mu(t)] &= \det\!\left(\mathbf{I} + 2\beta\sigma_0^2\mathbf{K}_t\right)^{-1/2} \\ &\quad\times \exp\left[-\frac{\beta}{2}\mathbf{z}^\top\boldsymbol{\Gamma}_t^{-1}\mathbf{z} + \beta^2\sigma_0^2\mathbf{h}_t^\top\left(\mathbf{I} + 2\beta\sigma_0^2\mathbf{K}_t\right)^{-1}\mathbf{h}_t\right].\end{aligned} \tag{55}$$

Since $\mathbf{z} \sim \mathcal{N}(0, \mathbf{\Gamma}_t)$, $d^{-1}\mathbf{z}^\top \mathbf{\Gamma}_t^{-1}\mathbf{z} \to 1$ in high dimension. Moreover, $\mathbf{h}_t$ has covariance $\mathbf{K}_t$. In the eigenbasis of $\mathbf{K}_t$, concentration gives

$$\frac{1}{d}\mathbf{h}_t^\top F(\mathbf{K}_t)\mathbf{h}_t \to \frac{1}{d}\mathrm{Tr}[\mathbf{K}_t F(\mathbf{K}_t)] \tag{56}$$

for bounded spectral observables F. If $\kappa_j(t)$ are the eigenvalues of $\mathbf{K}_t$, the background free-energy density is

$$g_t(\beta) = -\frac{1}{2d}\sum_{j=1}^{d}\left[\log\left(1 + 2\beta\sigma_0^2\kappa_j(t)\right) + \frac{\beta}{1 + 2\beta\sigma_0^2\kappa_j(t)}\right]. \tag{57}$$

The correct data point contributes $-E_1(t)/d \approx -1/2$. The background contribution contains the pattern entropy $\alpha = (\log n)/d$ and is $\alpha + g_t(1)$. Equating the reference and background free-energy densities gives

$$2\alpha + 1 = \frac{1}{d}\sum_{j=1}^{d}\left[\log\left(1 + 2\sigma_0^2\kappa_j(t_C)\right) + \frac{1}{1 + 2\sigma_0^2\kappa_j(t_C)}\right], \tag{58}$$

which is the implicit REM criterion used in the main text.

The trace identity fixes the volume-contraction part of the calculation. For $\mathbf{A} = (\mathbf{I} + \mathbf{Q})\mathbf{U}$,

$$\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{U}) + \mathrm{Tr}(\mathbf{QU}) = \mathrm{Tr}(\mathbf{U}), \tag{59}$$

because $\mathbf{U}$ is symmetric and $\mathbf{Q}$ is anti-symmetric. Therefore $\det(e^{-\mathbf{A}t}) = \exp[-t\mathrm{Tr}(\mathbf{U})]$ is independent of $\mathbf{Q}$. However, the REM criterion depends on the full spectrum of $\mathbf{K}_t = \mathbf{M}_t^\top \mathbf{\Gamma}_t^{-1}\mathbf{M}_t$. The full spectrum of $\mathbf{K}_t$ therefore allows t_C to depend on $\mathbf{Q}$ despite the trace identity.

E.2. Diagonal Reversible Limit

Restrict to $\mathbf{Q} = 0$ and $\mathbf{A} = \mathbf{U} = \mathrm{diag}(u_1, \ldots, u_d)$. Then

$$\mathbf{M}_t = \mathrm{diag}(e^{-u_1 t}, \ldots, e^{-u_d t}), \qquad \mathbf{\Gamma}_t = \mathrm{diag}\left(\frac{1 - e^{-2u_1 t}}{u_1}, \ldots, \frac{1 - e^{-2u_d t}}{u_d}\right). \tag{60}$$

