Learning Manifold Data with Flow Matching

Sophia Pi [*1] **Mingcheng Lu** [*] **Maojiang Su** [1] **Weimin Wu** [1] **Jerry Yao-Chieh Hu** [1] **Han Liu** [1 2]

Abstract

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

1. Introduction

We study the sample complexity of learning flow matching generative models when data lie on low-dimensional manifolds. Flow matching (Lipman et al., 2023; 2024) has emerged as a powerful framework for training continuous-time generative models. Instead of simulating sample trajectories or explicitly reversing a noising process as in diffusion models (Song & Ermon, 2019; Ho et al., 2020), flow matching learns a time-dependent velocity field that transports a simple source distribution to the target data distribution along a prescribed probability path. This simulation-free objective improves training stability and efficiency, and has become increasingly central to modern generative modeling. In parallel, Transformer architectures (Vaswani et al., 2017) have substantially advanced generative modeling, as exemplified by diffusion transformers operating on latent image patches (Peebles & Xie, 2023). These advances motivate a theoretical study of flow matching models equipped with expressive Transformer parameterizations, which we refer to as 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. For example, latent generative models compress data into lower-dimensional codes to simplify learning (Rombach et al., 2022). However, many existing guarantees for generative models still scale with the ambient dimension d_x, and therefore do not fully explain how these models benefit from low-dimensional structure. Recent work has started to address this: for score-based diffusion models, Chen et al. (2023) show that approximation and sampling errors can scale with the intrinsic dimension d_0 rather than d_x under a low-dimensional latent subspace assumption. However, analogous guarantees for flow matching methods are unexplored. In particular, it is unclear (i) whether flow-based models can achieve dimension-free statistical rates when data lie on a low-dimensional manifold, and (ii) how higher-order flow matching (which incorporates acceleration or higher derivatives) behaves in theory.

In this paper, we bridge these gaps by developing an end-to-end 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., 2023). We analyze both first-order flow matching (standard continuous flow models) and higher-order flow matching (augmented with Kth-order dynamics for $K \geq 2$) under this low-dimensional linear latent subspace assumption. Our analysis leverages a novel velocity decomposition: we split the generative velocity field into two components. One tangent to the data subspace and one orthogonal to it. Intuitively, this separates the dynamics that move points along the data manifold from those that push points onto the manifold. This decomposition holds for any order K of flow matching and is the key to tying model complexity to the intrinsic dimension d_0. Exploiting this decomposition, we prove that flow-matching transformer models indeed escape

[*]Equal contribution [1]Center for Foundation Models and Generative AI & Department of Computer Science, Northwestern University, Evanston, IL 60208, USA [2]Department of Statistics and Data Science, Northwestern University, Evanston, IL, USA. Correspondence to: Sophia Pi <sophiapi2026.1@u.northwestern.edu>, Mingcheng Lu <2860215400pp@gmail.com>, Maojiang Su <smj@u.northwestern.edu>, Weimin Wu <wwm@u.northwestern.edu>, Jerry Yao-Chieh Hu <jhu@u.northwestern.edu>, Han Liu <hanliu@northwestern.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

the curse of dimensionality: their approximation error, sample complexity, and distribution convergence rates depend only on d_0, not the ambient d_x.

Contributions. Our contributions are three-fold:

- **Explicit Tangent/Normal Velocity Decomposition.** We derive an explicit tangent/normal decomposition of the flow-matching velocity under the low-dimensional linear latent subspace assumption (Theorem 3.1). The tangent component captures transport within the data subspace, while the normal component is a simple linear term that pulls points toward the subspace.
- **Intrinsic-Dimension Statistical Guarantees.** Building on the decomposition, we establish the intrinsic-dimension statistical guarantees for flow-matching transformers. Specifically, we prove d_0-dependent rates for velocity approximation (Theorem 4.1), velocity estimation (Theorem 4.2), and distribution estimation under the 2-Wasserstein distance (Theorem 4.3). These rates avoid the ambient-dimensional d_x-dependence that appears in full-space analyses.
- **Extension to Higher-Order Flow Matching.** We show that our first-order flow velocity decomposition extends to higher-order flow matching models, which inherit d_0-dependent statistical rates.

Organization. Section 2 presents the mathematical flow matching foundation we build on. Section 3 presents our velocity decomposition trick for the first order and K-order flow matching. Section 4 presents our statistical analysis of first order flow matching transformers. Finally, we discuss our results and give concluding remarks in Section 5.

1.1. Related Work

Flow Matching for Generative Modeling. Flow matching trains continuous normalizing flows by regressing a time-dependent velocity field that transports a base distribution to the data along a prescribed probability path, thereby avoiding the trajectory simulation required by neural ODE training (Lipman et al., 2023; Chen et al., 2018). This simulation-free objective provides unbiased stochastic gradients, improves training stability and sample quality, and includes score-based diffusion as the special case induced by Gaussian noising paths (Lipman et al., 2023; 2024). Recent work augments the dynamics with acceleration or higher-order terms to refine trajectories for one-step and few-step generation (Gong et al., 2025; Chen et al., 2025), while complementary methods pursue iterative path correction and multi-step refinement (Haber et al., 2025). Flow matching has been scaled to a wide range of practical generative tasks, including high-resolution text-to-image synthesis with rectified-flow transformer, multilingual speech generation in Voicebox, unified audio generation in Audiobox, efficient video generation in Pyramid Flow, and 3D protein structure generation (Esser et al., 2024; Le et al., 2023; Vyas et al., 2023; Jin et al., 2025; Bose et al., 2024).

Statistical Theory for Flow Matching and Diffusion Models. The closest statistical foundation is the analysis of score-based diffusion under structural data assumptions. Chen et al. (2023) derive a tangent-normal decomposition of the score under a d_0-dimensional linear latent subspace. Their analysis gives approximation, estimation, and Wasserstein rates that depend on d_0 rather than the ambient dimension d_x, and thus circumvents the curse of dimensionality. Tang & Yang (2024) extend this perspective to smooth manifold support with adaptive minimax rates. Hu et al. (2024; 2025b) establish matching statistical and computational guarantees for diffusion transformers under related subspace assumptions. For flow matching in the full ambient space $\mathbb{R}^{d_x}$, Benton et al. (2024) bound Wasserstein error by L^2 velocity error, while Fukumizu et al. (2025); Kunkel & Trabs (2025) prove near-minimax-optimal Wasserstein convergence. Su et al. (2025) provide a unified framework and sharp rates for higher-order flow matching. These analyses, however, operate entirely in $\mathbb{R}^{d_x}$ and therefore scale with ambient dimension d_x. Jiao et al. (2024) study latent flow-matching transformers but excess risk of their encoder network's reconstruction still scales with the input dimension d_x. Whether flow matching escapes the curse of dimensionality under a manifold assumption has therefore remained open.

