THEORY OF OUTCOME BOUNDARY PREDICTION IN A DETERMINISTIC DIFFUSION ODE

Anonymous authors
Paper under double-blind review

ABSTRACT

Are there patterns in the unstructured noise from which a diffusion samples? Surprisingly, deterministic ODE samplers impose structure in the noise space. When the data distribution consists of separable gaussian clusters, a deterministic ODE-based sampler partitions the sampling noise based on the class the noise belongs to. The boundaries that separate these clusters can be mathematically characterized to represent decision thresholds in the noise space where the score field transitions between distinct modes, causing nearby seeds to produce ambiguous or hallucinated outputs. We formalize this correspondence and introduce a method that traces this mapping, which we call *pullback*. Using *pullback*, we can predict hallucinations, or classes in any mixture-of-Gaussian setting.

1 INTRODUCTION

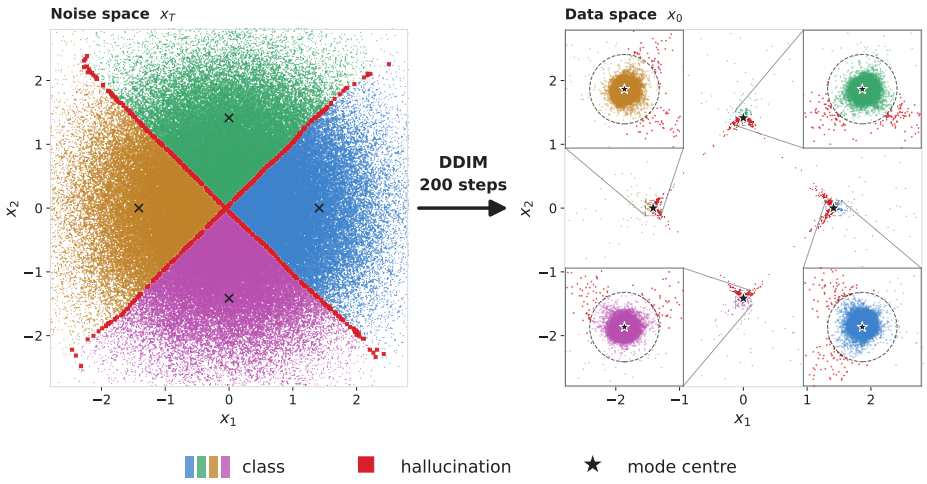


Figure 1: A demonstration of hallucinated lines (in red) and mode basins (assuming 4 modes) before and after running a determinstic DDIM. The colored points indicate samples reaching the true mode. The noise space (left) partitions the samples very clearly, enabling us to separate the hallucinated seeds from those that converge to a true mode and even predict the mode it would converge to.

ODE-based diffusion samplers have been strongly preferred over SDE samplers in recent years, owing to their sample quality and their lower cost at inference (Song et al., 2021a). Ashiq et al. (2026) prove analytically that the trade-off for this efficiency is higher hallucination rate, i.e., producing samples that fall between the modes of the data distribution (Aithal et al., 2024).

In this work, we advocate that the trade-off is more nuanced. Concretely, we show that hallucination in deterministic ODE-based samplers is surprisingly predictable. We demonstrate this for a mixture-of-Gaussian setup adopted from Aithal et al. (2024) in Fig. 1. The points that result in hallucinations and the points that converge to the modes are not scattered across the noise space. Rather, they

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

occupy contiguous regions. Furthermore, we observe that the same structure appears under different ODE-based samplers, suggesting that the structure is a property of the deterministic map rather than of a particular deterministic sampler.

Is this structure in the noise space accidental? No. We show this by demonstrating its existence for a deterministic ODE sampler. Because the reverse process is invertible, the mode into which a sample will approach is determined by its initialization and can be recovered by clustering the noise space alone, without running the reverse process. We call the the pre-image function that maps from the data space to the noise space *pullback*. For a gaussian mixture, this endpoint map takes the form of the Kim–Milman heat-flow map (Kim & Milman, 2012). The structure can be connected to phase boundaries in the noise space observed by Lobashev et al. (2025).

This structure has three consequences that we develop in the remainder of the paper:

- The class outcome for ODE-based methods can be predicted early. The invertible map can define the class boundaries in sampling noise.
- The normal can be predicted for the samples belonging to the boundary of each class. This predicted normal can be used to correct samples that were originally hallucinations.
- Having constructed the boundary, unconditional diffusion models can be trained and then sampled class by class, without access to a single label.

2 Related Work

Outcome structure in the initial noise. Initial noise contains predictive information for generation, which can be used for seed-selection to improve rare-concept generation or sample quality (Samuel et al., 2024; Xu et al., 2025; Li et al., 2025). Ban et al. (2025) show that initial noise can predict the locations of objects in the final image. Class-specific organization exists in the seed and can be concentrated in small structured directions. Yan et al. (2025) show that semantic variation is concentrated along a small number of structured directions in noise space. These works recover the structure after the fact, from generated samples and their labels, whereas its origin and its evolution along the reverse process, which we characterize here, have been left open. Wei et al. (2026) demonstrate such a classification exists using a learnt classifier, assigning varying levels of confidence to noise samples using the classifier score. We detail the distinction with their work in Sec. B.4.

Commitment of outcomes during denoising. A prominent line of research analyzes how generative outcomes crystallize dynamically along the reverse trajectory. Aithal et al. (2024) make this concrete studying the geometry between well separated data modes. Ashiq et al. (2026) build on this setting to show that there is a critical time after which the local dynamics become dominated by a nearby pair of modes, determining their eventual fate as hallucinations. Broadly, these observations fit into the theory of progressive commitment: Raya & Ambrogioni (2023) describe denoising through symmetry breaking, Biroli et al. (2024) identify a speciation transition toward coarse data structure. These works study trajectory signatures through the behaviour of reverse trajectories at different noise scales (Li & Chen, 2024; Sclocchi et al., 2025; Achilli et al., 2026), whereas we use the invertible map to analyse the seed directly and predict the outcome class from it.

Hallucination detection and mitigation. Prior work detects and corrects hallucinations after sampling has begun. Aithal et al. (2024) identify mode interpolation as a source of unsupported generations and detect them later in the reverse process. RODS (Tian et al., 2025) use the curvature of the learned score to flag unstable regions that are prone to hallucinations. Other methods modify the reverse trajectory itself: Laplacian Score Sharpening (Chandran C et al., 2025) suppresses interpolation by sharpening the score, and Dynamic Guidance (Triaridis et al., 2025) applies such corrections selectively to directions associated with invalid generations. HEaD+ (Betti et al., 2026) instead predicts failure from an incomplete trajectory and restarts generation. These methods differ in whether they detect or correct failures, but they all require observing a portion of the trajectory. We ask whether outcome regions can be identified directly in the initial-noise space, screening candidate seeds before generation begins.

Outlier detection with generative models. Identifying samples outside the mode cores is related to out-of-distribution (OOD) detection. Deep generative models can assign higher likelihood to OOD inputs than to training data (Nalisnick et al., 2019a), motivating typicality tests (Nalisnick et al., 2019b) and methods that combine likelihood with local intrinsic dimension (Kamkari et al., 2024). Diffusion-based detectors use reconstruction errors after noising or masking an input (Graham et al., 2023; Liu et al., 2023), score norms across noise scales (Mahmood et al., 2021), the rate of change and curvature of diffusion paths (Heng et al., 2024), and the spectrum of the denoising posterior covariance (Shoushtari et al., 2026). These methods assess a given sample through its response to noising and denoising. Closely related, Sobieski et al. (2026) connect structural hallucinations to local intrinsic dimension and introduce a correction during sampling. Our analysis characterizes the initial seeds whose outputs lie outside the mode cores by pulling the terminal boundaries back through the generative flow.

3 Problem Setup

Notation: For $n \in \mathbb{N}$, let $[n] := \{1, \dots, n\}$, and let d denote the ambient dimension. Bold lowercase letters denote vectors in $\mathbb{R}^d$, while unbolded letters denote scalars. We write $\langle \cdot, \cdot \rangle$ and $\| \cdot \|_2$ for the Euclidean inner product and norm, respectively, and $\boldsymbol{I}_d$ for the $d \times d$ identity matrix. For $\boldsymbol{c} \in \mathbb{R}^d$ and $a \geq 0$, let $B(\boldsymbol{c}, a) := \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{c}\|_2 \leq a\}$. We write $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the Gaussian density evaluated at $\boldsymbol{x}$. Gradients of scalar-valued functions are denoted by ∇, while Df denotes the Jacobian of a vector-valued map f. A complete list of notation is given in Sec. A, with a symbol table in Sec. I.

Deterministic generative transport: Consider a deterministic generative processes represented by a time-dependent vector field $\boldsymbol{F} : \mathbb{R}^d \times [0, 1] \to \mathbb{R}^d$, where $\tau = 1$ indicates the sample belongs to the initial noise and $\tau = 0$ indicates the sample belongs to the data distribution. Given a latent code $\boldsymbol{z} \sim \gamma := \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, the trajectory satisfies

$$\boldsymbol{X}_\tau = \boldsymbol{z} - \int_\tau^1 \boldsymbol{F}(\boldsymbol{X}_s, s)\, ds, \qquad \tau \in [0, 1]. \tag{1}$$

This formulation is kept general: it can be used to model probability-flow ODE(Song et al., 2021b) and flow-based generative models(Lipman et al., 2023). We call the function $\boldsymbol{F}$ the *sampling field*. We denote the flow map by $\Phi^{\boldsymbol{F}}_{\tau_a \to \tau_b}$ and its endpoint map by $\mathcal{G}_{\boldsymbol{F}} := \Phi^{\boldsymbol{F}}_{1 \to 0}$. The sampling field and endpoint map are denoted by $\boldsymbol{F}_\star$ and $\mathcal{G}_\star$, while $\boldsymbol{F}_\theta$ and $\mathcal{G}_\theta$ denote their learned counterparts.

A Gaussian mixture target. Consider a mixture of isotropic gaussian with weights w_i, μ_i and a covariance matrix $\sigma_0^2 \boldsymbol{I}_d$. Under a PF-ODE in Eq. (1), with schedule $\{\alpha_i\}_{i=1}^T$

$$\frac{d\boldsymbol{Y}_\sigma}{d\sigma} = \frac{\boldsymbol{Y}_\sigma - \boldsymbol{D}_\sigma(\boldsymbol{Y}_\sigma)}{\sigma}, \qquad \sigma : \sigma_a \downarrow \sigma_0, \tag{2}$$

where $\sigma(t)^2 = \sigma_0^2 \bar{\alpha}_t + (1 - \bar{\alpha}_t)$, $\boldsymbol{Y}_t := \boldsymbol{X}_t / \sqrt{\bar{\alpha}_t}$ and $\boldsymbol{D}_\sigma(\boldsymbol{y}) := \sum_i \pi_i(\boldsymbol{y}, \sigma) \boldsymbol{\mu}_i$ with $\sum_{i=1}^C \pi_i(\boldsymbol{y}, \sigma) = 1$. The term $\sigma : \sigma_a \downarrow \sigma_0$ indicates a noise scale that decreases during generation. We explain the full derivation of this reparameterization in Sec. B.

4 Theory of Pullback

Our analysis is focused on the commitment phase ($t < \tau$) of a diffusion trajectory to one of its C training classes. For each class $i \in C$, we define a spherical region called a ***mode core***, where diffusion is governed primarily by samples belonging to c. In data space, the mode core is centered at the class mean and has a size given by *critical radius* r_i. On applying pullback the ***mode core*** of radius r_i the boundary of the class can be generated.

To illustrate the formation of a mode core, we visualize the diffusion trajectory on a two mode gaussian mixture in Fig. 2 for T steps. Fig. 2(a) demonstrates 3 data samples lying on 3 confidence shells of $\{3\sigma, 4\sigma, 5\sigma\}$. The origins of these shells are semi-elliptical structures shown in Fig. 2(b). Fig. 2(c) highlights a snapshot at $\tau = 0.067T$, when the core becomes identifiable.

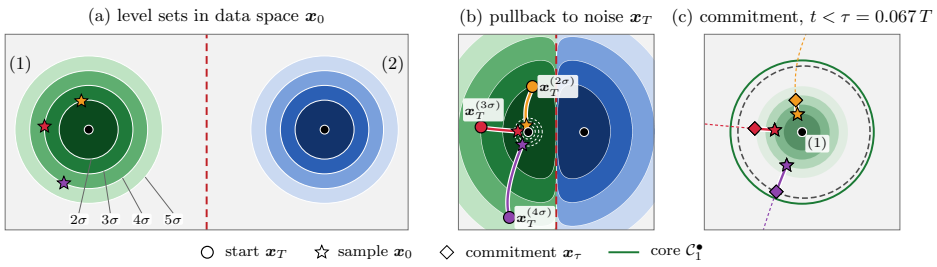


Figure 2: We demonstrate the motion of 3 trajectories, denoted by red, purple and orange. Figure (a) shows the samples in $\boldsymbol{x}_0$ and their confidence shells in the data space. Figure (b) demonstrates the result of applying *pullback* to the samples and their confidence shells. Figure (c) highlights the specific time τ at which the mode core C_i becomes clearly identifiable as a spherical ball.

4.1 Can we characterize the mode core?

The reverse process can be viewed in two separate phases. In the early stages, the trajectory is influenced by competing class-dependent contributions. Convergence to a specific class region emerges later Aithal et al. (2024); Achilli et al. (2026). This can be made explicit as the convergence to a specific class c, over all other training classes $\{1, \dots C\} - \{c\}$ and governed by the class mean $\boldsymbol{\mu}_c$ and its covariance $\boldsymbol{\Sigma}_c$. Together, they define the Gaussian approximation $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$ of class c.

More generally, consider $f_{c,0}$ to denote the population density constructed by the samples of class c. For $\eta > 0$, $f_{c,\eta} := f_{c,0} * \mathcal{N}(\mathbf{0}, \eta^2 \boldsymbol{I}_d)$ for its Gaussian smoothing, and $g_{c,\eta} := \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c + \eta^2 \boldsymbol{I}_d)$ for the reference at the same scale. We now argue the existence a the mode core and its size assuming an isotropic gaussian approximation $\boldsymbol{\Sigma}_c = \sigma_0^2 \boldsymbol{I}_d$.

Proposition 1 (Single Mode Dominance). *Suppose $f_{c,0}$ is positive and C^1, with $\|\boldsymbol{\Sigma}_c \nabla \log f_{c,0}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\mu}_c)\|_2 \le \kappa_c < \infty$ for every $\boldsymbol{x} \in \mathbb{R}^d$, and let $\lambda_{\min}(\boldsymbol{\Sigma}_c)$ denote the smallest eigenvalue of $\boldsymbol{\Sigma}_c$. Then, for every $\eta \ge 0$,*

$$\|\nabla \log f_{c,\eta}(\boldsymbol{y}) - \nabla \log g_{c,\eta}(\boldsymbol{y})\|_2 \le \frac{\kappa_c}{\lambda_{\min}(\boldsymbol{\Sigma}_c) + \eta^2}, \qquad \boldsymbol{y} \in \mathbb{R}^d. \tag{3}$$

Remark. The proposition above establishes a relationship between the score generated by $f_{c.\eta}$ and its gaussian reference $g_{c,n}$ with κ_c modulating the strength of the relationship. Along the principal axis of Σ_c, the relationship is governed by the covariance eigenvalues λ_k. The eigenvector corresponding to the smallest eigenvalue $\lambda_{\min}$ sets the maximum discrepancy and $\lambda_{\max}$ the least. For an isotropic gaussian, the bound is a departure from radial motion towards $\boldsymbol{\mu}_c$. The full derivation for the proposition can be found in Sec. C.

Proposition 2 (Critical Radius for Single Mode dominance). *For a tolerance $\delta \in (0, 1/2)$, mode i has responsibility at least $1 - \delta$ throughout the closed ball $B(\boldsymbol{\mu}_i, r_i(\sigma; \delta))$, provided $r_i(\sigma, \delta) > 0$:*

$$r_i(\sigma; \delta) := \min \left\{ \frac{d_i}{2}, \left[\frac{d_i}{2} - \frac{\sigma^2}{d_i} \log\left(\frac{(1-\delta)(1-w_i)}{\delta w_i} \right) \right]_+ \right\}, \qquad d_i := \min_{j \ne i} d_{ij}, \tag{4}$$

where $[a]_+ := \max\{a, 0\}$. Within this ball, the velocity differs from the radial field $(\boldsymbol{y} - \boldsymbol{\mu}_i)/\sigma$ by at most $\delta \max_{j \ne i} d_{ij}/\sigma$ guaranteeing a spherical level set centered at μ_i.

Proof Sketch. $\nabla \log g_{c,\eta}(\boldsymbol{y})$ is dominated by a of pull towards class means $\sum_j \pi_j (y - \boldsymbol{\mu}_j)/\sigma$. During single mode dominance by i, departure of $\pi_i (y - \boldsymbol{\mu}_i)/\sigma$ from $\sum_j \boldsymbol{\pi}_j (y - \boldsymbol{\mu}_j)/\sigma$ is bounded. The complete derivation is given in Sec. D.

Discussion. The radius $r_i(\sigma, \delta)$ is the minimum required separation between a mode core and its nearest neighbor, d_i. Smaller δ will require a stricter radius. Denoising can only enlarge the radius upto a ceiling of $d_i/2$. This definition makes the distinction between neighboring balls clear.

The critical radius is largest at the data endpoint, where $\sigma = \sigma_0$. Although added noise has been removed, the within-class spread remains, and competition from neighboring modes can still limit the radius below $d_i/2$. When the radius is zero, the bound cannot guarantee one-mode dominance

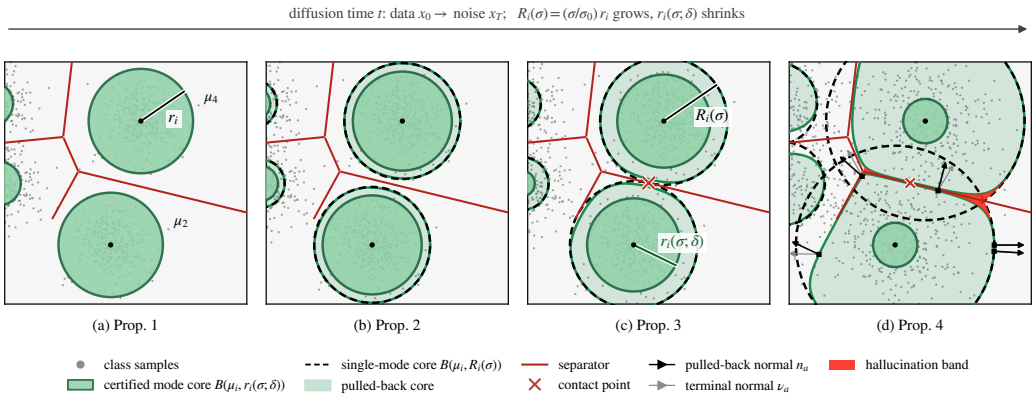


Figure 3: Mode cores traced from data toward noise in fixed-center coordinates; grey dots show class samples at each scale. (a) At the data endpoint, each core is a ball $\mathcal{C}_i^{\bullet} = B(\boldsymbol{\mu}_i, r_i)$. (b) Its single-mode reference boundary expands as $R_i(\sigma) = (\sigma/\sigma_0)r_i$, while the radius $r_i(\sigma;\delta)$ certifying dominance decreases (Prop. 2). (c) The reference spheres first touch at the midpoint on their equal-responsibility separator (Prop. 3). (d) Under the mixture flow, the pulled-back cores flatten near the separator, and their normals $\boldsymbol{n}_a$ differ from the terminal normals $\boldsymbol{\nu}_a$ (Prop. 4). The band between them represents seeds that generate samples outside the certified cores. These are the samples that become hallucinations

throughout a neighborhood at the chosen tolerance δ. Other modes may still contribute appreciably, so a single-mode description requires further justification at that noise level. These cases are detailed in Lemma 4.

4.2 Class Boundaries are constructed by an interplay between mode cores

Our focus till now has been to demonstrate the existence of the mode core. Now, we explore the interplay between a mode i and its neighboring class j as a function of class variance $\sigma(t)^2\boldsymbol{I}$. which creates its boundaries. Intuitively, consider Fig. 3 where arbitrarily placed modes are expanding. We illustrate first contact, and the subsequent expansion of regions during pullback.

Assuming that a single mode is dominant, the mode core C_i can be traced using *pullback*. The radius of the mode cores for classes, $\{i, j\}$ expand as $\{R_i(\sigma), R_j(\sigma)\} = \{(\sigma(t)/\sigma_0)r_i, (\sigma(t)/\sigma_0)r_j\}$. First contact, t_{ij} identifies where the two classes $\{i, j\}$ intersect. We explain the full derivation for the proposition in (Sec. D.2).

Proposition 3 (End of Single Mode Dominance)**.** *Consider an equally weighted pair (i, j) with separation d_{ij} and common endpoint critical radius $0 < r < d_{ij}/2$. As noise increases, the two single-mode reference boundaries first meet at time t_{ij} satisfying*

$$\bar{\alpha}_{t_{ij}} = \left[1 + \sigma_0^2 \left(\frac{d_{ij}^2}{4(r)^2} - 1\right)\right]^{-1}, \tag{5}$$

provided this time lies within the diffusion schedule. The contact point is given by $\sqrt{\bar{\alpha}_{t_{ij}}}\boldsymbol{m}_{ij}$. This hyperplane passes through this point is pierced by the line joining the two modes i and j.

Proof Sketch. First contact occurs when the sum of the reference radii equals the separation, $2(\sigma/\sigma_0)r = d_{ij}$. Substituting $\sigma^2 = \sigma_0^2 + (1 - \bar{\alpha}_t)/\bar{\alpha}_t$ gives Eq. (5). If $w_i = w_j$ the point of contact is the midpoint $(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)/2$, with opposing normals along the line joining the centers. Their common tangent is the perpendicular bisector, on which the competing modes have equal responsibility. The complete derivation is given in Sec. E.

Discussion. For $t < t_{ij}$, the reference regions are disjoint; at t_{ij}, they touch at the midpoint; and for $t > t_{ij}$, they overlap around the same separator. First contact therefore occurs where the two modes have equal responsibility, connecting the radial geometry around each center to the region of

competition between them. Fig. 3 illustrates this contact and the deformation of the cores under the full mixture flow, whose invertibility preserves their disjointness.

When remaining modes are sufficiently distant from this region, their responsibilities are exponentially suppressed and the local dynamics are approximated by those of the pair. Now, we describe the recovery of the normal and tangent created by pullback in original noise.

4.3 Perturbation in Class Boundaries Translates samples

The mode core $g_{c,0}$ has outward unit normal $\boldsymbol{\nu}_a := (\boldsymbol{x}_a - \boldsymbol{\mu}_i)/r_i$ and a tangent projection $P_a := \boldsymbol{I}_d - \boldsymbol{\nu}_a\boldsymbol{\nu}_a^\top$. The deformed mode core $g_{c,\eta}$ cannot be computed in closed form as it relies on J_a, the Jacobian of the diffusion process $\mathcal{G}_F$. Instead we characterize how perturbations on $g_{c,\eta}$ affect $g_{c,0}$. In Sec. 6.3, we characterize how these perturbations can be used to mitigate hallucinations.

Consider an anchor point $\boldsymbol{z}_a$ on $g_{c,\eta}$. We record core membership through $h_i^{\boldsymbol{F}}(\boldsymbol{z}) := \|\mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z}) - \boldsymbol{\mu}_i\|_2^2 - r_i^2$. Its sign indicates whether the generated sample lies inside or outside the core, while its magnitude measures its distance from the terminal boundary. The corresponding membership indicator is $\mathbb{I}(h_i^{\boldsymbol{F}}(\boldsymbol{z}) < 0)$.

Proposition 4 (Class Translation in Sampling Noise)**.** *Among unit seed perturbations, the inward normal* $-\boldsymbol{n}_a := -\frac{J_a^\top \boldsymbol{\nu}_a}{\|J_a^\top \boldsymbol{\nu}_a\|_2}$ *produces the greatest first-order decrease in terminal radius. The normal perturbations satisfy, as* $\varepsilon \to 0^+$*,*

$$h_i^{\boldsymbol{F}}\left(\boldsymbol{z}_a \pm \varepsilon \boldsymbol{n}_a\right) = \pm 2 r_i \left\| J_a^\top \boldsymbol{\nu}_a \right\|_2 \varepsilon + o(\varepsilon). \tag{6}$$

Proof Sketch. The chain rule can be applied to $\nabla h_i^{\boldsymbol{F}}(\boldsymbol{z}_a) = \boldsymbol{n}_a = 2 r_i J_a^\top \boldsymbol{\nu}_a$, to produces the pullback normal. The complete proof is given in Sec. F.

Remark. For $\langle \boldsymbol{n}_a, \boldsymbol{v} \rangle = 0$, $h_i^{\boldsymbol{F}}(\boldsymbol{z}_a + \varepsilon \boldsymbol{v}) = o(\varepsilon)$ and $\mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z}_a + \varepsilon \boldsymbol{v}) = \boldsymbol{x}_a + \varepsilon P_a J_a \boldsymbol{v} + o(\varepsilon)$. Thus the output moves in the terminal tangent space, preserving its radius to first order.

Discussion. Prop. 4 offers the most efficient direction for entering the **certified mode core** in sampling noise. The transport scales linearly with $\boldsymbol{r}_i$, $\boldsymbol{n}_a$ remains close to the terminal normal $\boldsymbol{\nu}_a$, so moving the seed along $-\boldsymbol{\nu}_a$ remains effective.

Remark. $-\boldsymbol{n}_a$ provides a local benchmark for comparing seed correction methods by measuring the perturbations efficiency relative to by defining the operator $\mathrm{Align}_a(\boldsymbol{v}) := \cos \angle(\boldsymbol{v}, -\boldsymbol{n}_a)$.

5 The Pullback Algorithm

Let us now translate our theory into actionable insights:

- Samples at the center closer to the mode core are more likely to belong to the true mode and samples at dominance radiusare more likely to hallucinate. When the sample has converged to a specific class c, the critical radius r_i is exactly equal to R_{99}.
- Samples close to the class mean are more likely to belong to the corresponding mode in the data space. The point that are equidistant to multiple modes belong to regions of low confidence where we cannot predict the corresponding mode in the data space. Therefore, a large number of samples must be generated in the low-confidence regions of the space to effectively characterize it.
- The normal direction of perturbation in sampling noise can result in the greatest translation in data space. This can be used to correct the outcome of samples that were originally hallucinating.

We first present our pullback algorithm. Based on the insights from this algorithm, we sample N points within R_{99} of every mode and label each point with the mode it lies within. We then sample $K/2$ points in the shell extending from R_{99} to 2σ, where σ^2 is the isotropic variance of each Gaussian in the mixture, and label these points as hallucinations. By discretizating the reverse ODE and integrating it from each point back to x_T, we track where each point originates in the noise space, and hence characterize where points belonging to a particular mode or to the hallucination

Algorithm 1 Pullback: characterize the seed space by transporting labeled data-space anchors back under the exact field F_*.

```
# mu[0..K-1] : mode centres, R99 : mode-ball radius
# n, m : ball and band anchors per mode, w : band width
# F_star : exact field, T : steps
# labels : k = mode k, -1 = hallucination

# 1. labeled anchors in data space
P, y = [], []
for k in range(K):
    u = uniform(n) ** (1 / d)
    P += [mu[k] + R99 * u * unit_vectors(n, d)]
    y += [k] * n
    r = R99 + w * uniform(m)
    P += [mu[k] + r * unit_vectors(m, d)]
    y += [-1] * m
P, y = concat(P), concat(y)

# 2. transport the anchors back to the seed space (Heun)
Z, dt = P, 1 / (T - 1)
for tau in time_grid(T)[:-1]:
    k1 = F_star(Z, tau)
    k2 = F_star(Z + dt * k1, tau + dt)
    Z = Z + 0.5 * dt * (k1 + k2)

# 3. labeled atlas of the seed space
atlas = (Z, y)
g = fit_classifier(Z, y)

return g
```

category come from. This gives a mapping from initial noise to a label that determines its outcome (either a specific mode or a hallucination). We use this mapping to train several classifiers that give a general characterization of the initial noise space, described in detail in Appendix G.

6 Experiments

In this section, we evaluate the performance of pullback on Gaussian mixtures up to 512 dimensions and extend it to the MNIST and Fashion-MNIST dataset. We first outline the experimental setup, and then follow it with targeted questions to highlight our algorithm's capabilities and limitations.

6.1 Experimental Setup

We choose pairs of $(C, d) \in \{2, 4, \dots 32\} \times \{2, 4, \dots 512\}$ and build a mixture of C isotropic Gaussian in $\mathbb{R}^d$. Rejection sampling is conducted to ensure the spacing between modes is: $\|\mu_i - \mu_j\| > 3R_{99}$, where R_{99} represents the radius around a gaussian mode μ_i. For image experiments, this constraint is enforced mildly using a separation penalty to ensure distinct classes are separated. Each image sample is embedded in a 32 dimensional latent space using the encoder. We sample N anchor points per mode across K modes, which leads to NK total samples. For each dimension and mode pair (d, K) we train a diffusion model on the GMM generated by $\{d, k\}$.

Anchor points transported via pullback, originate from the critical mode core, of K distinct modes. For each data sample, x, we can characterize the space using $d_p = 256$ independent directions $\boldsymbol{v}_i^T \boldsymbol{x} \; \forall \; i \in \{1, \dots d_p\}$. On this transformed space, we apply a degree 3 polynomial neural network. We refer to the neural network characterized in this form as Polar3. We use 3 baselines k-nearest neighbors (k-NN), weighted k-NN, and a degree-2 neural network. All neural networks are trained on C+1 classes, treating hallucinated samples as a distinct class. A full description of the experimental setup and classification algorithms is provided in Sec. G.

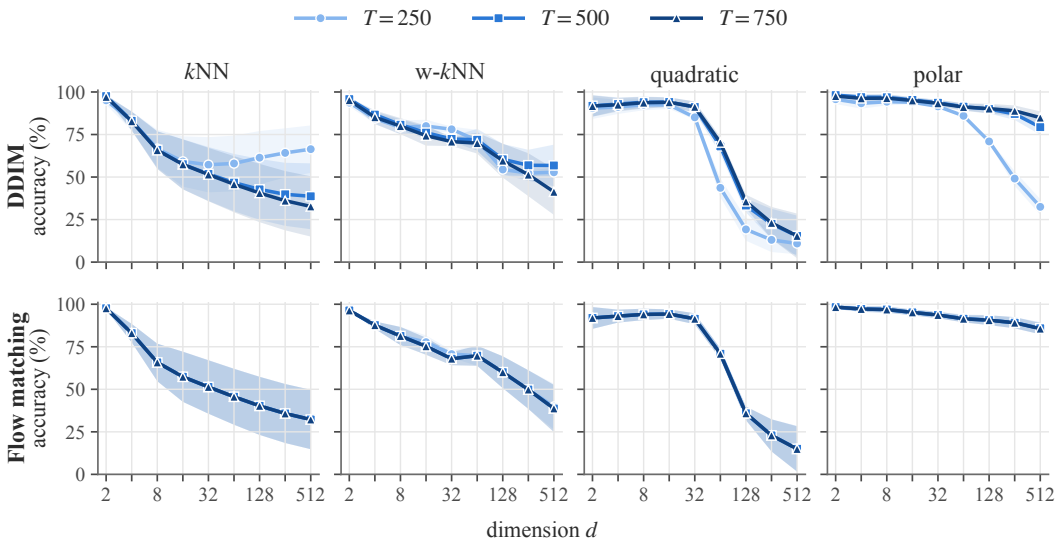


Figure 4: Accuracy of seed-space classifiers trained on the pullback atlas for DDIM (top) and flow matching (bottom), averaged over K from 2 to 16, with shading showing the spread over K. See Section 6.2 for the setup.

