Dynamical Regimes of Multimodal Diffusion Models

Emil Albrychiewicz*
Leinweber Institute for Theoretical Physics & Department of Physics
University of California, Berkeley, CA 94720-7300, USA
Theoretical Physics Group, Lawrence Berkeley National Laboratory
Berkeley, CA 94720-8162, USA
ealbrych@berkeley.edu

Andrés Franco Valiente*
Leinweber Institute for Theoretical Physics & Department of Physics
University of California, Berkeley, CA 94720-7300, USA
Theoretical Physics Group, Lawrence Berkeley National Laboratory
Berkeley, CA 94720-8162, USA
Department of Radiation Oncology
University of California, San Francisco, USA
andresfranco@berkeley.edu

Li-Ching Chen
Computational Precision Health
University of California, San Francisco
Mission Hall 550 16th St, Floor 5 San Francisco, CA 94158
liching_chen@berkeley.edu

Abstract

Diffusion based generative models have achieved unprecedented fidelity in synthesizing high dimensional data, yet the theoretical mechanisms governing multimodal generation remain poorly understood. Here, we present a theoretical framework for coupled diffusion models, using coupled Ornstein-Uhlenbeck processes as a tractable model. By using the nonequilibrium statistical physics of dynamical phase transitions, we demonstrate that multimodal generation is governed by a spectral hierarchy of interaction timescales rather than simultaneous resolution. A key prediction is the "synchronization gap", a temporal window during the reverse generative process where distinct eigenmodes stabilize at different rates, providing a theoretical explanation for common desynchronization artifacts. We derive analytical conditions for speciation and collapse times under both symmetric and anisotropic coupling regimes, establishing strict bounds for coupling strength to avoid unstable symmetry breaking. We show that the coupling strength acts as a spectral filter that enforces a tunable temporal hierarchy on generation. We support these predictions through controlled experiments with diffusion models trained on MNIST datasets and exact score samplers. These results motivate time dependent coupling schedules that target mode specific timescales, offering a potential alternative to ad hoc guidance tuning.

*Equal contribution.

Conference on Physics and AI at Stanford University (PAI 2026).

1 Introduction

Diffusion based generative models have transformed unsupervised learning, enabling synthesis of high dimensional data across image, audio, and video modalities Sohl-Dickstein et al. [2015], Song and Ermon [2019], Song et al. [2021b], Ho et al. [2020], Saharia et al. [2022], Peebles and Xie [2023]. While empirically successful, the theoretical mechanisms governing the reverse time generative process remain actively studied Ambrogioni [2023], Kadkhodaie et al. [2024]. Biroli et al. [2024] provide a statistical mechanics framework delineating three dynamical regimes for diffusion models: (I) a noise regime, (II) a speciation regime where trajectories converge to the data manifold, and (III) a collapse regime characterized by memorization of training data points.

Extending this framework to multimodal generation introduces nontrivial complexities. Multimodality imposes two constraints: intra-modal coherence and cross modal alignment Radford et al. [2021], Rombach et al. [2022]. We address these within a coupled diffusion framework that represents statistical dependencies between modalities as an interaction term constraining the joint probability manifold.

Contributions. We model the coupled forward process via coupled Ornstein–Uhlenbeck (OU) processes Uhlenbeck and Ornstein [1930] aligned with the variance preserving formulation of Song et al. [2021b]. By diagonalizing the symmetric coupling into common and difference eigenmodes, we show that speciation is governed by a sum of per mode signal-to-noise ratios (SNRs), inducing a synchronization gap, a temporal window where one eigenmode has transitioned into regime II while the other remains in regime I. We derive the collapse time via a Random Energy Model (REM) analysis Derrida [1981], Biroli et al. [2024], obtain a strict stability bound on coupling strength, and extend the analysis to anisotropic coupling relevant for conditional generation. Experiments on MNIST LeCun et al. [1998] with both deterministic (DDIM) and stochastic (DDPM) samplers confirm the predicted synchronization gap and its coupling dependence.

