WASSERSTEIN CONVERGENCE OF ODE-BASED SAMPLERS IN DECENTRALIZED DIFFUSION MODEL VIA VELOCITY FIELD DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models have achieved impressive empirical success, accompanied by substantial progress in understanding their sampling convergence. Motivated by privacy and scalability, decentralized diffusion architectures replace a single global model with local experts and a learned routing mechanism. Random expert switching introduces fluctuations beyond ordinary ODE discretization error, requiring an analysis of how local approximation errors and routing randomness jointly affect generation. We study the sampling convergence of decentralized diffusion models for finite-mixture distributions under linear flow interpolation, using a single-expert-per-step Euler sampler with normalized, state-dependent routing. Under compact support, bounded learned data predictions, and spatial and temporal regularity of the learned composite field, we establish a non-asymptotic Wasserstein-2 (W_2) guarantee based on time-integrated population L^2 errors. We measure expert error ϵ_v under the respective component marginals and router error ϵ_a under the mixture marginals. At a fixed cutoff $t = \delta > 0$, the N-step sampler admits a W_2 upper bound of order $\mathcal{O}(\epsilon_v + \epsilon_a + N^{-1/2})$ relative to the exact cutoff marginal. The key is to combine posterior-weighted error decomposition with cancellation of the experts' shared linear drift, allowing ordinary L^2 router error to enter the bound linearly without an additional spatial weight. Our analysis separates deterministic integration error from stochastic routing fluctuations and explicitly accounts for data-end truncation, connecting local population approximation accuracy to distributional guarantees for decentralized sampling.

1 INTRODUCTION

Diffusion probabilistic models (Ho et al., 2020; Song et al., 2020) and their recent generalization, Flow Matching models (Lipman et al., 2023; Albergo et al., 2023), have achieved unprecedented success in generative tasks, covering a wide range of modalities such as text-to-image generation (Rombach et al., 2022), video synthesis (Ho et al., 2022), 3D content creation (Poole et al., 2023), and image restoration (Saharia et al., 2022). Among various sampling strategies, the Probability Flow ODE (PF-ODE) (Song et al., 2020; Karras et al., 2022) formulation has become the de facto standard for efficient deployment. Unlike stochastic samplers, the PF-ODE defines a deterministic transport between the noise and data distributions while preserving marginal distributions, enabling the use of specialized ODE solvers to significantly reduce sampling steps without sacrificing quality (Lu et al., 2022; 2025). Given its centrality, the theoretical properties of the PF-ODE have been extensively studied, with a rich body of work establishing convergence guarantees in terms of Wasserstein-2 distance (W_2), KL divergence, and Total Variation (TV) distance (Lee et al., 2022; Kwon et al., 2022; Chen et al., 2023b;a; Gao & Zhu, 2024; Huang et al., 2025).

However, existing theoretical frameworks predominantly rely on an assumption: they presume access to a global score function (or velocity field) learned from the entire dataset. This assumption is increasingly challenged by the shift towards decentralized and modular learning paradigms. Driven by data privacy concerns, computational scaling bottlenecks, and the mixture nature of real-world distributions, recent works have explored *Decentralized Diffusion Model* architectures (McAllister et al., 2025; Chen et al., 2025). In these settings, one does not learn a single global velocity field. Instead, multiple "expert" fields are trained on local data partitions or clusters, and the global

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.

generation process is realized by dynamically routing or switching between these local experts. While such heuristic routing strategies have demonstrated empirical success, their stochastic updates require analysis beyond deterministic ODE bounds, even when the composite mean field satisfies Lipschitz regularity. Moreover, the routing decisions introduce stochastic fluctuations that act as an extrinsic noise term, complicating the analysis of the sampling process. Consequently, a critical theoretical gap emerges: *Does the ODE sampling process driven by a mixture of local experts rigorously converge to the target distribution? If so, how do the routing mechanism and local approximation errors impact the final generation quality?*

In this paper, we address this problem by establishing a convergence theory for decentralized sampling via the Probability Flow ODE, in contrast to previous works that focus on SDE-based samplers (Anonymous, 2026).To the best of our knowledge, this is the first rigorous convergence analysis of decentralized ODE sampling with state-dependent stochastic expert routing under linear flow interpolation. We study decentralized ODE sampling under linear flow interpolation, a setting that directly covers linear-interpolation flow matching. We characterize the global velocity field as a probability-weighted superposition of component fields. Image-generation and trajectory-error experiments in Appendix B provide empirical support for the resulting decentralized sampling framework and and characterize its behavior across sampling budgets, complementing the theoretical convergence analysis. Our main contributions are summarized as follows:

- **Theoretical Framework for Mixture Dynamics:** We formulate the Probability Flow ODE for mixture distributions and derive a velocity field decomposition that justifies the use of decentralized experts. We identify the specific conditions under which component-wise learning approximates the global velocity field.
- **Wasserstein-2 Convergence Guarantee:** We provide a rigorous convergence analysis for the proposed sampling algorithm. For linear flow interpolation under compact support, bounded learned data predictions, normalized routing, and spatial and temporal regularity of the learned composite field, we establish the **first** non-asymptotic convergence bound for the proposed sampling algorithm. Specifically, we bound the Wasserstein-2 distance between the generated distribution and the exact marginal at the truncation time $t = \delta$ by $\mathcal{O}(\epsilon_v + \epsilon_a + N^{-1/2})$ for fixed cutoff and regularity parameters. Crucially, our analysis explicitly characterizes the error coefficients' dependence on the data support radius R, dimension d, the number of mixture components L, the truncation time δ, and the learned-field regularity constants, thereby providing a rigorous theoretical baseline for the decentralized diffusion framework. The approximation budgets ϵ_v and ϵ_a are measured in time-integrated population L^2 norms: expert velocity errors under their respective component marginals, weighted by the mixture proportions, and routing error under the mixture marginals. The key is to combine the Bayes identity $a_l p_t = \pi_l p_t^l$ with cancellation of the experts' shared drift under normalized routing, allowing ordinary L^2 router error to enter the W_2 bound linearly without an additional spatial weight. The proof further separates the N^{-1} contribution of deterministic Euler discretization from the $N^{-1/2}$ contribution of random expert selection and includes the truncation term $\delta\sqrt{R^2 + d}$ when comparing with the data distribution.

2 Related Work

2.1 Convergence of SDE-based Samplers

Score-based generative models (SGMs) (Song et al., 2020) have achieved remarkable success, prompting extensive theoretical analysis of their sampling convergence (Lee et al., 2022; Kwon et al., 2022; Chen et al., 2023a; Benton et al., 2023). A significant body of work focuses on Stochastic Differential Equation (SDE) samplers. For instance, Chen et al. (2023b) established polynomial-time convergence guarantees in Total Variation (TV) distance under minimal data assumptions. However, these analyses typically rely on SDE-specific tools, such as Girsanov's theorem and log-Sobolev inequalities, which crucially exploit the stochasticity of Brownian motion. The injected noise effectively acts as a regularizer, facilitating distribution convergence. Furthermore, to obtain meaningful TV bounds for general distributions, techniques like early stopping (targeting a distribution at $t = \delta$) are often required to avoid singularities near $t = 0$. In our decentralized setting, we utilize the Probability Flow ODE, which is deterministic. The absence of Brownian noise removes the contractive effects found

in SDEs, making error accumulation more delicate and limiting the direct application of SDE-based proof techniques.

2.2 Convergence of ODE-based Samplers

The theoretical understanding of ODE-based samplers has Progressed rapidly. Recent works, such as Kwon et al. (2022)and Gao & Zhu (2024), have established rigorous Wasserstein-2 (W_2) convergence guarantees for ODE samplers in diffusion models and flow models. These analyses typically rely on Grönwall inequalities and assume the velocity field is deterministic and Lipschitz continuous. The choice of W_2 distance in these works, as well as in the present work, is fundamentally motivated by the nature of the sampling process as numerical integration. Unlike TV or KL divergence, which require controlling pointwise density evolution, W_2 naturally quantifies the mean square error of particle trajectories. However, existing ODE theories do not directly apply to our decentralized sampling framework. Although our underlying probability flow is deterministic, the *realized* velocity field in our decentralized setting involves stochastic routing among experts. This introduces a "routing noise" (modeled as martingale noise in our analysis) that is neither Gaussian diffusion nor simple discretization error. A direct application of standard ODE bounds would treat this stochastic switching as a worst-case perturbation, leading to loose error estimates. Our work extends deterministic ODE theory by modeling this routing variability as a martingale difference sequence, allowing for tighter convergence rates.

2.3 Decentralized Diffusion Models

The paradigm of decentralized learning for diffusion models is an emerging area of research (McAllister et al., 2025; Chen et al., 2025). Most relevant to our analysis is the Decentralized Diffusion Models (DDM) framework proposed by McAllister et al. (2025), which utilizes a routing mechanism to partition the generation process among multiple experts. Although McAllister et al. (2025) demonstrate the empirical effectiveness and scalability of this approach, their study focuses primarily on system design and implementation, and lacks a rigorous theoretical foundation for convergence. To the best of our knowledge, our work provides the first rigorous Wasserstein-2 (W_2) convergence guarantee for such mixture-based decentralized diffusion frameworks, explicitly accounting for the stochasticity induced by the routing mechanism.

3 Notations and Preliminaries

This section introduces the notation and mixture model, describes the flow interpolation and velocity decomposition, and defines the decentralized sampler analyzed in the following sections.

3.1 Notation and Mixture Model

We write $\|\cdot\|$ for the Euclidean norm and $\|U\|_2 = (\mathbb{E}\|U\|^2)^{1/2}$ for the L^2 norm of a square-integrable random vector U. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the probability measures on $\mathbb{R}^d$ with finite second moments. For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, their quadratic Wasserstein distance is defined by

$$W_2^2(\mu,\nu) = \inf_{\gamma\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2 \,\gamma(dx,dy), \tag{1}$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν.

We model the target data distribution as a finite mixture

$$p_0 = \sum_{l=1}^{L} \pi_l p_0^l, \qquad \pi_l > 0, \qquad \sum_{l=1}^{L} \pi_l = 1. \tag{2}$$

Each component p_0^l is a probability measure on $\mathbb{R}^d$. The components need not have densities or disjoint supports, and may in particular be empirical measures of finite datasets. To track the generating component, we introduce a latent label $C \in \{1,\dots,L\}$ with $\mathbb{P}(C=l) = \pi_l$ and draw $x_0 \mid C = l \sim p_0^l$, so that $x_0 \sim p_0$.

3.2 Flow Interpolation and Reverse Dynamics

To connect the data distribution to a Gaussian reference, we use the linear interpolation

$$x_t = (1-t)x_0 + t\epsilon, \qquad 0 \le t \le 1, \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I_d)$ is independent of (x_0, C). Write $p_t = \mathrm{Law}(x_t)$ and $p_t^l = \mathrm{Law}(x_t \mid C = l)$ for the marginal and component distributions. The mixture identity $p_t = \sum_{l=1}^{L} \pi_l p_t^l$ holds at every time, with p_0 as the data endpoint and $p_1 = p_1^l = \mathcal{N}(0, I_d)$ for every l. For $t > 0$, these distributions have strictly positive smooth densities, for which we use the same notation (Lemma A.1).