6.2 How effectively can the pullback describe the original structure?

For each dimension d and number of modes K from 2 to 16, we train a DDIM model and a flow matching model with 500 diffusion steps, and then run Algorithm 1 with $T \in \{250, 500, 750\}$ steps and 2,000 anchors per mode. Each classifier trained on the resulting atlas is evaluated against ground-truth labels from the learned model: we draw 50,000 seeds per mode, run the sampler, and label a seed with mode k if its sample lands within R_{99} of μ_k and as a hallucination otherwise. The atlas therefore uses 25 times fewer samples than the ground truth, and with the right classifier it still characterizes the seed space accurately (Fig. 4).

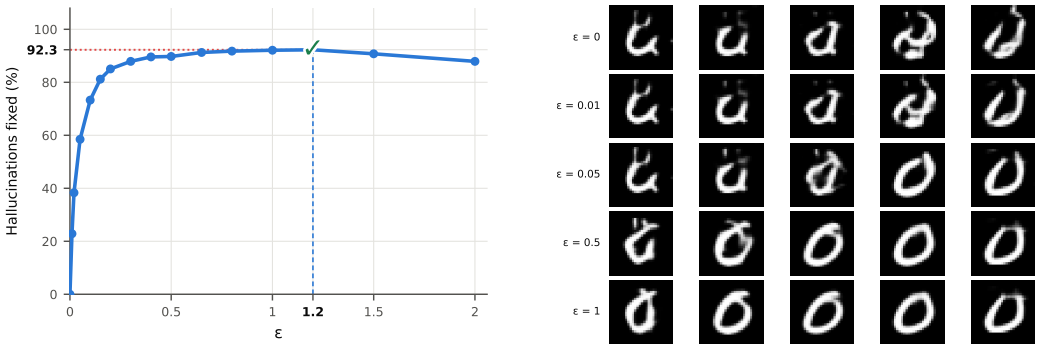
How well the atlas works depends strongly on the classifier. k-NN and weighted k-NN (w-kNN) are accurate at low d but degrade steadily as d grows. Weighted k-NN weights each anchor by its position in the data space, so anchors close to the class mean μ_c count heavily and anchors at the periphery count less. This raises accuracy to 75% at $d = 32$, but it still fails for $d > 128$, where 2,000 anchors per mode are too few for a nonparametric method to cover the space. The quadratic classifier stays near 95% up to $d = 32$ and then collapses. A degree-3 polynomial network on polar coordinates is the only classifier that holds up, staying above about 80% up to $d = 512$. For DDIM this requires $T \geq 500$; with $T = 250$ its accuracy falls off beyond $d = 128$.

Accuracy is nearly identical for DDIM and flow matching. Additional results for other ODE solvers are detailed in Sec. G.

6.3 Pullback for Hallucination Mitigation

Building on the theoretical exposition in Sec. 4.3, we demonstrate the task of hallucination mitigation. We define a hallucination as a mode interpolation, falling outside the R_{99} region of the class and on the line joining a neighboring class.

In Fig. 5, we apply a fix using the predicted normals $-\boldsymbol{n}_a := -\frac{J_a^\top \boldsymbol{\nu}_a}{\|J_a^\top \boldsymbol{\nu}_a\|_2}$ of varying magnitude ϵ to noise sample reporting the number of samples successfully fixed. In Fig. 5a, we observe that extremely small and large values of ϵ have a negative effect, with the optimal value becoming $\epsilon = 1.2$. In Fig. 5b, we demonstrate hallucination correction for digit 0 and its trajectory for varying times in the reverse process $t \in \{0, 0.25, 0.5, 0.75, 1\}$. Here, its evident that increasing the values of ϵ shows the a generation of the true sample much sooner. $\epsilon = 0.5$ produces a discretional digit at $\tau = 0.5$, in contrast to $\epsilon = 0.25$, when the structure is evident only at $\tau = 0.75$



(a) Hallucination rate vs ϵ

(b) Normal based perturbation on the MNIST dataset.

Figure 5: Hallucinations fixing on the MNIST digits. Among all the hallucinated samples given by mode interpolations 92.3% of them can be fixed by applying to $\epsilon = 1.2$ to the initial sample. Very large values of ϵ, can revisit the movement closer a boundary and thus increase hallucinations.

7 Conclusion

In this work, we study the connection between the geometry of the data distribution and the initial noise. We show that an unconditional diffusion model exhibits class-dependent structure without receiving class labels during generative training. Our method, *pullback*, makes this structure accessible by relating regions of seeds to their generated classes. Beyond selective sampling, this connection suggests a way to study how different models represent the same classes and where their outcomes differ. We hope that our results motivate new approaches to training and combining unconditional models, using their outcome geometry to align shared classes, retain complementary modes, and identify regions associated with unreliable generation.

8 Limitations and Future Work

Our analysis assumes well-separated Gaussian modes, and our image experiments use a structured latent representation. Natural images can contain several objects and shared attributes, so class membership alone may not distinguish valid compositions from hallucinations. Extending pullback to these distributions requires outcome regions that capture more complex properties, such as spatial relationships in images or temporal consistency in videos. These regions also depend on the sampling map; establishing correspondences between models and conditioning choices could support their reuse and model combination. Finally, stochastic samplers introduce randomness beyond the initial seed thus requiring a probabilistic description of outcome regions for these samplers. These directions would broaden pullback from class-based generation to more general control over the content and reliability of generated samples.

Ethics Statement

This work adheres to the ICLR Code of Ethics. This paper is a theoretical study of the reverse dynamics of deterministic diffusion samplers, supported by experiments on Gaussian mixtures and on MNIST-like datasets. This paper involves no human subjects, no personal data, and no newly collected dataset. The methods we develop identify, before sampling, which seeds a model carries into each mode and which lead to hallucinated samples. We see no direct path to harmful use: the same information is obtained by running the sampler, and our results reduce rather than create the ability to produce unintended outputs. We do not foresee fairness or privacy concerns beyond those inherent to the generative models the analysis is applied to.

Reproducibility Statement

All assumptions of the theory are stated in Sec. 3 and collected in Sec. A, and complete proofs of Props. 1 to 4 are given in Secs. C to F. The figures of Sec. 4 are computed from the exact mixture score and the schedule of Sec. B; the scripts that generate them, together with the mixture configurations and random seeds, are part of the supplementary material or will be provided in the camera-ready version of the paper. The Gaussian-mixture sweep of Sec. 6 is described in full in Sec. G, including the data, the score network, the samplers, the anchor placement, the three predictors, and the reported metrics, and the MNIST setup is given in Sec. H. Source code for training, sampling, and the pullback predictor is provided in the supplementary material or will be provided in the camera-ready version of the paper.

AI use statement

Large Language Models (LLMs) were utilized to assist in discovering and retrieving relevant literature, verifying proofs and mathematics, and polish overall prose for clarity and readability. All generated content, mathematical derivations, theoretical claims, and references from the literature were thoroughly reviewed, verified, and refined by human authors, who take full responsibility for the contents of this manuscript.

References

Beatrice Achilli, Marco Benedetti, Giulio Biroli, and Marc Mézard. Theory of speciation transitions in diffusion models with general class structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4):043304, 2026. doi: 10.1088/1742-5468/ae5c94.

Sumukh K. Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Muhammad H. Ashiq, Samanyu Arora, Abhinav N. Harish, Ishaan Kharbanda, Hung Yun Tseng, and Grigorios G. Chrysos. Why DDIM hallucinates more than DDPM: A theoretical analysis of reverse dynamics. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. The crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise. In *International Conference on Learning Representations*, 2025.

Federico Betti, Lorenzo Baraldi, Lorenzo Baraldi, Rita Cucchiara, and Nicu Sebe. Hallucination early detection in diffusion models. *International Journal of Computer Vision*, 134(1):35, 2026. doi: 10.1007/s11263-025-02622-0.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:9957, 2024. doi: 10.1038/s41467-024-54281-3.

Barath Chandran C, Srinivas Anumasa, and Dianbo Liu. Laplacian score sharpening for mitigating hallucination in diffusion models. *arXiv preprint arXiv:2511.07496*, 2025.

Mark S. Graham, Walter H. L. Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and M. Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2947–2956, 2023.

Alvin Heng, Alexandre H. Thiery, and Harold Soh. Out-of-distribution detection with a single unconditional diffusion model. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Hamidreza Kamkari, Brendan Leigh Ross, Jesse C. Cresswell, Anthony L. Caterini, Rahul G. Krishnan, and Gabriel Loaiza-Ganem. A geometric explanation of the likelihood OOD detection paradox. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22908–22935. PMLR, 2024.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Young-Heon Kim and Emanuel Milman. A generalization of Caffarelli's contraction theorem via (reverse) heat flow. *Mathematische Annalen*, 354(3):827–862, 2012. doi: 10.1007/s00208-011-0749-x.

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

Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. Enhancing compositional text-to-image generation with reliable random seeds. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=5BSlakturs.

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

Zhenzhen Liu, Jin Peng Zhou, Yufan Wang, and Kilian Q. Weinberger. Unsupervised out-of-distribution detection with diffusion inpainting. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*. PMLR, 2023.

Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40227–40252. PMLR, 2025.

Ahsan Mahmood, Junier Oliva, and Martin Styner. Multiscale score matching for out-of-distribution detection. In *International Conference on Learning Representations*, 2021.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019a.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using typicality. *arXiv preprint arXiv:1906.02994*, 2019b.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Walter Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 3 edition, 1976.

Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4695–4703, 2024. doi: 10.1609/aaai.v38i5.28270.

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

Shirin Shoushtari, Yi Wang, Xiao Shi, M. Salman Asif, and Ulugbek S. Kamilov. EigenScore: OOD detection using posterior covariance in diffusion models. In *International Conference on Learning Representations*, 2026.

Bartlomiej Sobieski, Matthew Tivnan, Dawid Płudowski, Michał Jan Włodarczyk, Pengfei Jin, Przemyslaw Biecek, and Quanzheng Li. Local intrinsic dimension unveils hallucinations in diffusion models. *arXiv preprint arXiv:2605.05026*, 2026.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yiqi Tian, Pengfei Jin, Mingze Yuan, Na Li, Bo Zeng, and Quanzheng Li. RODS: Robust optimization inspired diffusion sampling for detecting and reducing hallucination in generative models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kostas Triaridis, Alexandros Graikos, Aggelina Chatziagapi, Grigorios G. Chrysos, and Dimitris Samaras. Mitigating diffusion model hallucinations with dynamic guidance. *arXiv preprint arXiv:2510.05356*, 2025.

Wei Wei, Yizhou Zeng, Kuntian Chen, Sophie Langer, Mariia Seleznova, and Hung-Hsu Chou. Latent structure emergence in diffusion models via confidence-based filtering. *arXiv preprint arXiv:2602.06155*, 2026.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3024–3034, 2025.

Song Yan, Wei Zhai, Chenfeng Wang, Xinliang Bi, Jian Yang, Yancheng Cai, Yusen Zhang, Yunwei Lan, Tao Zhang, GuanYe Xiong, Min Li, and Zheng-Jun Zha. Determinism of randomness: Prompt-residual seed shaping for diffusion generation. *arXiv preprint arXiv:2511.07756*, 2025. Revised 2026.

CONTENTS OF THE APPENDIX

- In Sec. A, we provide a full list of notation and the flow assumptions under which every endpoint map $\mathcal{G}_{\boldsymbol{F}}$ is a C^1 diffeomorphism.
- In Sec. B, we derive the fixed-center coordinates and the reparameterized probability-flow ODEs in Eq. (2), and verify the flow assumptions for the Gaussian mixture.
- In Sec. C, we prove Prop. 1 and treat the empirical class score and the contribution of competing classes.
- In Sec. D, we derive the certified radius in Eq. (4), prove Prop. 2, and characterize its dependence on δ and σ and the transported boundary $R_i(\sigma)$.
- In Sec. E, we prove Prop. 3 and provide the two-mode reduction and segment contraction of Ashiq et al. (2026) in our coordinates.
- In Sec. F, we identify the pulled-back boundary, prove Prop. 4, and derive the alignment score for seed corrections.
- In Sec. H, we provide additional experiments.
- In Sec. G, we provide implementation details for our experiments.
- In Sec. I, we provide a table of symbols used throughout our paper.

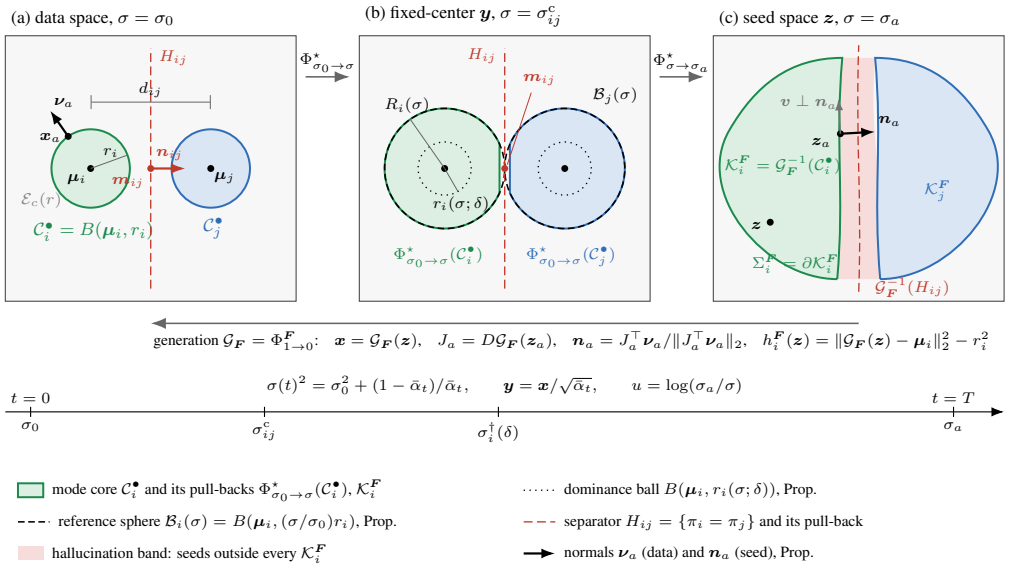


Figure S6: Full notation

A Notation and Flow Assumptions

A.1 Full Notation

Let d denote the ambient dimension and $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N}$. Vectors belong to $\mathbb{R}^d$ unless otherwise specified, and $\mathbf{0}$ is the zero vector. We use $\langle \boldsymbol{x}, \boldsymbol{y} \rangle := \boldsymbol{x}^\top \boldsymbol{y}$ and $\|\boldsymbol{x}\|_2$ for the Euclidean inner product and norm. For a matrix A, $A^\top$ denotes its transpose and $\|A\|_{\mathrm{op}} := \sup_{\|\boldsymbol{v}\|_2=1} \|A\boldsymbol{v}\|_2$ its induced norm. The identity matrix is $\boldsymbol{I}_d$, and $\mathrm{diag}(\lambda_1, \dots, \lambda_d)$ is the diagonal matrix with the given entries. For symmetric A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues, while $A \succeq 0$ and $A \succ 0$ denote positive semidefiniteness and positive definiteness.

Balls $B(\boldsymbol{c}, r) := \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{c}\|_2 \le r\}$ are closed. For a set $\mathcal{A}$ we write $\operatorname{int}\mathcal{A}$, $\overline{\mathcal{A}}$, and $\partial\mathcal{A}$ for its interior, closure, and boundary, $\operatorname{dist}(\boldsymbol{x}, \mathcal{A})$ for its Euclidean distance from $\boldsymbol{x}$, and $T_{\boldsymbol{z}}\Sigma$ for the tangent space of a smooth hypersurface Σ at $\boldsymbol{z}$. For scalar-valued f, ∇f and $\nabla^2 f$ denote its gradient and Hessian; for vector-valued f, Df denotes its Jacobian. A C^1 diffeomorphism is a continuously differentiable bijection with continuously differentiable inverse.

We distinguish the Gaussian law $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ from its density $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$. The symbol $*$ denotes convolution and $\mathcal{G}_{\#}\nu$ the pushforward of a measure ν by $\mathcal{G}$. We write $\mathbb{E}[\cdot \mid \cdot]$ for conditional expectation and Var, Cov for variance and covariance; $\mathrm{Cov}_\pi(\boldsymbol{\mu})$ is the covariance of a random center equal to $\boldsymbol{\mu}_i$ with probability π_i. A hat distinguishes an empirical class density from its population counterpart. The subscripts $\star$ and θ distinguish population and learned sampling fields.

Generation runs from $\tau = 1$ to $\tau = 0$. For the VP model, original time t runs from T to 0, with $\tau = t/T$, and the total component scale σ decreases from σ_a to σ_0. The parameter $u := \log(\sigma_a/\sigma)$ instead increases during denoising. The smoothing scale η is tied to t through $\eta(t)^2 = (1-\bar\alpha_t)/\bar\alpha_t$, except in Sec. C, where it is a free parameter $\eta \ge 0$. A dot denotes differentiation with respect to the displayed time variable. The process $\boldsymbol{W}_t$ is standard d-dimensional Brownian motion, and stochastic integrals are interpreted in the Itô sense.

We write $[a]_+ := \max\{a, 0\}$. In a stated limit, $f = O(g)$ means $\|f\| \le Cg$ for some constant independent of the limiting parameter, with $g > 0$; $f = o(g)$ means $\|f\|/g \to 0$, and $o(\varepsilon)$ is always understood as $\varepsilon \to 0^+$. Absolute values replace norms for scalars. Uniformity over additional variables is stated when required. The symbols specific to the model and its geometry are listed in Sec. I.

A.2 Flow regularity and invertibility

We assume that $\boldsymbol{F}$ is continuous in $(\boldsymbol{x}, \tau)$, continuously differentiable in $\boldsymbol{x}$ with continuous spatial derivative, and that there are finite constants L, M such that

$$\|D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{x},\tau)\|_{\mathrm{op}} \le L, \qquad \|\boldsymbol{F}(\boldsymbol{0},\tau)\|_2 \le M, \qquad (\boldsymbol{x},\tau) \in \mathbb{R}^d \times [0,1]. \tag{A.1}$$

Lemma 1 (Invertible generative flow)**.** *Under Eq.* (A.1)*, the trajectory through any point at any time in* $[0,1]$ *exists uniquely throughout the interval. Every map* $\Phi^{\boldsymbol{F}}_{a\to b}$ *is a* C^1 *diffeomorphism with inverse* $\Phi^{\boldsymbol{F}}_{b\to a}$*. In particular,* $\mathcal{G}_{\boldsymbol{F}} = \Phi^{\boldsymbol{F}}_{1\to 0}$ *is a* C^1 *diffeomorphism of* $\mathbb{R}^d$*.*

Proof. The mean-value formula gives

$$\boldsymbol{F}(\boldsymbol{x},\tau) - \boldsymbol{F}(\boldsymbol{y},\tau) = \int_0^1 D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{y} + s(\boldsymbol{x}-\boldsymbol{y}),\tau)(\boldsymbol{x}-\boldsymbol{y})\, ds, \tag{A.2}$$

$$\|\boldsymbol{F}(\boldsymbol{x},\tau) - \boldsymbol{F}(\boldsymbol{y},\tau)\|_2 \le L\|\boldsymbol{x}-\boldsymbol{y}\|_2, \qquad \|\boldsymbol{F}(\boldsymbol{x},\tau)\|_2 \le M + L\|\boldsymbol{x}\|_2. \tag{A.3}$$

Local existence and uniqueness therefore hold in either time direction. Along a solution starting from $\boldsymbol{x}$ at time a, the integral equation and Grönwall's inequality give

$$\|\Phi^{\boldsymbol{F}}_{a\to\tau}(\boldsymbol{x})\|_2 \le (\|\boldsymbol{x}\|_2 + M|\tau - a|)e^{L|\tau-a|}. \tag{A.4}$$

For $\tau < a$, the same estimate follows by reversing the time parameter. Thus no solution escapes to infinity in finite time, and the local solution extends to all of $[0,1]$.

Uniqueness implies the composition rule $\Phi^{\boldsymbol{F}}_{b\to c} \circ \Phi^{\boldsymbol{F}}_{a\to b} = \Phi^{\boldsymbol{F}}_{a\to c}$. Taking $c = a$ and reversing the roles of a, b proves that the forward and backward maps are mutual inverses. Continuous differentiability with respect to initial conditions follows from the continuous spatial derivative of $\boldsymbol{F}$. More explicitly, along $\boldsymbol{X}_\tau = \Phi^{\boldsymbol{F}}_{a\to\tau}(\boldsymbol{x})$, its Jacobian $J(\tau) := D_{\boldsymbol{x}}\Phi^{\boldsymbol{F}}_{a\to\tau}(\boldsymbol{x})$ satisfies the variational equation

$$\frac{dJ}{d\tau} = D_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{X}_\tau,\tau)J, \qquad J(a) = \boldsymbol{I}_d. \tag{A.5}$$

The matrix solution K of $K' = -KD_{\boldsymbol{x}}\boldsymbol{F}(\boldsymbol{X}_\tau,\tau)$, $K(a) = \boldsymbol{I}_d$, satisfies $(KJ)' = \boldsymbol{0}$, so $KJ = \boldsymbol{I}_d$ at every time. Thus J is invertible. The inverse flow has the same C^1 regularity, proving the assertion. □

Remark. These are sufficient conditions for a globally defined C^1 invertible flow over the sampling interval. They are verified for the Gaussian reference mixture in Lemma 3, with $\sigma_0 > 0$. For a general sampling field, the global flow statements are conditional on these assumptions. The local boundary calculation requires only a C^1 endpoint map near the anchor with $J_a^\top \boldsymbol{\nu}_a \neq \boldsymbol{0}$.

B Population dynamics in fixed-center coordinates

We use the Gaussian mixture, with $C \ge 2$ distinct centers, positive weights summing to one, and $\sigma_0 > 0$. Let $T > 0$ and let β be continuous and positive on $[0,T]$. The VP forward process is

$$d\boldsymbol{X}_t = -\frac{\beta_t}{2}\boldsymbol{X}_t\, dt + \sqrt{\beta_t}\, d\boldsymbol{W}_t, \qquad \bar\alpha_t := \exp\!\left(-\int_0^t \beta_s\, ds\right). \tag{B.1}$$

Time t increases toward higher noise; denoising runs from T to 0.

B.1 Original diffusion time and noise scale

The VP transition law is (Song et al., 2021b)

$$\boldsymbol{X}_t \mid \boldsymbol{X}_0 \sim \mathcal{N}\big(\sqrt{\bar\alpha_t}\boldsymbol{X}_0, (1-\bar\alpha_t)\boldsymbol{I}_d\big), \qquad t > 0. \tag{B.2}$$

Conditioning on component i gives mean $\sqrt{\bar\alpha_t}\boldsymbol{\mu}_i$ and covariance $(\bar\alpha_t\sigma_0^2 + 1 - \bar\alpha_t)\boldsymbol{I}_d$. Thus, for $t \in [0,T]$,

$$p_t^{\mathrm{VP}}(\boldsymbol{x}) = \sum_{i=1}^C w_i \mathcal{N}\big(\boldsymbol{x}; \sqrt{\bar\alpha_t}\boldsymbol{\mu}_i, v_t\boldsymbol{I}_d\big), \qquad v_t := \bar\alpha_t\sigma_0^2 + 1 - \bar\alpha_t. \tag{B.3}$$

Set $\boldsymbol{y} := \boldsymbol{x}/\sqrt{\bar{\alpha}_t}$. This is the change of variables that removes the signal decay of the VP process: by Eq. (B.2), $\boldsymbol{Y}_t := \boldsymbol{X}_t/\sqrt{\bar{\alpha}_t}$ satisfies

$$\boldsymbol{Y}_t \mid \boldsymbol{X}_0 \sim \mathcal{N}\left(\boldsymbol{X}_0, \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}\boldsymbol{I}_d\right), \tag{B.4}$$

so in the coordinates $\boldsymbol{y}$ the data are not shrunk but only convolved with isotropic noise, exactly as in a variance-exploding process (Song et al., 2021b, Sec. 3.4) or in the σ-parameterization of Karras et al. (2022), where the same rescaling appears as the scaling function $s(t) = \sqrt{\bar{\alpha}_t}$. Accordingly, define

$$\eta(t)^2 := \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}, \qquad \sigma(t)^2 := \frac{v_t}{\bar{\alpha}_t} = \sigma_0^2 + \eta(t)^2, \qquad \sigma_a := \sigma(T), \tag{B.5}$$

where the middle identity is $v_t/\bar{\alpha}_t = (\bar{\alpha}_t\sigma_0^2 + 1 - \bar{\alpha}_t)/\bar{\alpha}_t$ from Eq. (B.3). Here η is the added noise scale in the coordinates $\boldsymbol{y}$, and σ is the total standard deviation of a component in those coordinates: a component of variance σ_0^2 convolved with noise of variance $\eta(t)^2$. In particular, conditioning Eq. (B.4) on component i gives $\boldsymbol{Y}_t \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma(t)^2\boldsymbol{I}_d)$: the centers $\boldsymbol{\mu}_i$ do not move, and only the scale $\sigma(t)$ changes with t (Fig. S9).

Remark 1 (Schedule independence). The only hypothesis used above is that β is continuous and positive on $[0, T]$; no form of the schedule is assumed. Since the σ-parameterized flow Eq. (B.15) does not involve β, two VP schedules with the same terminal scale $\sigma_a = \sigma(T)$ have the same trajectories as curves and the same endpoint map $\mathcal{G}_\star$, and they differ only in the speed at which the scale is traversed. All statements of Sec. 4 are therefore properties of the mixture and of σ_a alone; the schedule enters only through the conversion of a scale to a time, as in Eq. (5), and through the discretization of the sampler.

Since $\bar{\alpha}_t = \exp(-\int_0^t \beta_s\, ds)$, we have $\dot{\bar{\alpha}}_t = -\beta_t\bar{\alpha}_t$, and differentiating $\sigma(t)^2 = \sigma_0^2 + \bar{\alpha}_t^{-1} - 1$ gives

$$2\sigma(t)\dot{\sigma}(t) = \frac{d}{dt}\bar{\alpha}_t^{-1} = -\bar{\alpha}_t^{-2}\dot{\bar{\alpha}}_t = \frac{\beta_t}{\bar{\alpha}_t}, \qquad \dot{\sigma}(t) = \frac{\beta_t}{2\bar{\alpha}_t\sigma(t)} > 0. \tag{B.6}$$

Consequently, σ is a C^1 strictly increasing function on $[0, T]$, hence a bijection onto $[\sigma(0), \sigma(T)] = [\sigma_0, \sigma_a]$, where $\sigma(0) = \sigma_0$ because $\bar{\alpha}_0 = 1$ gives $\eta(0) = 0$. This is what allows the trajectory to be indexed by σ instead of t in Lemma 2.

For $\sigma > 0$, write $\phi_i^\sigma(\boldsymbol{y}) := \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_i, \sigma^2\boldsymbol{I}_d)$. The fixed-center density and responsibilities are

$$p_\sigma := \sum_{i=1}^C w_i\phi_i^\sigma, \qquad \pi_i := \frac{w_i\phi_i^\sigma}{p_\sigma}, \qquad \boldsymbol{D}_\sigma := \sum_{i=1}^C \pi_i\boldsymbol{\mu}_i. \tag{B.7}$$

Using $v_t = \bar{\alpha}_t\sigma(t)^2$ gives

$$\begin{aligned}\mathcal{N}\big(\sqrt{\bar{\alpha}_t}\boldsymbol{y}; \sqrt{\bar{\alpha}_t}\boldsymbol{\mu}_i, v_t\boldsymbol{I}_d\big) &= (2\pi\bar{\alpha}_t\sigma(t)^2)^{-d/2}\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}_i\|_2^2}{2\sigma(t)^2}\right)\\ &= \bar{\alpha}_t^{-d/2}\phi_i^{\sigma(t)}(\boldsymbol{y}).\end{aligned} \tag{B.8}$$

Summing over components and differentiating with respect to $\boldsymbol{x}$,

$$p_t^{\mathrm{VP}}(\boldsymbol{x}) = \bar{\alpha}_t^{-d/2}p_{\sigma(t)}(\boldsymbol{x}/\sqrt{\bar{\alpha}_t}), \tag{B.9}$$

$$\nabla_{\boldsymbol{x}}\log p_t^{\mathrm{VP}}(\boldsymbol{x}) = \bar{\alpha}_t^{-1/2}(\nabla_{\boldsymbol{y}}\log p_{\sigma(t)})(\boldsymbol{x}/\sqrt{\bar{\alpha}_t}). \tag{B.10}$$

Writing π_i^{VP} for the component responsibilities in Eq. (B.3), the common density factor cancels:

$$\pi_i^{\mathrm{VP}}(\sqrt{\bar{\alpha}_t}\boldsymbol{y}, t) = \frac{w_i\bar{\alpha}_t^{-d/2}\phi_i^{\sigma(t)}(\boldsymbol{y})}{\bar{\alpha}_t^{-d/2}p_{\sigma(t)}(\boldsymbol{y})} = \pi_i(\boldsymbol{y}, \sigma(t)). \tag{B.11}$$

B.2 Probability flow and denoising parameterizations

The associated probability-flow trajectory satisfies (Song et al., 2021b)

$$\frac{d\boldsymbol{X}_t}{dt} = -\frac{\beta_t}{2}\left(\boldsymbol{X}_t + \nabla_{\boldsymbol{x}}\log p_t^{\mathrm{VP}}(\boldsymbol{X}_t)\right). \tag{B.12}$$

Lemma 2 (Fixed-center probability flow)**.** *Let $\boldsymbol{X}_t$ solve the probability-flow ODE Eq.* (B.12) *of the VP process on $[0,T]$, and set $\boldsymbol{Y}_t := \bar{\alpha}_t^{-1/2}\boldsymbol{X}_t$, $\sigma(t)^2 := \sigma_0^2 + (1-\bar{\alpha}_t)/\bar{\alpha}_t$, $\sigma_a := \sigma(T)$, and $u_f := \log(\sigma_a/\sigma_0)$. Then $\boldsymbol{Y}_t$ satisfies*

$$\frac{d\boldsymbol{Y}_t}{dt} = -\frac{\beta_t}{2\bar{\alpha}_t}\nabla \log p_{\sigma(t)}(\boldsymbol{Y}_t), \qquad t \in [0,T].$$

Reindexed by the noise scale $\sigma = \sigma(t)$, the same trajectory satisfies

$$\frac{d\boldsymbol{Y}_\sigma}{d\sigma} = -\sigma \nabla \log p_\sigma(\boldsymbol{Y}_\sigma) = \frac{\boldsymbol{Y}_\sigma - \boldsymbol{D}_\sigma(\boldsymbol{Y}_\sigma)}{\sigma}, \qquad \sigma \in [\sigma_0, \sigma_a],$$

and reindexed by $u = \log(\sigma_a/\sigma)$, with $\sigma_u := \sigma_a e^{-u}$,

$$\frac{d\boldsymbol{Y}_u}{du} = \boldsymbol{D}_{\sigma_u}(\boldsymbol{Y}_u) - \boldsymbol{Y}_u, \qquad u \in [0, u_f].$$

During generation t and σ decrease while u increases.