Generative Modeling on Low-Dimensional Manifolds. The manifold hypothesis posits that real-world data concentrate near a low-dimensional manifold embedded in a high-dimensional ambient space (Pope et al., 2021), motivating geometry-aware generative models. De Bortoli et al. (2022) define forward and reverse diffusions intrinsically on Riemannian manifolds. Chen & Lipman (2024) construct geometry-aware probability paths for Riemannian flow matching, enabling generation on curved spaces. On the theoretical side, the linear latent subspace model $x = Uh$ is the standard formalization for proving d_0-dependent rates (Chen et al., 2023; Hu et al., 2024; Jiao et al., 2024). Beyond that, Tang & Yang (2024); Zhang et al. (2026) study the more general smooth-manifold cases.

Our Contribution. We bridge these threads by deriving an explicit tangent-normal decomposition of the flow-matching velocity field under the linear latent subspace assumption, which is the flow-matching analogue of the score decomposition by Chen et al. (2023). This yields the first d_0-dependent approximation, estimation, and distribution-estimation guarantees for flow-matching transformers, in both first-order and higher-order settings.

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 $(X_t)_{0 \le t \le 1}$ 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

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

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 determines a unique flow ψ_t as the solution of an ordinary differential equation (ODE). In particular, ψ_t must satisfy

$$\frac{\mathrm{d}\psi_t}{\mathrm{d}t} = u_t(\psi_t(x)) \quad \text{with} \quad \psi_0(x) = x, \quad (2)$$

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

$$u_t(x) = \dot{\psi}_t(\psi_t^{-1}(x)), \quad \text{with} \quad \dot{\psi}_t = \frac{\mathrm{d}}{\mathrm{d}t}\psi_t, \quad (3)$$

which shows a theoretical method for computing u_t from ψ_t at the point x at time t. 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. 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

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t \sim U[0,1], X_t \sim p_t} \left[\| u_t^\theta(X_t) - u_t(X_t) \|_2^2 \right], \quad (4)$$

where $u_t^\theta(x)$ 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 corresponding conditional velocity fields $u_t(x|Z)$ and $p_t(x|Z)$, where $Z \in \mathbb{R}^m$ is an auxiliary random variable. To fit the original model, the marginal density and velocity should recover the original p_t and u_t via

$$p_t(x) = \int p_t(x|z) p_Z(z) \mathrm{d}z, \quad (5)$$

$$u_t(x) = \int u_t(x|z) \frac{p_t(x|z) p_Z(z)}{p_t(x)} \mathrm{d}z. \quad (6)$$

The Conditional Flow Matching (CFM) loss is defined as

$$\begin{aligned} &\mathcal{L}_{\mathrm{CFM}}(\theta) \quad (7) \\ &= \mathbb{E}_{t, Z \sim p_Z, X_t \sim p_{t|Z}(\cdot|Z)} \left[\| u_t^\theta(X_t) - u_t(X_t|Z) \|_2^2 \right]. \end{aligned}$$

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 path is constructed via the following interpolation between the source point x and the target sample x_1:

$$\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x, \quad (8)$$

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

$$\begin{aligned} &\mu_0 = \sigma_1 = 0, \mu_1 = \sigma_0 = 1, \text{ and} \\ &\dot{\mu}_t = \frac{\mathrm{d}\mu_t}{\mathrm{d}t} > 0, \dot{\sigma}_t = \frac{\mathrm{d}\sigma_t}{\mathrm{d}t} < 0 \text{ for } t \in (0,1). \quad (9) \end{aligned}$$

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

$$\begin{aligned} u_t(x|x_1) &= \dot{\psi}_t(\psi_t^{-1}(x|x_1)|x_1) \\ &= \frac{\dot{\sigma}_t(x - \mu_t x_1)}{\sigma_t} + \dot{\mu}_t x_1. \end{aligned}$$

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

$$\begin{aligned}&\mathcal{L}_{\mathrm{CFM}}(\theta) \qquad (10)\\ &= \mathop{\mathbb{E}}_{t,X_1\sim p_1,X_0\sim p_0}[\|u_t^\theta(\mu_t X_1+\sigma_t X_0)-(\dot\mu_t X_1+\dot\sigma_t X_0)\|_2^2].\end{aligned}$$

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:

$$\widehat{\mathcal{L}}_{\mathrm{CFM}}(u_\theta) := \frac{1}{n}\sum_{i=1}^n \int_{t_0}^T \frac{1}{T-t_0}\mathbb{E}_{X_0\sim N(0,I)}\left[l_t\right]\mathrm{d}t, \quad (11)$$

where

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

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. Higher-Order Flow Matching Preliminaries

We briefly recall the higher-order flow matching (HOFM) following (Su et al., 2025). Fix an integer $K \geq 2$. A K-order flow augments the usual state with time-derivatives of the trajectory:

$$\begin{aligned}y_t &:= \left(x_t^{(0)}, x_t^{(1)}, \ldots, x_t^{(K-1)}\right) \in \mathbb{R}^{Kd_x},\\ x_t^{(0)} &:= \psi_t(x), \quad x_t^{(k)} := \frac{\mathrm{d}^k}{\mathrm{d}t^k}\psi_t(x).\end{aligned}$$

The corresponding K-order velocity field $f_t(y_t) \in \mathbb{R}^{Kd_x}$ is

$$\frac{\mathrm{d}}{\mathrm{d}t}y_t = f_t(y_t) := \left(u_t^{(1)}(x_t^{(0)},t),\ldots,u_t^{(K)}(x_t^{(0)},t)\right), \quad (12)$$

where $u_t^{(k)}$ denotes the k-th order velocity (e.g. $u_t^{(1)}$ is the usual velocity, $u_t^{(2)}$ is acceleration, etc.). The ODE (12) imposes *total-derivative constraints* tying the orders together along trajectories ($k \geq 2$):

$$\begin{aligned}&u_t^{(k)}(x_t^{(0)},t) \qquad (13)\\ &= \frac{\mathrm{d}}{\mathrm{d}t}u_t^{(k-1)}(x_t^{(0)},t)\\ &= \partial_t u_t^{(k-1)}(x_t^{(0)},t) + \nabla_x u_t^{(k-1)}(x_t^{(0)},t)\cdot u_t^{(1)}(x_t^{(0)},t).\end{aligned}$$

The population K-order flow matching objective is the regression loss

$$\mathcal{L}_{\mathrm{FM}}^K(\Theta) := \mathop{\mathbb{E}}_{t,\,Y_t}\left[\left\|f_t(Y_t)-f_t^\Theta(Y_t)\right\|_2^2\right],$$

but f_t is intractable. As in first-order flow matching, we therefore train via a conditional objective using a tractable conditional path. In particular, for an affine conditional flow

