Caffarelli Regularity and Hierarchical Phase Boundaries in Diffusion Model Latent Space

Alexander Lobashev
Glam AI, San Francisco, USA
lobashevalexander@gmail.com

Dmitry Guskov
Glam AI, San Francisco, USA
guskov01dmitry@gmail.com

Victor Kawasaki-Borruat
EPFL, Lausanne, Switzerland
victor.borruat@epfl.ch

Maria Larchenko
Magicly AI, Dubai, UAE
mariia.larchenko@gmail.com

Abstract

Recent studies have shown phase-transition–like behavior in diffusion models, where a small perturbation of the initial Gaussian noise sample can cause an abrupt change in the generated image. The underlying mechanism of these transitions, however, remains theoretically underexplored. In this work, we investigate this phenomenon through the lens of the Riemannian metric on the latent space induced by the distance between CLIP embeddings. We observe a hierarchical emergence of phase boundaries: coarse boundaries appear in the early denoising steps, while finer boundaries progressively emerge within these regions as the diffusion process advances. These findings have practical implications for diffusion inversion–based image editing: images within the same Riemannian basin can be edited with only a few inversion steps, whereas images that are nearby in latent space but separated by a phase boundary require substantially more steps. To provide a theoretical foundation, we approximate the reverse diffusion dynamics by a discrete-time sequence of quadratic-cost optimal transport maps between successive noisy marginals. By employing Caffarelli's regularity theory, we demonstrate that discontinuities of the diffusion generative map are associated with mode-splitting, thereby giving rise to phase boundaries. This leads to a hierarchical, tree-like organization of data distribution modes, implying that distances between images in this geometry follow an ultrametric structure.

1 Introduction

Diffusion models are powerful generative models capable of achieving good mode coverage. Recently, their behavior has been studied from the perspective of statistical physics and the phenomenon of phase transitions has been revealed. Phase transitions in diffusion models can be divided into temporal and spatial ones.

Temporal phase transitions Diffusion sampling undergoes different regimes as time evolves from high noise to clean data, which are known as dynamical phase transitions Biroli et al. [2024]. In the high-noise regime, the distribution over noisy data remains unimodal. In the middle stage, the diffusion path chooses the data mode from which the sample will be generated, so the medium- and low-frequency details of the image are formed. In the final stage, only high-frequency details are altered, while the overall content of the generated image remains the same. It was also noted that the Lipschitz constant of diffusion diverges as time approaches zero Yang et al. [2023]; see also analyses of nonequilibrium and phase phenomena Sclocchi et al. [2025], Yu and Huang [2025].

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

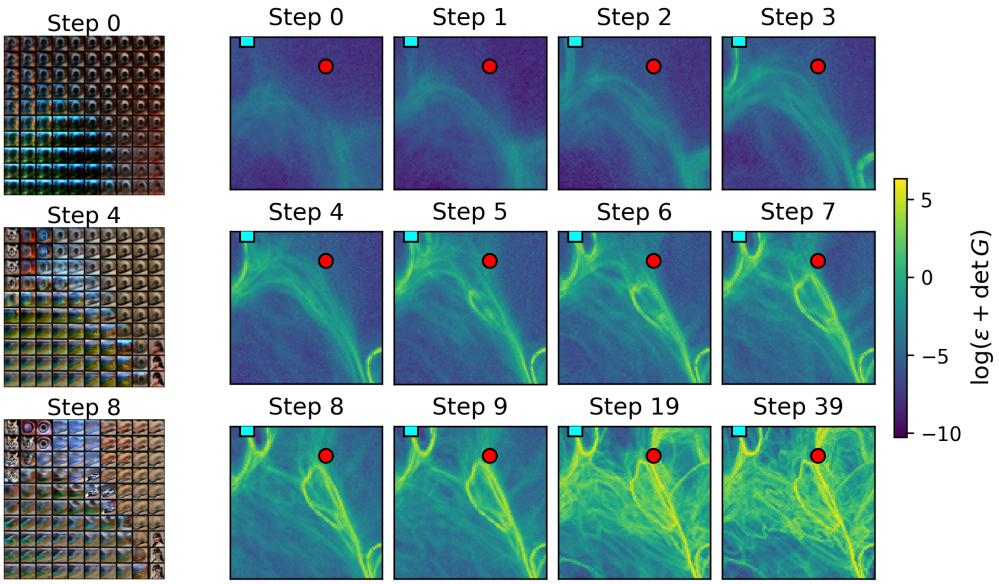

Figure 1: Left: a grid of predicted noiseless images, $\hat{x}_0(\alpha, \beta)$, over a 2D latent slice. Right: evolution of the determinant of the CLIP pullback metric, G, shown on a log scale as $\log(\varepsilon + \det G)$ (with $\varepsilon = 10^{-8}$) across denoising timesteps for the same slice. The determinant is proportional to the local volume element of the latent space; high values indicate regions where the features of the generated images change abruptly. The red circle and cyan square mark the two latent points used in Figure 3.

Spatial phase transitions The second line of research concerns how the generated image changes under small perturbations of the initial Gaussian noise. These studies are related to the geometry of the latent space. Empirically, it has been observed that generative models which learn primarily a single data mode — such as GANs or VAEs trained on human faces — exhibit latent geometry that is nearly flat, with vanishing curvature Shao et al. [2018], Wang and Ponce [2021]. In contrast, when models generate data from multiple distinct modes, diffusion models exhibit boundaries between phases Lobashev et al. [2025]. Within a single mode, the latent geometry is flat: a straight line in latent space corresponds to a geodesic interpolation. However, when the geodesic crosses a boundary, it curves, and the generated image changes abruptly. This indicates that diffusion models, as maps from the prior Gaussian distribution to the data distribution, exhibit a diverging Lipschitz constant (with respect to the spatial axis). To reduce sensitivity to latent perturbations Liu et al. [2021], Guo et al. [2024] promote smoothing the geometry of the latent space.

