

## 898 Contents

899	<b>1 Introduction</b>	<b>1</b>
900	<b>2 Background</b>	<b>2</b>
901	2.1 Diffusion models in continuous and discrete time . . . . .	2
902	2.2 Basics of fluctuation theory . . . . .	3
903	<b>3 Methodology</b>	<b>3</b>
904	<b>4 Demonstrations of our methodology</b>	<b>4</b>
905	4.1 Warm-up: Predicting convergence of the data distribution . . . . .	5
906	4.2 Class-conditional sampling . . . . .	6
907	4.3 Rare-class generation . . . . .	7
908	4.4 Zero-shot classification . . . . .	7
909	4.5 Zero-shot style transfer . . . . .	8
910	<b>5 Conclusion</b>	<b>9</b>
911	<b>A Notation used throughout the paper</b>	<b>25</b>
912	<b>B Theoretical contributions</b>	<b>26</b>
913	B.1 A concise primer on fluctuation theory . . . . .	26
914	B.2 On the validity of fluctuation theory . . . . .	29
915	B.3 Pull-back of data events along the PF-ODE . . . . .	29
916	B.4 From the empirical reverse process to the learned sampler . . . . .	30
917	B.5 Relation between cross-fluctuation and within-event fluctuation based thresholds .	31
918	B.6 Coupling and mixing for discrete Markov chains . . . . .	33
919	B.7 Fluctuation moments bound total variation distance . . . . .	33
920	B.8 Stochastic-flow formulation for the SDE view . . . . .	34
921	B.9 Extending the framework to certain non-Markovian samplers . . . . .	35
922	B.10 Mixing time of isotropic Gaussians under Brownian diffusion . . . . .	36
923	B.11 Higher-order fluctuations as a proof of concept . . . . .	37
924	B.12 Centred kernel alignment (CKA) . . . . .	37
925	B.13 Structural regularity bounds fluctuations . . . . .	38
926	B.14 Phases of diffusion-model dynamics . . . . .	38
927	B.15 A single forward Monte-Carlo sweep yields unbiased estimates of all terms in $\widetilde{\mathcal{M}}_p^{(n)}$	40
928	<b>C Experimental details and further results</b>	<b>42</b>
929	C.1 Compute and Reproducibility . . . . .	42
930	C.2 Convergence of data . . . . .	42
931	C.3 Class-conditional generation . . . . .	43

932	C.4 Rare Class Generation . . . . .	49
933	C.5 Zero-shot classification . . . . .	51
934	C.6 Binary classification via linear probes . . . . .	51
935	C.7 Zero Shot Style Transfer . . . . .	53
936	<b>D Philosophical background and related work</b>	<b>57</b>
937	<b>E Limitations and future work</b>	<b>58</b>
938	<b>F Related work</b>	<b>58</b>

Table 6: Global symbols

Symbol	Domain / type	Meaning
$\mathbf{x}_0 \sim p_0$	$\mathbb{R}^d$	Data sample from initial distribution $p_0$
$p_{\text{desired}}$	density on $\mathbb{R}^d$	Desired distribution
$\mathbf{x}_t$	$\mathbb{R}^d$	Forward-diffused variable at time $t$
$\beta(t), \beta_t$	$\mathbb{R}_{>0} / (0, 1)$	Continuous / discrete noise schedule
$J(t)$	$(0, 1]$	Signal attenuation factor (2.4)
$p_t$	density on $\mathbb{R}^d$	Marginal distribution of $\mathbf{x}_t$
$s_\theta(\mathbf{x}, t)$	$\mathbb{R}^d \rightarrow \mathbb{R}^d$	Learned score network
$\rho(\cdot)$	$\text{map } \Omega \rightarrow \mathbb{R}^m$	State operator (usually $\rho(\mathbf{x}) = \mathbf{x}$ ) (Section 3)
$F_\rho^{(n)}(\Omega)$	$\mathbb{R}_{\geq 0}$	$n^{\text{th}}$ centred fluctuation moment on event $\Omega$ (2.5)
$\widehat{F}_\rho^{(n)}(\Omega_i), \widehat{F}_{\rho_i}^{(n)}(\Omega_i)$	$\mathbb{R}_{\geq 0}$	$2n^{\text{th}}$ within-event fluctuation moment on $\Omega_i$ (2.7)
$G_\rho^{(n)}(\Omega_1, \Omega_2), G_{\rho_1, \rho_2}^{(n)}(\Omega_1, \Omega_2)$	$\mathbb{R}_{\geq 0}$	Unnormalised cross-fluctuation (2.6)
$\mathcal{M}_\rho^{(n)}(\Omega_1, \Omega_2), \mathcal{M}_{\rho_1, \rho_2}^{(n)}(\Omega_1, \Omega_2)$	$[0, 1]$	Normalised cross-fluctuation (2.8)
$\Omega_{k,0}$	event in $\Omega$	Class- $k$ source region ( $k = 1, \dots, K$ ) (Section 4.2)
$\Omega_{k,t}$	event	Image of $\Omega_{k,0}$ after $t$ forward steps
$\Sigma_{k,t}$	$\mathbb{S}_+^d$	Covariance of $\Omega_{k,t}$
$\lambda_k^{\max}(t)$	$\mathbb{R}_{\geq 0}$	Maximum eigenvalue of $\Sigma_{k,t}$
$\mathcal{M}_\rho(t)$	$[0, 1]$	Shorthand for $\mathcal{M}_\rho^{(2)}(\Omega_{1,t}, \Omega_{2,t})$
$i^*$	$\{0, \dots, T\}$	First index where $\mathcal{M}_\rho^{(n)}(t) = 1$ (merger) (Section 3)
$w(t)$	probability mass	Importance weight over timesteps (Section 4.4)
$\text{SNR}(t)$	$\mathbb{R}_{\geq 0}$	$\alpha_t^2 / (1 - \alpha_t^2)$ (signal-to-noise)
$\text{Tr}(\cdot)$	$\mathbb{R}$	Matrix trace operator

Table 7: Time- and index-specific symbols

Symbol	Type / range	Meaning
$t$	$\mathbb{R}_{\geq 0}$	Continuous diffusion time (PF-ODE or SDE)
$s, u$	$\mathbb{R}_{\geq 0}$	Generic continuous times used in flow composition ( $0 \leq s \leq t \leq u \leq T$ )
$i$	$\{0, \dots, n\}$	Discrete forward-process index ( $i = 0$ data; $i = n$ white noise)
$T$	positive real / integer	<b>Continuous</b> final time (SDE/ODE) <b>or discrete</b> horizon with schedule $\{\beta_t\}_{t=1}^T$
$n$	$\mathbb{N}$	Chosen number of forward (or reverse) <b>discrete</b> steps in an experiment (may be $n=T$ )
$\Delta t$	$\mathbb{R}_{>0}$	Integration step size in continuous-time numerical solvers
$\beta_t$	sequence on $\{1, \dots, T\}$	Discrete noise-schedule value at step $t$
$\beta(t)$	function $[0, T] \rightarrow \mathbb{R}_{>0}$	Continuous noise-schedule function
$i^*$	integer	Earliest discrete index where two events first merge (convergence index)
$t_{\text{conv}}$	$\mathbb{R}_{\geq 0}$	Same as $i^*$ but expressed on the continuous time axis
$t_{\text{merge},k}$	$\mathbb{R}_{\geq 0}$	First time class $k$ merges with any other class
$t_{\text{start},k}$	$\mathbb{R}_{\geq 0}$ or int	Lower bound of guidance / weighting window for class $k$
$t_{\text{stop},k}$	same type as above	Upper bound of the window for class $k$
$t_{u \rightarrow s}, t_{s \rightarrow c}$	$\mathbb{R}_{\geq 0}$	Thermodynamic phase-transition times (unbiased $\rightarrow$ speciation, speciation $\rightarrow$ condensation)
$t_{\text{mix}}(\varepsilon)$	$\mathbb{R}_{\geq 0}$ (Appendix B.14)	$\varepsilon$ -mixing time of the VP-SDE (Appendix B.6)
$t_{\text{cpl}}(\varepsilon)$	integer	$\varepsilon$ -coupling time for discrete Markov chains (Appendix B.6)
$t_{k\ell}^{\text{lat}}(\varepsilon)$	$\mathbb{R}_{\geq 0}$	First lattice-merger time between classes $k$ and $\ell$ with tolerance $\varepsilon$ (Appendix B.14)
$\eta_t$	$[0, 1]$	Interpolation schedule value used in Algorithm 3 (rare-class generation)

## B Theoretical contributions

### B.1 A concise primer on fluctuation theory

This subsection provides a concise overview of the fluctuation-theoretic framework that underpins the notation introduced in [Section 2.2](#).

**State operators.** In [Section 2.2](#), we model our dataset by a *state operator*

$$\rho: \Omega \longrightarrow (\mathcal{X}, \Sigma), \quad \omega \mapsto x(\omega),$$

where  $(\mathcal{X}, \Sigma)$  is the measurable space appropriate to the data—e.g.  $\mathcal{X} = \mathbb{R}^d$  with its Borel  $\sigma$ -algebra for vector-valued samples,  $\mathcal{X}$  for directions on the Poincaré sphere, or  $\mathcal{X}$  the set of (labelled) graphs equipped with the  $\sigma$ -algebra generated by subgraph counts. We then define its centered  $n^{\text{th}}$  moments by

$$F_\rho^{(n)}(\Omega) = \int_\Omega d(\rho(\omega), \mathbb{E}[\rho])^n P(d\omega),$$

where  $d$  is any metric (or norm) on  $\mathcal{X}$  and  $\mathbb{E}[\rho] \in \mathcal{X}$  denotes the appropriate “mean” of  $\rho$  (e.g. the Fréchet mean when  $\mathcal{X}$  is non-linear).

**Remark 3** More generally, if  $(\mathcal{X}, d)$  is a metric space, one defines the Fréchet mean

$$\mu = \arg \min_{y \in \mathcal{X}} \int_\Omega d(\rho(\omega), y)^2 P(d\omega),$$

and then set

$$F_\rho^{(n)}(\Omega) = \int_\Omega d(\rho(\omega), \mu)^n P(d\omega),$$

which recovers the usual  $\mathbb{E}[\|\rho - \mathbb{E}[\rho]\|^n]$  in linear settings.

In statistical mechanics, the map  $\rho$  is replaced by an observable  $\mathcal{A}$  taking values in a normed space  $(V, \|\cdot\|)$ :

1. **Classical distribution.** A probability density  $f: \Gamma \rightarrow [0, \infty)$  on phase space  $\Gamma = \{(\mathbf{x}, \mathbf{p})\} \subset \mathbb{R}^{2d}$ ,  $\int_\Gamma f = 1$ , with expectation  $\mathbb{E}_f[A] = \int_\Gamma A(\mathbf{x}, \mathbf{p}) f(\mathbf{x}, \mathbf{p}) d\mathbf{x} d\mathbf{p}$ .
2. **Quantum density operator.** A trace-one positive operator  $\hat{\rho}$  on a separable Hilbert space  $\mathcal{H}$ , with  $\text{Tr}(\hat{\rho}) = 1$ , and expectation  $\text{Tr}_{\hat{\rho}}[A] = \text{Tr}(\hat{\rho} A)$ .

The unified  $n^{\text{th}}$ -order fluctuation of  $A$  is

$$F_A^{(n)} = \begin{cases} \mathbb{E}_f[\|A - \mathbb{E}_f[A]\|^n], & \text{(classical),} \\ \text{Tr}_{\hat{\rho}}[|A - \text{Tr}_{\hat{\rho}}[A]|^n], & \text{(quantum),} \end{cases}$$

which recovers [\(2.5\)](#) when  $A = \rho$  and [\(2.6\)](#) when  $A = \hat{\rho}$ .

**Classical mechanics: Liouville, Boltzmann, and probability-flow ODEs.** Let  $f(t, \mathbf{x}, \mathbf{p})$  denote the phase-space density on  $\mathbb{R}^{2d}$  evolving under Hamilton’s equations

$$\dot{\mathbf{x}} = \nabla_{\mathbf{p}} H(\mathbf{x}, \mathbf{p}), \quad \dot{\mathbf{p}} = -\nabla_{\mathbf{x}} H(\mathbf{x}, \mathbf{p}).$$

Liouville’s theorem asserts the continuity (or collisionless) equation

$$\partial_t f + \{f, H\} = 0, \quad \{f, H\} = (\nabla_{\mathbf{p}} H) \cdot \nabla_{\mathbf{x}} f - (\nabla_{\mathbf{x}} H) \cdot \nabla_{\mathbf{p}} f.$$

A stationary solution of this is the Maxwell–Boltzmann distribution

$$f_{\text{MB}, \beta}(\mathbf{x}, \mathbf{p}) = \frac{\exp(-\beta H(\mathbf{x}, \mathbf{p}))}{Z(\beta)}, \quad Z(\beta) = \int_{\mathbb{R}^{2d}} \exp(-\beta H(\mathbf{x}, \mathbf{p})) d\mathbf{x} d\mathbf{p},$$

which satisfies  $\partial_t f_{\text{MB}, \beta} = 0$ .

967 **Remark 4** By contrast, the Boltzmann transport equation

$$\partial_t f + \{f, H\} = C[f],$$

968 includes the collision integral  $C[f]$ , modeling irreversible interactions that drive  $f$  toward  $f_{\text{MB},\beta}$   
969 under the molecular-chaos assumption.

970 In variance-preserving diffusion, one instead considers the marginal  $p_t(\mathbf{x})$  evolving under the  
971 probability-flow ODE

$$\dot{\mathbf{x}}_t = -\frac{1}{2}\beta(t)\mathbf{x}_t - \beta(t)\nabla_{\mathbf{x}}\log p_t(\mathbf{x}_t), \quad (\text{Eq. (2.2)})$$

972 whose push-forward densities satisfy the continuity equation

$$\partial_t p_t + \nabla_{\mathbf{x}} \cdot (p_t v_t) = 0, \quad v_t(\mathbf{x}) = -\frac{1}{2}\beta(t)\mathbf{x} - \beta(t)\nabla_{\mathbf{x}}\log p_t(\mathbf{x}).$$

973 One checks that if  $\beta(t) \equiv \beta$  is constant and

$$p_t(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\|\mathbf{x}\|^2\right),$$

974 then  $v_t \equiv 0$  and  $p_t$  is stationary for the ODE (i.e.  $\partial_t p_t = 0$ ). Thus, the probability-flow ODE is  
975 the gradient-flow analogue of Liouville’s equation—deterministic and reversible but encoding a  
976 coarse-grained score term—yet still falls short of the full irreversibility of the Boltzmann equation.  
977 The fluctuation formalism for Liouville applies verbatim to diffusion models (see [Song et al., 2021](#),  
978 App. D; [Biroli and Mézard, 2023](#); [Biroli et al., 2024](#); [Raya and Ambrogioni, 2024](#)).

979 **Quantum mechanics.** A quantum state is described by a density operator  $\hat{\rho}$  on a separable Hilbert  
980 space  $\mathcal{H}$ , satisfying  $\hat{\rho} \succeq 0$ , where  $\text{Tr } \hat{\rho} = 1$ . Under closed-system dynamics,  $\hat{\rho}(t)$  evolves by the von  
981 Neumann equation

$$\partial_t \hat{\rho}(t) = -i[\hat{H}, \hat{\rho}(t)],$$

982 so any  $\hat{\rho}$  commuting with  $\hat{H}$  is stationary. In particular, the Gibbs (Maxwell–Boltzmann) state at  
983 inverse temperature  $\beta = (k_B T)^{-1}$ ,

$$\hat{\rho}_{\text{MB},\beta} = \frac{e^{-\beta \hat{H}}}{Z(\beta)}, \quad Z(\beta) = \text{Tr}(e^{-\beta \hat{H}}),$$

984 satisfies  $\partial_t \hat{\rho}_{\text{MB},\beta} = 0$ . Open-system (Lindblad) dynamics add dissipative terms that drive  $\hat{\rho}$  toward  
985  $\hat{\rho}_{\text{MB},\beta}$ . All fluctuation measures  $F^{(n)}$  from [Section 2.2](#) carry over under the replacement  $\mathbb{E}_f[\cdot] \mapsto$   
986  $\text{Tr}_{\hat{\rho}}[\cdot]$ , with the discrete spectrum of  $\hat{H}$  giving rise to “lattice” transitions absent in the continuous  
987 classical setting (see [Appendix B.14](#)).

988 **Fluctuation–dissipation relations.** Let  $A$  and  $B$  be observables with conjugate fields  $h_A$  and  $h_B$ .  
989 Writing  $\Delta A = A - \langle A \rangle$ , the static (equilibrium) fluctuation–dissipation theorem(FDT) states

$$\langle (\Delta A)^2 \rangle = k_B T \chi_A, \quad \chi_A = \left. \frac{\partial \langle A \rangle}{\partial h_A} \right|_{h_A=0}. \quad (\text{B.1})$$

990 The dynamic (Kubo) relation for the linear response function  $R_{AB}(t)$  and correlation function  
991  $C_{AB}(t) = \langle \Delta A(t) \Delta B(0) \rangle$  is

$$R_{AB}(t) = \frac{1}{k_B T} \frac{d}{dt} C_{AB}(t). \quad (\text{B.2})$$

992 Together, [Eqs. \(B.1\) and \(B.2\)](#) imply that, at a phase transition, the static susceptibility  $\chi_A$  diverges  
993 precisely when the normalized cross-fluctuation

$$\mathcal{M}_{\rho}^{(1)}(\Omega_{i,t}, \Omega_{j,t}) = \frac{\langle \Delta A_i(t) \Delta A_j(t) \rangle}{\sqrt{\langle (\Delta A_i)^2 \rangle \langle (\Delta A_j)^2 \rangle}}$$

994 drops sharply, since the system becomes unable to distinguish the macrostates  $\Omega_i$  and  $\Omega_j$ .

995 **Connection to our framework.** Identifying

$$A \leftrightarrow \mathbb{1}_{\Omega_{i,t}}, \quad B \leftrightarrow \mathbb{1}_{\Omega_{j,t}},$$

996 in the dynamic FDT (B.2) gives

$$R_{AB}(t) = \frac{1}{k_B T} \frac{d}{dt} \langle \Delta A(t) \Delta B(0) \rangle = \frac{d}{dt} \mathbb{E}[\mathbb{1}_{\Omega_{i,t}} \mathbb{1}_{\Omega_{j,0}}] = G_\rho^{(1)}(\Omega_{i,t}, \Omega_{j,t}).$$

997 Consequently, the normalized cross-fluctuation

$$\mathcal{M}_\rho^{(1)}(\Omega_{i,t}, \Omega_{j,t}) = \frac{G_\rho^{(1)}(\Omega_{i,t}, \Omega_{j,t})}{\sqrt{F_\rho^{(2)}(\Omega_{i,t}) F_\rho^{(2)}(\Omega_{j,t})}}$$

998 acts as a correlation length, and the time  $t_{\text{merge}}$  at which  $\mathcal{M}^{(1)}$  drops sharply can be viewed as a finite-  
 999 size critical point along the diffusion trajectory. The same correspondence extends to higher-order  
 1000 fluctuations  $G_\rho^{(n)}$  and  $\mathcal{M}_\rho^{(n)}$ .

