LEARNING LONG-HORIZON FLOW MAPS IN HIGH DIMENSIONS: FROM A REGRESSION SURROGATE TO EFFICIENT STATISTICAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Establishing statistically efficient neural flow-map learning for one-step generation requires a valid training objective in the absence of direct flow-map supervision. To this end, we develop (i) a flow-map regression surrogate perspective and (ii) a dimension-efficient statistical learning theory that exploits intrinsic mixture structure of real data. Specifically, we introduce transport loss, which identifies the true flow map and controls its regression error, providing a common framework for MeanFlow, CM, and sCM objectives. We then investigate statistical learnability under this objective. While generic learning bounds suffer from the curse of dimensionality, we establish learning guarantees with polynomial dependence on intrinsic and latent dimensions under a low-rank Gaussian mixture assumption, using a layered residual network construction to control approximation and estimation errors. These insights motivate Gaussian Mixture Residual Flow (GMRFlow), which combines analytic Gaussian component flow maps with a learned neural residual correction. Experiments on CelebA80 and FFHQ64 show that GMRFlow achieves improved one-step FID over MeanFlow baselines.

1 INTRODUCTION

Diffusion and flow models generate samples by transporting noise toward data through learned SDE or ODE dynamics (Song et al., 2020; Lipman et al., 2022). Their reliance on repeated numerical integration motivates one-step generative models, including consistency models and MeanFlow (Song et al., 2023; Geng et al., 2025). These methods can be understood through neural flow-map learning: approximating flow maps that transport a state across a finite time interval in a single model evaluation (Boffi et al., 2024; 2025). We study the statistical learnability of such maps in high dimensions, beginning with the construction of a theoretically justified training objective.

The natural learning target is *flow-map regression*, which measures the squared discrepancy between a neural prediction and the true probability flow map. However, the true map is not directly available as a training label. Existing methods instead obtain training signals by enforcing properties of the underlying dynamics (Song et al., 2023; Geng et al., 2025; Lu & Song, 2024). To understand what these objectives achieve, we need to connect their training signals to the desired regression target. This leads to a foundational question: *Can we construct a surrogate objective that provably controls flow-map regression error without direct flow-map supervision?*

A valid surrogate does not by itself guarantee statistically efficient learning from finite data. Existing learning analyses for diffusion and consistency models highlight that learning arbitrary high-dimensional distributions typically suffers from the curse of dimensionality (Oko et al., 2023; Dou et al., 2024). The *manifold hypothesis* suggests that real data often concentrates near low-dimensional manifolds (Pope et al., 2021), motivating learning guarantees that exploit linear subspaces or low-rank Gaussian mixture structure (Chen et al., 2023; Yang et al., 2026). We therefore ask: *Can intrinsic data structure enable statistically efficient neural flow-map learning under such a surrogate objective?*

In this work, we develop a flow-map regression surrogate framework and use it to study how intrinsic data structure enables statistically efficient learning. These insights further motivate *Gaussian Mixture Residual Flow* (GMRFlow), a practical model for one-step generation.

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.

1.1 CONTRIBUTIONS

A flow-map regression surrogate perspective for training objectives. We construct a transport loss objective that provably identifies the true flow map and upper bounds its L_2 regression error, without further assumptions and access to the true map. We show that MeanFlow loss (Geng et al., 2025) and sCM loss (Lu & Song, 2024) are special cases of transport loss, while CM loss (Song et al., 2023) is the discretized version of transport loss.

Polynomial learnability under intrinsic Gaussian mixture structure. We first establish a generic learning bound to show that learning arbitrary high-dimensional flow maps typically suffers from the curse of dimensionality. We then model the latent distribution as a mixture of low-rank Gaussians and construct a layered residual network to establish a transport-loss learning bound with polynomial dependence on the intrinsic and latent dimensions.

Gaussian Mixture Residual Flow for one-step generation. Motivated by these insights, we propose GMRFlow, which combines analytic Gaussian component flow maps with a shared, cluster-conditioned neural residual correction to capture deviations from the fitted mixture prior. Experiments on CelebA80 and FFHQ64 show improved one-step FID over MeanFlow baselines.

2 RELATED WORK

Diffusion Models under Structured Distribution Assumptions. Existing work studies diffusion models under structured data assumptions to reduce the curse of dimensionality, with two notable directions: multi-modal modeling and low-dimensional modeling. Several works study mixture-of-Gaussians (MoG) data (Shah et al., 2023; Cui et al., 2023; Chen et al., 2024; Zhang et al., 2025). Pope et al. (2021); Brown et al. (2022); Kamkari et al. (2024) assume that the data distribution admits low-dimensional structure, motivated by evidence that image and text data often lie near low-dimensional manifolds. Related theoretical works study linear subspace assumptions (Chen et al., 2023; Yuan et al., 2023; Guo et al., 2024). Recently, Wang et al. (2025); Yang et al. (2026) combine the two approaches through low-dimensional Gaussian mixture assumptions.

Theoretical Analyses of Single-step Generative Models. Existing works have investigated general principles for designing flow-map learning models (Boffi et al., 2024; 2025). In contrast, theoretical analyses of learnability have focused primarily on the CM loss (Song et al., 2023). Dou et al. (2024) studied the statistical estimation error of the CM loss over an R-Lipschitz function class. Lyu et al. (2024); Chen et al. (2025) analyzed the sampling error induced by approximate self-consistency, while Li et al. (2024) related the number of consistency-training steps to the final sampling error.

3 PRELIMINARIES

Notation. We denote by D, d, d^* the ambient, latent, and intrinsic dimensions, respectively, and by p_{data} the data distribution. For a fixed $\epsilon > 0$, let $\mathcal{T}_\epsilon = \{(r,t) \mid \epsilon \le r \le t \le 1\}$ be the time domain. For a measurable map $F : \mathbb{R}^D \to \mathbb{R}^D$ and a distribution ρ, $F_\sharp \rho$ denotes the pushforward distribution of ρ under F. We write $B_R(x)$ for the Euclidean ball of radius R centered at x, and $\|\cdot\|_{\text{op}}$ for the spectral norm.

Flow Matching. Flow Matching (Lipman et al., 2022) is a framework for learning the velocity fields of probability flows that push a source distribution to a target distribution. Formally, let $x_0 \sim \rho_0 = p_{\text{data}}$ and $\varepsilon \sim \rho_1 = \mathcal{N}(0, I_D)$. Define the probability path $x_t = (1-t)x_0 + t\varepsilon$ with independent x_0 and ε. We write $\rho_t = \text{Law}(x_t)$ as the marginal distribution of x_t. The probability flow is defined by the PF-ODE

$$\frac{\mathrm{d}}{\mathrm{d}t} z_t = v(z_t, t), \quad z_0 \sim \rho_0.$$

Here $v(z,t)$ is defined as the conditional expectation of the time derivative of x_t

$$v(z,t) = \mathbb{E}\left[\dot{x}_t \mid x_t = z\right] = \mathbb{E}\left[\varepsilon - x_0 \mid x_t = z\right],$$

which averages the velocity of all sample interpolant paths that cross the point z at time t. In this way, z_t follows the same marginal distribution as x_t, i.e. $\mathrm{Law}(z_t) = \rho_t$. A neural network v_θ is trained to fit $v(z,t)$ by minimizing the flow matching loss

$$\mathbb{E}\left[\|v_\theta(x_t,t) - v(x_t,t)\|^2\right] = \mathbb{E}\left[\|v_\theta(x_t,t) - (\varepsilon - x_0)\|^2\right] + C,$$

where C is a constant independent of θ. Hence one can sample from p_{data} by sampling $z_1 \sim \mathcal{N}(0, I_D)$ and then solving the PF-ODE $z_0 = z_1 - \int_0^1 v_\theta(z_s, s)\mathrm{d}s$ using numerical solvers over discrete time steps $0 = t_0 < t_1 < \cdots < t_n = 1$. For example, Euler's method gives $z_{t_{i-1}} = z_{t_i} - (t_i - t_{i-1})v_\theta(z_{t_i}, t_i)$. In practice, time-truncation $t \in [\epsilon, 1]$ is often used to ensure numerical stability.

4 Surrogate Loss for Flow Map Regression

This section first introduces the flow map regression loss, which is the most natural objective for flow map learning. Then we propose transport loss as a surrogate for flow map regression and investigate its properties and connections to existing successful one-step generative losses.

4.1 Flow Map Regression Loss and Transport Loss

The probability flow map $X(z,r,t)$ maps a state at time t to the corresponding state at time r along the PF-ODE trajectory, i.e.

$$X(z_t, r, t) = z_t - \int_r^t v(z_s, s)\mathrm{d}s = z_r. \tag{1}$$

The flow map $X(z,r,t)$ is the learning target of many one-step generative models. If it is well learned by a neural network $X_\theta(z,r,t)$, one can sample from p_{data} by first sampling $z_1 \sim \mathcal{N}(0, I_D)$ and then applying $X_\theta(z_1, \epsilon, 1)$ once. The most natural learning objective is to minimize the L_2 **regression loss**

$$\mathcal{L}_{L^2}(\theta) = \mathbb{E}_{(r,t)\sim\nu, z_t\sim\rho_t}\left[\|X_\theta(z_t,r,t) - X(z_t,r,t)\|^2\right],$$

for some time distribution ν fully supported on $\mathcal{T}_\epsilon$. However, this loss is intractable since it requires the knowledge of the true flow map $X(z,r,t)$, which is unknown. To bridge this gap, we propose transport loss, which depends only on the velocity field $v(z,t)$. To this end, define transport derivative operator $\mathcal{D} = \partial_t + v(z,t)\cdot\nabla_z$, which is exactly the time derivative along the PF-ODE trajectory z_t.

$$\frac{\mathrm{d}}{\mathrm{d}t} f(z_t, t) = v(z_t,t)\cdot\nabla_z f(z_t,t) + \partial_t f(z_t,t) = \mathcal{D}f(z_t,t).$$

The key observation is that, for any fixed $r \in [\epsilon, 1]$, the exact flow map $X(z_t,r,t)$ is invariant along the PF-ODE trajectory z_t, i.e. $X(z_t,r,t) \equiv z_r$. Hence the true flow map satisfies the transport equation $\mathcal{D}X(z,r,t) = 0$. Through this property, we use the deviation of the learned flow map X_θ from the transport equation as the training signal. Define **transport loss**

$$\mathcal{L}_{\mathrm{tr}}(\theta) = \mathbb{E}_{(r,t)\sim\nu, z_t\sim\rho_t}\left[\|\mathcal{D}X_\theta(z_t,r,t)\|^2\right].$$

Theorem 4.1 (Transport Loss Identifies the true Flow Map)**.** *If $X_\theta(z,r,t)$ satisfies the boundary condition $X_\theta(z,t,t) = z$, then $\mathcal{L}_{\mathrm{tr}}(\theta) = 0$ if and only if $X_\theta(z,r,t) = X(z,r,t)$ throughout $\mathbb{R}^D \times \mathcal{T}_\epsilon$.*

This theorem further shows that transport loss $\mathcal{L}_{\mathrm{tr}}(\theta) = 0$ iff flow map regression loss $\mathcal{L}_{L^2}(\theta) = 0$, because the latter obviously identifies the true flow map. Another perspective is that $\mathcal{D}X_\theta$ is also the time derivative of the difference $X_\theta - X$. Hence along the PF-ODE trajectory we have

$$X_\theta(z_t,r,t) - X(z_t,r,t) = \int_r^t \mathcal{D}X_\theta(z_s,r,s)\,\mathrm{d}s$$

$$\xrightarrow{\text{Cauchy–Schwarz}} \|X_\theta(z_t,r,t) - X(z_t,r,t)\|^2 \le (t-r)\int_r^t \|\mathcal{D}X_\theta(z_s,r,s)\|^2\,\mathrm{d}s.$$

In this sense, transport loss $\mathcal{L}_{\mathrm{tr}}(\theta)$ is an upper bound on flow map regression loss $\mathcal{L}_{L^2}(\theta)$.

Theorem 4.2 (Transport Loss Upper Bounds Flow Map Regression Loss)**.** *Take $t \sim \mathrm{Unif}[\epsilon,1], r \sim \mathrm{Unif}[\epsilon,t], z_t \sim \rho_t$, for any continuously differentiable X_θ we have $\mathcal{L}_{L^2}(\theta) \le \frac{1}{4}\mathcal{L}_{\mathrm{tr}}(\theta)$.*

Corollary 4.3 (Transport Loss bounds Flow Map Sampling Error)**.** *Take $r = \epsilon, t \sim \mathrm{Unif}[\epsilon,1], z_t \sim \rho_t$, for any continuously differentiable X_θ we have $W_2^2(\rho_\epsilon, X_\theta(\cdot,\epsilon,1)_\sharp\rho_1) \le \mathcal{L}_{\mathrm{tr}}(\theta)$.*

With these results, we conclude that transport loss $\mathcal{L}_{\mathrm{tr}}(\theta)$ identifies the true flow map and upper bounds the flow map regression loss $\mathcal{L}_{L^2}(\theta)$, hence is a valid surrogate for $\mathcal{L}_{L^2}(\theta)$.

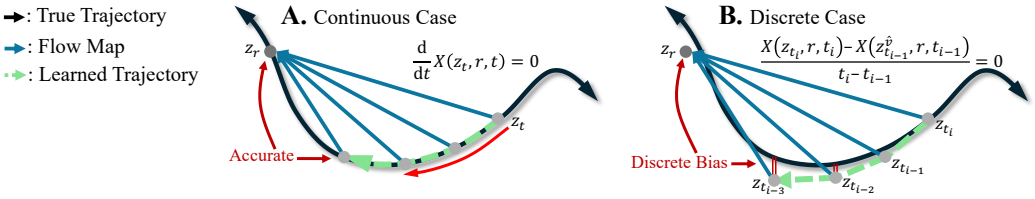


Figure 1: **Continuous and discrete transport constraints.** The continuous constraint preserves the flow-map value along a true trajectory, while enforcing consistency along discretized steps can introduce bias. A numerical experiment is conducted in Section 8.1 and illustrated in Figure 3.

4.2 Connection to existing losses

We relate transport loss to existing one-step generative losses. Figure 1 illustrates the distinction between enforcing the transport equation continuously and through finite differences.

MeanFlow Loss MeanFlow loss (Geng et al., 2025) models the average velocity of probability flows, which is defined as $u(z_t, r, t) = \frac{1}{t-r}\int_r^t v(z_s, s)\mathrm{d}s$. Through MeanFlow Identity $u(z_t, r, t) = v(z_t, t) - (t-r)\frac{\mathrm{d}}{\mathrm{d}t}u(z_t, r, t)$, MeanFlow loss is defined as

$$\mathcal{L}_{\mathrm{MF}}(\theta) = \mathbb{E}\left[\|u_\theta(z_t, r, t) - u_{\mathrm{tgt}}\|^2\right], \quad u_{\mathrm{tgt}} = v(z_t, t) - (t-r)(v(z_t, t)\partial_z u_\theta + \partial_t u_\theta).$$

With parametrization $X(z_t, r, t) = z_t - (t-r)u(z_t, r, t)$, MeanFlow loss is exactly equivalent to transport loss $\mathcal{L}_{\mathrm{tr}}(\theta)$.

CM Loss Consistency Model (CM) Loss (Song et al., 2023) models the consistency function $f(z_t, t) = X(z_t, \epsilon, t)$. It uses the property that $f(z_t, t)$ is invariant along the PF-ODE trajectory, i.e. $f(z_t, t) \equiv z_\epsilon$. With discretization $t_0 = \epsilon < t_1 < \cdots < t_n = 1$, and a given velocity teacher $\hat{v}$, CM loss is defined as

$$\mathcal{L}_{\mathrm{CM}}(\theta) = \mathbb{E}\left[\|f_\theta(z_{t_i}, t_i) - f_\theta(z^{\hat{v}}_{t_{i-1}}, t_{i-1})\|^2\right], \quad z^{\hat{v}}_{t_{i-1}} = z_{t_i} - (t_i - t_{i-1})\hat{v}(z_{t_i}, t_i).$$

CM loss is the discrete version of transport loss, because when divided by $(t_i - t_{i-1})^2$, CM loss becomes the MSE over the discrete transport derivative and converges to transport loss in the time limit. However CM loss introduces explicit discretization bias.

Theorem 4.4. *Take the probability flow as a linear ODE $\frac{\mathrm{d}x}{\mathrm{d}t} = Ax_t$, with $A^2 \neq 0$. Use uniform time steps. Assume the velocity teacher $\hat{v}$ is exact. Suppose the consistency function f_θ attains zero CM loss and satisfies $f_\theta(z, \epsilon) = z$, we have $\mathbb{E}\left[\|f_\theta(z_{t_i}, t_i) - X(z_{t_i}, \epsilon, t_i)\|^2\right] = \Theta(\Delta^2)$.*

While Lu & Song (2024) establish **sCM loss**, which is the continuous version of CM loss, we can show that both its objective and gradient update are equivalent to MeanFlow loss with stop gradient, hence recovering transport loss.

5 The Curse of Dimensionality for Generic Flow Maps

The regression surrogate objective identifies the true flow map, but does not by itself ensure efficient learnability. Indeed, learning a generic high-dimensional function typically suffers from the curse of dimensionality. We show below that generic flow maps exhibit the same phenomenon. Let $\Omega_R = B_R(0) \times \mathcal{T}_\epsilon$, endowed with the Euclidean norm.

Assumption 5.1 (Bounded support and velocity regularity)**.** $\forall t \in [\epsilon, 1]$, the marginal distribution ρ_t is supported on $B_R(0)$ for some $R > 0$. The velocity field $v(z, t)$ is B_v-bounded and L_v-Lipschitz over Ω_R.

Write the full Jacobian of a flow map $X : \Omega_R \to \mathbb{R}^D$ as $DX = [\nabla_z X \quad \partial_r X \quad \partial_t X] \in \mathbb{R}^{D\times(D+2)}$.

Definition 5.2 ($C^{1,1}$ Flow Map Class)**.** For $M_X, L_X > 0$, let $\mathcal{X}_{\mathrm{Lip}}(M_X, L_X)$ consist of continuously differentiable maps $X : \Omega_R \to \mathbb{R}^D$ satisfying $X(z, t, t) = z$, $\|DX(w)\|_{\mathrm{op}} \leq M_X$, and $\|DX(w) - DX(w')\|_{\mathrm{op}} \leq L_X\|w - w'\|_2$ for all $w, w' \in \Omega_R$.

Assume the true flow map $X^* \in \mathcal{X}_{\text{Lip}}(M_X, L_X)$. Let $\hat{X}_n$ minimize the empirical transport loss over this class, using n i.i.d. samples with $t \sim \text{Unif}[\epsilon, 1]$, $r \sim \text{Unif}[\epsilon, t]$, and $z_t \sim \rho_t$.

Theorem 5.3 (Learning bound for generic flow maps)**.** *Under Assumption 5.1, let $M_g = M_X(1+B_v)$, $L_g = (1+B_v)L_X + M_X L_v$, and $\Delta_\Omega = \text{diam}(\Omega_R) \le \sqrt{4R^2+2}$. There is a universal constant $C > 0$, such that for $\delta \in (0,1)$, with probability at least $1-\delta$ over the training samples, we have*

$$\mathbb{E}\left[\|\hat{X}_n - X^*\|_2^2\right] \le \frac{M_g L_g \Delta_\Omega}{2}\left(C\sqrt{D+2}n^{-1/(D+2)} + \sqrt{\frac{\log(1/\delta)}{2n}}\right)$$

The rate $n^{-1/(D+2)}$ in Theorem 5.3 deteriorates as the ambient dimension D increases. This motivates exploiting intrinsic data structure to obtain sharper learning guarantees. The proof is deferred to Appendix B.

6 Flow Map Learning with Low-Dimensional Structure

We study how intrinsic data structure can alleviate the generic learning difficulty in Section 5. A linear subspace assumption first reduces the unknown part of the flow map to latent coordinates. We then impose a low-rank Gaussian mixture model in that latent space to establish dimension-efficient approximation and estimation guarantees.

6.1 Low Dimensional Linear Subspace and Flow Map Decomposition

Chen et al. (2023) assume that p_{data} lies on a low-dimensional linear subspace of $\mathbb{R}^D$. We extend this assumption in the following way.

Assumption 6.1 (Manifold embedded in a linear subspace)**.** The data distribution p_{data} is supported on a d^*-dimensional manifold $\mathcal{M}$, which is embedded in a d-dimensional linear subspace V of $\mathbb{R}^D$.

Let $A \in \mathbb{R}^{D\times d}$ have orthonormal columns spanning V, so that $A^\top A = I_d$. The latent coordinates of $z \in \mathbb{R}^D$ are given by $A^\top z$, denoted as ξ. The manifold $\mathcal{M}$ may be nonlinear, and its intrinsic dimension d^* can be smaller than the latent dimension d of the containing linear subspace. We focus on the regime $d^* \le d \ll D$. While real data may be supported on an affine subspace, this case can be handled by centering the data.