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

the k-th order conditional target is the k-th time derivative,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}X_t = \mu_t^{(k)}X_1 + \sigma_t^{(k)}X_0,$$

where $\mu_t^{(k)} := \frac{\mathrm{d}^k\mu_t}{\mathrm{d}t^k}$ and $\sigma_t^{(k)} := \frac{\mathrm{d}^k\sigma_t}{\mathrm{d}t^k}$. This yields the K-order conditional flow matching loss

$$\begin{aligned}&\mathcal{L}_{\mathrm{CFM}}^K(\Theta) \qquad (14)\\ &:= \sum_{k=1}^K \mathop{\mathbb{E}}_{t,\,X_1,\,X_0}\left[\left\|u_\Theta^{(k)}(X_t,t)-\left(\mu_t^{(k)}X_1+\sigma_t^{(k)}X_0\right)\right\|_2^2\right].\end{aligned}$$

Moreover, when the dissimilarity is a Bregman divergence (in particular squared ℓ_2), the gradients of $\mathcal{L}_{\mathrm{FM}}^K$ and $\mathcal{L}_{\mathrm{CFM}}^K$ coincide, so (14) is a valid training surrogate for the higher-order FM objective.

2.3. Manifold Assumption

Following (Chen et al., 2023), 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)**.** *The target 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.*

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., 2023; Hu et al., 2024; 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. (2023) investigate the approximation, estimation, and distribution recovery of diffusion models under manifold assumption. Building on similar assumptions, Hu et al. (2024) analyze 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.

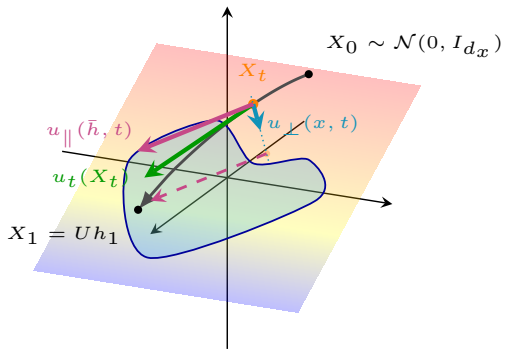

Figure 1. An illustration of the velocity decomposition in ambient dimension $d_x = 3$ with data manifold in a linear latent subspace with dimension $d_0 = 2$. We depict the flow path from X_0 to X_1 with the curved gray arrow. The green arrow $u_t(X_t)$ represents the velocity along the path at time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

2.4. Transformer Networks

We defer the standard definition of transformer networks to Section C 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., 2023; 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.

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 $x = Uh$ satisfy 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 (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 (10) admits the decomposition

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

where p_t^h is the marginal density of $\bar{h}$ and coefficients satisfy

$$\alpha_t := \kappa_t + \lambda_t/\mu_t, \quad \beta_t := \lambda_t \sigma_t^2/\mu_t.$$

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 A for a detailed proof. □

We visualize the decomposition of the velocity in Figure 1.

Remark 3.1 (Flow-Matching Analogue of Score Decomposition). *Theorem 3.1 is the flow-matching analogue of the score decomposition of (Chen et al., 2023). Two structural differences are worth noting. First, our result applies to any affine conditional flow satisfying (9), subsuming the Gaussian noising path used by diffusion. Second, the orthogonal component here is a closed-form linear function of $x_\perp$ with no learnable nonlinearity, so the entire statistical complexity of velocity matching concentrates on the d_0-dimensional latent score. This separation is what drives the intrinsic-dimension rates in later sections.*

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 (9). We define the following constants

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

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

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

where p_t^h is the marginal density of $\bar{h}$ and coefficients satisfy $\alpha_{k,t} := \kappa_{k,t} + \lambda_{k,t}/\mu_t$ and $\beta_{k,t} := \lambda_{k,t}\sigma_t^2/\mu_t$.

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

Remark 3.2. *Similar to Theorem 3.1, the decomposition in Theorem 3.2 shows that, for every order k, the k-th order velocity field $u_t^{(k)}$ splits into a tangent component in $\mathrm{span}(U)$ and a normal component in $\mathrm{span}(U)^\perp$:*

$$u_t^{(k)}(x) = u_\parallel^{(k)}(\bar{h}, t) + u_\perp^{(k)}(x, t),$$

where $u_\parallel^{(k)}(\bar{h}, t) \in \mathrm{span}(U)$ and $u_\perp^{(k)}(x,t) \in \mathrm{span}(U)^\perp$. We note that the normal term is always linear in the off-subspace coordinate $x_\perp = (I - UU^\top)x$, namely $u_\perp^{(k)}(x,t) = \kappa_{k,t} x_\perp$. Thus, higher-order flow matching inherits the same conceptual roles as first-order flow matching in that it is also composed of a statistically nontrivial on-subspace transport term and a simple off-subspace component.

4. Main Theoretical Results (Intrinsic Dimension Analysis)

In this section, we establish statistical rates of flow matching transformers under the manifold assumption Assumption 2.1. Specifically, Section 4.1 presents velocity approximation and 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.

Proof Strategy and Role of Velocity Decomposition. The derivation of our statistical rates (Theorem 4.1, Theorem 4.2, and Theorem 4.3) 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 $u_\parallel(\bar{h}, t)$ in the d_0-dimensional latent subspace. The dynamics in the orthogonal complement $u_\perp(x,t)$ 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 $u_\parallel(\bar{h}, t)$ 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.

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.

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 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 (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*

$$p_1^h(h_1) \le C_1 \exp\left(-C_2 \|h_1\|_2^2/2\right),$$

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 2.1.

Theorem 4.1 (Velocity Approximation with Transformers under manifold assumptions Assumption 2.1)**.** *Assume Assumption 2.1 , Assumption 4.1 and Assumption 4.2. For any $\beta > 0$ and sufficiently large $N \in \mathbb{N}$, there exists a transformer $\widehat{s} \in \mathcal{T}_R^{h,s,r}$ on $\mathbb{R}^{d_0} \times [t_0, T]$ such that the structured velocity field*

$$u_\theta(x,t) := U[\alpha_t U^\top x + \beta_t \widehat{s}(U^\top x, t)] + \kappa_t (I_{d_x} - UU^\top)x$$

satisfies the integrated approximation error bound

$$\int_{t_0}^{T} \int_{\mathbb{R}^{d_x}} \|u_t^\star(x) - u_\theta(x,t)\|_2^2 p_t(x) \mathrm{d}x \mathrm{d}t = O(B^2 N^{-\beta} (\log N)^{d_0 + \beta/2 + 1}).$$

Writing the input shape as $d_0 + 1 = d \times L$ (patch size times sequence length), the transformer parameter bounds satisfy

$$C_{KQ}, C_{KQ}^{2,\infty} = O((\log N)^{2d+1} N^{(4d+2)\beta}),$$
$$C_{OV}, C_{OV}^{2,\infty} = O(N^{-\beta}), C_F, \quad C_F^{2,\infty} = O((\log N) N^{\beta}),$$
$$C_E = O(1), \quad C_T = O(\sqrt{\log N}).$$

The constants in $O(\cdot)$ depend on d_0, β, B, C_1, C_2.