Thus

$$\kappa_j(t) = \frac{u_j e^{-2u_j t}}{1 - e^{-2u_j t}}. \tag{61}$$

Substitution into the general REM criterion gives

$$2\alpha + 1 = \frac{1}{d}\sum_{j=1}^{d}\left[\log\left(1 + 2\sigma_0^2\frac{u_j e^{-2u_j t_C}}{1 - e^{-2u_j t_C}}\right) + \frac{1}{1 + 2\sigma_0^2\frac{u_j e^{-2u_j t_C}}{1 - e^{-2u_j t_C}}}\right]. \tag{62}$$

The same expression can be checked component by component. Let $\lambda_j(t) = u_j/(1 - e^{-2u_j t})$ and write $z_j \sim \mathcal{N}(0, \lambda_j(t)^{-1})$. The one-dimensional contribution to the background moment is

$$\begin{aligned} M_j(\beta) &= \mathbb{E}_{\delta_j}\exp\left[-\frac{\beta}{2}\lambda_j(t)\left(e^{-u_j t}\delta_j + z_j\right)^2\right] \\ &= \left(1 + 2\sigma_0^2\beta\lambda_j(t)e^{-2u_j t}\right)^{-1/2}\exp\left[-\frac{\beta\lambda_j(t)z_j^2}{2\left(1 + 2\sigma_0^2\beta\lambda_j(t)e^{-2u_j t}\right)}\right], \end{aligned} \tag{63}$$

where $\delta_j \sim \mathcal{N}(0, 2\sigma_0^2)$. Replacing $\lambda_j(t)z_j^2$ by its typical value 1 in high dimension gives $g_t(\beta) = d^{-1}\sum_j \log M_j(\beta)$, which is the diagonal specialization of (57). This is the diagonal reversible implicit analytical criterion, with t_C obtained by root finding.

F. Experimental and Implementation Details

F.1. Gaussian-Mixture Setup

The Gaussian-mixture experiments use two equally weighted clusters with means $\pm\mathbf{m}$ and within-cluster covariance $\sigma^2\mathbf{I}$. In the paper figures, $d = 1024$, $\mathbf{m} = \mathbf{1}$, and the initial covariance distributes the symmetry-breaking signal across eigendirections. The diagonal potential profile is

$$\mathbf{U} = \mathrm{diag}(u_1, \ldots, u_d), \qquad u_j = 1 + \frac{8(j-1)}{d-1}. \tag{64}$$

This profile spans a controlled range of reversible relaxation rates.

The Simple perturbation family is

$$Q_{ij} = q \quad (i < j), \qquad Q_{ji} = -q, \qquad Q_{ii} = 0. \tag{65}$$

The Lelièvre curves use three auxiliary spectra in the construction algorithm. The labels $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ correspond respectively to an unshifted linear spectrum, the shifted spectrum used in the Remark 7 construction of Lelièvre et al., and a geometrically spaced spectrum. These families separate generic rotational perturbations from a geometry-aware perturbation designed to redistribute relaxation rates.