Under Assumption 6.1, the flow map $X(z,r,t)$ defined in (1) decomposes into an explicit ambient scaling term and a residual that depends on the latent coordinates.

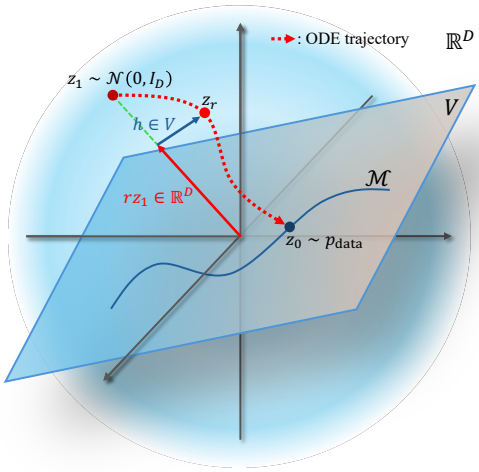


Figure 2: Flow map decomposition for a data manifold $\mathcal{M}$ in a linear subspace V: $z_r = rz_1 + h$, where $h \in V$.

Proposition 6.2 (Flow map decomposition)**.** *Under Assumption 6.1, there exists a map $h : \mathbb{R}^d \times \mathcal{T}_\epsilon \to V$ that only depends on the d-dimensional latent coordinates and satisfies the boundary condition $h(\xi, t, t) = 0$, such that*

$$X(z,r,t) = \frac{r}{t}z + h(A^\top z, r, t), \qquad (r,t) \in \mathcal{T}_\epsilon. \tag{2}$$

The decomposition, illustrated in Figure 2, reduces the effective dimension of flow map learning from D to d. Empirical evidence for approximate subspace concentration of real data and flow-map decomposition is given in Section 8.1.

6.2 Learning under Latent Gaussian Mixtures

Data modeling. The subspace decomposition alone does not ensure efficient learning of an arbitrary latent flow map. We therefore further model the latent distribution as a mixture of low-rank Gaussians. This additional structure retains sufficient flexibility to model real data (Yang et al., 2026), and allows us to establish a learning bound that depends polynomially on component rank and latent dimension.

Assumption 6.3 (Mixture of low-rank Gaussians in latent space)**.** The data distribution satisfies $p_{\text{data}} = A_\sharp q$, where

$$q = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k), \quad \Sigma_k = U_k \Lambda_k U_k^\top, \qquad \pi_k > 0, \quad \sum_{k=1}^{K} \pi_k = 1.$$

Here A is the embedding matrix defined in Section 6.1. For each k, $\mu_k \in \mathbb{R}^d$, $U_k \in \mathbb{R}^{d \times d^*}$ satisfies $U_k^\top U_k = I_{d^*}$, and Λ_k is a positive diagonal matrix.

Define the latent flow map $\Phi(\xi, r, t) = A^\top X(z, r, t)$, which satisfies $h = A(\Phi(\xi, r, t) - \frac{r}{t}\xi)$. Under Assumption 6.3, Φ is the flow map of the latent Gaussian mixture distribution. We construct a layered residual network Φ_θ to approximate Φ and analyze its approximation and estimation errors.

Construction. Let $C_{k,t} = t^2 I_d + (1-t)^2 \Sigma_k$ be the interpolant covariance of the k-th Gaussian component at time t. The k-th component velocity is

$$v_k(\xi, t) = \frac{\xi}{t} + \beta_k(\xi, t), \quad \beta_k(\xi, t) = -\frac{1-t}{t} \Sigma_k C_{k,t}^{-1} \xi - t C_{k,t}^{-1} \mu_k. \tag{3}$$

Note that β_k is a (d^*+1)-dimensional affine transformation for ξ in $\operatorname{span}\{U_k, \mu_k\}$. Let $\gamma_k(\xi, t)$ denote the posterior probability of the k-th component at point ξ and time t, and let $v(\xi, t)$ denote the overall velocity field. Then

$$\gamma_k(\xi, t) = \frac{\pi_k \mathcal{N}(\xi; (1-t)\mu_k, C_{k,t})}{\sum_{i=1}^{K} \pi_i \mathcal{N}(\xi; (1-t)\mu_i, C_{i,t})}, \quad v(\xi, t) = \sum_{k=1}^{K} \gamma_k v_k = \frac{\xi}{t} + \sum_{k=1}^{K} \gamma_k \beta_k.$$

In this way, the velocity field $v(\xi, t)$ admits a residual structure with a mixture of K low-rank affine transformations. We next construct a learnable Euler-type approximation to the corresponding flow map. Partition $[r, t]$ uniformly with step size $\Delta = (t-r)/m$, and let $\tau_i = r + i\Delta$. The network Φ_θ is parameterized by $\theta = \{\theta_i\}_{i=1}^m$ and consists of m residual blocks, where $\theta_i = \{\pi_{k,i}, \mu_{k,i}, U_{k,i}, \Lambda_{k,i}\}_{k=1}^K$ parameterizes the velocity field v_{θ_i} at the i-th block. Starting from $\Psi_m = \xi$, the network recursively applies Euler updates backward in time:

$$\Psi_{i-1} = \Psi_i - \Delta\, v_{\theta_i}(\Psi_i, \tau_i), \qquad \Phi_\theta(\xi, r, t) = \Psi_0.$$

Thus, Φ_θ provides a learnable Euler-type approximation to the flow map $\Phi(\xi, r, t)$ and automatically satisfies the boundary condition $\Phi_\theta(\xi, t, t) = \xi$.

Analysis. We first make the following regularity assumption on the parameter class Θ_m of the network Φ_θ.

Assumption 6.4 (Regularity of the parameter class Θ_m)**.** For any $\theta \in \Theta_m$, every layer $i \in [m]$ satisfies: (i) $\pi_{k,i} \geq 0$ and $\sum_{k=1}^K \pi_{k,i} = 1$; (ii) $U_{k,i}^\top U_{k,i} = I_{d^*}$; (iii) $\Lambda_{k,i} = \operatorname{diag}(\lambda_{k,i,1}, \ldots, \lambda_{k,i,d^*})$ with $0 \leq \lambda_{k,i,j} \leq R_\lambda$; (iv) $\|\mu_{k,i}\|_2 \leq R_\mu$; and (v) $D_\xi v_{\theta_i}(\xi, t) \succeq -L I_d$ for all $(\xi, t) \in \mathbb{R}^d \times [\epsilon, 1]$, where $R_\lambda, R_\mu, L > 0$ are constants independent of D, d, d^*, K, m.

Let $\theta^* = \{\pi_k^*, \mu_k^*, U_k^*, \Lambda_k^*\}_{k=1}^K$ denote the Gaussian-mixture parameters of the underlying data distribution q, and assume that $(\theta^*, \ldots, \theta^*) = \theta_{\text{or}} \in \Theta_m$. Denote the population transport loss for Φ_θ as $\mathcal{L}_{\text{tr}}(\theta) = \mathbb{E}\left[\|\mathcal{D}\Phi_\theta(\xi, r, t)\|^2\right]$, where $\mathcal{D}\Phi_\theta = \partial_t \Phi_\theta + D_\xi \Phi_\theta v(\xi, t)$. We consider n i.i.d. triples (ξ_i, r_i, t_i) with $(r_i, t_i) \sim \nu$ and $\xi_i \mid (r_i, t_i) \sim q_{t_i}$, where q_t is the noised latent marginal, and define the empirical transport loss as $\hat{\mathcal{L}}_{\text{tr},n}(\theta) = \frac{1}{n}\sum_{i=1}^n \|\mathcal{D}\Phi_\theta(\xi_i, r_i, t_i)\|^2$. Denote $A_m = \min_{\theta \in \Theta_m} \mathcal{L}_{\text{tr}}(\theta)$ as the minimum population loss, and $\hat{\theta} \in \arg\min_{\theta \in \Theta_m} \hat{\mathcal{L}}_{\text{tr},n}(\theta)$ as the empirical risk minimizer. We have the following approximation and overall learning bound for the soft conditioning architecture.

Theorem 6.5 (Intrinsic approximation bound)**.** *Under Assumption 6.4, there exists a constant C_{app} depending only on $\epsilon, R_\lambda, R_\mu, L$, such that the minimum population loss $A_m \leq C_{\text{app}}(1+d^*)^4/m^2$.*

Theorem 6.6 (Polynomial learning bound)**.** *Under Assumption 6.4, let $n \geq 2$, $0 < \delta < 1/2$ and let C be a constant depending only on $\epsilon, R_\lambda, R_\mu, L$. Set $P_m = mK(d+1)(d^*+1)$ and $S_n = 1 + d^* + \log(8n/\delta)$. With probability at least $1-\delta$, the empirical risk minimizer $\hat{\theta}$ satisfies*

$$\mathcal{L}_{\text{tr}}(\hat{\theta}) \leq \underbrace{C\frac{(1+d^*)^4}{m^2}}_{\text{approximation error}} + \underbrace{C\frac{S_n^3}{n}\left(P_m \log(CnP_mKS_n) + \log(8/\delta)\right)}_{\text{estimation error}}.$$

Proof sketch. Cancelling the common drift ξ/s confines each component correction to at most $1 + d^*$ dimensions. Stability of the oracle Euler network then gives the m^{-2} approximation. For estimation, low-rank moment bounds and a covering-number argument yield the term governed by P_m/n. Applying these bounds to the empirical risk minimizer establishes the learning guarantee. The detailed proof road map and technical lemmas are given in Appendix C. □

Remark 6.7. Theorem 6.6 implies, up to logarithmic factors, that

$$\mathcal{L}_{\mathrm{tr}}(\hat{\theta}) = \widetilde{O}\left(\frac{(d^*)^4}{m^2} + \frac{(d^*)^4 dmK}{n}\right).$$

The overall learning bound depends polynomially on the intrinsic dimension d^*, with only linear dependence on the latent dimension d. For a fixed target accuracy, choosing $m \asymp (1 + d^*)^2$ yields a sample complexity of $\widetilde{O}(K(d+1)(1+d^*)^6)$, in contrast to the $\exp(O(D \log D))$ complexity of the generic case established in Theorem 5.3. This separation shows how latent Gaussian mixture structure escapes the curse of dimensionality and enables efficient learning of long-horizon flow maps.

7 Gaussian Mixture Residual Flow

The structured learning analysis in Section 6 motivates a Gaussian-mixture prior for flow-map learning. We therefore propose Gaussian Mixture Residual Flow (GMRFlow), which combines analytic Gaussian component flow maps with a shared neural residual correction.

Algorithm Design. We consider the Gaussian mixture distribution q in $\mathbb{R}^D$, which has a hierarchical representation: first sample $k \sim \mathrm{Cat}(\pi_1, \ldots, \pi_K)$, then sample $z_0 \mid k \sim \mathcal{N}(\mu_k, \Sigma_k)$. Conditional on k, the component flow map $X_k^G(z_t, r, t) = z_t - \int_r^t v_k(z_s, s)\mathrm{d}s$ admits the following closed-form expression

$$X_k^{\mathrm{G}}(z, r, t) = (1 - r)\mu_k + C_{k,r}^{1/2} C_{k,t}^{-1/2}\big(z - (1 - t)\mu_k\big), \tag{4}$$

With fixed Gaussian estimates, we introduce a neural residual correction to make up the difference between the true flow map and the Gaussian component flow map

$$X_\theta(z, r, t; k) = X_k^{\mathrm{G}}(z, r, t) + \left(1 - \frac{r}{t}\right) c_{\mathrm{out}}(t) F_\theta(z, r, t; k), \tag{5}$$

Here F_θ is a shared network conditioned on k, and $c_{\mathrm{out}}(t) = t\sigma_x / \sqrt{(1-t)^2\sigma_x^2 + t^2}$ is an EDM-style output scale with fixed $\sigma_x > 0$. The factor $1 - r/t$ enforces $X_\theta(z, t, t; k) = z$; setting $F_\theta = 0$ recovers the Gaussian prior.

Training and sampling. During training, we first cluster the data into K groups and estimate $(\widehat{\mu}_k, \widehat{\Sigma}_k)$ and the empirical mixture weights $\widehat{\pi}_k$. The estimates are substituted into (4), and both the Gaussian parameters and cluster assignments remain fixed. For each labeled image (z_0, k), we sample (r, t) and independent $z_1 \sim \mathcal{N}(0, I_D)$, form $z_t = (1 - t)z_0 + tz_1$, and train F_θ using the cluster-conditioned transport loss $\mathcal{L}_{\mathrm{tr}}$, in its MeanFlow formulation. At sampling time, we independently draw $k \sim \mathrm{Cat}(\widehat{\pi}_1, \ldots, \widehat{\pi}_K)$ and $z_1 \sim \mathcal{N}(0, I_D)$, then generate $\widehat{z}_\epsilon = X_\theta(z_1, \epsilon, 1; k)$ in one flow-map evaluation.

8 Experiments

We first compare continuous and discrete objectives through MeanFlow and CM. We then validate the predicted subspace structure and flow-map decomposition on CelebA80, and finally evaluate GMRFlow on CelebA80 and FFHQ64.

8.1 Numerical Validation

Continuous and discrete objectives. Figure 3 compares MeanFlow and CM on a two-dimensional nonlinear dynamical system. MeanFlow samples more closely follow the target geometry, while the

CM samples and mapped grid exhibit visible distortions, illustrating that continuous objectives respect the transport equation while discrete objectives have explicit bias, as predicted by Theorem 4.4.

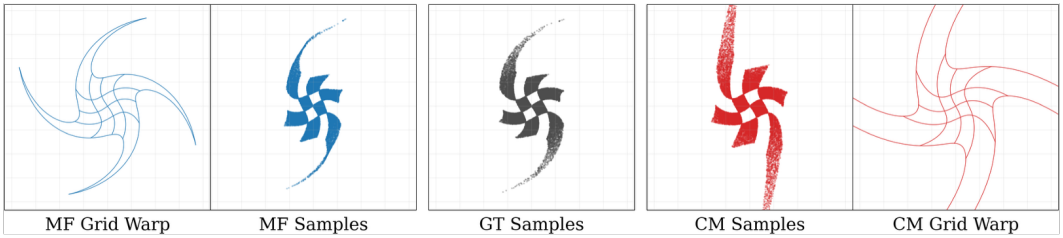


Figure 3: **Continuous versus discrete objectives.** MeanFlow samples (left) and CM samples (right), with the target samples in the center. The outer panels show the corresponding grid deformations.

Subspace assumption and flow-map decomposition. We examine the subspace assumption and flow map decomposition on 80×80 grayscale CelebA images ($D = 6400$) by PCA. Figure 4(a) shows that the leading $d = 256$ components, only 4% of the ambient dimension, retain 94.90% of the validation energy, supporting approximate concentration around a low-dimensional linear subspace.

Using the same PCA basis, we compare X and $h = X - (r/t)z_t$ from empirical and neural velocity fields on paired noisy images, integrating from $t = 0.5$ to $r = 0.25$. The neural field uses the released pretrained model of Kadkhodaie et al. (2024) without further training. The $d = 256$ subspace retains 96.38% and 98.55% of the empirical and neural residual h energy, respectively, versus 34.76% and 38.15% for X (Figure 4(b)). These observations support the residual concentration predicted by Proposition 6.2. Protocols and additional results are given in Appendix E.

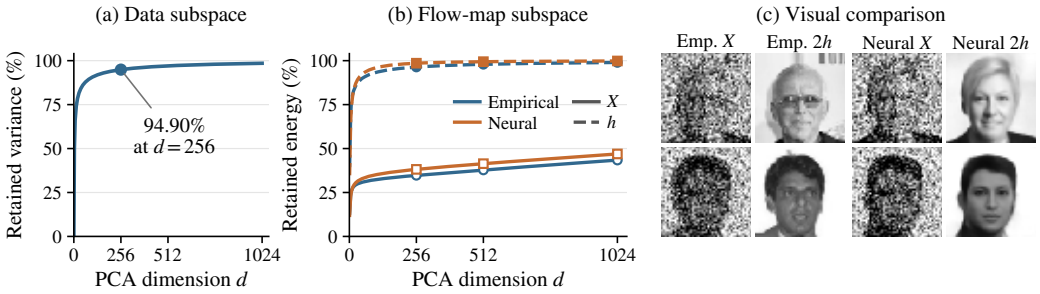


Figure 4: Subspace structure of CelebA data and flow maps. (a) Retained energy of validation images under data PCA. (b) Retained energy of X and h in the same subspaces after mean correction. (c) Two fixed queries (rows), with residuals displayed as $2h$ on the same grayscale scale as X.

8.2 Image Generation

Experimental setup. We evaluate GMRFlow in pixel space on CelebA80 (100K grayscale 80×80 images) using an 8.23M U-Net and FFHQ64 (70K RGB 64×64 images) using a 21.85M U-Net. The main comparisons share the same training recipe and the same MeanFlow objective. We report Fréchet Inception Distance (FID) for generation with one function evaluation (1-NFE) by default. Experimental details are given in Appendix F.

Performance comparison. We compare GMRFlow with (i) vanilla MeanFlow parameterization; (ii) a CTM-style preconditioned parameterization, which utilizes the flow map decomposition (2)

$$X_\theta(z, r, t) = \frac{r}{t} z + \left(1 - \frac{r}{t}\right) G_\theta(z, r, t);$$

and (iii) a hard-cluster-conditioned (HC) version of the CTM-style parametrization. HC and GMRFlow share the same cluster labels and neural architecture. The CelebA models are trained for 50K updates, and the FFHQ models are trained for 100K updates. Table 1 and Figure 5 show that

GMRFlow achieves the lowest terminal 1-NFE FID among the evaluated baselines on each dataset and faster convergence in the training curves.

Table 1: 1-NFE FID comparison.

Method	CelebA ↓	FFHQ ↓
MF	59.76	–
CTM	55.18	76.33
HC	50.55	70.72
GMRFlow	**40.65**	**53.90**

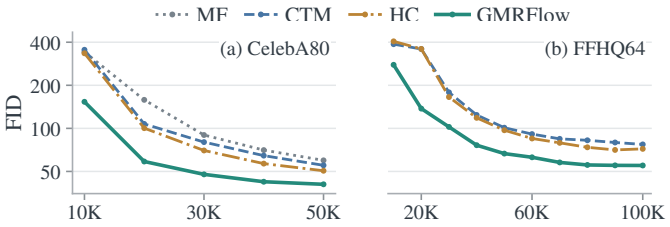


Figure 5: Training curves for 1-NFE FID.

Ablation studies. We study four GMRFlow design choices on FFHQ64 using the same 21.85M U-Net and iMF objective (Geng et al., 2026). Table 2 reports results at 50K updates. For each ablation, all other training settings are kept fixed. Ranks 16 and 32 yield similar one-step FID; increasing the rank to 128 raises FID to 57.06. At rank 64, $K = 4$ gives the lowest observed FID, close to $K = 8$.

DINO-based clustering outperforms the two pixel-based alternatives. For time sampling, an off-diagonal fraction of 50% yields the lowest one-step FID; increasing it to 75% worsens generation quality. Detailed ablation protocols and additional results are given in Appendix F.1.

Rank	1-NFE FID ↓
16	**33.43**
32	33.87
64	37.33

(a) **Gaussian rank.**

K	1-NFE FID ↓
4	**36.57**
8	37.33
16	45.72

(b) **Components.**

Clustering	1-NFE FID ↓
DINO	**37.33**
Pixel K-means	41.70
Pixel EM	41.68

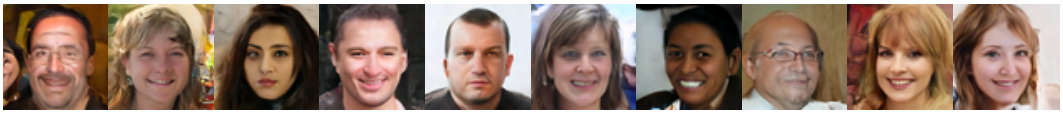
(c) **Clustering method.**

% of $r \neq t$	1-NFE FID ↓
25%	35.80
50%	**33.43**
75%	55.19

(d) **Time sampling.**

Table 2: **Ablations on FFHQ64.** EMA 1-NFE FID on 50K generated images after 50K updates.

We use the best-performing configuration from the ablations (rank 16, $K = 8$, DINO clustering, and a 50% off-diagonal fraction) to train GMRFlow for 100K updates. The selected model achieves 1-NFE FID 24.11 and 2-NFE FID 16.8. Figure 6 shows selected 2-NFE samples generated by this model on FFHQ64.



GMRFlow, 2-NFE FID = 16.80

Figure 6: Selected 2-NFE samples from GMRFlow on FFHQ64 after 100K training updates.