Proof. See Section E 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 (11). 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 $\mathbb{R}^{d_x}$). 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 $X_t \sim p_t$. 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 (11).

Subspace Recovery. In practice the matrix U is unknown, but we're capable of recovering the subspace from samples. Let $\widehat{\Sigma}_x := n^{-1} \sum_{i=1}^n x_i x_i^\top$ and let $\widehat{U} \in \mathbb{R}^{d_x \times d_0}$ collect its top-d_0 orthonormal eigenvectors. In Section F.1 we show that $\widehat{U}\widehat{U}^\top = UU^\top$ almost surely. Since $U^\top$ enters structured velocity only through the rotation-invariant pair $UU^\top$, replacing U by $\widehat{U}$ leaves the data-driven structured estimator

$$\widehat{u}_\theta(x,t) = \widehat{U}[\alpha_t \widehat{U}^\top x + \beta_t \widehat{s}_\theta(\widehat{U}^\top x, t)] + \kappa_t (I - \widehat{U}\widehat{U}^\top)x \tag{15}$$

in the same class as structured velocity. All subsequent rates are unaffected. Please see Section F.1 for details.

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, Assumption 4.1 and Assumption 4.2. Let $d \times L = d_0 + 1$ be the input shape used by the transformer. Suppose we choose the transformer as in Theorem 4.1. Then,*

for sample size $n \geq d_0$ it holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n}[\mathcal{R}(\widehat{u}_\theta)] = O\left(n^{-\frac{1}{16d+15}}(\log n)^{\max\{d_0+\frac{1}{2}\beta+1,(8d+16)/3\}}\right).$$

Proof. See Section F 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 Wasserstein-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 (11).

Theorem 4.3 (Distribution Estimation With Wasserstein Distance Under Assumption 2.1)**.** *Let $\widehat{P}_T$ be the distribution obtained at clipped time $[t_0, T]$ by running the flow driven by the learned velocity field $\widehat{u}_\theta$. Assume Assumption 2.1, Assumption 4.1 and Assumption 4.2. Then, for sample size $n \geq d_0$,*

$$\mathop{\mathbb{E}}_{\{x_i\}_{i=1}^n}\left[W_2\left(\widehat{P}_T, P_T\right)\right] = O\left(n^{-\frac{1}{32d+30}}(\log n)^{\frac{1}{2}\max\{d_0+\beta/2+1,\,(8d+16)/3\}}\right).$$

Proof. See Section G for a detailed proof. □

Remark 4.1 (Escaping the Curse of Dimensionality)**.** *Theorem 4.3 provides the first distribution-estimation guarantee for flow-matching transformers under the manifold assumption that is free of the ambient input dimension when the intrinsic dimension is fixed. The W_2 convergence rate depends on patch dimension d determined by the intrinsic dimension d_0 instead of the ambient dimension d_x. By contrast, existing flow-matching distribution-estimation analyses that operate in the full ambient space $\mathbb{R}^{d_x}$ (Fukumizu et al., 2025; Su et al., 2025) yield rates with exponents scaling in d_x, which deteriorates exponentially as d_x grows.*

Remark 4.2 (Extension to High-Order Flow Matching)**.** *Theorem 3.2 gives, for high-order flow matching satisfying $k \geq 2$, the same velocity structure holds as the first-order decomposition: a d_0-dimensional latent-transport term involving $\nabla_{\bar{h}} \log p_t^h$ and a closed-form orthogonal contraction in $x_\perp$. Consequently, after replacing u_t by $u_t^{(k)}$ and changing coefficients accordingly, the proofs of Theorems 4.1 to 4.3 carry over without modification. Therefore, the same d_0-dependent approximation and statistical estimation hold for high-order flow matching satisfying $k \geq 2$.*

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 explicit d_0-dependent 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 show that flow-matching transformers avoid ambient-dimensional dependence and mitigate the curse of dimensionality under the low-dimensional manifold assumption.

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), 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, 2024), is a key next step.

Finally, while our results analyze the statistical rates of higher-order flow matching models under Assumption 2.1, the precise statistical advantages of increasing order K remain an open question; potential benefits in sampling efficiency or numerical stability due to improved trajectory properties warrant further study.

Impact Statement

By the theoretical nature of this work, we do not anticipate any negative social impact.

References

Benton, J., Deligiannidis, G., and Doucet, A. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2024.

Bose, J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C., Nica, A., Korablyov, M., Bronstein, M., and Tong, A. Se (3)-stochastic flow matching for protein backbone generation. In *International Conference on Learning Representations*, volume 2024, pp. 22590–22621, 2024.

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, 2023.

Chen, R. T. Q. and Lipman, Y. Flow matching on general geometries. In Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., and Sun, Y. (eds.), *International Conference on Learning Representations*, volume 2024, pp. 47922–47945, 2024.

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.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 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. In *International Conference on Learning Representations*, volume 2025, pp. 27608–27640, 2025.

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.

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.

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

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)*, 2024.

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. In *International Conference on Learning Representations*, volume 2025, pp. 29634–29686, 2025a.

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)*, 2025b.

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

Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., and Lin, Z. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, volume 2025, pp. 23378–23402, 2025.

Kajitsuka, T. and Sato, I. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? In *International Conference on Learning Representations*, volume 2024, pp. 35177–35205, 2024.

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.

Le, M., Vyas, A., Shi, B., Karrer, B., Sari, L., Moritz, R., Williamson, M., Manohar, V., Adi, Y., Mahadeokar, J., et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

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.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 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.

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.

Su, M., Hu, J. Y.-C., Lee, Y.-C., Zhu, N., Chung, J.-H., Wu, S., Song, Z., Chen, M., and Liu, H. High-order flow matching: Unified framework and sharp statistical rates. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.-C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? In *International Conference on Learning Representations*, 2019.

Zhang, Z., Huang, K., Zhao, T., Wang, M., and Chen, M. Diffusion model for manifold data: Score decomposition, curvature, and statistical complexity. *arXiv preprint arXiv:2603.20645*, 2026.

Supplementary Material

LLM Usage Disclosure

We used large language models (LLMs) to aid and polish writing, such as improving clarity, grammar, and conciseness. We also used LLMs for retrieval and discovery, for example exhausting literature to identify potential missing related work. All technical content, proofs, experiments, and results are original contributions by the authors.

A. Proof of Theorem 3.1

In this section, we provide the detailed proofs of the velocity decomposition lemmas and theorems presented in section Section 3.

First, we introduce the Generalized Tweedie's Formula.