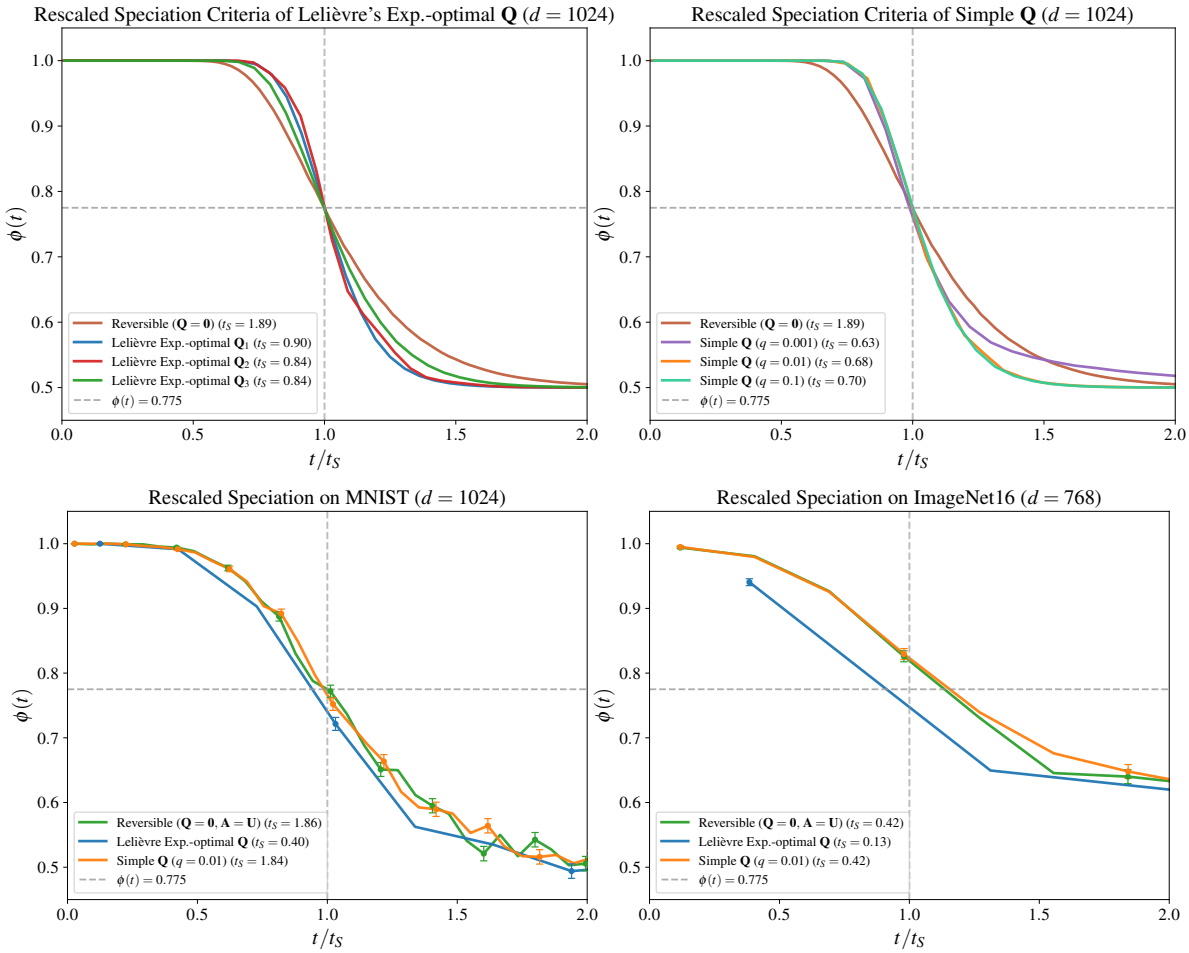

Figure 5. Rescaled speciation curves compare the matrix-computed t_S with trajectory cloning. Top: for Gaussian mixtures, rescaling by the t_S obtained from (7) aligns the curves at $(t/t_S, \phi) = (1, 0.775)$. Bottom: on MNIST and ImageNet16, the fixed-$\mathbf{U}$ reversible, Lelièvre, and Simple $\mathbf{Q}$ transitions likewise align near this point when rescaled by their independently matrix-computed t_S; Standard DDPM appears separately as the isotropic baseline with $\mathbf{U} = \mathbf{I}$. Finite data and learned-model variation produce small deviations among the real-data curves. Their near-alignment motivates $\phi(t) = 0.775$ as the empirical speciation threshold.

Speciation is evaluated by the trajectory-cloning probability $\phi(t)$. For each forward time t, a noisy state is sampled from the forward process, two reverse trajectories are initialized from that state, and $\phi(t)$ is the fraction of pairs that terminate in the same mixture component. The vertical dashed lines in the speciation plots mark the theoretical t_S values obtained by solving (7), including the propagated within-cluster covariance term in (38). The top row of Figure 5 displays the same Gaussian-mixture trajectories against t/t_S. The bottom row applies this rescaling to the saved MNIST and ImageNet16 cloning curves using the independently computed t_S values reported in their legends. The reversible anisotropic, Lelièvre, and Simple $\mathbf{Q}$ models share the same $\mathbf{U}$ and asymptotic noise covariance; Standard DDPM supplies the separate isotropic baseline.