2 Theoretical Framework

Coupled OU process. We model a coupled diffusion architecture via a system of two d dimensional OU processes $Z(t) = (X(t), Y(t)) \in \mathbb{R}^{2d}$ satisfying the Itô SDE

$$dZ(t) = MZ(t)\,dt + \Sigma_W\,dW(t), \tag{1}$$

driven by a $\mathbb{R}^{2d}$ valued Wiener process $W(t)$. For the symmetric case, we choose

$$M = \begin{pmatrix} -\beta & g \\ g & -\beta \end{pmatrix} \otimes I_d, \quad \Sigma_W = \sigma_W I_{2d}, \tag{2}$$

with stability condition $\beta > |g|$. The solution is a Gaussian process with mean $\mu(t) = e^{Mt}\mu(0)$ and transition covariance $Q(t) = \int_0^t e^{M\tau}\Sigma_W\Sigma_W^\top e^{M^\top\tau}\,d\tau$. The total diffusion kernel is $C(t) = S(t) + Q(t)$ where $S(t) = e^{Mt}\Sigma(0)e^{M^\top t}$ is the drifted initial covariance.

For the initial population distribution we take a two component Gaussian mixture with means $\pm\mu(0) \in \mathbb{R}^{2d}$ and diagonal covariance $\Sigma(0) = \sigma^2 I_{2d}$. By Anderson's time reversal theorem Anderson [1982], the reverse SDE involves the score $\nabla \log p_t$.

Spectral decomposition and speciation. The symmetric M diagonalizes with eigenvectors $v_\pm = \frac{1}{\sqrt{2}}(1, \pm 1)^\top$ and eigenvalues $\lambda_\pm = -\beta \pm g$. The common mode v_+ captures shared content; the difference mode v_- captures modality discrepancies. Since $|\lambda_+| < |\lambda_-|$ for $g > 0$, the common mode retains signal longer under forward noising.

The reverse drift for the population distribution admits a scalar self consistency equation for fixed points $u = \kappa(t)\tanh(u)$, where additional fixed points appear via pitchfork bifurcation at $\kappa(t) = 1$. We define the speciation time t_S by $\kappa(t_S) = 1$. In the eigenbasis, κ decomposes as

$$\kappa(t) = \sigma_W^2\big(\mathrm{SNR}_+(t) + \mathrm{SNR}_-(t)\big), \tag{3}$$

where $\mathrm{SNR}_\pm(t) = e^{2\lambda_\pm t}m_\pm^2 / \big[c_\pm(t)(\lambda_\pm c_\pm(t) + \sigma_W^2)\big]$, with $m_\pm^2 = \frac{1}{d}\|\mu_\pm(0)\|^2$ the per dimension projected norms and $c_\pm(t) = \sigma^2 e^{2\lambda_\pm t} + \sigma_W^2(e^{2\lambda_\pm t} - 1)/(2\lambda_\pm)$ the per mode diffusion kernel

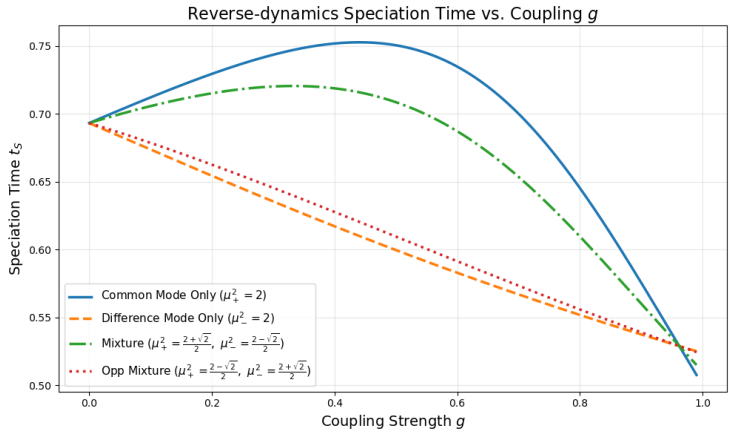

Figure 1: We show the impact of the coupling strength g on speciation time t_S for different modes using analytic and numerical solutions. Common and difference modes exhibit different sensitivities to g, creating a synchronization gap. Parameters: $\beta = 1$, $\sigma_W^2 = 2$, $\sigma^2 = 1$.