Proof. For this trajectory, set $\boldsymbol{Y}_t := \bar{\alpha}_t^{-1/2}\boldsymbol{X}_t$. Using Eq. (B.10),

$$\begin{aligned}\frac{d\boldsymbol{Y}_t}{dt} &= \bar{\alpha}_t^{-1/2}\frac{d\boldsymbol{X}_t}{dt} + \frac{\beta_t}{2}\boldsymbol{Y}_t \\ &= -\frac{\beta_t}{2}\boldsymbol{Y}_t - \frac{\beta_t}{2\bar{\alpha}_t}\nabla \log p_{\sigma(t)}(\boldsymbol{Y}_t) + \frac{\beta_t}{2}\boldsymbol{Y}_t \\ &= -\frac{\beta_t}{2\bar{\alpha}_t}\nabla \log p_{\sigma(t)}(\boldsymbol{Y}_t). \end{aligned} \tag{B.13}$$

Written as $d\boldsymbol{Y}_t/dt = -\frac{1}{2}g(t)^2\nabla \log p_{\sigma(t)}(\boldsymbol{Y}_t)$, the noise growth rate is $g(t)^2 = \beta_t/\bar{\alpha}_t = 2\sigma(t)\dot{\sigma}(t)$ by Eq. (B.6).

Since $\nabla\phi_i^\sigma = \phi_i^\sigma(\boldsymbol{\mu}_i - \boldsymbol{y})/\sigma^2$ and $\sum_i \pi_i = 1$,

$$\nabla \log p_\sigma(\boldsymbol{y}) = \sum_{i=1}^{C}\pi_i(\boldsymbol{y},\sigma)\frac{\boldsymbol{\mu}_i - \boldsymbol{y}}{\sigma^2} = \frac{\boldsymbol{D}_\sigma(\boldsymbol{y}) - \boldsymbol{y}}{\sigma^2}. \tag{B.14}$$

Reindexing by $\sigma = \sigma(t)$ therefore yields

$$\begin{aligned}\frac{d\boldsymbol{Y}_\sigma}{d\sigma} &= \frac{d\boldsymbol{Y}_t/dt}{\dot{\sigma}(t)} = \frac{-\beta_t/(2\bar{\alpha}_t)}{\beta_t/(2\bar{\alpha}_t\sigma)}\nabla \log p_\sigma(\boldsymbol{Y}_\sigma) \\ &= -\sigma\nabla \log p_\sigma(\boldsymbol{Y}_\sigma) = \frac{\boldsymbol{Y}_\sigma - \boldsymbol{D}_\sigma(\boldsymbol{Y}_\sigma)}{\sigma}, \qquad \sigma : \sigma_a \downarrow \sigma_0. \end{aligned} \tag{B.15}$$

This proves Eq. (2). For $u := \log(\sigma_a/\sigma)$, set $u_f := \log(\sigma_a/\sigma_0)$ and $\sigma_u := \sigma_a e^{-u}$. Then

$$\frac{d\boldsymbol{Y}_u}{du} = \frac{\boldsymbol{Y}_u - \boldsymbol{D}_{\sigma_u}(\boldsymbol{Y}_u)}{\sigma_u}(-\sigma_u) = \boldsymbol{D}_{\sigma_u}(\boldsymbol{Y}_u) - \boldsymbol{Y}_u, \qquad u \in [0, u_f]. \tag{B.16}$$

Thus u increases during denoising, while t and σ decrease. □

Posterior center and clean observation. The quantity $\boldsymbol{D}_\sigma(\boldsymbol{y})$ is the posterior mean of the component center. For $\boldsymbol{X}_0 \sim p_{\text{data}}$ and $\boldsymbol{Y} = \boldsymbol{X}_0 + \eta\boldsymbol{Z}$, with independent $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ and $\sigma^2 = \sigma_0^2 + \eta^2$, Gaussian conditioning instead gives

$$\begin{aligned}\mathbb{E}[\boldsymbol{X}_0 \mid \boldsymbol{Y} = \boldsymbol{y}] &= \sum_i \pi_i(\boldsymbol{y},\sigma)\left(\boldsymbol{\mu}_i + \frac{\sigma_0^2}{\sigma^2}(\boldsymbol{y} - \boldsymbol{\mu}_i)\right) \\ &= \frac{\sigma_0^2}{\sigma^2}\boldsymbol{y} + \frac{\eta^2}{\sigma^2}\boldsymbol{D}_\sigma(\boldsymbol{y}) = \boldsymbol{y} + \eta^2\nabla \log p_\sigma(\boldsymbol{y}). \end{aligned} \tag{B.17}$$

Relation to the main-text field. For the VP model, choosing $\tau = t/T$ in Eq. (1) gives

$$\boldsymbol{F}_\star(\boldsymbol{x},\tau) = -\frac{T\beta_{T\tau}}{2}\left(\boldsymbol{x} + \nabla_{\boldsymbol{x}} \log p_{T\tau}^{\text{VP}}(\boldsymbol{x})\right). \tag{B.18}$$

Its endpoint map is $\mathcal{G}_\star = \Phi_{1\to 0}^{\boldsymbol{F}_\star}$. Equivalently, scale the initial point by $\bar{\alpha}_T^{-1/2}$ and integrate Eq. (B.15) from σ_a to σ_0; the terminal coordinates agree because $\bar{\alpha}_0 = 1$. Replacing the true score in Eq. (B.18) by its learned estimate gives the corresponding VP field $\boldsymbol{F}_\theta$.

B.3 Regularity of the Gaussian mixture flow

Lemma 3 (Regularity of the mixture flow). *Let $D_{\max} := \max_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2$. For every $\sigma > 0$ and $\boldsymbol{y} \in \mathbb{R}^d$,*

$$D\boldsymbol{D}_\sigma(\boldsymbol{y}) = \frac{\mathrm{Cov}_{\pi(\boldsymbol{y},\sigma)}(\boldsymbol{\mu})}{\sigma^2}, \qquad \|D\boldsymbol{D}_\sigma(\boldsymbol{y})\|_{\mathrm{op}} \le \frac{D_{\max}^2}{4\sigma^2}, \tag{B.19}$$

where $\boldsymbol{\mu}$ takes value $\boldsymbol{\mu}_i$ with probability $\pi_i(\boldsymbol{y}, \sigma)$. The flow maps of Eqs. (B.12)*,* (B.15) *and* (B.16) *between any two times in their respective intervals are C^1 diffeomorphisms of $\mathbb{R}^d$.*

Proof. All component responsibilities $\pi_i(\boldsymbol{y}, \sigma)$ are continuously differentiable in $(\boldsymbol{y}, \sigma)$ for $\sigma > 0$. Using the score relation $\nabla_{\boldsymbol{y}} \log \phi_i^\sigma(\boldsymbol{y}) = -\frac{\boldsymbol{y}-\boldsymbol{\mu}_i}{\sigma^2}$ alongside Eq. (B.14), the gradient of the i-th responsibility with respect to $\boldsymbol{y}$ is given by:

$$\nabla \pi_i = \pi_i\big(\nabla \log \phi_i^\sigma - \nabla \log p_\sigma\big) = \frac{\pi_i}{\sigma^2}(\boldsymbol{\mu}_i - \boldsymbol{D}_\sigma). \tag{B.20}$$

Differentiating the posterior mean $\boldsymbol{D}_\sigma(\boldsymbol{y}) = \sum_i \pi_i(\boldsymbol{y}, \sigma)\boldsymbol{\mu}_i$ with respect to $\boldsymbol{y}$ yields the spatial Jacobian matrix:

$$D\boldsymbol{D}_\sigma = \sum_i \boldsymbol{\mu}_i (\nabla \pi_i)^\top = \frac{1}{\sigma^2}\left(\sum_i \pi_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top - \boldsymbol{D}_\sigma \boldsymbol{D}_\sigma^\top\right) = \frac{\mathrm{Cov}_\pi(\boldsymbol{\mu})}{\sigma^2}. \tag{B.21}$$

To bound the operator norm $\|D\boldsymbol{D}_\sigma\|_{\mathrm{op}}$, fix an arbitrary unit vector $\boldsymbol{v} \in \mathbb{R}^d$ ($\|\boldsymbol{v}\|_2 = 1$). The Rayleigh quotient evaluates to:

$$\sigma^2 \boldsymbol{v}^\top D\boldsymbol{D}_\sigma \boldsymbol{v} = \boldsymbol{v}^\top \mathrm{Cov}_\pi(\boldsymbol{\mu})\boldsymbol{v} = \mathrm{Var}_\pi(\langle \boldsymbol{v}, \boldsymbol{\mu}\rangle). \tag{B.22}$$

Let $a := \min_i \langle \boldsymbol{v}, \boldsymbol{\mu}_i\rangle$ and $b := \max_i \langle \boldsymbol{v}, \boldsymbol{\mu}_i\rangle$, so that the scalar random variable $\langle \boldsymbol{v}, \boldsymbol{\mu}\rangle$ is supported on $[a, b]$. Bounding the variance by the expected squared deviation from the interval midpoint $\frac{a+b}{2}$ (Popoviciu's inequality for variance) gives:

$$\mathrm{Var}_\pi(\langle \boldsymbol{v}, \boldsymbol{\mu}\rangle) \le \mathbb{E}_\pi\left[\left(\langle \boldsymbol{v}, \boldsymbol{\mu}\rangle - \frac{a+b}{2}\right)^2\right] \le \frac{(b-a)^2}{4}. \tag{B.23}$$

Let i^*, j^* denote the component indices attaining the maximum and minimum projections, respectively, such that $b - a = \langle \boldsymbol{v}, \boldsymbol{\mu}_{i^*} - \boldsymbol{\mu}_{j^*}\rangle$. By the Cauchy–Schwarz inequality:

$$b - a = \langle \boldsymbol{v}, \boldsymbol{\mu}_{i^*} - \boldsymbol{\mu}_{j^*}\rangle \le \|\boldsymbol{v}\|_2 \|\boldsymbol{\mu}_{i^*} - \boldsymbol{\mu}_{j^*}\|_2 \le 1 \cdot \max_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 = D_{\max}. \tag{B.24}$$

Hence, $\sigma^2 \boldsymbol{v}^\top D\boldsymbol{D}_\sigma \boldsymbol{v} \le \frac{D_{\max}^2}{4}$. Since $D\boldsymbol{D}_\sigma$ is symmetric positive semidefinite, taking the supremum over all unit vectors $\boldsymbol{v}$ yields:

$$\|D\boldsymbol{D}_\sigma\|_{\mathrm{op}} \le \frac{D_{\max}^2}{4\sigma^2}. \tag{B.25}$$

Next, consider the vector field $\boldsymbol{G}(\boldsymbol{y}, u) := \boldsymbol{D}_{\sigma_u}(\boldsymbol{y}) - \boldsymbol{y}$. By the triangle inequality for operator norms and the fact that $\|I_d\|_{\mathrm{op}} = 1$:

$$\begin{aligned} \|D_{\boldsymbol{y}}\boldsymbol{G}(\boldsymbol{y}, u)\|_{\mathrm{op}} &= \|D\boldsymbol{D}_{\sigma_u}(\boldsymbol{y}) - I_d\|_{\mathrm{op}} \\ &\le \|D\boldsymbol{D}_{\sigma_u}(\boldsymbol{y})\|_{\mathrm{op}} + \|I_d\|_{\mathrm{op}} \\ &\le 1 + \frac{D_{\max}^2}{4\sigma_u^2} \le 1 + \frac{D_{\max}^2}{4\sigma_0^2}, \end{aligned} \tag{B.26}$$

where the final inequality holds since $\sigma_u \ge \sigma_0 > 0$. Finally, applying the triangle inequality and the responsibility normalization $\sum_i \pi_i(\boldsymbol{0}, \sigma_u) = 1$ at $\boldsymbol{y} = \boldsymbol{0}$ gives:

$$\|\boldsymbol{G}(\boldsymbol{0}, u)\|_2 = \|\boldsymbol{D}_{\sigma_u}(\boldsymbol{0})\|_2 = \left\|\sum_i \pi_i(\boldsymbol{0}, \sigma_u)\boldsymbol{\mu}_i\right\|_2 \le \sum_i \pi_i(\boldsymbol{0}, \sigma_u)\|\boldsymbol{\mu}_i\|_2 \le \max_i \|\boldsymbol{\mu}_i\|_2. \tag{B.27}$$

□

The field $\boldsymbol{G}$ and $D_{\boldsymbol{y}}\boldsymbol{G}$ are jointly continuous in $(\boldsymbol{y}, u)$, and the two bounds above are Eq. (A.1) after normalizing $[0, u_f]$ to $[0, 1]$, so Lemma 1 gives a C^1 diffeomorphism between any two u-times. The C^1 bijections $t \leftrightarrow \sigma \leftrightarrow u$ of Eq. (B.6) and the linear map $\boldsymbol{x} = \sqrt{\bar{\alpha}_t}\,\boldsymbol{y}$ transfer this to the σ and VP flows, and hence to $\mathcal{G}_\star$.

The seed law. The exact VP endpoint map satisfies $(\mathcal{G}_\star)_\# p_T^{\mathrm{VP}} = p_{\mathrm{data}}$. At finite T, p_T^{VP} need not equal $\gamma = \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. By Eq. (B.3), p_T^{VP} converges in distribution to γ as $\bar{\alpha}_T \to 0$, which requires $\int_0^T \beta_s \, ds \to \infty$ when taking $T \to \infty$. The pullback boundaries and their pointwise normal and tangent formulas depend on the endpoint map; probabilities assigned to these regions also depend on the seed law.

B.4 Relation to confidence-based latent separation

Support assumptions and flow regularity. Wei et al. (2026) establish that a regular, invertible probability flow transports disjoint, connected class supports to disjoint, connected latent regions. The disjoint-support assumption does not hold for our Gaussian components, each of which has support $\mathbb{R}^d$. Their disjointness conclusion therefore does not directly apply to these component distributions. We instead identify disjoint subsets on which a single component has responsibility at least $1 - \delta$, and transport these certified cores into the initial noise space.

A separate distinction concerns their explanation of latent mixing. In Section 3.3, they discuss possible failure of Lipschitz regularity in low-density regions, leading to a loss of uniqueness and invertibility. Low density alone, however, does not imply such a failure. In our reference model, the mixture score is globally Lipschitz uniformly over the finite sampling interval $\sigma \in [\sigma_0, \sigma_a]$, with $\sigma_0 > 0$, and the endpoint map is a diffeomorphism (Lemmas 1 and 3). Consequently, seeds can generate low-density inter-mode outputs while the flow remains regular and the pulled-back cores remain disjoint. Loss of injectivity is therefore not required for the inter-mode outcomes studied here.

Preservation of separation and boundary geometry. Their transport theorem establishes preservation of connectedness and disjointness, but does not determine which subsets of overlapping class distributions admit a single-mode description, or quantify the geometry of their latent boundaries. Prop. 2 supplies a certified radius for such a subset. Its boundary is then transported through the endpoint map, and Prop. 4 identifies the normal direction and the first-order response to seed perturbations. These results characterize where core membership changes and how to move locally toward the core. The geometric object is the boundary of a certified mode region, rather than the support boundary of an entire class.

Classifier confidence and geometric certification. Their practical method assigns labels and confidence scores to generated samples using a pretrained classifier, then learns a predictor of these quantities from the initial seeds. Confidence filtering empirically improves class predictability, but their analysis does not establish that a confidence threshold certifies flow regularity or a particular distance to a class boundary. Our reference construction instead specifies the terminal region through a responsibility guarantee and obtains boundary anchors by pullback. The normal calculation is consequently tied to terminal core membership, without requiring classifier confidence to represent geometric distance. Predictors fitted to the transported anchors approximate this boundary; their approximation accuracy remains distinct from the guarantee defining the terminal core.

C Gaussian approximation of a class

C.1 Class geometry and Gaussian smoothing

Fix a class c with observations $\boldsymbol{x}_c^{(1)}, \ldots, \boldsymbol{x}_c^{(n_c)}$. Let $\boldsymbol{\mu}_c := \frac{1}{n_c} \sum_{i=1}^{n_c} \boldsymbol{x}_c^{(i)}$ and $\boldsymbol{\Sigma}_c := \frac{1}{n_c} \sum_{i=1}^{n_c} (\boldsymbol{x}_c^{(i)} - \boldsymbol{\mu}_c)(\boldsymbol{x}_c^{(i)} - \boldsymbol{\mu}_c)^\top$, and assume that $\boldsymbol{\Sigma}_c$ is nonsingular. For any $\boldsymbol{v} \in \mathbb{R}^d$,

$$\boldsymbol{v}^\top \boldsymbol{\Sigma}_c \boldsymbol{v} = \frac{1}{n_c} \sum_{i=1}^{n_c} \left\langle \boldsymbol{v}, \boldsymbol{x}_c^{(i)} - \boldsymbol{\mu}_c \right\rangle^2 \geq 0. \tag{C.1}$$

Thus $\boldsymbol{\Sigma}_c$ is positive semidefinite and, by nonsingularity, positive definite. For $r \geq 0$, define

$$\mathcal{E}_c(r) := \left\{ \boldsymbol{x} \in \mathbb{R}^d : (\boldsymbol{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_c) \leq r^2 \right\}. \tag{C.2}$$

Write $\boldsymbol{\Sigma}_c = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_k > 0$. For $\boldsymbol{z} := \boldsymbol{U}^\top(\boldsymbol{x} - \boldsymbol{\mu}_c)$,

$$(\boldsymbol{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c) = \boldsymbol{z}^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{z} = \sum_{k=1}^{d} \frac{z_k^2}{\lambda_k}. \tag{C.3}$$

For $r > 0$, $\mathcal{E}_c(r)$ is therefore a closed ellipsoidal region with principal directions given by the columns of $\boldsymbol{U}$ and semiaxes $r\sqrt{\lambda_k}$. For $r = 0$, it consists only of $\boldsymbol{\mu}_c$. If $\boldsymbol{\Sigma}_c = \sigma_0^2 \boldsymbol{I}_d$, then

$$\mathcal{E}_c(r) = \left\{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{\mu}_c\|_2^2 \le \sigma_0^2 r^2\right\} = B(\boldsymbol{\mu}_c, \sigma_0 r).$$

Let $f_{c,0}$ denote the population class density and set $g_{c,0}(\boldsymbol{y}) := \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. For $\eta > 0$, define

$$f_{c,\eta} := f_{c,0} * \varphi_\eta, \qquad g_{c,\eta} := g_{c,0} * \varphi_\eta, \qquad \varphi_\eta(\boldsymbol{z}) := \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \eta^2 \boldsymbol{I}_d). \tag{C.4}$$

At $\eta = 0$, use $f_{c,0}$ and $g_{c,0}$ directly. The reference parameters $\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c$ are held fixed throughout the proof. The assumption on $f_{c,0}$ is relative to these parameters; it does not follow from the empirical moment definitions.

Adding independent Gaussian noise adds its covariance, so $g_{c,\eta}(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_c, \boldsymbol{S}_c(\eta))$, where $\boldsymbol{S}_c(\eta) := \boldsymbol{\Sigma}_c + \eta^2 \boldsymbol{I}_d$. Consequently,

$$\log g_{c,\eta}(\boldsymbol{y}) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\det \boldsymbol{S}_c(\eta) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu}_c)^\top \boldsymbol{S}_c(\eta)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_c), \tag{C.5}$$

$$\nabla \log g_{c,\eta}(\boldsymbol{y}) = -\boldsymbol{S}_c(\eta)^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_c). \tag{C.6}$$

Writing $\boldsymbol{z} = \boldsymbol{U}^\top(\boldsymbol{y} - \boldsymbol{\mu}_c)$, the kth principal component of the score is $-z_k/(\lambda_k + \eta^2)$. For the same displacement along a principal axis, the inward score component is therefore larger when the corresponding variance is smaller. At zero smoothing, the Gaussian density on $\partial\mathcal{E}_c(r)$ is $(2\pi)^{-d/2}(\det \boldsymbol{\Sigma}_c)^{-1/2} e^{-r^2/2}$.

C.2 Proof of Proposition 1

Proof. Write $f := f_{c,0}$, $\boldsymbol{\mu} := \boldsymbol{\mu}_c$, $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_c$, and $\kappa := \kappa_c$. By assumption, f is positive and C^1, $\boldsymbol{\Sigma} \succ 0$, and the residual $\boldsymbol{R}(\boldsymbol{x}) := \boldsymbol{R}_c(\boldsymbol{x}) := \boldsymbol{\Sigma}\nabla \log f(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\mu})$ satisfies $\|\boldsymbol{R}(\boldsymbol{x})\|_2 \le \kappa$ for every $\boldsymbol{x} \in \mathbb{R}^d$.

We first justify differentiation and integration by parts. For $V(\boldsymbol{x}) := \log f(\boldsymbol{x}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, we have $\nabla V(\boldsymbol{x}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{R}(\boldsymbol{x})$. Integrating along the segment from $\boldsymbol{\mu}$ to $\boldsymbol{x}$ gives

$$\begin{aligned} V(\boldsymbol{x}) - V(\boldsymbol{\mu}) &= \int_0^1 \left\langle \boldsymbol{\Sigma}^{-1}\boldsymbol{R}(\boldsymbol{\mu} + s(\boldsymbol{x} - \boldsymbol{\mu})), \boldsymbol{x} - \boldsymbol{\mu} \right\rangle ds \\ &\le \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}} \kappa \|\boldsymbol{x} - \boldsymbol{\mu}\|_2. \end{aligned} \tag{C.7}$$

Since $V(\boldsymbol{\mu}) = \log f(\boldsymbol{\mu})$,

$$f(\boldsymbol{x}) \le f(\boldsymbol{\mu}) \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) + \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}} \kappa \|\boldsymbol{x} - \boldsymbol{\mu}\|_2 \right). \tag{C.8}$$

The quadratic term dominates the linear term as $\|\boldsymbol{x}\|_2 \to \infty$. Moreover,

$$\begin{aligned} \|\nabla f(\boldsymbol{x})\|_2 &= f(\boldsymbol{x}) \left\| \boldsymbol{\Sigma}^{-1}\big(\boldsymbol{R}(\boldsymbol{x}) - (\boldsymbol{x} - \boldsymbol{\mu})\big) \right\|_2 \\ &\le f(\boldsymbol{x}) \|\boldsymbol{\Sigma}^{-1}\|_{\mathrm{op}} \big(\kappa + \|\boldsymbol{x} - \boldsymbol{\mu}\|_2\big). \end{aligned} \tag{C.9}$$

Thus f, its first derivatives, and their products with polynomial factors are integrable and decay at infinity. These bounds justify differentiation under the convolution integral and ensure that the boundary terms below vanish.

Fix $\eta > 0$ and $\boldsymbol{y} \in \mathbb{R}^d$. Let $\boldsymbol{X} \sim f$ and $\boldsymbol{Y} = \boldsymbol{X} + \eta \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ is independent of $\boldsymbol{X}$. Write $f_\eta := f_{c,\eta}$ and $\boldsymbol{m}(\boldsymbol{y}) := \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}]$. Since $f_\eta(\boldsymbol{y}) > 0$, the conditional density is

$f(\boldsymbol{x})\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})/f_\eta(\boldsymbol{y})$. Differentiating the convolution yields

$$\begin{aligned}\nabla \log f_\eta(\boldsymbol{y}) &= \frac{1}{f_\eta(\boldsymbol{y})}\int_{\mathbb{R}^d} f(\boldsymbol{x})\nabla_{\boldsymbol{y}}\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})\,d\boldsymbol{x}\\ &= \frac{1}{f_\eta(\boldsymbol{y})}\int_{\mathbb{R}^d} \frac{\boldsymbol{x}-\boldsymbol{y}}{\eta^2} f(\boldsymbol{x})\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})\,d\boldsymbol{x}\\ &= \frac{\boldsymbol{m}(\boldsymbol{y})-\boldsymbol{y}}{\eta^2}. \end{aligned} \tag{C.10}$$

On the other hand, $\nabla_{\boldsymbol{y}}\varphi_\eta(\boldsymbol{y}-\boldsymbol{x}) = -\nabla_{\boldsymbol{x}}\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})$. Integration by parts in each coordinate gives

$$\begin{aligned}\nabla \log f_\eta(\boldsymbol{y}) &= -\frac{1}{f_\eta(\boldsymbol{y})}\int_{\mathbb{R}^d} f(\boldsymbol{x})\nabla_{\boldsymbol{x}}\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})\,d\boldsymbol{x}\\ &= \frac{1}{f_\eta(\boldsymbol{y})}\int_{\mathbb{R}^d} \nabla f(\boldsymbol{x})\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})\,d\boldsymbol{x}\\ &= \int_{\mathbb{R}^d} \nabla \log f(\boldsymbol{x})\frac{f(\boldsymbol{x})\varphi_\eta(\boldsymbol{y}-\boldsymbol{x})}{f_\eta(\boldsymbol{y})}\,d\boldsymbol{x}\\ &= \mathbb{E}[\nabla \log f(\boldsymbol{X}) \mid \boldsymbol{Y}=\boldsymbol{y}]. \end{aligned} \tag{C.11}$$

Combining these identities,

$$\begin{aligned}\mathbb{E}[\boldsymbol{R}(\boldsymbol{X}) \mid \boldsymbol{Y}=\boldsymbol{y}] &= \boldsymbol{\Sigma}\mathbb{E}[\nabla \log f(\boldsymbol{X}) \mid \boldsymbol{Y}=\boldsymbol{y}] + \mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y}=\boldsymbol{y}] - \boldsymbol{\mu}\\ &= \boldsymbol{\Sigma}\nabla \log f_\eta(\boldsymbol{y}) + \boldsymbol{m}(\boldsymbol{y}) - \boldsymbol{\mu}\\ &= \boldsymbol{\Sigma}\nabla \log f_\eta(\boldsymbol{y}) + \eta^2\nabla \log f_\eta(\boldsymbol{y}) + \boldsymbol{y} - \boldsymbol{\mu}\\ &= \boldsymbol{S}_c(\eta)\nabla \log f_\eta(\boldsymbol{y}) + (\boldsymbol{y}-\boldsymbol{\mu}). \end{aligned} \tag{C.12}$$

Using Eq. (C.6),

$$\nabla \log f_{c,\eta}(\boldsymbol{y}) - \nabla \log g_{c,\eta}(\boldsymbol{y}) = \boldsymbol{S}_c(\eta)^{-1}\mathbb{E}[\boldsymbol{R}(\boldsymbol{X}) \mid \boldsymbol{Y}=\boldsymbol{y}]. \tag{C.13}$$

The eigenvalues of $\boldsymbol{S}_c(\eta)$ are $\lambda_k(\boldsymbol{\Sigma}_c)+\eta^2$. Hence

$$\begin{aligned}\|\nabla \log f_{c,\eta}(\boldsymbol{y}) - \nabla \log g_{c,\eta}(\boldsymbol{y})\|_2 &\le \|\boldsymbol{S}_c(\eta)^{-1}\|_{\mathrm{op}}\,\|\mathbb{E}[\boldsymbol{R}(\boldsymbol{X}) \mid \boldsymbol{Y}=\boldsymbol{y}]\|_2\\ &\le \frac{\mathbb{E}[\|\boldsymbol{R}(\boldsymbol{X})\|_2 \mid \boldsymbol{Y}=\boldsymbol{y}]}{\lambda_{\min}(\boldsymbol{\Sigma}_c)+\eta^2}\\ &\le \frac{\kappa_c}{\lambda_{\min}(\boldsymbol{\Sigma}_c)+\eta^2} \le \frac{\kappa_c}{\lambda_{\min}(\boldsymbol{\Sigma}_c)}. \end{aligned} \tag{C.14}$$

At $\eta = 0$, the residual definition gives directly

$$\begin{aligned}\nabla \log f_{c,0}(\boldsymbol{y}) - \nabla \log g_{c,0}(\boldsymbol{y}) &= \nabla \log f_{c,0}(\boldsymbol{y}) + \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_c)\\ &= \boldsymbol{\Sigma}_c^{-1}\boldsymbol{R}_c(\boldsymbol{y}). \end{aligned} \tag{C.15}$$

Its norm is at most $\kappa_c/\lambda_{\min}(\boldsymbol{\Sigma}_c)$. This proves Eq. (3) at every smoothing scale. □

C.3 Approximate radial dynamics in the isotropic model

Corollary 1 (Isotropic model)**.** *Under the assumptions of Prop. 1, suppose $\boldsymbol{\Sigma}_c = \sigma_0^2\boldsymbol{I}_d$ with $\sigma_0 > 0$. For every $\eta \ge 0$ and $\boldsymbol{y} \in \mathbb{R}^d$,*

$$\left\| -\sigma\nabla \log f_{c,\eta}(\boldsymbol{y}) - \frac{\boldsymbol{y}-\boldsymbol{\mu}_c}{\sigma} \right\|_2 \le \frac{\kappa_c}{\sigma}, \qquad \sigma^2 := \sigma_0^2+\eta^2. \tag{C.16}$$

Here $-\sigma\nabla \log f_{c,\eta}$ is the class velocity in the σ parameterization.

Proof. Since $\boldsymbol{\Sigma}_c = \sigma_0^2\boldsymbol{I}_d$, $\boldsymbol{S}_c(\eta) = \sigma^2\boldsymbol{I}_d$ and $\nabla \log g_{c,\eta}(\boldsymbol{y}) = -(\boldsymbol{y}-\boldsymbol{\mu}_c)/\sigma^2$. The sharper inequality in Eq. (C.14) therefore gives

$$\left\| \nabla \log f_{c,\eta}(\boldsymbol{y}) + \frac{\boldsymbol{y}-\boldsymbol{\mu}_c}{\sigma^2} \right\|_2 \le \frac{\kappa_c}{\sigma_0^2+\eta^2} = \frac{\kappa_c}{\sigma^2}. \tag{C.17}$$

The density and score changes of variables in Sec. B apply to any class density. The class velocity in the σ parameterization is therefore $-\sigma\nabla\log f_{c,\eta}$, with $\eta=\sqrt{\sigma^2-\sigma_0^2}$. Multiplying by σ,

$$\left\|-\sigma\nabla\log f_{c,\eta}(\boldsymbol{y})-\frac{\boldsymbol{y}-\boldsymbol{\mu}_c}{\sigma}\right\|_2=\sigma\left\|\nabla\log f_{c,\eta}(\boldsymbol{y})+\frac{\boldsymbol{y}-\boldsymbol{\mu}_c}{\sigma^2}\right\|_2 \le\frac{\kappa_c}{\sigma},\qquad \sigma\ge\sigma_0>0. \tag{C.18}$$

□

For the isolated Gaussian reference, the trajectory satisfies $d\boldsymbol{Y}_\sigma/d\sigma=(\boldsymbol{Y}_\sigma-\boldsymbol{\mu}_c)/\sigma$. Consequently,

$$\frac{d}{d\sigma}\left(\frac{\boldsymbol{Y}_\sigma-\boldsymbol{\mu}_c}{\sigma}\right)=\frac{1}{\sigma}\frac{d\boldsymbol{Y}_\sigma}{d\sigma}-\frac{\boldsymbol{Y}_\sigma-\boldsymbol{\mu}_c}{\sigma^2}=\mathbf{0}.$$

For any reference scale $\sigma_b\ge\sigma_0$, integration gives

$$\boldsymbol{Y}_\sigma-\boldsymbol{\mu}_c=\frac{\sigma}{\sigma_b}(\boldsymbol{Y}_{\sigma_b}-\boldsymbol{\mu}_c). \tag{C.19}$$

For trajectories away from the center, the direction is preserved and the distance scales in proportion to σ. In particular, a sphere of radius R_0 at σ_0 has radius $R(\sigma)=(\sigma/\sigma_0)R_0$ under this isolated reference flow. It contracts during denoising and expands when traced toward higher noise. Corollary 1 bounds the instantaneous velocity error relative to this reference motion.

