Learning Manifold Data with Flow Matching

Anonymous Authors[1]

Abstract

We study flow-matching transformers when data lie on a low-dimensional manifold. Our key insight is a flow decomposition that splits motion along the manifold from motion off the manifold. The scheme works for first- and higher-order flow matching and ties model complexity to the intrinsic manifold dimension. Building on these, we establish tighter sample-complexity bounds for velocity approximation, velocity estimation, and distribution estimation. These bounds meet near-minimax rates for flow-matching transformers of any order. Our results show how flow-matching transformers escape the curse of dimensionality by utilizing intrinsic data structure.

1 Introduction

We study the sample complexity of learning flow matching generative models for data lying on low-dimensional manifolds. This theoretical analysis is of practical importance. Deep generative models have achieved remarkable success in modeling complex data distributions, with leading approaches including diffusion models (which learn to reverse a noising process) (Song & Ermon, 2019; Ho et al., 2020) and flow-based models (which learn invertible transformations) (Rezende & Mohamed, 2015). Flow matching is a recent flow-based paradigm that trains continuous normalizing flows by matching probability "flows"/vector fields rather than simulating sample paths (Lipman et al., 2023). Flow matching generalizes diffusion-type training objectives and is often observed to be stable and efficient in practice (Lipman et al., 2023; Liu et al., 2023). Modern architectures like Transformers further push generative modeling—e.g., diffusion transformers operating on latent image patches attain state-of-the-art results (Peebles & Xie, 2023). These advances motivate us to study the theoretical limits of flow matching models, especially when combined with powerful function approximators (we term such models *flow-matching transformers*).

A key question in high-dimensional generative modeling is how to mitigate the curse of dimensionality. The *manifold hypothesis* posits that although data lives in a high-dimensional ambient space (e.g. pixel space), it actually concentrates near a lower-dimensional manifold of intrinsic dimension $d_0 \ll d_x$ (Pope et al., 2021). This insight motivates many advances in representation learning and generative modeling (Loaiza-Ganem et al., 2024). For example, latent generative models compress data into lower-dimensional codes to simplify learning (Rombach et al., 2022). However, most theoretical guarantees for generative models scale poorly with d_x and do not explicitly exploit low-dimensional structure. Can generative models provably avoid exponential dependence on ambient dimension under manifold assumption? Recent work has started to address this: for score-based diffusion models, Chen et al. (2023a) showed that approximation and sampling errors can scale with the intrinsic dimension d_0 rather than d_x under a low-dimensional latent subspace assumption. Yet, analogous guarantees for flow matching methods are unexplored. In particular, it is unclear (i) how higher-order flow matching (which incorporates acceleration or higher derivatives in the flow) behaves in theory, and (ii) whether flow-based models can achieve dimension-free statistical rates when data lie on a manifold.

In this paper, we develop a theory of flow-matching transformers on manifold data. We focus on the setting where the data distribution is supported on a d_0-dimensional linear subspace of $\mathbb{R}^{d_x}$ (a special case of the manifold hypothesis) (Chen et al., 2023a). Our analysis introduces an explicit tangent/normal velocity decomposition that makes the population flow-matching risk *decompose pointwise*, which in turn yields *identifiability* of the two components at any global optimum. This same structure provides a *separation lower bound* that matches our upper bounds up to logarithmic/constant factors, a simple *stability/robustness* account via an orthogonal contraction back to the subspace, and a natural *two-head* architecture (a d_0-dimensional tangent head with a lightweight orthogonal head). Briefly, these ingredients also enable intrinsic-dimension rates for estimation and distributional (W_2) error that depend on d_0 rather than d_x, and they extend to higher-order flow matching

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

($K \geq 2$).

Contributions. Our contributions center on our decomposition of the flow velocity and its downstream implications:

- **Explicit tangent/normal velocity decomposition.** Under the manifold hypothesis, we give an explicit decomposition of the flow velocity into a tangent component that transports mass on-manifold and an orthogonal contraction off-manifold. The flow-matching risk decomposes pointwise across these components, yielding identifiability at any global optimum and stability via the orthogonal contraction. This naturally motivates a two-headed architecture.

- **Intrinsic-dimension statistical guarantees.** A key implication of our velocity decomposition is tighter statistical rates for flow matching transformers. We show that estimation and distributional error rates depend on the intrinsic dimension d_0 (and mild path regularity) rather than the ambient dimension d_x.

- **Near-minimax optimality.** Another profound implication of our velocity decomposition is that the achieved rates are near-minimax optimal. We prove matching (up to logs and constants) lower bounds adapted from worst-case density estimation on the latent space, showing that no method can substantially beat our d_0-dependent rates. This aligns with recent optimality results for flow matching in general settings (Fukumizu et al., 2024b).

- **Extension to higher-order flow matching.** We show that our first-order flow velocity decomposition extends naturally to higher-order flow matching models, which inherit identifiability, d_0-dependent rates, and near-minimax optimality.

Related Work. We defer the related work discussion to Section A due to page limits.

Organization. Section 2 presents the mathematical flow matching foundation we build on. Section 3 presents our flow decomposition trick for 1st order and K-order flow matching. Section 4 presents our sharp statistical analysis of first order flow matching transformers. We present an extension of this analysis to statistical rates of K-order flow matching transformers in Section L. Finally, we discuss our results and give concluding remarks Section 5.

2 Background

In this section, we provide a high-level overview of flow matching. We also describe the manifold hypothesis and our low-dimensional linear latent subspace assumption.

2.1 Flow Matching Framework

Flow-Based Generative Framework. A flow model transforms samples from a source distribution into samples from a target distribution by means of evolving flows over continuous time. Formally, let $X_0 = x_0 \in \mathbb{R}^{d_x}$ be a sample from a source distribution P_0 (e.g. a standard Gaussian), and $X_1 = x_1 \in \mathbb{R}^{d_x}$ be a sample from the target distribution P_1. A flow model is a model learning a time-dependent mapping $\psi_t : [0,1] \times \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$ sending (t,x) to $\psi_t(x)$. Then, with ψ_t we obtain a continuous-time process $(X_t)_{0 \leq t \leq 1}$ by evolving the initial point X_0 under this flow:

$$X_t = \psi_t(X_0), \quad t \in [0,1].$$

Namely, the distribution of X_t evolves according to

$$p_t(x) = [\psi_t]_* p_0(x) := p_0(\psi_t^{-1}(x)) \cdot \left| \det \left[\frac{\partial \psi_t^{-1}}{\partial x} \right] \right|, \tag{2.1}$$

where $[\psi_t]_* p_0$ denotes the pushforward distribution.