eigenvalues. When a single mode dominates, the speciation time admits the closed form

$$t_{S,\pm} = \frac{1}{2\lambda_\pm} \log\left(\frac{\sigma_W^2\left(m_\pm^2 - \sqrt{m_\pm^4 + B_\pm^2}\right)}{2\lambda_\pm B_\pm^2}\right), \quad B_\pm = \sigma^2 + \frac{\sigma_W^2}{2\lambda_\pm}. \tag{4}$$

Synchronization gap. Due to the exponential separation of eigenvalues, the common and difference modes reach $\kappa = 1$ at different times (Figure 1). This creates a synchronization gap a temporal window where the common mode has entered regime II while the difference mode remains in regime I. The coupling strength g acts as a spectral filter that controls the width of this gap, offering a principled explanation for desynchronization artifacts in multimodal generation.

Stability bound. For the reverse potential to remain confining, we require $\lambda_\pm + \sigma_W^2/c_\pm(t) > 0$ for all t. A sufficient condition is

$$\sigma^2 < \frac{\sigma_W^2}{\beta + |g|}, \tag{5}$$

beyond which unstable symmetry breaking causes sampling trajectories to diverge Yu et al. [2025].

Collapse time. The transition to memorization (regime III) is derived via a REM analysis Derrida [1981], Biroli et al. [2024]. Writing the partition function $Z_t(\tilde{z}) = \sum_i \exp(-2d\,\varepsilon_t(\tilde{z}, z_i))$ and using large deviation theory in the thermodynamic limit $d \to \infty$ with $\alpha = \frac{1}{2d}\log n$, the collapse time t_C satisfies

$$n^{2/d} = \prod_{s\in\{+,-\}} (1 + \chi_s), \quad \text{where} \quad \chi_s = \frac{\sigma^2}{\sigma_W^2}\frac{2\lambda_s}{1 - e^{-2\lambda_s t_C}}. \tag{6}$$

The per mode collapse times $t_{C,\pm}$ admit closed-form solutions. Unlike speciation, t_C is nearly constant as g increases, since χ_+ and χ_- compensate. This shows g tunes the generative hierarchy without risking joint system collapse. The curse of dimensionality is manifest, if $n^{2/d} = 1$, no finite collapse time exists.

Anisotropic coupling. For conditional generation, we consider a lower triangular M with one directional information flow from X to Y. The speciation condition $\kappa(t_S) = 1$ now depends on the alignment angle θ between conditioning and target means. Aligned directions $\theta \approx 0$ are sensitive to coupling at moderate g, entering a nonspeciation regime for strong coupling, while misaligned directions $\theta \approx \pi$ are more robust. This motivates treating coupling strength as a time dependent control analogous to classifier free guidance scales Ho and Salimans [2022].

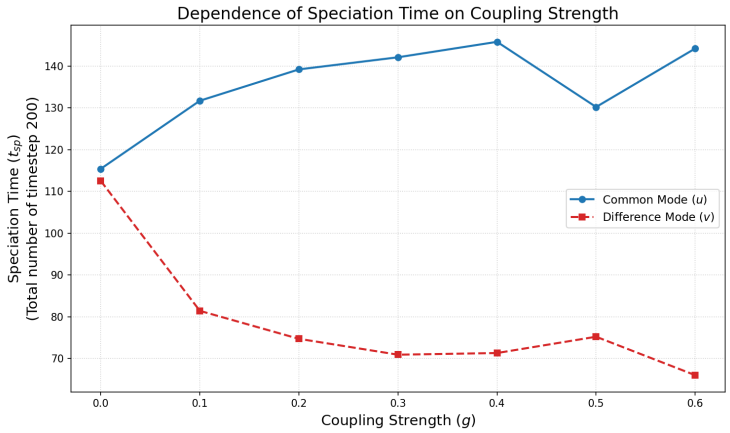

Figure 2: We demonstrate the dependence of the specification time on the coupling strength in the experiment with MNIST. We treat the speciation threshold crossing times for the common mode u and the difference mode v as a function of the coupling g.