9 Conclusion

We establish transport loss as a surrogate for flow-map regression and clarify its connections to existing one-step objectives. Under latent low-rank Gaussian mixture structure, we derive learning bounds with polynomial dependence on intrinsic and latent dimensions. These insights motivate GMRFlow, which combines analytic Gaussian flow maps with neural residuals and improves one-step FID on CelebA80 and FFHQ64. Extending the guarantees to approximate mixture structure and practical neural architectures remains an important direction.

AI Use Statement

Generative AI tools were used to assist with mathematical derivations, experimental design, method implementation, and interpretation of experimental results. They were also used for literature search and analysis, drafting and editing parts of the manuscript, LaTeX formatting, and software code

development and debugging. Other tasks requiring disclosure under the ICLR AI policy were either not performed with generative AI or were not applicable to this work.

All AI-assisted mathematical claims and proofs were independently checked by the authors, and AI-assisted code used in the experiments was reviewed and tested for correctness. Suggested references and literature-based claims were verified against the original sources. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including its theoretical claims, proofs, code, and experimental results.

Ethics Statement

This work studies the theory and design of flow-map models using existing image datasets. Models trained on face images may reproduce training examples, inherit demographic biases, or be misused for impersonation and misleading content. The reported image-quality metrics do not establish privacy or fairness guarantees. Applications should respect dataset usage terms and individual privacy, and include safeguards against deceptive or harmful uses of generated images.

Reproducibility Statement

The theoretical assumptions and model constructions are stated in the main text, with supporting derivations in the appendices. Appendix D describes the Gaussian component flow maps and residual parameterization. Appendices E, F, and G document data preprocessing, experimental protocols, training configurations, and evaluation settings.

References

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *arXiv preprint arXiv:2406.07507*, 2024.

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *arXiv preprint arXiv:2505.18825*, 2025.

Bradley CA Brown, Anthony L Caterini, Brendan Leigh Ross, Jesse C Cresswell, and Gabriel Loaiza-Ganem. Verifying the union of manifolds hypothesis for image data. *arXiv preprint arXiv:2207.02862*, 2022.

Minshuo Chen, Kaixuan Huang, Tuo Zhao, and Mengdi Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pp. 4672–4712. PMLR, 2023.

Sitan Chen, Vasilis Kontonis, and Kulin Shah. Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*, 2024.

Yiding Chen, Yiyi Zhang, Owen Oertell, and Wen Sun. Convergence of consistency model with multistep sampling under general data assumptions. *arXiv preprint arXiv:2505.03194*, 2025.

Hugo Cui, Florent Krzakala, Eric Vanden-Eijnden, and Lenka Zdeborová. Analysis of learning a flow-based generative model from limited sample complexity. *arXiv preprint arXiv:2310.03575*, 2023.

Zehao Dou, Minshuo Chen, Mengdi Wang, and Zhuoran Yang. Theory of consistency diffusion models: Distribution estimation meets fast sampling. In *Forty-first International Conference on Machine Learning*, 2024.

Nicolas Fournier. Convergence of the empirical measure in expected Wasserstein distance: non-asymptotic explicit bounds in $\mathbb{R}^d$. *ESAIM: Probability and Statistics*, 27:749–775, 2023.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30467–30476, 2026.

Yingqing Guo, Hui Yuan, Yukang Yang, Minshuo Chen, and Mengdi Wang. Gradient guidance for diffusion models: An optimization perspective. *Advances in Neural Information Processing Systems*, 37:90736–90770, 2024.

Zahra Kadkhodaie, Florentin Guth, Eero Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, volume 2024, pp. 46543–46567, 2024.

Hamidreza Kamkari, Brendan L Ross, Rasa Hosseinzadeh, Jesse C Cresswell, and Gabriel Loaiza-Ganem. A geometric view of data complexity: Efficient local intrinsic dimension estimation with diffusion models. *Advances in Neural Information Processing Systems*, 37:38307–38354, 2024.

Gen Li, Zhihan Huang, and Yuting Wei. Towards a mathematical theory for consistency training in diffusion models. *arXiv preprint arXiv:2402.07802*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024.

Junlong Lyu, Zhitang Chen, and Shoubo Feng. Sampling is as easy as keeping the consistency: convergence guarantee for consistency models. In *Forty-first International Conference on Machine Learning*, 2024.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pp. 26517–26582. PMLR, 2023.

Phillip Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. *arXiv preprint arXiv:2104.08894*, 2021.

Kulin Shah, Sitan Chen, and Adam Klivans. Learning mixtures of gaussians using the ddpm objective. *Advances in Neural Information Processing Systems*, 36:19636–19649, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 32211–32252, 2023.

Peng Wang, Huijie Zhang, Zekai Zhang, Siyi Chen, Yi Ma, and Qing Qu. Diffusion models learn low-dimensional distributions via subspace clustering. In *2025 IEEE 10th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 211–215. IEEE, 2025.

Ruofeng Yang, Yongcan Li, Bo Jiang, Cheng Chen, and Shuai Li. Multi-subspace multi-modal modeling for diffusion models: Estimation, convergence and mixture of experts. *arXiv preprint arXiv:2601.01475*, 2026.

Hui Yuan, Kaixuan Huang, Chengzhuo Ni, Minshuo Chen, and Mengdi Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. *Advances in Neural Information Processing Systems*, 36:60599–60635, 2023.

Zekai Zhang, Xiao Li, Xiang Li, Lianghe Shi, Meng Wu, Molei Tao, and Qing Qu. Generalization of diffusion models arises with a balanced representation space. *arXiv preprint arXiv:2512.20963*, 2025.

Appendix

CONTENTS

A Proofs in Section 4 **13**
A.1 Proof of Theorem 4.1 . . . 13
A.2 Proof of Theorem 4.2 . . . 13
A.3 Proof of Corollary 4.3 . . . 14
A.4 Proofs in Section 4.2 . . . 14

B Proofs in Section 5 **17**
B.1 Auxiliary Lemmas . . . 17
B.2 Proof of Theorem 5.3 . . . 18

C Proofs in Section 6 **19**
C.1 Proof of Proposition 6.2 . . . 19
C.2 Proof of Equation 3 . . . 19
C.3 Preliminary Estimates for Gaussian-Mixture Velocities . . . 20
C.4 Proof of Theorem 6.5 . . . 24
C.5 Proof of Theorem 6.6 . . . 26

D Details of Gaussian Mixture Residual Flow **31**
D.1 Proof of Equation 4 . . . 31
D.2 Residual Parameterization and Preconditioning . . . 32

E Additional Details of Numerical Validation **33**
E.1 Continuous and Discrete Objectives . . . 33
E.2 Data Subspace Structure . . . 33
E.3 Flow Map Decomposition . . . 33

F Additional Details and Results for Generative Experiments **35**
F.1 Data, Model Variants, and Training Settings . . . 35

G Experiments on ImageNet64 **36**

A PROOFS IN SECTION 4

A.1 PROOF OF THEOREM 4.1

Proof. For any continuously differentiable f, the chain rule along the PF-ODE gives

$$\frac{\mathrm{d}}{\mathrm{d}t} f(z_t, t) = \partial_t f(z_t, t) + v(z_t, t) \cdot \nabla_z f(z_t, t) = \mathcal{D} f(z_t, t).$$

Since $X(z_t, r, t) = z_r$ is constant along a trajectory for fixed r, the true flow map has zero transport loss. Now suppose that $\mathcal{L}_{\mathrm{tr}}(\theta) = 0$, we know $\mathcal{D}X_\theta(z_t, r, t) = 0$ holds almost surely for $(z_t, r, t) \sim \rho_t \otimes \nu$. Because $\mathcal{D}X_\theta$ is continuous and the sampling distribution has full support, we have $\mathcal{D}X_\theta(z, r, t) = 0$ throughout $\mathbb{R}^D \times \mathcal{T}_\epsilon$. Take any fixed $(z, r, t) \in \mathbb{R}^D \times \mathcal{T}_\epsilon$, and PF-ODE trajectory z_s with $z_t = z$, we have

$$\frac{\mathrm{d}}{\mathrm{d}s} X_\theta(z_s, r, s) = \mathcal{D}X_\theta(z_s, r, s) = 0.$$

This gives $X_\theta(z_t, r, t) = X_\theta(z_r, r, r)$. By the boundary condition, we have $X_\theta(z_r, r, r) = z_r$. So we conclude that $X_\theta(z_t, r, t) = z_r = X(z_t, r, t)$, which proves the theorem. □

A.2 PROOF OF THEOREM 4.2

Proof. The boundary condition $X_\theta(z_r, r, r) = z_r$ and the chain rule yield

$$X_\theta(z_t, r, t) - X(z_t, r, t) = \int_r^t \mathcal{D}X_\theta(z_s, r, s)\, \mathrm{d}s.$$

By Cauchy–Schwarz,

$$\|X_\theta(z_t, r, t) - X(z_t, r, t)\|^2 \le (t - r) \int_r^t \|\mathcal{D}X_\theta(z_s, r, s)\|^2\, \mathrm{d}s.$$

Taking expectation over $z_t \sim \rho_t$ gives

$$\mathbb{E}_{z_t \sim \rho_t}\left[\|X_\theta(z_t, r, t) - X(z_t, r, t)\|^2\right] \le (t - r) \int_r^t \mathbb{E}_{z_t \sim \rho_t}\left[\|\mathcal{D}X_\theta(z_s, r, s)\|^2\right]\, \mathrm{d}s.$$

Since the true flow map pushes ρ_s to ρ_t, the expectation over $z_t \sim \rho_t$ is equivalent to the expectation over $z_s \sim \rho_s$. This gives

$$\mathbb{E}_{z_t \sim \rho_t}\left[\|X_\theta(z_t, r, t) - X(z_t, r, t)\|^2\right] \le (t - r) \int_r^t \mathbb{E}_{z_s \sim \rho_s}\left[\|\mathcal{D}X_\theta(z_s, r, s)\|^2\right]\, \mathrm{d}s. \tag{6}$$

Now take expectation over $t \sim \mathrm{Unif}[\epsilon, 1], r \sim \mathrm{Unif}[\epsilon, t]$, and apply the Fubini–Tonelli theorem with $f_\theta(r, s) = \mathbb{E}_{z_s \sim \rho_s}\left[\|\mathcal{D}X_\theta(z_s, r, s)\|^2\right]$, we have

$$\begin{aligned}
&\mathbb{E}_{t \sim \mathrm{Unif}[\epsilon,1], r \sim \mathrm{Unif}[\epsilon,t], z_t \sim \rho_t}\left[\|X_\theta(z_t, r, t) - X(z_t, r, t)\|^2\right] \\
\le &\frac{1}{1-\epsilon} \int_\epsilon^1 \frac{1}{t-\epsilon} \int_\epsilon^t (t - r) \int_r^t f_\theta(r, s)\, \mathrm{d}s\, \mathrm{d}r\, \mathrm{d}t \\
= &\frac{1}{1-\epsilon} \int_\epsilon^1 \frac{1}{s-\epsilon} \int_\epsilon^s \int_s^1 \frac{(s-\epsilon)(t-r)}{t-\epsilon}\, \mathrm{d}t\, f_\theta(r, s)\, \mathrm{d}r\, \mathrm{d}s.
\end{aligned}$$

Now consider

$$\begin{aligned}
\int_s^1 \frac{(s-\epsilon)(t-r)}{t-\epsilon}\, \mathrm{d}t &= \left((1-s) - (r-\epsilon)\log\frac{1-\epsilon}{s-\epsilon}\right)(s-\epsilon) \\
&\le (s-\epsilon)(1-s).
\end{aligned}$$

Let $y = s - \epsilon \in [0, 1-\epsilon]$, we deduce

$$(s-\epsilon)(1-s) = y((1-\epsilon) - y) \le \frac{(1-\epsilon)^2}{4} \le \frac{1}{4}.$$

We conclude that

$$\begin{aligned}
&\mathbb{E}_{t\sim\mathrm{Unif}[\epsilon,1],r\sim\mathrm{Unif}[\epsilon,t],z_t\sim\rho_t}\left[\|X_\theta(z_t,r,t)-X(z_t,r,t)\|^2\right]\\
\leq&\frac{1}{4}\times\frac{1}{1-\epsilon}\int_\epsilon^1\frac{1}{s-\epsilon}\int_\epsilon^s f_\theta(r,s)\,\mathrm{d}r\,\mathrm{d}s\\
=&\frac{1}{4}\mathbb{E}_{s\sim\mathrm{Unif}[\epsilon,1],r\sim\mathrm{Unif}[\epsilon,s],z_s\sim\rho_s}\left[\|\mathcal{D}X_\theta(z_s,r,s)\|^2\right]
\end{aligned}$$

This proves the theorem. □

A.3 Proof of Corollary 4.3

Proof. By equation (6), we have

$$\begin{aligned}
\mathbb{E}_{z_1\sim\rho_1}\left[\|X_\theta(z_1,\epsilon,1)-X(z_1,\epsilon,1)\|^2\right]&\leq(1-\epsilon)\int_\epsilon^1\mathbb{E}_{z_s\sim\rho_s}\left[\|\mathcal{D}X_\theta(z_s,\epsilon,s)\|^2\right]\,\mathrm{d}s\\
&=(1-\epsilon)^2\mathcal{L}_{\mathrm{tr}}(\theta)
\end{aligned}$$

For distributions $\rho_\epsilon, X_\theta(\cdot,\epsilon,1)_\sharp\rho_1$, take a random variable $z_1\sim\rho_1$ and define a coupling

$$x=X_\theta(z_1,\epsilon,1),\quad y=X(z_1,\epsilon,1).$$

By the definition of Wasserstein distance

$$\begin{aligned}
W_2^2(\rho_\epsilon,X_\theta(\cdot,\epsilon,1)_\sharp\rho_1)&\leq\mathbb{E}_{z_1\sim\rho_1}\left[\|X_\theta(z_1,\epsilon,1)-X(z_1,\epsilon,1)\|^2\right]\\
&\leq(1-\epsilon)^2\mathcal{L}_{\mathrm{tr}}(\theta)\leq\mathcal{L}_{\mathrm{tr}}(\theta)
\end{aligned}$$

This proves the corollary. □

A.4 Proofs in Section 4.2

MeanFlow Loss is equivalent to Transport Loss The average velocity is $u(z_t,r,t)=(t-r)^{-1}\int_r^t v(z_s,s)\,\mathrm{d}s$, so $X(z,r,t)=z-(t-r)u(z,r,t)$. The MeanFlow identity $u=v-(t-r)\mathcal{D}u$ motivates the target and loss

$$u_{\mathrm{tgt}}=v(z_t,t)-(t-r)\mathcal{D}u_\theta(z_t,r,t),\qquad\mathcal{L}_{\mathrm{MF}}(\theta)=\mathbb{E}\left[\|u_\theta(z_t,r,t)-u_{\mathrm{tgt}}\|^2\right].$$

Proof. Substitute

$$\begin{aligned}
\mathcal{D}X_\theta(z_t,r,t)&=\partial_tX_\theta(z_t,r,t)+v(z_t,t)\cdot\nabla_zX_\theta(z_t,r,t)\\
&=\partial_t\big(z-(t-r)u_\theta(z,r,t)\big)\big|_{z=z_t}+v(z_t,t)\cdot\nabla_z\big(z-(t-r)u_\theta(z,r,t)\big)\big|_{z=z_t}\\
&=-u_\theta(z_t,r,t)-(t-r)\partial_tu_\theta(z_t,r,t)+v(z_t,t)-(t-r)v(z_t,t)\cdot\nabla_zu_\theta(z_t,r,t)\\
&=v(z_t,t)-u_\theta(z_t,r,t)-(t-r)\mathcal{D}u_\theta(z_t,r,t)\\
&=-\left[u_\theta(z_t,r,t)-v(z_t,t)+(t-r)\mathcal{D}u_\theta(z_t,r,t)\right].
\end{aligned}$$

Therefore, $\mathcal{L}_{\mathrm{MF}}(\theta)=\mathbb{E}\left[\|\mathcal{D}X_\theta(z_t,r,t)\|^2\right]=\mathcal{L}_{\mathrm{tr}}(\theta)$. □

CM Loss converges to Transport Loss

Proof. Recall $f_\theta(z,t)=X_\theta(z,\epsilon,t)$ and $z_{t_{i-1}}^{\hat{v}}=z_{t_i}-(t_i-t_{i-1})\hat{v}(z_{t_i},t_i)$. Define the per-step loss

$$\mathcal{L}_{\mathrm{CM}}^{(i)}(\theta)=\mathbb{E}_{z_{t_i}\sim\rho_{t_i}}\left[\|f_\theta(z_{t_i},t_i)-f_\theta(z_{t_{i-1}}^{\hat{v}},t_{i-1})\|^2\right],$$

We have $\mathcal{L}_{\mathrm{CM}}(\theta)=\frac{1}{n}\sum_{i=1}^n\mathcal{L}_{\mathrm{CM}}^{(i)}(\theta)$. Take uniform time step $t_i=\epsilon+i\Delta$ with $\Delta=(1-\epsilon)/n$, and let the velocity teacher be exact $\hat{v}=v$. We show that the rescaled loss $\Delta^{-2}\mathcal{L}_{\mathrm{CM}}(\theta)$ converges to transport loss $\mathcal{L}_{\mathrm{tr}}(\theta)$ as $n\to\infty$. Fix i, for $s\in[0,1]$ define

$$g_i(s)=f_\theta\left(z_{t_i}-s\Delta\hat{v}(z_{t_i},t_i),\,t_i-s\Delta\right).$$

We have $g_i(0) = f_\theta(z_{t_i}, t_i)$ and $g_i(1) = f_\theta(z^{\hat{v}}_{t_{i-1}}, t_{i-1})$. Suppose $f_\theta, \hat{v}$ are twice continuously differentiable, Taylor's theorem gives

$$\begin{aligned} f_\theta(z_{t_i}, t_i) - f_\theta(z^{\hat{v}}_{t_{i-1}}, t_{i-1}) &= -g_i'(0) + O(\Delta^2) \\ &= \Delta \left[\hat{v}(z_{t_i}, t_i) \cdot \nabla_z f_\theta(z_{t_i}, t_i) + \partial_t f_\theta(z_{t_i}, t_i)\right] + O(\Delta^2) \\ &= \Delta \, \mathcal{D} f_\theta(z_{t_i}, t_i) + O(\Delta^2). \end{aligned}$$

Taking squared norms and expectations,

$$\begin{aligned} \mathcal{L}^{(i)}_{\mathrm{CM}}(\theta) &= \mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[\| \Delta \, \mathcal{D} f_\theta(z_{t_i}, t_i) + O(\Delta^2)\|^2\right] \\ &= \Delta^2 \, \mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[\|\mathcal{D} f_\theta(z_{t_i}, t_i)\|^2\right] + 2\Delta \, \mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[\mathcal{D} f_\theta(z_{t_i}, t_i) O(\Delta^2)\right] + \mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[O(\Delta^4)\right] \\ &= \Delta^2 \, \mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[\|\mathcal{D} f_\theta(z_{t_i}, t_i)\|^2\right] + O(\Delta^3). \end{aligned}$$

Dividing by Δ^2, we obtain

$$\frac{1}{\Delta^2} \mathcal{L}^{(i)}_{\mathrm{CM}}(\theta) = \mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[\|\mathcal{D} f_\theta(z_{t_i}, t_i)\|^2\right] + O(\Delta).$$

Averaging over the time steps and letting $n \to \infty$ gives

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{\Delta^2} \mathcal{L}_{\mathrm{CM}}(\theta) &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left(\mathbb{E}_{z_{t_i} \sim \rho_{t_i}} \left[\|\mathcal{D} f_\theta(z_{t_i}, t_i)\|^2\right] + O(\Delta)\right) \\ &= \frac{1}{1-\epsilon} \int_\epsilon^1 \mathbb{E}_{z_t \sim \rho_t} \left[\|\mathcal{D} f_\theta(z_t, t)\|^2\right] \mathrm{d}t \\ &= \mathbb{E}_{t \sim \mathrm{Unif}[\epsilon, 1], z_t \sim \rho_t} \left[\|\mathcal{D} X_\theta(z_t, \epsilon, t)\|^2\right] = \mathcal{L}_{\mathrm{tr}}(\theta), \end{aligned}$$

This proves that CM loss converges to transport loss as the time step goes to zero. □

Proof of Theorem 4.4

Proof. For the linear ODE $\frac{\mathrm{d}z}{\mathrm{d}s} = Az$, the exact flow map is $X(z, r, t) = \exp\left((r-t)A\right) z$. Let the uniform time step be Δ. So the exact consistency function satisfies

$$X(z, \epsilon, t_i) = \exp\left(-\tau_i A\right) z, \qquad \tau_i := t_i - \epsilon = i\Delta.$$

With the exact velocity teacher $\hat{v}(z, t) = Az$, we have

$$z^v_{t_i} = z_{t_{i+1}} - \Delta A z_{t_{i+1}} = (I - \Delta A) z_{t_{i+1}}.$$

Since the consistency function f_θ is continuous and attains zero CM loss, we have

$$f_\theta(z, t_{i+1}) = f_\theta\left((I - \Delta A) z, t_i\right), \qquad \forall z \in \mathbb{R}^D.$$