The mixture structure yields a decomposition of the interpolation velocity into component fields. Under the compact-support condition in Assumption 4.1 below, define the posterior routing probabilities and component means, for $t > 0$, by

$$a_l(x,t) = \frac{\pi_l p_t^l(x)}{p_t(x)}, \qquad m_l(x,t) = \mathbb{E}[x_0 \mid x_t = x, C = l]. \tag{4}$$

We use the versions defined at every x by the Gaussian kernel formulas in Lemma A.1. The routing weights form a probability vector and satisfy

$$a_l(x,t)p_t(x) = \pi_l p_t^l(x), \qquad m(x,t) := \mathbb{E}[x_0 \mid x_t = x] = \sum_{l=1}^{L} a_l(x,t) m_l(x,t). \tag{5}$$

Conditioning the interpolation velocity $\epsilon - x_0$ gives

$$\begin{aligned} v_l(x,t) &:= \mathbb{E}[\epsilon - x_0 \mid x_t = x, C = l] = \frac{x - m_l(x,t)}{t}, \\ v(x,t) &:= \mathbb{E}[\epsilon - x_0 \mid x_t = x] = \sum_{l=1}^{L} a_l(x,t) v_l(x,t) = \frac{x - m(x,t)}{t}. \end{aligned} \tag{6}$$

These identities follow from Bayes' rule (Lemma A.2). In particular, the first identity in equation 5 will allow us to measure each expert's approximation error under its own component marginal.

Generation follows the interpolation marginals from the Gaussian endpoint toward the data. Fix a cutoff $\delta \in (0,1)$, set $T = 1 - \delta$, and use the increasing generation time $\tau = 1 - t$. The reverse velocity and its flow are

$$\widetilde{v}(x,\tau) = -v(x, 1-\tau), \qquad \dot{X}(\tau) = \widetilde{v}(X(\tau), \tau), \qquad X(0) = \zeta \sim \mathcal{N}(0, I_d). \tag{7}$$

Under Assumption 4.1, Proposition A.4 ensures a unique solution on $[0, T]$ and identifies its marginals:

$$\mathrm{Law}(X(\tau)) = p_{1-\tau}, \quad 0 \le \tau \le T, \qquad \mathrm{Law}(X(T)) = p_\delta. \tag{8}$$

The random interpolation x_t and the ODE flow $X(\tau)$ thus have the same marginals at corresponding times, but need not follow the same sample paths. Component velocities use the same reversal of time and sign; routing probabilities are evaluated at $t = 1 - \tau$.

3.3 Decentralized Sampling with Stochastic Expert Routing

Motivated by decentralized diffusion models, we consider local velocity experts associated with the mixture components and a learned router that selects one expert at each sampling step. We formulate this sampling mechanism for the linear flow interpolation introduced above.

To approximate this reverse flow, we replace the posterior means and routing probabilities by learned functions $m_{l,\theta}$ and $a_{l,\theta}$ on $\mathbb{R}^d \times [\delta, 1]$, where θ collects all expert and routing parameters. The routing outputs are nonnegative and sum to one. We define

$$\begin{aligned} v_{l,\theta}(x,t) &= \frac{x - m_{l,\theta}(x,t)}{t}, & \bar{m}_\theta(x,t) &= \sum_{l=1}^{L} a_{l,\theta}(x,t) m_{l,\theta}(x,t), \\ v_\theta(x,t) &= \sum_{l=1}^{L} a_{l,\theta}(x,t) v_{l,\theta}(x,t) = \frac{x - \bar{m}_\theta(x,t)}{t}. \end{aligned} \tag{9}$$

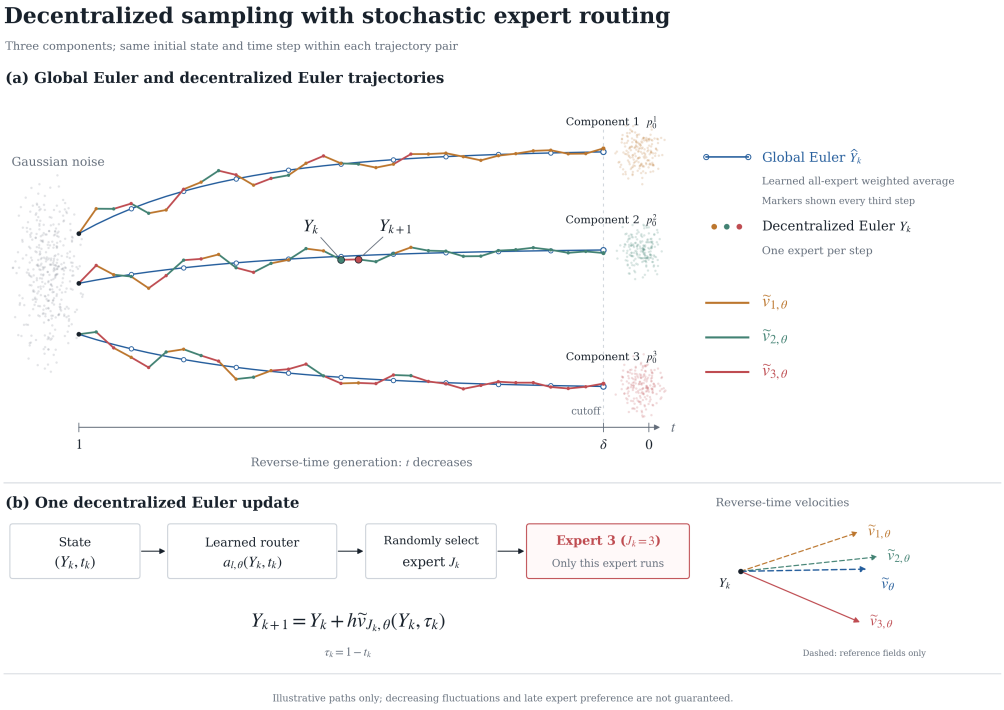

Figure 1: Decentralized sampling with stochastic expert routing. (a) Schematic trajectories of decentralized Euler Y_k and the all-expert mean-field Euler reference $\widehat{Y}_k$, sharing the initial state and time grid. Colors identify selected experts. (b) The router samples J_k with probabilities $a_{l,\theta}(Y_k, t_k)$, and only the selected expert is evaluated. Generation runs from $t = 1$ to the cutoff $t = \delta$; the component clouds depict the data distribution at $t = 0$. Paths are illustrative, not experimental trajectories or pathwise guarantees.

Write $\widetilde{v}_{l,\theta}(x, \tau) = -v_{l,\theta}(x, 1-\tau)$ and $\widetilde{v}_\theta(x, \tau) = -v_\theta(x, 1-\tau)$ for the corresponding reverse fields. The boundedness and regularity conditions used in the analysis are stated in the next section.

The sampler evaluates one expert at each step, selected using the learned routing probabilities at the current state and time. For an integer $N \geq 1$, let $h = T/N$, $\tau_k = kh$ and $t_k = 1 - \tau_k$, for $k = 0, \ldots, N$. Starting from $Y_0 = \zeta$, define the information available before step k and the conditional sampling rule by

$$\mathcal{F}_k = \sigma(\zeta, J_0, \ldots, J_{k-1}), \qquad \mathbb{P}(J_k = l \mid \mathcal{F}_k) = a_{l,\theta}(Y_k, t_k). \tag{10}$$

For $k = 0, \ldots, N-1$, the Euler update is

$$\begin{aligned} Y_{k+1} &= Y_k + h\widetilde{v}_{J_k,\theta}(Y_k, \tau_k) \\ &= \left(1 - \frac{h}{t_k}\right) Y_k + \frac{h}{t_k} m_{J_k,\theta}(Y_k, t_k). \end{aligned} \tag{11}$$

Since $t_k \geq \delta + h > h$ at each update, the second line is a convex combination, a property used later to control moments. We denote the output distribution by $q^\theta_{N,\delta} = \mathrm{Law}(Y_N)$. The convergence theorem below bounds its distance to the exact cutoff marginal p_δ and, after accounting for truncation, to the data distribution p_0. Figure 1 illustrates the stochastic expert update and compares it with the deterministic Euler reference driven by the learned composite field.

4 Convergence Theory

This section establishes a Wasserstein-2 error bound for the decentralized sampler.

4.1 ASSUMPTIONS

We introduce the following assumptions regarding the data distribution, neural networks, and their approximation errors. Throughout the analysis, the cutoff $\delta \in (0, 1)$ and the learned parameters θ are fixed. All learned functions below are defined on $\mathbb{R}^d \times [\delta, 1]$.

Assumption 4.1 (Compactly supported mixture)**.** There exists $R > 0$ such that the components in equation 2 satisfy

$$\operatorname{supp}(p_0^l) \subseteq \overline{B}(0, R), \qquad l = 1, \dots, L, \tag{12}$$

where $\overline{B}(0, R)$ is the closed Euclidean ball.

This condition allows empirical measures and overlapping components. It bounds the posterior means by R and yields the true-field regularity established in Lemmas A.1 and A.3.

Assumption 4.2 (Bounded data prediction and normalized routing)**.** The functions $m_{l,\theta}$ and $a_{l,\theta}$ are Borel measurable and, for every $(x, t) \in \mathbb{R}^d \times [\delta, 1]$ and $l = 1, \dots, L$, satisfy

$$\|m_{l,\theta}(x, t)\| \le R, \qquad a_{l,\theta}(x, t) \ge 0, \qquad \sum_{l=1}^{L} a_{l,\theta}(x, t) = 1. \tag{13}$$

In particular, $\|\bar{m}_\theta(x, t)\| \le R$. Combined with the parameterization in equation 9, this bound controls the growth of the learned fields and the moments of the sampling process.

Assumption 4.3 (Regularity of the learned composite field)**.** There exist finite constants $\Lambda, \Gamma \ge 0$ such that, for all $x, y \in \mathbb{R}^d$ and $s, t \in [\delta, 1]$,

$$\|v_\theta(x, t) - v_\theta(y, t)\| \le \Lambda \|x - y\|, \tag{14}$$

$$\|v_\theta(x, t) - v_\theta(x, s)\| \le \Gamma(1 + \|x\|)|t - s|. \tag{15}$$

Under Assumption 4.1, the true field v satisfies the same form of bounds (Lemma A.3), motivating this condition on the learned composite field. The constants Λ, Γ are separate assumption parameters and may depend on θ and δ. Spatial regularity controls stability; temporal regularity also enters the Euler discretization estimate.

Assumption 4.4 (Population approximation errors)**.** There exist finite $\epsilon_v, \epsilon_a \ge 0$ such that

$$\int_\delta^1 \sum_{l=1}^{L} \pi_l \int_{\mathbb{R}^d} \|v_{l,\theta}(x, t) - v_l(x, t)\|^2 \, p_t^l(dx) \, dt \le \epsilon_v^2, \tag{16}$$

$$\int_\delta^1 \int_{\mathbb{R}^d} \sum_{l=1}^{L} |a_{l,\theta}(x, t) - a_l(x, t)|^2 \, p_t(dx) \, dt \le \epsilon_a^2. \tag{17}$$

Consistent with the squared-error objective of flow matching, we measure each expert's population L^2 error under its component marginal. The routing error is measured under the mixture marginal. Both time integrals over $[\delta, 1]$ are unnormalized.

4.2 MAIN CONVERGENCE RESULTS

Our main result bounds the distance from the sampler output to both the cutoff marginal and the data distribution. It applies to every integer $N \ge 1$, without an additional step-size restriction.

Theorem 4.5 (Wasserstein-2 error bound)**.** *Suppose Assumptions 4.1–4.4 hold. For fixed $\delta \in (0, 1)$ and any integer $N \ge 1$, let $q_{N,\delta}^\theta$ be the output law of the sampler in equation 10–equation 11. Define*

$$\mathcal{E}_{N,\delta,\theta} = A_v \epsilon_v + A_a \epsilon_a + \frac{A_{\text{disc}}}{\sqrt{N}}, \tag{18}$$

where the finite, nonnegative coefficients

$$A_v = A_v(\delta, \Lambda), \qquad A_a = A_a(\delta, R, L, \Lambda), \qquad A_{\text{disc}} = A_{\text{disc}}(\delta, R, d, \Lambda, \Gamma) \tag{19}$$

are given explicitly in equation 42. Then

$$W_2(q^{\theta}_{N,\delta}, p_\delta) \le \mathcal{E}_{N,\delta,\theta}, \tag{20}$$

$$W_2(q^{\theta}_{N,\delta}, p_0) \le \delta\sqrt{R^2+d} + A_v\epsilon_v + A_a\epsilon_a + \frac{A_{\mathrm{disc}}}{\sqrt{N}}. \tag{21}$$

The coefficients are independent of N when their listed arguments are fixed.