Equivalently, we describe the time-dependent mapping ψ_t via a time-dependent velocity field $u : [0,1] \times \mathbb{R}^{d_x} \to \mathbb{R}^{d_x}$, where we write $u(t,x) = u_t(x)$. The velocity field u uniquely determines the flow ψ as the solution of an ODE. In particular, ψ must satisfy the ordinary differential equation (ODE)

$$\frac{\mathrm{d}\psi_t}{\mathrm{d}t} = u_t(\psi_t(x)) \quad \text{with initial conditions} \quad \psi_0(x) = x, \tag{2.2}$$

so that at each time, the point $X_t = \psi_t(X_0)$ moves with velocity $u_t(X_t)$. Likewise, due to the one-to-one relationship between ψ_t and u_t, for a given ψ_t there is a unique smooth velocity field u_t satisfying

$$u_t(x) = \dot{\psi}_t(\psi_t^{-1}(x)), \quad \text{with} \quad \dot{\psi}_t = \frac{\mathrm{d}}{\mathrm{d}t}\psi_t, \tag{2.3}$$

which shows a theoretical method for computing u_t from ψ_t at the point x in the original source distribution. In summary, the flow ψ_t and velocity field u_t provide two equivalent ways to describe a continuous transformation from P_0 to P_1: ψ_t moves points directly, while u_t specifies the instantaneous velocity at every point in space and time.

Flow Matching Objective. Flow Matching (FM) (Lipman et al., 2023; 2024) is a simulation-free strategy for training generative flow models. Namely, flow matching avoids the need to explicitly simulate the ODE during training. Importantly, this departs from standard maximum likelihood training of ODE flows that directly maximizes data log-likelihood (Chen et al., 2018). The key idea is to match the probability flow induced by the model to the desired flow transforming samples drawn from the distribution P_0 into samples following the distribution P_1. We align the model's velocity field $u_\theta(x,t)$ with the true velocity field $u_t(x)$ to achieve this. Formally, suppose u_t indeed generates a path of densities $(p_t)_{0\le t\le 1}$ from p_0 (the source) to p_1 (the target). Then we define the flow matching loss as

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t\sim U[0,1], X_t\sim p_t}\left[\|u_t^\theta(X_t) - u_t(X_t)\|_2^2\right], \tag{2.4}$$

where $u_{\theta,t}(x)$ is the model's learnable velocity field (e.g. a neural network with parameters θ) and the expectation is over a random time t uniform on $[0,1]$ and a sample X_t drawn from the true density p_t. In practice, we introduce the conditional velocity fields $u_t(x|Z)$ and $p_t(x|Z)$ corresponding to $u_t(x|Z)$, where $Z\in\mathbb{R}^m$ is an auxiliary random variable. To fit the original model, the marginal density and velocity should recover the origin p_t and u_t via

$$p_t(x) = \int p_t(x|z)p_Z(z)\mathrm{d}z, \tag{2.5}$$

$$u_t(x) = \int u_t(x|z)\frac{p_t(x|z)p_Z(z)}{p_t(x)}\mathrm{d}z. \tag{2.6}$$

The Conditional Flow Matching (CFM) loss is defined as

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,Z\sim p_Z, X_t\sim p_{t|Z}(\cdot|Z)}\left[\|u_t^\theta(X_t) - u_t(X_t|Z)\|_2^2\right]. \tag{2.7}$$

It holds that $\nabla_\theta\mathcal{L}_{\mathrm{FM}}(\theta) = \nabla_\theta\mathcal{L}_{\mathrm{CFM}}(\theta)$ and the minimizer of the Conditional Flow Matching loss is the marginal velocity $u_t(x)$. Therefore, by setting $Z = X_1\sim P_1$, we get $u_\theta(x,t)$ with selected start point and end point.

Affine Conditional Flow. The flow matching method and conditional flow matching loss are applicable to all constructions of conditional paths and conditional velocity field under mild assumptions, leaving room for picking certain accessible conditional flow. In this paper, we consider the affine conditional flow: we set $Z = X_1\sim P_1$, meaning that Z is the target sample itself. The paths is constructed via the following interpolation between the source point x and the target sample x_1:

$$\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x, \tag{2.8}$$

where μ_t and σ_t are smooth scalar schedules on $[0,1]$ satisfying the boundary and smooth conditions

$$\mu_0 = \sigma_1 = 0, \mu_1 = \sigma_0 = 1, \text{ and}$$
$$\dot\mu_t = \frac{\mathrm{d}\mu_t}{\mathrm{d}t} > 0, \dot\sigma_t = \frac{\mathrm{d}\sigma_t}{\mathrm{d}t} < 0 \text{ for } t\in(0,1). \tag{2.9}$$

The boundary conditions of μ_t and σ_t ensures that the smooth path starts at x and ends at x_1, the start and end points we select. Under this construction, we have $p_t(X_t|X_1) = N(\mu_t X_1, \sigma_t^2 I)$, and the velocity field takes the form

$$u_t(x|x_1) = \dot\psi_t(\psi_t^{-1}(x|x_1)|x_1)$$
$$= \frac{\dot\sigma_t(x-\mu_t x_1)}{\sigma_t} + \dot\mu_t x_1. \tag{2.10}$$

Further, substituting $X_t = \psi_t(X_0|X_1)$ we get

$$\mathcal{L}_{\mathrm{CFM}}(\theta) = \mathbb{E}_{t,X_1\sim p_1,X_0\sim p_0}[\|u_t^\theta(\mu_t X_1 + \sigma_t X_0) - (\dot\mu_t X_1 + \dot\sigma_t X_0)\|_2^2]. \tag{2.11}$$

In practice, given i.i.d. samples $\{x_i\}_{i=1}^n$ drawn from the target distribution P_1, the empirical loss function $\widehat{\mathcal{L}}_{\mathrm{CFM}}(u_\theta)$ for a neural network u_θ takes the form:

$$\widehat{\mathcal{L}}_{\mathrm{CFM}}(u_\theta) := \frac{1}{n}\sum_{i=1}^n\int_{t_0}^T \frac{1}{T-t_0}\mathbb{E}_{X_0\sim\mathcal{N}(0,I)}\left[\mathrm{DIF}\right]\mathrm{d}t, \tag{2.12}$$

where

$$\mathrm{DIF} := \|u_\theta(\mu_t x_i + \sigma_t X_0, t) - (\dot\mu_t x_i + \dot\sigma_t X_0)\|_2^2,$$

and $0 < t_0 < T < 1$. Note that since $\dot\mu$ and $\dot\sigma$ may blow up on the boundary, we use the interval $[t_0, T]$ instead of $[0,1]$ when integrating. By optimizing the empirical conditional flow matching loss, we push the learned u_θ towards the true optimal velocity, thereby simulating ψ_t and the whole generating process.