Previous work has mostly studied the map from noise to clean data, leaving the formation of intermediate boundaries underexplored. In contrast to this, our analysis covers both the temporal and spatial aspects of boundary formation. It relies on the theory of regularity of the optimal transport maps Caffarelli [1992], Figalli [2010], Luo et al. [2022].

Contributions In this work, we study in detail the formation of phase boundaries in the latent space of diffusion models. We show that boundaries emerge in a hierarchical way. As sampling progresses further into the low-noise regime, new finer boundaries appear within the regions separated at the previous level, Figure 1. We represent the denoising process as a series of optimal transport maps connecting probability distributions between consecutive timesteps and attribute the appearance of these boundaries to the singularity of optimal transport maps.

2 Background and Method

Regularity of transport maps was studied in the works of Caffarelli and Figalli. We refer to Figalli's regularity results for optimal maps between nonconvex planar domains for precise statements Figalli [2010].

Reverse diffusion path could be approximated by a sequence of optimal transport maps We give a derivation showing that, for a sufficiently smooth probability density ρ on $\mathbb{R}^d$, the heat (diffusion) interpolation

$$\rho_t := \rho * \gamma_t, \qquad \gamma_t = \mathcal{N}(0, tI_d), \tag{1}$$

is close to the quadratic-cost optimal-transport (displacement) interpolation between $\rho = \rho_0$ and $\rho_t = \rho * \gamma_t$ when time t is small. A precise statement of this result is provided in the appendix.

By leveraging this connection, we can model the reverse diffusion path as a sequence of optimal transport maps. Another way to see this is by using Jordan-Kinderlehrer-Otto (JKO) scheme Jordan et al. [1998] for the Fokker-Plank equation (FPE). Since solution of the FPE is the gradient flow of the Wasserstein metric then discretization of this flow gives the approximation of the reverse ODE solution by a composition of optimal transport maps. This framework then allows us to apply Figalli's regularity theorem Figalli [2010] to rigorously explain the appearance of growing singular sets during the denoising process.

Caffarelli regularity of optimal transport maps Our theoretical explanation of the appearance of phase boundaries (as illustrated in Figure 1) can be summarized in the following statement.

Theorem 2.1 (Hierarchical Emergence of Phase Boundaries)**.** *Let p_T be a high-noise Gaussian prior and p_0 be the target data distribution. Consider the discrete approximation of the probability-flow ODE given by a sequence of optimal transport maps $(T_k)_{k=1}^N$, where $T_k : \mathbb{R}^d \to \mathbb{R}^d$ is the Brenier map pushing p_{t_k} to $p_{t_{k-1}}$.*

1. ***(Existence and Regularity)** For each k, the map T_k exists, is unique, and is the gradient of a convex potential, $T_k = \nabla\phi_k$. Away from a singular set of measure zero, ϕ_k is smooth.*
2. ***(Formation of Singularities)** A singular set $S_k \subset supp(p_{t_k})$ for the potential ϕ_k emerges if the domain $supp(p_{t_{k-1}})$ is not convex with respect to the mapping from $supp(p_{t_k})$. Such a condition arises naturally during mode-splitting or due to non-convexity of the support of data distribution.*
3. ***(Propagation and Accumulation)** Let $\mathcal{T}_{k\to 0} = T_1 \circ T_2 \circ \cdots \circ T_k$ be the composed map from time t_k to t_0. The total singular set in the latent space at time T, corresponding to all phase boundaries, is given by the union of preimages:*

$$S_{\text{total}} = \bigcup_{k=1}^{N} (\mathcal{T}_{N\to k})^{-1}(S_k). \tag{2}$$

The complexity (number of singularities) of this set is non-decreasing as $t \to 0$.

3 Experiments

Phase boundaries evolution We investigate the formation of phase boundaries in the latent space of diffusion models. We use CLIP embeddings of generated images at different noise levels to estimate the metric tensor on a 2D latent slice. For each noise level, we use the data generation procedure described in Lobashev et al. [2025]. We fix three latent anchors $z_0, z_1, z_2 \in \mathbb{R}^k$ and parametrize a two–dimensional affine slice by

$$z(\alpha, \beta) \;=\; z_0 \;+\; \alpha\,(z_1 - z_0) \;+\; \beta\,(z_2 - z_0), \qquad (\alpha, \beta) \in [0, 1]^2. \tag{3}$$

At reverse time t, we run the diffusion sampler to obtain the predicted clean image $\hat{x}_0(\alpha, \beta; t)$ and embed it with CLIP (ViT-B/32) Radford et al. [2021] to obtain a feature

$$f(\alpha, \beta; t) \;=\; \phi\big(\hat{x}_0(\alpha, \beta; t)\big) \in \mathbb{R}^d. \tag{4}$$

The same sampler settings (guidance, steps, η) are used across the grid to ensure comparability across (α, β) and t.