The four terms in equation 21 account for data-end truncation, expert approximation, routing approximation, and finite-step sampling, respectively. The last term combines an ordinary Euler contribution of order N^{-1} with a contribution from stochastic expert routing of order $N^{-1/2}$. Their separate coefficients are retained in the proof and in Proposition A.10. For fixed cutoff and learned fields, increasing N reduces the finite-step term; the truncation and approximation terms remain. The dependence on δ and the learned regularity constants is explicit, so the theorem does not assert a joint rate as the cutoff and approximation errors tend to zero. The complete proof and expanded constants appear in Proposition A.11 and Appendix A.6.

5 Proof of the Main Result

We prove Theorem 4.5 by separating network approximation, ordinary Euler discretization, and stochastic expert routing. The following argument gives the main steps; Appendix A contains the complete estimates and explicit constants.

5.1 Coupled Processes and Moment Control

Alongside the true reverse flow X and the sampler Y, introduce the learned composite flow Z and its deterministic Euler approximation $\widehat{Y}$. With $f = \widetilde{v}_\theta$, set

$$\dot{Z}(\tau) = f(Z(\tau),\tau), \qquad \widehat{Y}_{k+1} = \widehat{Y}_k + hf(\widehat{Y}_k,\tau_k), \qquad X(0) = Z(0) = \widehat{Y}_0 = Y_0 = \zeta. \tag{22}$$

All four processes are coupled through the same Gaussian initial condition. Their well-posedness follows from Proposition A.4 and Lemma A.5. The terminal error splits as

$$\|Y_N - X(T)\|_2 \le \|X(T) - Z(T)\|_2 + \|Z(T) - \widehat{Y}_N\|_2 + \|\widehat{Y}_N - Y_N\|_2. \tag{23}$$

The data-prediction structure supplies a common moment bound. Since $t_k \ge \delta + h > h$, the updates of both discrete processes are convex combinations of the current state and a prediction bounded by R. The continuous flows have corresponding integrating-factor representations. Consequently, with $M = \max\{R, \sqrt{d}\}$,

$$\max\left\{\sup_{\tau\in[0,T]}\|X(\tau)\|_2,\ \sup_{\tau\in[0,T]}\|Z(\tau)\|_2,\ \max_{k\le N}\|\widehat{Y}_k\|_2,\ \max_{k\le N}\|Y_k\|_2\right\} \le M. \tag{24}$$

The suprema here are outside the expectations. The bound holds for every $N \ge 1$ and is independent of N and δ (Proposition A.4 and Lemma A.5).

5.2 From Component Errors to Flow Error

The first comparison connects the component-wise training marginals to the global field. For a probability measure μ, write $\|g\|_{L^2(\mu)} = (\int \|g(x)\|^2\,\mu(dx))^{1/2}$.

Lemma 5.1 (Composite velocity approximation)**.** *Under Assumptions 4.1, 4.2, and 4.4,*

$$\int_\delta^1 \|v_\theta(\cdot,t) - v(\cdot,t)\|_{L^2(p_t)}\,dt \le \sqrt{T}\left(\epsilon_v + \frac{R\sqrt{L}}{\delta}\epsilon_a\right). \tag{25}$$

Proof. At each (x,t), normalization of the two routing vectors gives

$$v_\theta - v = \sum_{l=1}^{L} a_l(v_{l,\theta} - v_l) - \frac{1}{t}\sum_{l=1}^{L}(a_{l,\theta} - a_l)m_{l,\theta}. \tag{26}$$

For the first term, Jensen's inequality with the true weights a_l and the identity $a_l p_t = \pi_l p_t^l$ imply

$$\left\| \sum_l a_l (v_{l,\theta} - v_l) \right\|^2_{L^2(p_t)} \le \sum_l \pi_l \|v_{l,\theta} - v_l\|^2_{L^2(p_t^l)}.$$

For the second term, the common x/t drift cancels because $\sum_l (a_{l,\theta} - a_l) = 0$. Bounded predictions and Cauchy–Schwarz give the pointwise bound

$$\left\| \frac{1}{t} \sum_l (a_{l,\theta} - a_l) m_{l,\theta} \right\| \le \frac{R\sqrt{L}}{t} \left(\sum_l |a_{l,\theta} - a_l|^2 \right)^{1/2}.$$

The triangle inequality, $t \ge \delta$, and Cauchy–Schwarz in time yield equation 25 from Assumption 4.4. Lemma A.6 gives the full pointwise-in-time and integrated estimates. □

To compare X and Z, evaluate the field approximation error along X and use the learned field's spatial Lipschitz constant for the position perturbation:

$$\widetilde{v}(X,\tau) - f(Z,\tau) = \{\widetilde{v}(X,\tau) - f(X,\tau)\} + \{f(X,\tau) - f(Z,\tau)\}.$$

Since $\mathrm{Law}(X(\tau)) = p_{1-\tau}$, Gronwall's inequality and Lemma 5.1 give

$$\|X(T) - Z(T)\|_2 \le e^{\Lambda T} \int_\delta^1 \|v_\theta - v\|_{L^2(p_t)}\, dt \le A_v \epsilon_v + A_a \epsilon_a. \tag{27}$$

Thus the population errors are used on the true flow marginals, exactly where they are assumed to be small. Proposition A.7 provides the stability argument.

5.3 Discretization and Stochastic Expert Routing

For the second comparison, interpolate $\widehat{Y}$ linearly between grid points. The spatial and temporal bounds in Assumption 4.3, together with equation 24, control the local error from freezing the field on each step. Integrating these local errors and applying stability yields

$$\|Z(T) - \widehat{Y}_N\|_2 \le \frac{A_E}{N}, \tag{28}$$

where A_E is given in equation 42. This is the step where the temporal constant Γ enters the error bound; see Proposition A.8.

For the last comparison, define the centered sampling noise at the current state by

$$\xi_k = \widetilde{v}_{J_k,\theta}(Y_k,\tau_k) - f(Y_k,\tau_k) = \frac{m_{J_k,\theta}(Y_k,t_k) - \bar{m}_\theta(Y_k,t_k)}{t_k}. \tag{29}$$

The conditional sampling rule and bounded predictions imply

$$\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\xi_k\|^2 \mid \mathcal{F}_k] \le \frac{4R^2}{t_k^2} \le \frac{4R^2}{\delta^2}. \tag{30}$$

In particular, the common linear drift cancels in the noise. With $D_k = Y_k - \widehat{Y}_k$, the error recursion is

$$D_{k+1} = D_k + h\{f(Y_k,\tau_k) - f(\widehat{Y}_k,\tau_k)\} + h\xi_k. \tag{31}$$

The first two terms are $\mathcal{F}_k$-measurable. Squaring and taking conditional expectation therefore removes the cross term with ξ_k, giving

$$\mathbb{E}[\|D_{k+1}\|^2 \mid \mathcal{F}_k] \le (1 + \Lambda h)^2 \|D_k\|^2 + \frac{4R^2 h^2}{\delta^2}. \tag{32}$$

Iterating from $D_0 = 0$ yields $\|Y_N - \widehat{Y}_N\|_2 \le A_S/\sqrt{N}$. The full summation appears in Proposition A.9. Only conditional centering is used; the labels need not be unconditionally independent.

5.4 Distributional Error and Truncation

Combining the three comparisons in equation 23 and using $N^{-1} \le N^{-1/2}$ gives

$$\|Y_N - X(T)\|_2 \le A_v\epsilon_v + A_a\epsilon_a + \frac{A_E}{N} + \frac{A_S}{\sqrt{N}} \le \mathcal{E}_{N,\delta,\theta}, \tag{33}$$

with $A_{\mathrm{disc}} = A_E + A_S$. Since $X(T) \sim p_\delta$, this coupling implies $W_2(q^\theta_{N,\delta}, p_\delta) \le \mathcal{E}_{N,\delta,\theta}$. Finally, the natural coupling from the original interpolation gives

$$W_2(p_\delta, p_0) \le \delta\big(\mathbb{E}\|\epsilon - x_0\|^2\big)^{1/2} = \delta\sqrt{d + \mathbb{E}\|x_0\|^2} \le \delta\sqrt{d + R^2}. \tag{34}$$

The equality uses independence and the zero mean of ϵ. The Wasserstein triangle inequality completes Theorem 4.5. Propositions A.10 and A.11 state the terminal mean-square and distributional estimates with all constants.

6 Discussion and Limitations

Theorem 4.5 gives a quantitative guarantee for the specified decentralized flow sampler with stochastic expert routing. Its assumptions and constant dependencies also delimit the current result.

Cutoff and stability. The total bound includes the explicit truncation term $\delta\sqrt{R^2 + d}$. Other coefficients contain inverse powers of δ and the stability factor $e^{\Lambda(1-\delta)}$. The learned regularity constants may themselves vary with the cutoff or the network. The bound therefore does not yield a joint vanishing-cutoff rate without additional control of these dependencies. Sharper stability estimates could improve this dependence; the present upper bound does not establish that such growth is unavoidable.

Population errors and model structure. The expert and routing errors are measured in population L^2 norms on the true interpolation marginals. This is closer to the form of training losses than a uniform pointwise bound, but the analysis does not supply a statistical guarantee connecting finite-sample training to these population errors. Bounded data predictions and global regularity of the learned composite field are explicit assumptions and require justification for a given implementation.

Algorithmic scope. The analysis concerns uniform-step Euler updates that select one learned expert according to the normalized routing probabilities. It provides upper bounds and does not establish optimality of the $N^{-1/2}$ term. Extensions to other interpolation schedules, unbounded data distributions, and alternative routing or integration schemes require separate analysis.

7 Conclusion

In this paper, We studied the theoretical convergence of probability flow ODE (PF-ODE) sampling for diffusion/flow-matching models in decentralized settings where the data distribution admits a mixture structure and the global generation dynamics are induced by combining multiple experts. To bridge the gap between practical mixture-of-experts samplers and classical ODE convergence theory (which typically assumes a single smooth global field), we introduced a velocity field decomposition framework that represents the global velocity as a posterior-weighted superposition of component-wise velocity fields.

Under compact support, bounded learned data predictions, normalized routing, and spatial and temporal regularity of the learned composite field, we established a non-asymptotic convergence guarantee based on time-integrated population L^2 errors. At a fixed cutoff $t = \delta > 0$, the N-step decentralized sampler satisfies a W_2 bound of order $\mathcal{O}(\epsilon_v + \epsilon_a + N^{-1/2})$ relative to p_δ. The analysis separates expert and router approximation errors, deterministic Euler discretization, and stochastic expert-selection fluctuations, and includes the truncation term $\delta\sqrt{R^2 + d}$ when comparing with the data distribution. The explicit constants depend on the cutoff and learned-field regularity, so this result does not by itself provide a joint vanishing-cutoff rate.

AI USE STATEMENT

The core theoretical ideas and proof strategies were developed by the authors. We used large language models to assist with language editing and polishing, checking and refining intermediate mathematical derivations, and improving the clarity and consistency of proof exposition. The authors reviewed all AI-assisted edits to ensure accuracy and consistency with the intended mathematical meaning. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

As a nearly pure theoretical work, we feel our work is less likely to suffer from ethical issues. The authors have read and comply with the ICLR Code of Ethics. This work primarily develops theoretical results and includes experiments on existing image benchmarks, including CelebA-HQ face images. No new human-subject data were collected. We acknowledge the privacy, bias, and misuse considerations associated with face-image generation. We do not foresee harmful or dual-use implications from the proposed methods. There are no conflicts of interest or undisclosed sponsorship.