3 Experiments

We validate the synchronization gap prediction using a standard VP diffusion model on MNIST. From each image x, we form a two-channel state $z_0 = (x_A, x_B)^\top$ via a deterministic pixel transform T (edge map, inversion, dilation, or blur), with random channel swaps enforcing A/B exchange symmetry (**MAIN** pairing). A **CTRL** pairing uses cross class samples to break content alignment while preserving marginals. In the (u, v) eigenbasis, coupling is implemented via anisotropic noise $\varepsilon_u \sim \mathcal{N}(0, (1-g)I)$, $\varepsilon_v \sim \mathcal{N}(0, (1+g)I)$, giving $\mathrm{SNR}_u/\mathrm{SNR}_v \propto (1+g)/(1-g)$.

Protocol I. Using DDIM sampling Song et al. [2021a], we track cosine similarity of each mode to its final state $\mathrm{Cos}_m(t) = \mathbb{E}[\cos(m(t), m_\star)]$. The synchronization gap $\Delta t(\tau) = t_v(\tau) - t_u(\tau)$ at threshold τ quantifies the temporal ordering. Results show $|\Delta t|$ is 2–3$\times$ larger in MAIN than CTRL across all (g, τ), confirming the effect is driven by cross channel alignment, not marginal dynamics. Intervention experiments injecting noise only into v at mid step of denoising produce persistent effects in MAIN but not CTRL, consistent with selective vulnerability during the gap.

Protocol II. Using DDPM on two class MAIN data, we branch independent clones from cached intermediate states and measure excess agreement $\phi_m^{\mathrm{ex}}(t)$, a baseline corrected probability that clones agree on class label. Figure 2 shows that increasing g introduces the separation between u- and v-mode speciation curves, directly confirming the coupling driven synchronization gap predicted by the theory.

4 Discussion

We have established a theoretical framework connecting multimodal diffusion generation to nonequilibrium phase transitions via coupled OU processes. The central prediction, the synchronization gap arising from the spectral hierarchy of interaction timescales, is confirmed experimentally. The coupling strength g serves as a robust control parameter, it tunes the temporal hierarchy of semantic emergence without significantly affecting collapse time, suggesting principled design of coupling schedules that respect intrinsic timescales rather than ad hoc guidance tuning.

Several extensions are natural. Time dependent couplings $M(t)$ would yield nonautonomous systems possible with perturbative analysis via Dyson-type expansions, where the signature transform Chen [1977], Chevyrev and Kormilitzin [2016] encodes the chronological ordering of information flow. Nonlinear mean reverting SDEs (e.g., cubic drift) would connect to nonequilibrium field theory techniques Zinn-Justin [2021], potentially modeling nonlinear coupling architectures such as cross-attention Vaswani et al. [2017]. Scaling to heterogeneous modality triplets should widen the spectral gap between eigenmodes, making the synchronization gap both more pronounced and more critical to control.

References

Luca Ambrogioni. Statistical mechanics of generative diffusion models. *arXiv:2304.05785*, 2023.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12: 313–326, 1982.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *arXiv:2402.18491*, 2024.

Kuo-Tsai Chen. Iterated path integrals. *Bull. AMS*, 1977.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv:1603.03788*, 2016.

Bernard Derrida. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B*, 24:2613, 1981.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *NeurIPS Workshop*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020.

Zahra Kadkhodaie et al. Generalization in diffusion models arises from geometry-adaptive harmonic representations. *ICLR*, 2024.

Yann LeCun et al. Gradient-based learning applied to document recognition. *Proc. IEEE*, 1998.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

Alec Radford et al. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Robin Rombach et al. High-resolution image synthesis with latent diffusion models. *CVPR*, 2022.

Chitwan Saharia et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 2022.

Jascha Sohl-Dickstein et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ICLR*, 2021a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 2019.

Yang Song et al. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021b.

George E Uhlenbeck and Leonard S Ornstein. On the theory of the Brownian motion. *Physical Review*, 36:823, 1930.

Ashish Vaswani et al. Attention is all you need. *NeurIPS*, 2017.

Zhendong Yu et al. Nonequilibrium phase transitions in diffusion models. *arXiv*, 2025.

Jean Zinn-Justin. Quantum field theory and critical phenomena. *Oxford University Press*, 2021.