2.2 Manifold Assumption

In this section, we formalize the manifold hypothesis and establish the central low-dimensional linear latent subspace assumption. We refer to the low-dimensional linear latent subspace assumption as the manifold assumption in the rest of the paper.

According to the manifold hypothesis, high-dimensional data (such as images or audio) concentrate near a much lower–dimensional set. Empirical studies confirm that common image datasets possess an intrinsic dimension one or two orders of magnitude smaller than the ambient pixel space in most cases (Pope et al., 2021). From a theoretical standpoint, recent works show that modern generative models automatically adapt to such low–dimensional structure. For instance, diffusion models provably attain manifold-dependent error rates (Tang & Yang, 2024). Also, score-based analyses demonstrate that sample complexity can scale with the intrinsic dimension rather than the ambient dimension (Chen et al., 2023a). These results show that incorporating the manifold structure do lead to sharper bounds and more efficient learning. Additionally, findings above justify adopting a low-dimensional data model when analyzing modern generative methods. Following (Chen et al., 2023a), we formalize the low-dimensional data assumption. We assume an intrinsic lower-dimensional representation generates the raw input $x\in\mathbb{R}^{d_x}$ in the following way.

Assumption 2.1 (Low-Dimensional Linear Latent Subspace)**.** Initial data point x have a latent representation given by $x = Uh$, where $U \in \mathbb{R}^{d_x \times d_0}$ is an unknown matrix with orthonormal columns. The latent variable $h \in \mathbb{R}^{d_0}$ follows distribution P_1^h with probability density function p_1^h.

Remark 2.1. "Linear Latent Space" means that each entry of a given latent vector is a linear combination of the corresponding input, i.e. $x = Uh$. Many recent theoretical works on generative modeling use this assumption (Chen et al., 2023a; Hu et al., 2024b; Jiao et al., 2024; Tang & Yang, 2024). Empirically, large-scale intrinsic-dimension studies confirm that image and audio datasets admit low linear dimension after suitable preprocessing (Pope et al., 2021).

Previous work proves the score decomposition theory of standard diffusion model under manifold assumption Assumption 2.1. (Chen et al., 2023a) investigates the approximation, estimation, and distribution recovery of diffusion models under manifold assumption. Building on similar assumptions, (Hu et al., 2024b) analyzes the statistical and computational limits of latent Diffusion Transformers. However, the effect of manifold assumption in flow matching model and related conclusions remain untouched in previous work. To bridge this gap, this paper introduces the velocity decomposition under manifold assumption in Section 3 and studies the statistical rates of flow matching model with Transformer network in Section 4 and Section L.

2.3 Transformer Networks

We defer the standard definition of transformer networks to Section E due to the page limit.

3 Velocity Decomposition

In this section, we show that for a low-dimensional data distribution, velocity function decomposes into two orthogonal components with distinct properties. Exploiting these properties enables an efficient approximation and estimation of the velocity function depending on the latent dimension d_0 instead of the ambient dimension d_x. See Section 4 for details of statistical rates of flow matching under manifold assumption Assumption 2.1.

The idea of separating "on-manifold" and "off-manifold" dynamics dates back to score-based analyses of diffusion models (Chen et al., 2023a; Tang & Yang, 2024). In the passages above, the score $\nabla_x \log p_t(x)$ decomposes into a latent part encoding intrinsic data geometry and an orthogonal part pulling points back towards the manifold. Our results below show an analogous decomposition for the velocity field of affine conditional flow (Lipman et al., 2023). We learn each component with complexity governed by the latent dimension d_0.

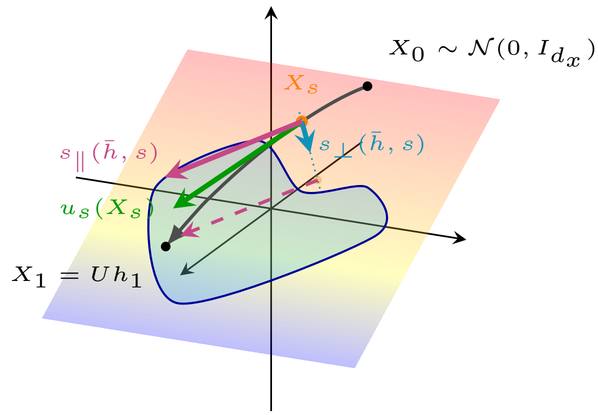

Figure 1. An illustration of the velocity decomposition in ambient dimension $d_x = 3$ with data manifold in a linear latent subspace with dimension $d_0 = 2$. We depict the flow path from X_0 to X_1 with the curved gray arrow. The green arrow $u_s(X_s)$ represents the velocity along the path at time $t = s$, and the purple and blue arrows ($s_{\|}$ and $s_{\perp}$) represent the on-support and orthogonal components of the velocity, respectively. $s_{\|}$ belongs to the linear latent subspace (as emphasized by the dashed purple arrow), while $s_{\perp}$ belongs to the orthogonal subspace.

3.1 Velocity Decomposition under Assumption 2.1

Under manifold assumption Assumption 2.1, we decompose the velocity into its on-support and orthogonal components.

Theorem 3.1 (Velocity Decomposition Under the Low Dimensional Linear Latent Subspace Assumption)**.** Let $x = Uh$ satisfies Assumption 2.1. Consider the affine conditional flow

$$X_t = \psi_t(X_0 \mid X_1) = \mu_t X_1 + \sigma_t X_0,$$

where $(X_1, X_0) \sim (q, N(0, I_{d_x}))$ with smooth coefficients $\mu_t, \sigma_t \in (0, 1)$ satisfying (2.9). We define the following constants

$$\kappa_t := \frac{\dot{\sigma}_t}{\sigma_t}, \quad \lambda_t := \dot{\mu}_t - \mu_t \kappa_t.$$

For every $x \in \mathbb{R}^{d_x}$, let $\bar{h} = U^\top x$. Then the optimal velocity field in the conditional flow–matching objective (2.11) admits the decomposition

$$u_t(x) = \underbrace{U\Big[\alpha_t \bar{h} + \beta_t \nabla_{\bar{h}} \log p_t^h(\bar{h})\Big]}_{u_{\|}(\bar{h}, t):\ \text{latent transport}} + \underbrace{\kappa_t (I - UU^\top) x}_{u_{\perp}(x, t):\text{orthogonal contraction}},$$

where p_t^h is the marginal density of $\bar{h}$ and coefficients satisfy $\alpha_t := \kappa_t + \lambda_t/\mu_t$,$\beta_t := \lambda_t \sigma_t^2/\mu_t$.