On a uniform $N \times N$ grid with $\alpha_i = i/(N-1)$, $\beta_j = j/(N-1)$, we estimate the Jacobian of the feature map by centered finite differences and one–sided differences at the boundary. Stacking the columns gives $J(\alpha, \beta; t) \in \mathbb{R}^{d\times 2}$, and the pullback metric on the slice is

$$G(\alpha, \beta; t) \;=\; J(\alpha, \beta; t)^\top J(\alpha, \beta; t) \;\in\; \mathbb{R}^{2\times 2}. \tag{5}$$

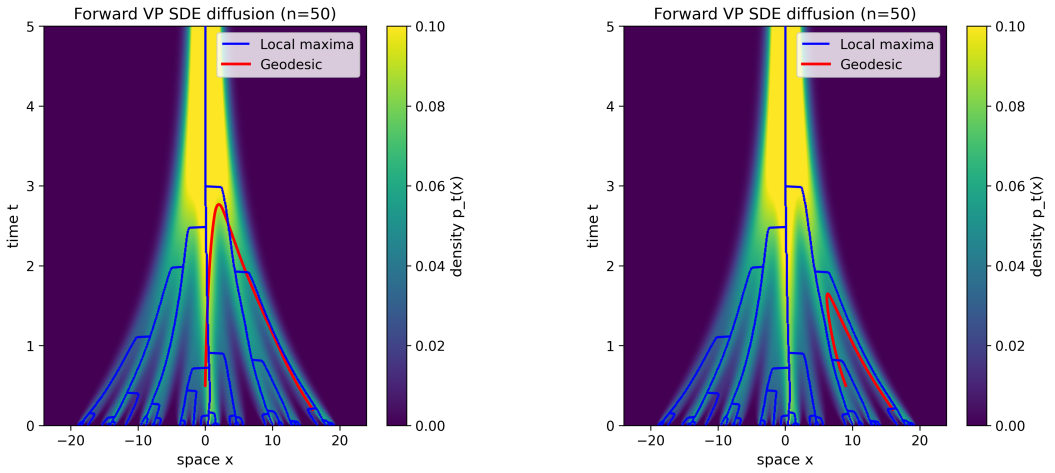

Figure 2: Spacetime density heatmap with linked maxima (blue) showing branching/merging and with spacetime geodesic (red) between chosen endpoints.

For each t, we render heatmaps of $S(\alpha, \beta; t) = \det G$ on $[0,1]^2$ with a shared color scale (see Figure 1). We observe that: (i) for large noise (early reverse time), $S \approx 0$ (flat metric); (ii) as t decreases, narrow high–S ridges appear and branch, partitioning the slice into phases with low interior sensitivity; (iii) at late t, the ridge pattern stabilizes and the mean of S approaches a plateau.

1D Diffusion We study how branching (emergence/merging of modes) arises over time and relate these changes to the intrinsic spacetime geometry by computing geodesics introduced in Karczewski et al. [2025]. We consider the 1D variance-preserving (VP) SDE. Let the initial data be an empirical distribution with atoms $\{x_{0,i}\}_{i=1}^{n}$ sampled uniformly on $[a, b]$. The forward marginal at time t is the equal-weight Gaussian mixture

$$p_t(x) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{N}\big(x;\ \alpha(t)\, x_{0,i},\ \sigma^2(t)\big). \tag{6}$$

We evaluate $p_t(x)$ on a spacetime grid to produce a density heatmap using exact forward marginals.

For each time slice $x \mapsto p_t(x)$, we detect local maxima using a relative-height threshold that adapts to the slice's scale. This yields a piecewise-linear "mode evolution tree" (blue line segments), revealing branching and merging as diffusion progresses.

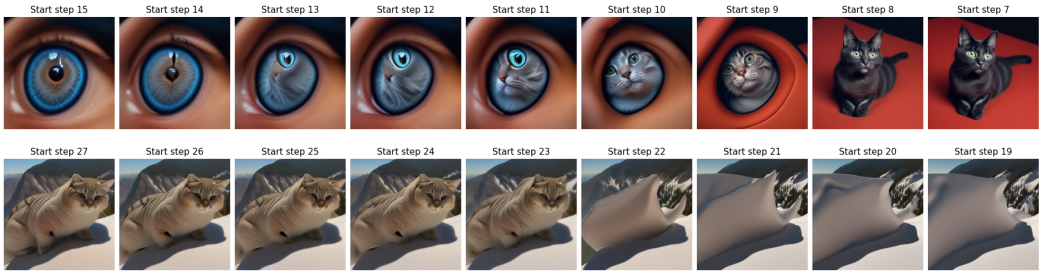

Figure 3: Prompt-guided edits: (top) an eye edited with 'a green eyed cat'; cyan square latent at Figure 1; (bottom) cat in mountains edited with 'snow mountains'; red circle latent at Figure 1

Following Karczewski et al. [2025], we compute spacetime geodesics by minimizing the discrete (μ, η)-energy along curves between fixed endpoints (see Karczewski et al. [2025] for details). We parameterize the curve with a cubic spline and optimize with Adam while softly constraining $t(s) \in [0, T]$. The resulting geodesic (red curve) reflects the intrinsic geometry of the diffusion spacetime (see Figure 2).

Image Editing Having images in one phase before a certain time means that their latents share features and their noiseless predictions point to similar images. We evaluate prompt-guided editing via partial DDIM inversion Mokady et al. [2023]. Given a source image x_{src} and source generic prompt, we invert to an intermediate step t_k to obtain a latent state x_{t_k}. We then resume reverse diffusion conditioned on a target prompt p_{tgt} to produce an edited image $\hat{x}_0$, holding sampler hyperparameters (guidance scale, steps, η) fixed across k. We observe that editing success rates correlate with mode-splitting times. Qualitative results are shown in Figure 3, where different edits require different numbers of steps: (top) for the cyan point $(\alpha, \beta) = (0.99, 0.11)$ we need to invert back to step 11, while for the red point $(\alpha, \beta) = (0.80, 0.62)$ (bottom) we need only step 23.