The Gaussian-mixture collapse diagnostic uses the same $d = 1024$ and diagonal $\mathbf{U}$. The reversible reference is compared with Simple $\mathbf{Q}$ at $q \in \{0.1, 0.5\}$ and with the Lelièvre Exp.-optimal perturbation. The plotted entropy quantity is a finite-dimensional proxy for the entropic criterion; the reported crossing is the last zero-crossing with $t < 1.0$, matching the rule used in the main-text collapse figures.

F.2. Real-Data DDPM Details

The real-data experiments train DDPMs on binary image tasks with the same network architecture and training budget across drifts. The anisotropic drift matrices are constructed in the full PCA basis of the training set and then rotated back to pixel space before training. The model operates in the original image coordinates, with PCA defining the direction-dependent OU rates. All reported models are trained for 350,000 steps. The Simple $\mathbf{Q}$ speciation runs use $q = 0.01$; the MNIST collapse diagnostic uses $q = 0.1$, and the ImageNet16 collapse diagnostic uses $q = 0.01$.

For MNIST, the binary task is digits 0 versus 7 with $n = 12{,}188$ and $d = 1024$. For ImageNet16, the task uses $n = 2000$ and $d = 768$. Speciation is evaluated by the same cloning protocol as in the Gaussian mixture: two reverse trajectories are initialized from a common noisy state, and their terminal classes are assigned by a ResNet-18 classifier. The empirical crossing is read using the cloning-probability threshold shown in the plotted figures.

Collapse is evaluated by two diagnostics. The first is the empirical excess entropy density

$$f^e(t) = s^{\mathrm{sep}}(t) - s^e(t), \qquad s^{\mathrm{sep}}(t) = \alpha + s_G(t), \tag{66}$$

where $s^e(t)$ is the entropy density of the empirical forward distribution and $s_G(t)$ is the entropy density of the corresponding single-Gaussian reference. The second is the nearest-neighbour switching proxy $\hat{t}_C$, defined as the last reverse-time point at which the nearest training index changes along a trajectory. These diagnostics track the collapse transition in finite-dimensional trained models.

For MNIST, the entropy zero-crossings with $t < 1$ are approximately 0.332, 0.420, and 0.112 for the reversible, Simple $\mathbf{Q}$, and Lelièvre curves; the corresponding nearest-neighbour estimates are

$$\hat{t}_C = 0.237 \pm 0.241, \quad 0.189 \pm 0.214, \quad 0.0084 \pm 0.0106. \tag{67}$$

For ImageNet16, the entropy zero-crossings are approximately 0.323, 0.337, and 0.116; the nearest-neighbour estimates are

$$\hat{t}_C = 0.214 \pm 0.141, \quad 0.214 \pm 0.186, \quad 0.011 \pm 0.013. \tag{68}$$

These values provide finite-dimensional diagnostics linked to the entropic definition of t_C.

Response to the Reviewers

Manuscript Title: *Steering Dynamical Regimes of Diffusion Models by Breaking Detailed Balance*

Venue: ICML 2026 Workshop on Foundations of Deep Generative Models

Manuscript Identifier: 110

Authors: Haiqi Lu and Ying Tang

General Response

We thank the Area Chair and all reviewers for their careful reading and constructive suggestions. Their comments identified places where the original submission was difficult to follow and brought a mathematical inconsistency in the speciation criterion to our attention. We rewrote the presentation around the diffusion-model question, defined the notation before use, and made the connection between the theory and experiments more direct.