Lemma A.1 (Generalized Tweedie's Formula)**.** *Let $x \in \mathbb{R}^d$ be with prior density $p(x)$. We consider the linear Gaussian channel:*

$$y = Ax + \sigma\epsilon,$$

where $A \in \mathbb{R}^{m\times d}$ has full column rank, $\epsilon \sim \mathcal{N}(0, I_m)$ is standard Gaussian noise, and $\sigma > 0$ is the noise level. If $p(y)$ is continuously differentiable, then the posterior mean $\mathbb{E}[x \mid y]$ satisfies

$$\mathbb{E}[x \mid y] = (A^\top A)^{-1}A^\top \left(y + \sigma^2 \nabla_y \log p(y)\right). \tag{16}$$

We also introduce the following helper lemma.

Lemma A.2 (Expectation Form of Flow Matching Velocity)**.** *Suppose $X_t = \psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$ is an affine conditional flow satisfying the conditions given in (9). Then the induced (marginal/optimal) velocity field can be written as*

$$u_t(x) = \dot{\mu}_t \mathbb{E}[X_1|X_t = x] + \dot{\sigma}_t \mathbb{E}[X_0|X_t = x].$$

Proof. According to (8), the affine path for a fixed target point x_1 is

$$\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x.$$

This induces the corresponding velocity field:

$$\begin{aligned} u_t(x|x_1) &= \frac{\mathrm{d}}{\mathrm{d}t}\psi_t(x|x_1) \\ &= \dot{\psi}_t(\psi_t^{-1}(x|x_1)|x_1) \\ &= \frac{\dot{\sigma}_t(x - \mu_t x_1)}{\sigma_t} + \dot{\mu}_t x_1. \end{aligned} \tag{17}$$

Now we consider $X_0 \sim \mathcal{N}(0, I_{d_x})$ and $X_1 \sim q$ as random variables and write

$$X_t = \mu_t X_1 + \sigma_t X_0.$$

For each pair (X_1, X_0), (17) gives

$$u_t(X_t|X_1) = \frac{\dot{\sigma}_t}{\sigma_t}(X_t - \mu_t X_1) + \dot{\mu}_t X_1. \tag{18}$$

Then we define the marginal velocity field

$$u_t(x) := \mathbb{E}[u_t(X_t|X_1)|X_t = x]. \tag{19}$$

Taking conditional expectations on both sides of (18) yields:

$$\begin{aligned} \mathbb{E}[u_t(X_t|X_1)|X_t = x] &= \dot{\sigma}_t \frac{\mathbb{E}[X_t - \mu_t X_1|X_t = x]}{\sigma_t} + \dot{\mu}_t \mathbb{E}[X_1|X_t = x], \\ u_t(x) &= \dot{\sigma}_t \frac{x - \mu_t \mathbb{E}[X_1|X_t = x]}{\sigma_t} + \dot{\mu}_t \mathbb{E}[X_1|X_t = x], \end{aligned} \tag{20}$$

where the LHS follows from (19) and the RHS follows from linearity of conditional expectations.

Recall $X_t = \mu_t X_1 + \sigma_t X_0$. Solving for X_0 yields

$$X_0 = \frac{1}{\sigma_t}(X_t - \mu_t X_1),$$

and hence

$$\mathbb{E}[X_0|X_t = x] = \frac{x - \mu_t \mathbb{E}[X_1|X_t = x]}{\sigma_t},$$

which is precisely what appears on the RHS of (20). Hence we obtain

$$u_t(x) = \dot{\sigma}_t \mathbb{E}[X_0|X_t = x] + \dot{\mu}_t \mathbb{E}[X_1|X_t = x],$$

as desired. This completes the proof. □

Now we prove our main 1st-order velocity decomposition theorem.

Theorem A.1 (Velocity Decomposition under the Low Dimensional Linear Latent Subspace Assumption, Theorem 3.1 Formal Version)**.** *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*

$$X_t = \psi_t(X_0 \mid X_1) = \mu_t X_1 + \sigma_t X_0, \qquad (X_1, X_0) \sim (q, N(0, I_{d_x})) \tag{21}$$

with smooth coefficients $\mu_t, \sigma_t \in (0, 1)$ that satisfy (9). Write $\dot{\mu}_t = \frac{\mathrm{d}}{\mathrm{d}t}\mu_t$ and $\dot{\sigma}_t = \frac{\mathrm{d}}{\mathrm{d}t}\sigma_t$, and define the constants

$$\kappa_t := \frac{\dot{\sigma}_t}{\sigma_t}, \qquad \lambda_t := \dot{\mu}_t - \mu_t \kappa_t. \tag{22}$$

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 velocity field that appears in the conditional flow–matching objective (10) admits the decomposition

$$u_t(x) = \underbrace{U\Big[\kappa_t \bar{h} + \lambda_t \,\mathbb{E}\left[h|\bar{h}\right]\Big]}_{u_\parallel(\bar{h},t)} + \underbrace{\kappa_t (I - UU^\top)x}_{u_\perp(x,t)}. \tag{23}$$

Moreover, by Tweedie's formula in latent space,

$$\mathbb{E}\left[h|\bar{h}\right] = \frac{1}{\mu_t}\Big(\bar{h} + \sigma_t^2 \nabla_{\bar{h}} \log p_t^h(\bar{h})\Big), \tag{24}$$

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

$$u_t(x) = \underbrace{U\Big[\alpha_t \bar{h} + \beta_t \nabla_{\bar{h}} \log p_t^h(\bar{h})\Big]}_{u_\parallel(\bar{h},t)\text{: on-support component}} + \underbrace{\kappa_t (I - UU^\top)x}_{u_\perp(x,t)\text{: orthogonal component}}, \tag{25}$$

with coefficients $\alpha_t := \kappa_t + \lambda_t/\mu_t$ and $\beta_t := \lambda_t \sigma_t^2/\mu_t$.

Proof. Observe that $X_0 \sim \mathcal{N}(0, I_{d_x})$ is independent of h and Gaussian. Then we write

$$\begin{aligned}
\bar{h} &= U^\top X_t \\
&= U^\top(\mu_t X_1 + \sigma_t X_0) \\
&= U^\top(\mu_t U h + \sigma_t X_0) \\
&= \mu_t U^\top U h + \sigma_t U^\top X_0 \\
&= \mu_t h + \sigma_t U^\top X_0 \\
&= \mu_t h + \sigma_t \epsilon,
\end{aligned} \tag{26}$$

where $\epsilon := U^\top X_0 \sim \mathcal{N}(0, I_{d_0})$. The first line is by definition, the second line is by (21), the third line is by Assumption 2.1, the fourth line is ordinary matrix multiplication, the fifth line is by Assumption 2.1 (in particular that U has orthonormal columns), and the sixth line is by the Gaussian-ness of X_0 (and that U has orthonormal columns).