C.4 Empirical class scores and population approximation

For the finite observations, define

$$\widehat{f}_{c,\eta}(\boldsymbol{y}):=\frac{1}{n_c}\sum_{i=1}^{n_c}\varphi_\eta(\boldsymbol{y}-\boldsymbol{x}_c^{(i)}),\qquad \eta>0. \tag{C.20}$$

Their normalized contributions and weighted center are

$$\alpha_{c,i}(\boldsymbol{y},\eta):=\frac{\varphi_\eta(\boldsymbol{y}-\boldsymbol{x}_c^{(i)})}{\sum_{j=1}^{n_c}\varphi_\eta(\boldsymbol{y}-\boldsymbol{x}_c^{(j)})},\qquad \bar{\boldsymbol{x}}_c(\boldsymbol{y},\eta):=\sum_{i=1}^{n_c}\alpha_{c,i}(\boldsymbol{y},\eta)\boldsymbol{x}_c^{(i)}. \tag{C.21}$$

Differentiating the finite sum,

$$\begin{aligned}\nabla\log\widehat{f}_{c,\eta}(\boldsymbol{y})&=\frac{\sum_i\varphi_\eta(\boldsymbol{y}-\boldsymbol{x}_c^{(i)})(\boldsymbol{x}_c^{(i)}-\boldsymbol{y})/\eta^2}{\sum_j\varphi_\eta(\boldsymbol{y}-\boldsymbol{x}_c^{(j)})}\\&=\sum_i\alpha_{c,i}(\boldsymbol{y},\eta)\frac{\boldsymbol{x}_c^{(i)}-\boldsymbol{y}}{\eta^2}\\&=\frac{\bar{\boldsymbol{x}}_c(\boldsymbol{y},\eta)-\boldsymbol{y}}{\eta^2}.\end{aligned} \tag{C.22}$$

To transfer the population bound, let $e_c(\boldsymbol{y},\eta):=\|\nabla\log\widehat{f}_{c,\eta}(\boldsymbol{y})-\nabla\log f_{c,\eta}(\boldsymbol{y})\|_2$. The triangle inequality gives

$$\begin{aligned}\left\|\nabla\log\widehat{f}_{c,\eta}(\boldsymbol{y})-\nabla\log g_{c,\eta}(\boldsymbol{y})\right\|_2&\le e_c(\boldsymbol{y},\eta)+\|\nabla\log f_{c,\eta}(\boldsymbol{y})-\nabla\log g_{c,\eta}(\boldsymbol{y})\|_2\\&\le e_c(\boldsymbol{y},\eta)+\frac{\kappa_c}{\lambda_{\min}(\boldsymbol{\Sigma}_c)+\eta^2}.\end{aligned} \tag{C.23}$$

The population assumption does not itself control e_c.

In particular, fix $\boldsymbol{y}$ with a unique nearest observation $\boldsymbol{x}_c^{(i_*)}$. For every $j\ne i_*$,

$$\frac{\alpha_{c,j}(\boldsymbol{y},\eta)}{\alpha_{c,i_*}(\boldsymbol{y},\eta)}=\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{x}_c^{(j)}\|_2^2-\|\boldsymbol{y}-\boldsymbol{x}_c^{(i_*)}\|_2^2}{2\eta^2}\right)\longrightarrow 0\qquad\text{as }\eta\downarrow 0. \tag{C.24}$$

Hence $\alpha_{c,i_*} \to 1$ and $\bar{\boldsymbol{x}}_c \to \boldsymbol{x}_c^{(i_*)}$. By Eq. (C.22),

$$\eta^2 \nabla \log \widehat{f}_{c,\eta}(\boldsymbol{y}) \longrightarrow \boldsymbol{x}_c^{(i_*)} - \boldsymbol{y}.$$

This recovers attraction toward an individual observation. If $\boldsymbol{y} \neq \boldsymbol{x}_c^{(i_*)}$, the empirical score norm diverges as $\eta \downarrow 0$, whereas

$$\nabla \log g_{c,\eta}(\boldsymbol{y}) \longrightarrow -\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_c).$$

Thus a general empirical score-error bound cannot remain uniformly finite down to zero smoothing.

C.5 Contribution of competing classes

At a fixed smoothing scale $\eta \geq 0$, define $p_\eta(\boldsymbol{y}) := \sum_{b=1}^{C} w_b f_{b,\eta}(\boldsymbol{y})$ and $\omega_b(\boldsymbol{y}, \eta) := w_b f_{b,\eta}(\boldsymbol{y})/p_\eta(\boldsymbol{y})$, where $w_b > 0$ and $\sum_b w_b = 1$. Suppressing $(\boldsymbol{y}, \eta)$ where unambiguous,

$$\nabla \log p_\eta = \frac{\sum_b w_b f_{b,\eta} \nabla \log f_{b,\eta}}{p_\eta} = \sum_b \omega_b \nabla \log f_{b,\eta}, \tag{C.25}$$

$$\nabla \log p_\eta - \nabla \log f_{c,\eta} = \sum_{b \neq c} \omega_b \big(\nabla \log f_{b,\eta} - \nabla \log f_{c,\eta} \big). \tag{C.26}$$

If $\omega_c(\boldsymbol{y}, \eta) \geq 1 - \delta$, then $\sum_{b \neq c} \omega_b(\boldsymbol{y}, \eta) \leq \delta$. Consequently,

$$\begin{aligned} \|\nabla \log p_\eta(\boldsymbol{y}) - \nabla \log f_{c,\eta}(\boldsymbol{y})\|_2 &\leq \sum_{b \neq c} \omega_b(\boldsymbol{y}, \eta) \, \|\nabla \log f_{b,\eta}(\boldsymbol{y}) - \nabla \log f_{c,\eta}(\boldsymbol{y})\|_2 \\ &\leq \delta \max_{b \neq c} \|\nabla \log f_{b,\eta}(\boldsymbol{y}) - \nabla \log f_{c,\eta}(\boldsymbol{y})\|_2 . \end{aligned} \tag{C.27}$$

This is the additional error from replacing the full score by one class score. The same calculation holds for empirical class densities at $\eta > 0$, using their corresponding responsibilities.

Class approximation and competition therefore contribute separate errors. Prop. 2 controls competition for the Gaussian reference mixture. Applying its certificate to non-Gaussian classes additionally requires control of their actual responsibilities; score approximation alone does not identify them with the reference responsibilities.

C.6 When is a class one Gaussian mode?

The reference of Sec. 4.1 replaces a class by $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$. Fig. S7 illustrates the geometry retained by this approximation, while Fig. S8 shows the effect of reducing the separation between classes.

The chain of approximations. The full covariance retains the second-moment orientation and anisotropy of the class, giving the ellipsoidal regions in Eq. (C.3). The diagonal fit sets the off-diagonal entries of $\boldsymbol{\Sigma}_c$ to zero and retains the variance along each coordinate, giving ellipsoids aligned with the coordinate axes. The isotropic fit replaces $\boldsymbol{\Sigma}_c$ by $(\operatorname{tr} \boldsymbol{\Sigma}_c/d)\boldsymbol{I}_d$ and retains only the average variance, giving spherical regions.

These fits describe progressively less of the class geometry. None need reproduce its non-Gaussian structure, as illustrated by the curved class in Fig. S7. Each covariance choice defines a different reference score and requires its own residual bound in Prop. 1. Matching moments alone does not establish that bound. For the isotropic mixture, the dominance radius additionally depends on the component spread, competing weights, and nearest-center separation through Eq. (4).

When the reference fails. The Gaussian approximation and the dominance certificate require separate conditions. The first requires control of the population score residual; a class being unimodal or visually concentrated does not by itself imply that κ_c is finite. The second requires sufficient separation relative to the component spread and competing weights at the chosen tolerance, as characterized by Lemma 4.

In Fig. S8, moving the classes closer leaves their within-class covariances unchanged but increases competition between them. Intersecting contours illustrate this overlap; they do not establish that every point lacks a dominant class. If Eq. (4) gives zero, the bound supplies no positive-radius core at that noise level and tolerance. This does not by itself invalidate the Gaussian approximation or rule out dominance outside the certified construction.

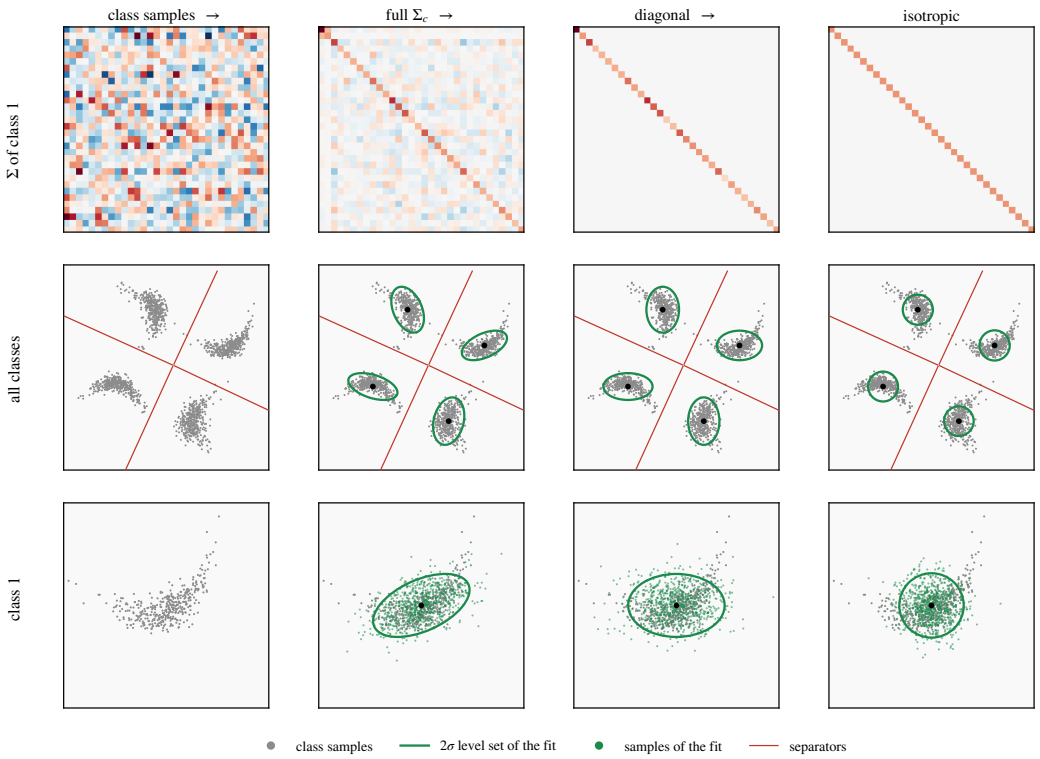


Figure S7: A class as a Gaussian mode. Columns show the class samples and Gaussian fits with full, diagonal, and isotropic covariance. The top row shows the observations of class 1 and the covariance matrices, with a common colour scale. The middle row shows four separated classes in the plane of their means, together with the displayed Gaussian contours and separators. The bottom row shows class 1 at its own scale, with samples drawn from each fit.

D Proof of Proposition 2

We use the Gaussian Mixture Model and fixed-center coordinates of Sec. B. Fix a component i, $\sigma > 0$, and $\delta \in (0, 1/2)$, and write $d_i := \min_{j \neq i} d_{ij}$ and $\Delta_i := \max_{j \neq i} d_{ij}$. The centers are distinct, the weights are positive and sum to one, and $C \geq 2$. The Gaussian-mixture score identity (Ashiq et al., 2026, Lemma G.1) gives $\nabla \log p_\sigma(\boldsymbol{y}) = (\boldsymbol{D}_\sigma(\boldsymbol{y}) - \boldsymbol{y})/\sigma^2$.

Proof. Let $0 \leq r \leq d_i/2$ and $\boldsymbol{y} \in B(\boldsymbol{\mu}_i, r)$. For every $j \neq i$,

$$\frac{\pi_j(\boldsymbol{y}, \sigma)}{\pi_i(\boldsymbol{y}, \sigma)} = \frac{w_j}{w_i} \exp\left(- \frac{\|\boldsymbol{y} - \boldsymbol{\mu}_j\|_2^2 - \|\boldsymbol{y} - \boldsymbol{\mu}_i\|_2^2}{2\sigma^2} \right), \tag{D.1}$$

$$\begin{aligned} \|\boldsymbol{y} - \boldsymbol{\mu}_j\|_2^2 - \|\boldsymbol{y} - \boldsymbol{\mu}_i\|_2^2 &= d_{ij}^2 - 2\langle \boldsymbol{y} - \boldsymbol{\mu}_i, \boldsymbol{\mu}_j - \boldsymbol{\mu}_i \rangle \\ &\geq d_{ij}^2 - 2r d_{ij} = d_{ij}(d_{ij} - 2r), \end{aligned} \tag{D.2}$$

$$d_{ij}(d_{ij} - 2r) - d_i(d_i - 2r) = (d_{ij} - d_i)(d_{ij} + d_i - 2r) \geq 0. \tag{D.3}$$

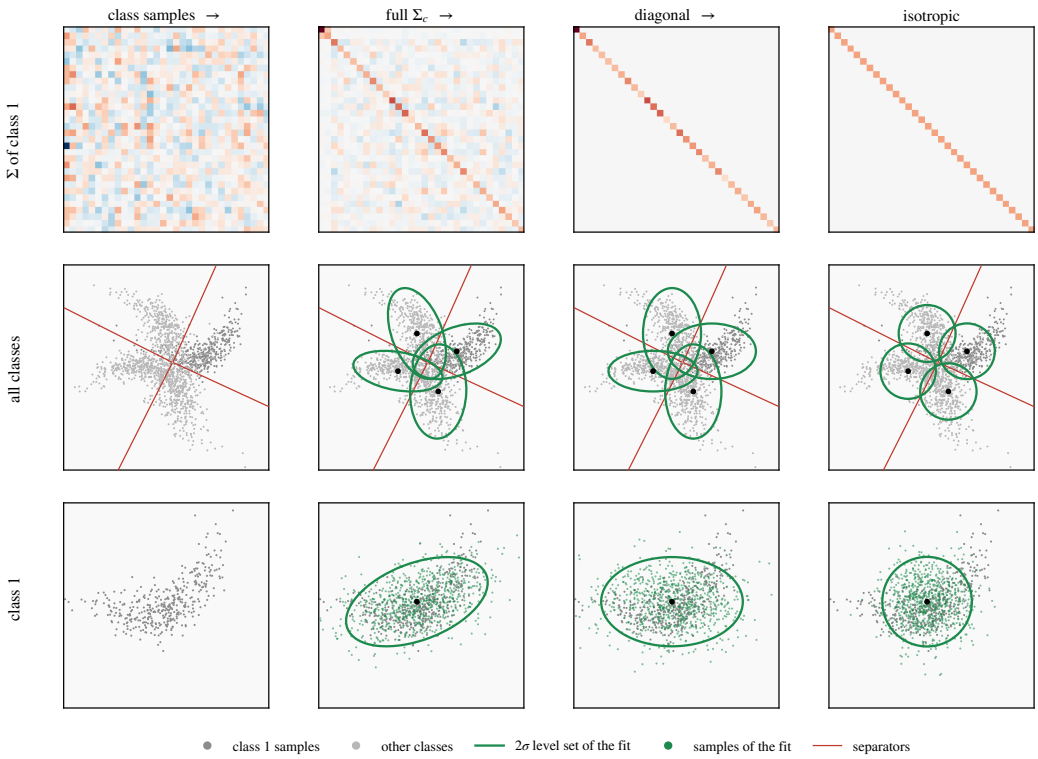


Figure S8: The same classes with reduced separation and unchanged within-class geometry. The displayed Gaussian contours now cross the separators and overlap those of neighboring classes.

Summing the odds and using $\sum_{j\neq i} w_j = 1 - w_i$,

$$\begin{aligned}\frac{1-\pi_i(\boldsymbol{y},\sigma)}{\pi_i(\boldsymbol{y},\sigma)} &= \sum_{j\neq i}\frac{\pi_j(\boldsymbol{y},\sigma)}{\pi_i(\boldsymbol{y},\sigma)} \\ &\leq \sum_{j\neq i}\frac{w_j}{w_i}\exp\left(-\frac{d_{ij}(d_{ij}-2r)}{2\sigma^2}\right) \\ &\leq \frac{1-w_i}{w_i}\exp\left(-\frac{d_i(d_i-2r)}{2\sigma^2}\right) =: \Psi_i(r,\sigma). \end{aligned} \tag{D.4}$$

Write $L_i(\delta) := \log((1-\delta)(1-w_i)/(\delta w_i))$. The sufficient odds condition is equivalent to

$$\begin{aligned}\Psi_i(r,\sigma) \leq \frac{\delta}{1-\delta} &\iff \log\frac{1-w_i}{w_i} - \frac{d_i(d_i-2r)}{2\sigma^2} \leq \log\frac{\delta}{1-\delta} \\ &\iff \frac{d_i(d_i-2r)}{2\sigma^2} \geq L_i(\delta) \\ &\iff r \leq \frac{d_i}{2} - \frac{\sigma^2}{d_i}L_i(\delta) =: \bar{r}_i(\sigma;\delta). \end{aligned} \tag{D.5}$$

Thus the radii certified by Eq. (D.4) form

$$\mathcal{R}_i(\sigma;\delta) = [0, d_i/2]\cap(-\infty, \bar{r}_i(\sigma;\delta)], \qquad r_i(\sigma;\delta) = \max\big(\mathcal{R}_i(\sigma;\delta)\cup\{0\}\big) = \min\left\{\frac{d_i}{2}, [\bar{r}_i(\sigma;\delta)]_+\right\}. \tag{D.6}$$

This is Eq. (4). If $\bar{r}_i < 0$, no radius is certified; if $\bar{r}_i = 0$, only the center is certified.

Suppose $r_i(\sigma;\delta) > 0$ and $\boldsymbol{y} \in B(\boldsymbol{\mu}_i, r_i(\sigma;\delta))$. Abbreviating $\pi_j := \pi_j(\boldsymbol{y},\sigma)$,

$$\frac{1-\pi_i}{\pi_i} \leq \frac{\delta}{1-\delta} \quad\Longrightarrow\quad \pi_i = \frac{1}{1+(1-\pi_i)/\pi_i} \geq 1-\delta. \tag{D.7}$$

Finally,

$$\left\| \frac{\boldsymbol{y} - \boldsymbol{D}_\sigma(\boldsymbol{y})}{\sigma} - \frac{\boldsymbol{y} - \boldsymbol{\mu}_i}{\sigma} \right\|_2 = \frac{1}{\sigma} \left\| \sum_{j \neq i} \pi_j (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) \right\|_2 \leq \frac{1}{\sigma} \sum_{j \neq i} \pi_j d_{ij} \leq \frac{\Delta_i}{\sigma}(1 - \pi_i) \leq \frac{\delta \Delta_i}{\sigma}. \tag{D.8}$$

□

D.1 Dependence on noise and tolerance

Lemma 4. *The radius $r_i(\sigma; \delta)$ is nondecreasing in δ and nonincreasing in σ. It is positive precisely when*

$$\frac{1 - w_i}{w_i} \exp\left(-\frac{d_i^2}{2\sigma^2}\right) < \frac{\delta}{1 - \delta}. \tag{D.9}$$

If $L_i(\delta) \leq 0$, then $r_i(\sigma; \delta) = d_i/2$. If $L_i(\delta) > 0$, then

$$r_i(\sigma; \delta) > 0 \quad \iff \quad \sigma < \sigma_i^\dagger(\delta) := \frac{d_i}{\sqrt{2L_i(\delta)}}. \tag{D.10}$$

For fixed $\sigma > 0$, the radius is zero for all sufficiently small $\delta > 0$.

Proof. By Eqs. (D.5) and (D.6),

$$r_i(\sigma; \delta) > 0 \iff \bar{r}_i(\sigma; \delta) > 0 \iff L_i(\delta) < \frac{d_i^2}{2\sigma^2} \iff \frac{1 - w_i}{w_i} \exp\left(-\frac{d_i^2}{2\sigma^2}\right) < \frac{\delta}{1 - \delta}.$$

Moreover,

$$L_i'(\delta) = -\frac{1}{\delta(1 - \delta)}, \qquad \frac{\partial \bar{r}_i}{\partial \delta} = \frac{\sigma^2}{d_i \delta(1 - \delta)} > 0,$$
$$\frac{\partial \bar{r}_i}{\partial \sigma} = -\frac{2\sigma}{d_i} L_i(\delta), \qquad L_i(\delta) \longrightarrow +\infty \quad \text{as } \delta \downarrow 0.$$

Clipping $\bar{r}_i$ to $[0, d_i/2]$ preserves monotonicity in δ and gives zero for all sufficiently small δ. If $L_i \leq 0$, then $\bar{r}_i \geq d_i/2$ and $r_i = d_i/2$. If $L_i > 0$, then $\bar{r}_i$ decreases with σ, and $\bar{r}_i > 0$ is equivalent to Eq. (D.10). □

At the center, the exact responsibility is

$$\pi_i(\boldsymbol{\mu}_i, \sigma) = \left[1 + \sum_{j \neq i} \frac{w_j}{w_i} \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right)\right]^{-1} < 1. \tag{D.11}$$

Thus, if $\delta < 1 - \pi_i(\boldsymbol{\mu}_i, \sigma)$, even the center fails the dominance condition. A zero certified radius alone does not imply this failure. Along the diffusion schedule, $\sigma \geq \sigma_0 > 0$, so the largest certified radius is $r_i := r_i(\sigma_0; \delta)$; the limit $\sigma \downarrow 0$ is not the data endpoint.

D.2 Terminal level sets and transported boundaries

For $\boldsymbol{\Sigma}_i = \sigma_0^2 \boldsymbol{I}_d$, Eq. (C.3) gives, for every $r \geq 0$,

$$\mathcal{E}_i(r/\sigma_0) = \left\{ \boldsymbol{x} : \frac{\|\boldsymbol{x} - \boldsymbol{\mu}_i\|_2^2}{\sigma_0^2} \leq \frac{r^2}{\sigma_0^2} \right\} = B(\boldsymbol{\mu}_i, r). \tag{D.12}$$

In particular, $\mathcal{C}_i^\bullet = \mathcal{E}_i(r_i/\sigma_0)$ when $r_i > 0$.

Lemma 5 (Isolated-mode transport)**.** *Under $d\boldsymbol{Y}_\sigma/d\sigma = (\boldsymbol{Y}_\sigma - \boldsymbol{\mu}_i)/\sigma$, the terminal core and its boundary map to $B(\boldsymbol{\mu}_i, R_i(\sigma))$ and its boundary, respectively, where*

$$R_i(\sigma) = \frac{\sigma}{\sigma_0} r_i. \tag{D.13}$$

Directions from $\boldsymbol{\mu}_i$ are preserved for trajectories away from the center.

Proof. The isolated solution in Eq. (C.19) gives

$$\boldsymbol{Y}_\sigma - \boldsymbol{\mu}_i = \frac{\sigma}{\sigma_0}(\boldsymbol{Y}_{\sigma_0} - \boldsymbol{\mu}_i),$$
$$\|\boldsymbol{Y}_\sigma - \boldsymbol{\mu}_i\|_2 = \frac{\sigma}{\sigma_0}\|\boldsymbol{Y}_{\sigma_0} - \boldsymbol{\mu}_i\|_2.$$

The scaling factor is positive and has inverse σ_0/σ; therefore it maps the ball and sphere onto the stated ball and sphere and preserves each nonzero radial direction. □

For $r_i > 0$ and $\sigma > \sigma_0$, Lemma 4 and Eq. (D.13) give

$$R_i(\sigma) = \frac{\sigma}{\sigma_0} r_i > r_i = r_i(\sigma_0; \delta) \geq r_i(\sigma; \delta). \tag{D.14}$$

Thus the transported reference sphere lies outside the ball certified at scale σ. Endpoint certification does not establish dominance along this transported boundary.

Remark. Eq. (D.14) says that dominance certified at the data end is not certified along the transported boundary. A trajectory ending at distance $\rho_0 \leq r_i$ from $\boldsymbol{\mu}_i$ is, under the isolated flow, at distance $(\sigma/\sigma_0)\rho_0$ at scale σ, inside $B(\boldsymbol{\mu}_i, r_i(\sigma; \delta))$ only while $(\sigma/\sigma_0)\rho_0 \leq r_i(\sigma; \delta)$; points ending near $\partial \mathcal{C}_i^\bullet$ leave the certified ball first.

D.3 Sharpness within the geometric cap

Lemma 6. *Let $j^\star \in \arg\min_{j \neq i} d_{ij}$, and define*

$$r_i^{\mathrm{ex}}(\sigma; \delta) := \sup\left(\left\{r \in [0, d_i/2] : \inf_{\boldsymbol{y} \in B(\boldsymbol{\mu}_i, r)} \pi_i(\boldsymbol{y}, \sigma) \geq 1 - \delta\right\} \cup \{0\}\right). \tag{D.15}$$

Then

$$r_i(\sigma; \delta) \leq r_i^{\mathrm{ex}}(\sigma; \delta) \leq \min\left\{\frac{d_i}{2}, \left[\bar{r}_i(\sigma; \delta) + \frac{\sigma^2}{d_i}\log\frac{1 - w_i}{w_{j^\star}}\right]_+\right\}. \tag{D.16}$$

For $C = 2$, $r_i^{\mathrm{ex}}(\sigma; \delta) = r_i(\sigma; \delta)$. The zero value in Eq. (D.15) *is a convention and need not certify the center.*

Proof. For $r \geq 0$, set $\boldsymbol{y}_r := \boldsymbol{\mu}_i + r(\boldsymbol{\mu}_{j^\star} - \boldsymbol{\mu}_i)/d_i$. By Eq. (D.1),

$$\begin{aligned}\frac{1 - \pi_i(\boldsymbol{y}_r, \sigma)}{\pi_i(\boldsymbol{y}_r, \sigma)} &\geq \frac{\pi_{j^\star}(\boldsymbol{y}_r, \sigma)}{\pi_i(\boldsymbol{y}_r, \sigma)} \\ &= \frac{w_{j^\star}}{w_i}\exp\left(-\frac{d_i(d_i - 2r)}{2\sigma^2}\right).\end{aligned} \tag{D.17}$$

If the radius r satisfies the exact dominance condition, then

$$\begin{aligned}\frac{w_{j^\star}}{w_i}\exp\left(-\frac{d_i(d_i - 2r)}{2\sigma^2}\right) &\leq \frac{\delta}{1 - \delta}, \\ r &\leq \frac{d_i}{2} - \frac{\sigma^2}{d_i}\log\frac{(1 - \delta)w_{j^\star}}{\delta w_i} \\ &= \bar{r}_i(\sigma; \delta) + \frac{\sigma^2}{d_i}\log\frac{1 - w_i}{w_{j^\star}}.\end{aligned}$$

Taking the supremum, with the zero convention, gives the upper bound in Eq. (D.16). The lower bound follows from Prop. 2. For $C = 2$, $w_{j^\star} = 1 - w_i$, so the upper bound equals $r_i(\sigma; \delta)$. □

For several competitors, Eq. (D.4) bounds each direction separately before replacing its distance by d_i. The aggregate responsibility need not be minimized in the direction of a nearest center. The cap $d_i/2$ is a geometric restriction; the lemma does not characterize the largest unrestricted ball.

D.4 A DISTANCE-DEPENDENT VELOCITY BOUND

Lemma 7. *Suppose $r := r_i(\sigma;\delta) > 0$ and $\sigma^2 \le d_i^2/2$. For every $\boldsymbol{y} \in B(\boldsymbol{\mu}_i, r)$,*

$$\left\| \frac{\boldsymbol{y} - \boldsymbol{D}_\sigma(\boldsymbol{y})}{\sigma} - \frac{\boldsymbol{y} - \boldsymbol{\mu}_i}{\sigma} \right\|_2 \le \frac{\delta d_i}{(1-\delta)\sigma}. \tag{D.18}$$

The noise condition follows from $r > 0$ whenever $L_i(\delta) \ge 1$.

Proof. For $x \ge d_i$, let $\phi(x) := x\exp(-x(x-2r)/(2\sigma^2))$. Since $r \le d_i/2$,

$$x(x-r) \ge d_i(d_i - r) \ge d_i^2/2 \ge \sigma^2,$$
$$\phi'(x) = \exp\left(-\frac{x(x-2r)}{2\sigma^2}\right)\left(1 - \frac{x(x-r)}{\sigma^2}\right) \le 0.$$

Using $\pi_j \le \pi_j/\pi_i$ and Eqs. (D.1) and (D.2),

$$\begin{aligned}\|\boldsymbol{D}_\sigma(\boldsymbol{y}) - \boldsymbol{\mu}_i\|_2 &\le \sum_{j\neq i} \pi_j d_{ij} \le \sum_{j\neq i} \frac{w_j}{w_i}\phi(d_{ij}) \\ &\le \frac{1-w_i}{w_i}\phi(d_i) = d_i\Psi_i(r,\sigma) \le \frac{\delta d_i}{1-\delta}.\end{aligned}$$

Divide by σ to obtain Eq. (D.18). Finally, $r > 0$ and $L_i(\delta) \ge 1$ imply $\sigma^2 < d_i^2/(2L_i(\delta)) \le d_i^2/2$. □

This improves the bound in Eq. (D.8) when $\Delta_i > d_i/(1-\delta)$; otherwise the original bound is no larger.

E PROOF OF PROPOSITION 3

Proof. Let $w_i = w_j$ and $r := r_i = r_j \in (0, d_{ij}/2)$. By Eq. (D.13), transporting each core under its own isolated-mode field gives

$$\mathcal{B}_k(\sigma) := B\left(\boldsymbol{\mu}_k, \frac{\sigma}{\sigma_0}r\right), \qquad k \in \{i,j\}, \qquad \sigma \ge \sigma_0. \tag{E.1}$$

Write $R(\sigma) := (\sigma/\sigma_0)r$. Any point in both balls satisfies $d_{ij} \le \|\boldsymbol{y} - \boldsymbol{\mu}_i\|_2 + \|\boldsymbol{y} - \boldsymbol{\mu}_j\|_2 \le 2R(\sigma)$; conversely, the midpoint belongs to both balls when $2R(\sigma) \ge d_{ij}$. Since R increases strictly, first contact occurs at

$$2R(\sigma_{ij}^{\mathrm{c}}) = d_{ij}, \qquad \sigma_{ij}^{\mathrm{c}} = \frac{d_{ij}\sigma_0}{2r} > \sigma_0. \tag{E.2}$$

At this scale, a point in the intersection must satisfy

$$\begin{aligned}2\|\boldsymbol{y} - \boldsymbol{m}_{ij}\|_2^2 + \frac{d_{ij}^2}{2} &= \|\boldsymbol{y} - \boldsymbol{\mu}_i\|_2^2 + \|\boldsymbol{y} - \boldsymbol{\mu}_j\|_2^2 \\ &\le 2R(\sigma_{ij}^{\mathrm{c}})^2 = \frac{d_{ij}^2}{2}.\end{aligned} \tag{E.3}$$

Hence

$$\mathcal{B}_i(\sigma_{ij}^{\mathrm{c}}) \cap \mathcal{B}_j(\sigma_{ij}^{\mathrm{c}}) = \{\boldsymbol{m}_{ij}\}, \qquad \boldsymbol{m}_{ij} \in \partial\mathcal{B}_i(\sigma_{ij}^{\mathrm{c}}) \cap \partial\mathcal{B}_j(\sigma_{ij}^{\mathrm{c}}). \tag{E.4}$$

By Eq. (B.5), $\sigma(t)$ increases continuously from σ_0 to σ_a. Thus a unique contact time $t_{ij} \in (0, T]$ exists if and only if $\sigma_{ij}^{\mathrm{c}} \le \sigma_a$. Substituting the contact scale gives

$$\begin{aligned}\sigma_0^2 + \frac{1-\bar{\alpha}_{t_{ij}}}{\bar{\alpha}_{t_{ij}}} &= \frac{d_{ij}^2\sigma_0^2}{4r^2}, \\ \bar{\alpha}_{t_{ij}} &= \left[1 + \sigma_0^2\left(\frac{d_{ij}^2}{4r^2} - 1\right)\right]^{-1}.\end{aligned} \tag{E.5}$$

This proves Eq. (5); the schedule condition is equivalently $\bar{\alpha}_{t_{ij}} \in [\bar{\alpha}_T, 1)$. In the original coordinates,

$$\sqrt{\bar{\alpha}_t}\,\mathcal{B}_k(\sigma(t)) = B\big(\sqrt{\bar{\alpha}_t}\boldsymbol{\mu}_k, \sqrt{\bar{\alpha}_t}\,R(\sigma(t))\big)\,. \tag{E.6}$$

The contact point is therefore $\sqrt{\bar{\alpha}_{t_{ij}}}\boldsymbol{m}_{ij}$.