Proof Sketch. The proof begins by expressing the marginal velocity field $u_t(x)$ in terms of conditional expectations $\mathbb{E}[X_1|X_t = x]$ and $\mathbb{E}[X_0|X_t = x]$. Crucially, the low-dimensional linear latent subspace assumption ($X_1 = Uh$) allows us to rewrite these expectations by conditioning on the latent projection $\bar{h} = U^\top x$ and the orthogonal component $x_\perp = (I - UU^\top)x$. This separates the dynamics into two parts. First part, the on-support component, depends on the score $\nabla_{\bar{h}} \log p_t^h(\bar{h})$ in the d_0-dimensional latent space via Tweedie's formula. The second part, the orthogonal component, is a simple linear function of $x_\perp = (I - UU^\top)x$. Please see Section B for a detailed proof. □

We visualize the decomposition of the velocity in Figure 1.

3.2 Higher Order Flow Matching Decomposition under Assumption 2.1

Higher-order flow objectives are attracting increasing attention for one-step and few-step generation (Chen et al., 2025; Gong et al., 2025). The next result extends the first-order decomposition to arbitrary order k, showing that each higher-order velocity $u_t^{(k)}$ enjoys the same latent/orthogonal splitting—and hence the same intrinsic-dimension benefits—as the base velocity. Under manifold assumption Assumption 2.1, we decompose the k-th order velocity into its on-support and orthogonal components.

Theorem 3.2 (k-th Order Velocity Decomposition Under the Low Dimensional Linear Latent Subspace Assumption)**.** Let $U \in \mathbb{R}^{d_x \times d_0}$ have orthonormal columns and suppose the data assumption $x = Uh$ with $h \sim P_1^h$ holds (Assumption 2.1). Consider the affine conditional flow

$$X_t = \psi_t(X_0 \mid X_1) = \mu_t X_1 + \sigma_t X_0,$$

where $(X_1, X_0) \sim (q, N(0, I_{d_x}))$ with smooth coefficients $\mu_t, \sigma_t \in (0, 1)$ that satisfy (2.9). Write $\mu_t^{(k)} = \frac{\mathrm{d}^k}{\mathrm{d}t^k}\mu_t$ and $\sigma_t^{(k)} = \frac{\mathrm{d}^k}{\mathrm{d}t^k}\sigma_t$, and define the constants

$$\kappa_{k,t} := \frac{\sigma_t^{(k)}}{\sigma_t}, \qquad \lambda_{k,t} := \mu_t^{(k)} - \mu_t \kappa_{k,t}.$$

For every realisation $x \in \mathbb{R}^{d_x}$ let $\bar{h} = U^\top x$ (latent coordinate) and $x_\perp = (I - UU^\top)x$ (orthogonal component). Then the optimal k-th order velocity field that appears in the k-th order conditional flow–matching objective (K.3) admits the decomposition

$$u_t^{(k)}(x) = \underbrace{U\Big[\kappa_{k,t}\bar{h} + \lambda_{k,t}\mathbb{E}\left[h\middle|\bar{h}\right]\Big]}_{s_\parallel^{(k)}(\bar{h},t)} + \underbrace{\kappa_{k,t}(I - UU^\top)x}_{s_\perp^{(k)}(x,t)}.$$

Moreover, by Tweedie's formula in latent space,

$$\mathbb{E}\left[h\middle|\bar{h}\right] = \frac{1}{\mu_t}\Big(\bar{h} + \sigma_t^2 \nabla_{\bar{h}} \log p_t^h(\bar{h})\Big),$$

where p_t^h is the marginal density of $\bar{h}$. This yields the equivalent "score–based" form

$$u_t^{(k)}(x) = \underbrace{U\Big[\alpha_{k,t}\bar{h} + \beta_{k,t}\nabla_{\bar{h}} \log p_t^h(\bar{h})\Big]}_{s_\parallel^{(k)}(\bar{h},t):\ k\text{-th order on-support component}} + \underbrace{\kappa_{k,t}(I - UU^\top)x}_{s_\perp^{(k)}(x,t):\ k\text{-th order orthogonal component}},$$

with coefficients $\alpha_{k,t} := \kappa_{k,t} + \lambda_{k,t}/\mu_t$ and $\beta_{k,t} := \lambda_{k,t}\sigma_t^2/\mu_t$.

Proof. Please see Section C for a detailed proof. □

3.3 Risk Decomposition and Identifiability

The explicit tangent/normal split implies that the population flow-matching risk is a sum of two squared errors—one for each component. This decoupling drives our identifiability result (the two components are learned independently at any global optimum) and underpins the intrinsic-dimension statistical consequences in Section 4.

Lemma 3.1 (Risk Decomposition)**.** Under Assumption 2.1, let $U \in \mathbb{R}^{d_x \times d_0}$ span the latent subspace and define the orthogonal projectors $P_U := UU^\top$ and $P_{U^\perp} := I - UU^\top$. For any measurable velocity $u : \mathbb{R}^{d_x} \times [0, 1] \to \mathbb{R}^{d_x}$, define the population flow-matching risk

$$R(u) := \mathbb{E}\|u(X_t, t) - u_t^\star(X_t)\|^2,$$

where (X_t, t) are drawn from the affine conditional path used for training and $u_t^\star$ is the oracle velocity field. Then the risk decomposes pointwise across the tangent and normal components:

$$\begin{aligned} R(u) =& \mathbb{E}\|P_U(u(X_t, t) - u_t^\star(X_t))\|^2 \\ &+ \mathbb{E}\|P_{U^\perp}(u(X_t, t) - u_t^\star(X_t))\|^2. \end{aligned}$$

Proof sketch. By orthogonality of the tangent and normal projectors, the Pythagorean identity yields a pointwise sum of squared errors for the two components; taking expectation over the training distribution of (X_t, t) gives the stated risk decomposition. See Section D for details. □

Proof. Write $a(x, t) := u(x, t) - u_t^\star(x)$. Since P_U and $P_{U^\perp}$ are orthogonal projectors with $P_U^\top = P_U$, $P_{U^\perp}^\top = P_{U^\perp}$, $P_U P_{U^\perp} = 0$, and $P_U + P_{U^\perp} = I$, we have for each (x, t):

$$a = (P_U + P_{U^\perp})a = P_U a + P_{U^\perp} a.$$

For any $a \in \mathbb{R}^{d_x}$, P_U and $P_{U^\perp}$ are orthogonal with $P_U P_{U^\perp} = 0$ and $P_U + P_{U^\perp} = I$. Hence, we have the