The revision also separates the entropic definition of collapse from its REM approximation and empirical diagnostics, corrects the propagated within-component covariance in the speciation criterion, and gives a fuller account of the image experiments. We added normalized cloning curves, a small MNIST generation-quality check, and clearer limits on what can be concluded from the binary tasks and the tested Simple **Q** family. We also checked the reference metadata flagged in the reviews.

The revision corrects the speciation criterion, makes the notation and transition criteria readable from the main text, and gives a clearer account of the experimental evidence. The central result remains that non-reversible forward dynamics can shift finite-time reverse regimes, with the direction and magnitude of the shift depending on perturbation geometry. The appendices provide the full derivations and implementation details.

Below, reviewer comments are shown in *italics*. Short excerpts from the revised manuscript, or concise summaries of those revisions, are shown in red boxes.

Summary of Changes

1. **Accessibility and terminology.**
   - Rewrote the abstract and introduction around the diffusion-model question of finite-time regime displacement.
   - Defined speciation, collapse, the asymptotic noise covariance, and probability currents when they first appear.
   - Replaced or explained physics-specific terminology, including the Random Energy Model approximation and trajectory-cloning probability.
2. **Mathematical notation and criteria.**
   - Defined symbols before use, including d, $\mathbf{I}$, $\mathbf{A}$, $\mathbf{U}$, $\mathbf{Q}$, $\mathbf{W}_t$, t_f, τ, and the forward marginal P_t.
   - Corrected the speciation criterion to include propagated within-component covariance.
   - Separated the collapse definition, the REM approximation, and the empirical collapse diagnostics.
3. **Experimental protocol and validation.**

- Described the matched drift-specific DDPM protocol, including the shared architecture, objective, and training budget.
- Added normalized MNIST and ImageNet16 cloning curves for comparison with the independently matrix-computed speciation times.
- Added an MNIST generation-quality check using a common evaluator, 1000 samples per method, class balance, confidence, feature-space nearest-neighbour distances, and fixed-index generated samples.

4. **Experimental scope.**
   - Restricted conclusions about collapse and Simple **Q** to the tested perturbations and binary-image experiments.
   - Separated the regime-timing results from the generation-quality check and noted the lack of multiclass evaluation.
5. **References and presentation.**
   - Checked reference metadata, corrected the entries highlighted by the reviewer, and replaced vague related-work wording with direct comparisons.
   - Expanded the main-text explanation of the two criteria and linked them to their appendix derivations.

Response to the Area Chair

Reviewer comment. *The mathematical symbols and terminology should be clarified to make the manuscript accessible to the broader diffusion-model community and FoGen attendees.*

Response. We thank the Area Chair for the careful assessment and constructive guidance on making the paper more accessible. We rewrote the opening sections for a diffusion-model audience. Speciation is introduced as the time when a reverse trajectory commits to a data mode; collapse is introduced as the later regime in which individual training samples become relevant attractors. The forward-process section now defines each matrix and time variable before use and explains the anti-symmetric perturbation in terms of probability currents and the preserved asymptotic noise covariance.

> **Revised manuscript:** Let the data distribution be supported on samples $\mathbf{a}_\mu \in \mathbb{R}^d$, where d is the data dimension. Let $\mathbf{I}$ be the $d \times d$ identity and $\mathbf{W}_t$ a standard d-dimensional Brownian motion. We use the drift $\mathbf{A} = (\mathbf{I} + \mathbf{Q})\mathbf{U}$, where $\mathbf{U} = \mathbf{U}^\top > 0$ is symmetric positive definite and $\mathbf{Q} = -\mathbf{Q}^\top$ is anti-symmetric.

Response. The main text now explains what each transition criterion measures. For speciation, it compares propagated class separation with local component blur. For collapse, it first gives the entropic definition, then introduces the REM approximation and the empirical diagnostics.

> **Revised manuscript:** The entropic equation defines collapse through entropy balance. The REM approximation yields a prediction in a tractable high-dimensional model, and the experiments track the transition using empirical excess entropy density and a nearest-neighbour switching proxy.