By the boundary condition $f_\theta(z, \epsilon) = z$, we can recursively compute f_θ at the grid points

$$f_\theta(z, t_i) = (I - \Delta A)^i z, \qquad i = 0, \ldots, n.$$

Hence

$$f_\theta(z, t_i) - X(z, \epsilon, t_i) = \left[(I - \Delta A)^i - \exp(-\tau_i A)\right] z.$$

Since $i = \tau_i / \Delta$, we have

$$(I - \Delta A)^i = \exp\left(\frac{\tau_i}{\Delta} \log(I - \Delta A)\right).$$

For sufficiently small Δ, apply the Taylor expansion under the spectral norm

$$\begin{aligned} \log(I - \Delta A) &= -\Delta A - \frac{\Delta^2}{2} A^2 + O(\Delta^3) \\ \frac{\tau_i}{\Delta} \log(I - \Delta A) &= -\tau_i A - \frac{\tau_i}{2} \Delta A^2 + O(\Delta^2) \\ (I - \Delta A)^i &= \exp(-\tau_i A) \exp\left(-\frac{\tau_i}{2} \Delta A^2 + O(\Delta^2)\right) \\ (I - \Delta A)^i &= \exp(-\tau_i A) \left(I - \frac{\tau_i}{2} \Delta A^2 + O(\Delta^2)\right) \\ (I - \Delta A)^i - \exp(-\tau_i A) &= -\frac{\tau_i}{2} \Delta \exp(-\tau_i A) A^2 + O(\Delta^2). \end{aligned}$$

Applying this identity to z yields

$$f_\theta(z,t_i) - X(z,\epsilon,t_i) = -\frac{\tau_i}{2}\Delta\, \exp(-\tau_i A)A^2 z + O(\Delta^2)\|z\|$$
$$\implies \|f_\theta(z,t_i) - X(z,\epsilon,t_i)\|^2 = \frac{\tau_i^2}{4}\Delta^2 \|\exp(-\tau_i A)A^2 z\|^2 + O(\Delta^3)\|z\|^2.$$

By $\rho_1 = \mathcal{N}(0, I_D)$ and linear ODE, we know that ρ_t is a Gaussian distribution with finite second moment, hence $\mathbb{E}\left[\|z\|^2\right] < \infty$. So taking expectation over $z_{t_i} \sim \rho_{t_i}$ gives

$$\mathbb{E}_{z_{t_i}\sim\rho_{t_i}}\left[\|f_\theta(z_{t_i},t_i) - X(z_{t_i},\epsilon,t_i)\|^2\right] = \frac{\tau_i^2}{4}\Delta^2\, \mathbb{E}_{z_{t_i}\sim\rho_{t_i}}\left[\|\exp(-\tau_i A)A^2 z_{t_i}\|^2\right] + O(\Delta^3).$$

Since $\exp(-\tau_i A)$ is invertible and $A^2 \neq 0$, the matrix $B := \exp(-\tau_i A)A^2$ is nonzero, and

$$\begin{aligned}\mathbb{E}_{z_{t_i}\sim\rho_{t_i}}\left[\|\exp(-\tau_i A)A^2 z_{t_i}\|^2\right] &= \mathbb{E}_{z_{t_i}\sim\rho_{t_i}}\left[z_{t_i}^\top B^\top B z_{t_i}\right] \\ &= \mathrm{Tr}\left(B^\top B\, \Sigma_i\right).\end{aligned}$$

Here $\Sigma_i = \mathrm{Cov}(z_{t_i})$. Because $\Sigma_i \succ 0$ and $B \neq 0$, we have $\mathrm{Tr}(B^\top B\, \Sigma_i) > 0$. Therefore the leading term in the expansion is nonzero, and we conclude that for any grid point with fixed $\tau_i > 0$ as $\Delta \to 0$,

$$\mathbb{E}_{z_{t_i}\sim\rho_{t_i}}\left[\|f_\theta(z_{t_i},t_i) - X(z_{t_i},\epsilon,t_i)\|^2\right] = \Theta(\Delta^2).$$

This completes the proof. □

Continuous CM Loss is equivalent to MeanFlow Loss with Stop Gradient

Proof. Recall the continuous CM (sCM) gradient, which is defined as

$$\nabla_\theta \mathbb{E}\left[w(t) f_\theta(z_t,t)^\top \frac{\mathrm{d} f_{\theta^-}(z_t,t)}{\mathrm{d}t}\right],$$

where $w(t)$ is a weighting function, and θ^- denotes the stop-gradient branch. The MeanFlow parametrization gives

$$f_\theta(z_t,t) = z_t - (t-\epsilon)u_\theta(z_t,\epsilon,t). \tag{7}$$

We write $J_u := \partial_\theta u_\theta(z_t,\epsilon,t)$ and $J_f := \partial_\theta f_\theta(z_t,t)$, so $J_f = -(t-\epsilon)J_u$. Since the second factor in the sCM expression is stopped, its gradient is

$$\begin{aligned}\nabla_\theta \mathbb{E}\left[w(t) f_\theta(z_t,t)^\top \frac{\mathrm{d} f_{\theta^-}(z_t,t)}{\mathrm{d}t}\right] &= \mathbb{E}\left[w(t) J_f^\top \mathcal{D}X_\theta(z_t,\epsilon,t)\right] \\ &= -\mathbb{E}\left[w(t)(t-\epsilon)J_u^\top \mathcal{D}X_\theta(z_t,\epsilon,t)\right].\end{aligned}$$

On the other hand, by the preceding proofs, stop-gradient through the entire MeanFlow target gives

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\mathrm{MF}}^{\mathrm{sg}}(\theta) &:= \nabla_\theta \mathbb{E}\left[\|u_\theta(z_t,\epsilon,t) - \mathrm{sg}[u_{\mathrm{tgt}}]\|^2\right] \\ &= 2\mathbb{E}\left[J_u^\top\left(u_\theta(z_t,\epsilon,t) - u_{\mathrm{tgt}}\right)\right] \\ &= -2\mathbb{E}\left[J_u^\top \mathcal{D}X_\theta(z_t,\epsilon,t)\right].\end{aligned}$$

Consequently, choosing

$$w(t) = \frac{2}{t-\epsilon}$$

makes the sCM gradient exactly equal to the stop-gradient MeanFlow gradient for $t > \epsilon$.

The sCM loss function is defined as

$$\mathcal{L}_{\mathrm{sCM}}(\theta) := \mathbb{E}\left[\frac{w(t)^2}{4}\left\|f_\theta(z_t,t) - f_{\theta^-}(z_t,t) + (t-\epsilon)\frac{\mathrm{d} f_{\theta^-}(z_t,t)}{\mathrm{d}t}\right\|^2\right],$$

such that

$$\nabla_\theta \mathcal{L}_{\mathrm{sCM}}(\theta) = \nabla_\theta \mathbb{E}\left[w(t) f_\theta(z_t,t)^\top \frac{\mathrm{d} f_{\theta^-}(z_t,t)}{\mathrm{d}t}\right].$$

Substituting the chosen weight $w(t) = 2/(t-\epsilon)$ gives

$$\mathcal{L}_{\mathrm{sCM}}(\theta) := \mathbb{E}\left[\frac{1}{(t-\epsilon)^2}\left\|f_\theta(z_t,t) - f_{\theta^-}(z_t,t) + (t-\epsilon)\frac{\mathrm{d}f_{\theta^-}(z_t,t)}{\mathrm{d}t}\right\|^2\right].$$

Hence the residual in the sCM loss satisfies

$$\begin{aligned}
&\frac{f_\theta(z_t,t) - f_{\theta^-}(z_t,t)}{t-\epsilon} + \frac{\mathrm{d}f_{\theta^-}(z_t,t)}{\mathrm{d}t} \\
&\quad = -\left(u_\theta(z_t,\epsilon,t) - u_{\theta^-}(z_t,\epsilon,t)\right) + \mathcal{D}X_{\theta^-}(z_t,\epsilon,t) && \text{(by 7)} \\
&\quad = -\left(u_\theta(z_t,\epsilon,t) - \mathrm{sg}[u_{\mathrm{tgt}}]\right). && \text{(by expanding } u_{\mathrm{tgt}})
\end{aligned}$$

Squaring and taking expectation therefore yields the exact identity

$$\mathcal{L}_{\mathrm{sCM}}(\theta) = \mathbb{E}\left[\left\|u_\theta(z_t,\epsilon,t) - \mathrm{sg}[u_{\mathrm{tgt}}]\right\|^2\right] = \mathcal{L}^{\mathrm{sg}}_{\mathrm{MF}}(\theta).$$

Thus the sCM loss and the stop-gradient MeanFlow loss have both the same forward value and the same gradient update. □

B Proofs in Section 5

Let μ denote the sampling distribution of $w = (z_t, r, t)$ specified in Section 5, and let $\hat{\mu}_n$ be its empirical measure on the n training samples. We write $\hat{\mathcal{L}}_{\mathrm{tr},n}(X) = \int \|\mathcal{D}X(w)\|_2^2 \,\mathrm{d}\hat{\mu}_n(w)$ for the empirical transport loss.

B.1 Auxiliary Lemmas

Lemma B.1 (Wasserstein control of regression error)**.** *Under the assumptions in Section 5, let $M_g = M_X(1+B_v)$ and $L_g = (1+B_v)L_X + M_X L_v$. For every realization of the training sample, the empirical risk minimizer $\hat{X}_n$ satisfies*

$$\mathbb{E}\left[\|\hat{X}_n - X^*\|_2^2\right] \le \frac{M_g L_g}{2} W_1(\mu, \hat{\mu}_n).$$

Proof. For $w = (z,r,t) \in \Omega_R$, define the transport residual and single-sample loss by

$$g_X(w) := \mathcal{D}X(z,r,t) = \partial_t X(w) + \nabla_z X(w) v(z,t), \qquad \ell_X(w) := \|g_X(w)\|_2^2.$$

We first show that ℓ_X is uniformly $2M_gL_g$-Lipschitz over $X \in \mathcal{X}_{\mathrm{Lip}}(M_X, L_X)$. Write $a(w) := (v(z,t)^\top, 0, 1)^\top \in \mathbb{R}^{D+2}$, so that $g_X(w) = DX(w)a(w)$. The velocity assumptions give

$$\|a(w)\|_2 \le 1 + B_v, \qquad \|a(w) - a(w')\|_2 \le L_v \|w - w'\|_2.$$

Consequently, the bounds on the full Jacobian imply

$$\|g_X(w)\|_2 \le M_X(1+B_v) = M_g,$$

and

$$\begin{aligned}
\|g_X(w) - g_X(w')\|_2 &\le \|DX(w) - DX(w')\|_{\mathrm{op}} \|a(w)\|_2 + \|DX(w')\|_{\mathrm{op}} \|a(w) - a(w')\|_2 \\
&\le \big((1+B_v)L_X + M_X L_v\big)\|w - w'\|_2 = L_g \|w - w'\|_2.
\end{aligned}$$

It follows that

$$\begin{aligned}
|\ell_X(w) - \ell_X(w')| &\le \big(\|g_X(w)\|_2 + \|g_X(w')\|_2\big)\|g_X(w) - g_X(w')\|_2 \\
&\le 2M_g L_g \|w - w'\|_2.
\end{aligned}$$

Let π be any coupling of μ and $\hat{\mu}_n$. By the preceding Lipschitz bound,

$$\begin{aligned}
\left|\mathcal{L}_{\mathrm{tr}}(X) - \hat{\mathcal{L}}_{\mathrm{tr},n}(X)\right| &= \left|\int \big(\ell_X(w) - \ell_X(w')\big)\,\mathrm{d}\pi(w,w')\right| \\
&\le 2M_g L_g \int \|w - w'\|_2 \,\mathrm{d}\pi(w,w').
\end{aligned}$$

Taking the infimum over couplings gives, simultaneously for every $X \in \mathcal{X}_{\mathrm{Lip}}(M_X, L_X)$,

$$\left|\mathcal{L}_{\mathrm{tr}}(X) - \hat{\mathcal{L}}_{\mathrm{tr},n}(X)\right| \leq 2M_g L_g W_1(\mu, \hat{\mu}_n).$$

In particular, this inequality applies to the sample-dependent minimizer $\hat{X}_n$.

The true flow map satisfies $\mathcal{D}X^* = 0$ along the probability flow, as established in Section 4. Since $X^* \in \mathcal{X}_{\mathrm{Lip}}(M_X, L_X)$ and $\hat{X}_n$ is an exact empirical risk minimizer, nonnegativity of the loss yields

$$0 \leq \hat{\mathcal{L}}_{\mathrm{tr},n}(\hat{X}_n) \leq \hat{\mathcal{L}}_{\mathrm{tr},n}(X^*) = 0.$$

Hence the uniform bound above gives $\mathcal{L}_{\mathrm{tr}}(\hat{X}_n) \leq 2M_g L_g W_1(\mu, \hat{\mu}_n)$. Applying Theorem 4.2 implies

$$\mathbb{E}\left[\|\hat{X}_n - X^*\|_2^2\right] \leq \frac{1}{4}\mathcal{L}_{\mathrm{tr}}(\hat{X}_n) \leq \frac{M_g L_g}{2} W_1(\mu, \hat{\mu}_n).$$

This proves the stated squared-error bound for each realization of the training sample. □

Lemma B.2 (Empirical Wasserstein distance on a compact domain)**.** *Let ν be a probability measure supported on a compact subset of $\mathbb{R}^q$ with diameter at most Δ, where $q \geq 3$. For i.i.d. samples $Y_1, \ldots, Y_n \sim \nu$, let $\hat{\nu}_n = n^{-1}\sum_{i=1}^n \delta_{Y_i}$. There is a constant $C_q \leq C\sqrt{q}$, where $C > 0$ is universal, such that*

$$\mathbb{E}\left[W_1(\nu, \hat{\nu}_n)\right] \leq C_q \Delta n^{-1/q}.$$

Moreover, for every $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$W_1(\nu, \hat{\nu}_n) \leq \Delta\left(C_q n^{-1/q} + \sqrt{\frac{\log(1/\delta)}{2n}}\right).$$

Proof. Let $W_{1,\infty}$ denote the Wasserstein distance with ground metric $\|x - y\|_\infty$. After translation and scaling, the compact-support specialization of Fournier (2023, Theorem 1(iii) and Section 2.4) gives

$$\mathbb{E}\left[W_{1,\infty}(\nu, \hat{\nu}_n)\right] \leq C\Delta n^{-1/q},$$

where C is universal: the maximum-norm coefficient in that result is uniformly bounded for integer $q \geq 3$. Since $\|x - y\|_2 \leq \sqrt{q}\|x - y\|_\infty$, we obtain the expectation bound with $C_q = C\sqrt{q}$.

Replacing one sample changes $W_1(\nu, \hat{\nu}_n)$ by at most Δ/n, by coupling the unchanged samples identically and applying the triangle inequality. Thus McDiarmid's inequality gives, with probability at least $1 - \delta$,

$$W_1(\nu, \hat{\nu}_n) \leq \mathbb{E}\left[W_1(\nu, \hat{\nu}_n)\right] + \Delta\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Combining the two bounds proves the lemma. □

B.2 Proof of Theorem 5.3

Proof. The sampling distribution μ is supported on $\Omega_R \subseteq \mathbb{R}^{D+2}$. Set $q = D + 2 > 2$ and $\Delta_\Omega = \mathrm{diam}(\Omega_R) \leq \sqrt{4R^2 + 2}$. Lemma B.2 gives, with probability at least $1 - \delta$ over the training sample,

$$W_1(\mu, \hat{\mu}_n) \leq \Delta_\Omega\left(C_{D+2} n^{-1/(D+2)} + \sqrt{\frac{\log(1/\delta)}{2n}}\right).$$

On this event, Lemma B.1 implies

$$\begin{aligned}\mathbb{E}\left[\|\hat{X}_n - X^*\|_2^2\right] &\leq \frac{M_g L_g}{2} W_1(\mu, \hat{\mu}_n) \\ &\leq \frac{M_g L_g \Delta_\Omega}{2}\left(C_{D+2} n^{-1/(D+2)} + \sqrt{\frac{\log(1/\delta)}{2n}}\right).\end{aligned}$$

This finishes the proof of the theorem. □

C Proofs in Section 6

C.1 Proof of Proposition 6.2

Proof. Let $P = AA^\top$ and $P_\perp = I_D - P$ be the orthogonal projections onto V and $V^\perp$, respectively. Since $x_0 \in \mathcal{M} \subseteq V$ almost surely, we can write $x_0 = A\xi_0$. Decompose the Gaussian noise as

$$\varepsilon = A\eta + \varepsilon_\perp, \qquad \eta = A^\top \varepsilon \sim \mathcal{N}(0, I_d), \qquad \varepsilon_\perp = P_\perp \varepsilon.$$

The variables η and $\varepsilon_\perp$ are jointly Gaussian with zero cross-covariance, so they are independent. Together with the independence of x_0 and ε, this implies that $\varepsilon_\perp$ is independent of (ξ_0, η). For each fixed $s \in [\epsilon, 1]$, the interpolant therefore satisfies

$$A^\top x_s = (1-s)\xi_0 + s\eta, \qquad P_\perp x_s = s\varepsilon_\perp.$$

In particular, $P_\perp x_s$ is independent of $(x_0, A^\top x_s)$. Observing x_s is equivalent to observing its two projections, and the orthogonal projection provides no information about x_0. Hence

$$\mathbb{E}\left[x_0 \mid x_s\right] = \mathbb{E}\left[x_0 \mid A^\top x_s, P_\perp x_s\right] = \mathbb{E}\left[x_0 \mid A^\top x_s\right].$$

Define $m(\xi, s) = \mathbb{E}\left[x_0 \mid A^\top x_s = \xi\right]$. Since $x_0 = A\xi_0$, we have $m(\xi, s) = A\mathbb{E}\left[\xi_0 \mid A^\top x_s = \xi\right] \in V$. Using $\varepsilon - x_0 = (x_s - x_0)/s$, the velocity field becomes

$$v(z, s) = \mathbb{E}\left[\varepsilon - x_0 \mid x_s = z\right] = \frac{z - m(A^\top z, s)}{s}.$$

Now fix $(r, t) \in \mathcal{T}_\epsilon$ and consider the PF-ODE trajectory with $z_t = z$. Its latent and orthogonal coordinates obey

$$\frac{\mathrm{d}}{\mathrm{d}s}(A^\top z_s) = \frac{A^\top z_s - A^\top m(A^\top z_s, s)}{s},$$
$$\frac{\mathrm{d}}{\mathrm{d}s}(P_\perp z_s) = \frac{P_\perp z_s}{s}.$$

The second equation gives $P_\perp z_s = (s/t)P_\perp z$. The first is a closed d-dimensional ODE: its right-hand side does not depend on $P_\perp z_s$. Write ξ_s for its solution with terminal value $\xi_t = \xi = A^\top z$. To obtain the claimed decomposition, differentiate z_s/s along the trajectory:

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{z_s}{s}\right) = \frac{v(z_s, s)}{s} - \frac{z_s}{s^2} = -\frac{m(A^\top z_s, s)}{s^2}.$$

Integrating from t to r yields

$$\frac{z_r}{r} - \frac{z}{t} = \int_r^t \frac{m(A^\top z_s, s)}{s^2}\,\mathrm{d}s,$$

and hence

$$X(z, r, t) = z_r = \frac{r}{t}z + r\int_r^t \frac{m(\xi_s, s)}{s^2}\,\mathrm{d}s.$$

Define the residual by

$$h(\xi, r, t) := r\int_r^t \frac{m(\xi_s, s)}{s^2}\,\mathrm{d}s.$$

Each integrand belongs to the linear subspace V, so $h(\xi, r, t) \in V$. The latent trajectory is determined by (ξ, t), which shows that this residual depends only on (ξ, r, t). Finally, when $r = t$ the integration interval is empty and $h(\xi, t, t) = 0$. This completes the proof. □

C.2 Proof of Equation 3

Proof. Fix a component k. Let $\xi_0 \sim \mathcal{N}(\mu_k, \Sigma_k)$ and $\eta \sim \mathcal{N}(0, I_d)$ be independent, and set $\xi_t = (1-t)\xi_0 + t\eta$. For $t \in [\epsilon, 1]$, the covariance $C_{k,t} = (1-t)^2\Sigma_k + t^2 I_d$ is positive definite, even when Σ_k is singular. Since $\mathrm{Cov}(\xi_0, \xi_t) = (1-t)\Sigma_k$, Gaussian conditioning gives

$$\mathbb{E}\left[\xi_0 \mid \xi_t = \xi\right] = \mu_k + (1-t)\Sigma_k C_{k,t}^{-1}\big(\xi - (1-t)\mu_k\big).$$

The identity $\eta - \xi_0 = (\xi_t - \xi_0)/t$ therefore implies

$$\begin{aligned} v_k(\xi, t) &= \mathbb{E}\left[\eta - \xi_0 \mid \xi_t = \xi\right] \\ &= \frac{\xi}{t} - \frac{1-t}{t}\Sigma_k C_{k,t}^{-1}\xi - \frac{1}{t}\big(I_d - (1-t)^2\Sigma_k C_{k,t}^{-1}\big)\mu_k \\ &= \frac{\xi}{t} - \frac{1-t}{t}\Sigma_k C_{k,t}^{-1}\xi - tC_{k,t}^{-1}\mu_k, \end{aligned}$$

where the last equality uses $I_d - (1-t)^2\Sigma_k C_{k,t}^{-1} = t^2 C_{k,t}^{-1}$. This is exactly (3). □

C.3 Preliminary Estimates for Gaussian-Mixture Velocities

Proof roadmap. The proofs of Theorems 6.5 and 6.6 share several estimates for Gaussian-mixture velocities. We first establish stability, posterior moment transfer, and directional derivative bounds. For Theorem 6.5, these estimates control the local transport defects of the oracle Euler network; accumulating the defects gives the m^{-2} approximation bound. For Theorem 6.6, we establish moment and covering-number bounds for the loss class, then combine empirical–population comparisons with empirical risk minimization.