1001 **Classical analytical techniques.** Classical methods such as the random-energy model (REM) and  
 1002 the renormalization group (RG) analyze phase transitions by extracting *critical exponents* that govern  
 1003 the divergence of correlation lengths and susceptibilities [Kivelson et al., 2024]. However, these  
 1004 techniques rely on continuum assumptions and do not directly extend to the lattice-type transitions  
 1005 encountered in quantum systems with discrete spectra [Sachdev, 1999]; accordingly, we do not  
 1006 deploy them here. Biroli et al. [2024], however, successfully applied the REM in a setting identical  
 1007 to our case study (Section 4.2), revealing three distinct classical phases. While those REM-derived  
 1008 phases differ in nature from the merger transitions we analyze below (see Appendix B.14), the two  
 1009 viewpoints remain fully compatible.

1010 **Remark 5 (Beyond equilibrium)** *Out-of-equilibrium systems can exhibit dynamical phase tran-*  
 1011 *sitions, signaled by nonanalytic behavior of large-deviation functions or thermodynamic poten-*  
 1012 *tials [Chaikin et al., 1995]. The fluctuation framework remains applicable in such settings: for*  
 1013 *instance, Biroli et al. [2024] demonstrated its efficacy on diffusion models. In particular, our merger*  
 1014 *statistic*

$$\mathcal{M}_\rho^{(n)}(\Omega_{i,t}, \Omega_{j,t})$$

1015 *precisely captures the lattice-type transitions that arise in these high-dimensional, nonequilibrium*  
 1016 *trajectories.*

1017 **Working with multiple states.** In Eqs. (2.6) and (2.8) we defined

$$G_{\rho_1, \rho_2}^{(n)} \quad \text{and} \quad \mathcal{M}_{\rho_1, \rho_2}^{(n)},$$

1018 allowing  $\rho_1$  and  $\rho_2$  to be *heterogeneous* (i.e.,  $\rho_1 \neq \rho_2$ ). As Example 1 in Section 2.2 illustrates,  
 1019 different state-operator choices can yield qualitatively different cross-fluctuation behaviour. This  
 1020 flexibility is crucial in quantum mechanics, where one often considers distinct pure-state projectors

$$\rho_i = |\psi_i\rangle \langle \psi_i|$$

1021 on a Hilbert space  $\mathcal{H}$ .<sup>6</sup> Their cross-fluctuations underpin tasks such as quantum state tomography  
 1022 and discrimination of non-orthogonal states Peres [1995], Nielsen and Chuang [2010].

1023 More generally, heterogeneous  $\rho_1, \rho_2$  detect an *alignment* between two evolving state spaces. For  
 1024 example, if text and image embeddings are both driven by the same diffusion dynamics, choosing  $\rho_1$   
 1025 and  $\rho_2$  to sample each modality reveals how a given concept manifests across them. We leave such  
 1026 multimodal extensions for future work.

---

<sup>6</sup>Mathematically, measurements project the density operator onto the eigenspace associated with the outcome; see Nielsen and Chuang [2010] for details.

## 1027 B.2 On the validity of fluctuation theory

1028 Fluctuations are defined through derivatives of the *characteristic function* (CF)  $\varphi_X(t) =$   
 1029  $\mathbb{E}[e^{itX}]$ ,  $t \in \mathbb{R}$ , which always exists and satisfies  $|\varphi_X(t)| \leq 1$ . If  $X$  has moments up to order  $n$ ,  
 1030 differentiation at the origin gives

$$\varphi_X^{(n)}(0) = i^n \mathbb{E}[X^n], \quad n \in \mathbb{N},$$

1031 so the centred fluctuation  $F_\rho^{(n)}(\Omega)$  of [Eq. \(2.5\)](#) is completely encoded by  $\varphi_X$ . Hence, the fluctuation  
 1032 tensor  $F_\rho^{(n)}(\Omega)$  of [\(2.5\)](#) is completely determined by  $\varphi_X$ .

### 1033 Foundational facts.

- 1034 1. **Uniqueness.** If  $\varphi_X = \varphi_Y$  then  $X \stackrel{d}{=} Y$  [[Lukacs, 1970](#)].
- 1035 2. **Inversion.** For continuous distribution functions,

$$\text{CDF}(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{\Im(e^{-itx} \varphi_X(t))}{t} dt \quad [\text{Dudley, 2018}].$$

- 1036 3. **Convolution.** For independent  $X, Y$ ,  $\varphi_{X+Y}(t) = \varphi_X(t) \varphi_Y(t)$ .

1037 **Theorem 6 (Bochner)** *A function  $\varphi: \mathbb{R} \rightarrow \mathbb{C}$  is a characteristic function if and only if it is posi-*  
 1038 *tive-definite, continuous at the origin, and  $\varphi(0) = 1$ .*

1039 *Proof.* See [Rudin \[1962, Thm. 15.2\]](#).

1040 **Remark 7** *We assume the random variable (or field)  $\rho$  has characteristic function  $\varphi_\rho(t) = \mathbb{E}[e^{i\langle t, \rho \rangle}]$ ,*  
 1041 *i.e. the Fourier transform of its law. For measures with Schwartz-class densities  $\mathcal{S}(\mathbb{R}^d)$  [[Blanchard](#)*  
 1042 *and [Brüning, 2015](#)], this holds automatically, a mild regularity satisfied by most machine-learning*  
 1043 *data distributions (images, text) and physical constructs (Onsager matrices, Wigner transforms)*  
 1044 *[[Bertini et al., 2002](#)]. Critically, characteristic functions exist even for purely discrete or singular*  
 1045 *continuous measures lacking classical densities [[Dudley, 2018](#), [Rudin, 1962](#)], making this assumption*  
 1046 *more fundamental than density existence and hence of greater utility. For examples of this broader*  
 1047 *utility, see [[Ansari et al., 2020](#), [Sriperumbudur et al., 2010](#)].*

1048 **Remark 8 (When derivatives fail)** *If  $\varphi_X$  is not differentiable at the origin, as for the Cauchy law*  
 1049 *with  $\varphi_X(t) = e^{-|t|}$ —classical moments, and hence ordinary fluctuations, do not exist. We can still*  
 1050 *define weak derivatives, for example*

$$\varphi_X^{(1,w)}(t) = -e^{-|t|} \text{sgn}(t),$$

1051 *but these belong to the tempered-distribution space  $\mathcal{S}$ . Extending the fluctuation framework to such*  
 1052 *cases requires a generalized (distribution-valued) moment theory, which we leave for future work.*

## 1053 B.3 Pull-back of data events along the PF-ODE

1054 Let  $\Phi_{s \rightarrow t}: \mathbb{R}^d \rightarrow \mathbb{R}^d$  be the *flow map* of the probability-flow ODE [\(2.2\)](#). Because the drift field is  
 1055 globally Lipschitz,  $\Phi_{s \rightarrow t}$  is a bijection for every  $0 \leq s < t \leq T$ . The continuity equation

$$\partial_t p_t + \nabla \cdot (p_t v_t) = 0$$

1056 implies that  $p_t$  is the push-forward of  $p_s$  under  $\Phi_{s \rightarrow t}$

$$p_t = (\Phi_{s \rightarrow t})_\# p_s, \quad P_t(B) = P_s(\Phi_{s \rightarrow t}^{-1}(B)) \quad \forall B \subseteq \mathbb{R}^d,$$

1057 where  $P_t$  is the law of  $p_t$ . Equivalently,

$$p_t(x) = p_s(\Phi_{t \rightarrow s}(x)) \exp\left(-\int_s^t \nabla \cdot v_u(\Phi_{u \rightarrow s}(x)) du\right).$$

1058 **From source events to time  $t$  images.** Fix two disjoint source events  $\Omega_{1,0}, \Omega_{2,0} \subseteq \text{supp}(p_0)$ , and  
 1059 define their time- $t$  pre-images

$$\Omega_{k,t} = \Phi_{0 \rightarrow t}^{-1}(\Omega_{k,0}), \quad k \in \{1, 2\}, t \in [0, T].$$

1060 Since  $\Phi_{0 \rightarrow t}$  is a diffeomorphism, each  $\Omega_{k,t}$  is measurable however may show overlap even though  
 1061 the sources are disjoint.

1062 **Event probabilities are preserved along the flow.** For a measurable map  $X : (\Omega, \mathcal{F}, P) \rightarrow (S, \mathcal{S})$   
 1063 write  $P_X(B) = P(X^{-1}(B))$ . With  $X = \Phi_{0 \rightarrow t}$  and  $B = \Omega_{k,0}$  we obtain

$$P_t(\Omega_{k,t}) = P_0(\Omega_{k,0}), \quad k \in \{1, 2\},$$

1064 so  $\{\Omega_{k,t}\}$  are well-defined events for the marginal density  $p_t$  at every  $t$ .

1065 **Monitoring macroscopic mixing.** Even when  $\Omega_{1,t} \cap \Omega_{2,t} \neq \emptyset$ , the centered cross-fluctuation  
 1066  $\mathcal{M}_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t})$  remains well defined. Its sharp rise toward one marks the moment the two  
 1067 macrostates become indistinguishable *in distribution*, i.e. the merger time  $t_{\text{merge}}$ .

#### 1068 **B.4 From the empirical reverse process to the learned sampler**

1069 During training, a diffusion model  $f_\theta$  minimizes the *simplified ELBO*:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{i \sim \text{Unif}\{0, \dots, T-1\}} \left\| f_\theta(\mathbf{x}_i(\mathbf{x}_0, \varepsilon), i) - \varepsilon \right\|_2^2, \quad (\text{B.3})$$

$$\mathbf{x}_0 \sim p_0, \varepsilon \sim \mathcal{N}(0, I)$$

1070 where

$$\mathbf{x}_i(\mathbf{x}_0, \varepsilon) = \sqrt{\alpha_i} \mathbf{x}_0 + \sqrt{1 - \alpha_i} \varepsilon \sim p_i \quad (\text{cf. Eq. (2.3)}).$$

1071 Up to an additive constant, (B.3) is a Monte Carlo estimate of the KL divergence between the true  
 1072 reverse kernel and the empirical Gaussian kernel parameterized by  $f_\theta$  [Ho et al., 2020]. At its global  
 1073 minimum ( $\mathcal{L}_{\text{simple}} = 0$ ), the network implements the exact empirical reverse process [Haussmann  
 1074 and Pardoux, 1986, Cattiaux et al., 2023], leading to perfect reconstruction and mode collapse.

1075 Since our merger time  $t_{\text{merge}}$  is defined by the cross-fluctuation  $\mathcal{M}_\rho^{(n)}$  (see Appendix B.1) for the  
 1076 empirical forward process, we must ensure the learned sampler tracks the true reverse chain accurately  
 1077 enough that its induced error does not spuriously trigger or mask a merger.

1078 **Finite-error regime and TV-tracking.** In practice  $\mathcal{L}_{\text{simple}} > 0$ . Let  $\varepsilon_\star = \sqrt{\mathcal{L}_{\text{simple}}}$  denote the  
 1079 root-mean-square score error.

1080 Under standard regularity—each true score  $\nabla_{\mathbf{x}} \log p_i$  is  $L$ -Lipschitz,  $p_0$  has finite second moment,  
 1081 and  $\text{KL}(p_0 \parallel \mathcal{N}) < \infty$ —the learned sampler tracks the true reverse chain in total variation:

1082 **Theorem 9 (Chen et al., 2022a, Thm. 1)** *If  $\varepsilon_\star \leq \varepsilon$  and one uses  $n = O(L^2 d / \varepsilon)$  reverse-chain*  
 1083 *steps of appropriate size, then for all  $i = 0, \dots, T$ ,*

$$\|p_i - \hat{p}_i\|_{\text{TV}} \leq \varepsilon,$$

1084 *where  $\hat{p}_i$  is the marginal produced by  $f_\theta$ .*

1085 Thus, a sufficiently accurate network provides a proxy for the ideal reverse SDE. By time-reversibility,  
 1086 one may equivalently analyze the *forward* chain as a computationally cheaper surrogate for detecting  
 1087 mergers. Indeed, letting  $\Phi_{0 \rightarrow t}$  denote the forward flow map (see Appendix B.3), we can pull back  
 1088 the empirical reverse steps and compute  $\mathcal{M}_\rho^{(n)}$  directly on the evolving events  $\{\Omega_{i,t}\}$ , avoiding full  
 1089 reverse simulation.

1090 As  $\varepsilon_\star \rightarrow 0$ , the model nears mode collapse. Nonetheless, in one-dimensional settings Chen [2025]  
 1091 show that even if  $f_\theta$  converges smoothly to an interpolated variant of the empirical score with near  
 1092 vanishing loss, it can be capable of generating diverse samples.



## 1093 B.5 Relation between cross-fluctuation and within-event fluctuation based thresholds

1094 Consider Eq. (3.1) which is defined as follows

$$\widetilde{\mathcal{M}}_\rho^{(n)}(i) = \begin{cases} \mathcal{M}_\rho^{(n)}(\Omega_{1,i}, \Omega_{2,i}), & d(\widehat{F}_\rho^{(2n)}(\Omega_{1,i}), \widehat{F}_\rho^{(2n)}(\Omega_{2,i})) > \varepsilon, \\ 1, & \text{otherwise.} \end{cases} \quad (\text{B.4})$$

1095 An alternative formulation for detecting a transition in the normalized cross-fluctuation can instead  
1096 utilize the value of  $\mathcal{M}_\rho^{(n)}(\Omega_{1,i}, \Omega_{2,i})$  itself

$$\widetilde{\mathcal{M}}_\rho^{(n)}(i) = \begin{cases} \mathcal{M}_\rho^{(n)}(\Omega_{1,i}, \Omega_{2,i}), & |\mathcal{M}_\rho^{(n)}(\Omega_{1,i}, \Omega_{2,i}) - 1| > \vartheta, \\ 1, & \text{otherwise,} \end{cases} \quad (\text{B.5})$$

1097 We show that Eqs. (B.4) and (B.5) are topologically equivalent for a broad class of distance metrics  
1098  $d(\cdot, \cdot)$  under Lipschitz state operators. However, (B.4) is computationally more efficient, making it  
1099 our preferred choice in this work.

1100 **Theorem 10** *Let the state operator  $\rho$  be Lipschitz and  $(\Gamma, d_{\text{fluc}})$  be a complete metric space over*  
1101 *events  $\Omega_i$  such that the metric  $d_{\text{fluc}}$  is a distance metric  $d$  over  $2n^{\text{th}}$  order within-event fluctuations*  
1102  *$\widehat{F}_\rho^{(2n)}(\Omega_i)$  equivalent to an  $L_p$ -norm over  $\widehat{F}_\rho^{(2n)}(\Omega_i)$  for some  $p > 1$ . Define the similarity metric*

$$d_{\text{sim}}(\Omega_1, \Omega_2) := |\mathcal{M}_\rho^{(n)}(\Omega_1, \Omega_2) - 1|.$$

1103 *Then, the metrics  $d_{\text{fluc}}$  and  $d_{\text{sim}}$  are topologically equivalent on  $\Gamma$ .*

1104 *Proof.* Suppose  $(X_m, Y_m) \subset \Gamma \times \Gamma$  is a sequence such that  $d_{\text{fluc}}(X_m, Y_m) \rightarrow 0$ . But then by  
1105 definition we get,

$$d(\widehat{F}_\rho^{(2n)}(X_m), \widehat{F}_\rho^{(2n)}(Y_m)) \rightarrow 0.$$

1106 By the equivalence of  $d$  to the  $L_p$  norm, it follows that

$$\|\widehat{F}_\rho^{(2n)}(X_m) - \widehat{F}_\rho^{(2n)}(Y_m)\|_{L_p} \rightarrow 0.$$

1107 Using Hölder's inequality for  $q = 1$ , we have

$$\|\widehat{F}_\rho^{(2n)}(X_m) - \widehat{F}_\rho^{(2n)}(Y_m)\|_{L_1} \leq \|\widehat{F}_\rho^{(2n)}(X_m) - \widehat{F}_\rho^{(2n)}(Y_m)\|_{L_p} \cdot \mu(S)^{\frac{p-1}{p}},$$

1108 where  $\mu(S)$  is the measure of the underlying space. Hence,

$$\|\widehat{F}_\rho^{(2n)}(X_m) - \widehat{F}_\rho^{(2n)}(Y_m)\|_{L_1} \rightarrow 0.$$

1109 By Jensen's inequality and for Lipschitz  $\rho$ ,

$$\|\widehat{F}_\rho^{(n)}(X_m) - \widehat{F}_\rho^{(n)}(Y_m)\|_{L_1} \leq \sqrt{\|\widehat{F}_\rho^{(2n)}(X_m) - \widehat{F}_\rho^{(2n)}(Y_m)\|_{L_1}} \rightarrow 0.$$

1110 Since

$$d_{\text{sim}}(X_m, Y_m) = |\mathcal{M}_\rho^{(n)}(X_m, Y_m) - 1|$$

1111 we conclude that

$$d_{\text{sim}}(X_m, Y_m) \rightarrow 0.$$

1112 The reverse direction follows from the continuity and Lipschitz property of  $\mathcal{M}_\rho^{(n)}$  in terms of  $\widehat{F}_\rho^{(n)}$   
1113 under the  $L_1$  norm, yielding topological equivalence between  $(\Gamma, d_{\text{fluc}})$  and  $(\Gamma, d_{\text{sim}})$ .

1114 Since the metric spaces  $(\Gamma, d_{\text{fluc}})$  and  $(\Gamma, d_{\text{sim}})$  are topologically equivalent, their open balls are  
1115 homeomorphic. Hence, for any  $\varepsilon$ -ball  $B_\varepsilon(x) \subseteq (\Gamma, d_{\text{fluc}})$ , there exists a corresponding  $\vartheta$ -ball  
1116  $B_\vartheta(y) \subseteq (\Gamma, d_{\text{sim}})$  such that the bijection  $f : x \mapsto y$  holds. Therefore, transitions detected via  
1117 thresholds on  $d_{\text{fluc}}$  in (B.4) are in principle identical to those detected via thresholds on  $d_{\text{sim}}$  in (B.5).

1118 We also show the following theorem that tracks the evolution of the geometry of the metric space of  
1119 fluctuations through time.

**Theorem 11 (Exponential contraction of fluctuations and MST construction)** Let  $\rho$  be a Lipschitz state operator, and consider the  $2n^{\text{th}}$ -order fluctuation embeddings  $\widehat{F}_\rho^{(2n)}(\Omega_i(t))$  of events  $\Omega_i$  evolving under a hypercontractive diffusion semigroup  $P_t$  on a complete metric space of events  $(\Gamma, d_{\text{fluc}})$  where  $d_{\text{fluc}}$  is defined identically to [Theorem 10](#). Then, there exist constants  $C, \lambda > 0$  such that for any  $i, j$  and  $t \geq 0$ ,

$$d_{\text{fluc}}(\Omega_i(t), \Omega_j(t)) \leq C e^{-\lambda t} d_{\text{fluc}}(\Omega_i(0), \Omega_j(0)).$$

Consequently, for any fixed  $\varepsilon > 0$  and a subspace  $\Gamma_{\text{sub}} \subseteq \mathcal{H}$  there exists  $T_\varepsilon(\Gamma_{\text{sub}}) > 0$  such that for all  $t \geq T_\varepsilon(\Gamma_{\text{sub}})$ , the graph on events with edges between pairs  $(\Omega_i, \Omega_j) \in \Gamma_{\text{sub}} \times \Gamma_{\text{sub}}$  satisfying

$$d_{\text{fluc}}(\Omega_i(t), \Omega_j(t)) \leq \varepsilon,$$

is connected and admits a well-defined minimal spanning tree (MST).