At contact, the outward sphere normals in fixed-center coordinates are

$$\frac{\boldsymbol{m}_{ij}-\boldsymbol{\mu}_i}{d_{ij}/2} = \boldsymbol{n}_{ij}, \qquad \frac{\boldsymbol{m}_{ij}-\boldsymbol{\mu}_j}{d_{ij}/2} = -\boldsymbol{n}_{ij}.$$

Their common tangent hyperplane is

$$H_{ij} := \{\boldsymbol{y} : \langle \boldsymbol{n}_{ij}, \boldsymbol{y}-\boldsymbol{m}_{ij}\rangle = 0\}\,. \tag{E.7}$$

Using $w_i = w_j$ in Eq. (D.1),

$$\begin{aligned}\log\frac{\pi_j(\boldsymbol{y},\sigma)}{\pi_i(\boldsymbol{y},\sigma)} &= \frac{\|\boldsymbol{y}-\boldsymbol{\mu}_i\|_2^2 - \|\boldsymbol{y}-\boldsymbol{\mu}_j\|_2^2}{2\sigma^2}\\ &= \frac{d_{ij}}{\sigma^2}\langle \boldsymbol{n}_{ij}, \boldsymbol{y}-\boldsymbol{m}_{ij}\rangle.\end{aligned} \tag{E.8}$$

Thus H_{ij} is the equal-responsibility separator at every scale. The density transformation in Eq. (B.9) gives $\pi_k^{\mathrm{VP}}(\sqrt{\bar{\alpha}_t}\boldsymbol{y}, t) = \pi_k(\boldsymbol{y}, \sigma(t))$. Consequently its original-coordinate form is

$$H_{ij}(t) := \big\{\boldsymbol{x} : \big\langle \boldsymbol{n}_{ij}, \boldsymbol{x} - \sqrt{\bar{\alpha}_t}\boldsymbol{m}_{ij}\big\rangle = 0\big\}\,. \tag{E.9}$$

At $t = t_{ij}$, this is the common tangent of the two scaled spheres. □

Equal endpoint radii. Equal weights give equal radii when $d_i = d_j$. For a mutually nearest pair, $d_i = d_j = d_{ij}$. Writing $L := \log((1-\delta)(1-w_i)/(\delta w_i)) > 0$, the positive radius in Eq. (4) satisfies

$$r = \frac{d_{ij}}{2} - \frac{\sigma_0^2 L}{d_{ij}} = \frac{d_{ij}}{2}(1-\epsilon_{ij}), \qquad \epsilon_{ij} := \frac{2\sigma_0^2 L}{d_{ij}^2} \in (0,1). \tag{E.10}$$

Hence

$$\sigma_{ij}^{\mathrm{c}} = \frac{\sigma_0}{1-\epsilon_{ij}}, \qquad \bar{\alpha}_{t_{ij}} = \big[1 + \sigma_0^2\big((1-\epsilon_{ij})^{-2} - 1\big)\big]^{-1}\,. \tag{E.11}$$

Thus $\sigma_{ij}^{\mathrm{c}}/\sigma_0 \to 1$ as $\epsilon_{ij} \to 0$. For an exact equal-weight pair, $L = \log((1-\delta)/\delta)$.

Scope of the contact statement. For $\sigma > \sigma_{ij}^{\mathrm{c}}$, the triangle inequality gives

$$B(\boldsymbol{m}_{ij}, R(\sigma) - d_{ij}/2) \subseteq \mathcal{B}_i(\sigma) \cap \mathcal{B}_j(\sigma).$$

These balls are transported by different isolated-mode fields. Their contact does not establish dominance along their boundaries or the onset of two-mode dominance. In particular, $R(\sigma) > r_i(\sigma;\delta)$ for $\sigma > \sigma_0$ by Eq. (D.14). The preimages of the disjoint terminal cores under a common endpoint map remain disjoint:

$$\mathcal{G}_\star^{-1}(\mathcal{C}_i^\bullet) \cap \mathcal{G}_\star^{-1}(\mathcal{C}_j^\bullet) = \mathcal{G}_\star^{-1}(\mathcal{C}_i^\bullet \cap \mathcal{C}_j^\bullet) = \emptyset.$$

E.1 Reduction under two-mode dominance

We use a distance margin as in Ashiq et al. (2026, Assumption 4.1), with explicit dependence on the mixture weights. At $(\boldsymbol{y}, \sigma)$, suppose that, for every $k \notin \{i,j\}$,

$$\|\boldsymbol{y}-\boldsymbol{\mu}_k\|_2^2 \geq \min_{a\in\{i,j\}} \|\boldsymbol{y}-\boldsymbol{\mu}_a\|_2^2 + 2\sigma^2\chi, \qquad \chi > 0. \tag{E.12}$$

Multiplication by $\bar{\alpha}_t$ gives the original-coordinate condition, with centers $\sqrt{\bar{\alpha}_t}\boldsymbol{\mu}_k$ and margin $2v_t\chi$. Abbreviate $\pi_k := \pi_k(\boldsymbol{y}, \sigma)$ and define

$$\rho_{ij} := \sum_{k\notin\{i,j\}} \pi_k, \qquad \widetilde{\pi}_a := \frac{\pi_a}{\pi_i + \pi_j}, \qquad \boldsymbol{D}_\sigma^{ij}(\boldsymbol{y}) := \widetilde{\pi}_i\boldsymbol{\mu}_i + \widetilde{\pi}_j\boldsymbol{\mu}_j, \tag{E.13}$$

where $a \in \{i,j\}$. Set

$$B_{ij} := \frac{1-w_i-w_j}{\min\{w_i,w_j\}}e^{-\chi}, \qquad \Delta_{ij} := \max_{k\notin\{i,j\}} \max\{d_{ik}, d_{jk}\}, \tag{E.14}$$

with $\Delta_{ij} := 0$ when $C = 2$.

Lemma 8 (Two-mode reduction)**.** *Under Eq.* (E.12)*,*

$$\rho_{ij} \le \frac{B_{ij}}{1+B_{ij}}, \qquad \|\boldsymbol{D}_\sigma(\boldsymbol{y}) - \boldsymbol{D}_\sigma^{ij}(\boldsymbol{y})\|_2 \le \Delta_{ij}\frac{B_{ij}}{1+B_{ij}}. \tag{E.15}$$

Consequently, the velocity error in the σ parameterization is

$$\left\| \frac{\boldsymbol{y}-\boldsymbol{D}_\sigma(\boldsymbol{y})}{\sigma} - \frac{\boldsymbol{y}-\boldsymbol{D}_\sigma^{ij}(\boldsymbol{y})}{\sigma} \right\|_2 \le \frac{\Delta_{ij}}{\sigma}\frac{B_{ij}}{1+B_{ij}}. \tag{E.16}$$

Proof. Choose $a \in \{i,j\}$ minimizing $\|\boldsymbol{y}-\boldsymbol{\mu}_a\|_2$. By Eqs. (D.1) and (E.12),

$$\begin{aligned} \frac{\pi_k}{\pi_a} &= \frac{w_k}{w_a}\exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{\mu}_k\|_2^2 - \|\boldsymbol{y}-\boldsymbol{\mu}_a\|_2^2}{2\sigma^2}\right) \le \frac{w_k}{w_a}e^{-\chi}, \qquad k \notin \{i,j\}, \\ \frac{\rho_{ij}}{1-\rho_{ij}} &= \frac{\sum_{k\notin\{i,j\}}\pi_k}{\pi_i+\pi_j} \le \sum_{k\notin\{i,j\}}\frac{\pi_k}{\pi_a} \le \frac{1-w_i-w_j}{w_a}e^{-\chi} \le B_{ij}. \end{aligned} \tag{E.17}$$

Thus $\rho_{ij} \le B_{ij}/(1+B_{ij})$. Moreover,

$$\begin{aligned} \boldsymbol{D}_\sigma - \boldsymbol{D}_\sigma^{ij} &= \sum_{k\notin\{i,j\}}\pi_k(\boldsymbol{\mu}_k - \boldsymbol{D}_\sigma^{ij}), \\ \|\boldsymbol{\mu}_k - \boldsymbol{D}_\sigma^{ij}\|_2 &\le \widetilde{\pi}_i d_{ik} + \widetilde{\pi}_j d_{jk} \le \Delta_{ij}, \\ \|\boldsymbol{D}_\sigma - \boldsymbol{D}_\sigma^{ij}\|_2 &\le \sum_{k\notin\{i,j\}}\pi_k\|\boldsymbol{\mu}_k - \boldsymbol{D}_\sigma^{ij}\|_2 \le \Delta_{ij}\rho_{ij} \le \Delta_{ij}\frac{B_{ij}}{1+B_{ij}}. \end{aligned} \tag{E.18}$$

Dividing by σ proves Eq. (E.16). For $C=2$, the excluded sum is empty and both errors vanish. □

E.2 Motion toward the pair segment

Let $S_{ij} := [\boldsymbol{\mu}_i, \boldsymbol{\mu}_j]$. In the increasing denoising parameter $u = \log(\sigma_a/\sigma)$ of Eq. (B.16), write

$$\frac{d\boldsymbol{Y}_u}{du} = \boldsymbol{D}_{\sigma_u}^{ij}(\boldsymbol{Y}_u) - \boldsymbol{Y}_u + \boldsymbol{e}_u, \qquad \boldsymbol{e}_u := \boldsymbol{D}_{\sigma_u}(\boldsymbol{Y}_u) - \boldsymbol{D}_{\sigma_u}^{ij}(\boldsymbol{Y}_u), \tag{E.19}$$

where $\sigma_u = \sigma_a e^{-u}$ and $u \in [0, u_f]$, with $u_f = \log(\sigma_a/\sigma_0)$.

Axial and transverse motion. For $w_i = w_j$, decompose $\boldsymbol{Y}_u = \boldsymbol{m}_{ij} + s_u\boldsymbol{n}_{ij} + \boldsymbol{z}_\perp(u)$, where $s_u = \langle \boldsymbol{n}_{ij}, \boldsymbol{Y}_u - \boldsymbol{m}_{ij}\rangle$ and $\boldsymbol{z}_\perp(u) \perp \boldsymbol{n}_{ij}$. By Eq. (E.8),

$$\begin{aligned} \widetilde{\pi}_j &= \frac{1}{1+\exp(-d_{ij}s_u/\sigma_u^2)}, \\ \boldsymbol{D}_{\sigma_u}^{ij}(\boldsymbol{Y}_u) &= \boldsymbol{m}_{ij} + \frac{d_{ij}}{2}(2\widetilde{\pi}_j - 1)\boldsymbol{n}_{ij} \\ &= \boldsymbol{m}_{ij} + \frac{d_{ij}}{2}\tanh\left(\frac{d_{ij}s_u}{2\sigma_u^2}\right)\boldsymbol{n}_{ij}. \end{aligned} \tag{E.20}$$

On $s_u = 0$, the conditional responsibilities equal $1/2$, while $\pi_i = \pi_j = (1-\rho_{ij})/2$. Projecting Eq. (E.19) gives the axial and transverse decomposition of Ashiq et al. (2026, Lemma G.3) in our coordinates:

$$\frac{ds_u}{du} = \frac{d_{ij}}{2}\tanh\left(\frac{d_{ij}s_u}{2\sigma_u^2}\right) - s_u + \langle \boldsymbol{n}_{ij}, \boldsymbol{e}_u\rangle, \tag{E.21}$$

$$\frac{d\boldsymbol{z}_\perp(u)}{du} = -\boldsymbol{z}_\perp(u) + (\boldsymbol{I}_d - \boldsymbol{n}_{ij}\boldsymbol{n}_{ij}^\top)\boldsymbol{e}_u. \tag{E.22}$$

An integrating factor and $\|\boldsymbol{I}_d - \boldsymbol{n}_{ij}\boldsymbol{n}_{ij}^\top\|_{\mathrm{op}} \le 1$ yield

$$\|\boldsymbol{z}_\perp(u)\|_2 \le e^{-(u-u_0)}\|\boldsymbol{z}_\perp(u_0)\|_2 + \int_{u_0}^{u} e^{-(u-v)}\|\boldsymbol{e}_v\|_2\,dv. \tag{E.23}$$

Lemma 9 (Contraction toward the pair segment)**.** *Let* $a(u) := \operatorname{dist}(\boldsymbol{Y}_u, S_{ij})$*. For arbitrary positive pair weights and* $0 \le u_0 \le u \le u_f$*,*

$$a(u) \le e^{-(u-u_0)} a(u_0) + \int_{u_0}^{u} e^{-(u-v)} \|\boldsymbol{e}_v\|_2 \, dv. \tag{E.24}$$

If Eq. (E.12) *holds at* $(\boldsymbol{Y}_v, \sigma_v)$ *for every* $v \in [u_0, u]$ *with the same* χ*, then*

$$a(u) \le e^{-(u-u_0)} a(u_0) + \Delta_{ij} \frac{B_{ij}}{1 + B_{ij}} \big(1 - e^{-(u-u_0)}\big). \tag{E.25}$$

Proof. Let $\boldsymbol{p}_0 \in S_{ij}$ satisfy $\|\boldsymbol{Y}_{u_0} - \boldsymbol{p}_0\|_2 = a(u_0)$. Integrating Eq. (E.19) gives

$$\boldsymbol{Y}_u = e^{-(u-u_0)} \boldsymbol{Y}_{u_0} + \int_{u_0}^{u} e^{-(u-v)} \big(\boldsymbol{D}^{ij}_{\sigma_v}(\boldsymbol{Y}_v) + \boldsymbol{e}_v\big) \, dv. \tag{E.26}$$

Since $\boldsymbol{D}^{ij}_{\sigma_v}(\boldsymbol{Y}_v) \in S_{ij}$ and $e^{-(u-u_0)} + \int_{u_0}^{u} e^{-(u-v)} \, dv = 1$, convexity gives

$$\boldsymbol{p}_u := e^{-(u-u_0)} \boldsymbol{p}_0 + \int_{u_0}^{u} e^{-(u-v)} \boldsymbol{D}^{ij}_{\sigma_v}(\boldsymbol{Y}_v) \, dv \in S_{ij}.$$

Therefore

$$\begin{aligned} a(u) &\le \|\boldsymbol{Y}_u - \boldsymbol{p}_u\|_2 \\ &= \left\| e^{-(u-u_0)} (\boldsymbol{Y}_{u_0} - \boldsymbol{p}_0) + \int_{u_0}^{u} e^{-(u-v)} \boldsymbol{e}_v \, dv \right\|_2 \\ &\le e^{-(u-u_0)} a(u_0) + \int_{u_0}^{u} e^{-(u-v)} \|\boldsymbol{e}_v\|_2 \, dv. \end{aligned} \tag{E.27}$$

Under the uniform margin, Lemma 8 bounds $\|\boldsymbol{e}_v\|_2 \le \Delta_{ij} B_{ij}/(1 + B_{ij})$ throughout the interval. Integration gives Eq. (E.25). □

F Proof of Proposition 4

Fix a mode i with $r_i > 0$ and write $\mathcal{G} := \mathcal{G}_{\boldsymbol{F}}$ and $\mathcal{C}_i^{\bullet} := B(\boldsymbol{\mu}_i, r_i)$. Under the assumptions of Lemma 1, $\mathcal{G}$ is a C^1 diffeomorphism of $\mathbb{R}^d$.

F.1 The pulled-back core and its boundary

Define

$$\mathcal{K}_i^{\boldsymbol{F}} := \mathcal{G}^{-1}(\mathcal{C}_i^{\bullet}), \tag{F.1}$$

$$h_i^{\boldsymbol{F}}(\boldsymbol{z}) := \|\mathcal{G}(\boldsymbol{z}) - \boldsymbol{\mu}_i\|_2^2 - r_i^2, \tag{F.2}$$

$$\Sigma_i^{\boldsymbol{F}} := \mathcal{G}^{-1}(\partial \mathcal{C}_i^{\bullet}). \tag{F.3}$$

Lemma 10 (Boundary of the pulled-back core)**.** *The set* $\mathcal{K}_i^{\boldsymbol{F}}$ *is compact,* $h_i^{\boldsymbol{F}} \in C^1(\mathbb{R}^d)$*, and*

$$\begin{aligned} \mathcal{K}_i^{\boldsymbol{F}} &= \{h_i^{\boldsymbol{F}} \le 0\}, \\ \operatorname{int} \mathcal{K}_i^{\boldsymbol{F}} &= \{h_i^{\boldsymbol{F}} < 0\}, \\ \partial \mathcal{K}_i^{\boldsymbol{F}} &= \Sigma_i^{\boldsymbol{F}} = \{h_i^{\boldsymbol{F}} = 0\}. \end{aligned} \tag{F.4}$$

Proof. The first identity follows from

$$\mathcal{G}(\boldsymbol{z}) \in \mathcal{C}_i^{\bullet} \iff \|\mathcal{G}(\boldsymbol{z}) - \boldsymbol{\mu}_i\|_2 \le r_i \iff h_i^{\boldsymbol{F}}(\boldsymbol{z}) \le 0.$$

Since $\mathcal{G}$ is a homeomorphism, it preserves interiors and boundaries. Hence

$$\begin{aligned} \operatorname{int} \mathcal{K}_i^{\boldsymbol{F}} &= \mathcal{G}^{-1}(\operatorname{int} \mathcal{C}_i^{\bullet}) = \{h_i^{\boldsymbol{F}} < 0\}, \\ \partial \mathcal{K}_i^{\boldsymbol{F}} &= \mathcal{G}^{-1}(\partial \mathcal{C}_i^{\bullet}) = \{h_i^{\boldsymbol{F}} = 0\}. \end{aligned}$$

The inverse map $\mathcal{G}^{-1}$ is continuous, so its image of the compact ball $\mathcal{C}_i^{\bullet}$ is compact. Finally, $h_i^{\boldsymbol{F}}$ is the composition of $\mathcal{G}$ with the polynomial $\boldsymbol{x} \mapsto \|\boldsymbol{x} - \boldsymbol{\mu}_i\|_2^2 - r_i^2$. □

F.2 Normal directions of the pulled-back boundary

Fix $\boldsymbol{z}_a \in \Sigma_i^{\boldsymbol{F}}$ and write $\boldsymbol{x}_a := \mathcal{G}(\boldsymbol{z}_a)$, $J_a := D\mathcal{G}(\boldsymbol{z}_a)$, $\boldsymbol{\nu}_a := (\boldsymbol{x}_a - \boldsymbol{\mu}_i)/r_i$, and $P_a := \boldsymbol{I}_d - \boldsymbol{\nu}_a\boldsymbol{\nu}_a^\top$. Since $\|\boldsymbol{x}_a - \boldsymbol{\mu}_i\|_2 = r_i$, we have $\|\boldsymbol{\nu}_a\|_2 = 1$. Let $\rho_i(\boldsymbol{x}) := \|\boldsymbol{x} - \boldsymbol{\mu}_i\|_2$ denote the terminal radius.

Proof of Prop. 4. By the chain rule,

$$\begin{aligned} \nabla h_i^{\boldsymbol{F}}(\boldsymbol{z}) &= 2D\mathcal{G}(\boldsymbol{z})^\top(\mathcal{G}(\boldsymbol{z}) - \boldsymbol{\mu}_i), \\ \nabla h_i^{\boldsymbol{F}}(\boldsymbol{z}_a) &= 2r_i J_a^\top \boldsymbol{\nu}_a \neq \boldsymbol{0}. \end{aligned} \tag{F.5}$$

The last inequality follows from the invertibility of J_a and $\|\boldsymbol{\nu}_a\|_2 = 1$. The implicit function theorem (Rudin, 1976, Chapter 9) therefore gives a C^1 boundary with tangent space

$$\begin{aligned} T_{\boldsymbol{z}_a}\Sigma_i^{\boldsymbol{F}} &= \ker Dh_i^{\boldsymbol{F}}(\boldsymbol{z}_a) \\ &= \{\boldsymbol{v} : \langle J_a^\top \boldsymbol{\nu}_a, \boldsymbol{v}\rangle = 0\} = \boldsymbol{n}_a^\perp, \qquad \boldsymbol{n}_a := \frac{J_a^\top \boldsymbol{\nu}_a}{\|J_a^\top \boldsymbol{\nu}_a\|_2}. \end{aligned} \tag{F.6}$$

For every fixed $\boldsymbol{v} \in \mathbb{R}^d$, differentiability gives

$$\begin{aligned} h_i^{\boldsymbol{F}}(\boldsymbol{z}_a + \varepsilon\boldsymbol{v}) &= h_i^{\boldsymbol{F}}(\boldsymbol{z}_a) + \varepsilon\langle \nabla h_i^{\boldsymbol{F}}(\boldsymbol{z}_a), \boldsymbol{v}\rangle + o(\varepsilon) \\ &= 2r_i\varepsilon\langle J_a^\top \boldsymbol{\nu}_a, \boldsymbol{v}\rangle + o(\varepsilon). \end{aligned} \tag{F.7}$$

Taking $\boldsymbol{v} = \pm\boldsymbol{n}_a$ yields

$$h_i^{\boldsymbol{F}}(\boldsymbol{z}_a \pm \varepsilon\boldsymbol{n}_a) = \pm 2r_i\|J_a^\top \boldsymbol{\nu}_a\|_2\varepsilon + o(\varepsilon), \qquad \varepsilon \to 0^+, \tag{F.8}$$

which proves Eq. (6). Since $2r_i\|J_a^\top \boldsymbol{\nu}_a\|_2 > 0$, for all sufficiently small $\varepsilon > 0$,

$$h_i^{\boldsymbol{F}}(\boldsymbol{z}_a - \varepsilon\boldsymbol{n}_a) < 0 < h_i^{\boldsymbol{F}}(\boldsymbol{z}_a + \varepsilon\boldsymbol{n}_a). \tag{F.9}$$

Thus $\boldsymbol{n}_a$ is the outward normal, and the two perturbations map strictly inside and outside the terminal core, respectively.

To establish optimality for the terminal radius, note that $\boldsymbol{x}_a \neq \boldsymbol{\mu}_i$ and $\nabla\rho_i(\boldsymbol{x}_a) = \boldsymbol{\nu}_a$. Consequently,

$$\begin{aligned} \nabla(\rho_i \circ \mathcal{G})(\boldsymbol{z}_a) &= J_a^\top \boldsymbol{\nu}_a, \\ \rho_i\big(\mathcal{G}(\boldsymbol{z}_a + \varepsilon\boldsymbol{v})\big) &= r_i + \varepsilon\langle J_a^\top \boldsymbol{\nu}_a, \boldsymbol{v}\rangle + o(\varepsilon). \end{aligned} \tag{F.10}$$

For $\|\boldsymbol{v}\|_2 = 1$, Cauchy–Schwarz gives

$$-\|J_a^\top \boldsymbol{\nu}_a\|_2 \le \langle J_a^\top \boldsymbol{\nu}_a, \boldsymbol{v}\rangle \le \|J_a^\top \boldsymbol{\nu}_a\|_2. \tag{F.11}$$

Equality on the left holds precisely when $\boldsymbol{v} = -\boldsymbol{n}_a$. Hence $-\boldsymbol{n}_a$ uniquely maximizes the first-order decrease in terminal radius among unit seed perturbations. □

F.3 Tangent motion along the pulled-back boundary

Lemma 11 (First-order tangent motion)**.** *For every fixed $\boldsymbol{v} \in \boldsymbol{n}_a^\perp$, as $\varepsilon \to 0$,*

$$h_i^{\boldsymbol{F}}(\boldsymbol{z}_a + \varepsilon\boldsymbol{v}) = o(\varepsilon), \tag{F.12}$$

$$\mathcal{G}(\boldsymbol{z}_a + \varepsilon\boldsymbol{v}) = \boldsymbol{x}_a + \varepsilon P_a J_a \boldsymbol{v} + o(\varepsilon), \tag{F.13}$$

$$\rho_i\big(\mathcal{G}(\boldsymbol{z}_a + \varepsilon\boldsymbol{v})\big) = r_i + o(\varepsilon). \tag{F.14}$$

Proof. Since $\boldsymbol{v} \in \boldsymbol{n}_a^\perp$,

$$\langle \boldsymbol{\nu}_a, J_a\boldsymbol{v}\rangle = \langle J_a^\top \boldsymbol{\nu}_a, \boldsymbol{v}\rangle = 0.$$

Equations Eqs. (F.7) and (F.10) give Eqs. (F.12) and (F.14). Moreover,

$$P_a J_a \boldsymbol{v} = J_a\boldsymbol{v} - \boldsymbol{\nu}_a\langle \boldsymbol{\nu}_a, J_a\boldsymbol{v}\rangle = J_a\boldsymbol{v}.$$

Differentiating $\mathcal{G}$ at $\boldsymbol{z}_a$ therefore gives

$$\begin{aligned} \mathcal{G}(\boldsymbol{z}_a + \varepsilon\boldsymbol{v}) &= \boldsymbol{x}_a + \varepsilon J_a\boldsymbol{v} + o(\varepsilon) \\ &= \boldsymbol{x}_a + \varepsilon P_a J_a \boldsymbol{v} + o(\varepsilon), \end{aligned} \tag{F.15}$$

which proves Eq. (F.13). □

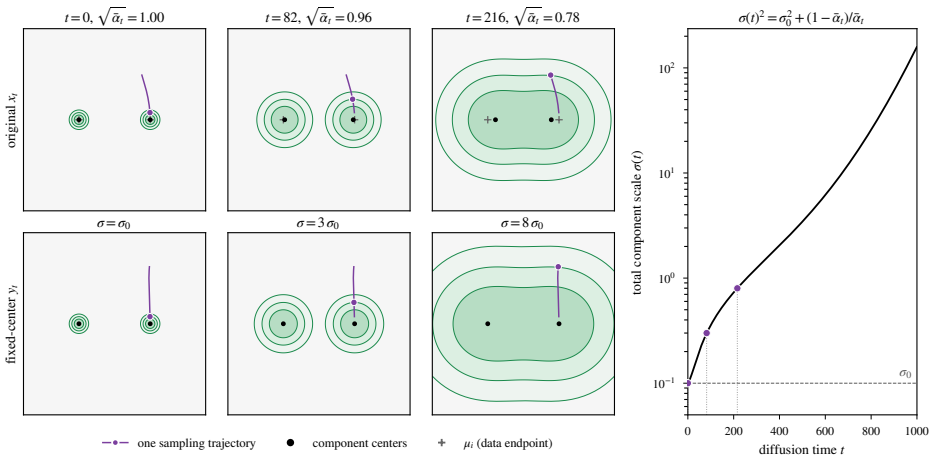


Figure S9: We demonstrate a correspondence between original time and normalized time for a reverse process

Remark 2. A straight tangent step preserves the terminal radius only to first order. For example, if $\mathcal{G}$ is the identity and $\boldsymbol{v} \perp \boldsymbol{\nu}_a$, then

$$h_i^{\boldsymbol{F}}(\boldsymbol{z}_a + \varepsilon \boldsymbol{v}) = \|r_i \boldsymbol{\nu}_a + \varepsilon \boldsymbol{v}\|_2^2 - r_i^2 = \varepsilon^2 \|\boldsymbol{v}\|_2^2. \tag{F.16}$$

Under C^1 regularity, the remainders above are $o(\varepsilon)$. A locally Lipschitz Jacobian suffices to strengthen them to $O(\varepsilon^2)$; in particular, this holds if $\mathcal{G}$ is C^2 near $\boldsymbol{z}_a$.

F.4 Alignment of seed corrections

For $\|\boldsymbol{v}\|_2 = 1$, define the first-order inward radial gain

$$\gamma_a(\boldsymbol{v}) := \lim_{\varepsilon \downarrow 0} \frac{r_i - \rho_i\big(\mathcal{G}(\boldsymbol{z}_a + \varepsilon \boldsymbol{v})\big)}{\varepsilon}. \tag{F.17}$$

Lemma 12 (Alignment)**.** *For every unit $\boldsymbol{v}$, the limit in Eq.* (F.17) *exists and satisfies*

$$\gamma_a(\boldsymbol{v}) = \|J_a^\top \boldsymbol{\nu}_a\|_2 \langle \boldsymbol{v}, -\boldsymbol{n}_a \rangle. \tag{F.18}$$

Consequently,

$$\mathrm{Align}_a(\boldsymbol{v}) := \frac{\gamma_a(\boldsymbol{v})}{\max_{\|\boldsymbol{u}\|_2 = 1} \gamma_a(\boldsymbol{u})} = \langle \boldsymbol{v}, -\boldsymbol{n}_a \rangle = \cos \angle(\boldsymbol{v}, -\boldsymbol{n}_a). \tag{F.19}$$

Proof. By Eq. (F.10),

$$\begin{aligned} \gamma_a(\boldsymbol{v}) &= -\langle J_a^\top \boldsymbol{\nu}_a, \boldsymbol{v} \rangle \\ &= \|J_a^\top \boldsymbol{\nu}_a\|_2 \langle \boldsymbol{v}, -\boldsymbol{n}_a \rangle. \end{aligned}$$

Since $\|\boldsymbol{n}_a\|_2 = 1$,

$$\max_{\|\boldsymbol{u}\|_2 = 1} \gamma_a(\boldsymbol{u}) = \gamma_a(-\boldsymbol{n}_a) = \|J_a^\top \boldsymbol{\nu}_a\|_2 > 0.$$

Dividing gives Eq. (F.19). □

Alignment equals 1 for the optimal inward direction, 0 for tangent directions, and is negative for directions with an outward component. This comparison is local and uses equal Euclidean step lengths in seed space.

Table S1: In this result, we demonstrate that pullback recovers the K mode basins with 3 different seeds through varying backtracking resolution T. Seed-fate full accuracy (%): the fraction of held-out seeds whose predicted fate (one of the K modes, or hallucination) matches the fate of the sampler started at that seed, for the learned DDIM sampler and learned flow-matching sampler. The two subtables are the two processes on the same (K, d) grid, so a row reads across as the same cell under either process. Within each sub-table the columns inside a predictor block are $T \in \{100, 200, 500\}$. η is that sub table's sampler's own hallucination rate (%) on the cell, which is independent of T. Polar classifier is the best among all the classifiers. Each predictor is reported at its best anchor budget for that cell; best predictor per (K, d, T) in bold, within its own sub-table.