Throughout the following proofs, $C > 0$ denotes a generic constant depending only on $\epsilon, R_\lambda, R_\mu, L$. Different occurrences of C may represent different values. When additional dependence on moment orders is allowed, we indicate it explicitly: C_a may also depend on a, while $C_{a,p}$ may also depend on both a and p. All these constants are independent of D, d, d^*, K, m and of the smallest positive mixture weight or covariance eigenvalue. Let $\xi_0 \sim q$ and $\eta = A^\top\varepsilon \sim \mathcal{N}(0, I_d)$ be independent, and denote the law of $\xi_s = (1-s)\xi_0 + s\eta$ by q_s. We write P_ν for the law of (ξ_t, r, t) with $(r, t) \sim \nu$ and $\xi_t \mid (r, t) \sim q_t$. The transport derivative always uses the true latent velocity:

$$\mathcal{D}\Phi_\theta(\xi, r, t) = \partial_t\Phi_\theta(\xi, r, t) + D_\xi\Phi_\theta(\xi, r, t)v(\xi, t).$$

Component structure and stability. Write $\beta = \sum_k \gamma_k\beta_k = v - \xi/s$. For each component, let R_k have orthonormal columns spanning $W_k = \operatorname{span}\{U_k, \mu_k\}$, so that $\dim W_k \le 1 + d^*$. Define the *component projection scale* ℓ_k and the *posterior-weighted a-th moment* M_a by

$$\ell_k(\xi) = 1 + \|R_k^\top\xi\|, \qquad M_a(\xi, s) = \sum_{k=1}^K \gamma_k(\xi, s)\ell_k(\xi)^a. \tag{8}$$

Equation (3) shows that β_k takes values in W_k and depends only on $R_k^\top\xi$. Let $\varphi_{k,s}(\xi) = \mathcal{N}(\xi; (1-s)\mu_k, C_{k,s})$ and $c(s) = (1-s)/s$. The component score satisfies

$$\nabla_\xi \log\varphi_{k,s} = -C_{k,s}^{-1}\big(\xi - (1-s)\mu_k\big) = -\frac{\xi}{s^2} - c(s)\beta_k.$$

Differentiating the normalized posterior weights therefore gives

$$\nabla_\xi\gamma_k = \gamma_k\left(\nabla_\xi \log\varphi_{k,s} - \sum_j \gamma_j\nabla_\xi \log\varphi_{j,s}\right) = -c(s)\gamma_k(\beta_k - \beta). \tag{9}$$

Since $\sum_k \nabla_\xi\gamma_k = 0$, the common drift drops out when differentiating the mixture:

$$\begin{aligned} D_\xi v &= \sum_k \gamma_k D_\xi v_k + \sum_k \beta_k(\nabla_\xi\gamma_k)^\top \\ &= \sum_k \gamma_k D_\xi v_k - c(s)\left(\sum_k \gamma_k\beta_k\beta_k^\top - \beta\beta^\top\right) \\ &= \sum_k \gamma_k D_\xi v_k - c(s)\mathrm{Cov}_\gamma(\beta_k) \preceq s^{-1}I_d. \end{aligned} \tag{10}$$

Here Cov_γ denotes covariance over the component label with weights γ_k. The last inequality follows because $D_\xi v_k = s^{-1}I_d - (1-s)s^{-1}\Sigma_k C_{k,s}^{-1}$ is symmetric and bounded above by $s^{-1}I_d$. Together with Assumption 6.4, this yields, for $\epsilon \le a \le b \le 1$,

$$c(s)\mathrm{Cov}_\gamma(\beta_k) \preceq (L + \epsilon^{-1})I_d, \tag{11}$$

$$\frac{a}{b}I_d \preceq I_d - (b-a)D_\xi v(\xi, b) \preceq \big(1 + (b-a)L\big)I_d. \tag{12}$$

Thus an Euler step from b to a is $(1+(b-a)L)$-Lipschitz, and products of Euler-step Jacobians over $[r,t]$ have norm at most

$$\prod_{i=1}^{m}(1+\Delta L) \le e^{Lm\Delta} = e^{L(t-r)} \le e^{L}.$$

Moreover, $-(L+s^{-1})I_d \preceq D_\xi\beta \preceq 0$, so $\|D_\xi\beta\|_{\mathrm{op}} \le L+\epsilon^{-1}$. These identities hold for every admissible parameter as well as for the oracle parameter θ_{or}. The estimates below apply to any admissible layer, with all mixture-dependent quantities defined using that layer's parameters.

For a random vector Y, we write $\|Y\|_{L^p} := (\mathbb{E}\,[\|Y\|_2^p])^{1/p}$. The standard Gaussian norm estimate $\|G_j\|_{L^p} \le C\sqrt{j+p}$ for $G_j \sim \mathcal{N}(0, I_j)$ gives

$$\|\xi_0\|_{L^p} \le R_\mu + \sqrt{R_\lambda}\,\|G_{d^*}\|_{L^p} \le C\sqrt{1+d^*+p}, \qquad p \ge 2. \tag{13}$$

The same bound holds for every admissible layer's prior, since it holds for each component before averaging over its label.

Numerical states need not follow the corresponding noised marginal. The next lemma transfers posterior moment bounds to such states through a coupling.

Lemma C.1 (Posterior moment transfer)**.** *Fix $a \ge 1$, $p \ge 2$, and $s \in [\epsilon, 1]$. Let $\widetilde{\xi}$ be an arbitrary $\mathbb{R}^d$-valued random vector. For any coupling of $\xi_s \sim q_s$ and $\widetilde{\xi}$, we have*

$$\|M_a(\widetilde{\xi}, s)\|_{L^p} \le C_a\left(1 + d^* + p + \|\widetilde{\xi} - \xi_s\|_{L^{ap}}^2\right)^{a/2}. \tag{14}$$

The estimate also holds for any admissible layer and its own noised marginal, including layers with zero weights or zero eigenvalues.

Proof. We first control posterior moments under the matching marginal. Bayes' identity gives, for nonnegative functions f_k,

$$\mathbb{E}_{\xi_s\sim q_s}\left[\sum_k \gamma_k(\xi_s, s) f_k(\xi_s)\right] = \sum_k \pi_k \mathbb{E}_{\xi_s\sim\mathcal{N}((1-s)\mu_k, C_{k,s})}\left[f_k(\xi_s)\right].$$

Under the k-th component, $R_k^\top \xi_s$ has dimension at most $1+d^*$ and mean norm at most R_μ. Its covariance is $R_k^\top C_{k,s} R_k = s^2 I_{\dim W_k} + (1-s)^2 R_k^\top \Sigma_k R_k$. Since R_k has orthonormal columns and $\|\Sigma_k\|_{\mathrm{op}} \le R_\lambda$,

$$\|R_k^\top C_{k,s} R_k\|_{\mathrm{op}} \le s^2 + (1-s)^2\|\Sigma_k\|_{\mathrm{op}} \le 1 + R_\lambda.$$

Representing the projected Gaussian under component k using $G_{\dim W_k} \sim \mathcal{N}(0, I_{\dim W_k})$, the mean and covariance bounds give

$$\begin{aligned}\ell_k(\xi_s)^2 &\le \left(1 + R_\mu + \sqrt{1+R_\lambda}\,\|G_{\dim W_k}\|\right)^2 \\ &\le 2(1+R_\mu)^2 + 2(1+R_\lambda)\|G_{\dim W_k}\|^2.\end{aligned}$$

Thus, for $\lambda_0 > 0$,

$$e^{\lambda_0 \ell_k(\xi_s)^2} \le e^{2\lambda_0(1+R_\mu)^2} e^{2\lambda_0(1+R_\lambda)\|G_{\dim W_k}\|^2}.$$

For a sufficiently small constant $\lambda_0 > 0$ such that $4\lambda_0(1+R_\lambda) \le 1/2$, the *posterior exponential moment*

$$\mathcal{A}_s(\xi) := \sum_k \gamma_k(\xi, s) e^{\lambda_0 \ell_k(\xi)^2}$$

satisfies

$$\begin{aligned}\mathbb{E}\left[\mathcal{A}_s(\xi_s)\right] &= \sum_k \pi_k \mathbb{E}\left[e^{\lambda_0\ell_k(\xi_s)^2} \mid k\right] \\ &\le e^{2\lambda_0(1+R_\mu)^2}\sum_k \pi_k \mathbb{E}\left[e^{2\lambda_0(1+R_\lambda)\|G_{\dim W_k}\|^2}\right] \\ &= e^{2\lambda_0(1+R_\mu)^2}\sum_k \pi_k \big(1 - 4\lambda_0(1+R_\lambda)\big)^{-\dim W_k/2} \\ &\le e^{2\lambda_0(1+R_\mu)^2} 2^{(1+d^*)/2} \le e^{C(1+d^*)}.\end{aligned}$$

Here the second equality uses the Gaussian moment identity $\mathbb{E}\left[e^{u\|G_j\|^2}\right] = (1-2u)^{-j/2}$ for $0 \le u < 1/2$. For every $u \ge 0$, Markov's inequality gives

$$\begin{aligned}\Pr\big(\log \mathcal{A}_s(\xi_s) > C_0(1+d^*) + u\big) &= \Pr\big(\mathcal{A}_s(\xi_s) > e^{C_0(1+d^*)+u}\big) \\ &\le e^{-C_0(1+d^*)-u}\mathbb{E}\left[\mathcal{A}_s(\xi_s)\right] \le e^{-u}.\end{aligned}$$

Writing $x_+ = \max\{x, 0\}$, the tail integral formula therefore yields, for every real $p \ge 1$,

$$\begin{aligned}\mathbb{E}\left[\big(\log \mathcal{A}_s(\xi_s) - C_0(1+d^*)\big)_+^p\right] &= p\int_0^\infty u^{p-1}\Pr\big(\log \mathcal{A}_s(\xi_s) > C_0(1+d^*) + u\big)\,\mathrm{d}u \\ &\le p\int_0^\infty u^{p-1}e^{-u}\,\mathrm{d}u \le (2p)^p.\end{aligned}$$

For the last bound, integration by parts and maximization of $u^p e^{-u/2}$ give

$$\begin{aligned}p\int_0^\infty u^{p-1}e^{-u}\,\mathrm{d}u &= \int_0^\infty u^p e^{-u}\,\mathrm{d}u \\ &\le \left(\sup_{u\ge 0} u^p e^{-u/2}\right)\int_0^\infty e^{-u/2}\,\mathrm{d}u = 2(2p/e)^p \le (2p)^p.\end{aligned}$$

Since $\mathcal{A}_s(\xi_s) \ge 1$, its logarithm is nonnegative. Using $\log \mathcal{A}_s(\xi_s) \le C_0(1+d^*) + (\log \mathcal{A}_s(\xi_s) - C_0(1+d^*))_+$ and the L^p triangle inequality, we conclude

$$\begin{aligned}\|1 + \log \mathcal{A}_s(\xi_s)\|_{L^p} &\le 1 + C_0(1+d^*) + \|(\log \mathcal{A}_s(\xi_s) - C_0(1+d^*))_+\|_{L^p} \\ &\le 1 + C_0(1+d^*) + 2p \le C(1+d^*+p), \qquad p \ge 1.\end{aligned} \tag{15}$$

The constant C is independent of p.

We next control how the posterior changes between two inputs. Let Π be the joint distribution of the component label and ξ_0, and define

$$F_s(\xi) = \log \int \exp\left(\frac{1-s}{s^2}\langle \xi, x\rangle - \frac{(1-s)^2}{2s^2}\|x\|^2\right)\,\mathrm{d}\Pi(k, x).$$

The associated tilted measure is the posterior Π_s^ξ of (k, ξ_0) given $\xi_s = \xi$. If $m_s(\xi) = \mathbb{E}\left[\xi_0 \mid \xi_s = \xi\right]$, differentiation under the integral and $v = (\xi - m_s)/s$ give

$$\nabla F_s = \frac{1-s}{s^2}m_s, \qquad 0 \preceq \nabla^2 F_s = c(s)(s^{-1}I_d - D_\xi v) \preceq H I_d, \quad H = \frac{1-\epsilon}{\epsilon}(\epsilon^{-1} + L).$$

For two inputs $\xi, \widetilde{\xi}$, put $e = \widetilde{\xi} - \xi$. The exponential-family density ratio therefore satisfies

$$\int \left(\frac{\mathrm{d}\Pi_s^{\widetilde{\xi}}}{\mathrm{d}\Pi_s^{\xi}}\right)^2 \mathrm{d}\Pi_s^{\xi} = e^{F_s(\xi+2e) - 2F_s(\xi+e) + F_s(\xi)} \le e^{H\|e\|^2}.$$

Taking conditional expectation given the label and applying Jensen's inequality yields

$$\sum_{k:\pi_k>0} \frac{\gamma_k(\widetilde{\xi}, s)^2}{\gamma_k(\xi, s)} \le e^{H\|\widetilde{\xi}-\xi\|^2}. \tag{16}$$

All densities here are relative to Π, so singular component covariances cause no difficulty. At $s = 1$, the label posterior equals the prior and the same inequality holds.

Since $\ell_k(\widetilde{\xi}) \le \ell_k(\xi) + \|e\|$, Cauchy–Schwarz and (16), with $\lambda = \lambda_0/4$, give

$$\begin{aligned}\sum_k \gamma_k(\widetilde{\xi}, s)e^{\lambda \ell_k(\widetilde{\xi})^2} &\le e^{2\lambda\|e\|^2}\left(\sum_{k:\pi_k>0}\frac{\gamma_k(\widetilde{\xi}, s)^2}{\gamma_k(\xi, s)}\right)^{1/2}\mathcal{A}_s(\xi)^{1/2} \\ &\le e^{(2\lambda + H/2)\|e\|^2}\mathcal{A}_s(\xi)^{1/2}.\end{aligned}$$

Applying the exponential Markov inequality under the posterior label distribution and integrating its tail now gives

$$M_a(\widetilde{\xi}, s) \le C_a\big(1 + \log \mathcal{A}_s(\xi) + \|\widetilde{\xi} - \xi\|^2\big)^{a/2}.$$

Substitute the coupled random input ξ_s for ξ, take the L^p norm, and apply (15) with moment order $ap/2$. Since a is fixed, this proves (14). □

The local defect estimates require first and second derivatives along a fixed direction. The following bound keeps the common drift separate from the component-dependent terms.

Lemma C.2 (Directional derivative bounds)**.** *Fix $w \in \mathbb{R}^d$ during differentiation, and put*

$$\mathcal{L}_w = \partial_s + w \cdot \nabla_\xi, \qquad \mathcal{R} = \sqrt{1+d^*} + M_8(\xi, s)^{1/8} + \|w - \xi/s\|.$$

For every admissible mixture velocity,

$$\|\mathcal{L}_w v(\xi, s)\| \le C\mathcal{R}^3, \qquad \|\mathcal{L}_w^2 v(\xi, s)\| \le C\chi(s)\mathcal{R}^4, \quad \chi(s) = 1 + (1-s)^{-1/2}. \tag{17}$$

The first inequality holds for $s \in [\epsilon, 1]$ and the second for $s \in [\epsilon, 1)$. The function χ is integrable on $[\epsilon, 1]$.

Proof. Zero-weight components may be omitted. After removing the common isotropic log density from every component, write $\gamma_k = \exp(\log \pi_k + g_k)/\sum_j \exp(\log \pi_j + g_j)$ for positive weights, where

$$\begin{aligned} g_k(\xi, s) = & -\frac{1}{2}\sum_{j=1}^{d^*} \log\left(1 + \frac{(1-s)^2}{s^2}\lambda_{k,j}\right) - \frac{1}{2}\xi^\top (C_{k,s}^{-1} - s^{-2} I_d)\xi \\ & + (1-s)\mu_k^\top C_{k,s}^{-1}\xi - \frac{1}{2}(1-s)^2 \mu_k^\top C_{k,s}^{-1}\mu_k. \end{aligned} \tag{18}$$

Here $\lambda_{k,j}$ is the j-th diagonal entry of Λ_k. We first bound the component derivatives. For $a_\lambda(s) = s^2 + (1-s)^2\lambda$, the spectral decomposition gives

$$C_{k,s}^{-1} = s^{-2}(I_d - U_k U_k^\top) + U_k \mathrm{diag}\big(a_{\lambda_{k,j}}(s)^{-1}\big) U_k^\top.$$

Since $a_\lambda \ge \epsilon^2$, $a_\lambda' = 2s - 2(1-s)\lambda$, and $a_\lambda'' = 2(1+\lambda)$, the identities

$$(a_\lambda^{-1})' = -\frac{a_\lambda'}{a_\lambda^2}, \qquad (a_\lambda^{-1})'' = \frac{2(a_\lambda')^2}{a_\lambda^3} - \frac{a_\lambda''}{a_\lambda^2}$$

bound the first two time derivatives uniformly for $0 \le \lambda \le R_\lambda$. Consequently, if $\Gamma_{k,s} = C_{k,s}^{-1} - s^{-2} I_d$, then

$$\|\partial_s^j C_{k,s}^{-1}\|_{\mathrm{op}} + \|\partial_s^j \Gamma_{k,s}\|_{\mathrm{op}} + \|\partial_s^j (\Sigma_k C_{k,s}^{-1})\|_{\mathrm{op}} \le C, \qquad 0 \le j \le 2.$$

The last two matrices are supported on $\mathrm{span}(U_k)$, and $(\partial_s^j C_{k,s}^{-1})\mu_k$ lies in W_k. In particular, differentiating the component residual explicitly gives

$$\partial_s^j \beta_k = -\sum_{\ell=0}^{j} \binom{j}{\ell} \partial_s^{j-\ell}\left(\frac{1-s}{s}\right) \partial_s^\ell (\Sigma_k C_{k,s}^{-1})\xi - \partial_s^j (s C_{k,s}^{-1})\mu_k.$$

Its norm is bounded by $C(1 + \|U_k^\top \xi\|) \le C\ell_k$.