Response to Reviewer TDuF

Reviewer comment. *The writing is a significant barrier to evaluation. Symbols are used before definition, and physics terminology such as "Random Energy Model," "Mahalanobis signal-to-noise spectrum," "trajectory-cloning probability," and "Regime I/II" is insufficiently explained.*

Response. We are grateful to the reviewer for the close reading and detailed suggestions on the paper's presentation. We rewrote these passages and introduced symbols before their first equations. The revision removes the regime-number terminology and replaces "Mahalanobis signal-to-noise spectrum" with "covariance-normalized separation between training samples." It also defines "terminal component" as the final mixture component or class label and explains the finite-time behavior of non-normal drifts directly.

> **Revised manuscript:** The matrix $\mathbf{K}_t$ converts a difference between two training samples into a squared distance measured in units of the forward noise covariance $\mathbf{\Gamma}_t$. Thus its eigenvalues $\kappa_j(t)$ are direction-wise signal-to-noise ratios for separating training examples.

Response. The trajectory-cloning probability is now defined through the simulation procedure used to measure it.

> **Revised manuscript:** To estimate speciation experimentally, two reverse trajectories are continued from the same noisy state at time t, and $\phi(t)$ is the probability that their final class labels agree. Values near 1/2 indicate no class commitment, while values near 1 indicate that the shared noisy state has already selected a class.

Reviewer comment. *The speciation and collapse criteria are poorly explained in the main body, while a fuller explanation appears only in the appendix.*

Response. The main text now defines every object in the speciation criterion before displaying it and explains the crossing as a competition between local blur and propagated between-mode signal. The corrected expression includes the within-component covariance. The collapse section separately presents the entropy balance, the REM approximation, and the empirical proxies. The appendices retain the full derivations.

> **Revised manuscript:** Let $\mathbf{\Sigma}_{\mathrm{B}}$ be the between-mode covariance, which records how component means differ, and let $\mathbf{\Sigma}_{\mathrm{W}}$ be the within-mode covariance. We write $\mathbf{\Sigma}_{\mathrm{loc}}(t)$ for the covariance of one propagated component after adding forward noise. The speciation time satisfies $\lambda_{\min}(\mathbf{\Sigma}_{\mathrm{loc}}(t_S) - e^{-\mathbf{A}t_S}\mathbf{\Sigma}_{\mathrm{B}}e^{-\mathbf{A}^\top t_S}) = 0$, where $\mathbf{\Sigma}_{\mathrm{loc}}(t) = \mathbf{\Sigma}_{\mathrm{sto}}(t) + e^{-\mathbf{A}t}\mathbf{\Sigma}_{\mathrm{W}}e^{-\mathbf{A}^\top t}$.

Reviewer comment. *The paper should connect speciation and collapse to generalization and memorization, and it would be useful to measure image quality between drifts.*

Response. The introduction now connects speciation with semantic mode commitment and collapse with the onset of sample-specific attraction. We also added an MNIST generation-quality check using 1000 samples per method, one common ResNet-18 evaluator, predicted class balance, confidence, generated-to-test and generated-to-train nearest-neighbour distances, and fixed-index samples.

On this binary MNIST task, the four methods have similar class balance, confidence, generated-to-test distance, and visual recognizability. Generated-to-train distance is reported separately because

it measures training-set proximity. Broader perceptual-quality and memorization questions require additional evaluations.

Revised manuscript: Across class balance, confidence, and feature distance, the modified dynamics remain close to the Standard DDPM baseline on this binary task. Broader evaluations are required to resolve perceptual-quality differences and memorization behavior.

Reviewer comment. *Experiments consider small image datasets limited to a binary-class setting.*

Response. The binary tasks were chosen because they allow the mode-commitment transition to be measured directly. The Discussion now limits the experimental conclusions to the tested MNIST and ImageNet16 settings and identifies multiclass evaluation as the next step.