*Proof.* Consider the diffusion semigroup  $P_t$  acting on a suitable function space  $L_2(\mu)$ , where  $\mu$  is the invariant measure. The spectral gap inequality [Bakry et al. \[2014\]](#) states that there exists  $\lambda > 0$  such that for any centered function  $f$  (i.e.,  $\mathbb{E}[f] = 0$ ),

$$\|P_t f\|_{L_2} \leq e^{-\lambda t} \|f\|_{L_2}.$$

This implies that the semigroup contracts deviations from the mean exponentially fast in the  $L_2$ -norm. Since the fluctuation operators  $\widehat{F}_\rho^{(2n)}(\Omega_i(t))$  can be identified with such centered functions, their differences decay exponentially over time. By [Gross's logarithmic Sobolev inequality Gross \[1975\]](#), the semigroup  $P_t$  is hypercontractive, which implies that for  $p > 2$ , there exists a constant  $C = C(p, t)$  such that

$$\|P_t f\|_{L_p} \leq \|f\|_{L_2}.$$

Combining this with the spectral gap inequality, it follows that for some positive constants  $C_*, \lambda_*$

$$\|P_t f - \mathbb{E}[f]\|_{L_p} \leq C_* e^{-\lambda_* t} \|f - \mathbb{E}[f]\|_{L_p}.$$

Since the fluctuation distance  $d_{\text{fluc}}$  is assumed to be equivalent to an  $L_p$ -norm difference of fluctuation operators, we obtain for a fixed  $i, j$

$$d_{\text{fluc}}(\Omega_i(t), \Omega_j(t)) \leq C e^{-\lambda t} d_{\text{fluc}}(\Omega_i(0), \Omega_j(0)).$$

Clearly then for any fixed  $\varepsilon > 0$ , there exists a time  $T_\varepsilon$  such that for all  $t \geq T_\varepsilon$ ,

$$d_{\text{fluc}}(\Omega_i(t), \Omega_j(t)) \leq \varepsilon,$$

for every pair of events  $(\Omega_i, \Omega_j)$ . Define the graph  $G_t = (V, E_t)$  with vertices  $V = \Gamma_{\text{sub}}$  representing events in a subspace and edges

$$E_t = \{(i, j) : d_{\text{fluc}}(\Omega_i(t), \Omega_j(t)) \leq \varepsilon \text{ where } (\Omega_i, \Omega_j) \in \Gamma_{\text{sub}}^2 \setminus \text{diag}(\Gamma_{\text{sub}}^2)\}.$$

Here for the sake of clarity, we use the notation  $\Gamma_{\text{sub}}^2$  for  $\Gamma_{\text{sub}} \times \Gamma_{\text{sub}}$  and the diagonal  $\text{diag}(\Gamma_{\text{sub}}^2)$  refers to tuples of identical events, thus we only consider edges between distinct events at  $t = 0$  to avoid loops. Now, define

$$T_\varepsilon(\Gamma_{\text{sub}}) = \sup_{(\Omega_i, \Omega_j) \in \Gamma_{\text{sub}}^2 \setminus \text{diag}(\Gamma_{\text{sub}}^2)} \{T_\varepsilon \mid \forall t \geq T_\varepsilon, d_{\text{fluc}}(\Omega_i(t), \Omega_j(t)) \leq \varepsilon\}.$$

Then, for  $t \geq T_\varepsilon(\Gamma_{\text{sub}})$ , since all pairwise distances are below  $\varepsilon$ ,  $G_t$  is a complete graph and therefore connected. Because  $d_{\text{fluc}}$  is a metric, we can obtain a well-defined minimal spanning tree (MST) on  $G_t$ . [Kruskal \[1956\]](#).

From [Theorem 10](#), the above theorem extends naturally to the cross-fluctuation metric. Effectively, [Theorem 11](#) shows that considering sub-collections of events (i.e., sub- $\sigma$ -algebras under the probability measure  $p_0$ ), mergers induce a form of path-connectivity. This leads to two key implications:

1. We can refine partitions of the event space progressively down to singletons, enabling increasingly fine resolution on the same desired distribution.
2. Tracking the evolution of  $G_t$  until the time  $T_\varepsilon(\Gamma_{\text{sub}})$  when the graph becomes complete corresponds to connecting neighborhoods in the metric subspace  $(\Gamma_{\text{sub}}, d_{\text{fluc}})$ . This relates closely to neighborhood clustering methods such as t-SNE [van der Maaten and Hinton \[2008\]](#) and UMAP [McInnes et al. \[2018\]](#), we leave the exploration of this connection in more detail to future work.

## 1155 B.6 Coupling and mixing for discrete Markov chains

1156 We recall (and slightly adapt) the terminology of [Aldous and Fill \[2002\]](#), [Levin and Peres \[2017\]](#) so  
1157 that it aligns with the notation used in the main text.

1158 **Single chain and mixing time.** Let  $\mathcal{S} = \{S_0, S_1, \dots\}$  be the marginal sequence of an *ergodic*  
1159 Markov chain on a finite state space  $X$ , written in row-vector form. Its transition matrix is  $\Pi \in$   
1160  $[0, 1]^{X \times X}$  and the (unique) stationary distribution satisfies  $\Pi S_\infty = S_\infty$ . (The symbol  $\infty$  is used  
1161 instead of  $m$  to emphasise that the limit need not be reached in finitely many steps.)

1162 For two probability vectors  $p, q$  on  $X$  define the *total-variation* distance

$$d_{\text{TV}}(p, q) = \frac{1}{2} \sum_{x \in X} |p(x) - q(x)| = \sup_{A \subseteq X} |P(A) - Q(A)|.$$

1163 The  $\varepsilon$ -*mixing time* of the chain is

$$t_{\text{mix}}(\varepsilon) = \min\{t \geq 0 : d_{\text{TV}}(S_t, S_\infty) \leq \varepsilon\}. \quad (\text{B.6})$$

1164 **Multiple chains and coupling time.** Let  $\Lambda$  be a finite index set. For every  $\lambda \in \Lambda$  fix an *initial event*  
1165  $\Omega_{\lambda,0} \subseteq X$  with  $\{\Omega_{\lambda,0}\}_\lambda$  pairwise disjoint and  $\sum_\lambda S_0(\Omega_{\lambda,0}) = 1$ . Run the *same* transition matrix  
1166 on each event to obtain the collection of chains  $\mathcal{S}_\lambda = \{S_{\lambda,t}\}_{t \geq 0}$ , where  $S_{\lambda,t}$  is  $S_t$  restricted to  $\Omega_{\lambda,0}$   
1167 and renormalised. Because the dynamics are identical,  $S_{\lambda,\infty} = S_\infty$  for every  $\lambda$ , but for finite  $t$  the  
1168 restricted marginals differ. Given  $\varepsilon > 0$ , we define the *coupling time*

$$t_{\text{cpl}}(\varepsilon) = \min\left\{t \geq 0 : \max_{\alpha, \beta \in \Lambda} d_{\text{TV}}(S_{\alpha,t}, S_{\beta,t}) \leq \varepsilon\right\}. \quad (\text{B.7})$$

1169 Hence  $t_{\text{cpl}}$  measures when *all* class-conditioned chains have coalesced in distribution, whereas  $t_{\text{mix}}$   
1170 monitors convergence to the common stationary measure.

1171 **Limitations in continuous state spaces.** The total-variation distance is well-behaved only for  
1172 discrete  $X$ ; for absolutely continuous measures, it degenerates into a trivial bound (see [Bhattacharyya](#)  
1173 [et al., 2024](#), [Tao et al., 2024](#)). Consequently, definitions (B.6)–(B.7) do *not* extend to diffusion models,  
1174 whose state space is  $\mathbb{R}^d$ .

1175 For continuous densities whenever second moments exist, we therefore define the *generalised*  
1176 *coupling time*

$$t_{\text{gen}}(\varepsilon) = \min\left\{t \geq 0 : \max_{\alpha, \beta \in \Lambda} \mathcal{M}_\rho^{(n)}(\Omega_{\alpha,t}, \Omega_{\beta,t}) \geq 1 - \varepsilon\right\},$$

1177 which generalizes the coupling time to Euclidean diffusion processes and coincides with (B.7) when  
1178  $X$  is finite.

## 1179 B.7 Fluctuation moments bound total variation distance

1180 We show that asymptotically utilizing fluctuation theory to understand mergers is equivalent to  
1181 probing the similarity of probability distributions using the Total Variation Distance.

1182 **Proposition 12 (Moment–TV inequality)** Fix an integer  $n \geq 2$ . Let  $p, q$  be probability densities  
1183 on  $\mathbb{R}$  that

- 1184 (i) are of bounded variation;
- 1185 (ii) admit centred moments  $\mu_p^{(k)}, \mu_q^{(k)}$  for  $k = 1, \dots, n+1$ ;
- 1186 (iii) obey the moment proximity bound  $|\mu_p^{(k)} - \mu_q^{(k)}| \leq M$  for  $k = 1, \dots, n$ ;
- 1187 (iv) satisfy the uniform first- and second-moment bound  $|\mu_\bullet^{(1)}|, \mu_\bullet^{(2)} \leq B$ ;
- 1188 (v) have matching tails:  $\lim_{R \rightarrow \infty} \int_{|x| > R} (p - q) = 0$ .

1189 Then

$$d_{\text{TV}}(p, q) \leq C_n(M^2 + B),$$

1190 where one may take  $C_n = c_0(1 + n!)(2^n + 48)$  and  $c_0 > 0$  is an absolute constant.

1191 *Proof. 1. Characteristic-function bound.* Let  $f_p(t) = \mathbb{E}_p[e^{itX}]$  and  $f_q(t) = \mathbb{E}_q[e^{itX}]$ . Since  $p, q$  are  
 1192 absolutely continuous and of bounded variation,  $f_p, f_q$  are bounded by 1 and possess derivatives up  
 1193 to order  $n + 1$  at the origin. Write the order- $n$  Taylor expansions with integral remainder

$$f_p(t) = \sum_{k=0}^n \frac{(it)^k}{k!} \mu_p^{(k)} + \frac{(it)^{n+1}}{n!} \int_0^1 (1-s)^n \mu_p^{(n+1)} e^{istX} ds.$$

1194 Subtract the analogous expression for  $f_q$ , use hypothesis (iii), and take absolute values:

$$|f_p(t) - f_q(t)| \leq \sum_{k=1}^n \frac{|t|^k}{k!} M + \frac{|t|^{n+1}}{(n+1)!} (\mu_p^{(n+1)} + \mu_q^{(n+1)}). \quad (\text{B.8})$$

1195 2. *Bounding the  $(n+1)$ st moments.* By Jensen's inequality,  $\mu_{\bullet}^{(n+1)} \leq (\mu_{\bullet}^{(2)})^{(n+1)/2} \leq B^{(n+1)/2}$ .  
 1196 Insert this in (B.8) to get

$$|f_p(t) - f_q(t)| \leq a_n M |t| + b_n B^{(n+1)/2} |t|^{n+1}, \quad (\text{A}_n)$$

1197 where  $a_n := \sum_{k=1}^n \frac{1}{k!}$ ,  $b_n := \frac{2}{(n+1)!}$ .

1198 3. *Esseen's smoothing inequality.* For any  $T > 0$  (Ibragimov., 1975, Thm. 1.5.4),

$$d_{\text{TV}}(p, q) \leq \frac{1}{2\pi} \int_{-T}^T \left| \frac{f_p(t) - f_q(t)}{t} \right| dt + \frac{24}{\pi T} (\text{Var}(p) + \text{Var}(q)). \quad (\text{B.9})$$

1199 *Integral term:* divide (A<sub>n</sub>) by  $|t|$  and integrate,

$$\frac{1}{2\pi} \int_{-T}^T \left| \frac{f_p - f_q}{t} \right| \leq a_n M T + \frac{b_n}{n+1} B^{(n+1)/2} T^{n+1}.$$

1200 *Variance term:* hypothesis (iv) yields  $\text{Var}(p), \text{Var}(q) \leq B + M^2$ , so the second term in (B.9) is  
 1201 bounded by  $48(B + M^2)/(\pi T)$ .

1202 4. *Choice of  $T$ .* Set  $T = 1$ . (A different  $T$  only rescales the constant.) The bounds become

$$\begin{aligned} d_{\text{TV}}(p, q) &\leq [a_n + b_n] M + [a_n + b_n] B^{(n+1)/2} + \frac{48}{\pi} (B + M^2) \\ &\leq C_n (M^2 + B), \end{aligned}$$

1203 where the last line uses  $M \leq M^2 + 1$  and  $B^{(n+1)/2} \leq 2^n B$  for  $B \geq 1$ , and absorbs all numeric  
 1204 factors into  $C_n = c_0(1 + n!)(2^n + 48)$  with a universal  $c_0$ .

1205 **Remark 13** If  $p, q$  are sub-Gaussian (or sub-exponential) [Vershynin, 2018], all moments exist and  
 1206 satisfy  $\mu_p^{(k)} = O((\sqrt{B})^k)$ ; the conditions of Proposition 12 are then automatically satisfied on  $\mathbb{R}^d$ .

## 1207 B.8 Stochastic-flow formulation for the SDE view

1208 Appendix B.3 defined  $\Omega_{k,t} = \Phi_{0 \rightarrow t}^{-1}(\Omega_{k,0})$  via the deterministic flow  $\Phi_{s \rightarrow t}$  of the PF-ODE (2.2). We  
 1209 now show that the same construction works pathwise for the stochastic forward SDE

$$d\mathbf{x}_t = -\frac{1}{2}\beta(t) \mathbf{x}_t dt + \sqrt{\beta(t)} d\mathbf{w}_t,$$

1210 whose solution map  $x_0 \mapsto x_t$  depends on the Wiener path  $\omega \in \Omega_{\text{prob}}$ .

1211 **Kunita’s stochastic flow of diffeomorphisms.** Let  $\varphi_{s,t}(\omega, \cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ,  $0 \leq s \leq t \leq T$ , denote  
 1212 the *Kunita flow* generated by (2.1) [Kunita, 1990, Ch. 4]. For every fixed  $\omega$ , the map  $x \mapsto \varphi_{s,t}(\omega, x)$   
 1213 is a  $C^1$  diffeomorphism and

$$\varphi_{s,u}(\omega, \cdot) = \varphi_{t,u}(\omega, \varphi_{s,t}(\omega, \cdot)), \quad 0 \leq s \leq t \leq u \leq T.$$

1214 Hence the pathwise inverse  $\varphi_{0,t}^{-1}(\omega, \cdot)$  exists almost surely.

1215 Given two disjoint data events  $\Omega_{1,0}, \Omega_{2,0} \subset \mathbb{R}^d$  set

$$\Omega_{k,t}(\omega) := \varphi_{0,t}^{-1}(\omega, \Omega_{k,0}), \quad k \in \{1, 2\}.$$

1216 The map  $\omega \mapsto \mathbb{1}_{\Omega_{k,t}(\omega)}(x)$  is  $\mathcal{F}_t$ -measurable (Kunita’s measurability theorem), so  $\Omega_{k,t}$  is a *random*  
 1217 *closed set*. Its law equals the push-forward of  $p_0$  by the SDE:

$$\mathbb{P}\{x_t \in A\} = \mathbb{E}_\omega \mathbb{P}\{\varphi_{0,t}(\omega, x_0) \in A\}, \quad x_0 \sim p_0,$$

1218 and we again write  $p_t = \mathcal{L}(x_t)$ .

1219 **Annealed cross-fluctuations.** Fix  $n \leq 4$ . Because  $\varphi_{0,t}^{-1}(\omega, \cdot)$  is  $C^1$  and  $p_0$  has finite  $n$ -th moments,  
 1220 the pathwise fluctuation  $F_\rho^{(n)}(\Omega_{k,t}(\omega))$  exists. Define the *annealed* quantities

$$\overline{F}_\rho^{(n)}(\Omega_{k,t}) := \mathbb{E}_\omega [F_\rho^{(n)}(\Omega_{k,t}(\omega))], \quad \overline{\mathcal{M}}_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t}) := \mathbb{E}_\omega [\mathcal{M}_\rho^{(n)}(\Omega_{1,t}(\omega), \Omega_{2,t}(\omega))].$$

1221 Expectation commutes with the finite sums and integrals that define moments, so every identity from  
 1222 **Section 2.2** remains valid after adding bars. Note that the inverse map CDF:  $\Omega_{k,0} \mapsto \Omega_{k,t}$  is a  
 1223 *random distribution*  $\omega \mapsto \varphi_{0,t}^{-1}(\omega, \cdot)$ ; formally it lives in the Schwartz space  $\mathcal{S}(\mathbb{R}^d)$  [Friedlander and  
 1224 Joshi, 1998]. We keep the shorthand CDF but interpret it as the measurable family  $\{\varphi_{0,t}^{-1}(\omega, \cdot)\}_\omega$ .

1225 Thus, all merger-time results proved that the deterministic PF-ODE holds *pathwise* for the SDE:  
 1226 for almost every Brownian path, the indicator in (3.1) is well defined. Taking  $\mathbb{E}_\omega$  recovers the  
 1227 deterministic statistics used in **Algorithm 2** and **Algorithm 4**, so no implementation change will be  
 1228 needed.

## 1229 B.9 Extending the framework to certain non-Markovian samplers

1230 For a *non-Markovian* latent chain the marginal  $p_i$  depends on the entire future tail  
 1231  $\{p_{i+1}, p_{i+2}, \dots, p_n\}$ . Hence the pull-back  $\Omega_{k,0} \mapsto \Omega_{k,i}$  is well defined only if every conditional  
 1232 kernel beyond step  $i$  is known. This hurdle disappears when the latent family belongs to a *natural*  
 1233 *exponential family* (NEF).

1234 **Tail-statistic Markovization.** An NEF on  $\mathbb{R}^d$  has densities  $p_\theta(x) = h(x) \exp(\langle \theta, T(x) \rangle - A(\theta))$ ,  
 1235 with sufficient statistic  $T$  and log-partition function  $A$ . For an *independent* sequence  $\{X_i\}_{i=1}^n$  drawn  
 1236 from an NEF, define the *tail statistic*

$$G_i := \sum_{t=i}^n T(X_t), \quad i = 1, \dots, n.$$

1237 The Pitman–Koopman–Darmois theorem gives

1238 **Theorem 14 (Tail-statistic Markov property Pitman, 1936)**

- 1239 (i) If  $\{X_i\}$  are i.i.d. from an NEF, the conditional sequence  $\{\mathcal{L}(X_i \mid G_i)\}_{i=1}^n$  is first-order Markov.  
 1240 (ii) Conversely, if a statistic sequence  $\{G_i\}$  makes  $\{\mathcal{L}(X_i \mid G_i)\}$  Markov for all  $n$ , then the  
 1241 marginals must form an NEF.

1242 Thus the random vector  $G_i$  captures *all* future information relevant at step  $i$ .

1243 **Injecting the tail statistic into fluctuations.** Fix disjoint initial events  $\Omega_{1,0}, \Omega_{2,0}$ . Condition on  
 1244  $G_i = g$  and apply the deterministic PF-ODE of [Appendix B.3](#) inside the fibre  $\{X_i \mid G_i = g\}$ :

$$\Omega_{k,i}(g) := \{x \in \mathbb{R}^d : \varphi_{0 \rightarrow i}^{-1}(g, x) \in \Omega_{k,0}\}, \quad k \in \{1, 2\}.$$

1245 Because the conditioned kernels are Markov ([Theorem 14](#)), this construction mirrors the purely  
 1246 Markovian case. Averaging over the law of  $G_i$  yields the *annealed* cross-fluctuation

$$\widetilde{\mathcal{M}}_\rho^{(n)}(\Omega_{1,i}, \Omega_{2,i}) := \int \mathcal{M}_\rho^{(n)}(\Omega_{1,i}(g), \Omega_{2,i}(g)) d\mathbb{P}_{G_i}(g),$$

1247 so every algebraic identity from [Section 2.2](#) carries over with  $\mathcal{M}_\rho^{(n)}$  replaced by its tail-averaged  
 1248 counterpart. Star-DDPM [Okhotin et al. \[2023\]](#) is a recent work that uses a similar formulation to  
 1249 obtain non-Markovian diffusion generative models.