REPRODUCIBILITY STATEMENT

The main contributions of this work are theoretical. The main text specifies the mathematical framework, assumptions, and convergence guarantees. Detailed derivations and proofs are provided in the main text and appendices to support independent verification of the results.

REFERENCES

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Anonymous. Convergence theory of decentralized diffusion models via pseudo-non-markov analysis, 2026. URL https://openreview.net/forum?id=8N7vPCQ7IQ.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Hansheng Chen, Kai Zhang, Hao Tan, Zexiang Xu, Fujun Luan, Leonidas Guibas, Gordon Wetzstein, and Sai Bi. Gaussian mixture flow matching models. *arXiv preprint arXiv:2504.05304*, 2025.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *International Conference on Learning Representations*, 2023b.

Xuefeng Gao and Lingjiong Zhu. Convergence analysis for general probability flow odes of diffusion models in wasserstein distances. *arXiv preprint arXiv:2401.17958*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Dohyun Kwon, Ying Fan, and Kangwook Lee. Score-based generative modeling secretly minimizes the wasserstein distance. *Advances in Neural Information Processing Systems*, 35:20205–20217, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882, 2022.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *11th International Conference on Learning Representations, ICLR 2023*, 2023.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pp. 1–22, 2025.

David McAllister, Matthew Tancik, Jiaming Song, and Angjoo Kanazawa. Decentralized diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23323–23333, 2025.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.02502.

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.

A Complete Proofs for Flow Sampling

A.1 Notation and Comparison Processes

This appendix proves Theorem 4.5 and gives explicit expressions for its coefficients. For convenience, we recall the notation from Section 3 and the assumptions from Section 4.1; each result below specifies the conditions it needs. We use the Euclidean norm $\|\cdot\|$, its induced operator norm $\|\cdot\|_{\mathrm{op}}$, and $\|g\|_{L^2(\mu)} = (\int \|g(x)\|^2 \mu(dx))^{1/2}$.

Mixture and fields. Let $p_0 = \sum_{l=1}^{L} \pi_l p_0^l$, where $\pi_l > 0$ and $\sum_l \pi_l = 1$. Draw C with $\mathbb{P}(C = l) = \pi_l$, $x_0 \mid C = l \sim p_0^l$, and an independent $\epsilon \sim \mathcal{N}(0, I_d)$. Set $x_t = (1-t)x_0 + t\epsilon$, $p_t = \mathrm{Law}(x_t)$, and $p_t^l = \mathrm{Law}(x_t \mid C = l)$. For $t > 0$, using the same notation for their densities, define

$$a_l(x,t) = \frac{\pi_l p_t^l(x)}{p_t(x)}, \qquad m_l(x,t) = \mathbb{E}[x_0 \mid x_t = x, C = l], \qquad v_l(x,t) = \frac{x - m_l(x,t)}{t}.$$

The posterior versions are those in Lemma A.1. Write $m = \sum_l a_l m_l$ and $v = \sum_l a_l v_l$. The learned predictors $m_{l,\theta}$ and routing weights $a_{l,\theta}$ approximate m_l and a_l, respectively, and define

$$v_{l,\theta}(x,t) = \frac{x - m_{l,\theta}(x,t)}{t}, \qquad \bar{m}_\theta = \sum_l a_{l,\theta} m_{l,\theta}, \qquad v_\theta = \sum_l a_{l,\theta} v_{l,\theta}.$$

Fix $\delta \in (0,1)$ and θ, and set $T = 1 - \delta$. For $N \geq 1$ steps, let $h = T/N$, $\tau_k = kh$, and $t_k = 1 - kh$, $0 \leq k \leq N$. The sampler defined below has output law $q_{N,\delta}^\theta = \mathrm{Law}(Y_N)$ at $t_N = \delta$.

Assumptions recalled. The four assumptions have the following content, with all learned functions defined on $\mathbb{R}^d \times [\delta, 1]$. *(i) Compact support (Assumption 4.1):* $\mathrm{supp}(p_0^l) \subseteq \overline{B}(0,R)$ for every l, for some $R > 0$. *(ii) Bounded predictions and normalized routing (Assumption 4.2):* $m_{l,\theta}$ and $a_{l,\theta}$ are Borel measurable, with $\|m_{l,\theta}(x,t)\| \leq R$, $a_{l,\theta}(x,t) \geq 0$, and $\sum_l a_{l,\theta}(x,t) = 1$. *(iii) Composite-field regularity (Assumption 4.3):* for finite $\Lambda, \Gamma \geq 0$, all $x, y \in \mathbb{R}^d$, and $s, t \in [\delta, 1]$,

$$\|v_\theta(x,t) - v_\theta(y,t)\| \leq \Lambda \|x - y\|, \qquad \|v_\theta(x,t) - v_\theta(x,s)\| \leq \Gamma(1 + \|x\|)|t - s|.$$

(iv) Population approximation errors (Assumption 4.4): for finite $\epsilon_v, \epsilon_a \geq 0$,

$$\int_\delta^1 \sum_l \pi_l \|v_{l,\theta}(\cdot,t) - v_l(\cdot,t)\|^2_{L^2(p_t^l)}\, dt \leq \epsilon_v^2, \qquad \int_\delta^1 \sum_l \|a_{l,\theta}(\cdot,t) - a_l(\cdot,t)\|^2_{L^2(p_t)}\, dt \leq \epsilon_a^2.$$

These time integrals are unnormalized.

Reverse time and the sampler. Let $\tau = 1 - t$ and set

$$\begin{aligned} \widetilde{v}(x,\tau) &= -v(x, 1-\tau), \\ f(x,\tau) = \widetilde{v}_\theta(x,\tau) &= -v_\theta(x, 1-\tau), \\ \widetilde{v}_{l,\theta}(x,\tau) &= -v_{l,\theta}(x, 1-\tau). \end{aligned} \tag{35}$$

For any integer $N \geq 1$, define

$$h = T/N, \qquad \tau_k = kh, \qquad t_k = 1 - \tau_k, \qquad 0 \leq k \leq N. \tag{36}$$

To separate the errors due to network approximation, time discretization, and stochastic expert routing, we introduce four processes: the true reverse flow X, the learned composite flow Z, its deterministic Euler approximation $\widehat{Y}$, and the decentralized sampler Y. We couple them through a common Gaussian initial condition $\zeta \sim \mathcal{N}(0, I_d)$:

$$\dot{X}(\tau) = \widetilde{v}(X(\tau), \tau), \qquad X(0) = \zeta, \tag{37}$$

$$\dot{Z}(\tau) = f(Z(\tau), \tau), \qquad Z(0) = \zeta, \tag{38}$$

$$\widehat{Y}_{k+1} = \widehat{Y}_k + h f(\widehat{Y}_k, \tau_k), \qquad \widehat{Y}_0 = \zeta. \tag{39}$$

For the decentralized sampling process, let

$$\mathcal{F}_k = \sigma(\zeta, J_0, \dots, J_{k-1}), \qquad \mathbb{P}(J_k = l \mid \mathcal{F}_k) = a_{l,\theta}(Y_k, t_k), \tag{40}$$

and set $Y_0 = \zeta$ and

$$\begin{aligned} Y_{k+1} &= Y_k + h\widetilde{v}_{J_k,\theta}(Y_k, \tau_k) \\ &= \left(1 - \frac{h}{t_k}\right) Y_k + \frac{h}{t_k} m_{J_k,\theta}(Y_k, t_k). \end{aligned} \tag{41}$$

Write $q^{\theta}_{N,\delta} = \mathrm{Law}(Y_N)$. The interpolation variable x_t in equation 3 defines reference marginals; it is not an ODE trajectory. It may be constructed separately from ζ and the sampling labels.

Constants. We use the following explicit, non-optimized coefficients:

$$\begin{aligned} &M = \max\{R, \sqrt{d}\}, \quad B_\delta = \frac{\Lambda(M+R)}{\delta} + \Gamma(1+M), \\ &A_v = e^{\Lambda T}\sqrt{T}, \qquad A_a = e^{\Lambda T}\frac{R\sqrt{LT}}{\delta}, \\ &A_E = \frac{T^2 e^{\Lambda T}}{2} B_\delta, \qquad A_S = \frac{2RTe^{\Lambda T}}{\delta}, \qquad A_{\mathrm{disc}} = A_E + A_S. \end{aligned} \tag{42}$$

The refined and combined terminal error bounds are

$$\mathcal{E}^{\mathrm{ref}}_{N,\delta,\theta} = A_v\epsilon_v + A_a\epsilon_a + \frac{A_E}{N} + \frac{A_S}{\sqrt{N}}, \tag{43}$$

$$\mathcal{E}_{N,\delta,\theta} = A_v\epsilon_v + A_a\epsilon_a + \frac{A_{\mathrm{disc}}}{\sqrt{N}}. \tag{44}$$

Since $N^{-1} \le N^{-1/2}$ for $N \ge 1$, the refined bound never exceeds the combined bound.

A.2 Posterior Identities and the True Flow

Lemma A.1 (Gaussian smoothing and posterior means)**.** *Under Assumption 4.1, for every $t \in (0, 1]$, the measures p^l_t, p_t have strictly positive smooth densities. With*

$$g_t(x, y) = \frac{1}{(2\pi t^2)^{d/2}} \exp\left(-\frac{\|x - (1-t)y\|^2}{2t^2}\right), \tag{45}$$

these densities and versions of the posterior means are

$$p^l_t(x) = \int g_t(x, y) p^l_0(dy), \qquad p_t(x) = \int g_t(x, y) p_0(dy) = \sum_l \pi_l p^l_t(x), \tag{46}$$

$$m_l(x, t) = \frac{\int y g_t(x, y) p^l_0(dy)}{p^l_t(x)}, \qquad m(x, t) = \frac{\int y g_t(x, y) p_0(dy)}{p_t(x)}. \tag{47}$$

They are defined for every x and satisfy $\|m_l(x, t)\|, \|m(x, t)\| \le R$. The kernel integrals can be differentiated locally in space and time for $t > 0$.

Proof. Conditionally on $x_0 = y$, the variable x_t is Gaussian with mean $(1-t)y$ and covariance $t^2 I_d$. Integrating this conditional density gives equation 46, by Tonelli's theorem. The kernel is strictly positive, so each density is strictly positive.

For any compact set of (x, t) with t bounded away from zero, every fixed-order derivative of $g_t(x, y)$ is bounded uniformly over that set and $\|y\| \le R$. The same holds after multiplication by y or a fixed polynomial in y. These bounds are integrable against each probability measure p^l_0 and p_0. Dominated convergence therefore justifies each local differentiation under the integral. The kernel formula also extends to all $t > 0$, so derivatives at $t = 1$ can be understood through this extension or as one-sided derivatives on the interpolation interval.