Response to Reviewer T8rS

Reviewer comment. *Multiclass testing is preferred.*

Response. We appreciate the reviewer's careful reading and practical suggestions for strengthening the empirical evaluation. We did not add a multiclass experiment to the camera-ready version. The Discussion now states that the image experiments cover binary tasks. We instead added normalized MNIST and ImageNet16 cloning curves, which compare the observed transitions with independently matrix-computed t_S values.

Reviewer comment. *Can the authors measure generated-sample quality and connect regime shifts to generative metrics?*

Response. We added the MNIST generation-quality check described above. The tested methods have similar class balance, confidence, generated-to-test distance, and fixed-index digit recognizability. The Discussion now identifies precision/recall-style diversity metrics, FID, and multiclass evaluation as the appropriate tests for broader generation effects.

Reviewer comment. *The claim that breaking detailed balance alone does not imply an earlier collapse transition is tested using a highly structured Simple* $\mathbf{Q}$*; more general experiments are desirable.*

Response. The tested Simple $\mathbf{Q}$ is a weak dense comparison family whose orientation is independent of the data geometry. We now describe it that way and limit the empirical conclusion to the tested perturbations and binary-image tasks. The broader result is that suitable non-reversible perturbations can shift finite-time regimes, while the collapse response depends on how the perturbation changes covariance-normalized sample separation.

Revised manuscript: Together with the MNIST results, the ImageNet16 results agree with the REM interpretation: substantial changes in the covariance-normalized separation spectrum move the collapse diagnostics, with direction and magnitude varying across the binary-image experiments.

Response to Reviewer yajQ

Reviewer comment. *The Gaussian mixture should reflect the within-cluster covariance, and the main speciation criterion omits the propagated $\sigma^2 e^{-\mathbf{A}t} e^{-\mathbf{A}^\top t}$ term mentioned in the appendix.*

Response. We thank the reviewer for the meticulous reading and for identifying the missing covariance term. The reviewer is correct: the main criterion omitted the propagated within-component covariance. We corrected the main text and appendix by separating within-component covariance from between-component signal. For a symmetric two-component mixture, $\mathbf{\Sigma}_{\mathrm{W}} = \sigma^2 \mathbf{I}$ and $\mathbf{\Sigma}_{\mathrm{B}} = \mathbf{m}\mathbf{m}^\top$. The propagated local covariance is

$$\mathbf{\Sigma}_{\mathrm{loc}}(t) = \mathbf{\Sigma}_{\mathrm{sto}}(t) + e^{-\mathbf{A}t} \mathbf{\Sigma}_{\mathrm{W}} e^{-\mathbf{A}^\top t},$$

which includes the highlighted $\sigma^2 e^{-\mathbf{A}t} e^{-\mathbf{A}^\top t}$ contribution. The scalar reversible limit has also been updated to include the within-mode variance w_Λ.

Reviewer comment. *The references contain inaccurate metadata and DOI entries, and the wording "A closely related work is Albrychiewicz et al." should be revised.*

Response. We checked the highlighted entries against publisher and venue records and revised the related-work comparison. The Biroli et al. entry now contains the complete author list and correct Nature Communications metadata. The highlighted DOI entries were corrected or verified as `10.1088/1742-5468/ade136`, `10.1038/s41467-024-54281-3`, `10.1007/s10955-013-0769-x`, and `10.1088/1742-5468/acf8ba`. We also replaced the vague "closely related" wording with a direct technical comparison between coupled-OU synchronization gaps and our anti-symmetric forward-generator intervention.

Reviewer comment. *The phrase "This overlap can be read directly" should clarify what overlap means.*

Response. The revised discussion names the Simple **Q** and reversible trajectory-cloning curves directly and explains their near-overlap through the concentration of class-separation signal in a low-dimensional slow sector. We also replaced other vague references to "overlap" with the observable and methods being compared.

Revised manuscript: The near-overlap between the Simple **Q** and reversible cloning curves follows from the signal term in the speciation criterion.