1250 Thus, whenever a non-Markovian diffusion admits a finite-dimensional *tail statistic*—a property guar-  
 1251 anteed for exponential-family latents—conditioning on that statistic restores the Markov property and  
 1252 lets the fluctuation framework operate unchanged. Identifying broader classes of tail-Markovizable  
 1253 samplers is a promising direction for future work.

## 1254 **B.10 Mixing time of isotropic Gaussians under Brownian diffusion**

1255 We quantify how long the forward VP-SDE ([2.1](#)) needs to *forget* an isotropic sub-Gaussian input and  
 1256 become  $\varepsilon$ -close (in total variation) to its Gaussian limit. Write

$$J(t) = \exp\left(-\frac{1}{2} \int_0^t \beta(s) ds\right),$$

1257 the deterministic attenuation factor from ([2.4](#)).

1258 **Proposition 15 (Mixing time for sub-Gaussian data)** Let  $\mathbf{y} \in \mathbb{R}^d$  have i.i.d. mean-zero, unit-  
 1259 variance components that are  $\sigma^2$ -sub-Gaussian:  $\Pr(|y_i| > t) \leq 2e^{-t^2/2\sigma^2}$ . Evolve  $\mathbf{y}$  with the  
 1260 VP-SDE ([2.1](#)) and denote  $\mathcal{L}(\mathbf{x}_t) = p_t$ . For every  $\varepsilon \in (0, 1)$  the  $\varepsilon$ -mixing time

$$t_{\text{mix}}(\varepsilon) := \inf\{t \geq 0 : d_{\text{TV}}(p_t, \mathcal{N}(0, I_d)) \leq \varepsilon\}$$

1261 satisfies

$$J(t_{\text{mix}}(\varepsilon)) \leq \frac{2}{d} \left(1 + O(\sigma^2 \log \frac{1}{\varepsilon})\right), \quad \text{and} \quad J(t_{\text{mix}}(e^{-1})) = \Theta(d^{-1}).$$

1262 *Proof. Step 1: tails of the input.* Sub-Gaussianity yields  $\mathbb{E}\|\mathbf{y}\|_2^2 = d$  and  $\text{Var}\|\mathbf{y}\|_2^2 \leq Cd$  (for  
 1263  $C = C(\sigma)$ ). Bernstein’s inequality [Vershynin \[2018\]](#) gives  $\|\mathbf{y}\|_2^2 = d \pm O(\sqrt{d})$  w.h.p.

1264 *Step 2: second moment under the SDE.* Conditioned on  $\mathbf{y}$ ,  $\mathbb{E}\|\mathbf{x}_t\|_2^2 = J(t)^2\|\mathbf{y}\|_2^2 + d(1 - J(t)^2)$ , so  
 1265 averaging produces  $\mathbb{E}\|\mathbf{x}_t\|_2^2 = d + J(t)^2 O(\sqrt{d})$ .

1266 *Step 3: bounding  $\chi^2$ .* Pinsker’s inequality [Cover and Thomas \[2006\]](#) gives  $d_{\text{TV}}^2 \leq \frac{1}{2}\chi^2$ . For Gaussians  
 1267 with equal means,  $\chi^2 = (\det \Sigma)^{-1/2} \exp(\frac{1}{2} \text{tr}(I - \Sigma^{-1})) - 1$ . With  $\Sigma = J(t)^2 I_d + (1 - J(t)^2) I_d$ ,

$$d_{\text{TV}}^2(p_t, \mathcal{N}) \leq \frac{1}{2} d \frac{J(t)^4}{1 - J(t)^2} (1 + O(d^{-1/2})).$$

1268 *Step 4: solve for  $J(t)$ .* Setting the rhs to  $\varepsilon^2$  and solving yields the claimed bound.

1269 **Closed form for a linear schedule.** With  $\beta(t) = \beta_0 + (\beta_T - \beta_0)t/T$ ,

$$J(t) = \exp\left(-\frac{1}{2}\beta_0 t - \frac{1}{4}(\beta_T - \beta_0) \frac{t^2}{T}\right).$$

1270 Taking  $\varepsilon = e^{-1}$  in [Theorem 15](#),  $\log J(t_{\text{mix}}) \simeq -\log d + \log 2$ , so  $t_{\text{mix}}$  solves a quadratic. For the  
 1271 DDPM defaults  $(\beta_0, \beta_T, T) = (10^{-4}, 0.02, 1000)$ :

$$t_{\text{mix}} + 0.0995 t_{\text{mix}}^2 = 5000 \log(d/2).$$

Data dimension	Pred. $t_{\text{mix}}/T$	Obs. $i^*/T$
$3 \times 32 \times 32$ (CIFAR-10)	0.602	0.60
$1 \times 28 \times 28$ (MNIST)	0.543	0.60
$4 \times 32 \times 32$ (ImageNet latents)	0.614	0.70

Table 8: Theoretical mixing index vs. empirical convergence index.

Table 8 shows theory versus measured convergence indices; the  $\Theta(d^{-1})$  scaling persists on real data. Thus, it is possible to estimate the mixing time for sub-gaussian data analytically, in fact due to concentration bounds on high-dimensional sub-gaussians Vershynin [2018] it could be shown that the above time manifests physically as a symmetry breaking transition, leading to an estimate consistent with Raya and Ambrogioni [2024] for spherically symmetric distributions and in Biroli et al. [2024] for gaussian mixtures. Interestingly, ImageNet latent representations align closely with these theoretical estimates. We hypothesize that this occurs because the compressed space corresponds to the latent space of a Variational Autoencoder (VAE) trained to approximate a Gaussian distribution, potentially making these latents effectively sub-Gaussian. Verification of this hypothesis and related questions remain future work.

### B.11 Higher-order fluctuations as a proof of concept

To consider tracking  $\mathcal{M}_\rho^n$  beyond  $n = 2$ , we adopt a tractable approximation

- (a) Squash high-dimensional inputs to a single vector whose coordinates are treated as i.i.d. samples.
- (b) Track the scalar moment identity

$$\hat{F}_\rho^{(n)}(\Omega_i)(t) = J(t)^n \hat{F}_\rho^{(n)}(\Omega_i)(0) + (1 - J(t)^n) \hat{F}_\rho^{(n)}(\mathcal{N}),$$

where  $\mathcal{N}$  is the isotropic Gaussian. By Isserlis/Wick [Isserlis, 1918, Wick, 1950],  $\hat{F}_\rho^{(n)}(\mathcal{N})$  is a polynomial in  $\hat{F}_\rho^{(2)}(\mathcal{N})$  and vanishes for odd  $n$ .

Here, the understanding is that  $\mathcal{N}$  is treated effectively as a single "event" similar to Section 4.1. This rule lets us draw *higher-order generative diagrams* such as Figures 1-2 for CIFAR10 and MNIST. Compared with the second-order diagram, high-order curves fan out more widely at early times—evidence that non-linear features dominate, but they collapse sooner, consistent with the rapid  $J(t)^n$  decay.



Figure 1: Fourth order generative diagram for CIFAR10. We show the emergence of classes using fourth-order correlations.

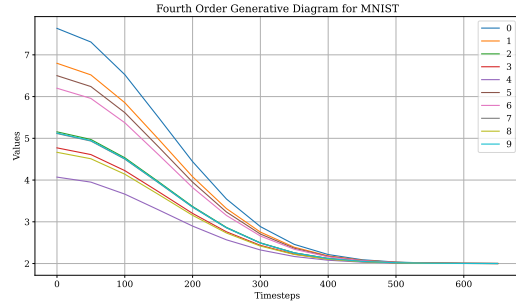


Figure 2: Fourth order generative diagram for MNIST. We show the emergence of classes using fourth-order correlations.

### B.12 Centred kernel alignment (CKA)

Kernel alignment measures how similarly two Gram matrices embed the *same* data. For  $K, L \in \mathbb{R}^{n \times n}$  the uncentred score is the cosine in  $\mathbb{R}^{n^2}$  [Cristianini et al., 2001]:

$$A(K, L) = \frac{\langle K, L \rangle_F}{\|K\|_F \|L\|_F}, \quad \langle K, L \rangle_F := \sum_{ij} K_{ij} L_{ij}.$$



1296 Because  $A$  reacts to mean shifts, Cortes et al. [2012] introduced *centred kernel alignment*

$$A_c(K, L) := A(HKH, HLH), \quad H := I_n - \frac{1}{n} \mathbf{1}\mathbf{1}^\top.$$

1297  $A_c$  is invariant to any feature-space translation and, for linear kernels, to orthogonal mixing and  
1298 isotropic scaling.

1299 **CKA for covariance kernels.** Section 4.2 sets  $\tilde{K} = \Sigma_{i,t}$ ,  $\tilde{L} = \Sigma_{j,t}$ . For symmetric matrices

$$\langle \tilde{K}, \tilde{L} \rangle_F = \sum_{m=1}^d \lambda_{i,t,m} \lambda_{j,t,m}, \quad (B.10)$$

$$\|\tilde{K}\|_F^2 = \sum_{m=1}^d \lambda_{i,t,m}^2, \quad \|\tilde{L}\|_F^2 = \sum_{m=1}^d \lambda_{j,t,m}^2, \quad (B.11)$$

1300 where  $\{\lambda_{i,t,m}\}$  are the eigenvalues of  $\Sigma_{i,t}$  and *mutatis mutandis* for  $\{\lambda_{j,t,m}\}$ . Under the VP flow  
1301 they contract as  $\lambda_{i,t,m} = \lambda_{i,0,m} J(t)^2 + (1 - J(t)^2)$  [Biroli et al., 2024]. The largest eigenvalue  
1302 dominates once  $J(t)^2 \ll \lambda_{i,0,2}/\lambda_{i,0,1}$ , so

$$A_c(\Sigma_{i,t}, \Sigma_{j,t}) \uparrow 1 \iff \lambda_{i,t,1} \simeq \lambda_{j,t,1}.$$

1303 Therefore tracking the top eigenvalues  $\lambda_{i,t,1}, \lambda_{j,t,1}$  gives an efficient proxy for merger of  $\Omega_{i,t}$  and  
1304  $\Omega_{j,t}$  in Section 4.2.

### 1305 B.13 Structural regularity bounds fluctuations

1306 Eq. (4.7) in Section 4.5 bounds the  $L^2$  Fourier distance between the source density  $p_0$  and its  
1307 style-transferred variant  $p^*$ :

$$\|p_0 - p^*\|_{L^2} \leq \delta.$$

1308 Centred moments are weak derivatives of the characteristic function at 0 (Appendix B.2); by Parseval,

$$|\mu_{p_0}^{(k)} - \mu_{p^*}^{(k)}| \leq C_k \delta, \quad k \geq 1,$$

1309 for constants  $C_k$  depending only on the order. Consequently the centred fluctuations satisfy

$$|F_\rho^{(n)}(p_0) - F_\rho^{(n)}(p^*)| \leq C'_n \delta,$$

1310 so the merger schedule computed on  $p_0$  transfers to  $p^*$  with  $O(\delta)$  accuracy. As  $\mathcal{T}_{style}$ , the map  
1311 between the supports of  $p$  and  $p^*$  (Section 4.5) is bijective this extends easily to events  $\Omega \subseteq p_0$  as  
1312 desired. This therefore, justifies the zero-shot style-transfer procedure of Section 4.5.

### 1313 B.14 Phases of diffusion-model dynamics

1314 We distinguish two kinds of time-indexed phase transition for a dynamical system process with an  
1315 evolving phase-space  $\Phi_t$  indexed by time:

1316 (1) **Thermodynamic.** A discontinuity in the  $n$ -th classical derivative  $\Phi^{(n)}$  at some  $t_0$  ( $\Phi \in C^n$   
1317 assumed). Such transitions are *exclusive* in the sense that atmost *one* transition is possible with  
1318 respect to the entire phase space at a time  $t$ .

1319 (2) **Lattice.** For a threshold  $\tau > 0$  if  $\exists \varepsilon > 0$  such that for the  $n$ -th finite differences,  $\text{LD}_\tau^{(n)}, \text{RD}_\tau^{(n)}$ ,  
1320 the below equation holds true then  $\Phi$  undergoes a *lattice* transition at  $t_0$  of order  $n$ .

$$\|\text{LD}_\tau^{(n)} \Phi(t_0) - \text{RD}_\tau^{(n)} \Phi(t_0)\| \geq \varepsilon.$$

1321 Lattice transitions include thermodynamic ones in the limit  $\tau \rightarrow 0$  but are generally *non-exclusive*  
1322 since the threshold  $\tau$  is independent of  $n$ .

1323 **Theorem 16**  $\widetilde{\mathcal{M}}_\rho^{(n)}$  as defined in Eq. (3.1) undergoes a lattice transition at the merger time



1324 *Proof.* Consider two initially disjoint events  $\Omega_{i,0}, \Omega_{j,0} \subset \Omega_0$ . Let  $t$  denote the merger time as given  
 1325 by Eq. (3.1). From the topological equivalence proved in Theorem 10, there exists a constant  $\vartheta > 0$   
 1326 such that

$$|\widetilde{\mathcal{M}}_\rho^{(n)}(\Omega_{i,t}, \Omega_{j,t}) - 1| \leq \vartheta.$$

1327 By selecting a suitable  $\epsilon > 0$  in the within-event fluctuation metric  $d(\cdot, \cdot)$ , we can ensure  $\vartheta < \frac{1}{2}$ . Since  
 1328 by definition  $\widetilde{\mathcal{M}}_\rho^{(n)}(\cdot, \cdot)$  takes values in the interval  $[0, 1]$ , it follows that  $\widetilde{\mathcal{M}}_\rho^{(n)}(\Omega_{i,t}, \Omega_{j,t}) \geq 1 - \vartheta$ .  
 1329 Fix a small radius  $r > 0$  with  $r \ll 1$ , and sample  $\tau$  uniformly from the ball

$$\tau \sim B_{\frac{r}{2}}(r),$$

1330 where  $B_a(b)$  denotes the ball centered at  $b$  with radius  $a$ . Due to the hypercontractivity of the  
 1331 underlying Brownian motion, the distance between initially disjoint events  $\Omega_{i,t}$  and  $\Omega_{j,t}$  contracts  
 1332 over time. Hence  $\widetilde{\mathcal{M}}_\rho^{(n)}(\Omega_{i,t}, \Omega_{j,t})$  is monotonically increasing in  $t$ . This monotonicity implies that  
 1333 the  $n$ -th order finite differences satisfy

$$\text{RD}_\tau^{(n)}(\widetilde{\mathcal{M}}_\rho^{(n)}) \leq \frac{\vartheta}{\tau}, \quad \text{and} \quad \text{LD}_\tau^{(n)}(\widetilde{\mathcal{M}}_\rho^{(n)}) \geq \frac{1 - \vartheta}{\tau}.$$

1334 Applying the reverse triangle inequality, we get

$$\|\text{LD}_\tau^{(n)}(\widetilde{\mathcal{M}}_\rho^{(n)}) - \text{RD}_\tau^{(n)}(\widetilde{\mathcal{M}}_\rho^{(n)})\| \geq \frac{1 - 2\vartheta}{\tau} > 0,$$

1335 which establishes the existence of a lattice transition at the merger time. Notice that as  $\tau, \vartheta$  are  
 1336 arbitrary, this transition is *not* thermodynamic.

1337 **Thermodynamic phases from prior works.** For class-conditioned VP diffusion, Biroli et al.  
 1338 [2024] proved two thermodynamic boundaries  $t_{u \rightarrow s}$  (unbiased  $\rightarrow$  speciation) and  $t_{s \rightarrow c}$  (speciation  
 1339  $\rightarrow$  condensation):

$$\text{unbiased } [0, t_{u \rightarrow s}) \subset \text{speciation } (t_{u \rightarrow s}, t_{s \rightarrow c}) \subset \text{condensation } (t_{s \rightarrow c}, T].$$

1340 **Relation of class conditional lattice mergers to thermodynamic phases.** For two classes  $k \neq \ell$   
 1341 define the centred cross-fluctuation  $\mathcal{M}_{k\ell}(t)$  ((4.5) in Section 4.2). Its  $\varepsilon$ -merger time is

$$t_{k\ell}^{\text{lat}}(\varepsilon) := \inf\{t \geq 0 : \mathcal{M}_{k\ell}(t) \geq 1 - \varepsilon\}, \quad \varepsilon \in (0, 1).$$

1342 **Lemma 17 (Merger times lie inside the speciation phase)** For all  $k \neq \ell$  and  $\varepsilon \in (0, 1)$ ,

$$t_{u \rightarrow s} < t_{k\ell}^{\text{lat}}(\varepsilon) \leq t_{s \rightarrow c}.$$

1343 *Proof. Unbiased phase.* If  $t < t_{u \rightarrow s}$  then  $p_{k,t} = p_{l,t}$ , so  $\mathcal{M}_{k\ell}(t) = 1$ ; no upward crossing can occur.

1344 *Condensation phase.* For  $t > t_{s \rightarrow c}$  each covariance (4.3) satisfies  $\lambda_{\max}(\Sigma_{k,t}) \leq e^{-c(t - t_{s \rightarrow c})}$  [Biroli  
 1345 et al., 2024, Prop. 4]. Using (B.11),  $1 - \mathcal{M}_{k\ell}(t) = O(e^{-c(t - t_{s \rightarrow c})})$ ; hence  $\mathcal{M}_{k\ell}(t) \geq 1 - \varepsilon$  for all  
 1346 large  $t$ . No new crossing can start after  $t_{s \rightarrow c}$ .

1347 *Speciation phase.* Because a crossing cannot start before  $t_{u \rightarrow s}$  or after  $t_{s \rightarrow c}$ , any merger time must  
 1348 lie in  $(t_{u \rightarrow s}, t_{s \rightarrow c}]$ .

1349 **Lattice transitions in other contexts for diffusion models.** Note also that the ELBO loss regime  
 1350 observed in the training of diffusion models (Appendix B.4) can also be understood in the context of  
 1351 lattice and thermodynamic transitions,

1352 The curve  $\varepsilon \mapsto \|p_i - \hat{p}_i\|_{\text{TV}}$  exhibits two critical errors:

- 1353 1. a *thermodynamic* transition at  $\varepsilon = 0$ , where the learned and empirical processes coincide,  
 1354 but diversity vanishes;
- 1355 2. a *lattice* transition at some finite  $\varepsilon_{\text{lat}} > 0$ , below which the two kernels are indistinguishable  
 1356 at machine precision.

Our experiments operate in the slab  $\varepsilon_{\text{lat}} < \varepsilon_* \ll 1$ : the model is close enough to justify forward-process analysis ([Theorem 9](#)) yet far enough from collapse to yield novel, high-quality samples. Thus, the simplified ELBO acts as a quantitative knob—through  $\varepsilon_*$ —that tunes the distance between the trained sampler and the empirical reverse kernel; in modern diffusion models this regime ensures that the dynamics can be understood through the empirical forward trajectory, which faithfully mirrors the merger dynamics. Theoretically, the use of fluctuation theory to detect such behavior would be identical to our use case of understanding sampling; however, we leave such investigations to future work.