(a) DDIM

			kNN			w-kNN			quadratic			polar		
K	d	η	100	200	500	100	200	500	100	200	500	100	200	500
2	2	0.9	93.9	98.6	**99.6**	**98.9**	**98.9**	99.0	92.6	97.8	99.0	93.7	98.5	**99.6**
2	4	0.9	86.5	96.0	97.0	81.2	89.1	92.1	**86.7**	97.4	99.2	85.1	**97.5**	**99.5**
2	8	1.1	**81.0**	88.7	87.5	54.2	77.0	90.6	67.2	**95.7**	99.0	64.3	95.2	**99.4**
2	16	1.1	**87.2**	83.8	81.0	51.9	70.9	84.3	37.3	91.8	99.0	38.9	**92.7**	**99.3**
2	32	1.0	**91.0**	84.8	78.3	53.6	60.9	86.6	7.2	73.7	98.9	15.7	**85.5**	**99.3**
2	64	1.0	**91.8**	**86.4**	75.3	63.0	62.7	80.3	1.1	21.9	98.2	4.5	69.5	**99.2**
4	2	1.1	82.8	**90.1**	**99.1**	**84.1**	**90.1**	98.6	80.8	89.2	94.1	82.7	90.0	99.0
4	4	1.1	**80.3**	89.8	94.2	74.4	85.6	92.4	78.3	**90.5**	95.0	79.3	90.1	**98.6**
4	8	1.1	77.8	80.0	78.3	58.3	83.8	85.1	**78.7**	**93.5**	96.3	76.3	93.1	**98.7**
4	16	1.0	**80.6**	71.3	67.3	50.3	81.3	83.4	60.3	**94.8**	98.0	55.4	**94.8**	**98.9**
4	32	1.0	**83.0**	71.0	61.8	51.0	65.0	79.9	14.2	87.0	98.0	23.8	**89.9**	**98.5**
4	64	1.1	**85.4**	74.3	57.2	50.9	57.1	82.0	1.4	48.2	97.2	10.0	**76.4**	**98.0**
8	2	1.2	**93.1**	**95.9**	**98.3**	92.9	95.6	97.9	85.8	89.0	91.1	93.0	**95.9**	98.2
8	4	1.1	80.2	87.0	92.7	76.6	85.0	91.9	78.8	**88.0**	92.4	**80.4**	**88.0**	**97.7**
8	8	1.2	74.1	73.3	72.2	68.8	82.3	84.6	83.3	92.3	95.2	**83.7**	**92.8**	**97.9**
8	16	1.1	**71.4**	60.0	56.1	51.1	78.0	77.8	70.0	92.8	96.2	65.9	**93.3**	**97.9**
8	32	1.1	**76.3**	58.4	48.6	51.2	79.9	76.8	24.9	90.9	97.1	31.9	**92.0**	**97.8**
8	64	1.1	**79.5**	62.7	42.8	50.8	60.5	75.9	1.9	72.7	96.2	12.9	**81.9**	**96.8**
16	4	1.3	84.0	88.6	92.0	83.8	88.4	91.6	83.0	87.2	89.7	**85.0**	**90.6**	**96.8**
16	8	1.4	68.4	67.4	66.5	70.2	79.2	80.6	81.2	88.9	92.4	**82.3**	**89.6**	**96.6**
16	16	1.2	62.9	50.6	48.1	60.8	75.3	73.8	**77.8**	92.2	94.9	75.3	**92.9**	**96.7**
16	32	1.1	**68.7**	46.7	38.3	50.6	78.6	72.7	38.8	91.7	**96.0**	39.6	**92.1**	95.2
16	64	1.2	**73.5**	51.1	31.9	51.7	66.4	72.1	2.6	82.7	**95.6**	11.9	**85.0**	94.9

(b) Flow matching

			kNN			w-kNN			quadratic			polar		
K	d	η	100	200	500	100	200	500	100	200	500	100	200	500
2	2	0.9	**99.6**	**99.6**	**99.6**	99.0	99.0	99.0	99.1	99.1	99.1	**99.6**	**99.6**	**99.6**
2	4	0.9	96.9	96.9	96.9	93.9	95.3	95.3	99.3	99.3	99.3	**99.6**	**99.6**	**99.6**
2	8	1.0	87.6	87.6	87.6	91.4	91.4	91.4	99.2	99.2	99.2	**99.5**	**99.5**	**99.5**
2	16	0.9	81.2	81.2	81.2	86.8	86.8	86.8	99.1	99.1	99.1	**99.3**	**99.3**	**99.3**
2	32	0.9	77.9	77.9	77.9	84.0	85.7	85.7	99.0	99.0	99.0	**99.4**	**99.4**	**99.4**
2	64	0.8	74.2	74.2	74.2	84.1	84.3	84.3	98.5	98.5	98.5	**99.3**	**99.3**	**99.3**
4	2	1.1	**99.2**	**99.2**	**99.2**	98.7	98.7	98.7	94.2	94.2	94.2	99.1	99.1	99.1
4	4	1.1	94.2	94.2	94.2	92.0	91.1	91.6	95.1	95.1	95.1	**99.0**	**99.0**	**99.0**
4	8	1.1	78.3	78.3	78.3	88.8	88.7	88.7	96.3	96.3	96.3	**98.9**	**98.9**	**98.9**
4	16	0.9	67.3	67.3	67.3	84.0	84.0	84.0	98.1	98.1	98.1	**98.8**	**98.8**	**98.8**
4	32	1.0	61.6	61.5	61.5	81.2	81.2	81.2	97.9	97.9	97.9	**98.6**	**98.6**	**98.6**
4	64	0.9	56.0	56.0	56.0	78.9	79.1	79.2	97.4	97.4	97.4	**98.4**	**98.4**	**98.4**
8	2	1.1	**98.5**	**98.5**	**98.5**	98.1	98.1	98.1	91.3	91.3	91.3	98.4	**98.5**	**98.5**
8	4	1.2	92.7	92.7	92.7	91.9	91.9	91.9	92.4	92.4	92.4	**98.4**	**98.4**	**98.4**
8	8	1.3	72.0	71.9	71.9	84.9	84.9	84.7	95.2	95.2	95.2	**98.1**	**98.1**	**98.1**
8	16	1.0	56.2	56.1	56.1	77.8	77.8	77.7	96.2	96.2	96.2	**97.8**	**97.8**	**97.8**
8	32	1.0	48.6	48.5	48.5	76.5	76.5	76.5	97.1	97.1	97.1	**97.9**	**97.9**	**97.9**
8	64	0.8	42.0	41.9	41.9	74.7	74.7	74.8	96.6	96.6	96.6	**97.7**	**97.7**	**97.7**
16	4	1.3	92.1	92.0	92.0	92.0	91.1	92.0	89.9	89.9	89.9	**97.5**	**97.5**	**97.5**
16	8	1.4	66.3	66.2	66.2	81.7	81.7	81.7	92.4	92.4	92.4	**97.0**	**97.0**	**97.0**
16	16	1.1	47.9	47.9	47.9	74.5	74.1	74.0	94.7	94.7	94.7	**96.6**	**96.6**	**96.6**
16	32	1.0	38.1	38.0	38.0	71.9	71.9	71.7	**95.9**	**95.9**	**95.9**	95.0	94.9	95.0
16	64	0.8	31.4	31.4	31.4	71.5	71.4	71.4	**96.2**	**96.2**	**96.2**	95.4	95.5	95.5

G Experimental Details

This appendix specifies the experiments of Sec. 5, in which pullback predicts, before sampling: whether the seed hallucinates or whether it reaches some mode core $\mathcal{C}_k$. We describe the Gaussian-mixture targets and their mode cores (Sec. G.2), the pullback anchors and the evaluation protocol (Sec. G.1), the learned sampling fields whose outcomes are predicted (Sec. G.2), and the anchor classifiers (Sec. G.3). Sec. H extends the setup to the latent space of an image autoencoder.

Target distribution. The data distribution is a mixture of K isotropic Gaussians in $\mathbb{R}^d$,

$$p_{\text{data}}(\boldsymbol{x}) = \sum_{i=1}^{K} w_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \sigma_0^2 \boldsymbol{I}_d), \qquad \sigma_0 = 0.1. \tag{G.1}$$

Let $R_{99} = \sigma_0 \sqrt{\chi^2_d(0.99)}$ be the radius of the ball around a center that holds 99% of the mass of its component, where $\chi^2_d(0.99)$ is the 0.99-quantile of the chi-square distribution with d degrees of freedom. Centers are placed one at a time, uniformly on the sphere of radius $2\sqrt{d/2}$, and a proposal is kept only if it is at least $3R_{99}$ from every center already placed, so that $d_{ij} \geq 3R_{99}$ for every pair. In $d = 2$ at most 13 centers fit on the circle at this separation, so $(K, d) = (16, 2)$ is not evaluated. In the unweighted mixture $w_i = 1/K$; in the weighted one the K weights are drawn once from $U[0.3, 0.8]$ with a fixed seed and normalized, and each run assigns them to the centers in its own random order. The two variants use separately trained models.

Mode cores and outcomes. Following Sec. 6, the core of component k is $\mathcal{C}_k^{\bullet} = B(\boldsymbol{\mu}_k, R_{99})$; the separation $3R_{99}$ keeps the cores disjoint. The noised mixture is again a Gaussian mixture, so the exact sampling field $\boldsymbol{F}_\star$ has a closed form: for DDIM and the solvers that reuse its network it is the probability-flow field of the VP process of Sec. B with the exact score, and for flow matching it is the marginal velocity of the optimal-transport path of Sec. G.2. A trained network gives the learned field $\boldsymbol{F}_\theta$. Every sampler is deterministic, so a seed $\boldsymbol{z} \sim \gamma$ has one sample $\boldsymbol{x} = \mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z})$,

$\boldsymbol{F} \in \{\boldsymbol{F}_\star, \boldsymbol{F}_\theta\}$, and we label the seed by its outcome

$$y(\boldsymbol{z}) = \begin{cases} k, & \mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z}) \in \mathcal{C}_k^\bullet, \\ -1, & \mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z}) \notin \bigcup_k \mathcal{C}_k^\bullet \quad \text{(hallucination)}. \end{cases} \tag{G.2}$$

Thus $y(\boldsymbol{z}) = k$ exactly when $\boldsymbol{z} \in \mathcal{K}_k^{\boldsymbol{F}} = \mathcal{G}_{\boldsymbol{F}}^{-1}(\mathcal{C}_k^\bullet)$, and the hallucinating seeds are the band outside every $\mathcal{K}_k^{\boldsymbol{F}}$. Predicting the outcome from the seed is a $(K+1)$-class problem in seed space whose decision boundaries are the pulled-back boundaries $\partial\mathcal{K}_k^{\boldsymbol{F}}$.

G.1 Anchor Classification

Pullback anchors. The pulled-back cores have no closed form, so we describe them by *anchors*: seeds of known outcome, obtained by pullback of labelled points in data space (Algorithm 1). For a budget of b anchors per mode, we draw $n = b$ points uniformly in each mode core $\mathcal{C}_k^\bullet$ (label k) and $m = b/2$ points in the data-space band $R_{99} < \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2 \le R_{99} + w$ around it (label -1), with band width $w = 2\sigma_0$. Each point is $\boldsymbol{x} = \boldsymbol{\mu}_k + r\,\boldsymbol{g}/\|\boldsymbol{g}\|_2$, where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ makes the direction uniform on the unit sphere and the radius r is drawn with $U \sim U[0,1]$ as

$$r = R_{99}\,U^{1/d} \quad \text{(core)}, \qquad r = R_{99} + w\,U \quad \text{(band)}. \tag{G.3}$$

The exponent $1/d$ makes the core points uniform in volume. The band points are uniform in radius, which keeps them spread across the band in high dimension, where a volume-uniform draw would concentrate them at its outer edge. They supply hallucination labels directly outside the core boundaries $\partial\mathcal{C}_k^\bullet$, whose pullbacks $\partial\mathcal{K}_k^{\boldsymbol{F}_\star}$ the classifiers must locate. The band has a fixed width while R_{99} grows with d, so it narrows relative to the core, from $w/R_{99} = 0.66$ at $d = 2$ to 0.08 at $d = 512$. The offsets of the anchors from the centres are drawn with the same seed in every run.

The labelled points are pulled back by integrating $\boldsymbol{F}_\star$ from $\tau = 0$ to $\tau = 1$ on a grid of T time points, i.e. in $T - 1$ steps (Figure S10, top). For flow matching, the exact velocity is integrated with Heun's method on the uniform grid, so T changes only the discretisation. For DDIM, each step is a DDIM update whose noise prediction averages the exact $\boldsymbol{\epsilon}$ at both ends of the step, a Heun-type correction that makes the pullback second order (the learned DDIM sampler uses the plain first-order update). The DDIM pullback uses the linear schedule $\beta \in [10^{-4}, 0.02]$ discretised into T noise levels, $\bar\alpha_t = \prod_{s \le t}(1 - \beta_s)$ with $\beta_1, \dots, \beta_T$ equally spaced, so its terminal $\bar\alpha_T$ equals that of the learned sampler (the same schedule with 500 levels) only for $T = 500$ ($\bar\alpha_T = 8.0 \times 10^{-2}$, 6.4×10^{-3} and 5.1×10^{-4} for $T = 250, 500, 750$). Pullback never queries the learned model.

Reference outcomes. The test set consists of NK seeds $\boldsymbol{z} \sim \gamma$, $N = 50{,}000$, labelled by their outcome under the learned endpoint map $\mathcal{G}_\theta$ (Figure S10, bottom). The anchors come from $\boldsymbol{F}_\star$ and the reference outcomes from $\boldsymbol{F}_\theta$, so the evaluation measures how well the pulled-back cores of the exact field predict the trained sampler. The choice of T affects only the exact-field side of the evaluation: the anchors and the exact-field labels of the w-kNN calibration seeds (Sec. G.3).

Metrics. We report the accuracy over the $K+1$ outcome classes, the F1 score macro-averaged over the mode classes, and the F1 score of the hallucination class. Hallucinations are rare: averaged over the (K, d) grid and the runs, the learned samplers send 1.0% (DDIM) and 0.8% (flow matching) of the test seeds outside every mode core, close to the 1% of its mass that p_{data} itself places outside the cores, and 1.6–1.8% under Heun, RK45 and DPM-Solver++(2M), which reuse the DDIM network (up to 3.7% in a single run at $d = 512$). Accuracy is therefore dominated by seeds inside the pulled-back cores, and hallucination F1 is the relevant measure of how well the hallucination band is recovered. All results are means over three runs (seeds 0, 100 and 200). Each run draws its own centres, weight assignment, training set, network initialisation and test seeds, while the anchor offsets and the classifier initialisations use the same seeds in every run. The test seeds and the centres depend only on (K, d) and the run, so all samplers and both mixture variants are evaluated on the same seeds and centres, and comparisons between them are paired.

G.2 Diffusion Model on GMM

Diffusion with Different Solvers We use two generative processes, and three further solvers for one of them. All samplers are deterministic discretisations of Eq. (1), so each seed has a single outcome. As in Sec. 3.1, the seed is $\boldsymbol{z}$ at $\tau = 1$ and the sample is $\boldsymbol{x} = \mathcal{G}_\theta(\boldsymbol{z})$ at $\tau = 0$.

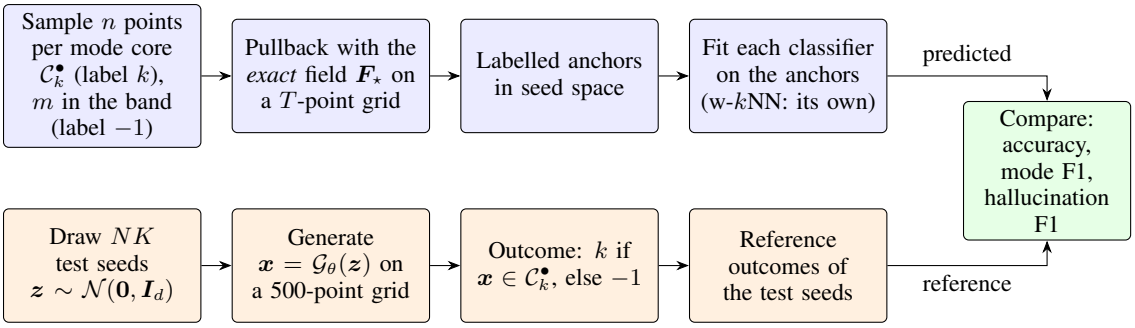


Figure S10: **Evaluation protocol.** Top: points in and around each mode core are pulled back to seed space with the exact field $\boldsymbol{F}_\star$, and the classifiers are fit on the resulting anchors (w-kNN on its own pulled-back, mode-labelled anchors, with its threshold calibrated on seeds labelled by $\mathcal{G}_\star$; Sec. G.3). Bottom: test seeds are labelled by their outcome under the learned endpoint map $\mathcal{G}_\theta$. The classifiers are evaluated on the same test seeds.

Algorithm 2 Evaluation protocol for one configuration (sampler, mixture variant, K, d, T, b).

```
# mu[0..K-1], w_mix : mode centres and weights, R99 : mode-core radius
# d_min : minimum centre separation, N : test seeds per mode, n_runs : runs
# F_star / F_theta : exact / learned field, G_theta : endpoint map of F_theta
# n, m : core and band anchors per mode, w : band width, T : pullback steps
# n_train, n_steps : training samples and steps, n_probe : probe seeds
# hall_max : retraining threshold, s_re : retraining step factor

def outcome(x, mu, R99):                          # k if x in C_k, -1 = hallucination
    D = cdist(x, mu)
    return where(min(D, axis=1) <= R99, argmin(D, axis=1), -1)

for run in range(n_runs):
    mu, w_mix = draw_centres(K, d, d_min), draw_weights(K)
    x_train   = sample_gmm(mu, w_mix, n_train)          # fixed training set
    F_theta   = train_field(x_train, n_steps)           # Heun/RK45/DPM++: DDIM's
    if hallucination_rate(G_theta, randn(n_probe, d)) > hall_max:
        F_theta = train_field(x_train, s_re * n_steps)   # once, same init

    z     = randn(N * K, d)                             # test seeds
    y_ref = outcome(G_theta(z), mu, R99)                # reference outcomes

    Z, y  = pullback(mu, R99, n, m, w, F_star, T)       # Alg. 1
    for clf in classifiers:                             # w-kNN: its own anchors
        y_hat = clf.fit(Z, y).predict(z)
        scores[run, clf] = metrics(y_hat, y_ref)        # accuracy, mode F1, hall. F1

return mean(scores, axis=0)
```

DDIM. The variance-preserving process of Sec. B, $\boldsymbol{X}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{X}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$, with β linear in $[10^{-4}, 0.02]$ over 500 noise levels. The network is trained with the standard $\boldsymbol{\epsilon}$-prediction objective at levels drawn uniformly from the 500 and sampled with the deterministic DDIM update. This update is the Euler discretisation of the probability-flow ODE, Eq. (B.12), in the fixed-centre coordinates $\boldsymbol{y} = \boldsymbol{x}/\sqrt{\bar{\alpha}_t}$ of Sec. B with the added noise scale $\eta_t = \sqrt{(1-\bar{\alpha}_t)/\bar{\alpha}_t}$ as time, where it reads $\mathrm{d}\boldsymbol{y}/\mathrm{d}\eta = \hat{\boldsymbol{\epsilon}}_\theta$: Eq. (2) reparameterised by η, with the network's noise prediction in place of the exact one. The sampler takes 499 steps between the 500 levels, from $\bar{\alpha} = 6.4 \times 10^{-3}$ ($\eta = 12.5$) at the seed to $\bar{\alpha} = 1 - 10^{-4}$ ($\eta = 0.01$), and returns the sample there, without a final denoising step.

Flow matching. The optimal-transport conditional path of Lipman et al. (2023) with $\sigma_{\min} = 10^{-4}$, written in model time, $\boldsymbol{X}_\tau = (1-\tau)\,\boldsymbol{x} + (\sigma_{\min} + (1-\sigma_{\min})\tau)\,\boldsymbol{z}$. The network regresses the conditional velocity $\boldsymbol{x} - (1-\sigma_{\min})\,\boldsymbol{z} = -\partial_\tau \boldsymbol{X}_\tau$ at $\tau \sim U[0,1]$, so it outputs $-\boldsymbol{F}_\theta$, and is sampled with 499 Euler steps on a uniform grid of 500 points in τ. Its step-index input is $\mathrm{round}(499(1-\tau))$, which is 0 at the seed.

Heun, RK45 and DPM-Solver++(2M). The trained DDIM network, integrated between the same 500 noise levels (499 steps) with a different solver. All three integrate $\mathrm{d}\boldsymbol{y}/\mathrm{d}\eta = \hat{\boldsymbol{\epsilon}}_\theta$ from $\eta = 12.5$ to 0.01: Heun's method (a DDIM predictor with a trapezoidal corrector; second order, 2 network evaluations per step), fixed-step Dormand–Prince RK45 (fifth order, 6 evaluations per step) and DPM-Solver++(2M) (Lu et al., 2022; a second-order multistep method on the data prediction in $\lambda = -\log\eta$ whose first step is a DDIM step; 1 evaluation per step). Heun and DPM-Solver++(2M) query the network only at the 500 levels it was trained on, whereas the four intermediate RK45 stages query it at fractional step indices, interpolated linearly in η between neighbouring levels. These samplers are not retrained: they share the DDIM network, its anchors and its calibration seeds, and differ from DDIM only in the reference outcomes, which isolates the effect of discretising $\boldsymbol{F}_\theta$ on the pulled-back cores.

Network. Every learned field uses the residual MLP architecture from Figure S11 each containing four pre-LayerNorm residual blocks with LeakyReLU(0.2) activations and sinusoidal step-index embeddings. The hidden width is $n_h = 256$ for $d \le 64$ and $n_h = 1024$ for $d \ge 128$. Training them requires a mean-squared-error objective with Adam optimizer with cosine learning-rate decay, and exponential moving average (EMA) weight tracking; probing and sampling are performed on the EMA weights (Table S2). The retraining rule is provided after Algorithm 2.

Table S2: **Learned field $\boldsymbol{F}_\theta$: architecture and training.** Identifiers are those of Figure S11 and Alg. 1.

Setting	Identifier	Value
Hidden width	n_h	256 for $d \le 64$; 1024 for $d \ge 128$
Residual blocks		4
Time embedding		sinusoidal, dimension 128, of the step index
Activation	ϕ	LeakyReLU(0.2)
Output		$\hat{\boldsymbol{\epsilon}}_\theta$ (DDIM) or $-\boldsymbol{F}_\theta$ (flow matching)
Training set	`n_train`	10^5 samples from p_{data}, fixed per run
Optimiser		Adam, learning rate 10^{-3}, cosine decay to 10^{-5}
Batch size		512
Gradient clipping		norm 1
Weight EMA		decay $\min(0.999, (1+s)/1001)$ at step s
Training steps	`n_steps`	$\lfloor 30{,}000\,(1+d/16)(1+K/16) \rfloor$
Probe seeds	`n_probe`	10^4, independent of the test seeds
Retraining threshold	`hall_max`	probe hallucination rate $> 1.5\%$
Retraining	`s_re`	once, same initialisation, $1.7\times$ the steps
Sampling grid		500 points (499 steps)

G.3 Algorithm Details

We compare four classifiers that map a seed $\boldsymbol{z}$ to one of the $K+1$ outcomes, given in Algorithms 3–7 and configured as in Table S3. kNN, the quadratic classifier and the polar classifier are fit on the core and band anchors of Sec. G.1. Weighted kNN (w-kNN) builds its own anchors by pulling back

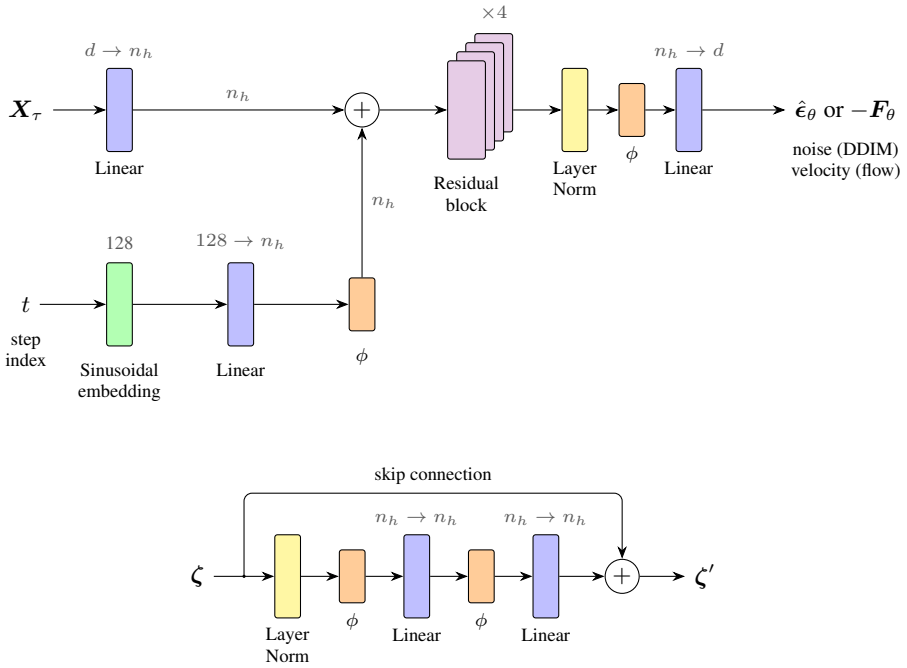


Figure S11: **Network of the learned field $\boldsymbol{F}_\theta$.** Top: the full network with hidden width n_h. The step index t is embedded sinusoidally (dimension 128), mapped to $\mathbb{R}^{n_h}$ by a linear layer and ϕ, and added to the input projection; for DDIM it is the level $t \in \{0, \dots, 499\}$ ($t = 499$ at the seed, fractional at the intermediate RK45 stages), and for flow matching it is $\mathrm{round}(499(1-\tau))$ (0 at the seed). Bottom: one pre-LayerNorm residual block acting on a hidden activation $\boldsymbol{\zeta} \in \mathbb{R}^{n_h}$. Box colours give the layer type: linear (blue), LayerNorm (yellow), activation $\phi = \mathrm{LeakyReLU}(0.2)$ (orange), embedding (green), residual block (violet); grey text gives the dimensions.

points on 5 concentric spheres of radii $0.3R_{99}, 0.6R_{99}, \ldots, 1.5R_{99}$ around each $\boldsymbol{\mu}_k$, all labelled with the mode k and none with -1 (the outer two spheres lie outside the core $\mathcal{C}_k^\bullet$), and detects hallucinations by the low confidence of its weighted vote. The quadratic and polar classifiers differ only in their logits and 6); both are trained with cross-entropy over the $K+1$ classes, as ensembles whose log-probabilities are summed (Algorithm 7).

Algorithm 3 kNN: seed $\to$ outcome.

```
# Z, y : pulled-back anchors, labels in {-1, 0..K-1};  z : test seeds
# n_nb : neighbours, eps : numerical floor
Y      = one_hot(y + 1, K + 1)                      # column 0 = hallucination
idx    = topk(cdist(z, Z), n_nb, largest=False)     # n_nb nearest anchors
lp     = log(mean(Y[idx], axis=1) + eps)            # log vote shares
margin = lp[:, 0] - max(lp[:, 1:], axis=1)          # hallucination margin
y_hat  = where(margin > 0, -1, argmax(lp[:, 1:], axis=1))
```

Algorithm 4 Weighted kNN (w-kNN) over mode-labelled anchors: seed $\to$ outcome.

```
# mu, R99, n, F_star, T : as in Alg. 1;  outcome, G_star : as in the protocol
# n_sph : spheres per mode, r_max : outermost radius (units of R99)
# w_min : weight of the outermost sphere, n_nb : neighbours, temp : temperature
# n_cal : calibration seeds;  n_q, [q_lo, q_hi] : threshold candidates
P, y, om = [], [], []
for k in range(K):
    for j in range(1, n_sph + 1):
        P  += [mu[k] + (j / n_sph) * r_max * R99 * unit_vectors(n // n_sph, d)]
        y  += [k]
        om += [1 - (1 - w_min) * j / n_sph]         # linear decay to w_min
Za = transport(P, F_star, T)                        # integration loop of Alg. 1

def vote(z):
    idx   = topk(cdist(z, Za), n_nb, largest=False)
    s     = sum(om[idx] * one_hot(y[idx], K), 1) / sum(om[idx], 1)
    probs = softmax(s / temp)
    conf  = 1 - entropy(probs) / log(K)
    return probs, conf

# calibration seeds, independent of the test seeds, labelled by the exact map
z_cal = randn(n_cal, d)
y_cal = outcome(G_star(z_cal), mu, R99)
_, c  = vote(z_cal)
c_thr = argmax_over(quantiles(c, q_lo .. q_hi, n_q),
                    lambda th: hall_f1(c < th, y_cal == -1))

probs, conf = vote(z)
y_hat = where(conf < c_thr, -1, argmax(probs, axis=1))
```

The calibration seeds are the same for every T and b of a run and are labelled by $\mathcal{G}_\star$ on the same T-point grid as the pullback; `c_thr` is recalibrated for every run, T and b.

Algorithm 5 Quadratic classifier: logits of a seed.

```
# z : seeds;  output : K + 1 logits, column 0 = hallucination
def forward(z):
    iu  = triu_indices(d, d)                        # pairs i <= j
    psi = concat([z, z[:, iu[0]] * z[:, iu[1]]])    # d + d(d+1)/2 features
    return Linear[d + d * (d + 1) // 2 -> K + 1](psi)
```

Each epoch reshuffles the anchors and drops the last incomplete batch, so a network takes between 3,000 optimiser steps (1,500 epochs at $K = 2$, $b = 2{,}000$) and 7,020 (30 epochs at $K = 16$, $b = 20{,}000$); there is no validation split or early stopping, and the final weights are used. The one-cycle schedule has the PyTorch defaults: the learning rate rises from `lr`$/25$ to `lr` over the

Algorithm 6 Polar classifier: logits of a seed.

```
# z : seeds;  n_hid : hidden size, deg : polynomial degree;  output : K + 1 logits
# alpha0, alpha1, log_beta : learned scalars of the hallucination logit
def forward(z):
    r0    = sqrt(d)                                  # typical seed radius
    r     = norm(z, axis=1, keepdims=True)
    r_off = r - r0                                   # radial offset
    v     = z / r                                    # direction
    hid   = Linear[d + 1 -> n_hid]([v * sqrt(r0), r_off]) / sqrt(r0 + 1)

    phi = concat([hid ** i for i in range(1, deg + 1)])
    l   = Linear[deg * n_hid -> K](phi)              # mode logits

    t1, t2 = top2(l)
    l_h = alpha0 + alpha1 * r_off - exp(log_beta) * (t1 - t2)   # hallucination logit
    return concat([l_h, l])
```

Algorithm 7 Training and prediction of the quadratic and polar classifiers.

```
# Z, y : pulled-back anchors;  arch in {quadratic, polar};  z : test seeds
# n_ens : ensemble size, n_batch : batch size, lr, wd : learning rate and weight decay
# e_min, s_min : minimum epochs and minimum optimiser steps
nets = []
for e in range(n_ens):
    net    = Classifier(d, K, arch, seed=e)
    epochs = max(e_min, ceil(s_min / (len(Z) // n_batch)))
    train(net, Z, y + 1, loss=cross_entropy, epochs=epochs, batch=n_batch,
          opt=AdamW(lr=lr, weight_decay=wd), sched="one-cycle")
    nets += [net]

lp     = sum(log_softmax(net(z)) for net in nets)
margin = lp[:, 0] - max(lp[:, 1:], axis=1)           # hallucination margin
y_hat  = where(margin > 0, -1, argmax(lp[:, 1:], axis=1))
```

first 30% of the steps and then decays to $\texttt{lr}/(2.5 \times 10^5)$, both along cosines. The linear layers have the PyTorch default initialisation, the scalars `alpha0`, `alpha1` and `log_beta` start at 0, and ensemble member $e \in \{0, 1, 2\}$ seeds its initialisation and batch order with e in every run.