We visualize pairs of nearby points (α, β) on the latent grid and their predicted noiseless images. Before two points are separated by a phase boundary (high-S ridge), their latents yield similar noiseless predictions; once a boundary intervenes, the predictions diverge.

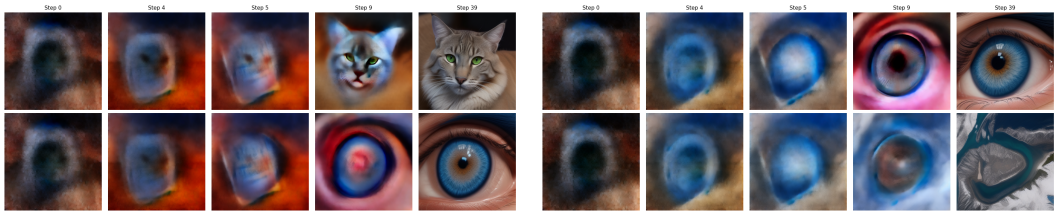

Figure 4: Nearby-grid latent pairs: left $(\alpha, \beta) = (1.00, 0.10)$ vs $(1.00, 0.11)$; right $(\alpha, \beta) = (0.91, 0.24)$ vs $(0.91, 0.25)$. Before crossing a boundary, nearby latents yield similar noiseless predictions; after crossing, they diverge.

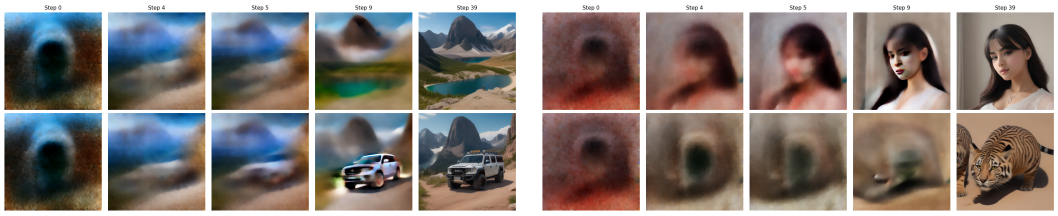

Figure 5: Nearby-grid latent pairs: left $(\alpha, \beta) = (0.51, 0.50)$ vs $(0.51, 0.51)$; right $(\alpha, \beta) = (0.22, 1.00)$ vs $(0.23, 1.00)$. Similarity holds before a boundary; it breaks after.

4 Discussion and Conclusion

In this work, we observe the hierarchical emergence of phase boundaries in the latent space of a diffusion model. Whereas prior work examined only the boundaries present at the final timestep, we show that regions with large Lipschitz constants appear even in the very first denoising steps. These coarse boundaries sharpen as the denoising process proceeds, and new, finer boundaries appear within basins delineated by the coarser ones from earlier steps. One may ask whether this behavior is an artifact of the diffusion backbone architecture or an inherent property of the data distribution. We argue for the latter: we show theoretically that such boundaries arise for multimodal data with disjoint support, making sharp boundaries an intrinsic feature of the generative modeling problem.

Our main technical tool is an approximation of the reverse probability-flow ODE (equivalently, the corresponding Fokker–Planck equation) by a composition of quadratic optimal transport maps. This composition can also be obtained via a discretization of the JKO scheme, which formulates the Fokker–Planck equation as the gradient flow of the entropy with respect to the Wasserstein distance. By applying the Caffarelli–Figalli regularity theory to the individual transport maps in this reverse-ODE approximation, we demonstrate the hierarchical accumulation of singular sets. From a practical perspective, our findings call into question current diffusion-transformer architectures that do not explicitly account for such singularities, and suggest introducing more flexible layers capable of modeling them, especially for diffusion distillation and few-step generation.

5 Acknowledgements

VKB was funded by the Swiss National Science Foundation (SNSF), grant number SNF 200021-232277.

References

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Luis A Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

Lawrence C Evans. *Partial differential equations*, volume 19. American mathematical society, 2022.

Alessio Figalli. Regularity properties of optimal maps between nonconvex domains in the plane. *Communications in Partial Differential Equations*, 35(3):465–479, 2010.

Thomas Hakon Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.

Jiayi Guo, Xingqian Xu, Yifan Pu, Zanlin Ni, Chaofei Wang, Manushree Vasu, Shiji Song, Gao Huang, and Humphrey Shi. Smooth diffusion: Crafting smooth latent spaces in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7548–7558, 2024.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Rafał Karczewski, Markus Heinonen, Alison Pouplin, Søren Hauberg, and Vikas Garg. Spacetime geometry of denoising in diffusion models. *arXiv preprint arXiv:2505.17517*, 2025.

Yahui Liu, Enver Sangineto, Yajing Chen, Linchao Bao, Haoxian Zhang, Nicu Sebe, Bruno Lepri, Wei Wang, and Marco De Nadai. Smoothing the disentangled latent style space for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10785–10794, 2021.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models. *arXiv preprint arXiv:2506.10632*, 2025.

Zhongxuan Luo, Wei Chen, Na Lei, Yang Guo, Tong Zhao, Jiakun Liu, and Xianfeng Gu. The singularity set of optimal transportation maps. *Computational Mathematics and Mathematical Physics*, 62(8):1313–1330, 2022.

Lykon. Dreamshaper-8. `https://huggingface.co/Lykon/dreamshaper-8`, 2023.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6038–6047, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/radford21a.html`.

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

Hang Shao, Abhishek Kumar, and P Thomas Fletcher. The riemannian geometry of deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 315–323, 2018.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Binxu Wang and Carlos R Ponce. A geometric analysis of deep generative image models and its applications. In *International Conference on Learning Representations*, 2021.