For fixed (x, t), the probability kernel

$$\rho_{x,t}(dy) = \frac{g_t(x, y) p_0(dy)}{p_t(x)}$$

is a version of the conditional distribution of x_0 given $x_t = x$. Indeed, multiplication by $p_t(x)$ and integration in x verifies the defining conditional-expectation identity for all bounded Borel test functions. The component case is identical with p_0^l. Taking the mean gives equation 47. Each posterior is supported in $\overline{B}(0, R)$, and convexity of the norm yields the asserted bounds. □

Lemma A.2 (Bayes identities and velocity decomposition)**.** *Under Assumption 4.1, for every $x \in \mathbb{R}^d$ and $t \in (0, 1]$,*

$$a_l \geq 0, \qquad \sum_l a_l = 1, \qquad a_l(x,t)p_t(x) = \pi_l p_t^l(x), \tag{48}$$

and

$$m(x,t) = \sum_l a_l(x,t)m_l(x,t), \qquad v(x,t) = \sum_l a_l(x,t)v_l(x,t). \tag{49}$$

For every nonnegative Borel function F_l,

$$\int F_l(x)a_l(x,t)p_t(dx) = \pi_l \int F_l(x)p_t^l(dx). \tag{50}$$

Moreover, the fields in equation 6 are versions of the conditional interpolation velocities:

$$v_l(x,t) = \mathbb{E}[\epsilon - x_0 \mid x_t = x, C = l], \qquad v(x,t) = \mathbb{E}[\epsilon - x_0 \mid x_t = x].$$

Proof. Positivity and normalization follow from equation 46 and the definition of a_l. Multiplying that definition by p_t gives equation 48, and integration proves equation 50, including when the integrals are infinite. The numerator defining m is the π_l-weighted sum of the component numerators. Dividing by p_t proves the first identity in equation 49; substituting the velocity definitions and using $\sum_l a_l = 1$ proves the second. Finally,

$$\epsilon - x_0 = \frac{x_t - x_0}{t}.$$

Conditioning this identity gives the velocity formulas. Integrability follows from Assumption 4.1 and the finite Gaussian moments. □

Lemma A.3 (Growth and regularity of the true fields)**.** *Under Assumption 4.1, define $\Sigma(x,t) = \mathrm{Cov}(x_0 \mid x_t = x)$ using the posterior kernel in Lemma A.1. For every $x \in \mathbb{R}^d$ and $t \in [\delta, 1]$,*

$$\|\Sigma(x,t)\|_{\mathrm{op}} \leq R^2, \qquad \nabla_x m(x,t) = \frac{1-t}{t^2}\Sigma(x,t), \tag{51}$$

$$\|v(x,t)\| \leq \frac{\|x\| + R}{t}, \qquad \|\nabla_x v(x,t)\|_{\mathrm{op}} \leq \frac{1}{t} + \frac{(1-t)R^2}{t^3}, \tag{52}$$

and

$$\|\partial_t v(x,t)\| \leq \frac{\|x\| + R}{t^2} + \frac{2R^2\|x\| + 2R^3}{t^4}. \tag{53}$$

Consequently, the constants

$$\Lambda_*^\delta = \delta^{-1} + R^2\delta^{-3}, \qquad \Gamma_*^\delta = (1+R)\delta^{-2} + (2R^2 + 2R^3)\delta^{-4} \tag{54}$$

satisfy

$$\|v(x,t) - v(y,t)\| \leq \Lambda_*^\delta \|x - y\|, \tag{55}$$

$$\|v(x,t) - v(x,s)\| \leq \Gamma_*^\delta(1 + \|x\|)|t - s|. \tag{56}$$

The same estimates hold for each v_l and for the corresponding reverse-time fields on $[0, T]$.

Proof. Write $\mathbb{E}_{x,t}$ for expectation under $\rho_{x,t}$ and let y denote its coordinate variable. For any unit vector u,

$$u^\top \Sigma u = \mathrm{Var}_{x,t}(u^\top y) \leq \mathbb{E}_{x,t}(u^\top y)^2 \leq R^2.$$

Taking the supremum over unit vectors proves the covariance bound. Differentiation of a normalized kernel integral gives

$$\partial_{x_j} m_i = \mathrm{Cov}_{x,t}\big(y_i, \partial_{x_j} \log g_t(x,y)\big).$$

Since $\nabla_x \log g_t(x,y) = -x/t^2 + (1-t)y/t^2$, the covariance with the constant term vanishes, proving equation 51. The growth bound follows from $\|m\| \le R$; differentiating $v = (x-m)/t$ then proves equation 52.

For the time derivative, direct differentiation of equation 45 gives

$$\partial_t \log g_t(x,y) = -\frac{d}{t} + \frac{\|x\|^2 - (2-t)\langle x,y\rangle + (1-t)\|y\|^2}{t^3}. \tag{57}$$

The same normalized-integral rule yields

$$\partial_t m(x,t) = -\frac{2-t}{t^3}\Sigma(x,t)x + \frac{1-t}{t^3}\,\mathrm{Cov}_{x,t}(y, \|y\|^2). \tag{58}$$

Here the vector covariance is $\mathbb{E}_{x,t}[(y-m)\|y\|^2]$, because $\mathbb{E}_{x,t}(y-m) = 0$. Hence

$$\big\|\mathrm{Cov}_{x,t}(y,\|y\|^2)\big\| \le \mathbb{E}_{x,t}\big[\|y-m\|\|y\|^2\big] \le 2R^3.$$

Using $\|\Sigma\|_{\mathrm{op}} \le R^2$, $2-t \le 2$, and $1-t \le 1$, we obtain

$$\|\partial_t m(x,t)\| \le \frac{2R^2\|x\| + 2R^3}{t^3}.$$

The identity $\partial_t v = -(x-m)/t^2 - (\partial_t m)/t$ now proves equation 53. Integrating the spatial derivative along the segment from y to x, and the time derivative from s to t, gives the stated Lipschitz bounds with equation 54. The component proofs use the same support bound. Finally, replacing t by $1-\tau$ and negating the field preserves the norms and the absolute time difference. □

Proposition A.4 (The reverse true flow and its marginals)**.** *Under Assumption 4.1, the reverse ODE in equation 37 has a unique solution on $[0,T]$ for every deterministic initial value. Its flow maps are measurable in the initial value. For $X(0) = \zeta \sim \mathcal{N}(0, I_d)$,*

$$\mathrm{Law}(X(\tau)) = p_{1-\tau} \quad (0 \le \tau \le T). \tag{59}$$

The pathwise bound holds almost surely, simultaneously for all $\tau \in [0,T]$, and implies the following uniform L^2 bound:

$$\|X(\tau)\| \le (1-\tau)\|\zeta\| + \tau R, \qquad \sup_{0\le\tau\le T} \|X(\tau)\|_2 \le M. \tag{60}$$

Proof. Lemma A.3 gives continuity in time, a uniform global spatial Lipschitz bound, and at most linear growth for $\widetilde{v}$ on the truncated interval. The Picard–Lindelöf theorem gives local existence and uniqueness. The growth bound and Gronwall's inequality prevent explosion on the finite interval, yielding global solutions there. Applied to the difference of two solutions, Gronwall also gives continuous dependence on the initial value, and hence measurability of the flow maps.

The reverse drift has the form $(m(X(\tau), 1-\tau) - X(\tau))/(1-\tau)$. Differentiating $X(\tau)/(1-\tau)$ and integrating gives

$$X(\tau) = (1-\tau)\zeta + (1-\tau)\int_0^\tau \frac{m(X(s), 1-s)}{(1-s)^2}\,ds. \tag{61}$$

Since $\|m\| \le R$ and $(1-\tau)\int_0^\tau (1-s)^{-2} ds = \tau$, the pathwise bound follows. Minkowski's inequality and $\|\zeta\|_2 = \sqrt{d}$ give $\|X(\tau)\|_2 \le (1-\tau)\sqrt{d} + \tau R \le M$.

We next identify the marginals. For $\varphi \in C_c^\infty(\mathbb{R}^d)$, differentiating the expectation along the interpolation is justified by the integrable bound $\|\nabla\varphi\|_\infty(\|\epsilon\| + R)$. Thus

$$\frac{d}{dt}\int \varphi(x)p_t(dx) = \mathbb{E}\langle \nabla\varphi(x_t), \epsilon - x_0\rangle \tag{62}$$

$$= \int \langle \nabla\varphi(x), v(x,t)\rangle p_t(dx). \tag{63}$$

In reverse time, $\mu_\tau = p_{1-\tau}$ therefore satisfies

$$\frac{d}{d\tau}\int \varphi(x)\mu_\tau(dx) = \int \langle \nabla\varphi(x), \widetilde{v}(x,\tau)\rangle \mu_\tau(dx), \qquad \mu_0 = \mathcal{N}(0, I_d). \tag{64}$$

The curve μ_τ is weakly continuous: this follows directly from the almost sure continuity of $x_{1-\tau}$ and dominated convergence for bounded continuous test functions.

For completeness, the characteristic argument identifies this solution of the continuity equation. Let $\Phi_{s,r}(x)$ denote the position at time r of the reverse ODE started from x at time s. Solutions exist both forward and backward between any two times in $[0,T]$. Smoothness of the true field implies differentiability of these flow maps. Fix $r \in [0,T]$ and $\varphi \in C_c^\infty$ and put $u(s,x) = \varphi(\Phi_{s,r}(x))$ for $0 \le s \le r$. The flow identity implies

$$\partial_s u(s,x) + \langle \widetilde{v}(x,s), \nabla_x u(s,x) \rangle = 0, \qquad u(r,x) = \varphi(x).$$

The function u has compact spatial support uniformly in s: if $\Phi_{s,r}(x)$ lies in a fixed compact set, the linear-growth estimate applied to the backward ODE bounds $\|x\|$ uniformly over $s \in [0,r]$. The weak identity equation 64 also applies to this time-dependent test function. This follows by a time partition and the product rule; compact support and continuity of the derivatives justify the limit. Consequently,

$$\frac{d}{ds} \int u(s,x) \mu_s(dx) = 0.$$

Integrating from 0 to r proves

$$\int \varphi(x) \mu_r(dx) = \int \varphi(\Phi_{0,r}(z)) \mu_0(dz).$$

These test functions determine the measure, so $\mu_r = (\Phi_{0,r})_\# \mu_0 = \mathrm{Law}(X(r))$. This proves equation 59. The argument identifies one-time marginals and does not identify the interpolation paths with ODE trajectories. □

A.3 Learned Processes and Moment Control

Lemma A.5 (Well-posedness and uniform second moments)**.** *Under Assumptions 4.2–4.3, the learned ODE has a unique solution on $[0,T]$ for each deterministic initial value. The two discrete processes are well defined and $\widehat{Y}_k, Y_k$ are $\mathcal{F}_k$-measurable. With the common Gaussian initialization, almost surely*

$$\|Z(\tau)\| \le (1-\tau)\|\zeta\| + \tau R \qquad (0 \le \tau \le T), \tag{65}$$

$$\|\widehat{Y}_k\|,\ \|Y_k\| \le t_k \|\zeta\| + (1-t_k) R \qquad (0 \le k \le N). \tag{66}$$

In particular,

$$\sup_{0 \le \tau \le T} \|Z(\tau)\|_2 \le M, \qquad \max_{0 \le k \le N} \{\|\widehat{Y}_k\|_2, \|Y_k\|_2\} \le M. \tag{67}$$

The discrete assertions require only Assumption 4.2 and the specified sampling rule. All these bounds hold for every integer $N \ge 1$.

Proof. Assumption 4.3 implies joint continuity of f: for convergent (x_n, τ_n), split its increment into a spatial increment and a time increment at the fixed limiting position. It also gives a uniform spatial Lipschitz constant Λ. Assumption 4.2 gives the linear-growth bound

$$\|f(x,\tau)\| \le \frac{\|x\| + R}{1-\tau} \le \frac{\|x\| + R}{\delta}. \tag{68}$$

Existence, uniqueness, global continuation on $[0,T]$, and measurable dependence on the initial value follow as in Proposition A.4. With $b(x,\tau) = \bar{m}_\theta(x, 1-\tau)$, the integrating-factor identity is

$$Z(\tau) = (1-\tau)\zeta + (1-\tau) \int_0^\tau \frac{b(Z(s),s)}{(1-s)^2}\, ds.$$

Since $\|b\| \le R$, this proves equation 65.

The sampling rule can be realized on a probability space carrying ζ and independent uniform random variables $U_0, \dots, U_{N-1}$ on $[0,1)$, independent of ζ. Recursively set

$$J_k = \min \left\{ l : U_k < \sum_{j=1}^{l} a_{j,\theta}(Y_k, t_k) \right\}.$$

The set is nonempty because its final cumulative probability is one. Borel measurability gives a measurable recursion. The fresh uniform variable is independent of $\mathcal{F}_k$, so this construction satisfies equation 40. Induction shows that Y_k is $\mathcal{F}_k$-measurable. The mean Euler process is a measurable function of ζ, and is therefore also $\mathcal{F}_k$-measurable.