Pythagorean identity

$$\|a\|^2 = \|P_U a\|^2 + \|P_{U_\perp} a\|^2.$$

Taking squared norms and expanding with the inner product $\langle \cdot, \cdot \rangle$, we have

$$\|a\|^2 = \|P_U a\|^2 + \|P_{U^\perp} a\|^2 + 2\langle P_U a, P_{U^\perp} a\rangle.$$

The cross term vanishes pointwise: by self-adjointness and $P_U P_{U^\perp} = 0$,

$$\langle P_U a, P_{U^\perp} a\rangle = \langle a, P_U P_{U^\perp} a\rangle = \langle a, 0\rangle = 0.$$

Hence for every (x, t), we have the pointwise identity

$$\|a(x,t)\|^2 = \|P_U a(x,t)\|^2 + \|P_{U^\perp} a(x,t)\|^2.$$

Now we evaluate at the random pair (X_t, t) drawn by the training path and take expectations. Since both terms on the right are nonnegative and by assumption $\mathbb{E}\|a(X_t,t)\|^2 < \infty$, Tonelli's theorem justifies exchanging expectation with the sum. Thus,

$$\begin{aligned} R(u) &= \mathbb{E}\|a(X_t,t)\|^2 \\ &= \mathbb{E}\Big(\|P_U a(X_t,t)\|^2 + \|P_{U^\perp} a(X_t,t)\|^2\Big) \\ &= \mathbb{E}\|P_U a(X_t,t)\|^2 \ + \ \mathbb{E}\|P_{U^\perp} a(X_t,t)\|^2. \end{aligned}$$

This is the claimed decomposition after substituting back $a = u - u_t^\star$. □

Remark 3.1. Write $R(u) = R_\parallel(u) + R_\perp(u)$ with $R_\parallel(u) := \mathbb{E}\|P_U(u - u_t^\star)\|^2 \geq 0$ and $R_\perp(u) := \mathbb{E}\|P_{U^\perp}(u - u_t^\star)\|^2 \geq 0$. Note that if $\widehat{u}$ is a global minimizer of R, then necessarily $\widehat{u} \in \arg\min R_\parallel$ and $\widehat{u} \in \arg\min R_\perp$. Indeed, if say $R_\parallel(\widehat{u})$ were not minimal, then there exists v with $R_\parallel(v) < R_\parallel(\widehat{u})$. Keeping the orthogonal component unchanged (so $R_\perp(v) = R_\perp(\widehat{u})$) yields $R(v) < R(\widehat{u})$, which contradicts optimality. Thus the tangent and normal components are optimized independently at any global minimum, which is the functional "no interference" property used in the identifiability theorem. This decoupling of components also admits a natural geometric interpretation, which we visualize in Figure 2.

Theorem 3.3 (Identifiability of Tangent and Normal Components)**.** Under Assumption 2.1, let $U \in \mathbb{R}^{d_x \times d_0}$ span the latent subspace and define $P_U := UU^\top$, $P_{U^\perp} := I - UU^\top$. For any measurable velocity field $u : \mathbb{R}^{d_x} \times [0,1] \to \mathbb{R}^{d_x}$ with $\mathbb{E}\|u(X_t,t)\|^2 < \infty$ and $\mathbb{E}\|u_t^\star(X_t)\|^2 < \infty$, consider the population flow-matching risk

$$R(u) := \mathbb{E}\|u(X_t,t) - u_t^\star(X_t)\|^2.$$

If $\widehat{u} \in \arg\min_u R(u)$ (where the minimization ranges over all measurable, square-integrable velocity fields), then for the training distribution of (X_t, t) we have, almost surely,

$$P_U \widehat{u}(\cdot, t) = P_U u_t^\star(\cdot),$$
$$P_{U^\perp} \widehat{u}(\cdot, t) = P_{U^\perp} u_t^\star(\cdot).$$

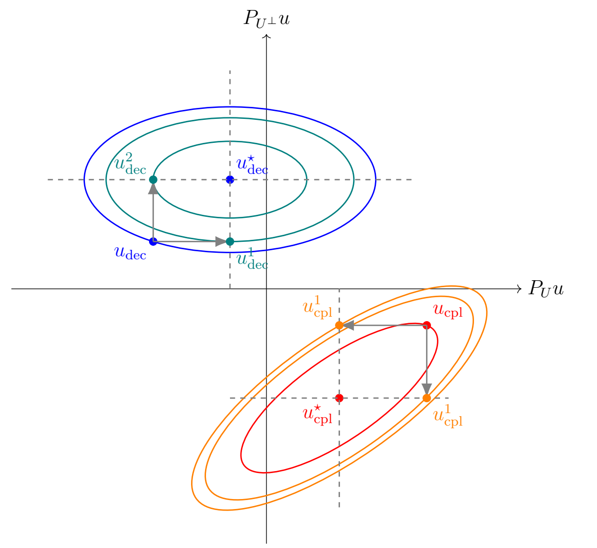

Figure 2. Decoupled vs. coupled loss landscapes for tangent/normal components. In the decoupled case, level sets are axis-aligned ellipses around the oracle $u^\star_{\mathrm{dec}}$; a unilateral axis-aligned step from u_{dec} toward the oracle (to u^1_{dec} or u^2_{dec}) always lands on a weakly better level set. In the coupled case, level sets are rotated ellipses around $u^\star_{\mathrm{cpl}}$; a unilateral axis-aligned step from u_{cpl} toward the oracle (to u^1_{cpl} or u^2_{cpl}) can move to a worse level set. This conceptually illustrates that the tangent and normal components cannot be optimized independently.

Proof sketch. By Lemma 3.1, write the population risk as $R(u) = R_\parallel(u) + R_\perp(u)$. If the tangent component at a global minimizer $\widehat{u}$ were suboptimal, replacing only that component with the optimal one (and keeping the orthogonal component fixed) would strictly reduce R, a contradiction; the orthogonal case is symmetric. See Section D for details. □

Remark 3.2. The explicit velocity decomposition and identifiability result suggest a natural architectural design: parameterize $u_\theta(x,t)$ as the sum of two *heads*, one constrained to the tangent subspace and one to the orthogonal subspace. For instance,

$$u_\theta(x,t) = U f_\theta(U^\top x, t) + g_\phi((I - UU^\top)x, t),$$

where f_θ is a d_0-dimensional network predicting the tangent (transport) component, and g_ϕ is a lightweight (possibly parametric or fixed) function predicting the normal (contractive) component. This structure is analogous to the functional decoupling in the risk: each head targets its component independently, with the tangent head carrying the statistical complexity and the orthogonal head promoting stability.