Hence, we can write the density of $\bar{h}$ as the convolution

$$p_t^h := \mathcal{N}(\mu_t h, \sigma_t^2 I_{d_0}) * P_1^h.$$

Now we consider $\mathbb{E}[h|\bar{h}]$. From (26), $\bar{h} = \mu_t h + \sigma_t \epsilon$, so by Lemma A.1,

$$\begin{aligned}
\mathbb{E}[h|\bar{h}] &= ((\mu_t I_{d_0})^\top (\mu_t I_{d_0}))^{-1} (\mu_t I_{d_0})^\top (\bar{h} + \sigma_t^2 \nabla_{\bar{h}} \log p_t^h(\bar{h})) \\
&= (\mu_t^{-1} I_{d_0})(\bar{h} + \sigma_t^2 \nabla_{\bar{h}} \log p_t^h(\bar{h})) \\
&= \frac{1}{\mu_t}(\bar{h} + \sigma_t^2 \nabla_{\bar{h}} \log p_t^h(\bar{h})).
\end{aligned} \tag{27}$$

Note that $x_\perp$ is independent of h. This is an intuitively obvious fact that we show explicitly:

$$\begin{aligned}
x_\perp &= (I_{d_x} - UU^\top) X_t \\
&= (I_{d_x} - UU^\top)(\mu_t X_1 + \sigma_t X_0) \\
&= (I_{d_x} - UU^\top)\mu_t U h + (I_{d_x} - UU^\top)\sigma_t X_0 \\
&= \mu_t U h - \mu_t U U^\top U h + \sigma_t (I_{d_x} - UU^\top) X_0 \\
&= \sigma_t (I_{d_x} - UU^\top) X_0,
\end{aligned}$$

where the first line is by definition, the second line is by (21), the third line is by Assumption 2.1, the fourth line is ordinary matrix multiplication, and the fifth line is by the Assumption 2.1 (and in particular that U has orthonormal columns).

Next, we consider $\mathbb{E}[X_1|X_t = x]$ and $\mathbb{E}[X_0|X_t = x]$. Conditioning on the full vector x is equivalent to conditioning on the pair $(\bar{h}, x_\perp)$ (by definition), and since $x_\perp$ is independent of h, we have $p(h|x) = p(h|\bar{h})$.

Hence

$$\begin{aligned}\mathbb{E}[X_1|X_t = x] &= \mathbb{E}[Uh|X_t = x] \\ &= U\mathbb{E}[h|X_t = x] \\ &= U\mathbb{E}[h|\bar{h}],\end{aligned} \tag{28}$$

where the first line is by Assumption 2.1, the second line is by linearity of expectations, and the third line is by the fact that $p(h|\bar{h}) = p(h|x)$. Also,

$$\begin{aligned}\mathbb{E}[X_0|X_t = x] &= \mathbb{E}\left[\frac{1}{\sigma_t}X_t - \frac{\mu_t}{\sigma_t}X_1 \middle| X_t = x\right] \\ &= \frac{x}{\sigma_t} - \frac{\mu_t}{\sigma_t}\mathbb{E}[X_1|X_t = x] \\ &= \frac{U\bar{h} + x_\perp}{\sigma_t} - \frac{\mu_t}{\sigma_t}U\mathbb{E}[h|\bar{h}] \\ &= \frac{x_\perp}{\sigma_t} + U\left[\frac{\bar{h}}{\sigma_t} - \frac{\mu_t}{\sigma_t}\mathbb{E}[h|\bar{h}]\right],\end{aligned} \tag{29}$$

where the first line is by rearranging (21) to solve for X_0, the second line is by linearity of expectations, the third line is by definition of $\bar{h}$ and $x_\perp$ (in particular that $x = U\bar{h} + x_\perp$) and Assumption 2.1, and the fourth line is simply by collecting common factors.

Now, by Lemma A.2, we have

$$u_t(x) = \dot{\mu}_t\mathbb{E}[X_1|X_t = x] + \dot{\sigma}_t\mathbb{E}[X_0|X_t = x].$$

Plugging in (29) and (28), we obtain:

$$\begin{aligned}u_t(x) &= \dot{\mu}_t U\mathbb{E}[h|\bar{h}] + \dot{\sigma}_t\left(\frac{x_\perp}{\sigma_t} + U\left[\frac{\bar{h}}{\sigma_t} - \frac{\mu_t}{\sigma_t}\mathbb{E}[h|\bar{h}]\right]\right) \\ &= U\left[\dot{\mu}_t\mathbb{E}[h|\bar{h}] + \frac{\dot{\sigma}_t}{\sigma_t}\bar{h} - \frac{\dot{\sigma}_t}{\sigma_t}\mu_t\mathbb{E}[h|\bar{h}]\right] + \frac{\dot{\sigma}_t}{\sigma_t}x_\perp \\ &= U\left[\left(\dot{\mu}_t - \frac{\dot{\sigma}_t}{\sigma_t}\mu_t\right)\mathbb{E}[h|\bar{h}] + \frac{\dot{\sigma}_t}{\sigma_t}\bar{h}\right] + \frac{\dot{\sigma}_t}{\sigma_t}x_\perp \\ &= U\left[\left(\dot{\mu}_t - \kappa_t\mu_t\right)\mathbb{E}[h|\bar{h}] + \kappa_t\bar{h}\right] + \kappa_t x_\perp \\ &= U\left[\lambda_t\mathbb{E}[h|\bar{h}] + \kappa_t\bar{h}\right] + \kappa_t x_\perp,\end{aligned}$$

where the first line is by (20) and (28), the second and third lines are simply collecting common terms, and the fourth and fifth lines are by definition of κ_t and λ_t in (22).

Now, plugging in $\mathbb{E}[h|\bar{h}]$ from (27), we arrive at the final decomposition

$$\begin{aligned}u_t(x) &= U\left[\lambda_t\left(\frac{1}{\mu_t}(\bar{h} + \sigma_t^2\nabla_{\bar{h}}\log p_t^h(\bar{h}))\right) + \kappa_t\bar{h}\right] + \kappa_t x_\perp \\ &= U\left[\left(\kappa_t + \frac{\lambda_t}{\mu_t}\right)\bar{h} + \frac{\lambda_t\sigma_t^2}{\mu_t}\nabla_{\bar{h}}\log p_t^h(\bar{h})\right] + \kappa_t x_\perp \\ &= U\left[\left(\kappa_t + \frac{\lambda_t}{\mu_t}\right)\bar{h} + \frac{\lambda_t\sigma_t^2}{\mu_t}\nabla_{\bar{h}}\log p_t^h(\bar{h})\right] + \kappa_t(I_{d_x} - UU^\top)x \\ &= U\left[\alpha_t\bar{h} + \beta_t\nabla_{\bar{h}}\log p_t^h(\bar{h})\right] + \kappa_t(I_{d_x} - UU^\top)x,\end{aligned}$$

where $\alpha_t := \kappa_t + \lambda_t/\mu_t$ and $\beta_t := \lambda_t\sigma_t^2/\mu_t$. The first line follows from direct substitution of (27), the second line is simply collecting common factors, the third line is from the definition of $x_\perp$, and the fourth line is from the definition of α_t and β_t.