For the log-determinant term in (18), use

$$\partial_s \log \frac{a_\lambda}{s^2} = \frac{a_\lambda'}{a_\lambda} - \frac{2}{s}, \qquad \partial_s^2 \log \frac{a_\lambda}{s^2} = \frac{a_\lambda''}{a_\lambda} - \frac{(a_\lambda')^2}{a_\lambda^2} + \frac{2}{s^2}.$$

Each summand and its derivatives are uniformly bounded, so summing costs only a factor d^*. The spatial derivatives are

$$\nabla_\xi \partial_s^j g_k = -(\partial_s^j \Gamma_{k,s})\xi + \partial_s^j\big((1-s)C_{k,s}^{-1}\mu_k\big), \qquad D_\xi^2 \partial_s^j g_k = -\partial_s^j \Gamma_{k,s}.$$

The quadratic and linear terms thus depend only on $R_k^\top \xi$. Together, these calculations yield

$$\begin{aligned} &\|\partial_s^j \beta_k\| \le C\ell_k, \qquad \|D_\xi \partial_s^j \beta_k\|_{\mathrm{op}} \le C && (0 \le j \le 2), \\ &|\partial_s^j g_k| \le C(1 + d^* + \ell_k^2) && (j = 1, 2), \\ &\|\nabla_\xi \partial_s^j g_k\| \le C\ell_k, \qquad \|D_\xi^2 \partial_s^j g_k\|_{\mathrm{op}} \le C && (0 \le j \le 2). \end{aligned}$$

Put $b = w - \xi/s$, so that $\|R_k^\top w\| \le \epsilon^{-1}\ell_k + \|b\|$. Because w is frozen and β_k is affine in ξ,

$$\mathcal{L}_w\beta_k = \partial_s\beta_k + D_\xi\beta_k\, w, \qquad \mathcal{L}_w^2\beta_k = \partial_s^2\beta_k + 2D_\xi\partial_s\beta_k\, w,$$
$$\dot g_k := \mathcal{L}_w g_k = \partial_s g_k + \nabla_\xi g_k \cdot w, \qquad \ddot g_k := \mathcal{L}_w^2 g_k = \partial_s^2 g_k + 2\nabla_\xi\partial_s g_k \cdot w + w^\top D_\xi^2 g_k\, w.$$

It follows that

$$\|\mathcal{L}_w^j\beta_k\| \le C(\ell_k + \|b\|) \quad (0 \le j \le 2), \qquad |\dot g_k| + |\ddot g_k| \le C\big(1 + d^* + (\ell_k + \|b\|)^2\big).$$

Write $\langle a_k\rangle_\gamma = \sum_k \gamma_k a_k$ and $\|a\|_{\gamma,p} = \langle \|a_k\|^p\rangle_\gamma^{1/p}$. Since $\|\ell\|_{\gamma,8} = M_8^{1/8}$, these bounds give

$$\|\dot g\|_{\gamma,4} + \|\ddot g\|_{\gamma,2} \le C\big(1 + d^* + M_8^{1/4} + \|b\|^2\big) \le C\mathcal{R}^2.$$

For $\widetilde g_k = \dot g_k - \langle \dot g\rangle_\gamma$, $\|\widetilde g\|_{\gamma,4} \le 2\|\dot g\|_{\gamma,4}$, and

$$\|\beta_k\|_{\gamma,2} + \|\mathcal{L}_w\beta_k\|_{\gamma,2} + \|\mathcal{L}_w^2\beta_k\|_{\gamma,1} \le C\mathcal{R}.$$

There is no factor K because these are averages under a probability distribution. Direct differentiation of the softmax gives

$$\mathcal{L}_w\gamma_k = \gamma_k\widetilde g_k, \qquad \mathcal{L}_w^2\gamma_k = \gamma_k\big(\ddot g_k - \langle \ddot g\rangle_\gamma + \widetilde g_k^2 - \langle \widetilde g^2\rangle_\gamma\big).$$

Applying the product rule to $\beta = \sum_k \gamma_k\beta_k$ now yields

$$\mathcal{L}_w\beta = \langle \mathcal{L}_w\beta_k\rangle_\gamma + \mathrm{Cov}_\gamma(\beta_k, \dot g_k),$$
$$\mathcal{L}_w^2\beta = \langle \mathcal{L}_w^2\beta_k\rangle_\gamma + 2\langle (\mathcal{L}_w\beta_k)\widetilde g_k\rangle_\gamma + \mathrm{Cov}_\gamma(\beta_k, \ddot g_k + \widetilde g_k^2).$$

Cauchy–Schwarz bounds the first covariance by $2\|\beta_k\|_{\gamma,2}\|\dot g\|_{\gamma,2} \le C\mathcal{R}^3$. For the second covariance, let f_k be any scalar component quantity. For every unit vector u,

$$|u^\top \mathrm{Cov}_\gamma(\beta_k, f_k)| \le \big(u^\top \mathrm{Cov}_\gamma(\beta_k)u\big)^{1/2}\,\mathrm{Var}_\gamma(f_k)^{1/2}.$$

Taking the supremum over u and applying (11) gives, for $s < 1$,

$$\|\mathrm{Cov}_\gamma(\beta_k, f_k)\| \le \sqrt{\frac{L + \epsilon^{-1}}{c(s)}}\,\|f\|_{\gamma,2} \le C\chi(s)\|f\|_{\gamma,2}.$$

This uses the covariance operator norm without summing over coordinates. In particular, $\|\ddot g + \widetilde g^2\|_{\gamma,2} \le \|\ddot g\|_{\gamma,2} + \|\widetilde g\|_{\gamma,4}^2 \le C\mathcal{R}^4$, so

$$\|\mathcal{L}_w\beta\| \le C(\mathcal{R} + \mathcal{R}^3) \le C\mathcal{R}^3, \qquad \|\mathcal{L}_w^2\beta\| \le C(\mathcal{R} + \mathcal{R}^3 + \chi(s)\mathcal{R}^4) \le C\chi(s)\mathcal{R}^4.$$

Finally, $\mathcal{L}_w(\xi/s) = b/s$ and $\mathcal{L}_w^2(\xi/s) = -2b/s^2$. Combining these identities proves (17). The actual derivatives are smooth at $s = 1$; only the upper envelope χ is singular there. □

C.4 Proof of Theorem 6.5

Fix $(r, t) \in \mathcal{T}_\epsilon$, set $\Delta = (t - r)/m$, $\tau_i = r + i\Delta$, and write Ψ_i for the states of the oracle network:

$$\Psi_m = \xi, \qquad \Psi_{i-1} = \Psi_i - \Delta v(\Psi_i, \tau_i), \quad i = m, \dots, 1.$$

Let $\bar\xi_s$ be the true characteristic with terminal input $\bar\xi_t = \xi \sim q_t$. Then $\bar\xi_s \sim q_s$, since q_s solves the continuity equation for the globally Lipschitz velocity v. For $0 \le u \le \Delta$, define $\Psi_{i,u} = \Psi_i - uv(\Psi_i, \tau_i)$. All L^p norms in this subsection are over $\xi \sim q_t$.

Lemma C.3 (Moments and local defects of the oracle Euler network)**.** *For fixed $a \ge 1$ and $p \ge 2$, uniformly in i, m, r, t, u,*

$$\|M_a(\Psi_i, \tau_i)\|_{L^p} + \|M_a(\Psi_{i,u}, \tau_i - u)\|_{L^p} \le C_{a,p}(1 + d^*)^{a/2}. \tag{19}$$

Let

$$v_i = v(\Psi_i, \tau_i), \quad a_i = (\partial_s v + D_\xi v\, v)(\Psi_i, \tau_i), \quad T_i = v_i - \Delta a_i - v(\Psi_{i-1}, \tau_{i-1}).$$

Then

$$\|a_i\|_{L^2} \le C(1 + d^*)^{3/2}, \qquad \|T_i\|_{L^2} \le C\Delta(1 + d^*)^2 \int_{\tau_{i-1}}^{\tau_i} \chi(s)\,\mathrm{d}s. \tag{20}$$

Proof. Step 1: moments of the numerical states. The conditional-mean identity $\beta(\xi, s) = -s^{-1}\mathbb{E}[\xi_0 \mid \xi_s = \xi]$ and (13) imply $\|\beta(\bar{\xi}_s, s)\|_{L^\ell} \le C_\ell\sqrt{1+d^*}$ for each fixed $\ell \ge 2$. Integrating the common drift exactly gives

$$\bar{\xi}_a = \frac{a}{b}\bar{\xi}_b - a\int_a^b \frac{\beta(\bar{\xi}_s, s)}{s}\,\mathrm{d}s.$$

The Euler step from $\bar{\xi}_b$ is $(a/b)\bar{\xi}_b - (b-a)\beta(\bar{\xi}_b, b)$. Their L^ℓ difference is at most $C_\ell(b-a)\sqrt{1+d^*}$. Thus (12) gives

$$\|\Psi_{i-1} - \bar{\xi}_{\tau_{i-1}}\|_{L^\ell} \le (1+\Delta L)\|\Psi_i - \bar{\xi}_{\tau_i}\|_{L^\ell} + C_\ell\Delta\sqrt{1+d^*}.$$

The terminal discrepancy is zero. Iterating this inequality, and then applying the same comparison over a partial step, yields

$$\|\Psi_i - \bar{\xi}_{\tau_i}\|_{L^\ell} + \|\Psi_{i,u} - \bar{\xi}_{\tau_i - u}\|_{L^\ell} \le C_\ell\sqrt{1+d^*}.$$

Lemma C.1 now proves (19). In particular, $\|\beta(\Psi_i, \tau_i)\|_{L^\ell} \le C_\ell\sqrt{1+d^*}$, since $\|\beta\| \le CM_1$.

Step 2: local transport defects. Apply Lemma C.2 at (Ψ_i, τ_i) with frozen direction $w = v_i$. Here $w - \Psi_i/\tau_i = \beta(\Psi_i, \tau_i)$, and the preceding moment bounds give $\mathbb{E}\left[\mathcal{R}^8\right] \le C(1+d^*)^4$. The first inequality in (17) therefore implies $\|a_i\|_{L^2} \le C(1+d^*)^{3/2}$.

For the remainder, set $F_i(u) = v(\Psi_i - uv_i, \tau_i - u)$. Along this segment, with $s = \tau_i - u$,

$$v_i - \frac{\Psi_i - uv_i}{s} = \frac{\tau_i}{s}\beta(\Psi_i, \tau_i).$$

Consequently, (19) gives the same bound on $\mathbb{E}\left[\mathcal{R}^8\right]$ along the segment. The second derivative estimate yields $\|F_i''(u)\|_{L^2} \le C(1+d^*)^2\chi(\tau_i - u)$. Since $F_i'(0) = -a_i$, Taylor's formula with integral remainder gives

$$T_i = -\int_0^\Delta (\Delta - u)F_i''(u)\,\mathrm{d}u.$$

Minkowski's inequality proves the second bound in (20). When $\tau_i = 1$, the integral remains finite because χ is integrable. □

Proof of Theorem 6.5. Step 1: the oracle and its differentiated recursion. By assumption, $\theta_{\mathrm{or}} = (\theta^*, \dots, \theta^*) \in \Theta_m$, so it suffices to bound $\mathcal{L}_{\mathrm{tr}}(\theta_{\mathrm{or}})$. Fix r and differentiate the oracle states with respect to t along the true characteristic at the terminal input. Put

$$\omega_i = \frac{i}{m}, \quad Q_i = \partial_t\Psi_i + D_\xi\Psi_i\, v(\xi, t), \quad E_i = Q_i - \omega_i v(\Psi_i, \tau_i), \quad G_i = I_d - \Delta D_\xi v(\Psi_i, \tau_i).$$

Then $E_m = 0$ and $E_0 = \mathcal{D}\Phi_{\theta_{\mathrm{or}}}$. Since $\partial_t\Delta = 1/m$ and $\partial_t\tau_i = \omega_i$, direct differentiation of the Euler update gives

$$Q_{i-1} = G_iQ_i - \frac{1}{m}v_i - \Delta\omega_i\partial_s v(\Psi_i, \tau_i).$$

Subtracting $\omega_{i-1}v_{i-1}$ and using the definition of T_i yields

$$\begin{aligned} E_{i-1} &= G_iE_i + \omega_{i-1}(v_i - v_{i-1}) - \Delta\omega_i a_i \\ &= G_iE_i - \frac{\Delta}{m}a_i + \omega_{i-1}T_i. \end{aligned} \tag{21}$$

Step 2: accumulation of the local defects. All products of G_i have norm at most e^L by (12). Unrolling (21) and applying Lemma C.3 gives

$$\begin{aligned} \|\mathcal{D}\Phi_{\theta_{\mathrm{or}}}\|_{L^2(q_t)} &\le e^L\left(\frac{\Delta}{m}\sum_{i=1}^m \|a_i\|_{L^2} + \sum_{i=1}^m \|T_i\|_{L^2}\right) \\ &\le C\Delta\left((1+d^*)^{3/2} + (1+d^*)^2\int_r^t \chi(s)\,\mathrm{d}s\right) \le \frac{C(1+d^*)^2}{m}. \end{aligned}$$

At $r = t$, direct differentiation gives $\partial_t\Phi_{\theta_{\mathrm{or}}} = -v$ and $D_\xi\Phi_{\theta_{\mathrm{or}}} = I_d$, so the transport residual is zero. Squaring the last bound and integrating over $(r, t) \sim \nu$ proves

$$A_m \le \mathcal{L}_{\mathrm{tr}}(\theta_{\mathrm{or}}) \le C_{\mathrm{app}}\frac{(1+d^*)^4}{m^2}.$$

□

C.5 Proof of Theorem 6.6

For $\theta \in \Theta_m$, define $f_\theta(\xi, r, t) = \|\mathcal{D}\Phi_\theta(\xi, r, t)\|^2$ and let $\mathcal{F}_m = \{f_\theta : \theta \in \Theta_m\}$. We first bound the moments of each loss using the component rank, then cover the loss class on a bounded input set.

Lemma C.4 (Intrinsic moments of the transport loss)**.** *For every $p \geq 2$,*

$$\sup_{\theta \in \Theta_m} \|f_\theta\|_{L^p(P_\nu)} \leq C(1 + d^* + p)^3. \tag{22}$$

The constant is independent of the moment order p.

Proof. Fix θ, r, t and realize the terminal input as $\xi = (1-t)\xi_0 + t\eta$, where $\xi_0 \sim q$ and $\eta \sim \mathcal{N}(0, I_d)$ are independent. Write Ψ_i^θ for the network states and $\beta_{\theta_i}(y, s) = v_{\theta_i}(y, s) - y/s$. Norms in this proof are over (ξ_0, η) and any independent layer-prior variables introduced in the couplings.

Step 1: moments along the network. For each layer, take a sample $\xi_0^{(i)}$ from that layer's Gaussian mixture, independent of η, and set $\xi_s^{(i)} = (1-s)\xi_0^{(i)} + s\eta$. The conditional-mean identity and Jensen's inequality give

$$\|\beta_{\theta_i}(\xi_s^{(i)}, s)\|_{L^p} \leq \epsilon^{-1}\|\xi_0^{(i)}\|_{L^p} \leq C\sqrt{1 + d^* + p}.$$

Using the uniform Lipschitz bound for β_{θ_i} and the same coupling, we obtain

$$\|\beta_{\theta_i}(s\eta, s)\|_{L^p} \leq \|\beta_{\theta_i}(\xi_s^{(i)}, s)\|_{L^p} + C(1-s)\|\xi_0^{(i)}\|_{L^p} \leq C\sqrt{1 + d^* + p}. \tag{23}$$

Compare the network states with the common noise path by putting $e_i = \Psi_i^\theta - \tau_i\eta$. Euler stability and (23) imply

$$\|e_{i-1}\|_{L^p} \leq (1 + \Delta L)\|e_i\|_{L^p} + C\Delta\sqrt{1 + d^* + p}, \qquad e_m = (1-t)\xi_0.$$

Since $m\Delta \leq 1$, iteration and (13) give

$$\max_{0 \leq i \leq m} \|\Psi_i^\theta - \tau_i\eta\|_{L^p} \leq C\sqrt{1 + d^* + p}. \tag{24}$$

This is a maximum of individual norms, not a moment of the maximum.

Let M_{a,θ_i} denote the posterior moment for layer i. Couple Ψ_i^θ to $\xi_{\tau_i}^{(i)} = (1-\tau_i)\xi_0^{(i)} + \tau_i\eta$. Their L^{8p} distance is at most $C\sqrt{1 + d^* + p}$ by (24) and (13). Lemma C.1 therefore yields

$$\|M_{8,\theta_i}(\Psi_i^\theta, \tau_i)\|_{L^p} \leq C(1 + d^* + p)^4.$$

Define

$$\mathcal{B}_{\theta_i}(y, s) = \partial_s\beta_{\theta_i}(y, s) + D_y\beta_{\theta_i}(y, s)\frac{y}{s}.$$

At the point (y, s), take the frozen direction $w = y/s$ in Lemma C.2. The derivative of the common drift vanishes at that point, so

$$\|\mathcal{B}_{\theta_i}(y, s)\| \leq C\big(\sqrt{1 + d^*} + M_{8,\theta_i}(y, s)^{1/8}\big)^3.$$

Together with the Lipschitz bound for β_{θ_i}, this proves

$$\begin{aligned} \|\beta_{\theta_i}(\Psi_i^\theta, \tau_i)\|_{L^p} &\leq C(1 + d^* + p)^{1/2}, \\ \|\mathcal{B}_{\theta_i}(\Psi_i^\theta, \tau_i)\|_{L^p} &\leq C(1 + d^* + p)^{3/2}. \end{aligned} \tag{25}$$

Step 2: cancellation of the common drift. Before substituting the random terminal input, define

$$Q_i = \partial_t\Psi_i^\theta + D_\xi\Psi_i^\theta\, v(\xi, t), \qquad \widetilde{E}_i = Q_i - \omega_i\frac{\Psi_i^\theta}{\tau_i}, \qquad G_i^\theta = I_d - \Delta D_y v_{\theta_i}(\Psi_i^\theta, \tau_i), \quad \omega_i = i/m.$$

Then $\widetilde{E}_m = \beta(\xi, t)$ and $\widetilde{E}_0 = \mathcal{D}\Phi_\theta$. Differentiate the residual form of the Euler update,

$$\Psi_{i-1}^\theta = \frac{\tau_{i-1}}{\tau_i}\Psi_i^\theta - \Delta\beta_{\theta_i}(\Psi_i^\theta, \tau_i).$$

For this calculation, abbreviate $\beta_i = \beta_{\theta_i}(\Psi_i^\theta, \tau_i)$, with its derivatives evaluated at the same point. Since $\tau_{i-1}\omega_i - \tau_i\omega_{i-1} = r/m$, we have

$$\partial_t \frac{\tau_{i-1}}{\tau_i} = -\frac{r}{m\tau_i^2}, \qquad G_i^\theta = \frac{\tau_{i-1}}{\tau_i} I_d - \Delta D_y \beta_i.$$

Applying $\partial_t + D_\xi(\,\cdot\,)v(\xi, t)$ to the update gives

$$Q_{i-1} = G_i^\theta Q_i - \frac{r}{m\tau_i^2}\Psi_i^\theta - \frac{1}{m}\beta_i - \Delta\omega_i \partial_s \beta_i.$$

Substitute $Q_i = \widetilde{E}_i + \omega_i \Psi_i^\theta/\tau_i$ and subtract $\omega_{i-1}\Psi_{i-1}^\theta/\tau_{i-1}$. The coefficient of Ψ_i^θ simplifies according to

$$\frac{\omega_i}{\tau_i} G_i^\theta - \frac{r}{m\tau_i^2} I_d - \frac{\omega_{i-1}}{\tau_i} I_d = -\Delta \frac{\omega_i}{\tau_i} D_y \beta_i.$$

The resulting recursion is

$$\widetilde{E}_{i-1} = G_i^\theta \widetilde{E}_i - \Delta\omega_i \left(\partial_s \beta_i + D_y \beta_i \frac{\Psi_i^\theta}{\tau_i} \right) + \left(-\frac{1}{m} + \frac{\Delta\omega_{i-1}}{\tau_{i-1}} \right) \beta_i.$$

Finally, $\tau_{i-1} = r + m\Delta\omega_{i-1}$ implies $-m^{-1} + \Delta\omega_{i-1}/\tau_{i-1} = -r/(m\tau_{i-1})$. Recognizing $\mathcal{B}_{\theta_i}$ yields

$$\widetilde{E}_{i-1} = G_i^\theta \widetilde{E}_i - \frac{r}{m\tau_{i-1}} \beta_{\theta_i}(\Psi_i^\theta, \tau_i) - \Delta\omega_i \mathcal{B}_{\theta_i}(\Psi_i^\theta, \tau_i). \tag{26}$$

This differentiation follows the true velocity at the terminal input; it is not differentiation along the straight interpolation $(1-t)\xi_0 + t\eta$.

Step 3: propagation of the transport residual. The true conditional-mean identity gives $\|\widetilde{E}_m\|_{L^p} \le C\sqrt{1 + d^* + p}$. Products of G_i^θ have norm at most e^L, and

$$\sum_{i=1}^m \frac{r}{m\tau_{i-1}} \le 1, \qquad \sum_{i=1}^m \Delta\omega_i \le 1.$$

Unrolling (26) and applying (25) gives

$$\|\mathcal{D}\Phi_\theta(\xi, r, t)\|_{L^p} \le C(1 + d^* + p)^{3/2}.$$

The recursions also hold at $r = t$ by smooth extension. Apply the last bound with moment order $2p$, square, and average over $(r, t) \sim \nu$ to obtain (22). □

The supremum in (22) is outside the L^p norm. We will therefore use a uniform comparison in one direction and a comparison for one fixed population minimizer in the other direction. First, we construct a finite cover of the loss class.

Lemma C.5 (Envelope and covering numbers)**.** *For every $\theta \in \Theta_m$,*

$$f_\theta(\xi, r, t) \le C\big(1 + d^* + (1 + \|\xi\|)^2\big)^3. \tag{27}$$

For $R \ge 1$, let $B_R = 1 + d^* + (1+R)^2$ and $\|f\|_{\infty,R} = \sup_{\|\xi\| \le R,\, (r,t) \in \mathcal{T}_\epsilon} |f(\xi, r, t)|$. Then, for $0 < \zeta \le 1$,

$$\log \mathcal{N}(\zeta, \mathcal{F}_m, \|\cdot\|_{\infty,R}) \le C P_m \log(C P_m K B_R/\zeta), \qquad P_m = mK(d+1)(d^*+1). \tag{28}$$

The covering centers can be chosen in $\mathcal{F}_m$.