4 Statistical Rates Analysis

In this section, we establish sharp statistical rates of flow matching transformers under the manifold assumption Assumption 2.1. We show that flow matching models achieve approximation and learning rates depending only on d_0, not the ambient d_x. In particular, we show the model is expressive enough to fit the decomposed velocity. Then, we establish sample complexity bounds for learning these velocity from data. Lastly, we bound the generative distribution error. All results reflect intrinsic-dimension dependence, and we confirm they are statistically near-optimal. Specifically, Section 4.1 presents velocity approximation under a generic Hölder smoothness assumption. Section 4.2 utilizes these approximation results to develop velocity estimation bounds. Section 4.3 then develops distribution estimation rates under the 2-Wasserstein metric. Finally, Section 4.4 establishes the nearly minimax optimality of flow matching transformers.

Proof Strategy and Role of Velocity Decomposition. The derivation of our statistical rates (Theorem 4.1, Theorem 4.2, Theorem 4.3, and Proposition 4.1) hinges on the velocity decomposition presented in Theorem 3.1. This decomposition is crucial, as it concentrates the complexity of the dynamics on the on-support component $s_{\|}(\bar{h}, t)$ in the d_0-dimensional latent subspace. The dynamics in the orthogonal complement $s_{\perp}(x, t)$ are linear and simpler to model.

Our proof strategy involves:

1. **Approximation (Theorem 4.1):** We show that a transformer can efficiently approximate the decomposed velocity. The critical on-support component $s_{+}(\bar{h}, t)$ is approximated as a function on the d_0-dimensional latent space. This allows the approximation error to depend on d_0 rather than the ambient d_x.
2. **Estimation (Theorem 4.2):** We adapt standard empirical risk minimization arguments. Observing that the intricate part of the target velocity function is at most d_0-dimensional, we quantify the complexity of the learned function class via covering numbers.
3. **Distribution Estimation (Theorem 4.3):** The error in estimating the data distribution (in W_2 distance) is then bounded by the velocity estimation error, propagating the d_0-dimensional scaling.
4. **Minimax Optimality (Proposition 4.1):** Finally, we demonstrate the d_0-dependent rates match fundamental lower bounds for density estimation on d_0-dimensional manifolds, establishing the optimality of flow-matching transformers under manifold assumption Assumption 2.1.

Thus, the velocity decomposition is instrumental in circumventing the curse of dimensionality by tying the statistical complexity to the intrinsic dimension d_0.

4.1 Velocity Approximation under Assumption 2.1

Establishing our statistical theory starts with approximating the velocity using transformers. We present the velocity approximation theory under the Hölder smoothness assumption on the initial data (Fu et al., 2024). This theory ensures our approximation rate adapts to the initial data's smoothness. We first introduce the definition of Hölder space and Hölder ball.

Definition 4.1 (Hölder Space)**.** Let $\alpha \in \mathbb{Z}_{+}^{d_0}$, and let $\beta = k_1 + \gamma$ denote the smoothness parameter, where $k_1 = \lfloor \beta \rfloor$ and $\gamma \in [0, 1)$. For a function $f : \mathbb{R}^{d_0} \to \mathbb{R}$, the Hölder space $\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}\right)$ is defined as the set of α-differentiable functions satisfying: $\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}\right) := \left\{ f : \mathbb{R}^{d_0} \to \mathbb{R} \mid \|f\|_{\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}\right)} < \infty \right\}$, where the Hölder norm $\|f\|_{\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}\right)}$ satisfies:

$$\|f\|_{\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}\right)} := \max_{\alpha : \|\alpha\|_1 < k_1} \sup_{x} |\partial^{\alpha} f(x)| + \max_{\alpha : \|\alpha\|_1 = k_1} \sup_{x \neq x'} \frac{|\partial^{\alpha} f(x) - \partial^{\alpha} f(x')|}{\|x - x'\|_{\infty}^{\gamma}}.$$

Also, we define the Hölder ball of radius B by

$$\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}, B\right) := \left\{ f : \mathbb{R}^{d_0} \to \mathbb{R} \mid \|f\|_{\mathcal{H}^{\beta}\left(\mathbb{R}^{d_0}\right)} < B \right\}.$$

Before presenting the main result of velocity approximation, we first need to impose two assumptions: (i) the Generic Hölder Smooth assumption on the latent target distribution $p_1^h(h_1)$, and (ii) a regularity assumption on the first derivative of path coefficients.

Assumption 4.1 (Generic Hölder Smooth Data)**.** The true latent density function p_1^h belongs to Hölder ball of radius $B > 0$ (Definition 4.1), denoted by $p_1^h \in \mathcal{H}^{\beta}(\mathbb{R}^{d_0}, B)$. Also, there exist constants $C_1, C_2 > 0$ such that $p_1^h(h_1) \leq C_1 \exp\left(-C_2 \|h_1\|_2^2 / 2\right)$ for all $h_1 \in \mathbb{R}^{d_0}$.

Assumption 4.2 (Path Regularity)**.** Consider the affine conditional flow $\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$. The first derivative of the path coefficients $\dot{\sigma}_t$ and $\dot{\mu}_t$ are continuous on $[t_0, T]$, where $t_0, T \in (0, 1)$.

We now present the velocity approximation for flow matching transformers under Assumption 4.1 and Assumption 2.1.

Theorem 4.1 (Velocity Approximation with Transformers under manifold assumptions Assumption 2.1)**.** Assume Assumption 2.1, Assumption 4.1 (for p_1^h) and Assumption 4.2.

For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$. Then, for all $t \in [t_0, T]$ with $t_0, T \in (0, 1)$, there exists a transformer $u_\theta(x,t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{\mathbb{R}^{d_x}} \|u_t(x) - u_\theta(x,t)\|_2^2 p_t(x)\mathrm{d}x = O\left(B^2 N^{-\beta}(\log N)^{d_0 + d_x/2 + \beta/2 + 1}\right).$$

Furthermore, the parameter bounds in the transformer network $\mathcal{T}_R^{h,s,r}$ satisfy (with $\epsilon_{\mathrm{tf}} = N^{-\beta}$):

$$\begin{aligned} C_{KQ}, C_{KQ}^{2,\infty} &= O\left((\log N)^{2d+1} N^{\beta(4d+2)}\right), \\ C_{OV}, C_{OV}^{2,\infty} &= O(N^{-\beta}), \\ C_F, C_F^{2,\infty} &= O\left((\log N) N^{\beta}\right), \\ C_E &= O(1), \\ C_{\mathcal{T}} &= O(\sqrt{\log N}). \end{aligned}$$

The $O(\cdot)$ hides polynomial factors depending on $d_x, d_0, d, L, \beta, C_1, C_2$, and constants from domain definitions.