For $k < N$,

$$t_k = \delta + (N-k)h \geq \delta + h > h.$$

Both updates are convex combinations of the current state and a vector of norm at most R; for the mean Euler update this vector is $\bar{m}_\theta(\widehat{Y}_k, t_k)$. For either process V_k, suppose $\|V_k\| \leq t_k\|\zeta\| + (1-t_k)R$. Then

$$\begin{aligned}\|V_{k+1}\| &\leq \left(1 - \frac{h}{t_k}\right)\left[t_k\|\zeta\| + (1-t_k)R\right] + \frac{h}{t_k}R \\ &= t_{k+1}\|\zeta\| + (1-t_{k+1})R.\end{aligned}$$

The initial inequality is equality, proving equation 66 by induction. Taking L^2 norms proves equation 67. The supremum in that statement is outside the expectation. □

A.4 From Population Errors to Flow Error

Lemma A.6 (Approximation of the composite field)**.** *Under Assumptions 4.1–4.2, define, for $t \in [\delta, 1]$,*

$$\eta_v(t)^2 = \sum_l \pi_l \int \|v_{l,\theta}(x,t) - v_l(x,t)\|^2 p_t^l(dx), \tag{69}$$

$$\eta_a(t)^2 = \int \sum_l |a_{l,\theta}(x,t) - a_l(x,t)|^2 p_t(dx). \tag{70}$$

Then

$$\|v_\theta(\cdot,t) - v(\cdot,t)\|_{L^2(p_t)} \leq \eta_v(t) + \frac{R\sqrt{L}}{t}\eta_a(t). \tag{71}$$

If Assumption 4.4 also holds, then

$$\int_\delta^1 \|v_\theta(\cdot,t) - v(\cdot,t)\|_{L^2(p_t)}\, dt \leq \sqrt{T}\left(\epsilon_v + \frac{R\sqrt{L}}{\delta}\epsilon_a\right). \tag{72}$$

Proof. Adding and subtracting $\sum_l a_l v_{l,\theta}$ gives

$$v_\theta - v = \sum_l a_l(v_{l,\theta} - v_l) + \sum_l (a_{l,\theta} - a_l)v_{l,\theta}.$$

Both routing vectors sum to one. Substituting $v_{l,\theta} = (x - m_{l,\theta})/t$ therefore cancels the common drift:

$$v_\theta - v = \underbrace{\sum_l a_l(v_{l,\theta} - v_l)}_{U(x,t)} - \underbrace{\frac{1}{t}\sum_l (a_{l,\theta} - a_l)m_{l,\theta}}_{V(x,t)}. \tag{73}$$

By the weighted Cauchy–Schwarz inequality and equation 50,

$$\begin{aligned}\|U(\cdot,t)\|^2_{L^2(p_t)} &\leq \sum_l \int a_l(x,t)\|v_{l,\theta}(x,t) - v_l(x,t)\|^2 p_t(dx) \\ &= \sum_l \pi_l \int \|v_{l,\theta}(x,t) - v_l(x,t)\|^2 p_t^l(dx) = \eta_v(t)^2.\end{aligned}$$

For the other term, Assumption 4.2 and Cauchy–Schwarz over the expert index give

$$\|V(x,t)\| \leq \frac{R}{t}\sum_l |a_{l,\theta}(x,t) - a_l(x,t)| \leq \frac{R\sqrt{L}}{t}\left(\sum_l |a_{l,\theta}(x,t) - a_l(x,t)|^2\right)^{1/2}.$$

The triangle inequality in $L^2(p_t)$ proves equation 71. All these quantities are measurable in t by the Borel kernel formulas and Tonelli's theorem. They are finite: Assumptions 4.1–4.2 imply $\|v_{l,\theta} - v_l\| \le 2R/t$, and the routing vectors are bounded. Finally, use $t^{-1} \le \delta^{-1}$ and Cauchy–Schwarz in time:

$$\int_\delta^1 \eta_v(t)\,dt \le \sqrt{T}\,\epsilon_v, \qquad \int_\delta^1 \eta_a(t)\,dt \le \sqrt{T}\,\epsilon_a.$$

This proves equation 72. □

Proposition A.7 (Comparison of the true and learned flows)**.** *Under Assumptions 4.1–4.4, with the shared Gaussian initialization,*

$$\begin{aligned}\|X(T) - Z(T)\|_2 &\le e^{\Lambda T}\int_\delta^1 \|v_\theta(\cdot,t) - v(\cdot,t)\|_{L^2(p_t)}\,dt \\ &\le A_v\epsilon_v + A_a\epsilon_a.\end{aligned} \tag{74}$$

Proof. The processes are well defined by Proposition A.4 and Lemma A.5. Let $r(\tau) = \widetilde{v}(X(\tau),\tau) - f(X(\tau),\tau)$. Subtracting the two integral equations yields

$$X(\tau) - Z(\tau) = \int_0^\tau r(s)\,ds + \int_0^\tau \big[f(X(s),s) - f(Z(s),s)\big]\,ds.$$

All terms are integrable in L^2, by the growth and moment bounds. With $e(\tau) = \|X(\tau) - Z(\tau)\|_2$, Minkowski's inequality and Assumption 4.3 give

$$e(\tau) \le \int_0^\tau \|r(s)\|_2\,ds + \Lambda\int_0^\tau e(s)\,ds.$$

The functions involved are integrable, and e is continuous. The integral form of Gronwall's inequality gives

$$e(T) \le \int_0^T e^{\Lambda(T-s)}\|r(s)\|_2\,ds \le e^{\Lambda T}\int_0^T \|r(s)\|_2\,ds.$$

By equation 59,

$$\|r(s)\|_2 = \|v_\theta(\cdot,1-s) - v(\cdot,1-s)\|_{L^2(p_{1-s})}.$$

Changing variables $t = 1 - s$ and applying Lemma A.6 proves the result with the constants in equation 42. The approximation error is evaluated along the true flow. Once the processes are well defined, this comparison uses only the spatial constant Λ from Assumption 4.3. □

A.5 Deterministic and Random Euler Errors

Proposition A.8 (Euler error for the learned composite field)**.** *Under Assumptions 4.2–4.3, define, for $\tau \in [\tau_k, \tau_{k+1}]$,*

$$\widehat{Y}(\tau) = \widehat{Y}_k + (\tau - \tau_k)f(\widehat{Y}_k,\tau_k). \tag{75}$$

The local freezing error satisfies

$$\|f(\widehat{Y}(\tau),\tau) - f(\widehat{Y}_k,\tau_k)\|_2 \le B_\delta(\tau - \tau_k). \tag{76}$$

For every integer $N \ge 1$,

$$\|Z(T) - \widehat{Y}_N\|_2 \le \frac{Th}{2}e^{\Lambda T}B_\delta = \frac{A_E}{N}. \tag{77}$$

Proof. The interpolant is continuous at grid points by the Euler recursion. The moment and growth estimates give

$$\|f(\widehat{Y}_k,\tau_k)\|_2 \le \frac{M+R}{\delta}, \qquad \big(\mathbb{E}(1+\|\widehat{Y}_k\|)^2\big)^{1/2} \le 1 + M.$$

Assumption 4.3 is preserved by the reverse-time transformation. Splitting the field increment at $(\widehat{Y}_k, \tau)$ thus yields

$$\begin{aligned}
&\|f(\widehat{Y}(\tau), \tau) - f(\widehat{Y}_k, \tau_k)\|_2 \\
&\quad \le \Lambda \|\widehat{Y}(\tau) - \widehat{Y}_k\|_2 + \Gamma(1+M)(\tau - \tau_k) \\
&\quad \le \left[\frac{\Lambda(M+R)}{\delta} + \Gamma(1+M)\right](\tau - \tau_k),
\end{aligned}$$

which is equation 76.

Let $\kappa(s) = \tau_k$ for $s \in [\tau_k, \tau_{k+1})$. The interpolant satisfies $\widehat{Y}(\tau) = \zeta + \int_0^\tau f(\widehat{Y}(\kappa(s)), \kappa(s))\, ds$. Subtracting this identity from the integral equation for Z, and using the spatial Lipschitz condition, gives

$$\|Z(\tau) - \widehat{Y}(\tau)\|_2 \le \Lambda \int_0^\tau \|Z(s) - \widehat{Y}(s)\|_2\, ds + B_\delta \int_0^\tau (s - \kappa(s))\, ds.$$

The endpoint value assigned to $\kappa(T)$ is immaterial. Gronwall's inequality and equation 76 imply

$$\begin{aligned}
\|Z(T) - \widehat{Y}_N\|_2 &\le e^{\Lambda T} B_\delta \sum_{k=0}^{N-1} \int_{\tau_k}^{\tau_{k+1}} (s - \tau_k)\, ds \\
&= \frac{e^{\Lambda T} B_\delta N h^2}{2} = \frac{e^{\Lambda T} B_\delta T^2}{2N}.
\end{aligned}$$

The argument uses neither differentiability of the learned networks nor second derivatives of their composite field. □

Proposition A.9 (Mean-square error from stochastic expert routing)**.** *Suppose Assumption 4.2 holds, together with the spatial condition equation 14 and the sampling rule equation 40. Define*

$$D_k = Y_k - \widehat{Y}_k, \qquad \xi_k = \widetilde{v}_{J_k,\theta}(Y_k, \tau_k) - f(Y_k, \tau_k). \tag{78}$$

Then, for $k < N$,

$$\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0, \qquad \mathbb{E}[\|\xi_k\|^2 \mid \mathcal{F}_k] \le \frac{4R^2}{t_k^2} \le \frac{4R^2}{\delta^2}, \tag{79}$$

and

$$\mathbb{E}[\|D_{k+1}\|^2 \mid \mathcal{F}_k] \le (1 + \Lambda h)^2 \|D_k\|^2 + \frac{4R^2 h^2}{\delta^2}. \tag{80}$$

In particular,

$$\mathbb{E}\|Y_N - \widehat{Y}_N\|^2 \le \frac{4R^2 T^2 e^{2\Lambda T}}{\delta^2 N} = \frac{A_S^2}{N}. \tag{81}$$

Neither Assumption 4.4 nor the time regularity condition is required.

Proof. All discrete variables have finite second moments by the discrete part of Lemma A.5. Conditionally on $\mathcal{F}_k$, the position Y_k is fixed and the law of J_k is the routing vector at (Y_k, t_k). Hence

$$\mathbb{E}[\widetilde{v}_{J_k,\theta}(Y_k, \tau_k) \mid \mathcal{F}_k] = \sum_l a_{l,\theta}(Y_k, t_k) \widetilde{v}_{l,\theta}(Y_k, \tau_k) = f(Y_k, \tau_k).$$

Moreover, cancellation of the common state-dependent drift gives

$$\xi_k = \frac{m_{J_k,\theta}(Y_k, t_k) - \bar{m}_\theta(Y_k, t_k)}{t_k}. \tag{82}$$

Both data predictions have norm at most R, so $\|\xi_k\| \le 2R/t_k$, proving equation 79.

Subtracting the two Euler recursions gives

$$D_{k+1} = D_k + h\big[f(Y_k, \tau_k) - f(\widehat{Y}_k, \tau_k)\big] + h\xi_k. \tag{83}$$

Let G_k denote the sum of the first two terms on the right. It is $\mathcal{F}_k$-measurable and $\|G_k\| \le (1+\Lambda h)\|D_k\|$. Conditional centering gives

$$\mathbb{E}[\|D_{k+1}\|^2 \mid \mathcal{F}_k] = \|G_k\|^2 + 2h\langle G_k, \mathbb{E}[\xi_k \mid \mathcal{F}_k]\rangle + h^2\mathbb{E}[\|\xi_k\|^2 \mid \mathcal{F}_k]$$
$$\le (1+\Lambda h)^2\|D_k\|^2 + \frac{4R^2h^2}{\delta^2}.$$

Thus the vanishing cross term follows from conditional expectation; the labels are not assumed to be unconditionally independent.