Zhantao Yang, Ruili Feng, Han Zhang, Yujun Shen, Kai Zhu, Lianghua Huang, Yifei Zhang, Yu Liu, Deli Zhao, Jingren Zhou, et al. Lipschitz singularities in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Zhendong Yu and Haiping Huang. Nonequilbrium physics of generative diffusion models. *Physical Review E*, 111(1):014111, 2025.

A Additional illustrations

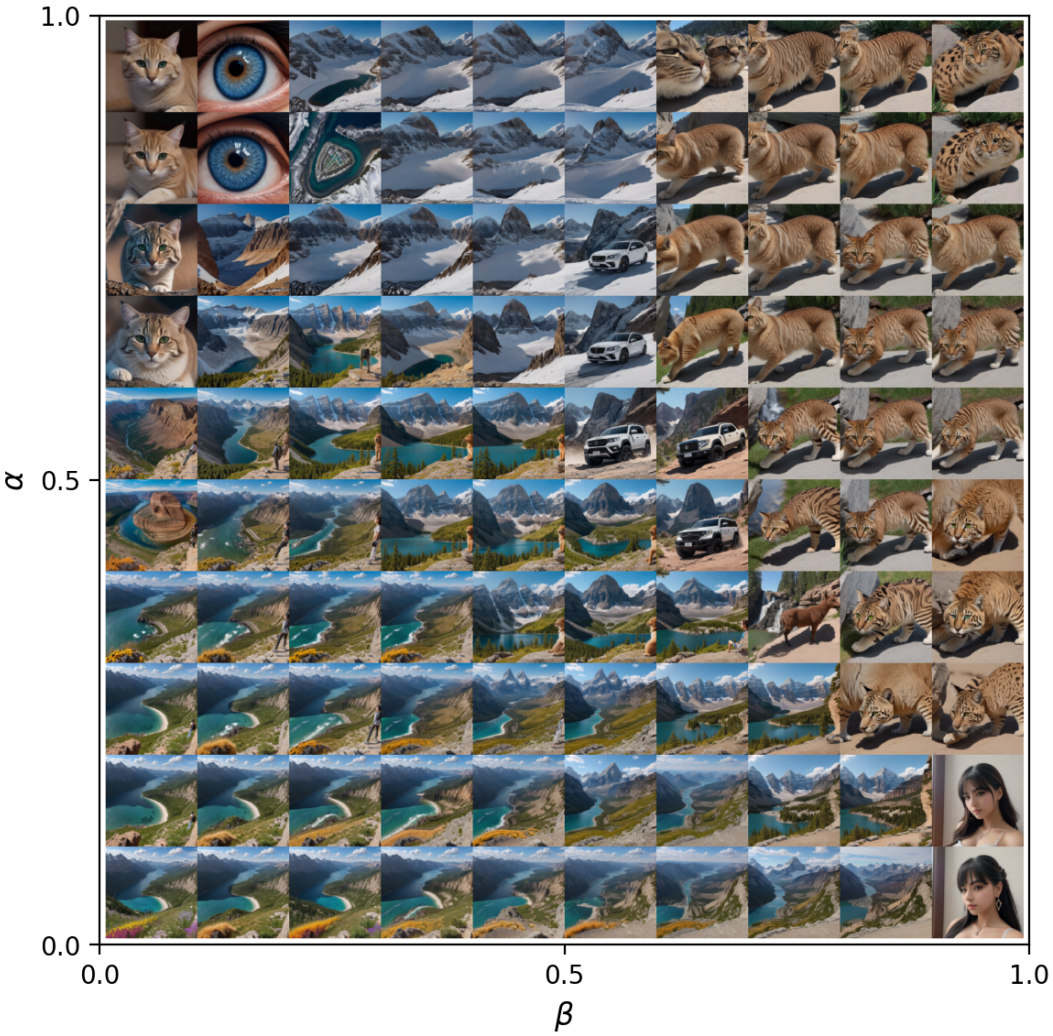

Figure 6: Annotated determinant grid with (α, β) axes: origin at lower-left; α increases upward; β increases to the right.

Figure 6 shows a detailed grid of generated images for a two-dimensional section of the latent space of the SD-1.5 Dreamshaper model Lykon [2023]. The three corners $(0, 0)$, $(0, 1)$, and $(1, 0)$ correspond to anchor latent vectors, and the entire grid is obtained by linear interpolation with standard-deviation normalization.

B Proof of theoretical results

We begin with the standing assumptions and notation. Let $\rho : \mathbb{R}^d \to (0, \infty)$ be a probability density of class $C^4(\mathbb{R}^d)$ with derivatives of at most polynomial growth. We assume there exist constants $0 < m \leq M < \infty$ such that

$$m \leq \rho(x) \leq M, \qquad \forall x \in \mathbb{R}^d, \tag{7}$$

and ρ decays sufficiently at infinity so that all integrations by parts used below are legitimate. For $t \geq 0$ we denote by γ_t the centred Gaussian with covariance tI_d, namely

$$\gamma_t(y) = (2\pi t)^{-d/2} \exp\Big(- \frac{||y||^2}{2t} \Big), \tag{8}$$

and we define the heat flow regularisation of ρ by

$$\rho_t := \rho * \gamma_t, \qquad \mu_t := \rho_t(x)\,dx. \tag{9}$$

With this normalization ρ_t solves the heat equation

$$\partial_t \rho_t = \tfrac{1}{2}\Delta \rho_t, \qquad t > 0. \tag{10}$$

We denote by W_2 the quadratic Wasserstein distance and by T_ε the Brenier map transporting $\mu = \rho\,dx$ to $\mu_\varepsilon = \rho * \gamma_\varepsilon\,dx$. The corresponding displacement interpolation is

$$\mu_t^{\mathrm{OT}} = \Big((1 - \tfrac{t}{\varepsilon})\mathrm{Id} + \tfrac{t}{\varepsilon} T_\varepsilon\Big)_{\#}\mu, \qquad t \in [0, \varepsilon]. \tag{11}$$