Proof. See Section G for the proof. □

4.2 Velocity Estimation Under Assumption 2.1

In this section, we study the statistical estimation problems and develop sample complexity results based on the established approximation results in Section 4.1. Specifically, we present the estimation error bound of flow matching transformers in Theorem 4.2.

Velocity Estimation. Building on the transformer-based velocity approximation, we evaluate the performance of the velocity estimator u_θ by optimizing the empirical loss (2.12). To quantify this, we define the flow matching risk:

Definition 4.2 (Flow Matching Risk)**.** Let the latent target sample be $H_1 \sim p_1^h$ (density in $\mathbb{R}^{d_0}$) and the visible target sample be $X_1 \sim p_1$ (push-forward density in $\mathbb{R}^{d_0}$). For $t \in [t_0, T]$, the affine conditional flow $\psi_t(x|X_1) = \mu_t X_1 + \sigma_t x$ induces the visible-space path density p_t and its true velocity field $u_t(\cdot)$. Given a velocity estimator $u_\theta : \mathbb{R}^{d_x} \times [t_0, T] \to \mathbb{R}^{d_x}$, we define the flow matching risk $\mathcal{R}(u_\theta)$ as the expectation of the mean-squared difference between u_θ and the ground truth velocity u_t:

$$\mathcal{R}(u_\theta) := \frac{1}{T - t_0} \int_{t_0}^{T} \mathbb{E}_{x_t \sim p_t}\left[\|u_\theta(x_t, t) - u_t(x_t)\|_2^2\right] \mathrm{d}t.$$

The expectation is taken over the latent-generated visible sample $X_t = U\bar{h}_t \sim p_t$. The estimator u_θ will be learned from the i.i.d. training set $\{x_i = Uh_i\}_{i=1}^n$ by minimizing the empirical loss (2.12).

Let $\widehat{u}_\theta$ be the trained velocity estimator with i.i.d. samples $\{x_i\}_{i=1}^n$. Then the following theorem presents upper bounds in the expectation of $\mathcal{R}(\widehat{u}_\theta)$ w.r.t. training samples $x_{i}{}_{i=1}^n$, where $x_i \sim p_1$.

Theorem 4.2 (Velocity Estimation with Transformer Under manifold assumption Assumption 2.1)**.** Assume Assumption 2.1. Let $\nu := 16\beta d + 12\beta$, where $d \times L = d_x$ is the (patch-size $\times$ sequence-length) input shape used by the transformer. Suppose we choose the transformer as in Theorem 4.1 and assume Assumption 4.1 and Assumption 4.2. Then, by taking $N = n^{1/(\nu + 3\beta)}$, it holds

$$\mathbb{E}_{\{x_1\}_{i=1}^n}\left[\mathcal{R}(\widehat{u}_\theta)\right] = O\left(n^{-\frac{1}{16d+15}} (\log n)^{\max\{d_0 + \frac{1}{2}\beta + 1, 8d+17\}}\right).$$

Proof. See Section H for a detailed proof. □

4.3 Distribution Estimation Under Assumption 2.1

Applying the velocity estimation rates from Section 4.2, we further analyze the distribution estimation rate for the velocity estimator $\widehat{u}_\theta$ through the 2-Wasserstein distance between estimated and true distributions. The 2-Wasserstein distance is defined as follows:

Definition 4.3 (2-Wasserstein Distance)**.** Let X and Y be two random variables with marginal densities μ_x and μ_y respectively. We define the 2-Wasserstein distance by:

$$W_2(\mu_x, \mu_y) := \left(\inf_{\pi \in \mathcal{M}(\mu_x, \mu_y)} \int \|x - y\|^2 \mathrm{d}\pi(x,y)\right)^{\frac{1}{2}},$$

where $\mathcal{M}(\mu_x, \mu_y)$ denotes the set of joint measures π with marginals μ_x and μ_y.

Based on the velocity estimation results in Section 4.2, the next theorem presents upper bounds on the Wassertstein-2 distance between the target distribution and the estimated distribution induced by the velocity estimator $\widehat{u}_\theta$ trained from optimizing the empirical conditional loss (2.12).

Theorem 4.3 (Distribution Estimation With Wasserstein Distance Under Assumption 2.1)**.** Let $\widehat{P}_T$ be the distribution obtained at (clipped) terminal time $T = C_\alpha \log N$ by running the reverse flow driven by the learned velocity field $\widehat{u}_\theta$. Assume Assumption 2.1, Assumption 4.1 and Assumption 4.2. Then, for any sample size n,

$$\mathbb{E}_{\{h_i\}_{i=1}^n}\left[W_2(\widehat{P}_T, P_T)\right] = O\left(n^{-\frac{1}{32d+30}} (\log n)^{\max\left\{\frac{d_0}{2} + \frac{\beta}{4} + \frac{1}{2}, 4d + \frac{17}{2}\right\}}\right).$$

Proof. See Section I for a detailed proof. □

4.4 Minimax Optimal Estimation Under Assumption 2.1

In Theorem 4.3, we present a fine-grained analysis of distribution estimation. In this section, we further show that the derived estimation rates match the minimax lower bounds in Hölder space under the 2-Wasserstein metric under specific settings. We begin by recalling the minimax optimal rate for distribution estimation over Hölder smooth function classes.

Lemma 4.1 (Minimax lower bound in the latent space, Modified from Theorem 3 of (Niles-Weed & Berthet, 2019))**.** Let $\mathcal{P}_h := \{p_1^h(h) : p_1^h \in \mathcal{H}^\beta([0,1]^{d_0}, B), p_1^h(h) \geq C, \int p_1^h = 1\}$, where $d_0 \geq 1$, $B, C > 0$ and $\beta > 0$. For every $r \geq 1$ and every estimator $\widehat{P}_h$ based on n i.i.d. samples $\{H_i\}_{i=1}^n \sim (p_1^h)^{\otimes n}$, we have

$$\inf_{\widehat{P}_h} \sup_{p_1^h \in \mathcal{P}_h} \mathbb{E}_{\{H_i\}} \left[W_r\left(\widehat{P}_h, P_1^h\right) \right] \gtrsim n^{-\frac{\beta+1}{d_0+2\beta}}.$$

Proof. Please see Section J for the proof. □

We now show that flow matching transformers match minimax optimal rate under specific conditions.