Thus, lattice merger times give a fine-grained view inside the speciation window, precisely where class structure exists, but pre-condensation collapse has not yet begun. They are invisible to purely derivative-based (thermodynamic) criteria, underscoring the value of lattice diagnostics for finite-precision generative modelling.

**Remark 18** *The existence of lattice transitions stems from the fundamental fact that no computation can achieve infinite precision, a concept rooted in quantum mechanics [Vopson \[2025\]](#). Although diffusion processes are classically continuous, their simulation in physical systems inherently introduces discretization, leading to lattice transitions.*

### 1373 **B.15 A single forward Monte-Carlo sweep yields unbiased estimates of all terms in $\widetilde{\mathcal{M}}_\rho^{(n)}$**

**Theorem 19 (One-sweep unbiasedness)** *Let  $\Omega_{1,0}, \Omega_{2,0} \subseteq \Omega$  be disjoint with probabilities  $p_1, p_2 > 0$ . Simulate once  $N$  i.i.d. forward trajectories  $\{\mathbf{x}_t^{(i)}\}_{t=0}^T$ ,  $i = 1, \dots, N$  from the VP process. Define  $Z_k^{(i)} := \mathbb{1}_{\Omega_{k,0}}(\mathbf{x}_0^{(i)})$ ,  $k \in \{1, 2\}$ , and  $f_t^{(n)}(\mathbf{x}) := \|\rho(\mathbf{x}) - \mathbb{E}[\rho]\|^n$ . For each  $t$  set*

$$\overline{\mathcal{F}}_\rho^{(n)}(\Omega_{k,t}) := \frac{1}{Np_k} \sum_{i=1}^N Z_k^{(i)} f_t^{(n)}(\mathbf{x}_t^{(i)}), \quad k \in \{1, 2\}, \quad (\text{B.12})$$

$$\overline{\mathcal{G}}_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t}) := \frac{1}{N(N-1)p_1p_2} \sum_{\substack{i,j \leq N \\ i \neq j}} Z_1^{(i)} Z_2^{(j)} f_t^{(n)}(\mathbf{x}_t^{(i)}) f_t^{(n)}(\mathbf{x}_t^{(j)}). \quad (\text{B.13})$$

Then  $\mathbb{E}[\overline{\mathcal{F}}_\rho^{(n)}(\Omega_{k,t})] = \widehat{F}_\rho^{(n)}(\Omega_{k,t})$ ,  $\mathbb{E}[\overline{\mathcal{G}}_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t})] = G_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t})$ . Hence, the plug-in ratio

$$\overline{\mathcal{M}}_\rho^{(n)}(t) := \frac{\overline{\mathcal{G}}_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t})}{\sqrt{\overline{\mathcal{F}}_\rho^{(2n)}(\Omega_{1,t}) \overline{\mathcal{F}}_\rho^{(2n)}(\Omega_{2,t})}}$$

satisfies

$$\mathbb{E}[\overline{\mathcal{M}}_\rho^{(n)}(t)] = \mathcal{M}_\rho^{(n)}(\Omega_{1,t}, \Omega_{2,t}) + O(N^{-1}),$$

i.e. it is unbiased up to the usual  $O(N^{-1})$  Monte-Carlo error (delta method).

*Proof.* All randomness is with respect to an appropriate product measure  $\mathbb{P}^{\otimes N}$  that generates the  $N$  i.i.d. forward trajectories<sup>7</sup>. We write  $\mathbb{E}_{\mathbb{P}}$  for expectation under that measure.

*Step 1: unbiasedness of  $\overline{\mathcal{F}}_\rho^{(2n)}(\Omega_{k,t})$ .* Fix  $k \in \{1, 2\}$  and a trajectory index  $i$ . Let  $A_k := \Omega_{k,0}$  for brevity, and recall  $Z_k^{(i)} = \mathbb{1}_{A_k}(x_0^{(i)})$ . Conditional on  $x_0^{(i)} = x_0$ , the time- $t$  state  $x_t^{(i)}$  has density

<sup>7</sup>For the PF-ODE or Liouville equation case, the measure is a standard product measure over  $d$ -dimensional Gaussians for each time step up to  $t$ , covering  $N$  trajectories. Each state is independent of the others, making this a product measure over independent distributions.

For the SDE case, the measure is the  $N$ -fold product of the  $d$ -dimensional **Wiener measure**  $W^d$  [Karatzas and Shreve \[1991\]](#), defined over all continuous paths in  $\mathbb{R}^d$ . This captures the full dependence of each path across time, making it fundamentally different from the independent Gaussian product in the PF-ODE

1385  $p_t(\cdot \mid x_0)$  (the transition kernel of the forward process). Therefore

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[Z_k^{(i)} f_t^{(n)}(x_t^{(i)})] &= \int_{A_k} p_0(x_0) \int_{\mathbb{R}^d} f_t^{(n)}(x) p_t(x \mid x_0) dx dx_0 \\ &= p_k \frac{1}{p_k} \int_{A_k} \mathbb{E}_{\mathbb{P}}[f_t^{(n)}(x_t) \mid x_0] p_0(x_0) dx_0 \\ &= p_k \widehat{F}_{\rho}^{(n)}(\Omega_{k,t}),\end{aligned}$$

1386 where the last equality is the definition of the centred  $n$ -th moment restricted to  $\Omega_{k,t}$ . Summing over  
1387  $i = 1, \dots, N$  and dividing by  $Np_k$  gives  $\mathbb{E}_{\mathbb{P}}[\widehat{\mathcal{F}}_{\rho}^{(2n)}(\Omega_{k,t})] = \widehat{F}_{\rho}^{(n)}(\Omega_{k,t})$ .

1388 *Step 2: unbiasedness of  $\overline{\mathcal{G}}_{\rho}^{(n)}(\Omega_{1,t}, \Omega_{2,t})$ .* For  $i \neq j$ , trajectories  $(x_{\bullet}^{(i)}, x_{\bullet}^{(j)})$  are independent, so  
1389  $\mathbb{E}_{\mathbb{P}}[Z_1^{(i)} Z_2^{(j)}] = \mathbb{E}_{\mathbb{P}}[Z_1^{(i)}] \mathbb{E}_{\mathbb{P}}[Z_2^{(j)}] = p_1 p_2$ . Hence

$$\begin{aligned}\mathbb{E}_{\mathbb{P}}[\overline{\mathcal{G}}_{\rho}^{(n)}] &= \frac{1}{N(N-1)p_1 p_2} \sum_{i \neq j} \mathbb{E}_{\mathbb{P}}[Z_1^{(i)} f_t^{(n)}(x_t^{(i)})] \mathbb{E}_{\mathbb{P}}[Z_2^{(j)} f_t^{(n)}(x_t^{(j)})] \\ &= \left( \frac{N(N-1)}{N(N-1)} \right) \frac{\widehat{F}_{\rho}^{(n)}(\Omega_{1,t}) \widehat{F}_{\rho}^{(n)}(\Omega_{2,t})}{p_1 p_2} = G_{\rho}^{(n)}(\Omega_{1,t}, \Omega_{2,t}),\end{aligned}$$

1390 where  $G_{\rho}^{(n)}$  is the true cross-moment.

1391 *Step 3: bias of the ratio.* Let  $U_N := \overline{\mathcal{G}}_{\rho}^{(n)}(\Omega_{1,t}, \Omega_{2,t})$ ,  $V_N^{(k)} := \overline{\mathcal{F}}_{\rho}^{(2n)}(\Omega_{k,t})$ ,  $k = 1, 2$ , and  
1392  $u := \mathbb{E}_{\mathbb{P}}[U_N]$ ,  $v_k := \mathbb{E}_{\mathbb{P}}[V_N^{(k)}]$ . By Steps 1–2,  $u$  and  $v_k$  are finite and strictly positive (constants  
1393 independent of  $N$ ). Because  $\{U_N, V_N^{(1)}, V_N^{(2)}\}$  are averages of  $O(N)$  i.i.d. random variables with  
1394 finite variance,  $\text{Var}(U_N) = O(N^{-1})$ ,  $\text{Var}(V_N^{(k)}) = O(N^{-1})$ . Consider the smooth mapping  
1395  $g(x, y_1, y_2) = x/\sqrt{y_1 y_2}$  on a neighbourhood of  $(u, v_1, v_2)$ . A multivariate second-order Taylor  
1396 expansion about  $(u, v_1, v_2)$  gives

$$g(U_N, V_N^{(1)}, V_N^{(2)}) = g(u, v_1, v_2) + \sum_r g'_r \Delta_r + \frac{1}{2} \sum_{r,s} g''_{rs} \Delta_r \Delta_s,$$

1397 where  $\Delta_r$  are the centred deviations and  $g'_r, g''_{rs}$  are first and second partials evaluated at the mean  
1398 point. Taking expectations, the linear term vanishes (as  $\overline{\mathcal{F}}_{\rho}^{(2n)}(\Omega_{k,t})$ , are unbiased), and  $\mathbb{E}[\Delta_r \Delta_s] =$   
1399  $O(N^{-1})$ , whence

$$\mathbb{E}_{\mathbb{P}}[\overline{\mathcal{M}}_{\rho}^{(n)}(t)] = g(u, v_1, v_2) + O(N^{-1}) = \mathcal{M}_{\rho}^{(n)}(\Omega_{1,t}, \Omega_{2,t}) + O(N^{-1}).$$

1400

1401 **Remark 20** All sums in (B.12)–(B.13) use the same trajectory set; no reverse-process simulation  
1402 or second sweep is required. In principle, one stores the  $N(T+1)$  feature vectors once and re-uses  
1403 them for every time index  $t$ .

## 1404 C Experimental details and further results

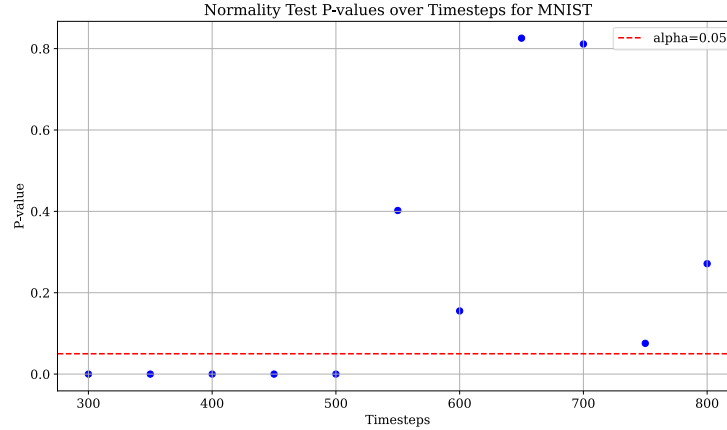
### 1405 C.1 Compute and Reproducibility

1406 We conducted all experiments using a single Nvidia A100 GPU and provided sample code for  
1407 reproducibility. Our implementation builds on open-source code from Hugging Face (diffusers  
1408 library) and publicly available code from Kynkäänniemi et al. [2024], Li et al. [2023], Peebles and  
1409 Xie [2022]. Our method is plug-and-play, requiring simple hyperparameter adjustments for these  
1410 techniques without any major code modifications. For zero-shot style transfer (Section 4.5), we used  
1411 the Img2Img transfer pipeline in diffusers (Meng et al. [2021], Rombach et al. [2022]), fixing the  
1412 VP/DDPM schedule and adjusting the strength parameter.

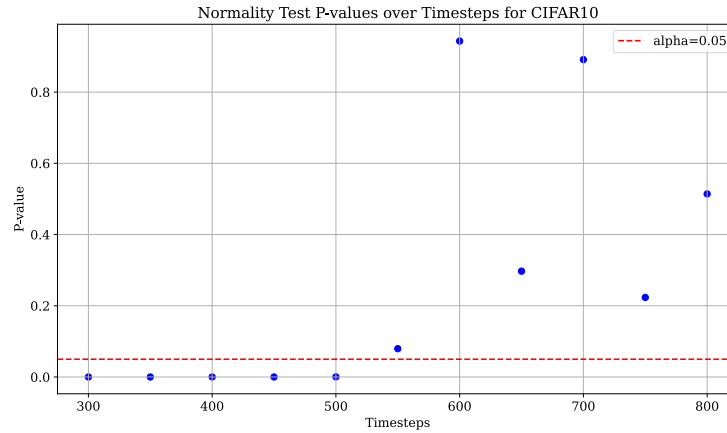
1413 Baseline experiments using grid search were computationally intensive, typically requiring 24–48  
1414 hours of GPU time. Fluctuation computations (Section 3) were primarily constrained by covariance  
1415 matrix calculations and eigendecompositions. These can be efficiently optimized using multi-  
1416 processing, though we used a single-process approach in this work. Our method is directly compatible  
1417 with any standard implementation of the baselines. Preliminary small-scale experiments verified our  
1418 theoretical framework but were excluded from the final paper.

### 1419 C.2 Convergence of data

We present images visualizing the normality tests as stated in Section 4.1

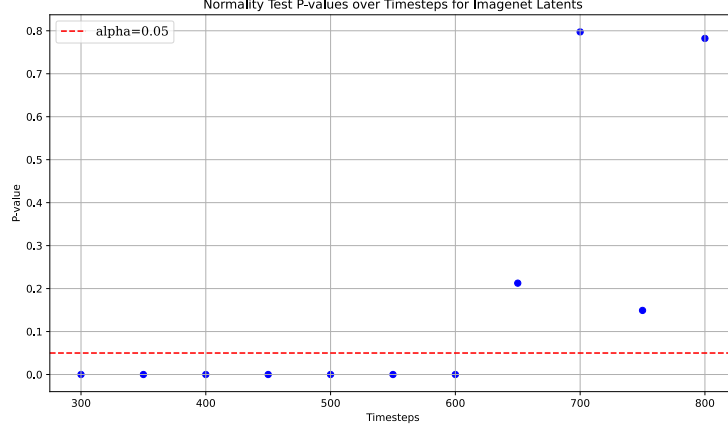


a Normality test p-values over time for MNIST



b Normality test p-values over time for CIFAR10

1420



a Normality test p-values over time for Imagenet

Figure 4: Normality test p-values for the DDPM schedule

### 1421 C.3 Class-conditional generation

1422 We compare two guidance schedules:

- 1423 1. a *grid-search baseline* that follows Interval Guidance (IG) [Kynkäänniemi et al., 2024] with
- 1424 a *single dataset-level interval* found by brute force<sup>8</sup>;
- 1425 2. our *merger-aware schedule*, in which each class  $k$  receives its own window  $(t_{\text{start},k}, t_{\text{end},k})$
- 1426 derived from fluctuation theory (Sections 4.1 and 4.2).

1427 **Interval Guidance baseline.** Let  $w > 0$  be the classifier-free guidance Ho and Salimans [2022b]  
 1428 (CFG) weight, and let  $T$  be the full diffusion horizon. During reverse sampling we switch CFG on  
 1429 only for  $t \in (t_{\text{end},c}, t_{\text{start},c})$ :

---

**Algorithm 2** Interval Guidance (class  $c$ )

---

**Require:** latent  $x_T \sim \mathcal{N}(0, I)$ , CFG weight  $w$

```

1: for  $t = T - 1, \dots, 0$  do
2:   if  $t_{\text{end},c} < t < t_{\text{start},c}$  then
3:      $x_t \leftarrow \text{CFG}_w(x_{t+1}, c)$ 
4:   else
5:      $x_t \leftarrow \text{CFG}_0(x_{t+1}, c)$ 
6:   end if
7: end for
8: return  $x_0$ 
```

---

1430 **Hyper-parameters and Grid-search baseline.** Following Peebles and Xie [2022] we set  $w = 1.5$   
 1431 for DiT-XL/2. Stable Diffusion requires stronger guidance; we use a fixed  $w \in [3.5, 4.5]$  per dataset.  
 1432 For every dataset we sweep

$$t_{\text{end},c} \in \{0.1T, 0.2T, \dots, 0.8T\}, \quad t_{\text{start},c} \in \{0.2T, \dots, T\},$$

1433 under the constraint  $t_{\text{start},c} > t_{\text{end},c}$ , yielding 44 admissible pairs. On ImageNet the best pair is  
 1434  $(0.8T, 0.2T)$  (lowest FID); MNIST and CIFAR-10 select  $(0.6T, 0.1T)$  and  $(0.7T, 0.1T)$ , respectively.

---

<sup>8</sup>In Kynkäänniemi et al. [2024], it is argued that a class-level or even a sample-level search is preferable. However, both of these settings require at least  $10^3 - 10^6 \times$  the baseline compute, making them infeasible for us. Our primary objective is to demonstrate that leveraging finer hierarchical levels can compensate for compute limitations. Specifically, we reason that for class-level brute-force search, it is more practical to consider the transitions of a fine-grained hierarchy derived from the representations of a semi/self-supervised learning model Chen et al. [2020], Grill et al. [2020], He et al. [2021]. A theoretical backing for the same is found in Theorem 11. We leave such an extension to future work. We note that asymptotically, our method converges to the same output as a per-sample-level grid search, as ultimately each data sample can be treated as a distinct singleton event  $\Omega_{k,0}$ .

For Stable Diffusion we replace FID by CLIP similarity and obtain  $(0.8T, 0.1T)$  for both ImageNet and Oxford-IIIT Pet. One exhaustive sweep on DiT-XL/2 costs  $4100 \text{ GFLOPs} \times 50\,000 \text{ samples} \times 44 \text{ configs} \approx 9.0 \text{ PFLOPs}$ ; five repeats per pair multiply the cost five-fold. Results for Imagenet, MNIST and CIFAR are in Table 2 in the main paper while that for Imagenet and Oxford-IIITPets using Stable Diffusion is in Table 9. Note that for the case of Imagenet, identical intervals are used for both settings as the empirical forward process trajectory is independent of the model choice.

**Merger-aware schedule (ours).** For each class  $k$ ,  $t_{\text{start},k} = i^*$ ,  $t_{\text{end},k} = t_{\text{merge},k}$ , where  $i^*$  is the global convergence index (Section 4.1) and  $t_{\text{merge},k}$  is the first  $t$  at which  $M_p^{(2)}(\Omega_{k,t}, \Omega_{\ell,t}) = 1$  for some  $\ell \neq k$  (Section 4.2). No search is required; windows differ automatically across classes.

Model/Dataset	CLIP Similarity ( $\uparrow$ )	Precision ( $\uparrow$ )	Recall ( $\uparrow$ )	Density ( $\uparrow$ )	Coverage ( $\uparrow$ )
SD (Imagenet, IG baseline)	$0.26 \pm 0.03$	$0.75 \pm 0.04$	$0.18 \pm 0.02$	$0.80 \pm 0.05$	$0.30 \pm 0.03$
SD (Imagenet, IG Ours)	<b><math>0.31 \pm 0.02</math></b>	<b><math>0.78 \pm 0.02</math></b>	<b><math>0.23 \pm 0.01</math></b>	<b><math>0.88 \pm 0.03</math></b>	<b><math>0.34 \pm 0.02</math></b>
SD (OxfordIIITPet, IG baseline)	$0.28 \pm 0.02$	$0.79 \pm 0.03$	$0.21 \pm 0.03$	$0.84 \pm 0.06$	$0.33 \pm 0.05$
SD (OxfordIIITPet, IG Ours)	<b><math>0.34 \pm 0.03</math></b>	<b><math>0.81 \pm 0.01</math></b>	<b><math>0.26 \pm 0.04</math></b>	<b><math>0.89 \pm 0.01</math></b>	<b><math>0.36 \pm 0.02</math></b>

Table 9: Class conditional generation using Stable Diffusion

**Generative diagrams.** Figure 10 plots the leading eigenvalues  $\lambda_{\max}(\Sigma_{k,t})$  of the class covariances  $\Sigma_{k,t}$  for ImageNet, CIFAR-10, and MNIST, highlighting merger points (proofs in Appendix B.12). Additional zoom-ins for ImageNet appear in Figure 9. For long-tail datasets used in Section 4.3, analogous diagrams are given in Figures 11a and 11c. We also plot the subplots for the 10 classes for Imagenet and OxfordIIITPet, having the greatest magnitude of principal eigenvalues in Figure 6.