Taking expectations and iterating from $D_0 = 0$ yields

$$\mathbb{E}\|D_N\|^2 \le \frac{4R^2h^2}{\delta^2}\sum_{j=0}^{N-1}(1+\Lambda h)^{2j}.$$

For every $0 \le j < N$, $(1+\Lambda h)^{2j} \le e^{2\Lambda hj} \le e^{2\Lambda T}$. Since $Nh^2 = T^2/N$, this proves equation 81, also when $\Lambda = 0$. □

A.6 Terminal Coupling and Total Distributional Error

Proposition A.10 (Combined terminal strong error)**.** *Under Assumptions 4.1–4.4 and the prescribed common initialization and sampling rule, for every integer $N \ge 1$,*

$$\mathbb{E}\|Y_N - X(T)\|^2 \le \left(A_v\epsilon_v + A_a\epsilon_a + \frac{A_E}{N} + \frac{A_S}{\sqrt{N}}\right)^2 = \left(\mathcal{E}^{\mathrm{ref}}_{N,\delta,\theta}\right)^2 \le \mathcal{E}^2_{N,\delta,\theta}. \tag{84}$$

The expectation includes the Gaussian initialization and the random expert selections.

Proof. All four processes are constructed on the common probability space described above. The triangle inequality in L^2 gives

$$\|Y_N - X(T)\|_2 \le \|X(T) - Z(T)\|_2 + \|Z(T) - \widehat{Y}_N\|_2 + \|\widehat{Y}_N - Y_N\|_2$$
$$\le A_v\epsilon_v + A_a\epsilon_a + \frac{A_E}{N} + \frac{A_S}{\sqrt{N}},$$

where the last line uses Propositions A.7, A.8, and A.9. Squaring the nonnegative quantities and using $N^{-1} \le N^{-1/2}$ proves the claim. This is a terminal mean-square estimate for the specified coupling. □

Proposition A.11 (Distributional error and data-end truncation)**.** *Under Assumption 4.1 and the interpolation definition,*

$$W_2(p_\delta, p_0) \le \delta\sqrt{\mathbb{E}\|x_0\|^2 + d} \le \delta\sqrt{R^2 + d}. \tag{85}$$

Under Assumptions 4.1–4.4, the prescribed sampler satisfies

$$W_2(q^\theta_{N,\delta}, p_\delta) \le \mathcal{E}^{\mathrm{ref}}_{N,\delta,\theta} \le \mathcal{E}_{N,\delta,\theta}, \tag{86}$$
$$W_2(q^\theta_{N,\delta}, p_0) \le \delta\sqrt{R^2 + d} + \mathcal{E}^{\mathrm{ref}}_{N,\delta,\theta} \tag{87}$$
$$\le \delta\sqrt{R^2 + d} + A_v\epsilon_v + A_a\epsilon_a + \frac{A_{\mathrm{disc}}}{\sqrt{N}}. \tag{88}$$

The coefficients are given by equation 42. For fixed $\delta, R, d, L, \Lambda, \Gamma$, they are independent of N.

Proof. The pair (x_δ, x_0) from equation 3 is a coupling of p_δ and p_0, and $x_\delta - x_0 = \delta(\epsilon - x_0)$. Independence and $\mathbb{E}\epsilon = 0$ imply

$$\mathbb{E}\|\epsilon - x_0\|^2 = \mathbb{E}\|\epsilon\|^2 + \mathbb{E}\|x_0\|^2 - 2\mathbb{E}\langle\epsilon, x_0\rangle = d + \mathbb{E}\|x_0\|^2 \le d + R^2.$$

The coupling bound equation 1 proves equation 85.

By Proposition A.4, $X(T)$ has law p_δ. The pair $(Y_N, X(T))$ therefore provides a coupling with finite second moment, by Lemma A.5. Proposition A.10 gives

$$W_2(q^\theta_{N,\delta}, p_\delta) \le \|Y_N - X(T)\|_2 \le \mathcal{E}^{\mathrm{ref}}_{N,\delta,\theta}.$$

The triangle inequality for W_2, followed by equation 85 and $\mathcal{E}^{\mathrm{ref}} \le \mathcal{E}$, proves the remaining claims. □

Fully expanded coefficients. Substituting $T = 1 - \delta$ and $M = \max\{R, \sqrt{d}\}$ gives the refined bound

$$\begin{aligned} W_2(q^{\theta}_{N,\delta}, p_0) \le{}& \delta\sqrt{R^2 + d} \\ &+ e^{\Lambda(1-\delta)}\Bigg[\sqrt{1-\delta}\,\epsilon_v + \frac{R\sqrt{L(1-\delta)}}{\delta}\epsilon_a \\ &\quad + \frac{(1-\delta)^2}{2N}\left\{\frac{\Lambda(\max\{R,\sqrt{d}\}+R)}{\delta} + \Gamma(1+\max\{R,\sqrt{d}\})\right\} \\ &\quad + \frac{2R(1-\delta)}{\delta\sqrt{N}}\Bigg]. \end{aligned} \tag{89}$$

For the form with a single N-dependent term, the full expression is

$$\begin{aligned} W_2(q^{\theta}_{N,\delta}, p_0) \le{}& \delta\sqrt{R^2 + d} \\ &+ e^{\Lambda(1-\delta)}\Bigg[\sqrt{1-\delta}\,\epsilon_v + \frac{R\sqrt{L(1-\delta)}}{\delta}\epsilon_a \\ &\quad + \frac{1}{\sqrt{N}}\left\{\frac{(1-\delta)^2}{2}\left(\frac{\Lambda(\max\{R,\sqrt{d}\}+R)}{\delta} + \Gamma(1+\max\{R,\sqrt{d}\})\right)\right. \\ &\qquad \left.+ \frac{2R(1-\delta)}{\delta}\right\}\Bigg]. \end{aligned} \tag{90}$$

Thus no unspecified stability constant is hidden in the conclusion. The explicit dependencies are $A_v = A_v(\delta, \Lambda)$, $A_a = A_a(\delta, R, L, \Lambda)$, and $A_{\text{disc}} = A_{\text{disc}}(\delta, R, d, \Lambda, \Gamma)$. The learned regularity parameters themselves may vary with the network, mixture, and truncation interval. The estimates above hold for each fixed choice of these quantities. They do not by themselves assert a joint rate as $\delta \to 0$ and the networks vary.

B Training Details and Experimental Results

We first explain how our DDIM implementation corresponds to the decentralized flow-matching sampler in the preliminaries and derive the posterior-routing training objective. We then present results on CIFAR-10 and CelebA-HQ.

B.1 High-Order Training Objectives

Our training framework uses the temporal evolution of posterior routing probabilities as an additional supervision signal.

B.1.1 DDIM Equivalence and Connection to the Main Analysis

We use the linear flow-matching notation of equation 3–equation 6:

$$x_t = (1-t)x_0 + t\epsilon, \qquad v_l(x,t) = \frac{x - m_l(x,t)}{t}, \qquad v(x,t) = \sum_{l=1}^{L} a_l(x,t)v_l(x,t). \tag{91}$$

Here a_l and m_l are the posterior probabilities and conditional data means in equation 4; generation uses $\tau = 1 - t$ and $\widetilde{v}(x,\tau) = -v(x, 1-\tau)$.

To relate the implementation to this formulation, write its noisy input as $z_s = \alpha_s x_0 + \sigma_s \epsilon$. For nonnegative α_s, σ_s, set

$$c_s = \alpha_s + \sigma_s > 0, \qquad t(s) = \frac{\sigma_s}{c_s}, \qquad \frac{z_s}{c_s} = (1 - t(s))x_0 + t(s)\epsilon. \tag{92}$$

Assume $t(s)$ is increasing. Thus the rescaled input is exactly the interpolation in the preliminaries. Conditioning on $z_s = z$ is equivalent to conditioning on $x_{t(s)} = z/c_s$, so the same posterior probabilities and data means define a_l and m_l. Their learned versions in these coordinates are $a_{l,\theta}$ and $m_{l,\theta}$ from equation 9.

For a selected expert, let $Y_k = Z_k/c_{s_k}$, $t_k = t(s_k)$, and $\widehat{m}_k = m_{J_k,\theta}(Y_k, t_k)$. The deterministic DDIM update (Song et al., 2021) is

$$Z_{k+1} = \alpha_{s_{k+1}}\widehat{m}_k + \sigma_{s_{k+1}}\frac{Z_k - \alpha_{s_k}\widehat{m}_k}{\sigma_{s_k}}. \tag{93}$$

Dividing by $c_{s_{k+1}}$ and setting $h_k = t_k - t_{k+1} > 0$ gives

$$\begin{aligned} Y_{k+1} &= (1 - t_{k+1})\widehat{m}_k + \frac{t_{k+1}}{t_k}\big(Y_k - (1 - t_k)\widehat{m}_k\big) \\ &= \left(1 - \frac{h_k}{t_k}\right) Y_k + \frac{h_k}{t_k} m_{J_k,\theta}(Y_k, t_k) = Y_k - h_k v_{J_k,\theta}(Y_k, t_k). \end{aligned} \tag{94}$$

This is exactly the decentralized update in equation 11. With the normalized routing law $\mathbb{P}(J_k = l \mid \mathcal{F}_k) = a_{l,\theta}(Y_k, t_k)$ from equation 10,

$$\mathbb{E}[Y_{k+1} - Y_k \mid \mathcal{F}_k] = -h_k \sum_l a_{l,\theta} v_{l,\theta}(Y_k, t_k) = -h_k v_\theta(Y_k, t_k), \tag{95}$$

where the routing weights are evaluated at (Y_k, t_k). Hence the implementation has the same learned composite field and zero-mean expert-selection fluctuations used in the main convergence proof. With $Y_0 = \zeta \sim \mathcal{N}(0, I_d)$ and $t_k = 1 - k(1 - \delta)/N$, its output law is precisely $q^\theta_{N,\delta}$ defined in the preliminaries. General DDIM grids instead induce nonuniform h_k; identifying the implementation with the theorem's sampler requires matching its grid and endpoints.

The training targets also agree. For a component-l training pair,

$$\|v_{l,\theta}(x_t, t) - (\epsilon - x_0)\|^2 = \frac{1}{t^2}\|m_{l,\theta}(x_t, t) - x_0\|^2. \tag{96}$$

Thus data-denoising regression is flow-matching regression with a time-dependent weight, with the same conditional-mean target m_l. For a noise-prediction expert $\widehat{\epsilon}_{l,\theta}(z_s, s)$, the corresponding data prediction is $(z_s - \sigma_s\widehat{\epsilon}_{l,\theta})/\alpha_s$ at $\alpha_s > 0$; noise MSE equals $(1 - t)^2$ times the velocity MSE above. Likewise, $\mathbb{E}[e_C \mid x_t, t] = a(x_t, t)$, so label MSE and cross-entropy target the same posterior router. These objectives learn the expert and router quantities entering ϵ_v and ϵ_a in the main analysis. The original time sampling and weights are retained; the convergence guarantee uses the population errors and regularity assumptions stated there.