This completes the proof. □

B. Proof of Theorem 3.2

Under Assumption 2.1, we decompose the k-th order velocity into its on-support and orthogonal components.

We introduce the following helper lemma.

Lemma B.1 (Expectation Form of k-th Order Flow Matching Velocity)**.** *Suppose $X_t = \psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$ is an affine conditional flow satisfying the conditions given in (9). Then the induced marginal k-th order velocity field can be written as*

$$u_t^{(k)}(x) := \mathbb{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}t^k} X_t \middle| X_t = x\right] = \mu_t^{(k)} \mathbb{E}[X_1|X_t = x] + \sigma_t^{(k)} \mathbb{E}[X_0|X_t = x].$$

Proof. The proof is exactly analogous to the proof of Lemma A.2. Since μ_t and σ_t are scalar functions of t,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\mu_t X_1) = \mu_t^{(k)} X_1,$$

and similarly,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k}(\sigma_t X_0) = \sigma_t^{(k)} X_0.$$

Thus, since all mixed derivatives $\binom{k}{j}\mu_t^{(j)} \frac{\mathrm{d}^{k-j}}{\mathrm{d}t^{(k-j)}} X_1$ vanish for $j < k$,

$$\frac{\mathrm{d}^k}{\mathrm{d}t^k} X_t = \mu_t^{(k)} X_1 + \sigma_t^{(k)} X_0.$$

Then we define the marginal velocity field

$$u_t^{(k)}(x) := \mathbb{E}\left[\frac{\mathrm{d}^k}{\mathrm{d}t^k} X_t \middle| X_t = x\right] = \mu_t^{(k)} \mathbb{E}[X_1|X_t = x] + \sigma_t^{(k)} \mathbb{E}[X_0|X_t = x].$$

This completes the proof. □

Now we prove our main k-th order velocity decomposition theorem.

Theorem B.1 (k-th Order Velocity Decomposition under the Low Dimensional Linear Latent Subspace Assumption, Theorem 3.2 Formal Version)**.** *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*

$$X_t = \psi_t(X_0 \mid X_1) = \mu_t X_1 + \sigma_t X_0, \qquad (X_1, X_0) \sim (q, N(0, I_{d_x})) \tag{30}$$

with smooth coefficients $\mu_t, \sigma_t \in (0,1)$ that satisfy (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

$$\kappa_{k,t} := \frac{\sigma_t^{(k)}}{\sigma_t}, \qquad \lambda_{k,t} := \mu_t^{(k)} - \mu_t \kappa_{k,t}. \tag{31}$$

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 (14) admits the decomposition

$$u_t^{(k)}(x) = \underbrace{U\Big[\kappa_{k,t}\bar{h} + \lambda_{k,t}\,\mathbb{E}\left[h\middle|\bar{h}\right]\Big]}_{u_\parallel^{(k)}(\bar{h},t)} + \underbrace{\kappa_{k,t}(I - UU^\top)x}_{u_\perp^{(k)}(x,t)}. \tag{32}$$

Moreover, by Tweedie's formula in latent space,

$$\mathbb{E}\left[h\middle|\bar{h}\right] = \frac{1}{\mu_t}\Big(\bar{h} + \sigma_t^2 \nabla_{\bar{h}} \log p_t^h(\bar{h})\Big), \tag{33}$$

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

$$u_t^{(k)}(x) = \underbrace{U\Big[\alpha_{k,t}\bar{h} + \beta_{k,t}\nabla_{\bar{h}} \log p_t^h(\bar{h})\Big]}_{u_\parallel^{(k)}(\bar{h},t)\text{: } k\text{-th order on-support component}} + \underbrace{\kappa_{k,t}(I - UU^\top)x}_{u_\perp^{(k)}(x,t)\text{: } k\text{-th order orthogonal component}}, \tag{34}$$

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. The proof builds off much of the proof of Theorem 3.1. From Lemma B.1, we have

$$u_t^{(k)}(x) = \mu_t^{(k)}\mathbb{E}[X_1|X_t = x] + \sigma_t^{(k)}\mathbb{E}[X_0|X_t = x].$$

Plugging in (28) and (29) from the proof of Theorem 3.1, we obtain:

$$\begin{aligned}
u_t^{(k)}(x) &= \mu_t^{(k)} U\mathbb{E}[h|\bar{h}] + \sigma_t^{(k)}\left(\frac{x_\perp}{\sigma_t} + U\left[\frac{\bar{h}}{\sigma_t} - \frac{\mu_t}{\sigma_t}\mathbb{E}[h|\bar{h}]\right]\right) \\
&= U\left[\mu_t^{(k)}\mathbb{E}[h|\bar{h}] + \frac{\sigma_t^{(k)}}{\sigma_t}\bar{h} - \frac{\sigma_t^{(k)}}{\sigma_t}\mu_t\mathbb{E}[h|\bar{h}]\right] + \frac{\sigma_t^{(k)}}{\sigma_t}x_\perp \\
&= U\left[\left(\mu_t^{(k)} - \frac{\sigma_t^{(k)}}{\sigma_t}\mu_t\right)\mathbb{E}[h|\bar{h}] + \frac{\sigma_t^{(k)}}{\sigma_t}\bar{h}\right] + \frac{\sigma_t^{(k)}}{\sigma_t}x_\perp \\
&= U\left[\left(\mu_t^{(k)} - \kappa_{k,t}\mu_t\right)\mathbb{E}[h|\bar{h}] + \kappa_{k,t}\bar{h}\right] + \kappa_{k,t}x_\perp \\
&= U\left[\lambda_{k,t}\mathbb{E}[h|\bar{h}] + \kappa_{k,t}\bar{h}\right] + \kappa_{k,t}x_\perp,
\end{aligned}$$

where the first line is by (20) and (28), the second and third lines are simply collecting common terms, and the fourth and fifth lines are by definition of $\kappa_{k,t}$ and $\lambda_{k,t}$ in (31).