Theorem B.1 (Reverse diffusion path could be approximated by a sequence of optimal transport maps)**.** *Under the standing assumptions, there exist constants $C, C' > 0$ (depending only on ρ and d) such that for all sufficiently small $\varepsilon > 0$ one has*

$$\sup_{t\in[0,\varepsilon]} \mathsf{W}_2(\mu_t, \mu_t^{\mathrm{OT}}) \le C\,\varepsilon^2, \tag{12}$$

and

$$\Big\| T_\varepsilon - \Big(\mathrm{Id} - \tfrac{\varepsilon}{2}\nabla \log \rho\Big)\Big\|_{L^2(\mu)} \le C'\varepsilon^2. \tag{13}$$

Consequently the heat interpolation $(\mu_t)_{0\le t\le\varepsilon}$ is a second-order accurate approximation in ε to the displacement interpolation, and the Brenier map admits the first-order expansion

$$T_\varepsilon = \mathrm{Id} - \tfrac{\varepsilon}{2}\nabla \log \rho + O(\varepsilon^2) \qquad \textit{in } L^2(\mu). \tag{14}$$

Proof. Let us apply a fourth-order Taylor expansion of ρ around x:

$$\rho(x + \sqrt{t}Z) = \rho(x) + \sum_{|\alpha|=1}^{4} \frac{(\sqrt{t}Z)^\alpha}{\alpha!}\,\partial^\alpha \rho(x) + R_5(x, Z, t), \tag{15}$$

where we sum over a multi-index α. Then we will average this equation with respect to the Gaussian $Z \sim \mathcal{N}(0, I_d)$. Due to the Wick's theorem for gaussian random variables, the odd moments will vanish and the even moments will lead to

$$\mathbb{E}\left[\tfrac{t}{2}\sum_{i=1}^{d} Z_i^2\,\partial_{ii}\rho(x)\right] = \tfrac{t}{2}\Delta\rho(x). \tag{16}$$

for the second-order term. The forth-order term will become

$$\mathbb{E}\left[\tfrac{t^2}{24}\sum_{i,j,k,\ell} Z_i Z_j Z_k Z_\ell\,\partial_{ijkl}\rho(x)\right] = \tfrac{t^2}{8}\Delta^2\rho(x). \tag{17}$$

So for $\rho_t(x) = \mathbb{E}_{Z\sim\mathcal{N}(0,I_d)}[\rho(x + \sqrt{t}Z)]$ we have obtained

$$\rho_t(x) = \rho(x) + \frac{t}{2}\Delta\rho(x) + \frac{t^2}{8}\Delta^2\rho(x) + R_t(x), \tag{18}$$

with the remainder term $\|R_t\|_\infty \le Ct^3$. Now we will show that ρ_t is indeed obeys the heat equations. By truncating our expansion

$$\rho_\varepsilon(x) = \rho(x) + \frac{\varepsilon}{2}\Delta\rho(x) + O(\varepsilon^2). \tag{19}$$

Differentiating under convolution gives $\partial_t\rho_t = \rho * \partial_t\gamma_t$ and since $\partial_t\gamma_t = \frac{1}{2}\Delta\gamma_t$ we obtain the heat equation $\partial_t\rho_t = \frac{1}{2}\Delta\rho_t$. Using the log-derivative trick $\Delta\rho_t = \nabla\cdot\,(\rho_t\nabla\log\rho_t)$, we get the continuity equation

$$\partial_t\rho_t + \nabla\cdot\,(\rho_t v_t^{\mathrm{heat}}) = 0, \qquad v_t^{\mathrm{heat}} = -\tfrac{1}{2}\nabla\log\rho_t. \tag{20}$$

Now let's divide the expansion

$$\rho_\varepsilon(x) = \rho(x) + \frac{\varepsilon}{2}\Delta\rho(x) + O(\varepsilon^2). \tag{21}$$

by ρ to get

$$\frac{\rho_\varepsilon(x)}{\rho(x)} = \frac{\rho(x)}{\rho(x)} + \frac{1}{\rho(x)}\frac{\varepsilon}{2}\Delta\rho(x) + O(\varepsilon^2) = 1 + \frac{\varepsilon}{2}\frac{\Delta\rho(x)}{\rho(x)} + O(\varepsilon^2). \tag{22}$$

By applying $\log$ to both parts we get $\log\rho_t = \log\rho + \log(1 + \frac{t}{2}\frac{\Delta\rho}{\rho} + O(t^2))$. Then apply the gradient ∇ and by expanding the logarithm we obtain

$$\nabla\log\rho_t = \nabla\log\rho + \frac{t}{2}\nabla\Big(\frac{\Delta\rho}{\rho}\Big) + O(t^2), \tag{23}$$

hence

$$v_t^{\mathrm{heat}} = -\tfrac{1}{2}\nabla\log\rho - \frac{t}{4}\nabla\Big(\frac{\Delta\rho}{\rho}\Big) + O(t^2). \tag{24}$$