The polar classifier uses the concentration of $\|\boldsymbol{z}\|_2$ around $\sqrt{d}$ for $\boldsymbol{z} \sim \gamma$, so the outcome depends mainly on the direction $\boldsymbol{v} = \boldsymbol{z}/\|\boldsymbol{z}\|_2$. Its hallucination logit grows as the two largest mode log-its approach each other, so hallucinations are predicted between two pulled-back cores, in a band whose width depends on the radial offset. This matches the geometry of Sec. 3.3, where the class boundaries arise from the contact of neighbouring cores.

Table S3: **Classifier settings.** Identifiers are those of Algorithms 3–7.

Classifier	Setting	Identifier	Value
kNN	neighbours	`n_nb`	10
	numerical floor	`eps`	10^{-9}
w-kNN	neighbours	`n_nb`	10
	spheres per mode	`n_sph`	5, radii $0.3j\,R_{99}$, $j = 1, \ldots, 5$
	anchors per sphere	`n // n_sph`	$b/5$ (Kb in total; $1.5Kb$ for the others)
	outermost radius	`r_max`	1.5 (units of R_{99})
	anchor weights	`w_min`	0.2; sphere j has weight $1 - 0.8\,j/5$
	softmax temperature	`temp`	0.1
	calibration seeds	`n_cal`	5,000, labelled by $\mathcal{G}_\star$
	threshold candidates	`n_q`, `q_lo`, `q_hi`	100 quantiles in $[0.5\%, 50\%]$
Quadratic	features		z_i and $z_i z_j$, $i \le j$
Polar	hidden size, degree	`n_hid`, `deg`	256, 3
Quadratic and polar	optimiser	`lr`	AdamW, one-cycle, peak 3×10^{-3}
	weight decay	`wd`	10^{-4}
	batch size	`n_batch`	2,048
	epochs, optimiser steps	`e_min`, `s_min`	at least 30 and at least 3,000
	ensemble size	`n_ens`	3

G.4 Result Tables

Note. We ran more experiments than fit in this appendix. The complete set of results is available, organized by experiment, in our repository.

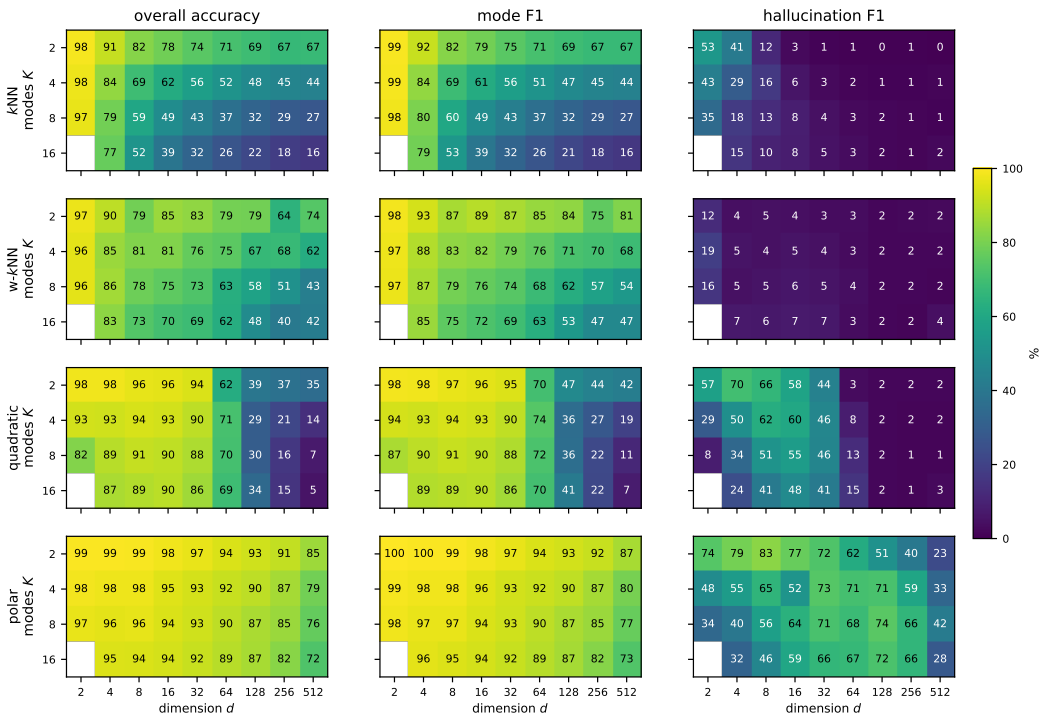


Figure S12: **DDIM, Weighted mixture,** $T = 500$**, 2,000 anchors per mode.** Full accuracy, mode F1 and hallucination F1 (%, mean over 3 seeds) per predictor, over K (rows) and d (columns). Empty cells were not run.

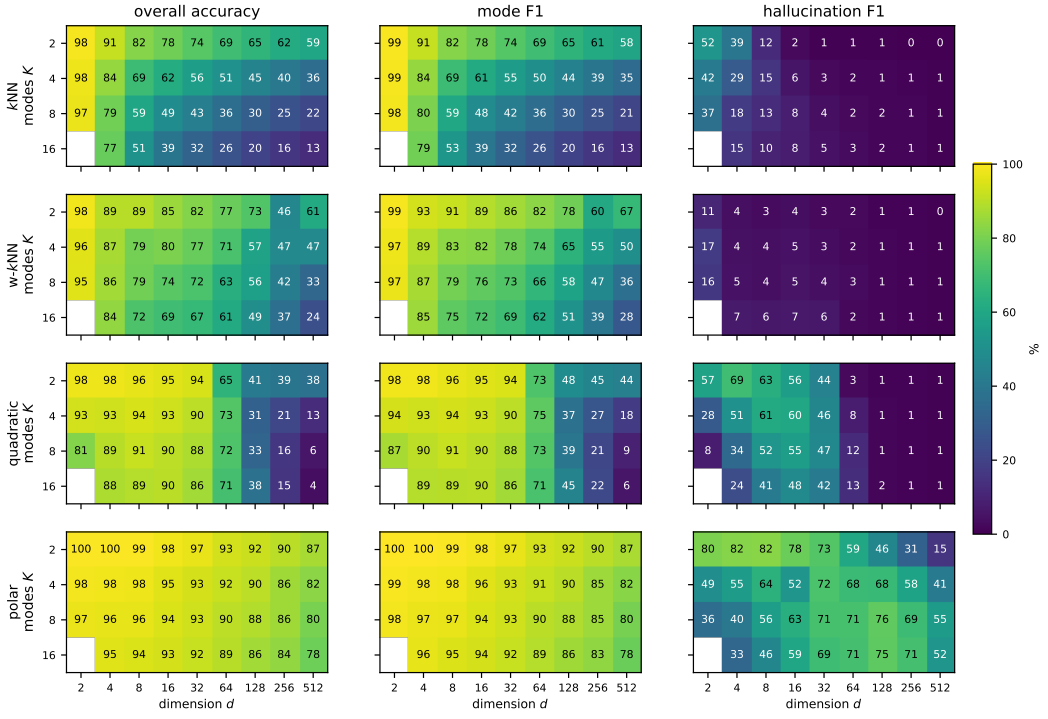


Figure S13: **Flow matching, Weighted mixture,** $T = 500$**, 2,000 anchors per mode.** Full accuracy, mode F1 and hallucination F1 (%, mean over 3 seeds) per predictor, over K (rows) and d (columns). Empty cells were not run.

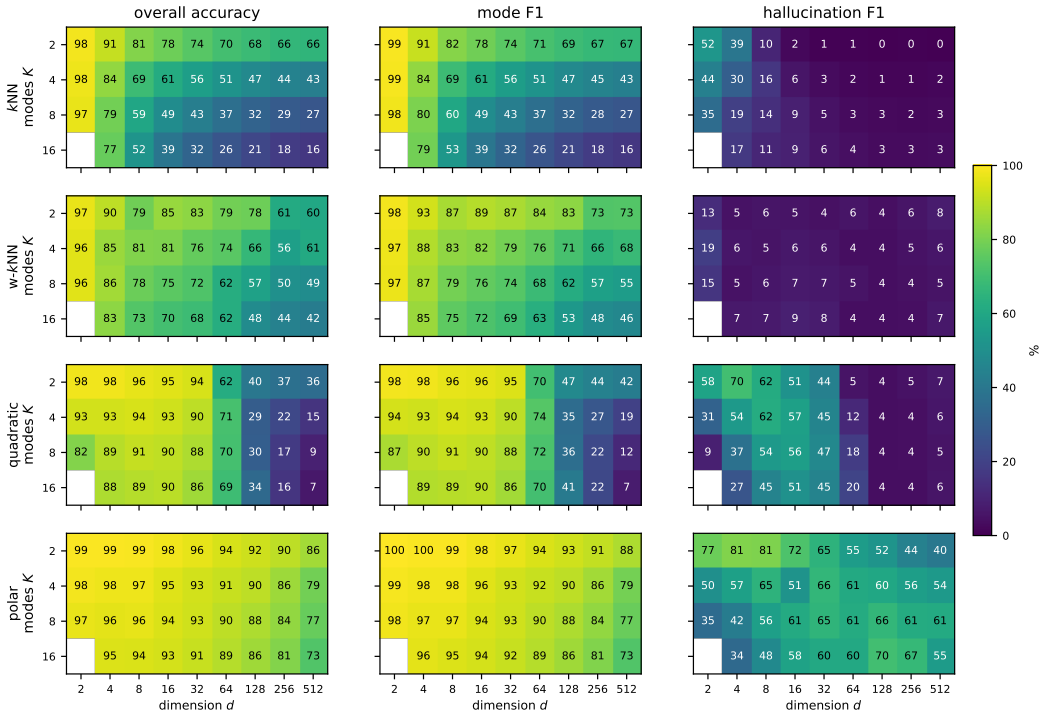


Figure S14: **dpmpp2m, Weighted mixture,** $T = 500$**, 2,000 anchors per mode.** Full accuracy, mode F1 and hallucination F1 (%, mean over 3 seeds) per predictor, over K (rows) and d (columns). Empty cells were not run.

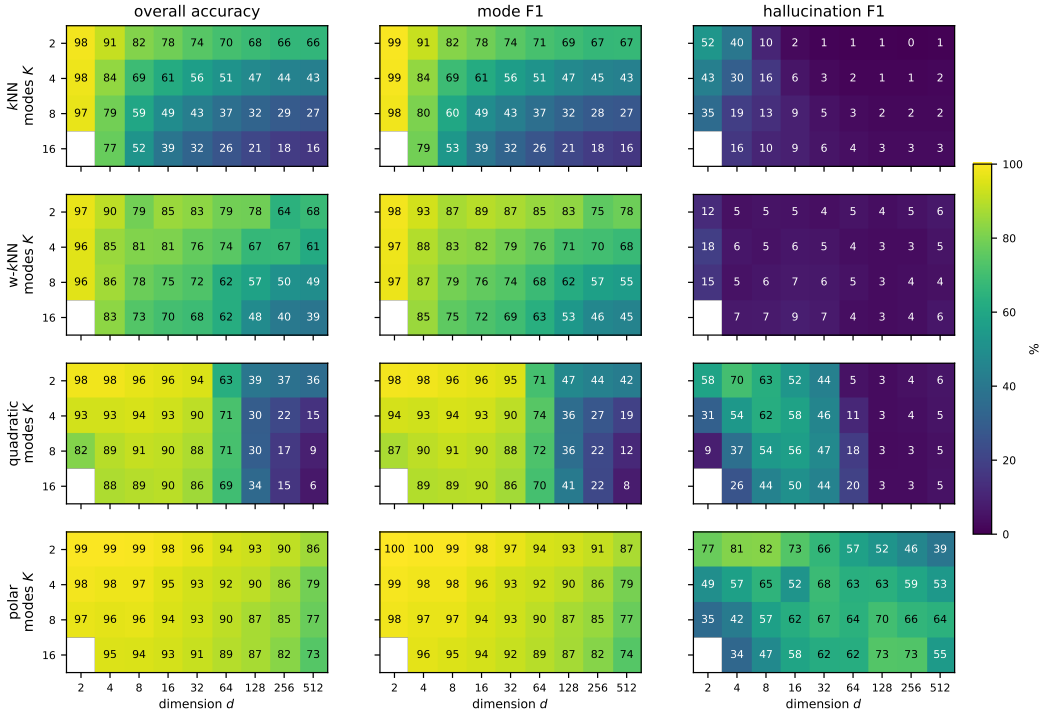


Figure S15: **Huen, Weighted mixture,** $T = 500$**, 2,000 anchors per mode.** Full accuracy, mode F1 and hallucination F1 (%, mean over 3 seeds) per predictor, over K (rows) and d (columns). Empty cells were not run.

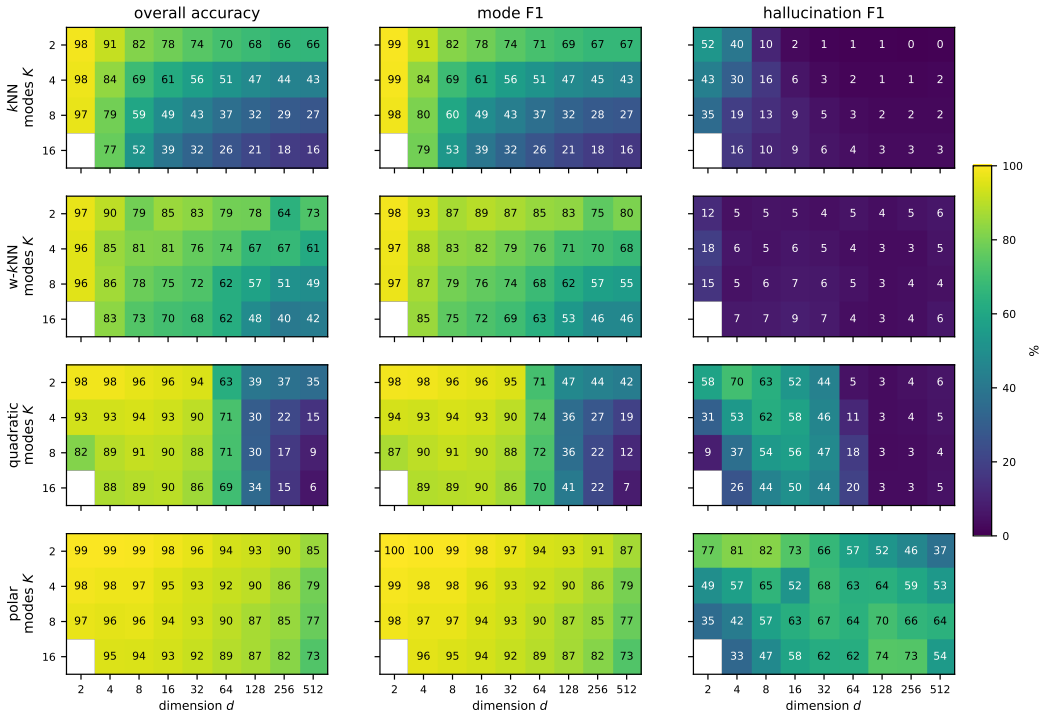


Figure S16: **rk45, Weighted mixture,** $T = 500$**, 2,000 anchors per mode.** Full accuracy, mode F1 and hallucination F1 (%, mean over 3 seeds) per predictor, over K (rows) and d (columns). Empty cells were not run.

Table S4: **Hallucination repair on the learned DDIM sampler.** Ground truth: the seeds, out of 200,000 per cell, that the sampler hallucinates on; *interp*: the endpoint lies within R_{99} of the segment joining two modes, *invalid*: near no such segment. Ours moves the seed by $\epsilon = 1.2$ along the pulled-back normal; IQ applies its guidance for $\tau \leq 0.6$; RODS corrects in the window $[0.1, 0.5]$ of the run. For ours and IQ we show the setting that leaves the fewest hallucinations summed over all cells. IQ's λ and RODS's ρ are chosen per cell.

		ground truth			**Ours** $\epsilon = 1.2$			**IQ** $\tau \leq 0.6$			**RODS-SAS**			**RODS-CAS**		
d	K	total	interp	invalid	left	int fix	inv fix	left	int fix	inv fix	left	int fix	inv fix	left	int fix	inv fix
2	2	2,034	921	1,113	37	917	1,080	1,597	434	3	649	506	879	579	580	875
2	4	2,221	1,527	694	27	1,523	671	1,724	453	44	462	1,243	516	405	1,298	518
2	8	2,216	1,817	399	31	1,807	378	1,519	692	5	265	1,666	285	277	1,651	288
4	2	2,035	694	1,341	65	690	1,280	1,480	419	136	500	515	1,020	456	560	1,019
4	4	2,282	1,298	984	72	1,276	934	1,233	971	78	382	1,169	731	386	1,169	727
4	8	2,332	1,636	696	84	1,606	642	1,476	815	41	336	1,499	497	346	1,493	493
4	16	2,694	2,187	507	88	2,152	454	1,968	725	1	480	1,881	333	459	1,900	335
8	2	2,210	503	1,707	56	499	1,655	1,827	198	185	454	457	1,299	449	463	1,298
8	4	2,181	1,047	1,134	89	1,039	1,053	1,640	409	132	453	913	815	436	928	817
8	8	2,372	1,488	884	104	1,467	801	1,879	383	110	505	1,261	606	459	1,300	613
8	16	2,618	1,950	668	91	1,933	594	2,035	582	1	602	1,622	394	575	1,646	397
16	2	2,065	434	1,631	57	434	1,574	1,509	356	200	452	375	1,238	391	431	1,243
16	4	2,106	773	1,333	79	773	1,254	1,472	506	128	523	633	950	381	742	983
16	8	2,250	1,124	1,126	91	1,121	1,038	1,715	420	115	521	970	759	420	1,053	777
16	16	2,357	1,507	850	49	1,504	804	2,037	235	85	582	1,222	553	419	1,368	570
32	2	1,965	346	1,619	43	345	1,577	1,449	195	321	341	300	1,324	277	346	1,342
32	4	2,019	628	1,391	63	628	1,328	1,510	289	220	517	452	1,050	266	624	1,129
32	8	2,098	987	1,111	84	984	1,030	1,731	260	107	666	693	739	277	983	838
32	16	2,285	1,263	1,022	71	1,263	951	1,910	295	80	716	951	618	320	1,244	721
64	2	2,067	340	1,727	61	340	1,666	1,499	220	348	482	280	1,305	344	340	1,383
64	4	2,019	543	1,476	49	543	1,427	1,448	320	251	504	437	1,078	303	542	1,174
64	8	2,175	774	1,401	49	774	1,352	1,661	353	161	691	569	915	271	766	1,138
64	16	2,402	1,071	1,331	71	1,071	1,260	1,804	449	149	939	713	750	346	1,052	1,004
128	2	1,736	249	1,487	24	249	1,463	1,074	211	451	266	209	1,261	215	249	1,272
128	4	1,473	325	1,148	24	325	1,124	923	233	317	211	291	971	164	323	986
128	8	1,470	501	969	32	501	937	943	312	215	250	421	799	154	498	818
128	16	1,316	542	774	24	542	750	805	338	173	196	473	647	116	535	665
256	2	1,803	230	1,573	29	230	1,544	997	222	584	271	213	1,319	233	230	1,340
256	4	1,517	306	1,211	18	306	1,193	824	263	430	223	274	1,020	184	303	1,030
256	8	1,576	417	1,159	31	417	1,128	915	314	347	386	330	860	195	410	971
256	16	1,315	423	892	25	421	869	704	359	252	230	375	710	136	420	759
total		63,209	27,851	35,358	**1,718**	27,680	33,811	**45,308**	12,231	5,670	**14,055**	22,913	26,241	**10,239**	25,447	27,523

Per method: left = ground-truth total − int fix − inv fix. int fix = interpolation hallucinations corrected; inv fix = invalid hallucinations corrected (sent into a mode's R_{99} ball).

H Additional Experiments

H.1 Hallucination Mitigation Table (for all d, K)

Mitigation strength and hallucination reduction. The one parameter of our correction is the step ϵ along $-\boldsymbol{n}_a$. In Tabs. S4 and S5 the columns for our method increase ϵ from left to right, and the share of corrected hallucinations grows with it in every cell, reaching 97.3% at $\epsilon = 1.2$ under both samplers. We separate the hallucinations into two kinds: interpolations, which lie within R_{99} of the segment joining two modes, and invalid samples, which lie near no such segment. The two kinds need very different steps. At $\epsilon = 0.02$, 62.6% of the interpolations are already corrected under DDIM and 61.9% under flow matching, and 90% of them by $\epsilon \approx 0.1$. Invalid samples need about seven times more: 6.0% and 3.6% at $\epsilon = 0.02$, and $\epsilon \approx 0.75$ for 90%. The step each kind needs varies less across (d, K) than between the two kinds, so what changes with the cell is their mix. Under DDIM, interpolations are 65.9% of the hallucinations at $d = 2$ and 22.2% at $d = 256$, 23.4% at $K = 2$ and 59.7% at $K = 16$; accordingly the share corrected at $\epsilon = 0.1$ falls from 70.9% at $d = 2$ to 38.3% at $d = 256$. That interpolations are the easy case follows from Sec. 4.2: an interpolating seed sits in the thin band between two neighbouring pulled-back cores, which the endpoint map sends to the tube around the segment between their centres, so a short step along $-\boldsymbol{n}_a$ moves it into a core.

Flow matching versus DDIM. Flow matching hallucinates less than DDIM, on 0.93% of the seeds against 1.05% on the cells both samplers share, so its hallucinating seeds form a thinner band. One might expect a smaller ϵ to suffice for it. It does not. On the 28 shared cells the share corrected

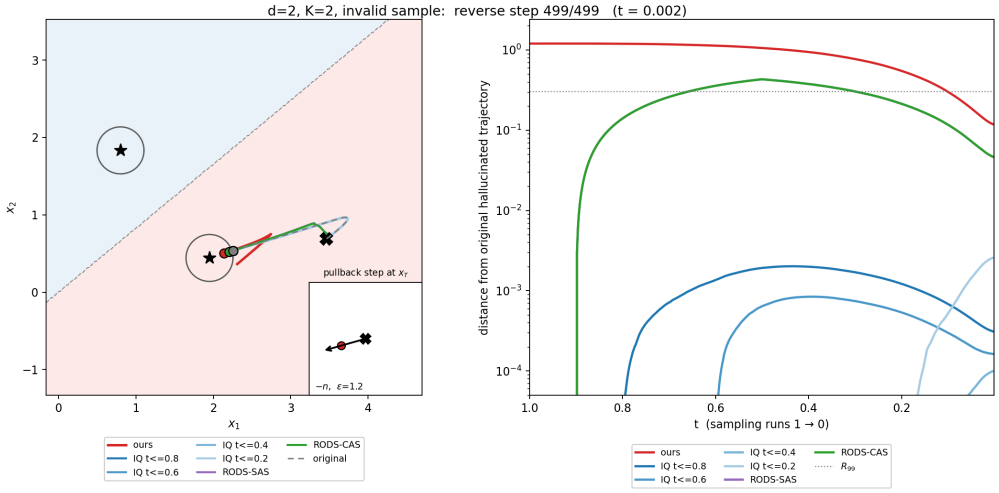
Table S5: **Hallucination repair on the learned flow matching sampler.** Ground truth: the seeds, out of 200,000 per cell, that the sampler hallucinates on; *interp*: the endpoint lies within R_{99} of the segment joining two modes, *invalid*: near no such segment. Ours moves the seed by $\epsilon = 1.2$ along the pulled-back normal; IQ applies its guidance for $\tau \leq 0.8$; RODS corrects in the window $[0.1, 0.5]$ of the run. For ours and IQ we show the setting that leaves the fewest hallucinations summed over all cells. IQ's λ and RODS's ρ are chosen per cell.

		ground truth			**Ours** $\epsilon = 1.2$			**IQ** $\tau \leq 0.8$			**RODS-SAS**			**RODS-CAS**		
d	K	total	interp	invalid	left	int fix	inv fix	left	int fix	inv fix	left	int fix	inv fix	left	int fix	inv fix
2	2	1,986	909	1,077	37	906	1,043	1,616	361	9	1,138	338	510	1,143	332	511
2	4	1,971	1,331	640	25	1,324	622	1,194	776	1	725	947	299	703	969	299
2	8	2,320	1,865	455	38	1,849	433	1,516	804	0	999	1,169	152	849	1,319	152
4	2	1,868	590	1,278	50	589	1,229	1,537	310	21	850	354	664	811	390	667
4	4	2,095	1,127	968	85	1,103	907	1,409	674	12	773	880	442	777	874	444
4	8	2,412	1,745	667	76	1,718	618	1,525	884	3	789	1,368	255	814	1,337	261
4	16	2,497	2,013	484	86	1,976	435	1,435	1,058	4	930	1,385	182	782	1,527	188
8	2	1,910	491	1,419	48	487	1,375	1,713	178	19	851	342	717	837	350	723
8	4	2,131	913	1,218	85	907	1,139	1,770	360	1	1,040	526	565	742	818	571
8	8	2,566	1,524	1,042	105	1,499	962	2,161	405	0	1,419	750	397	831	1,324	411
8	16	2,805	2,043	762	121	2,010	674	2,302	503	0	1,537	1,029	239	878	1,682	245
16	2	1,828	374	1,454	40	374	1,414	1,750	30	48	896	180	752	713	354	761
16	4	2,029	799	1,230	64	797	1,168	1,866	156	7	1,139	329	561	716	742	571
16	8	2,145	1,110	1,035	76	1,102	967	1,840	298	7	1,323	392	430	694	1,008	443
16	16	2,185	1,410	775	56	1,407	722	1,886	296	3	1,435	445	305	587	1,288	310
32	2	1,675	287	1,388	41	287	1,347	1,560	61	54	739	186	750	637	283	755
32	4	1,903	513	1,390	56	512	1,335	1,781	84	38	913	342	648	750	489	664
32	8	1,970	840	1,130	61	837	1,072	1,860	93	17	1,035	461	474	685	794	491
32	16	1,892	1,042	850	54	1,037	801	1,754	137	1	1,070	494	328	594	960	338
64	2	1,740	231	1,509	31	231	1,478	1,548	114	78	711	188	841	670	222	848
64	4	1,831	399	1,432	43	398	1,390	1,641	134	56	786	327	718	712	390	729
64	8	1,680	539	1,141	37	539	1,104	1,497	153	30	756	416	508	624	527	529
64	16	1,717	764	953	50	759	908	1,580	124	13	843	485	389	588	729	400
128	2	1,082	124	958	11	124	947	968	59	55	387	112	583	371	123	588
128	4	1,121	209	912	19	209	893	996	89	36	397	194	530	374	209	538
128	8	1,143	326	817	21	326	796	983	135	25	398	291	454	348	323	472
256	4	814	102	712	6	102	706	707	50	57	280	96	438	272	102	440
256	8	943	198	745	14	198	731	791	100	52	334	191	418	322	197	424
total		52,259	23,818	28,441	**1,436**	23,607	27,216	**43,186**	8,426	647	**24,493**	14,217	13,549	**18,824**	19,662	13,773

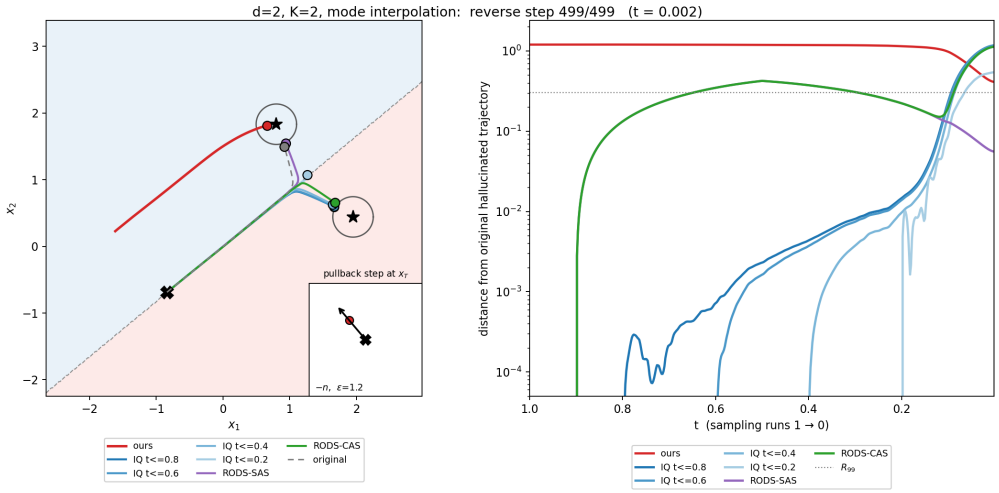
Per method: left = ground-truth total − int fix − inv fix. int fix = interpolation hallucinations corrected; inv fix = invalid hallucinations corrected (sent into a mode's R_{99} ball).

differs by at most 1.9 points between the samplers at every ϵ, and by at most 2.7 points within either kind. The step that corrects 90% of the hallucinations is $\epsilon = 0.550$ under DDIM and 0.555 under flow matching; for interpolations 0.098 and 0.106; for invalid samples 0.751 and 0.752. Where the two differ, DDIM is slightly ahead at small ϵ on invalid samples (6.1% against 3.6% at $\epsilon = 0.02$), and flow matching corrects the larger share at $\epsilon = 0.1$ in only 6 of the 28 cells. The step a seed needs is set by the kind of hallucination, not by the sampler. The baselines, which act on the trajectory rather than on the seed, do depend on the sampler: with the same parameter grids, RODS-SAS and RODS-CAS correct 53.1% and 64.0% of the flow-matching hallucinations against 77.3% and 83.4% under DDIM, and IQ corrects 3.2–17.4% against 20.4–27.2%, falling as its window shrinks.

Trajectory divergence from the original with time Our correction acts only at the beginning before the reverse process is done at $\tau = 1$. The sample it changes is reached only at $\tau = 0$. To see where along the run the corrected sample leaves the hallucinated one, let $D(\tau) := \|\boldsymbol{X}_\tau(\boldsymbol{z} - \epsilon\boldsymbol{n}) - \boldsymbol{X}_\tau(\boldsymbol{z})\|_2$ be the distance between the corrected trajectory and the original hallucinated one, so that $D(1) = \epsilon$. Figure S17 follows two seeds of the $d = 2$, $K = 2$ cell at $\epsilon = 1.2$, one from each class. In both, D stays at ϵ for most of the run: the sampler neither amplifies nor erases the seed step until the last fifth, and it is there that the seed step decides on which side of the separator the trajectory commits. What differs is what happens there. For the mode interpolation (Figure S17b) the step is amplified and D ends well above ϵ; for the invalid sample (Figure S17a) it is not as extreme and D ends below it. This is showing how in (Table S5) invalid samples need a step roughly seven times larger to be properly corrected. IQ and RODS leave the trajectory unchanged until their guidance switches on, at $\tau = w$ for IQ and at $\tau = 0.9$ for RODS, and must make the whole correction in the rest of the run. On the log axis IQ's perturbation stays orders of magnitude below ours throughout, which is consistent with IQ correcting only 3.2% of the flow-matching hallucinations at $w = 0.2$.



(a) An invalid sample: D is damped over the run, ending below ϵ.



(b) A mode interpolation: D is amplified, ending above ϵ.

Figure S17: **Where along the run the corrected trajectory leaves the hallucinated one.** Two seeds of the $d = 2$, $K = 2$ cell at $\epsilon = 1.2$, one of each class. Left: both trajectories in data space, with the pullback step at $\boldsymbol{x}_T$ inset. Right: $D(\tau)$ on a log axis, with R_{99} marked. Our step is applied at $\tau = 1$; IQ and RODS leave the trajectory untouched until their guidance switches on, at $\tau = w$ and $\tau = 0.9$ respectively.