B.1.2 Analytical Form of the Posterior Derivative

For the empirical mixture $p_0^l = M_l^{-1}\sum_{i=1}^{M_l}\delta_{y_i^l}$, define

$$\begin{aligned} V_i^l(x, t) &= \exp\left(-\frac{\|x - (1 - t)y_i^l\|^2}{2t^2}\right), \\ w_i^l(x, t) &= \frac{(\pi_l/M_l)V_i^l(x, t)}{\sum_{r=1}^{L}(\pi_r/M_r)\sum_{j=1}^{M_r} V_j^r(x, t)}, \qquad a_l = \sum_i w_i^l. \end{aligned} \tag{97}$$

Here M_l counts component data, while N counts sampling steps. The corresponding means are

$$u_i^l = \frac{w_i^l}{a_l}, \qquad m_l = \sum_i u_i^l y_i^l, \qquad m = \sum_l a_l m_l. \tag{98}$$

Write $a = (a_1, \ldots, a_L)^\top$ and let e_l be the one-hot component label. The derivative along the flow ODE is

$$\mathcal{D}_v f = \partial_t f + \nabla_x f\, v, \tag{99}$$

where $\nabla_x f$ is the Jacobian for vector-valued f. Define

$$\epsilon_i^l = \frac{x - (1 - t)y_i^l}{t}, \qquad G_i^l = -\frac{\langle \epsilon_i^l, y_i^l - m\rangle}{t^2}. \tag{100}$$

Using $v = (x - m)/t$, direct differentiation gives $\mathcal{D}_v \epsilon_i^l = (y_i^l - m)/t^2$ and $\mathcal{D}_v V_i^l = V_i^l G_i^l$. The quotient rule then yields $\mathcal{D}_v w_i^l = w_i^l (G_i^l - \sum_{r,j} w_j^r G_j^r)$, hence

$$\mathcal{D}_v a(x,t) = \sum_{l=1}^{L} \sum_{i=1}^{M_l} w_i^l(x,t) G_i^l(x,t) \big(e_l - a(x,t)\big). \tag{101}$$

This is the derivative along the mixture flow, including state motion, rather than along an individual interpolation path.

B.1.3 Regularization Loss

Let $d_\phi(x,t)$ estimate $\mathcal{D}_v a(x,t)$ and let $G(x,t,y)$ denote the expression for G_i^l with $y_i^l = y$. For training tuples $C \sim \text{Categorical}(\pi)$, $x_0 \mid C = l \sim p_0^l$, and independent Gaussian ϵ, equation 101 gives

$$\mathbb{E}[G(x_t,t,x_0)(e_C - a(x_t,t)) \mid x_t, t] = \mathcal{D}_v a(x_t,t). \tag{102}$$

This identity justifies derivative regression for the same posterior router as in the main model. Replacing m by $\bar{m}_\theta$ in G defines G_θ, and the implemented loss is expressed as

$$\mathcal{L}_{\text{reg}} = \mathbb{E}\Big[\omega(t) \, \|d_\phi(x_t,t) - G_\theta(x_t,t,x_0)(e_C - a_\theta(x_t,t))\|^2\Big], \qquad t \sim \rho, \tag{103}$$

where ρ is the time law induced by the implementation under equation 92. For an unweighted derivative loss in the original clock s, the chain rule gives

$$d_\phi(x,t(s)) = \frac{\widehat{d}_\phi(c_s x, s)}{t'(s)}, \qquad \omega(t(s)) = (t'(s))^2, \qquad t'(s) > 0. \tag{104}$$

The sample target scales by the same factor, preserving the original loss. Direct regression in t instead uses $\omega(t) = 1$. Learned posterior quantities make the target an approximation to equation 102.

Along the reverse flow $X(\tau)$ in equation 7,

$$\frac{d}{d\tau} a(X(\tau), 1-\tau) = -\mathcal{D}_v a(X(\tau), 1-\tau), \tag{105}$$

so generation uses the derivative estimate $-d_\phi(x, 1-\tau)$. The regularizer therefore supervises the evolution of the posterior routing probabilities used by the decentralized sampler.

B.2 Experimental Results on Image Datasets

We evaluate our method on two benchmark datasets: CIFAR-10 (Krizhevsky, 2009) and CelebA-HQ (Karras et al., 2017). CIFAR-10 contains 60,000 natural images across 10 object categories, while CelebA-HQ provides 30,000 high-resolution face images with attribute labels at 256×256 resolution.

B.2.1 Training Configuration

Due to numerical stability considerations during early training, we adopt a two-stage training strategy. Here j denotes the original discrete diffusion timestep, distinct from the linear-flow time t:

- **Stage 1 (0–50k steps):** Combined cross-entropy and MSE loss to stabilize category prediction
- **Stage 2 (50k steps onward):** MSE loss with the regularization term from Equation (103). Training continues until at least 300k total optimization steps have been completed. To address the heterogeneous distribution of temporal derivatives across the reverse process, we train three separate d_ϕ networks initialized with the same weights but trained independently on different time intervals:
  - **Network 1 ($j \in [0, 300]$):** Handles the low-noise, near-data regime
  - **Network 2 ($j \in [300, 700]$):** Captures the transition regime with rapidly varying temporal derivatives

Table 1: FID scores (↓) on CIFAR-10 and CelebA-HQ with varying number of sampling steps compared to the baseline(unconditional DDIM) method

Method	CIFAR-10				CelebA-HQ			
	10	25	50	100	10	25	50	100
DDIM (Uncond)	21.31	10.70	7.74	6.08	20.54	13.45	9.33	6.60
+ Naive	**15.97**	**8.83**	**6.54**	**5.27**	10.76	7.00	5.97	5.23
+ MoG-a net	17.93	9.63	7.35	6.21	11.45	7.41	6.23	5.45
+ MoG-da net (ours)	16.50	9.16	6.96	5.75	**10.61**	**6.68**	**5.73**	**5.10**



(a) 10 steps (b) 25 steps (c) 50 steps (d) 100 steps

Figure 2: Generated samples on CIFAR-10 with varying numbers of sampling steps. The examples compare visual quality at 10, 25, 50, and 100 sampling steps.

– **Network 3 ($j \in [700, 1000]$):** Models the high-noise regime at the beginning of sampling

During inference, the appropriate network is selected based on the current timestep. This time-conditional specialization is intended to improve training stability and prediction accuracy across noise levels.

All experiments use the following hyperparameters: learning rate $\eta = 10^{-4}$, total batch size $B = 128$, linear scheduler, and Adam optimizer (Kingma, 2014) with default momentum parameters $\beta_1 = 0.9, \beta_2 = 0.999$. The neural network architecture follows the standard U-Net (Ronneberger et al., 2015) design commonly used in diffusion models. Training is performed on 8 NVIDIA A800 GPUs (80GB memory each) for approximately 2 days per dataset.

B.2.2 Quantitative Comparison

Table 1 presents FID scores comparing our method against three baselines:

- **DDIM (Uncond baseline):** Standard unconditional DDIM (Song et al., 2021)
- **+ Naive:** Direct category prediction without temporal derivative
- **+ MoG-a net:** Training with only the posterior probability network a_θ
- **+ MoG-da net (ours):** Our full method with both a_θ and d_ϕ networks

Our method achieves lower FID than unconditional DDIM and MoG-a at every reported sampling budget on both datasets. It also outperforms the Naive method on CelebA-HQ, while Naive obtains the lowest FID on CIFAR-10. These results illustrate the empirical benefit of derivative-assisted routing relative to the posterior-only variant.

B.2.3 Qualitative Results

Figures 2 and 3 show generated samples from both datasets with varying numbers of sampling steps. Our method maintains high visual quality and categorical consistency even with very few steps, demonstrating the effectiveness of the temporal derivative regularization.



(a) 10 steps (b) 25 steps (c) 50 steps (d) 100 steps

Figure 3: Generated samples on CelebA-HQ with varying numbers of sampling steps. Our method produces high-quality faces even with as few as 10 steps, demonstrating effective category tracking on structured data.

B.3 Trajectory Error Analysis

We examine the numerical behavior of our decentralized sampler by measuring its endpoint discrepancy from a fine-step deterministic DDIM reference on CIFAR-10. The DDIM–flow correspondence above provides the connection between this implementation and the flow formulation.

B.3.1 Experimental Setup

We generate 10,000 reference images using 1000-step deterministic DDIM sampling, denoted by $\{\widehat{x}_{\text{ref}}^{(i)}\}_{i=1}^{10000}$. For each $N \in \{10, 25, 50, 100, 200, 500, 1000\}$, we generate 10,000 images with our method from the same initial noise realizations, denoted by $\{\widehat{x}_N^{(i)}\}_{i=1}^{10000}$. These symbols denote the evaluated image outputs. The mean squared endpoint discrepancy is

$$E(N) = \frac{1}{10000} \sum_{i=1}^{10000} \|\widehat{x}_N^{(i)} - \widehat{x}_{\text{ref}}^{(i)}\|^2. \tag{106}$$

This measures the discrepancy from a numerical reference, rather than from an exact ODE solution. If the reference uses a different model or routing rule, that difference also contributes to $E(N)$.

B.3.2 Error Scaling Analysis

Figure 4 plots $E(N)$ against the number of sampling steps. Fitting $E(N) = cN^{-\beta} + b$ gives

$$E(N) = 0.10 \cdot N^{-1.2916} + 0.0042, \qquad R^2 = 0.999963. \tag{107}$$

The fitted exponent $\beta \approx 1.29$ describes the decay of the squared discrepancy above the fitted offset over the measured sampling budgets. The square root of this excess scales as $N^{-0.6458}$, while the total root mean squared discrepancy retains the fitted floor. The main convergence result bounds $W_2(q_{N,\delta}^{\theta}, p_\delta)$ by $\mathcal{O}(\epsilon_v + \epsilon_a + N^{-1/2})$ at fixed δ under its stated assumptions. The fitted exponent describes the measured squared discrepancy from a numerical reference; it does not directly measure this distributional convergence bound.

The fit summarizes the observed decrease in endpoint discrepancy as the sampling budget increases. Its high R^2 indicates close agreement with the measured values, but does not by itself prove an asymptotic rate or isolate the effect of derivative supervision.

B.3.3 Practical Implications

The results illustrate a trade-off between sampling cost and agreement with the fine-step reference. Increasing the step budget reduces the observed discrepancy, with diminishing improvements as it approaches the fitted offset. Together with the FID comparisons, these measurements provide empirical support for decentralized sampling at limited step budgets, complementing the main convergence analysis.

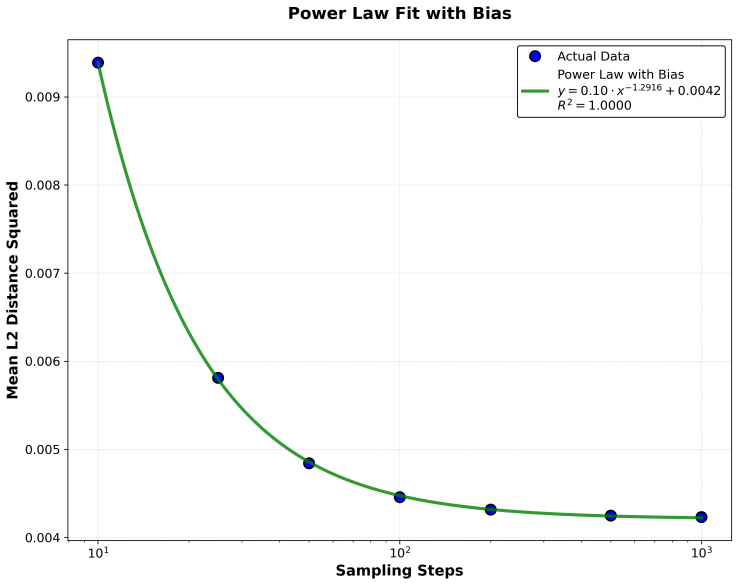

Figure 4: Endpoint discrepancy $E(N)$ versus the number of sampling steps on CIFAR-10. The reported fit is $E(N) = 0.10 \cdot N^{-1.2916} + 0.0042$, with $R^2 = 0.999963$.