We now compute the initial OT velocity. Let v_t^{OT} be the velocity field of the OT geodesic between ρ and ρ_ε, parameterized on $[0,\varepsilon]$. Since $\rho_\varepsilon - \rho = \frac{\varepsilon}{2}\Delta\rho + O(\varepsilon^2)$, taking the first variation at $t = 0$ gives

$$\partial_t\rho_t\big|_{t=0} = \frac{1}{\varepsilon}\big(\rho_\varepsilon - \rho\big) + o(1) = \frac{1}{2}\Delta\rho + o(1). \tag{25}$$

From the continuity equation for the Wasserstein geodesic we also have $\partial_t\rho_t|_{t=0} = -\nabla\cdot\,(\rho v_0^{\mathrm{OT}})$, hence

$$-\nabla\cdot\,(\rho v_0^{\mathrm{OT}}) = \frac{1}{2}\Delta\rho. \tag{26}$$

The vector field $v(x) = -\frac{1}{2}\nabla\log\rho(x)$ satisfies this identity, because $-\nabla\cdot\,(\rho(-\frac{1}{2}\nabla\log\rho)) = \frac{1}{2}\Delta\rho$. By elliptic uniqueness in the weighted space $L^2(\rho)$, one concludes

$$v_0^{\mathrm{OT}} = -\tfrac{1}{2}\nabla\log\rho = v_0^{\mathrm{heat}}. \tag{27}$$

We next expand the Brenier map T_ε, which solves the Monge optimal transport problem between ρ and ρ_ε for quadratic transport cost. Writing its convex potential as $\varphi_\varepsilon(x) = \frac{1}{2}|x|^2 + \varepsilon\psi_\varepsilon(x)$ we have

$$T_\varepsilon(x) = x + \varepsilon\nabla\psi_\varepsilon(x). \tag{28}$$

The pushforward relation $\rho(x) = \rho_\varepsilon(T_\varepsilon(x))\det(\nabla T_\varepsilon(x))$, together with the expansions $\rho_\varepsilon = \rho + \frac{\varepsilon}{2}\Delta\rho + O(\varepsilon^2)$ and $\det(I + \varepsilon\nabla^2\psi_\varepsilon) = 1 + \varepsilon\,\mathrm{tr}(\nabla^2\psi_\varepsilon) + O(\varepsilon^2)$, yields

$$-\nabla\cdot\,(\rho\nabla\psi_\varepsilon) = \tfrac{1}{2}\Delta\rho + r_\varepsilon, \qquad \|r_\varepsilon\|_{H^{-1}} = O(\varepsilon). \tag{29}$$

Letting $\varepsilon \downarrow 0$ gives the limit equation

$$-\nabla\cdot\,(\rho\nabla\psi_0) = \tfrac{1}{2}\Delta\rho. \tag{30}$$

Choosing $\psi_0 = -\frac{1}{2}\log\rho$ solves this exactly, since $\rho\nabla\psi_0 = -\frac{1}{2}\nabla\rho$. By Lax–Milgram theorem Evans [2022] in the weighted space with weight ρ and standard elliptic regularity,

$$\|\nabla(\psi_\varepsilon - \psi_0)\|_{L^2(\rho)} = O(\varepsilon). \tag{31}$$

Therefore

$$\|T_\varepsilon - (\mathrm{Id} - \tfrac{\varepsilon}{2}\nabla\log\rho)\|_{L^2(\mu)} \le \varepsilon\|\nabla(\psi_\varepsilon - \psi_0)\|_{L^2(\rho)} + O(\varepsilon^2) = O(\varepsilon^2), \tag{32}$$

which is the first claimed estimate, and it makes explicit that the Brenier map is approximated by the gradient of the potential $-\frac{1}{2}\log\rho$.

We finally compare the interpolating curves $t \mapsto \mu_t$ and $t \mapsto \mu_t^{\mathrm{OT}}$. Let π_t be an optimal coupling between μ_t and μ_t^{OT}. A standard estimate for absolutely continuous curves in $(\mathcal{P}_2, \mathsf{W}_2)$ gives

$$\frac{d}{dt}\frac{1}{2}\mathsf{W}_2^2(\mu_t, \mu_t^{\mathrm{OT}}) \le \int_{\mathbb{R}^d\times\mathbb{R}^d} (x-y)\cdot\big(v_t^{\mathrm{heat}}(x) - v_t^{\mathrm{OT}}(y)\big)\,d\pi_t(x,y). \tag{33}$$

Using Cauchy–Schwarz and splitting $v_t^{\mathrm{OT}}(y) - v_t^{\mathrm{OT}}(x)$ yields

$$\frac{d}{dt}\mathsf{W}_2(\mu_t, \mu_t^{\mathrm{OT}}) \le \|v_t^{\mathrm{heat}} - v_t^{\mathrm{OT}}\|_{L^2(\mu_t)} + L\,\mathsf{W}_2(\mu_t, \mu_t^{\mathrm{OT}}), \tag{34}$$

where L is a uniform Lipschitz bound for v_t^{OT} on $[0,\varepsilon]$, available by smoothness of ρ and regularity of OT maps. To bound the first term, note that $v_t^{\mathrm{heat}} = -\frac{1}{2}\nabla \log \rho - \frac{t}{4}\nabla((\Delta\rho)/\rho) + O(t^2)$, while the Eulerian OT velocity near $t = 0$ is obtained by pulling back the constant Lagrangian velocity $(T_\varepsilon - \mathrm{Id})/\varepsilon = \nabla\psi_\varepsilon$ along the interpolation map $x \mapsto (1 - \frac{t}{\varepsilon})x + \frac{t}{\varepsilon}T_\varepsilon(x)$. Since this map equals $\mathrm{Id} + O(t)$ in C^1, composition perturbs $L^2(\mu_t)$-norms by $O(t)$ and, using $\|\nabla\psi_\varepsilon - \nabla\psi_0\|_{L^2(\rho)} = O(\varepsilon)$ with $t \le \varepsilon$, one obtains

$$\|v_t^{\mathrm{heat}} - v_t^{\mathrm{OT}}\|_{L^2(\mu_t)} \le C_0\, t \qquad \text{for } t \in [0,\varepsilon]. \tag{35}$$