Figure 5: Stable-Diffusion samples for four Oxford-IIIT-Pet classes, generated with naïve interval guidance (top of each column) versus our method (bottom).

Figure 8 compares ImageNet samples from our merger-aware IG with the grid-search baseline, using a guidance weight of 4.5 for visual clarity; Figure 5 does the same for Oxford-IIIT Pet under Stable Diffusion. Our method yields crisper details and fewer artefacts—despite eliminating  $\approx 9 \text{ PFLOPs}$



1452 of search for Imagenet. Thus, class-wise guidance windows obtained from cross-fluctuation mergers  
 1453 can match or exceed the quality of an *exhaustive* dataset-level search, while slashing computational  
 1454 cost by orders of magnitude.

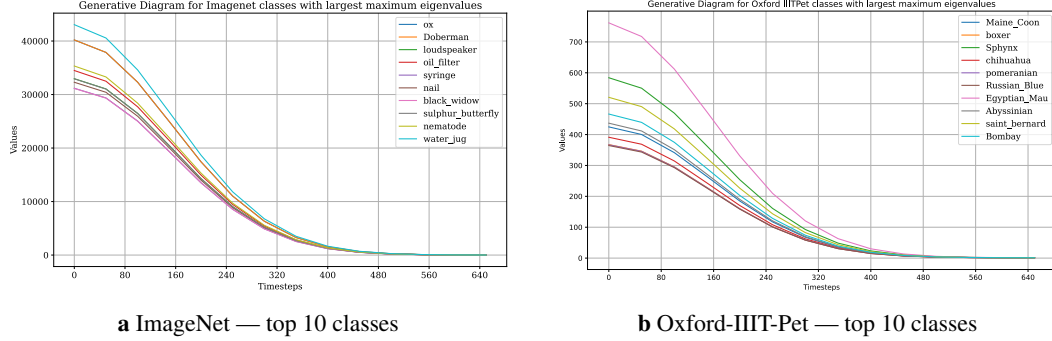


Figure 6: Merger transitions are upper-bounded by those of the ten classes with the largest principal-eigenvalue magnitudes in their covariance matrices.

$\varepsilon$	Merge prob. ( $t = 0$ )	FID ( $\downarrow$ )
200	0.027	$3.06 \pm 0.19$
100	0.013	<b><math>2.86 \pm 0.15</math></b>
80	0.010	$2.92 \pm 0.17$
67	0.009	$2.90 \pm 0.11$
20	0.002	$2.88 \pm 0.14$

Table 10: Effect of the MAE threshold  $\varepsilon$  (ImageNet).

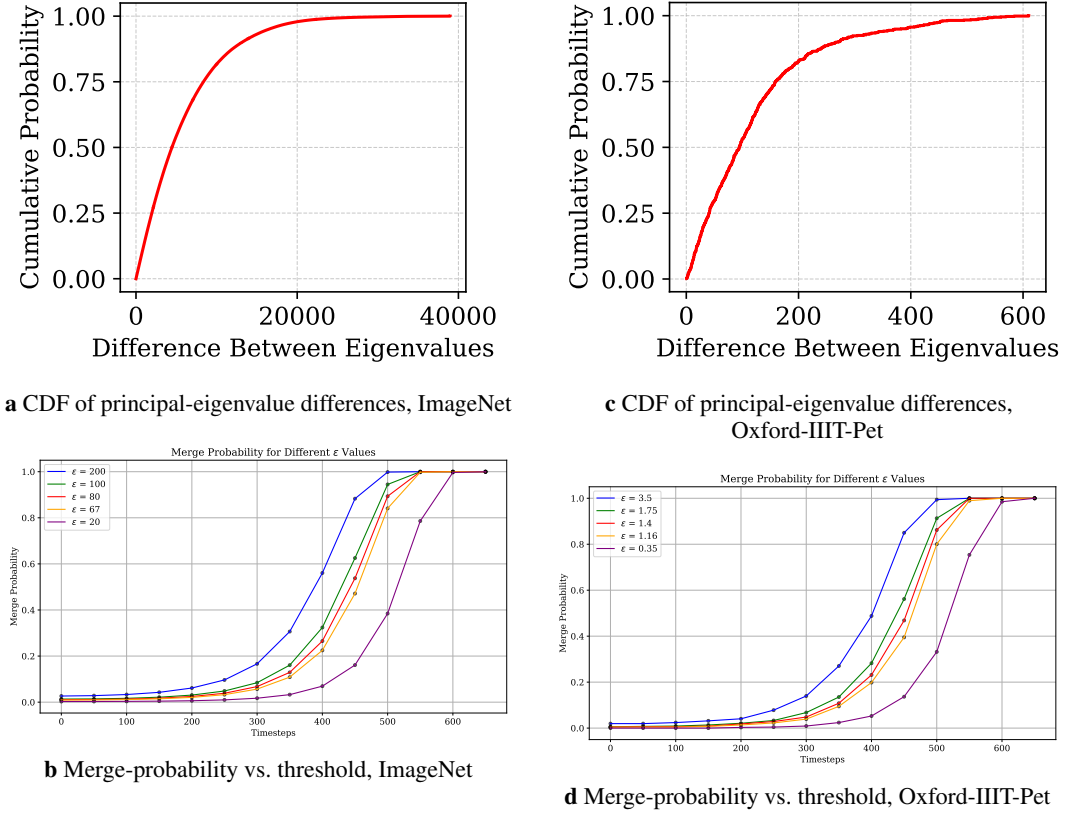


Figure 7: Eigenvalue-difference distributions and their associated merge-probability curves for ImageNet and Oxford-IIIT-Pet.

1455 To understand the sensitivity of the parameter  $\varepsilon$ , we plot the distribution of the difference in eigen-  
 1456 values (CDF) and evolution of merge probabilities for different choices of  $\varepsilon$  (Figure 7). Our default  
 1457 choice  $\varepsilon = 100$  has a merge probability  $\approx 0.01$  at  $t = 0$ . We show corresponding FIDs in Table 10.



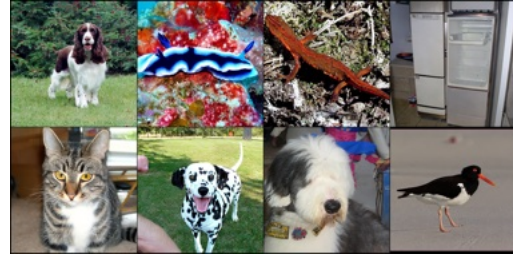
**a** Naive interval guidance for Imagenet



**b** Our optimized interval guidance for Imagenet



**c** Naive interval guidance for Imagenet



**d** Our optimized interval guidance for Imagenet



**e** Naive interval guidance for Imagenet



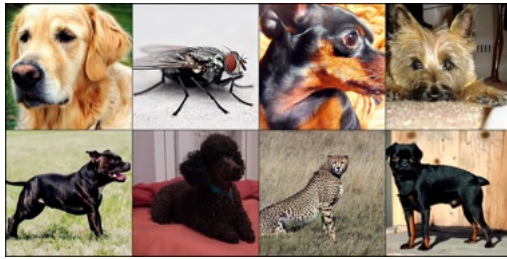
**f** Our optimized interval guidance for Imagenet



**g** Naive interval guidance for Imagenet



**h** Our optimized interval guidance for Imagenet



**i** Naive interval guidance for Imagenet



**j** Our optimized interval guidance for Imagenet

Figure 8: Visual Comparison of guidance for the Imagenet dataset [Deng et al., 2009, Ryu, 2024]. All samples were originally generated in  $512 \times 512$  resolution.



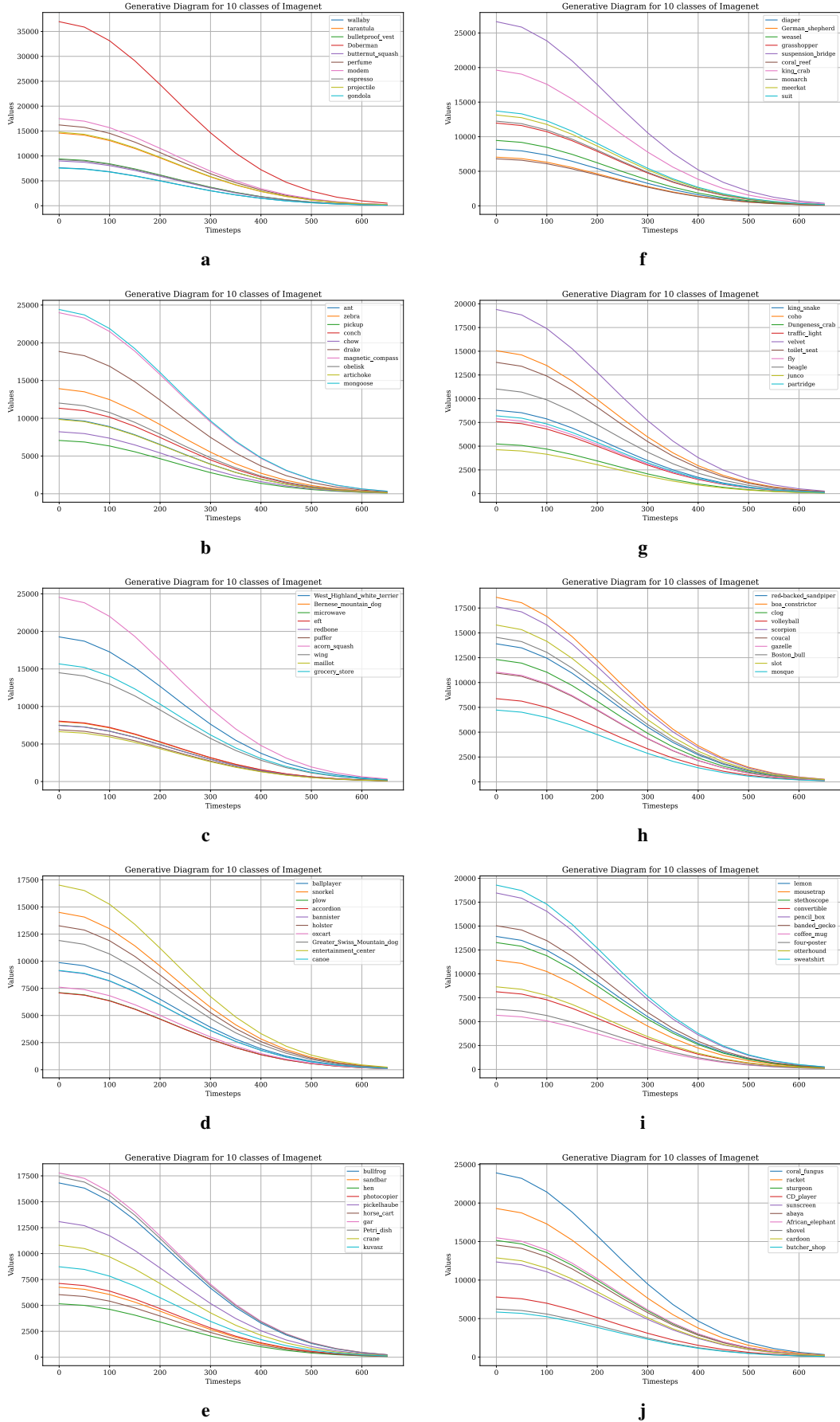
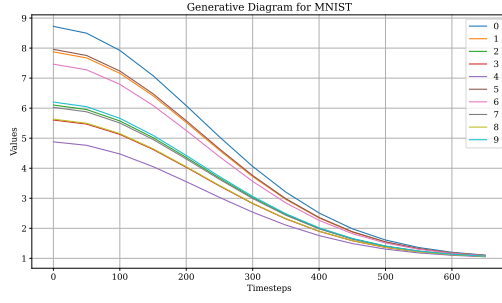
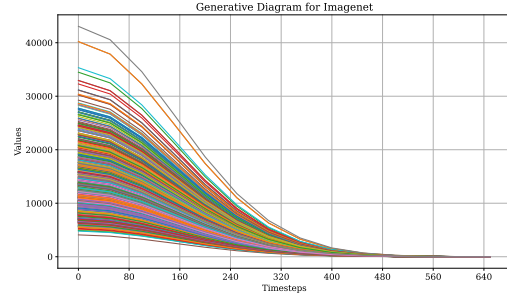


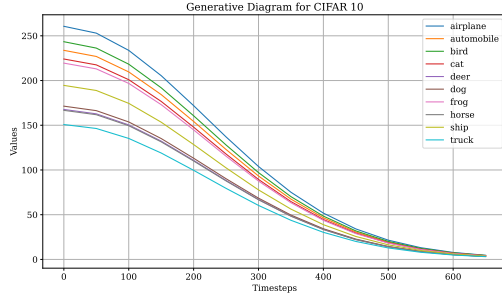
Figure 9: Merger-transition subplots for ImageNet (ten classes shown in two parallel columns).



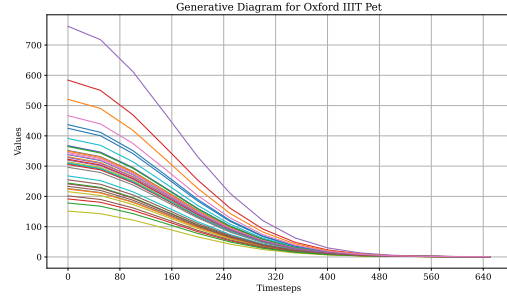
**a** Generative diagram of MNIST



**c** Generative diagram of ImageNet

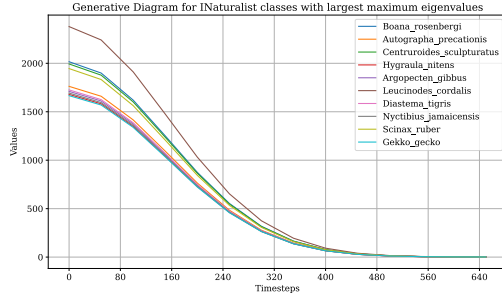


**b** Generative diagram of CIFAR-10

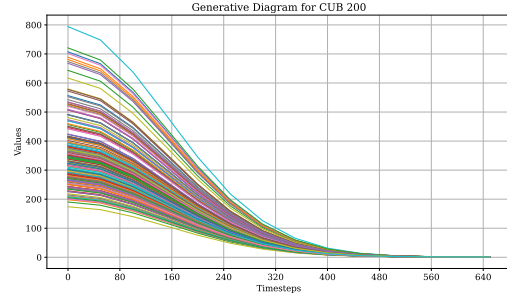


**d** Generative diagram of Oxford-IIIT Pet

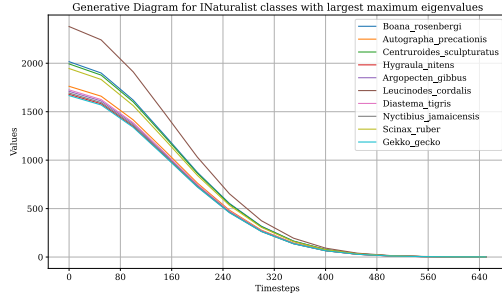
Figure 10: Merger-transition measure obtained from the intersection time of the principal eigenvalues of the class-covariance matrices.



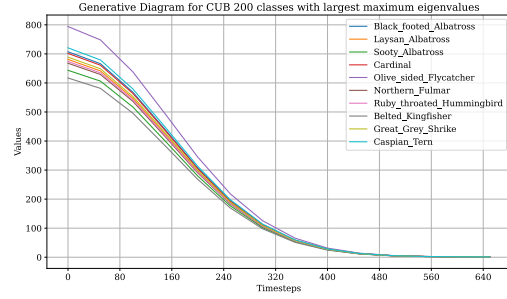
**a** iNaturalist — all classes



**c** CUB-200 — all classes



**b** iNaturalist — top 10 classes



**d** CUB-200 — top 10 classes

Figure 11: Merger-transition measures obtained from the intersection times of the principal eigenvalues for iNaturalist and CUB-200. The bottom row shows that the classes with the ten largest eigenvalue magnitudes upper-bound all other transitions.

## 1458 C.4 Rare Class Generation

1459 **Interpolation-based interval guidance.** In Section 4.3 we observed that augmenting Algorithm 2  
 1460 with an *interpolation correction*—akin to ILVR [Choi et al., 2021]—gives the best fidelity for CUB-  
 1461 200 and iNaturalist-2019. The full procedure is listed in Algorithm 3; it differs from standard Interval  
 1462 Guidance only in lines 5–7.

---

### Algorithm 3 Interpolation-based interval guidance (class $c$ )

---

**Require:** •  $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  *latent to be denoised*  
 •  $\hat{x}_0 \sim p_0$  with  $\text{shape}(x_T) = \text{shape}(\hat{x}_0)$  *class- $c$  exemplar*  
 • guidance weights  $\text{CFG}_w, \text{CFG}_0$   
 • interpolation schedule  $\eta = \{\eta_t\}_{t=0}^{T-1}$ , where  $0 \leq \eta_t \leq 1$   
 1: **for**  $t = T - 1, \dots, 0$  **do**  
 2:   **if**  $t_{\text{start},c} < t < t_{\text{end},c}$  **then**  $\triangleright$  guidance window  
 3:      $x_t \leftarrow \text{CFG}_w(x_{t+1}, c)$   
 4:      $\hat{x}_t \leftarrow \text{FWD}(\hat{x}_0, t)$   
 5:      $x_t \leftarrow \eta_t x_t + (1 - \eta_t) \hat{x}_t$   
 6:   **else**  
 7:      $x_t \leftarrow \text{CFG}_0(x_{t+1}, c)$   
 8:   **end if**  
 9: **end for**  
 10: **return**  $x_0$

---

1463 Note that in Algorithm 3,  $\text{CFG}_w$  applies classifier-free guidance with strength  $w$ ;  $\text{CFG}_0$  disables  
 1464 conditioning.  $\text{FWD}(\hat{x}_0, t)$  generates the *forward-noised* version of the exemplar at step  $t$ . The convex  
 1465 update in line 5 nudges the current latent towards the exemplar’s trajectory, counteracting class drift.

1466 **Choosing the interpolation schedule  $\eta$ .** Because guidance corrections are most valuable late in  
 1467 the reverse chain, we derive  $\eta_t$  from the noise schedule  $\beta_t$ :  $\eta_t = s(\beta_t / \max_{u < T} \beta_u)$ ,  $0 < s \leq 1$ .  
 1468 The scale factor  $s$  is found by a binary search over  $[10^{-4}, 10^{-2}]$ : larger values degrade sharpness,  
 1469 while smaller ones have negligible impact.