Proof. Step 1: the envelope. The component formula gives $\|v_{\theta_i}(y, s)\| \le C(1 + \|y\|)$, so the Euler recursion implies $1 + \|\Psi_i^\theta\| \le C(1+R)$ whenever $\|\xi\| \le R$. Let $H_i = \partial_t \Psi_i^\theta$ with the terminal input held fixed. Differentiating the Euler update gives

$$H_{i-1} = G_i^\theta H_i - \frac{1}{m} v_{\theta_i}(\Psi_i^\theta, \tau_i) - \Delta\omega_i \partial_s v_{\theta_i}(\Psi_i^\theta, \tau_i), \qquad H_m = 0. \tag{29}$$

The component and posterior formulas imply

$$\|\partial_s v_{\theta_i}(y, s)\| \le C(1 + \|y\|)\big(1 + d^* + (1 + \|y\|)^2\big).$$

Since the forcing coefficients in (29) have bounded sums and the Jacobian products have norm at most e^L, we obtain $\|H_i\| + \|\mathcal{D}\Phi_\theta\| \le C(1+R)B_R$. Squaring gives (27).

Step 2: covering one layer, including small weights. All states lie in a ball of radius $\bar{R} = C(1+R)$. For a layer, consider the three functions v_{θ_i}, $D_y v_{\theta_i}$, and $\partial_s v_{\theta_i}$ on this ball and $s \in [\epsilon, 1]$, using uniform vector and operator norms. Write its component logit as $\log \pi_{k,i} + g_{k,i}$. Equation (18) gives

$$|g_{k,i}| + |\partial_s g_{k,i}| \le CB_R, \qquad \|\nabla_y g_{k,i}\| \le C\sqrt{B_R}.$$

We make the dependence on the geometric parameters explicit. Suppress the layer and component indices, and compare two components whose parameters satisfy $\|\mu - \widetilde{\mu}\| + \|U - \widetilde{U}\|_F + \|\lambda - \widetilde{\lambda}\| \le \alpha$, where λ is the vector of diagonal entries of Λ. Expanding $U\Lambda U^\top - \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top$ gives

$$\|\Sigma - \widetilde{\Sigma}\|_{\mathrm{op}} \le 2R_\lambda \|U - \widetilde{U}\|_F + \|\lambda - \widetilde{\lambda}\| \le C\alpha.$$

For $C_s = s^2 I_d + (1-s)^2 \Sigma$, the identities

$$C_s^{-1} - \widetilde{C}_s^{-1} = C_s^{-1}(\widetilde{C}_s - C_s)\widetilde{C}_s^{-1}, \qquad \partial_s C_s^{-1} = -C_s^{-1}\big(2sI_d - 2(1-s)\Sigma\big)C_s^{-1}$$

and $C_s, \widetilde{C}_s \succeq \epsilon^2 I_d$ imply

$$\|C_s^{-1} - \widetilde{C}_s^{-1}\|_{\mathrm{op}} + \|\partial_s C_s^{-1} - \partial_s \widetilde{C}_s^{-1}\|_{\mathrm{op}} \le C\alpha.$$

The time-derivative bound follows by subtracting the three-factor products and changing one factor at a time. For the determinant term, the mean value theorem gives

$$\left| \sum_{j=1}^{d^*} \log \frac{a_{\lambda_j}(s)}{s^2} - \sum_{j=1}^{d^*} \log \frac{a_{\widetilde{\lambda}_j}(s)}{s^2} \right| \le C \sum_{j=1}^{d^*} |\lambda_j - \widetilde{\lambda}_j| \le C\sqrt{d^*}\,\alpha.$$

The same estimate holds after one time derivative, since $\partial_\lambda \partial_s \log(a_\lambda(s)/s^2)$ is uniformly bounded. The quadratic and linear terms in (18) cost at most $\bar{R}^2$ and $\bar{R}$, respectively. Substituting into the component formulas therefore bounds the differences of each of $v_k, D_y v_k, \partial_s v_k, g_k, \nabla_y g_k, \partial_s g_k$ by $CB_R^3\alpha$. This is a common, nonoptimal bound for these six quantities.

To pass from components to the mixture, write $\langle a_k \rangle_\gamma = \sum_k \gamma_k a_k$. Differentiating the posterior average gives

$$D_y v = \langle D_y v_k \rangle_\gamma + \langle v_k (\nabla_y g_k)^\top \rangle_\gamma - v \langle \nabla_y g_k \rangle_\gamma^\top,$$
$$\partial_s v = \langle \partial_s v_k \rangle_\gamma + \langle v_k \partial_s g_k \rangle_\gamma - v \langle \partial_s g_k \rangle_\gamma.$$

We use these formulas together with

$$\|\langle a_k \rangle_\gamma - \langle \widetilde{a}_k \rangle_{\widetilde{\gamma}}\| \le \max_k \|a_k - \widetilde{a}_k\| + \max_k \|\widetilde{a}_k\|\, \|\gamma - \widetilde{\gamma}\|_1.$$

The component bounds $\|v_k\| + \|\partial_s v_k\| \le C(1+\bar{R})$ and $\|D_y v_k\|_{\mathrm{op}} \le C$ ensure that all products above have polynomial bounds in B_R.

To handle arbitrarily small weights, encode each weight by $\alpha_{k,i} = \max\{\log \pi_{k,i}, -T\}$, with $\log 0 = -\infty$. Replacing $\pi_{k,i}$ by $\max\{\pi_{k,i}, e^{-T}\}$ changes the posterior weights in ℓ_1 norm by at most $2Ke^{-T+2CB_R}$: the original normalizer is at least e^{-CB_R} and its increase is at most Ke^{-T+CB_R}. The mixture derivative formulas above then bound the change in $v_{\theta_i}, D_y v_{\theta_i}, \partial_s v_{\theta_i}$ by $CB_R^6 K e^{-T+2CB_R}$. For a parameter resolution $0 < \alpha \le 1$, choose

$$T = CB_R + \log(CKB_R^6/\alpha),$$

with a sufficiently large first constant. Then flooring changes these three functions by at most α.

For a line segment of logits $h_k(u)$, its softmax weights satisfy

$$\partial_u \gamma_k = \gamma_k\big(\partial_u h_k - \langle \partial_u h \rangle_\gamma\big), \qquad \sum_k |\partial_u \gamma_k| \le 2 \max_k |\partial_u h_k|.$$

Thus the softmax is 2-Lipschitz from the maximum norm to the ℓ_1 norm. Two encoded layers whose geometric parameters and clipped log weights are within α consequently have posterior weights

within $CB_R^3\alpha$. Applying the weighted-average inequality to the displayed mixture formulas, and including both flooring errors, gives the common bound $CB_R^6\alpha$ for the original velocities and their first derivatives. Flooring is used only for this comparison; it need not preserve stability. We choose covering representatives from the original admissible parameter class.

Step 3: propagation and counting. Suppose two admissible networks have corresponding layer velocities and first derivatives within u on the state ball. With $e_i = \|\Psi_i^\theta - \Psi_i^{\theta'}\|$, Euler stability gives

$$e_{i-1} \le (1+\Delta L)e_i + \Delta u, \qquad e_m = 0, \qquad \max_i e_i \le e^L \sum_{i=1}^m \Delta u \le Cu.$$

We also need spatial Lipschitz bounds for the first derivatives. For commuting space or time derivatives $\mathcal{D}_1, \mathcal{D}_2$, differentiation of the softmax yields

$$\begin{aligned}\mathcal{D}_2\mathcal{D}_1\gamma_k = \gamma_k\big[&(\mathcal{D}_1 g_k - \langle \mathcal{D}_1 g\rangle_\gamma)(\mathcal{D}_2 g_k - \langle \mathcal{D}_2 g\rangle_\gamma) \\ &+ \mathcal{D}_2\mathcal{D}_1 g_k - \langle \mathcal{D}_2\mathcal{D}_1 g\rangle_\gamma - \mathrm{Cov}_\gamma(\mathcal{D}_1 g_k, \mathcal{D}_2 g_k)\big].\end{aligned}$$

For unit vectors a, b, let $\partial_a = a\cdot\nabla_y$ and $\partial_b = b\cdot\nabla_y$. The component derivative bounds in Lemma C.2 imply

$$\sum_k |\partial_a\gamma_k| \le C\sqrt{B_R}, \qquad \sum_k |\partial_a\partial_b\gamma_k| \le CB_R,$$
$$\sum_k |\partial_s\gamma_k| \le CB_R, \qquad \sum_k |\partial_a\partial_s\gamma_k| \le CB_R^{3/2}.$$

For example, the last bound uses $|\partial_s g_k| \le CB_R$, $|\partial_a g_k| + |\partial_a\partial_s g_k| \le C\sqrt{B_R}$ in the mixed derivative formula. Since v_k is affine in y,

$$D_y^2 v[a,b] = \sum_k \big((\partial_a\partial_b\gamma_k)v_k + (\partial_a\gamma_k)D_y v_k\, b + (\partial_b\gamma_k)D_y v_k\, a\big),$$
$$D_y\partial_s v[a] = \sum_k \big((\partial_a\partial_s\gamma_k)v_k + (\partial_s\gamma_k)D_y v_k\, a + (\partial_a\gamma_k)\partial_s v_k + \gamma_k D_y\partial_s v_k\, a\big).$$

Taking the supremum over unit directions therefore gives

$$\|D_y^2 v_{\theta_i}\|_{\mathrm{op}} \le CB_R^{3/2}, \qquad \|D_y\partial_s v_{\theta_i}\|_{\mathrm{op}} \le CB_R^2.$$

Let δv_i, $\delta(D_y v)_i$, and $\delta(\partial_s v)_i$ denote the differences between corresponding layer quantities evaluated at their respective states (Ψ_i^θ, τ_i) and $(\Psi_i^{\theta'}, \tau_i)$. The state bound and the preceding estimates give

$$\|\delta v_i\| \le Cu, \qquad \|\delta(D_y v)_i\|_{\mathrm{op}} + \|\delta(\partial_s v)_i\| \le CB_R^2 u, \qquad \|G_i^\theta - G_i^{\theta'}\|_{\mathrm{op}} \le C\Delta B_R^2 u.$$

Here the first inequality also uses the uniform Lipschitz bound on v_{θ_i} from (10). Write $J_i^\theta = D_\xi\Psi_i^\theta$ and $H_i^\theta = \partial_t\Psi_i^\theta$, and use $\delta J_i = J_i^\theta - J_i^{\theta'}$ and $\delta H_i = H_i^\theta - H_i^{\theta'}$. Subtracting the recursions gives

$$\delta J_{i-1} = G_i^\theta \delta J_i + (G_i^\theta - G_i^{\theta'})J_i^{\theta'}, \qquad \delta J_m = 0,$$
$$\delta H_{i-1} = G_i^\theta \delta H_i + (G_i^\theta - G_i^{\theta'})H_i^{\theta'} - \frac{1}{m}\delta v_i - \Delta\omega_i\delta(\partial_s v)_i, \qquad \delta H_m = 0.$$

Since $\|J_i^{\theta'}\|_{\mathrm{op}} \le e^L$ and $\|H_i^{\theta'}\| \le C(1+R)B_R$ by Step 1, unrolling these recursions and using the same Jacobian-product bound yields

$$\|\delta J_0\|_{\mathrm{op}} \le C\sum_{i=1}^m \Delta B_R^2 u \le CB_R^2 u,$$
$$\|\delta H_0\| \le C\sum_{i=1}^m \left(\Delta(1+R)B_R^3 + \frac{1}{m}\right) u \le C(1+R)B_R^3 u.$$

Both $\sum_i \Delta = t - r \le 1$ and $\sum_i m^{-1} = 1$, so no additional factor growing with depth appears. The true velocity satisfies $\|v(\xi,t)\| \le C(1+R)$ on the input ball. Hence

$$\|\mathcal{D}\Phi_\theta - \mathcal{D}\Phi_{\theta'}\| = \|\delta H_0 + \delta J_0 v(\xi,t)\| \le C(1+R)B_R^3 u.$$

Using the residual envelope from Step 1 to compare squared norms,

$$\|f_\theta - f_{\theta'}\|_{\infty,R} \le C(1+R)^2 B_R^4 u \le C B_R^5 u.$$

Combined with Step 2, this is at most $CB_R^{11}\alpha$ for two sufficiently close parameter codes.

Set $\alpha = \zeta/(CB_R^{11})$. The code for the entire network has $mK(dd^* + d + d^* + 1) = P_m$ scalar coordinates, each in an interval of length at most $C(1+T)$. Partition the resulting box into cubes of side $\alpha/\sqrt{P_m}$, and select one original admissible parameter from each nonempty cell. The selected losses form a ζ-cover of size at most

$$\left(1 + \frac{C\sqrt{P_m}(1+T)}{\alpha}\right)^{P_m}.$$

Taking logarithms and substituting α and T gives

$$\log \mathcal{N} \le P_m \log\left(1 + \frac{C\sqrt{P_m}(1+T)}{\alpha}\right) \le C P_m \log(C P_m K B_R/\zeta),$$

which proves (28). □

For the statistical comparison, write $Pf = \mathbb{E}_{P_\nu}[f]$ and $P_n f = n^{-1}\sum_{j=1}^n f(\xi_j, r_j, t_j)$. Recall $S_n = 1 + d^* + \log(8n/\delta)$, and put

$$r_{n,m} = \frac{S_n^3}{n}\left(P_m \log(CnP_m K S_n) + \log(8/\delta)\right), \tag{30}$$

where C is sufficiently large.

Lemma C.6 (Empirical–population comparisons)**.** *Fix a deterministic $\theta^\circ \in \Theta_m$. With probability at least $1-\delta$, simultaneously,*

$$\mathcal{L}_{\mathrm{tr}}(\theta) \le \hat{\mathcal{L}}_{\mathrm{tr},n}(\theta) + C\sqrt{\mathcal{L}_{\mathrm{tr}}(\theta)\, r_{n,m}} + C r_{n,m}, \quad \theta \in \Theta_m, \tag{31}$$

$$\hat{\mathcal{L}}_{\mathrm{tr},n}(\theta^\circ) \le \mathcal{L}_{\mathrm{tr}}(\theta^\circ) + C\sqrt{\mathcal{L}_{\mathrm{tr}}(\theta^\circ)\, r_{n,m}} + C r_{n,m}. \tag{32}$$

Proof. Step 1: truncation for the analysis. Let $u = 4\log(8n/\delta)$ and choose a sufficiently large fixed C_0. Define

$$B = C_0(1 + d^* + u)^3, \qquad R_n = 1 + R_\mu + \sqrt{1+R_\lambda}\big(\sqrt{d} + \sqrt{2u}\big),$$

and $g_\theta = (f_\theta \wedge B)\mathbf{1}_{\{\|\xi\| \le R_n\}}$. Lemma C.4, with moment order u, implies $\sup_\theta P(f_\theta > B) \le e^{-u}$. The Gaussian norm tail bound applied componentwise to q_t gives $P(\|\xi\| > R_n) \le e^{-u}$. Cauchy–Schwarz and the L^2 loss bound therefore yield, uniformly over θ,

$$0 \le Pf_\theta - Pg_\theta \le C(1+d^*)^3 e^{-u/2} \le C S_n^3/n^2. \tag{33}$$

Moreover, $0 \le g_\theta \le B \le C S_n^3$ and $\mathrm{Var}(g_\theta) \le B\, Pg_\theta$. The truncation changes only the quantities used in this proof.

Step 2: Bernstein's inequality on a finite cover. Take a $\zeta = n^{-2}$ cover from Lemma C.5 on the input ball of radius R_n. Capping is 1-Lipschitz, and the indicator is the same for every parameter, so the resulting g_θ form a global uniform cover. Its cardinality N satisfies

$$\begin{aligned}\log N &\le C P_m \log\big(Cn^2 P_m K(d + S_n)\big)\\ &\le C P_m \log(CnP_m K S_n).\end{aligned}$$

Here $B_{R_n} \le C(d + S_n)$ and $P_m \ge d$. Let $x_N = \log(4N/\delta)$. Bernstein's inequality at the covering centers, followed by a union bound, gives with probability at least $1 - \delta/2$,

$$|P_n g_\theta - Pg_\theta| \le C\sqrt{(Pg_\theta)Bx_N/n} + C(Bx_N/n + \zeta), \qquad \theta \in \Theta_m. \tag{34}$$

Indeed, each center has variance at most B times its mean. Replacing a center by a loss within ζ changes both means by at most ζ; the resulting square-root error is absorbed by $Bx_N/n + \zeta$. The choices above give

$$Bx_N/n + \zeta + CS_n^3/n^2 \le Cr_{n,m}.$$

The input radius enters the covering-number logarithm, while the bounded-loss scale in Bernstein's inequality is CS_n^3.

Step 3: comparison with the original empirical loss. For every θ, $P_n g_\theta \le P_n f_\theta$ deterministically. Combining this inequality with (33)–(34) gives

$$Pf_\theta \le P_n f_\theta + C\sqrt{(Pf_\theta) r_{n,m}} + Cr_{n,m},$$

which proves (31). For the fixed comparator θ°, the two tail bounds in Step 1 give

$$\Pr\big(P_n f_{\theta^\circ} \ne P_n g_{\theta^\circ}\big) \le 2ne^{-u} \le \delta/2.$$

On the complementary event, (34) and $Pg_{\theta^\circ} \le Pf_{\theta^\circ}$ yield (32). Intersecting the two events proves the lemma. □

Proof of Theorem 6.6. Fix a deterministic population minimizer θ° with $\mathcal{L}_{\mathrm{tr}}(\theta^\circ) = A_m$, and apply Lemma C.6. On its probability-$1-\delta$ event, empirical optimality of $\hat\theta$ gives

$$\begin{aligned}
\mathcal{L}_{\mathrm{tr}}(\hat\theta) &\le \hat{\mathcal{L}}_{\mathrm{tr},n}(\hat\theta) + C\sqrt{\mathcal{L}_{\mathrm{tr}}(\hat\theta) r_{n,m}} + Cr_{n,m} \\
&\le \hat{\mathcal{L}}_{\mathrm{tr},n}(\theta^\circ) + C\sqrt{\mathcal{L}_{\mathrm{tr}}(\hat\theta) r_{n,m}} + Cr_{n,m} \\
&\le A_m + C\sqrt{A_m r_{n,m}} + C\sqrt{\mathcal{L}_{\mathrm{tr}}(\hat\theta) r_{n,m}} + Cr_{n,m}.
\end{aligned}$$

Set $e = \mathcal{L}_{\mathrm{tr}}(\hat\theta) - A_m \ge 0$. Using $\sqrt{A_m + e} \le \sqrt{A_m} + \sqrt{e}$ and $C\sqrt{e r_{n,m}} \le e/2 + C^2 r_{n,m}/2$, we obtain

$$\mathcal{L}_{\mathrm{tr}}(\hat\theta) - A_m \le C\sqrt{A_m r_{n,m}} + Cr_{n,m}.$$

Finally, $\sqrt{A_m r_{n,m}} \le (A_m + r_{n,m})/2$ and Theorem 6.5 imply

$$\mathcal{L}_{\mathrm{tr}}(\hat\theta) \le C\frac{(1+d^*)^4}{m^2} + C\frac{S_n^3}{n}\left(P_m \log(CnP_m K S_n) + \log(8/\delta)\right).$$

This is the claimed bound for the original empirical risk minimizer. □

D Details of Gaussian Mixture Residual Flow

D.1 Proof of Equation 4

Proof. For the fitted component, linear interpolation gives the marginal $\mathcal{N}(\widehat{m}_{k,s}, \widehat{C}_{k,s})$, where $\widehat{m}_{k,s} = (1-s)\widehat{\mu}_k$ and $\widehat{C}_{k,s} = s^2 I_D + (1-s)^2\widehat{\Sigma}_k$. For $s > 0$, the covariance is positive definite, and the component velocity is

$$\widehat{v}_k^{\mathrm{G}}(z,s) = -\widehat{\mu}_k + \frac{1}{2}(\partial_s \widehat{C}_{k,s})\widehat{C}_{k,s}^{-1}(z - \widehat{m}_{k,s}).$$

All covariance matrices share the eigenvectors of $\widehat{\Sigma}_k$. Differentiating their symmetric square roots in this common eigenbasis gives

$$\frac{\mathrm{d}}{\mathrm{d}s}\widehat{C}_{k,s}^{1/2} = \frac{1}{2}(\partial_s \widehat{C}_{k,s})\widehat{C}_{k,s}^{-1/2}.$$

Consequently,

$$Y_s = \widehat{m}_{k,s} + \widehat{C}_{k,s}^{1/2}\widehat{C}_{k,t}^{-1/2}(z - \widehat{m}_{k,t})$$

satisfies $\mathrm{d}Y_s/\mathrm{d}s = \widehat{v}_k^{\mathrm{G}}(Y_s, s)$ and $Y_t = z$. Evaluating at $s = r \ge \epsilon > 0$ proves the formula, including when $\widehat{\Sigma}_k$ is singular. In particular,

$$\widehat{X}_k^{\mathrm{G}}(z_1, \epsilon, 1) = (1-\epsilon)\widehat{\mu}_k + \left[\epsilon^2 I_D + (1-\epsilon)^2\widehat{\Sigma}_k\right]^{1/2} z_1.$$

Thus, independently drawing $z_1 \sim \mathcal{N}(0, I_D)$ and a component with the fitted mixture weights samples the time-ϵ marginal of the fitted Gaussian mixture. The formula describes the flow conditional on a fixed component label. □

D.2 Residual Parameterization and Preconditioning

The decomposition in Proposition 6.2 motivates separating the explicit scaling term from the residual. Before introducing the fitted Gaussian prior, we use a CTM-style parameterization

$$X_\theta(z, r, t) = \frac{r}{t} z + \left(1 - \frac{r}{t}\right) G_\theta(z, r, t),$$

which enforces $X_\theta(z, t, t) = z$ for every choice of G_θ. EDM-inspired preconditioning adapted to the linear interpolation path gives

$$G_\theta(z, r, t) = c_{\text{skip}}(t) z + c_{\text{out}}(t) F_\theta(z, r, t),$$

where

$$s_t^2 = (1-t)^2 \sigma_x^2 + t^2, \qquad c_{\text{skip}}(t) = \frac{(1-t)\sigma_x^2}{s_t^2}, \qquad c_{\text{out}}(t) = \frac{t\sigma_x}{s_t}.$$

For centered reference data with covariance $\sigma_x^2 I_D$ and independent standard Gaussian noise, $c_{\text{skip}}(t) z_t$ is the optimal scalar linear predictor of the clean data, and $c_{\text{out}}(t)^2$ is its per-coordinate residual variance. In GMRFlow, $\sigma_x > 0$ is a chosen normalization scale.