Applying Grönwall's inequality Gronwall [1919] gives

$$\mathsf{W}_2(\mu_t, \mu_t^{\mathrm{OT}}) \le C_0 \int_0^t e^{L(t-s)} s\, ds \le \frac{C_0}{2} e^{Lt}\, t^2 \le C\, t^2, \tag{36}$$

and hence

$$\sup_{t\in[0,\varepsilon]} \mathsf{W}_2(\mu_t, \mu_t^{\mathrm{OT}}) \le C\, \varepsilon^2. \tag{37}$$

Together with the $L^2(\mu)$ estimate for T_ε above, this proves the theorem and, in particular, makes explicit that the approximation is by the gradient field $-\frac{1}{2}\nabla \log \rho$. □

Theorem B.2 (Hierarchical Emergence of Phase Boundaries, informal)**.** *Let p_T be a high-noise Gaussian prior and p_0 be the target data distribution. Consider the discrete approximation of the probability-flow ODE given by a sequence of optimal transport maps $(T_k)_{k=1}^N$, where $T_k : \mathbb{R}^d \to \mathbb{R}^d$ is the Brenier map pushing p_{t_k} to $p_{t_{k-1}}$.*

1. ***(Existence and Regularity)** For each k, the map T_k exists, is unique, and is the gradient of a convex potential, $T_k = \nabla\phi_k$. Away from a singular set of measure zero, ϕ_k is smooth.*
2. ***(Formation of Singularities)** A singular set $S_k \subset supp(p_{t_k})$ for the potential ϕ_k emerges if the domain $supp(p_{t_{k-1}})$ is not convex with respect to the mapping from $supp(p_{t_k})$. Such a condition arises naturally during mode-splitting or due to non-convexity of the support of data distribution.*
3. ***(Propagation and Accumulation)** Let $\mathcal{T}_{k\to 0} = T_1 \circ T_2 \circ \cdots \circ T_k$ be the composed map from time t_k to t_0. The total singular set in the latent space at time T, corresponding to all phase boundaries, is given by the union of preimages:*

$$S_{\mathrm{total}} = \bigcup_{k=1}^{N} (\mathcal{T}_{N\to k})^{-1}(S_k). \tag{38}$$

The complexity (number of singularities) of this set is non-decreasing as $t \to 0$.

Proof. The existence and uniqueness of each local map $T_k = \nabla\phi_k$ is a direct consequence of Brenier's theorem Brenier [1991], Villani et al. [2009]. Since the diffusion process ensures that p_t is a smooth, strictly positive density for all $t > 0$, the conditions for the theorem are satisfied at each step $p_{t_k} \to p_{t_{k-1}}$.

The regularity of the potential ϕ_k is governed by Caffarelli's regularity theory for the Monge-Ampère equation Caffarelli [1992], Figalli [2010]. The theory states that if the densities p_{t_k} and $p_{t_{k-1}}$ are bounded above and below on uniformly convex domains, then the potential ϕ_k is globally $C^{2,\alpha}$. However, in generative processes, we are interested in the case where $p_{t_{k-1}}$ develops a complex, non-convex support, for example, when a single mode of p_{t_k} splits into two distinct modes in $p_{t_{k-1}}$. In this scenario, the convexity condition on the target domain is violated. The theory then implies that the potential ϕ_k is no longer globally smooth. A singular set $S_k \subset \mathrm{supp}(p_{t_k})$ emerges, corresponding precisely to the preimage of the interface where the convexity of the target support fails. This set S_k is where ϕ_k fails to be strictly convex, and it constitutes the phase boundary at step k.

We now consider the propagation of these singularities. Let $\mathcal{T}_{N\to k} : \mathbb{R}^d \to \mathbb{R}^d$ denote the composed map from the initial prior space to the space at time t_k, $\mathcal{T}_{N\to k} = T_k \circ \cdots \circ T_N$. Each map T_j is a diffeomorphism away from its singular set S_j. The singular set observed in the final generated sample (at t_0) is the union of the images of all singular sets formed throughout the trajectory. Correspondingly, the set of points in the initial latent space p_T that will eventually pass through a singularity is the

union of the preimages. The set of points in the domain of p_{t_k} that lead to a singularity at a future step $j < k$ is given by the preimage $(\mathcal{T}_{k\to j})^{-1}(S_j)$. The total set of phase boundaries S_{total} as measured in the prior space is therefore the accumulation of all such preimages.

Finally, the emergence of these boundaries is monotonic and hierarchical. At high noise levels (large t_k), p_{t_k} is very smooth and close to a Gaussian. The maps T_k are smooth, and few, if any, singularities appear. These early-stage singularities correspond to coarse, large-scale features of the data distribution. As $t_k \to 0$, p_{t_k} progressively incorporates finer details from the target p_0, leading to more frequent mode-splitting and shape complexity. This necessarily creates new singular sets. Since singularities, once formed, are propagated backward in time via composition, the number of singular interfaces is non-decreasing. This establishes a hierarchical structure where phase transitions corresponding to fine-grained details appear at later stages (lower noise levels) of the reverse process. This completes the proof. □