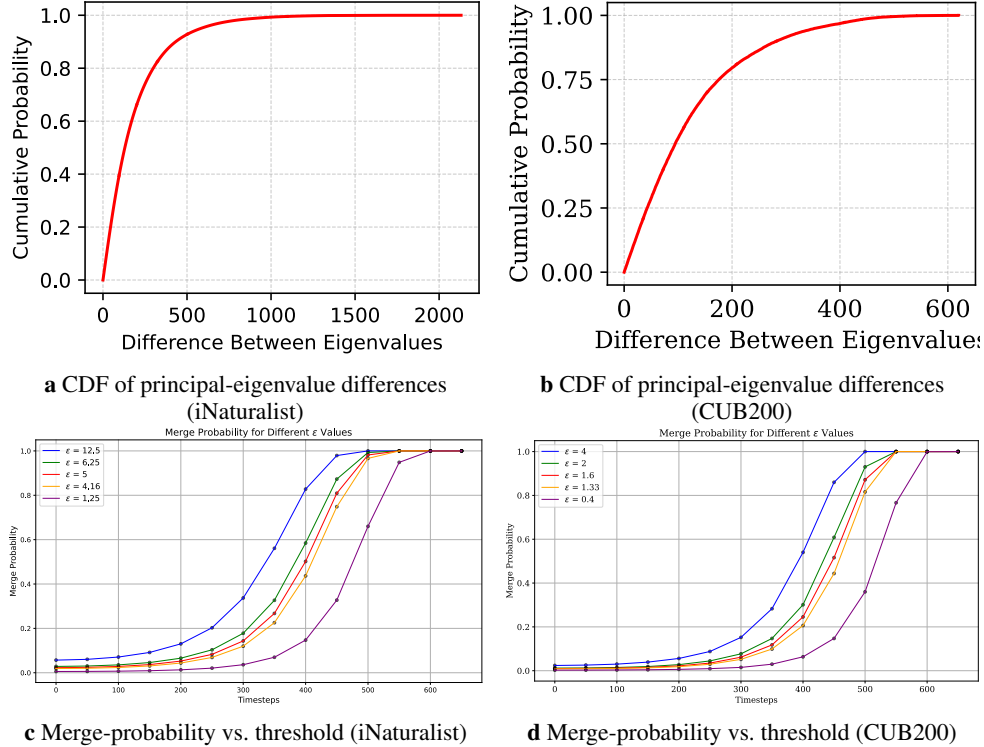


Figure 12: Comparison of eigenvalue-difference distributions and merge-probability curves for iNaturalist vs. CUB200.

1470 **Experimental settings.** For all runs, we fix  $\epsilon = 5$  on iNaturalist and  $\epsilon = 2$  on CUB-200 when  
 1471 computing merger statistics. **Figures 11a** and **11c** visualise per-class merger probabilities, with  
 1472 complementary CDF and eigenvalue trajectories in **Figures 12a** to **12d**. Top-10 eigenvalue plots  
 1473 appear in **Figures 11b** and **11d**. Qualitative comparisons between **Algorithm 3** and naïve Stable  
 1474 Diffusion, using identical random seeds, is given in **Figures 13** and **14**.



Figure 13: Visual comparison of generation algorithms for stable diffusion for the iNaturalist class prompt “clouded leopard walking”.

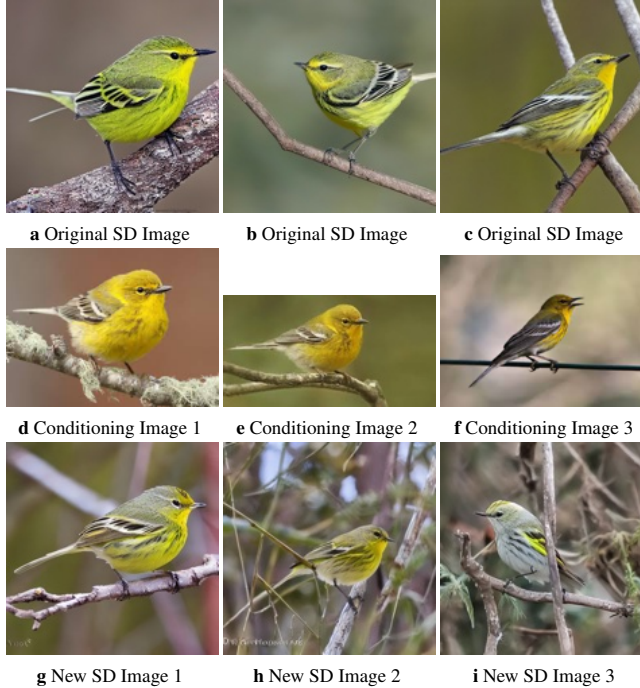


Figure 14: Visual comparison of generation algorithms for stable diffusion for the CUB-200 prompt “pine warbler.”

## 1475 C.5 Zero-shot classification

1476 In Li et al. [2023], a pretrained *class-conditional* diffusion network  $f_\theta$  is repurposed as a classifier by  
 1477 averaging softmax logits over forward-diffused replicas of the query image. We keep that spirit but  
 1478 (i) restrict the average to the *class-specific guidance window*  $[t_{\text{start},\lambda}, t_{\text{stop},\lambda}]$  identified in Section 4.2,  
 1479 and (ii) attach an importance weight  $w(t)$  to every timestep  $t$ . Setting  $w(t) = 1/T$  and the bounds  
 1480  $t_{\text{start},\lambda} = 0, t_{\text{stop},\lambda} = T$  recovers the estimator of Li et al. [2023].

1481 Throughout this section,  $\Lambda = \{1, \dots, K\}$  is the label set;  $\varepsilon_n \stackrel{\text{iid}}{\sim} \mathcal{N}(0, I_d)$  are noise seeds ( $n =$   
 1482  $1, \dots, N$ );  $\text{FWD}_t(x, \varepsilon)$  denotes the *forward* map that produces the noisy latent  $z_t$  at step  $t$ ;  $w(t) \geq 0$   
 1483 is a probability mass on the integer interval  $[t_{\text{start},\lambda}, t_{\text{stop},\lambda}]$ .

---

### Algorithm 4 Zero-shot class probability $p_\lambda(z)$

---

**Require:** query  $z \sim p_0$ , class label  $\lambda \in \Lambda$ , time window  $t_{\text{start},\lambda} \leq t \leq t_{\text{stop},\lambda}$ , weights  $\{w(t)\}$  summing to 1,  
 noise seeds  $\{\varepsilon_n\}_{n=1}^N$   
 1:  $p_\lambda \leftarrow 0$   
 2: **for**  $t = t_{\text{start},\lambda}, \dots, t_{\text{stop},\lambda}$  **do**  
 3:   **for**  $n = 1, \dots, N$  **do**  
 4:      $z_t \leftarrow \text{FWD}_t(z, \varepsilon_n)$   
 5:      $s_\lambda \leftarrow -\|f_\theta(z_t, \varepsilon_n, \lambda) - \varepsilon_n\|^2$   
 6:      $p_\lambda \leftarrow p_\lambda + \frac{w(t)}{N} \frac{\exp(s_\lambda)}{\sum_{\mu \in \Lambda} \exp(-\|f_\theta(z_t, \varepsilon_n, \mu) - \varepsilon_n\|^2)}$   
 7:   **end for**  
 8: **end for**  
 9: **return**  $p_\lambda$

---

1484 Let  $\alpha_t$  be the signal coefficient of the VP process (2.4). The signal-to-noise ratio is  $\text{SNR}(t) =$   
 1485  $\alpha_t^2 / (1 - \alpha_t^2)$ . We consider three discrete weight laws: *Uniform*:  $w(t) = 1/(t_{\text{stop},\lambda} - t_{\text{start},\lambda} + 1)$ .  
 1486 *Inverse-SNR*:  $w(t) \propto \text{SNR}(t)^{-1}$  with  $t_{\text{start},\lambda} = 0$ . *Truncated inverse-SNR*: same as previous but  
 1487 restricted to  $t \geq 20$  (empirically, very early steps degrade performance as the score model is not  
 1488 exact at these time scales [Chen et al., 2022a, Karras et al., 2022]). We fix  $N = 250$  for all of our  
 1489 experiments following Li et al. [2023].

## 1490 C.6 Binary classification via linear probes

1491 To understand why non-uniform weights help, we study binary accuracy along the forward chain  
 1492 with *no* diffusion model involved. Pick two ImageNet classes  $\{\lambda, \mu\}$  at random and sample  
 1493  $\min\{\text{card}(\lambda), \text{card}(\mu), 10000\}$  number of images for each class. At each step  $t$ , we extract the  
 1494 noisy embedding  $z_t$  (VP schedule) and train a linear MLP on 80% of the embeddings; the rest forms  
 1495 the test set. We repeat the experiment 20 times and report mean  $\pm$  s.d.

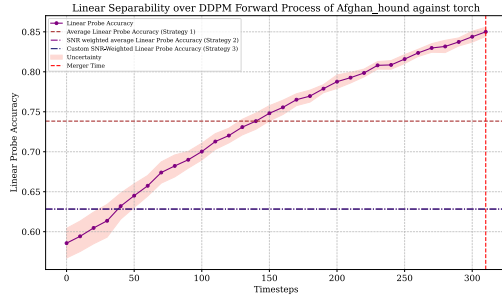
1496 Figure 15 shows that test accuracy rises sharply and peaks near the *merger time* of  $\lambda$  and  $\mu$ , after which  
 1497 it is not defined since the embeddings corresponding to  $\lambda, \mu$  become practically indistinguishable.  
 1498 Averaging the accuracies with the three weight laws yields the means in Table 11; using the single  
 1499 best  $t$  (the merger time) achieves the highest score but is undefined for the multi-class case, however,  
 1500 the general trend for the other strategies that are valid for the multi-class setting remains the same.

Weighting strategy	Avg. binary accuracy $\uparrow$
Uniform	$0.72 \pm 0.03$
Inverse-SNR	$0.79 \pm 0.06$
Trunc. inv. SNR	$0.82 \pm 0.04$
Merge-time probe	<b><math>0.90 \pm 0.02</math></b>

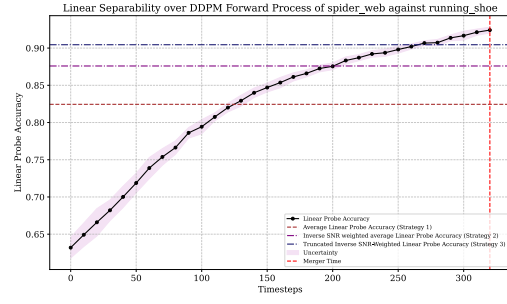
Table 11: Binary accuracy for random ImageNet pairs.

1501 Each forward step convolves the data with a Gaussian kernel, progressively smoothing non-linear  
 1502 features. Just before the classes merge, the representation is *simpler* yet still separates the two  
 1503 manifolds, making linear decision boundaries easiest to learn. This explains why inverse-SNR  
 1504 weighting, which emphasises mid-to-late timesteps—beats uniform weighting, and provides an  
 1505 additional reason why further truncation gives a small extra gain. Thus, merger-aware weighting  
 1506 focuses the zero-shot estimator on the most discriminative region of the diffusion trajectory, narrowing  
 1507 the accuracy gap to dedicated representation learners such as CLIP.

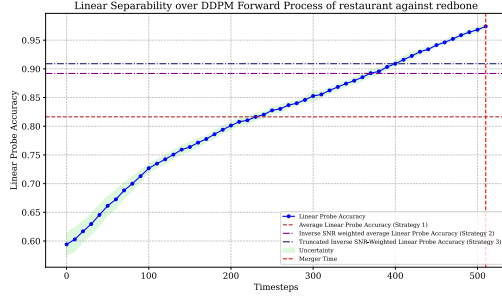




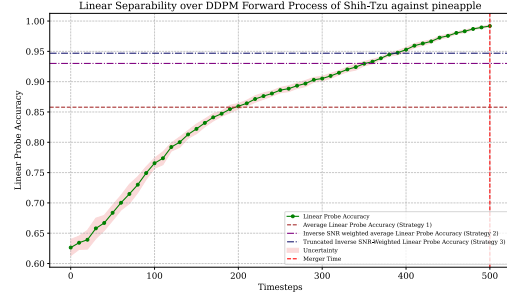
a



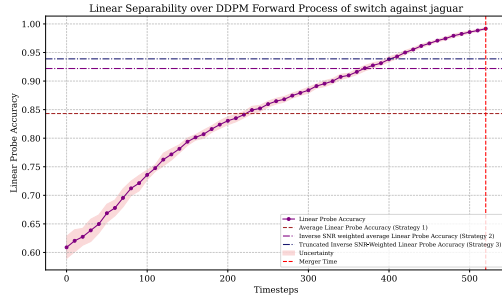
f



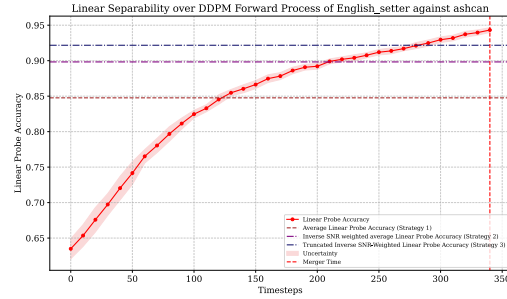
b



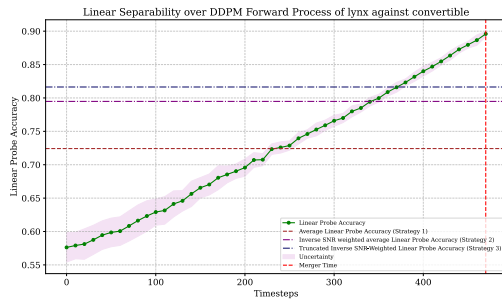
g



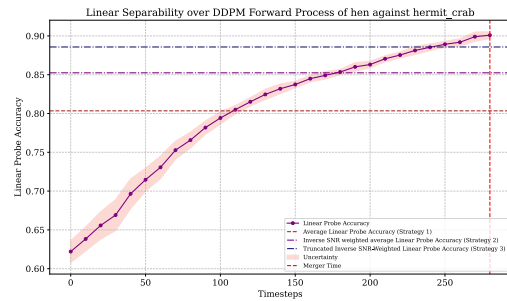
c



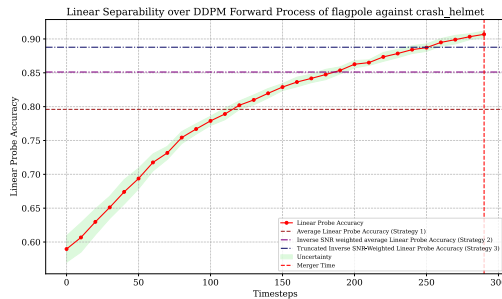
h



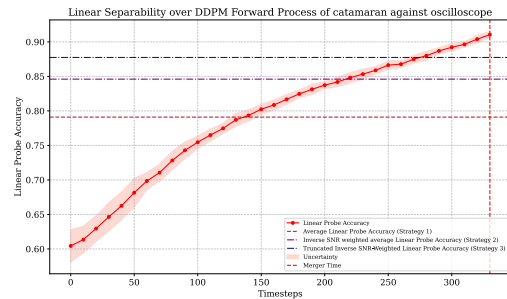
d



i



e



j

Figure 15: Linear-probe accuracy through the forward diffusion process. Later timesteps hold greater discriminative information between the two classes.

## C.7 Zero Shot Style Transfer

In [Meng et al., 2021] the parameter  $t_{\text{stop},z}$  is estimated by a grid search over the interval  $0.1T, 0.2T, \dots, T$  for an entire dataset for each style which we follow for our baseline implementation using the PSNR metric.  $t_{\text{stop},z}$  estimated in this way for both datasets were observed to lie in the range  $0.3T, 0.4T, 0.5T$  across styles. For our implementation  $t_{\text{stop},z}$  is estimated class-wise as the smallest merger time of the class(es) to which  $z$  belongs.

---

### Algorithm 5 Zero-shot Style Transfer

---

**Require:** query  $z \sim p_0$ , noise seed  $\epsilon_n$ , maximum time  $t_{\text{stop},z}$

- 1:  $z_{t_{\text{stop},z}} \leftarrow \text{FWD}_{t_{\text{stop},z}}(z, \epsilon_n)$
  - 2: **return**  $z^* \leftarrow \text{BWD}_{\theta, t_{\text{stop},z}}(z_{t_{\text{stop},z}})$
- 

We use open-source fine-tuned stable diffusion models available on Hugging Face for all of our experiments. These models are trained on the artistic styles of Studio Ghibli Miyazaki and Ghibli [2014], Van Gogh Roojen [2019] and the styles of the animation game Elden Ring Software and Entertainment [2022] and series Arcane Production and Games [2021]<sup>9</sup>. Note that some samples from the OxfordIIITPet Dataset may be randomly censored by the automatic filter present in these models. We observed that this happens mostly for dog images, where some breeds have samples with their mouth open. This can be mitigated to some extent by center cropping and realigning; however, if preprocessing fails, we discard such images. Table 5 in Section 4.5 of the main paper has results for the Studio Ghibli and Van Gogh styles while Table 12 has results for the Elden Ring and Arcane styles.

Style	Elden Ring		Arcane	
Models/Metrics	PSNR ( $\uparrow$ )	MSE ( $\downarrow$ )	PSNR ( $\uparrow$ )	MSE ( $\downarrow$ )
SD Edit (OxfordIIITPets)	$24.19 \pm 0.72$	$0.09 \pm 0.006$	$25.17 \pm 0.42$	$0.09 \pm 0.005$
<b>Ours (OxfordIIITPets)</b>	<b><math>28.14 \pm 0.69</math></b>	<b><math>0.03 \pm 0.002</math></b>	<b><math>28.35 \pm 0.72</math></b>	<b><math>0.03 \pm 0.006</math></b>
SD Edit (AFHQv2)	$26.08 \pm 0.37$	$0.06 \pm 0.003$	$26.49 \pm 0.34$	$0.05 \pm 0.004$
<b>Ours (AFHQ v2)</b>	<b><math>27.56 \pm 0.37</math></b>	<b><math>0.04 \pm 0.009</math></b>	<b><math>28.23 \pm 0.18</math></b>	<b><math>0.03 \pm 0.001</math></b>

Table 12: Style Transfer results for Elden Ring and Arcane styles

Visual results are in Figures 16 and 17 for the AFHQv2 and OxfordIIITPet datasets, respectively. We also show visual proof of our core assumption from Section 4.5 elaborated in Appendix B.13 through visual plots of the evolution of the fourier transforms through the forward process of images mainly differing only in style, in Figure 18. Here we set a gap of  $0.1T$  between the two images. These plots empirically show that for images differing only in style but with the same essential structure, their fourier transforms are *close* and this fact extends to their noised versions as well due to the nonlinear decay in Brownian Motion (Appendix B.12).

---

<sup>9</sup>Disclaimer: All referenced trademarks, copyrighted characters, and original artworks remain the property of their respective owners. Algorithm 5 was utilized for research purposes only and is not intended for commercial use or to infringe upon the intellectual property rights of the original creators.