Proposition 4.1 (Minimax Optimality of Flow Matching Transformers under Assumption 2.1)**.** Assume the conditions of Theorem 4.3 hold, specifically Assumption 4.1 for the latent density $p_1^h \in \mathcal{H}^\beta([0,1]^{d_0}, B)$ and Assumption 4.2. Let d be the transformer's internal feature dimension (as in the definition of ν in Theorem 4.2). Under the setting where

$$(32d + 30)(\beta + 1) = d_0 + 2\beta, \tag{4.1}$$

the distribution estimation rate of flow matching transformers, as given in Theorem 4.3, matches the minimax lower bound for estimating the β-Hölder smooth latent distribution P_1^h (from Lemma 4.1) in 2-Wasserstein distance, up to logarithmic factors.

Proof. Please see Section J for the proof. □

5 Conclusion and Discussion

In this work, we provide a rigorous theoretical analysis of flow-matching transformers operating on data concentrated on low-dimensional linear latent subspaces. We introduce a novel velocity field decomposition (Theorem 3.1) that separates dynamics along the latent subspace from those orthogonal to it. This decomposition, applicable to both first-order and K-th order flow matching (Theorem 3.2), is the cornerstone for deriving statistical guarantees depending on the intrinsic data dimension d_0 rather than the ambient dimension d_x.

Specifically, we establish sharp rates for velocity field approximation (Theorem 4.1), velocity estimation (Theorem 4.2), and distribution estimation in 2-Wasserstein distance (Theorem 4.3) for first-order flow-matching transformers under Assumption 2.1. These results demonstrate that flow-matching transformers mitigate the curse of dimensionality. Furthermore, we prove that these d_0-dependent rates are near minimax-optimal (Proposition 4.1), establishing the statistical efficiency of these models in the low-dimensional regime.

Extension to Higher Order Flow Matching Models. Our framework and analysis also extend to higher order flow matching models (Chen et al., 2025; Gong et al., 2025) (Section L), demonstrating the benefits of exploiting low-dimensional structure for higher-order dynamics. These findings provide strong theoretical backing for the empirical success of flow-matching transformers on high-dimensional data possessing low intrinsic dimensionality.

Limitations. Our analysis is currently grounded in the Low-Dimensional Linear Latent Subspace assumption. Extending these theoretical guarantees to general non-linear Riemannian manifolds, building upon initial efforts like Riemannian Flow Matching (Chen & Lipman, 2023), is a key next step. Furthermore, we assume the subspace matrix U is known, whereas in practice, its estimation or concurrent learning (e.g., via autoencoders) introduces error propagation that merits investigation.

Broader Impact

To be filled.

Acknowledgments

JH would like to thank Dino Feng and Andrew Chen for enlightening discussions on related topics, the Red Maple Family for support, and Jiayi Wang for facilitating experimental deployments. The authors would like to thank the anonymous reviewers and program chairs for constructive comments.

JH is partially supported by the Walter P. Murphy Fellowship. HL is partially supported by NIH R01LM1372201. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

Benton, J., Deligiannidis, G., and Doucet, A. Error bounds for flow matching methods. *arXiv preprint arXiv:2305.16860*, 2023.

Chen, B., Gong, C., Li, X., Liang, Y., Sha, Z., Shi, Z., Song, Z., and Wan, M. High-order matching for one-step shortcut diffusion models. *arXiv preprint arXiv:2502.00688*, 2025.

Chen, M., Huang, K., Zhao, T., and Wang, M. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 202:4672–4712, 2023a.

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

Chen, R. T. and Lipman, Y. Riemannian flow matching on general geometries, 2023.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y. W., and Doucet, A. Riemannian score-based generative modelling. *Advances in neural information processing systems*, 35:2406–2422, 2022.

Frans, K., Hafner, D., Levine, S., and Abbeel, P. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024.

Fu, H., Yang, Z., Wang, M., and Chen, M. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory, 2024.

Fukumizu, K., Suzuki, T., Isobe, N., Oko, K., and Koyama, M. Flow matching achieves almost minimax optimal convergence. *arXiv preprint arXiv:2405.20879*, 2024a.

Fukumizu, K., Suzuki, T., Isobe, N., Oko, K., and Koyama, M. Flow matching achieves almost minimax optimal convergence, 2024b.

Gong, C., Li, X., Liang, Y., Long, J., Shi, Z., Song, Z., and Tian, Y. Theoretical guarantees for high order trajectory refinement in generative flows. *arXiv preprint arXiv:2503.09069*, 2025.

Gronwall, T. H. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919. ISSN 0003486X, 19398980.

Haber, E., Ahamed, S., Siddiqui, M. S. R., Zakariaei, N., and Eliasof, M. Iterative flow matching–path correction and gradual refinement for enhanced generative modeling. *arXiv preprint arXiv:2502.16445*, 2025.

Hairer, E., Norsett, S., and Wanner, G. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8. Springer-Verlag, 01 1993. ISBN 978-3-540-56670-0. doi: 10.1007/978-3-540-78862-1.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Holderrieth, P., Havasi, M., Yim, J., Shaul, N., Gat, I., Jaakkola, T., Karrer, B., Chen, R. T. Q., and Lipman, Y. Generator matching: Generative modeling with arbitrary markov processes, 2025.

Hu, J. Y.-C., Wang, W.-P., Gilani, A., Li, C., Song, Z., and Liu, H. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. *arXiv preprint arXiv:2411.16525*, 2024a.

Hu, J. Y.-C., Wu, W., Li, Z., Pi, S., Song, Z., and Liu, H. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Hu, J. Y.-C., Wu, W., Lee, Y.-C., Huang, Y.-C., Chen, M., and Liu, H. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. In *International Conference on Learning Representations (ICLR)*, 2025.

Jiao, Y., Lai, Y., Wang, Y., and Yan, B. Convergence analysis of flow matching in latent space with transformers, 2024.

Kajitsuka, T. and Sato, I. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? *arXiv preprint arXiv:2307.14023*, 2023.

Kim, J., Kim, M., and Mozafari, B. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2022.

Kunkel, L. and Trabs, M. On the minimax optimality of flow matching through the connection to kernel density estimation. *arXiv preprint arXiv:2504.13336*, 2025.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling, 2023.

Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R. T. Q., Lopez-Paz, D., Ben-Hamu, H., and Gat, I. Flow matching guide and code, 2024.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*. OpenReview.net, 2023.

Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., and Cresswell, J. C. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *Transactions on Machine Learning Research (TMLR)*, 2024.

Niles-Weed, J. and Berthet, Q. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 2019.

Park, S., Lee, J., Yun, C., and Shin, J. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on learning theory*, pp. 3627–3661. PMLR, 2021.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., and Goldstein, T. The intrinsic dimension of images and its impact on learning. *International Conference on Learning Representations (ICLR)*, 2021.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Tang, R. and Yang, Y. Adaptivity of diffusion models to manifold structures. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1648–1656. PMLR, 2024.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019.

Supplementary Material