H.2 Additional MNIST Results

We test whether pullback predicts outcomes on real data, whose distribution is only approximately a Gaussian mixture. We train an autoencoder whose latent space is shaped like the mixtures of, fit a Gaussian mixture to the latents, run a latent DDIM in that space, and predict the outcomes of its seeds as before.

Note: The full set of MNIST experiments is too extensive to include here and is available in our repository.

Latent space. MNIST has $C = 10$ classes and we use a latent dimension $d = 32$. A convolutional autoencoder with encoder Enc and decoder Dec is trained for 300 epochs. The encoder sees only images; the class labels c_p enter only through the loss, which imposes the geometry of on the latents $\boldsymbol{x}_p = \mathrm{Enc}(\mathrm{img}_p)$ of a minibatch $p = 1, \dots, B$:

$$\begin{aligned}\mathcal{L} = \frac{1}{B}\sum_{p=1}^{B} \|\mathrm{img}_p - \mathrm{Dec}(\boldsymbol{x}_p + \boldsymbol{\xi}_p)\|_2^2 + \lambda_{\mathrm{sep}}\, \mathbb{E}_{c_p \neq c_q}\Big[\max\Big(0, \tfrac{d_{\mathrm{sep}} - \|\boldsymbol{x}_p - \boldsymbol{x}_q\|_2}{\sigma_0}\Big)^2\Big] \\ + \lambda_{\mathrm{cov}}\, \mathbb{E}_c \Big\| \tfrac{\boldsymbol{\Sigma}_c}{\sigma_0^2} - \boldsymbol{I}_d \Big\|_F^2 + \lambda_{\mathrm{norm}}\, \mathbb{E}_c\Big[\max\Big(0, \tfrac{\|\boldsymbol{\mu}_c\|_2 - r_\mu}{\sigma_0}\Big)^2\Big],\end{aligned} \tag{H.1}$$

where the pixels are scaled to $[-1, 1]$, $\boldsymbol{\xi}_p \sim \mathcal{N}(\mathbf{0}, (0.15\,\sigma_0)^2 \boldsymbol{I}_d)$, and $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the minibatch mean and unbiased covariance of the latents of class c. The separation term averages over the cross-class pairs that violate the margin, the covariance term over the classes with at least $d + 2$ members in the minibatch, and the norm term over the classes present. The constants are those of the Gaussian mixtures, $\sigma_0 = 0.1$, $r_\mu = 2\sqrt{d/2}$ and $d_{\mathrm{sep}} = 3R_{99}$ with $R_{99} = \sigma_0\sqrt{\chi^2_d(0.99)}$; the weights $\lambda_{\mathrm{sep}} = 5$, $\lambda_{\mathrm{cov}} = 2$ and $\lambda_{\mathrm{norm}} = 1$ are ramped in linearly over the first 5 epochs. The three terms ask that latents of different classes be at least d_{sep} apart, that each class have covariance $\sigma_0^2 \boldsymbol{I}_d$, and that the class means lie within radius r_μ; the means themselves are not prescribed.

Fitted mixture and mode cores. The latent distribution is only approximately Gaussian, so we fit a mixture to the encoded training set without labels. k-means with $K = 10$ clusters (k-means++ initialization, 4 restarts) gives the centers $\widehat{\boldsymbol{\mu}}_k$, the cluster fractions give the weights $\widehat{w}_k$, and the mean over clusters of the per-dimension within-cluster variance gives $\widehat{\sigma}_0^2$. Each cluster has its own radius $\widehat{R}_{99,k} = \widehat{\sigma}_k\sqrt{\chi^2_d(0.99)}$, with $\widehat{\sigma}_k^2$ its per-dimension variance, and we use the largest, $\widehat{R}_{99} = \max_k \widehat{R}_{99,k}$ (at most 0.62 over the three runs, against $R_{99} = 0.73$ for the target geometry). Labels are used only afterwards, to name each cluster by its majority class. Every training latent lies inside its own cluster's ball, while up to 2.3% of the encoded test images lie outside every ball. We therefore enlarge the cores to $\widehat{\mathcal{C}}_k^{\bullet} = B(\widehat{\boldsymbol{\mu}}_k, \widehat{r})$ with $\widehat{r} = s_{\mathrm{core}}\widehat{R}_{99}$ and $s_{\mathrm{core}} = 1.25$, so that samples just outside a fitted ball are not counted as hallucinations. The outcome of a seed is k if its sample lies in $\widehat{\mathcal{C}}_k^{\bullet}$ and -1 otherwise.

Diffusion model. The learned field $\boldsymbol{F}_\theta$ is a DDIM whose linear schedule $\beta \in [10^{-4}, 0.02]$ is discretized into 100 noise levels (99 sampling steps, terminal $\bar{\alpha} = 0.36$), so the seeds $\boldsymbol{z} \sim \gamma$ are not the terminal marginal of the forward process. It uses the network of Fig. S11 ($n_h = 256$), trained on the encoded training images rather than on samples of the fitted mixture. We train it for a tenth of the budget of Sec. G.2, $3{,}000\,(1 + d/16)(1 + K/16) = 14{,}625$ steps, without retraining, so that it hallucinates often enough for detection to be measured: on 4.3% of the test seeds under $\mathcal{G}_\theta$, against 0.004% under the exact endpoint map of the fitted mixture.

Anchors, reference outcomes and evaluation. Anchors are sampled as in Sec. G.1, with $n = b$ points in each enlarged core and $m = b/2$ in the band of width $w = 2\widehat{\sigma}_0$ around it; the w-kNN spheres extend to $r_{\max}\widehat{r}$. The exact field of the data is unknown, so the anchors are pulled back by inverting the learned DDIM sampler: each reverse step starts from the noise prediction at the current point and refines it with $n_{\mathrm{fp}} = 2$ fixed-point iterations (3 network evaluations per step), so that sampling from an anchor seed returns the anchor up to the fixed-point residual. The w-kNN threshold is calibrated on $5{,}000$ seeds labelled by $\mathcal{G}_\theta$. The test set consists of $NK = 100{,}000$ seeds ($N = 10{,}000$ per mode) labelled by their outcome under $\mathcal{G}_\theta$, and the classifiers and metrics are those of Secs. G.1 and G.3. As a baseline, each test seed is also mapped with the exact endpoint map $\widehat{\mathcal{G}}_\star$ of the fitted mixture, integrated with Heun-type DDIM steps on the linear schedule discretized into 200 levels (terminal $\bar{\alpha} = 0.13$), and its outcome is compared with that under $\mathcal{G}_\theta$; this reaches 85.6% accuracy but a hallucination F1 of only 0.002, since $\widehat{\mathcal{G}}_\star$ almost never hallucinates. As an independent check, a CNN classifier trained on the ten classes reads the decoded samples and verifies that a sample in $\widehat{\mathcal{C}}_k^{\bullet}$ decodes to an image of the class of cluster k. Alg. 8 summarizes the procedure and Tab. S6 lists its constants.

Decoded samples. To show what the predicted seed regions look like, we draw $100{,}000$ fresh seeds, fit each classifier on the anchors at the largest budget, and predict the outcome of every

seed. For each predicted class we decode the samples $\mathrm{Dec}(\mathcal{G}_\theta(\boldsymbol{z}))$ of up to 10 seeds: for a cluster k, the first seeds in seed order predicted as k whose samples lie within $0.7\,\widehat{R}_{99}$ of $\widehat{\boldsymbol{\mu}}_k$, and for the hallucination class, the first seeds predicted as hallucinations whose samples lie beyond $\widehat{r}$ from every center. For each classifier we also report the precision and recall of its hallucination predictions on these seeds and the distance of the flagged samples to their nearest center in units of $\widehat{r}$. A classifier flagging fewer than 200 seeds would have its highest-margin seeds added until 200 are flagged; every classifier flagged at least 637.

H.3 How do we know a low rank boundary exists?

We have shown that the decision boundary can be characterized via pullback. In this section, we provide empirical evidence for this structure by analyzing the effective rank of samples hallucinated by the learned score function. Because the class boundary consists of points at the periphery of these hallucinations, a low-rank structure in the hallucinated samples implies a low-rank structure in the class boundary itself.

Algorithm 8 Evaluation protocol in the MNIST latent space, for one anchor budget b.

```
# images, labels : MNIST, C = K = 10;  d : latent dimension;  N : test seeds per mode
# s_core : core enlargement;  n, m : core and band anchors per mode
# n_fp : fixed-point iterations;  n_steps : training steps of F_theta
# n_cal : calibration seeds;  n_runs : runs

for run in range(n_runs):
    Enc, Dec = train_autoencoder(images, labels, d)      # loss L above
    x_lat    = Enc(images)                               # latent training set
    mu_hat, w_hat, sigma0_hat = kmeans_mixture(x_lat, K) # no labels
    r_hat    = s_core * max(R99_per_cluster(x_lat, mu_hat))
    F_theta  = train_field(x_lat, n_steps)               # DDIM, 100 levels

    z      = randn(N * K, d)                             # test seeds
    y_ref  = outcome(G_theta(z), mu_hat, r_hat)          # reference outcomes
    y_base = outcome(G_star_hat(z), mu_hat, r_hat)       # exact map, 200 levels
    scores[run, "exact"] = metrics(y_base, y_ref)

    z_cal = randn(n_cal, d)                              # w-kNN calibration
    y_cal = outcome(G_theta(z_cal), mu_hat, r_hat)
    Z, y  = pullback(mu_hat, r_hat, n, m, 2 * sigma0_hat, invert(G_theta, n_fp))
    for clf in classifiers:                              # w-kNN: its own anchors
        y_hat = clf.fit(Z, y).predict(z)
        scores[run, clf] = metrics(y_hat, y_ref)

return mean(scores, axis=0)
```

Note: The full set of MNIST experiments is too extensive to include here and is available in our repository.

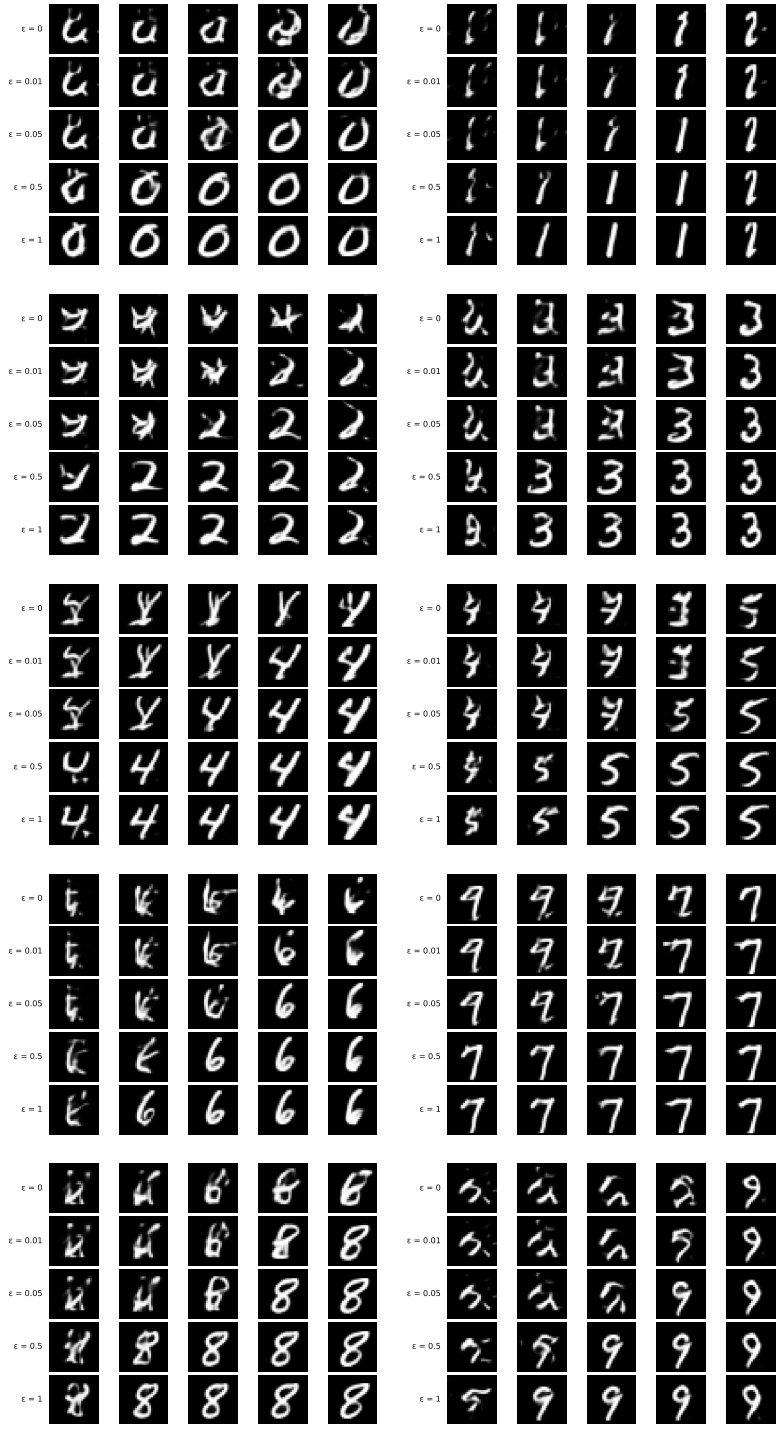


Figure S18: DDIM trajectories on MNIST for classes 0 through 9 (left to right, top to bottom).

Table S6: **Settings in the MNIST latent space.** Identifiers are those of Algorithm 8.

Setting	Identifier	Value
Dataset		MNIST ($C = 10$ classes)
Latent dimension	`d`	32
Autoencoder		convolutional, 300 epochs, AdamW, lr 8×10^{-4}, cosine decay
Autoencoder minibatch	B	$\max(512,\ 2K(d+2)) = 680$
Fitted mixture		k-means, $K = 10$; $\hat{\sigma}_0$ from within-cluster variance
Core enlargement	`s_core`	1.25
Learned sampler	$\mathcal{G}_\theta$	DDIM, 100 levels (99 steps, terminal $\bar{\alpha} = 0.36$)
Exact-field baseline	$\hat{\mathcal{G}}_\star$	Heun-type DDIM, 200 levels (terminal $\bar{\alpha} = 0.13$)
Network training	`n_steps`	$3{,}000\,(1 + d/16)(1 + K/16) = 14{,}625$, no retraining
Pullback	`n_fp`	inversion of $\mathcal{G}_\theta$, 2 fixed-point iterations per step
Anchors per mode	`n` $= b$	10,000, 50,000, 100,000
Band anchors per mode	`m`	$b/2$, band width $2\hat{\sigma}_0$
Test seeds per mode	`N`	10,000 (100,000 in total)
Calibration seeds	`n_cal`	5,000, labelled by $\mathcal{G}_\theta$
Runs	`n_runs`	3 (seeds 0, 100, 200)

I SYMBOL TABLE

Class data and Gaussian approximation

C	Number of classes or reference components; component i represents class $c = i$.
$\boldsymbol{x}_c^{(i)}$, n_c	The ith observation of class c and the number of observations in the class.
w_c, $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$	Class weight, empirical class mean, and empirical class covariance (denominator n_c), the latter assumed nonsingular (Sec. C). In Sec. H the same symbols denote the unbiased minibatch estimates used by the latent-space training loss.
$\boldsymbol{U}$, $\boldsymbol{\Lambda}$, λ_k	Eigendecomposition $\boldsymbol{\Sigma}_c = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$.
$\lambda_{\min}(\boldsymbol{\Sigma}_c)$, $\lambda_{\max}(\boldsymbol{\Sigma}_c)$	Smallest and largest eigenvalues of $\boldsymbol{\Sigma}_c$.
$\mathcal{E}_c(r)$	Class sublevel region $\{\boldsymbol{x} : (\boldsymbol{x} - \boldsymbol{\mu}_c)^\top \boldsymbol{\Sigma}_c^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_c) \le r^2\}$ of dimensionless Mahalanobis radius r (Sec. C); the ball $B(\boldsymbol{\mu}_c, \sigma_0 r)$ when $\boldsymbol{\Sigma}_c = \sigma_0^2 \boldsymbol{I}_d$, so that $\mathcal{C}_i^\bullet = \mathcal{E}_i(r_i/\sigma_0)$.
φ_η	Gaussian smoothing kernel $\mathcal{N}(\cdot; \boldsymbol{0}, \eta^2 \boldsymbol{I}_d)$ for $\eta > 0$.
$f_{c,0}$, $f_{c,\eta}$	Population class density and its Gaussian smoothing $f_{c,0} * \varphi_\eta$ for $\eta > 0$.
$\widehat{f}_{c,\eta}$	Empirical class density smoothed at scale $\eta > 0$.
$g_{c,\eta}$	Gaussian reference density with mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{S}_c(\eta)$; $g_{c,0} = \mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$.
$\boldsymbol{S}_c(\eta)$	Smoothed reference covariance $\boldsymbol{\Sigma}_c + \eta^2 \boldsymbol{I}_d$.
$\boldsymbol{R}_c(\boldsymbol{x})$	Covariance-scaled score residual $\boldsymbol{\Sigma}_c \nabla \log f_{c,0}(\boldsymbol{x}) + (\boldsymbol{x} - \boldsymbol{\mu}_c)$.
κ_c	Uniform bound on $\|\boldsymbol{R}_c\|_2$ in Prop. 1.
$\boldsymbol{m}(\boldsymbol{y})$	Conditional mean $\mathbb{E}[\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}]$ for $\boldsymbol{X} \sim f_{c,0}$ and $\boldsymbol{Y} = \boldsymbol{X} + \eta \boldsymbol{Z}$, used locally in the proof of Prop. 1.
$\alpha_{c,i}$, $\bar{\boldsymbol{x}}_c$	Within-class sample responsibility and the resulting weighted center in Sec. C.4.
$e_{c,n}$	Norm of the difference between the empirical and population class scores.
p_η, ω_b	Population mixture of class densities $\sum_b w_b f_{b,\eta}$ and the class responsibility $w_b f_{b,\eta}/p_\eta$ in it.

Generative transport

τ	Model time, decreasing from the source endpoint 1 to the data endpoint 0 during generation.
γ, $\boldsymbol{z}$	Standard Gaussian seed law $\mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$ and seed $\boldsymbol{z} \sim \gamma$.
$\boldsymbol{F}$	Deterministic sampling field in Eq. (1).
L, M	Bounds on the spatial derivative of $\boldsymbol{F}$ and on $\|\boldsymbol{F}(\mathbf{0}, \tau)\|_2$ in Eq. (A.1).
$\Phi^{\boldsymbol{F}}_{\tau_a \to \tau_b}$	Flow map of $\boldsymbol{F}$ from time τ_a to time τ_b.
$\mathcal{G}_{\boldsymbol{F}}$	Source-to-data endpoint map $\Phi^{\boldsymbol{F}}_{1\to 0}$.
$\boldsymbol{F}_\star$, $\mathcal{G}_\star$	Population sampling field and its endpoint map.
$\boldsymbol{F}_\theta$, $\mathcal{G}_\theta$	Learned sampling field and its endpoint map.
$\Phi^\star_{\sigma\to\sigma'}$	Population flow of Eq. (2) from scale σ to σ' in fixed-center coordinates; $\Phi^\star_{\sigma_0\to\sigma}(\mathcal{C}_i^\bullet)$ is the pull-back of the core to scale σ.

Diffusion coordinates and noise scales

t, T	Original diffusion time and its terminal forward-time value; denoising runs from T to 0, and $\tau = t/T$ for the VP model.
β_t, $\bar\alpha_t$	VP noise schedule and signal coefficient $\bar\alpha_t = \exp(-\int_0^t \beta_s\, ds)$.
$\boldsymbol{X}_t$, $\boldsymbol{Y}_t$	Original-coordinate state and fixed-center state $\boldsymbol{Y}_t = \boldsymbol{X}_t/\sqrt{\bar\alpha_t}$.
$\eta(t)$	Added smoothing standard deviation in fixed-center coordinates, $\eta(t)^2 = (1-\bar\alpha_t)/\bar\alpha_t$.
σ_0, $\sigma(t)$	Endpoint and current component standard deviations in fixed-center coordinates; $\sigma(t)^2 = \sigma_0^2 + \eta(t)^2$.
σ_a	Initial denoising scale $\sigma(T)$.
v_t	Component variance in original coordinates, $\bar\alpha_t\sigma_0^2 + 1 - \bar\alpha_t = \bar\alpha_t\sigma(t)^2$.
u, u_f	Increasing denoising parameter $\log(\sigma_a/\sigma)$ and its endpoint value $\log(\sigma_a/\sigma_0)$.
σ_u	Total component standard deviation $\sigma_a e^{-u}$.

Gaussian mixture and center geometry

p_{data}	Target Gaussian mixture of Sec. 3.
p_t^{VP}, p_σ	Mixture densities in original and fixed-center coordinates.
w_i, $\boldsymbol{\mu}_i$	Weight and center of component i.
ϕ_i^σ	Component density $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_i, \sigma^2 \boldsymbol{I}_d)$.
$\pi_i(\boldsymbol{y}, \sigma)$	Posterior responsibility $w_i \phi_i^\sigma(\boldsymbol{y}) / p_\sigma(\boldsymbol{y})$.
$\pi_i^{\text{VP}}(\boldsymbol{x}, t)$	Component responsibility in original VP coordinates; $\pi_i^{\text{VP}}(\sqrt{\bar{\alpha}_t}\boldsymbol{y}, t) = \pi_i(\boldsymbol{y}, \sigma(t))$.
$\boldsymbol{D}_\sigma(\boldsymbol{y})$	Posterior mean of the component center, $\sum_i \pi_i(\boldsymbol{y}, \sigma)\boldsymbol{\mu}_i$.
d_{ij}	Distance $\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\|_2$ between centers.
$\boldsymbol{n}_{ij}$, $\boldsymbol{m}_{ij}$	Unit direction $(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)/d_{ij}$ and midpoint $(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j)/2$.
d_i, Δ_i	Nearest and farthest center separations from component i.
$j^\star$	A nearest competitor of i, $j^\star \in \arg\min_{j \neq i} d_{ij}$.
$D_{\max}$	Maximum separation $\max_{i,j} d_{ij}$ over all centers.

One-mode dominance and reference contact

δ	Tolerance on the combined responsibility of competing modes.
$L_i(\delta)$	Logarithmic factor $\log((1-\delta)(1-w_i)/(\delta w_i))$ in the radius bound.
$r_i(\sigma;\delta)$	Certified radius of single-mode dominance in Eq. (4), obtained from the nearest-separation bound; a lower bound on $r_i^{\mathrm{ex}}(\sigma;\delta)$.
$\bar{r}_i(\sigma;\delta)$	Unclipped radius $d_i/2 - \sigma^2 L_i(\delta)/d_i$.
$\Psi_i(r,\sigma)$	Uniform upper bound on the competing responsibility odds over $B(\boldsymbol{\mu}_i, r)$ for $0 \le r \le d_i/2$.
$\mathcal{R}_i(\sigma;\delta)$	Set of radii in $[0, d_i/2]$ with $\Psi_i(r,\sigma) \le \delta/(1-\delta)$.
$r_i^{\mathrm{ex}}(\sigma;\delta)$	Supremum of radii satisfying exact dominance within the cap $d_i/2$, with the zero convention of Eq. (D.15).
$\sigma_i^{\dagger}(\delta)$	Birth scale $d_i/\sqrt{2L_i(\delta)}$ below which the certified radius is positive when $L_i(\delta) > 0$.
$r_i, \mathcal{C}_i^{\bullet}$	Endpoint radius $r_i(\sigma_0;\delta)$, possibly zero, and mode core $B(\boldsymbol{\mu}_i, r_i)$, nontrivial when $r_i > 0$.
$R(\sigma)$, $R_i(\sigma)$	Radius $(\sigma/\sigma_0)R_0$ of a sphere of endpoint radius R_0 under isolated-mode transport; $R_i(\sigma) = (\sigma/\sigma_0)r_i$ for the core of mode i.
r	Common endpoint radius $r_i = r_j$ of an equally weighted pair in Prop. 3.
$\mathcal{B}_k(\sigma)$	Transported reference ball $B(\boldsymbol{\mu}_k, (\sigma/\sigma_0)r)$ of a pair member $k \in \{i,j\}$; its boundary is the reference sphere.
t_{ij}, σ_{ij}^{c}	Original time and fixed-center scale of first contact between the two reference spheres.
H_{ij}, $H_{ij}(t)$	Equal-weight pair bisector in fixed-center and original coordinates; the equal-responsibility separator and the common tangent of the reference spheres at contact.
ε_{ij}	Relative radius shortfall $1 - 2r/d_{ij} = 2\sigma_0^2 L_i(\delta)/d_{ij}^2 \in (0,1)$ for the equally weighted, mutually nearest pair with positive common endpoint radius in Prop. 3.

Two-mode reduction

χ	Dimensionless distance margin in Eq. (E.12).
ρ_{ij}	Total responsibility outside the pair, $\sum_{k\notin\{i,j\}} \pi_k$.
$\widetilde{\pi}_i$, $\widetilde{\pi}_j$	Responsibilities conditional on belonging to the pair, $\widetilde{\pi}_a = \pi_a/(\pi_i+\pi_j)$.
$\boldsymbol{D}_\sigma^{ij}$	Conditional pair center $\widetilde{\pi}_i\boldsymbol{\mu}_i + \widetilde{\pi}_j\boldsymbol{\mu}_j$.
B_{ij}	Upper bound $((1-w_i-w_j)/\min\{w_i,w_j\})e^{-\chi}$ on the odds $\rho_{ij}/(1-\rho_{ij})$ of the excluded responsibility, so that $\rho_{ij} \le B_{ij}/(1+B_{ij})$.
Δ_{ij}	Maximum distance from an excluded center to either pair center; zero when $C=2$.
S_{ij}	Closed segment $[\boldsymbol{\mu}_i, \boldsymbol{\mu}_j]$.
$\boldsymbol{e}_u$	Center error $\boldsymbol{D}_{\sigma_u}(\boldsymbol{Y}_u) - \boldsymbol{D}^{ij}_{\sigma_u}(\boldsymbol{Y}_u)$.
s_u, $\boldsymbol{z}_\perp(u)$	Axial and transverse coordinates in $\boldsymbol{Y}_u = \boldsymbol{m}_{ij} + s_u\boldsymbol{n}_{ij} + \boldsymbol{z}_\perp(u)$.
$a(u)$	Distance $\mathrm{dist}(\boldsymbol{Y}_u, S_{ij})$ to the pair segment.

Pullback boundary geometry

$\mathcal{K}_i^{\boldsymbol{F}}$	Pulled-back core $\mathcal{G}_{\boldsymbol{F}}^{-1}(\mathcal{C}_i^\bullet)$.
$h_i^{\boldsymbol{F}}(\boldsymbol{z})$	Level-set function $\|\mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z})-\boldsymbol{\mu}_i\|_2^2 - r_i^2$; negative in the interior of $\mathcal{K}_i^{\boldsymbol{F}}$, zero on its boundary, positive outside. Its magnitude is not a distance.
$\Sigma_i^{\boldsymbol{F}}$	Seed-space boundary $\partial\mathcal{K}_i^{\boldsymbol{F}} = \mathcal{G}_{\boldsymbol{F}}^{-1}(\partial\mathcal{C}_i^\bullet) = \{\boldsymbol{z} : h_i^{\boldsymbol{F}}(\boldsymbol{z}) = 0\}$.
$\rho_i(\boldsymbol{x})$	Terminal radius $\|\boldsymbol{x}-\boldsymbol{\mu}_i\|_2$ of a generated sample.
$\boldsymbol{z}_a$, $\boldsymbol{x}_a$	Boundary anchor $\boldsymbol{z}_a \in \Sigma_i^{\boldsymbol{F}}$ and its output $\boldsymbol{x}_a = \mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z}_a) \in \partial\mathcal{C}_i^\bullet$.
J_a	Endpoint Jacobian $D\mathcal{G}_{\boldsymbol{F}}(\boldsymbol{z}_a)$.
$\boldsymbol{\nu}_a$	Outward terminal unit normal $(\boldsymbol{x}_a-\boldsymbol{\mu}_i)/r_i$.
P_a	Terminal tangent projection $\boldsymbol{I}_d - \boldsymbol{\nu}_a\boldsymbol{\nu}_a^\top$.
$\boldsymbol{n}_a$	Outward seed-space unit normal $J_a^\top\boldsymbol{\nu}_a/\|J_a^\top\boldsymbol{\nu}_a\|_2$, so that $\nabla h_i^{\boldsymbol{F}}(\boldsymbol{z}_a) = 2r_i\|J_a^\top\boldsymbol{\nu}_a\|_2\,\boldsymbol{n}_a$; $-\boldsymbol{n}_a$ is the inward normal of Prop. 4.
$\|J_a^\top\boldsymbol{\nu}_a\|_2$	Gain: first-order change of the terminal radius per unit seed step along $\boldsymbol{n}_a$.
ε, $\boldsymbol{v}$	Perturbation magnitude and seed direction; $\|\boldsymbol{v}\|_2 = 1$ when comparing intervention efficiency.
$\gamma_a(\boldsymbol{v})$	First-order inward radial gain of a unit seed direction, $\|J_a^\top\boldsymbol{\nu}_a\|_2\langle\boldsymbol{v}, -\boldsymbol{n}_a\rangle$ (Lemma 12).
$\mathrm{Align}_a(\boldsymbol{v})$	Alignment $\cos\angle(\boldsymbol{v}, -\boldsymbol{n}_a)$ with the locally optimal inward direction; equal to $\gamma_a(\boldsymbol{v})/\max_{\|\boldsymbol{u}\|_2=1}\gamma_a(\boldsymbol{u})$ for unit $\boldsymbol{v}$.

Experiments

K, d	Number of modes and ambient dimension of a Gaussian-mixture configuration; $K = C$ for the theoretical reference.
R_{99}	Radius $\sigma_0\sqrt{\chi^2_d(0.99)}$ of the ball around a mode center holding 99% of the mass of that mode.
$\mathcal{C}_k^\bullet$ (experiments)	Experimental mode core $B(\boldsymbol{\mu}_k, R_{99})$ of Sec. 6, with R_{99} in place of the certified radius $r_k(\sigma_0;\delta)$; a Gaussian-mixture sample outside every such ball is labelled a hallucination.
N	Number of test seeds per mode; the test set has NK seeds drawn from γ.
T (experiments)	Number of pullback grid points, i.e. $T-1$ integration steps; unrelated to the terminal time T above.
b, n, m, w	Anchor budget per mode, number of core anchors $n = b$, number of band anchors $m = b/2$, and width $w = 2\sigma_0$ of the data-space band $R_{99} < \|\boldsymbol{x} - \boldsymbol{\mu}_k\|_2 \le R_{99} + w$.
k	Number of neighbours in the kNN and weighted kNN classifiers.
η (tables)	Hallucination rate (%) of a sampler on a cell in Sec. G; unrelated to the smoothing scale η above.
$\widehat{\boldsymbol{\mu}}_k$, $\widehat{w}_k$, $\widehat{\sigma}_0$	Centers, weights, and per-dimension standard deviation of the mixture fitted to the latent space in Sec. H.
$\widehat{R}_{99,k}$, $\widehat{R}_{99}$	Fitted 99% radius of latent cluster k and their maximum.
$\widehat{r}$, s_{core}, $\widehat{\mathcal{C}}_k^\bullet$	Enlarged latent core radius $\widehat{r} = s_{\text{core}}\widehat{R}_{99}$ with $s_{\text{core}} = 1.25$, and the enlarged core $B(\widehat{\boldsymbol{\mu}}_k, \widehat{r})$ used to label latent outcomes.