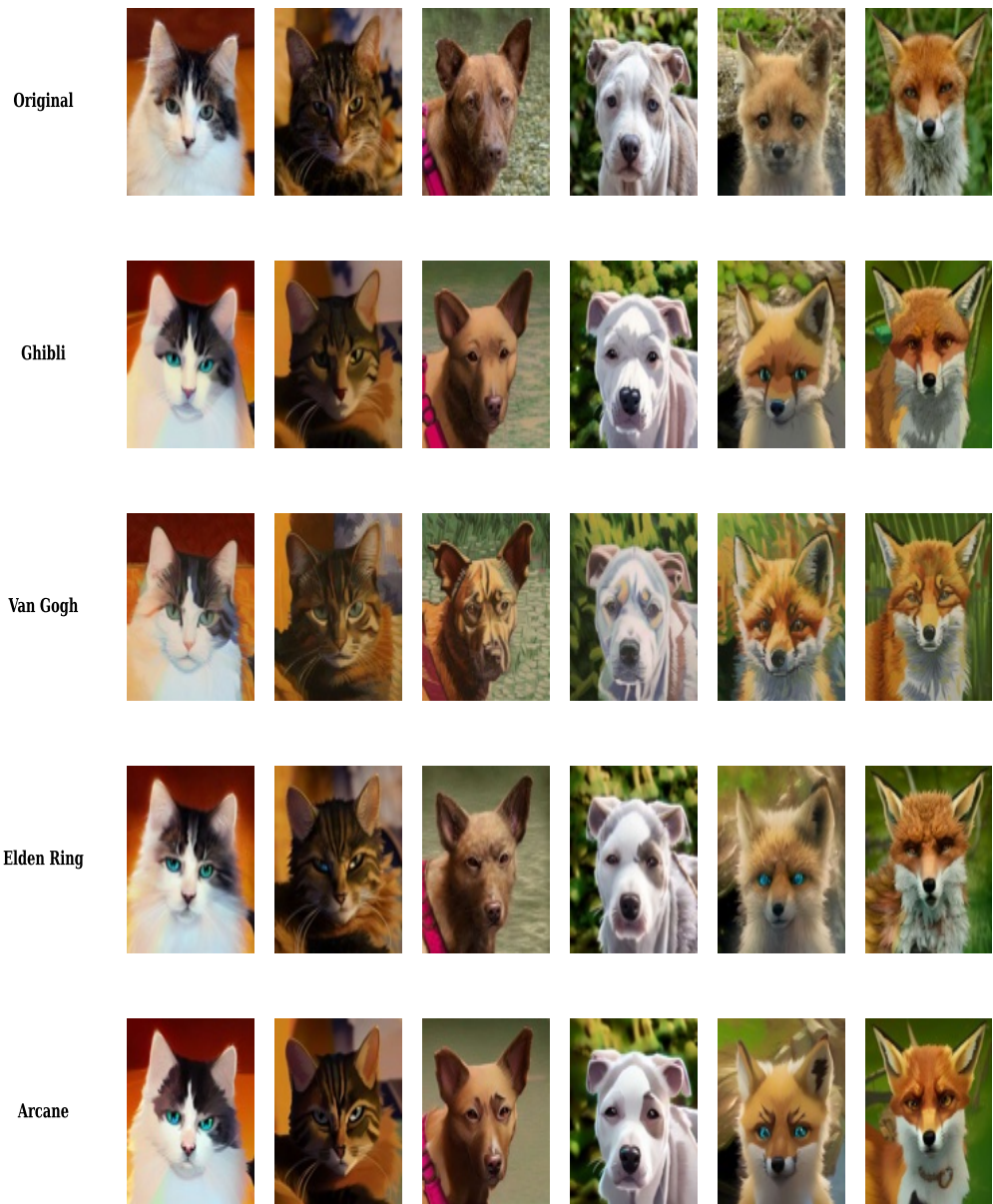


Figure 16: Zero Shot Transfer results on AFHQ v2



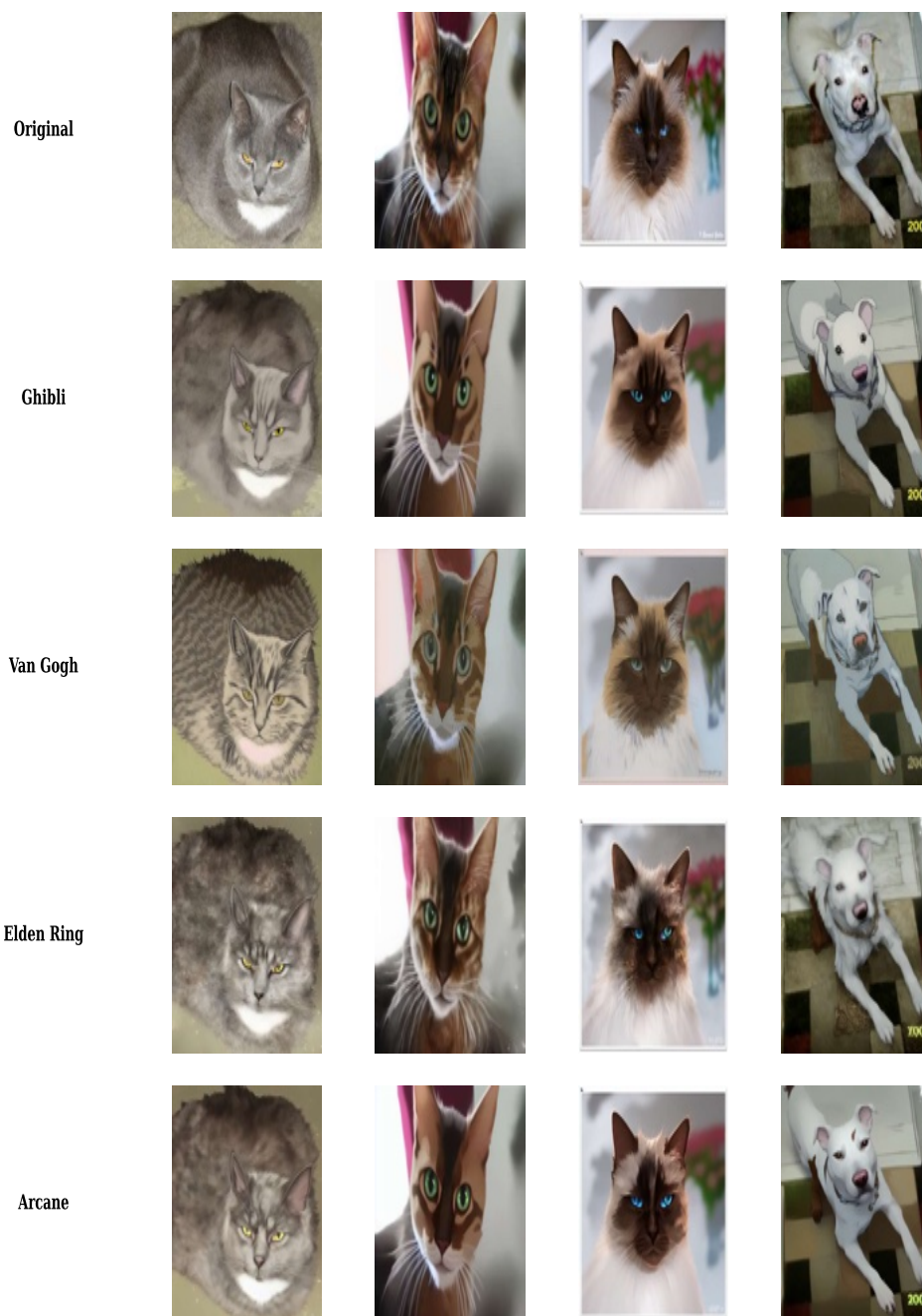


Figure 17: Zero Shot Transfer results on OxfordIIITPet

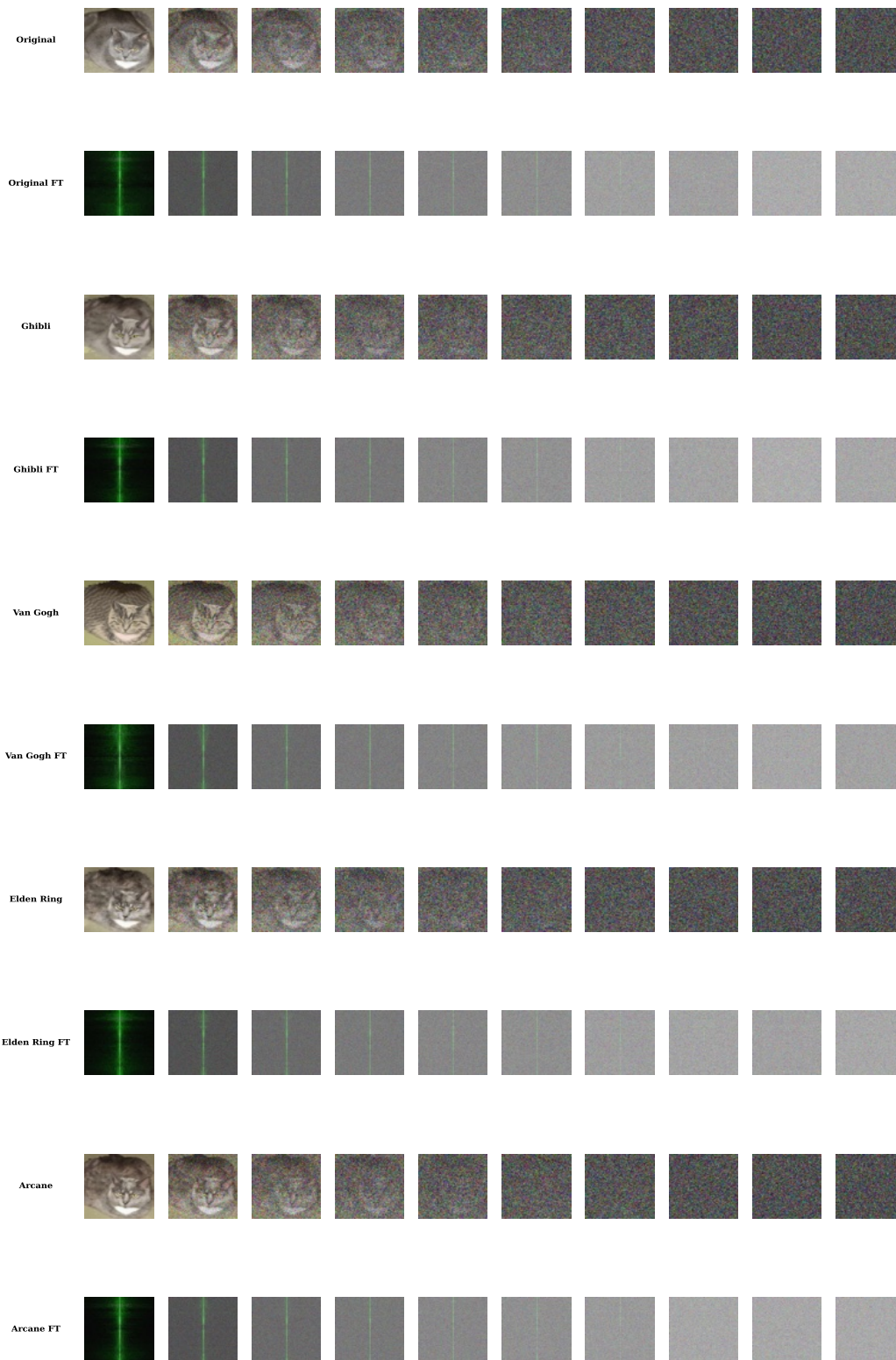


Figure 18: Fourier transforms and decay in Fourier spectra are close for structurally similar data

## 1531 D Philosophical background and related work

1532 “Nature uses only the longest threads to weave her patterns, so that each small  
1533 piece of her fabric reveals the organization of the entire tapestry.”

1534 —Richard P. Feynman

### 1535 Reductionism in science

1536 The idea that we can understand complex systems by analysing simpler building blocks is central to  
1537 modern science. Classic texts—from Schrödinger’s picture of life as an aperiodic crystal [Schrödinger,  
1538 1944] to Feynman’s lectures on the character of physical law [Feynman, 1965]—argue that clear, quan-  
1539 titative models emerge when we focus on fundamental interactions. Milestones such as Maxwell’s  
1540 theory of electromagnetism [Maxwell, 1873], Hilbert’s axioms for geometry [Hilbert, 1902], and  
1541 Shannon’s information theory [Shannon and Weaver, 1949] shows how this strategy unlocks new  
1542 domains of inquiry.

### 1543 Emergence and complexity

1544 Yet reductionism alone does not explain why new, collective behaviour appears when many parts inter-  
1545 act. Works on self-organisation and complexity—e.g., Thompson’s study of biological form [Thomp-  
1546 son, 1917], Hofstadter’s ideas on recursive structure [Hofstadter, 1979], and Mitchell’s overview of  
1547 complex systems [Mitchell, 2009]—highlight phenomena that are absent at smaller scales. Statisti-  
1548 cal physics formalises these ideas through phase transitions and fluctuations [Chaikin et al., 1995,  
1549 Landau and Lifshitz, 1980, Pathria and Beale, 2011], showing how tiny perturbations can produce  
1550 macroscopic order.

### 1551 AI as a new test case

1552 Modern generative models push the limits of our understanding of emergence. Diffusion, transformer,  
1553 and reinforcement-learning agents demonstrate impressive capabilities but remain opaque [Rudin,  
1554 2019, Hendrycks et al., 2023, Sharkey et al., 2025]. The call for interpretable AI echoes the broader  
1555 scientific quest to reduce apparent complexity to intelligible mechanisms.

### 1556 Related technical work

1557 **Overview.** Several strands inform our approach. Classical Markov-Chain analysis studies mixing  
1558 times and coupling [Aldous and Fill, 2002, Levin and Peres, 2017], while recent work connects  
1559 diffusion processes to stochastic differential equations and score matching [Sohl-Dickstein et al.,  
1560 2015, Ho et al., 2020].

1561 **The probabilistic method.** Our formalism for probing desired distributions is inspired by the  
1562 *probabilistic method* Alon and Spencer [2016], pioneered by Paul Erdős. This technique constructs  
1563 suitable probability distributions and uses moment bounds and concentration inequalities to prove the  
1564 existence or impossibility of combinatorial propositions. At a high level, our approach can be said to  
1565 apply the *probabilistic method* to analyze the properties of stochastic processes.

1566 **Lattice transitions and associated phenomena.** Our methods are rooted in statistical physics,  
1567 particularly the role of discretization in quantum systems like lattice quantum chromodynamics,  
1568 which explores complex phenomena such as the internal structure of nucleons. Classical quantum  
1569 chromodynamics is intractable for nucleons due to the strong nuclear force’s persistence over many  
1570 interacting particles, necessitating discretization via lattice models. This approach varies lattice  
1571 spacing to interpolate towards thermodynamic limits. However, in some scenarios, discretization is  
1572 inherent and does not vanish in the limit, such as in semiconductor energy states, quantum critical  
1573 transitions, and fundamental questions about the universe’s minimum resolvable length (*the Planck*  
1574 *length*).

1575 **Our Position.** In summary, while there have been attempts to probe model internals, including  
1576 feature visualisation, attribution, and mechanistic interpretability, a systematic account of emergent

bias during sampling is still lacking; we hope that our technique helps in answering some of these open questions. We view diffusion sampling as an ideal laboratory: the process is continuous, the initial distribution is known, and outputs can be inspected at any step. By framing desirable outcomes as events and applying fluctuation theory, we aim to extend the reductionist programme into the domain of large-scale generative AI.

## E Limitations and future work

**Modelling assumptions.** Our analysis is built on three structural premises: (i) a variance-preserving (VP) noise schedule, (ii) isotropic Brownian perturbations, and (iii) the Fourier-regularity bound of (4.7). These hold for the canonical SDE/ODE pairs [Song et al., 2021, Ho et al., 2020], but have not yet been extended to non-VP schedules (e.g. EDM [Karras et al., 2022]) or to anisotropic diffusions. Relaxing any of the three assumptions remains open.

**Moment truncation.** Empirically, the first four centered moments suffice to expose the phase boundaries, yet difficult data (heavy-tailed, multimodal) may demand higher orders. Reliable estimation of tensor moments beyond order four is computationally expensive and lacks sharp concentration inequalities.

**Memory overhead.** Computing covariances  $\Sigma_{k,t} \in \mathbb{R}^{d \times d}$  for many classes incurs  $O(Kd^2)$  memory. Random projections, sketching, or block-diagonal approximations could mitigate this without degrading detection power.

**Modalities beyond 2-D images.** All experiments target natural images. How merger-based guidance interacts with 3-D diffusion, video or audio, or with text diffusion is unexplored.

**Training-time usage.** We use cross-fluctuations *post hoc*. Injecting merger times into the loss—as a curriculum on noise levels or as a regulariser enforcing class separation—may accelerate or stabilise training; we leave this to future work.

## F Related work

**Statistical physics of diffusion models** Thermodynamic transitions in diffusion models have been studied in prior work, such as Raya and Ambrogioni [2024], which identifies a mixing time transition akin to ours in Section 4.1, but for a narrow class of initial distributions. This is similar to the Curie-Weiss transition in ferromagnetic systems Kivelson et al. [2024], with an analytical estimate. We provide an alternative derivation of this dependency for a broader class of sub-Gaussian data in Appendix B.10, but our approach in Section 4.1 is more general, relying solely on the assumption that the supports of the data and the isotropic Gaussian are *essentially disjoint*. Biroli and Mézard [2023], Biroli et al. [2024] extend Raya and Ambrogioni [2024] by modeling data as a Gaussian mixture and show three distinct *phases* in a class-conditional setup similar to Section 4.2. Our framework further builds on these ideas in Appendix B.14, where we showcase a general framework relying on discrete *lattice transitions* more common in quantum systems.

**Fast sampling for diffusion models.** DDIM [Song et al., 2020], IDDPM [Nichol and Dhariwal, 2021], DPM-Solver [Lu et al., 2022], and EDM [Karras et al., 2022] reduce the reverse step count; early-exit criteria such as ours are orthogonal and in principle could be combined with any of them. We leave such extensions to future work. As discussed earlier the occurrence of this criterion has also been demonstrated by [Raya and Ambrogioni, 2024, Biroli and Mézard, 2023, Biroli et al., 2024] through a different analysis based on symmetry breaking of a potential function, we show that our framework extends such considerations in Appendix B.14.

**Theoretical lenses on diffusion.** Hyper-contractivity [Saloff-Coste, 1994, Chen et al., 2022a], mixing-time bounds [Levin and Peres, 2017], and classical coupling [Aldous and Fill, 2002] traces back to Markov-chain theory. We recast these ideas as *cross-fluctuation mergers*, bridging discrete and continuous settings in Appendix B.6.

1623 **Conditional guidance.** Score distillation [Poole et al., 2022], ILVR [Choi et al., 2021], classifier-  
1624 free guidance [Ho and Salimans, 2022a], and Interval Guidance (IG) [Kynkäänniemi et al., 2024]  
1625 dominate conditional generation. Our merger-aware IG trims the interval search from per-sample to  
1626 per-class with no extra hyper-parameters with details in Section 4.2 and Appendix C.3

1627 **Zero-shot classification using diffusion networks.** Li et al. [2023] showed diffusion backbones  
1628 encode class information. Importance weighting by intermediate times with cutoff at merger times  
1629 boosts their zero-shot accuracy at equal compute (Section 4.4). We also demonstrate that near the  
1630 merger times diffusion process demonstrate near-perfect discriminability Appendix C.6.

1631 **Rare-class synthesis.** Long-tail generation typically relies on fine-tuning [Bansal et al., 2023,  
1632 Samuel et al., 2024]. Our fluctuation-based guidance is tuning-free and alleviates mode dropping at  
1633 inference time by utilizing optimized IG based guidance (Section 4.3, Appendix C.4)

1634 **Zero-shot style transfer using diffusion models** The intriguing property of style transfer by  
1635 learning a trained VP-SDE only on the target style after utilizing the forward process to corrupt inputs  
1636 from the source style was initially shown in Meng et al. [2021] and has become standard practice  
1637 across setups Rombach et al. [2022]. To our knowledge we are the first to provide a theoretical  
1638 foundation for this approach in terms of fourier regularity that manifests in observed transitions  
1639 (Section 4.5, Appendix B.13) while showing that by setting the noising parameter based on merger  
1640 times can boost fidelity (Section 4.5, Appendix C.7)