Substituting this preconditioning into the boundary-preserving parameterization yields

$$X_\theta(z, r, t) = \left[\frac{r}{t} + \left(1 - \frac{r}{t}\right) c_{\text{skip}}(t)\right] z + \left(1 - \frac{r}{t}\right) c_{\text{out}}(t) F_\theta(z, r, t).$$

We replace the analytic term multiplying z with $\widehat{X}_k^{\mathrm{G}}(z, r, t)$ and condition the shared network on k, giving (5). The covariance rank is a fitting hyperparameter; the fitted component subspaces may differ across clusters, and the neural correction takes values in the full ambient space. Since $\widehat{X}_k^{\mathrm{G}}(z, t, t) = z$ and $1 - r/t = 0$ at $r = t$, the resulting model preserves the boundary condition. Setting $F_\theta = 0$ recovers the fitted Gaussian component flow, so the independent component and noise sampling in Section 7 recovers the time-ϵ marginal of the fitted mixture in this case.

E Additional Details of Numerical Validation

E.1 Continuous and Discrete Objectives

Dynamical system and data. For Figure 3, we use the two-dimensional autonomous system

$$v(z) = a\left(1 - \frac{\|z\|^2}{R^2}\right) z + (\omega_0 + \beta\|z\|^2) Jz, \qquad Jz = (-z_2, z_1),$$

with $a = 0.8$, $R = 2.5$, $\omega_0 = 1$, $\beta = 0.2$, and $\epsilon = 0.05$. The terminal distribution ρ_1 is uniform on the eight black cells of a 4×4 checkerboard in $[-1.9, 1.9]^2$. For $\Delta = t - r$, put $b = R^2 - \|z\|^2 + \|z\|^2 e^{-2a\Delta}$. Writing R_φ for planar rotation by angle φ, the exact backward flow is

$$X(z, r, t) = \frac{Re^{-a\Delta}}{\sqrt{b}}\, \mathsf{R}_{-\omega_0\Delta - \frac{\beta R^2}{2a}\log(R^2/b)} z.$$

The denominator satisfies $b > 0$ on the sampled trajectories. We generate $z_t = X(z_1, t, 1)$ from fresh $z_1 \sim \rho_1$ and use the same analytic flow as the evaluation reference.

Objectives and training. We fix the endpoint at $r = \epsilon$. Both methods use an MLP with four hidden layers of width 128, SiLU activations, and the raw time t as an input. Their parameterizations are $X_\theta(z, \epsilon, t) = z - (t - \epsilon)u_\theta(z, t)$ for MeanFlow and $f_\theta(z, t) = z + (t - \epsilon)F_\theta(z, t)$ for CM. MeanFlow samples $t \sim \mathrm{Unif}[\epsilon, 1]$ and minimizes squared error against the stop-gradient target $v(z_t) - (t - \epsilon)(\partial_t u_\theta + D_z u_\theta\, v)(z_t, t)$, computed using a Jacobian–vector product. CM uniformly samples adjacent pairs from eight equally spaced time nodes, giving seven intervals of length $h = (1 - \epsilon)/7$. Its teacher update is the reverse-time explicit Euler step $z^{\mathrm{teacher}}_{t_{i-1}} = z_{t_i} - h\, v(z_{t_i})$, and its stopped target uses an EMA network with decay 0.99. Both methods use the exact velocity and train for 30K updates with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 1024, gradient clipping at norm 1.

Evaluation metric. We measure the squared error of the one-step map from 1 to ϵ:

$$E_{1\to\epsilon} = \mathbb{E}_{z_1\sim\rho_1}\left[\|X_\theta(z_1, \epsilon, 1) - X(z_1, \epsilon, 1)\|_2^2\right].$$

Here $X_\theta(z, \epsilon, t) = f_\theta(z, t)$ for CM. We estimate this expectation using 16,384 fresh terminal samples. MeanFlow achieves $E_{1\to\epsilon} = 0.00956$, compared with 0.27665 for CM.

E.2 Data Subspace Structure

Data and PCA. We use 100K training images and 10K disjoint validation images. Each aligned CelebA image is cropped to the pixel box $(9, 29, 169, 189)$, resized to 80×80 with bicubic interpolation, and converted to grayscale by averaging the RGB channels. Pixel intensities are represented in $[0, 1]$. We compute the training mean μ and covariance and obtain the leading 1024 eigenvectors using a symmetric eigensolver. Let A_d contain the leading d orthonormal eigenvectors. For every validation image, we use the same training mean and basis to reconstruct

$$\widehat{x}^{(d)} = \mu + A_d A_d^\top (x - \mu).$$

Metrics and visualization. For the validation set $\mathcal{I}_{\mathrm{val}}$, the retained energy, measured relative to the training mean μ, is

$$R_{\mathrm{val}}(d) = 1 - \frac{\sum_{i\in\mathcal{I}_{\mathrm{val}}} \|x_i - \widehat{x}_i^{(d)}\|^2}{\sum_{i\in\mathcal{I}_{\mathrm{val}}} \|x_i - \mu\|^2}.$$

We also report training retained energy and validation mean squared error (MSE), averaged over images and pixels, shown in Table 3. All metrics use unclipped reconstructions; clipping to $[0, 1]$ is used only for display. Figure 7 shows three reconstructions.

E.3 Flow Map Decomposition

Data and paired queries. We use the same 100K training images, training mean μ, and nested PCA bases A_d as in Appendix E.2. We sample 512 images from the validation set. The velocity fields start from the same independently sampled $z_t^{(i)} = (1 - t)x_i + t\varepsilon_i$ at $t = 0.5$. All computations use pixel intensities in $[0, 1]$.

Table 3: PCA results on CelebA. Training and validation columns report retained energy relative to the training mean; validation MSE is measured on pixel intensities.

d	d/D	Training (%)	Validation (%)	Validation MSE
32	0.5%	82.63	82.76	0.011473
128	2%	91.85	91.91	0.005381
256	4%	94.87	94.90	0.003397
512	8%	97.05	97.04	0.001969
1024	16%	98.55	98.52	0.000988

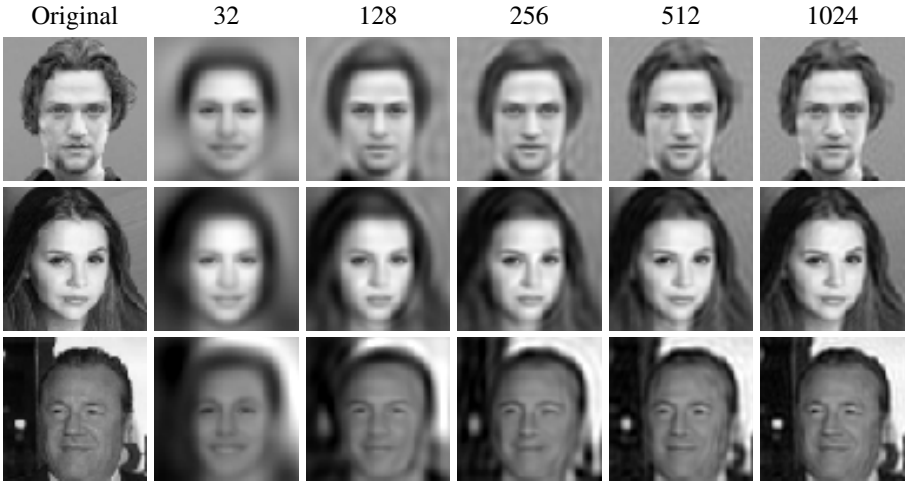


Figure 7: Complete PCA reconstruction grid for the three fixed validation examples. Columns show the original image and reconstructions with $d = 32, 128, 256, 512, 1024$ training principal components.

Velocity fields. The empirical field uses all n =100K training images at every evaluation:

$$v_{\text{emp}}(z,s) = \frac{z - \sum_{j=1}^{n} w_j(z,s)x_j}{s}, \qquad w_j(z,s) = \frac{\exp\big(-\|z-(1-s)x_j\|^2/(2s^2)\big)}{\sum_{k=1}^{n}\exp(-\|z-(1-s)x_k\|^2/(2s^2))}.$$

The neural predictor uses the group-A blind noise-prediction U-Net and its pretrained weights released in the official repository of Kadkhodaie et al. (2024), with all weights kept fixed. With its predicted additive noise f_θ, we set

$$D_\theta(y) = y - f_\theta(y), \qquad v_\theta(z,s) = \frac{z - D_\theta(z/(1-s))}{s}.$$

The interval $s \in [0.25, 0.5]$ corresponds to additive noise levels $s/(1-s) \in [1/3, 1]$.

Integration and mean correction. We integrate from $t = 0.5$ to $r = 0.25$ using fixed-step RK4 with 128 steps and retain the intermediate state at $r = 0.375$. To avoid subtracting large terms when computing the residual, the integrated state is

$$c_s = X_s - \frac{s}{t}z_t - \left(1 - \frac{s}{t}\right)\mu, \qquad \frac{\mathrm{d}c_s}{\mathrm{d}s} = \frac{c_s - [m(X_s,s) - \mu]}{s}, \quad c_t = 0,$$

where m is the empirical posterior mean or the neural denoising output.

Subspace energy and visualization. Let $P_d = A_d A_d^\top$. To account for the affine offset of uncentered images, we measure the energy of

$$X_c^{(i)} = X(z_t^{(i)}, r, t) - (1-r)\mu, \qquad h_c^{(i)} = X(z_t^{(i)}, r, t) - \frac{r}{t}z_t^{(i)} - \left(1 - \frac{r}{t}\right)\mu$$

inside the fixed data subspace, using

$$E_d(Y) = \frac{\sum_{i=1}^{512} \|P_d Y^{(i)}\|^2}{\sum_{i=1}^{512} \|Y^{(i)}\|^2}, \qquad Y \in \{X_c, h_c\}.$$

For visibility, residual images show $h/(1 - r/t)$, which equals $2h$ at $r = 0.25$.

Table 4: Paired subspace-energy results (%).

r	Field	$E_{256}(X_c)$	$E_{256}(h_c)$	$E_{1024}(X_c)$	$E_{1024}(h_c)$
0.375	Empirical	16.75	96.40	27.39	98.98
0.375	Neural	18.31	98.95	28.96	99.88
0.25	Empirical	34.76	96.38	43.43	98.97
0.25	Neural	38.15	98.55	46.93	99.83

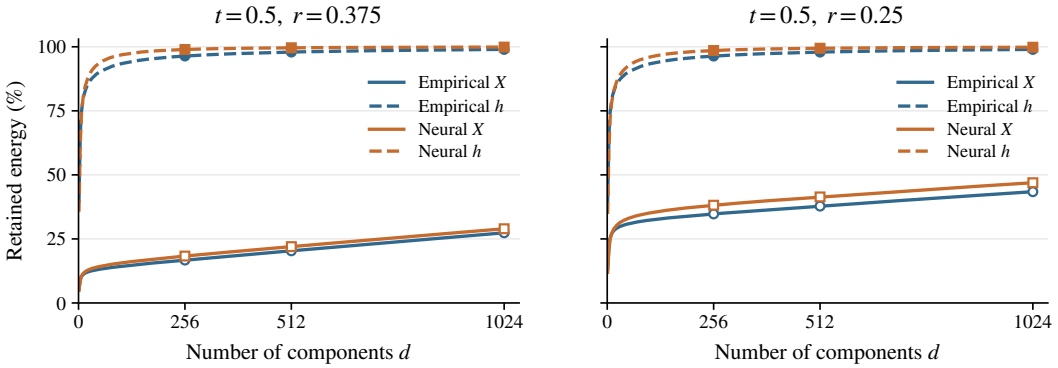


Figure 8: Energy retained by the same data PCA subspaces at two integration endpoints.

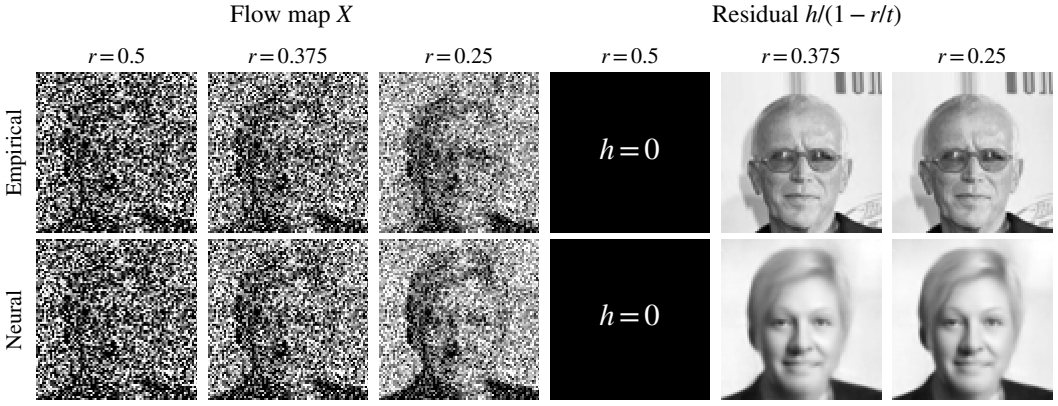


Figure 9: Paired flow-map and residual trajectories. Residuals are displayed as $h/(1 - r/t)$ for $r < t$. At the initial time $r = t = 0.5$, the residual is exactly zero, shown as a black panel.

F Additional Details and Results for Generative Experiments

F.1 Data, Model Variants, and Training Settings

Data. CelebA80 uses fixed 100K training images and disjoint 10K validation images. Images are cropped to the box $(9, 29, 169, 189)$, resized to 80×80 with bicubic interpolation, and converted to grayscale by averaging the RGB channels. FFHQ64 uses all 70K images, downsampled to 64×64 with Lanczos interpolation, for training and as the real reference set for FID evaluation. Training takes place in pixel space with image values scaled to $[-1, 1]$.

Model variants. The MF baseline directly predicts the average velocity, $X_\theta(z, r, t) = z - (t - r)F_\theta(z, r, t)$. The CTM baseline uses the residual parameterization in Appendix D, with $\sigma_x = 0.5$; HC adds a learned embedding of a fixed cluster label. GMRFlow uses the same cluster label and parameterization 5. All models use the same U-Net, with 8.23M parameters for CelebA80 and 21.85M for FFHQ64.

Training settings. All models in the main comparisons in Table 1 share the same MeanFlow objective and training settings. The FFHQ64 ablations in Table 2 use the iMF objective, with all other settings held fixed within each ablation. Both sets of experiments use global batch size 64, Adam with zero weight decay, gradient norm clipping at 1, and a 1K-update warmup. We use EMA decay 0.9999, without data augmentation or classifier-free guidance. Table 5 summarizes the training settings for the main comparisons and ablations.

Table 5: Training settings for the main comparisons and ablations. $\mathrm{LN}(\mu, \sigma)$ denotes the distribution of $\mathrm{sigmoid}(a)$ for $a \sim \mathcal{N}(\mu, \sigma^2)$. Off-diagonal pairs satisfy $r < t$.

Setting	Main comparisons	Ablations
Objective	MeanFlow	iMF
Dataset	CelebA80 / FFHQ64	FFHQ64
Reported updates	50K / 100K	50K
Peak learning rate	2×10^{-4}	10^{-4}
Adam (β_1, β_2)	$(0.9, 0.999)$	$(0.9, 0.95)$
After warmup	Constant	Cosine to 10^{-5} at 100K
Diagonal times	$\mathrm{LN}(-2, 2)$	$\mathrm{LN}(-0.4, 1)$
Off-diagonal times	$\mathrm{LN}(-2, 2)$	$\mathrm{LN}(-0.4, 1.5)$
Off-diagonal fraction	0.75	0.5

Unless varied, the ablations use DINO clustering, $K = 8$ components, covariance rank $R = 64$, and an off-diagonal fraction of 0.5; the time-ratio ablation fixes $R = 16$.

G Experiments on ImageNet64

Experimental setup. We use a 100-class subset of ImageNet-1K, with 1K training images per class (100K total), processed at 64×64 RGB resolution, scaled to $[-1, 1]$, and centered using the training-set mean. All models operate in pixel space with the same 21.85M U-Net as FFHQ64, trained for 100K updates with global batch size 64. EMA 1-NFE FID uses 10K generated images and the same 5K validation images from the selected classes' validation set.

Model variants. We compare the same CTM, HC and GMRFlow parameterizations as in the main comparisons of CelebA80 and FFHQ64, but with class conditioning. Clustering is performed within each class, and the same cluster labels are used for HC and GMRFlow.

Results. Table 6 compares the fixed 100K checkpoints. HC and GMRFlow reduce 1-NFE FID relative to MF by 5.88 and 5.42, respectively. HC obtains a slightly lower 1-NFE FID than GMRFlow, but GMRFlow has lower 2-NFE and 4-NFE FID. The training curves are shown in Figure 10. The MF configuration uses EMA warmup, whereas HC and GMRFlow use fixed EMA decay, so MF has lower EMA FID through 30K updates. The advantage of GMRFlow is less pronounced than in CelebA80 and FFHQ64, likely due to the smaller number of training images per class.

Table 6: ImageNet64 results on the selected 100 classes after 100K updates, with EMA FID at 1, 2, and 4 NFE.

Configuration	1-NFE FID ↓	2-NFE FID ↓	4-NFE FID ↓
MF	72.86	48.33	47.59
HC	**66.98**	36.80	32.47
GMRFlow	67.44	**34.94**	**31.94**

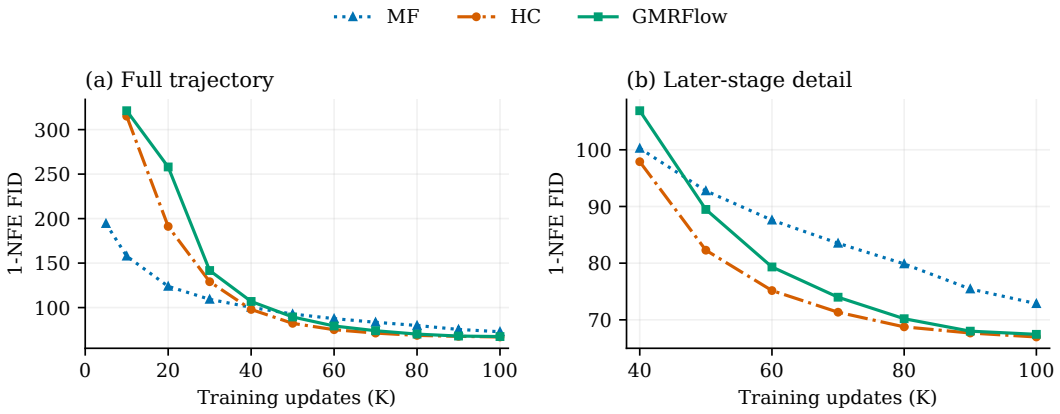


Figure 10: EMA 1-NFE FID on the selected ImageNet classes. (a) Full measured training trajectories. (b) The same trajectories over 40K–100K updates.