Now, plugging in $\mathbb{E}[h|\bar{h}]$ from (27), we arrive at the final decomposition

$$\begin{aligned}
u_t^{(k)}(x) &= U\left[\lambda_{k,t}\left(\frac{1}{\mu_t}(\bar{h} + \sigma_t^2\nabla_{\bar{h}}\log p_t^h(\bar{h}))\right) + \kappa_{k,t}\bar{h}\right] + \kappa_{k,t}x_\perp \\
&= U\left[\left(\kappa_{k,t} + \frac{\lambda_{k,t}}{\mu_t}\right)\bar{h} + \frac{\lambda_{k,t}\sigma_t^2}{\mu_t}\nabla_{\bar{h}}\log p_t^h(\bar{h})\right] + \kappa_{k,t}x_\perp \\
&= U\left[\left(\kappa_{k,t} + \frac{\lambda_{k,t}}{\mu_t}\right)\bar{h} + \frac{\lambda_{k,t}\sigma_t^2}{\mu_t}\nabla_{\bar{h}}\log p_t^h(\bar{h})\right] + \kappa_{k,t}(I_{d_x} - UU^\top)x \\
&= U\left[\alpha_{k,t}\bar{h} + \beta_{k,t}\nabla_{\bar{h}}\log p_t^h(\bar{h})\right] + \kappa_{k,t}(I_{d_x} - UU^\top)x,
\end{aligned}$$

where $\alpha_{k,t} := \kappa_{k,t} + \lambda_{k,t}/\mu_t$ and $\beta_{k,t} := \lambda_{k,t}\sigma_t^2/\mu_t$. The first line follows from direct substitution of (27), the second line is simply collecting common factors, the third line is from the definition of $x_\perp$, and the fourth line is from the definition of $\alpha_{k,t}$ and $\beta_{k,t}$.

This completes the proof. □

C. Supplementary Background: Transformer Block

Throughout this paper, we assume a standard transformer architecture that consists primarily of simple transformer blocks. In this section, we introduce the transformer block and define its constituent pieces. Following common architectures of diffusion transformers (Peebles & Xie, 2023; Hu et al., 2024; 2025b), we also adopt the reshape layer to complete our definition of the flow matching transformer.

Our notation follows (Hu et al., 2025b).

Reshape Layer. We consider square image inputs of dimension $d_{\text{input}} = i \times i$. Following common diffusion transformer architectures (Peebles & Xie, 2023; Hu et al., 2024; 2025b), we divide image inputs into square patches to convert them to matrix input format. The inputs $x \in \mathbb{R}^{d_{\text{input}}}$ are serialized according to patch size p through the reshape layer $R : \mathbb{R}^{d_{\text{input}}} \to \mathbb{R}^{d \times L}$, where $d = p^2$ and $L = (i/p)^2$.

Definition C.1 (Flow Matching Reshape Layer)**.** *Let $R(\cdot) : \mathbb{R}^{d_{\text{input}}} \to \mathbb{R}^{d \times L}$ be a reshape layer that transforms the d_{input}-dimensional input into a $d \times L$ matrix. Given any $d_{\text{input}} = i \times i$ image input, $R(\cdot)$ converts it into a sequence representation with feature dimension $d := p^2$ (where $p \geq 2$) and sequence length $L := (i/p)^2$. Symmetrically, let the reverse reshape (flattening) layer $R^{-1}(\cdot) : \mathbb{R}^{d \times L} \to \mathbb{R}^{d_{\text{input}}}$ be the inverse of $R(\cdot)$.*

Feed-Forward Layer. Let r be the width of the hidden layer in the MLP. For input matrix $A \in \mathbb{R}^{d \times L}$, weight matrices $W_1 \in \mathbb{R}^{r \times d}$ and $W_2 \in \mathbb{R}^{d \times r}$, and bias vectors $b_1 \in \mathbb{R}^r$ and $b_2 \in \mathbb{R}^d$, we define the feed-forward layer $\mathcal{F}^{(FF)} : \mathbb{R}^{d \times L} \mapsto \mathbb{R}^{d \times L}$:

$$\mathcal{F}^{(\text{FF})}(A) := A + W_2 \,\text{ReLU}\left(W_1 A + b_1 \mathbf{1}_L^\top\right) + b_2 \mathbf{1}_L^\top. \tag{35}$$

Self-Attention Layer. Let h and s denote the number of heads and the hidden dimension of the self-attention layer, respectively. For input matrix $A \in \mathbb{R}^{d \times L}$, query matrices $\left\{W_Q^i \in \mathbb{R}^{s \times d}\right\}_{i \in [h]}$, key matrices $\left\{W_K^i \in \mathbb{R}^{s \times d}\right\}_{i \in [h]}$, value matrices $\left\{W_V^i \in \mathbb{R}^{s \times d}\right\}_{i \in [h]}$, and output matrices $\left\{W_O^i \in \mathbb{R}^{s \times d}\right\}_{i \in [h]}$, we define the self-attention block $\mathcal{F}^{(SA)} : \mathbb{R}^{d \times L} \mapsto \mathbb{R}^{d \times L}$:

$$\mathcal{F}^{(\text{SA})}(A) := A + \sum_{i=1}^{h} W_O^i \cdot \left(W_V^i A\right) \text{Softmax}\left[\left(W_K^i A\right)^\top \left(W_Q^i A\right)\right]. \tag{36}$$

Here Softmax denotes the column-wise softmax function.

Definition C.2 (Transformer Block)**.** *We construct a single transformer as a composition of a h-headed, s-hidden dimension, self-attention layer with a r-dimensional MLP feed-forward layer applied to input matrix $Z \in \mathbb{R}^{d \times L}$ with positional encoding matrix $E \in \mathbb{R}^{d \times L}$:*

$$\mathcal{F}^{h,s,r}(Z) := \mathcal{F}^{(\text{FF})}\left(\mathcal{F}^{(\text{SA})}(Z + E)\right) : \mathbb{R}^{d \times L} \mapsto \mathbb{R}^{d \times L}.$$

We define the function class of transformer networks as the set of all compositions of such transformer blocks.

Definition C.3 (Transformer Network Function Class)**.** *Let $\mathcal{T}^{h,s,r}$ denote the transformer network function class where each function $f \in \mathcal{T}^{h,s,r}$ is a composition of transformer blocks $\mathcal{F}^{h,s,r}$, i.e.,*

$$\mathcal{T}^{h,s,r} := \left\{f : \mathbb{R}^{d \times L} \mapsto \mathbb{R}^{d \times L} \mid f = \mathcal{F}^{h,s,r} \circ \cdots \circ \mathcal{F}^{h,s,r}\right\}.$$

Finally, we define the function class of transformer networks with reshape and flattening layers. We parameterize this function class with the following norm bounds to enable characterization of the network class complexity. For ease of notation, let $W_{KQ} := (W_K)^\top W_Q$ and $W_{OV} := W_O W_V$.

Definition C.4 (Transformer Network Function Class with Reshape Layer $\mathcal{T}_R^{h,s,r}$)**.** *The transformer network class with reshape layer $\mathcal{T}_R^{h,s,r}\left(C_{\mathcal{T}}, C_{KQ}^{2,\infty}, C_{KQ}, C_{OV}^{2,\infty}, C_{OV}, C_E, C_F^{2,\infty}, C_F, L_{\mathcal{T}}\right)$ satisfies:*

- $\mathcal{T}_R^{h,s,r} := \left\{R^{-1} \circ f_{\mathcal{T}} \circ R : \mathbb{R}^{d_{\text{input}}} \mapsto \mathbb{R}^{d_{\text{input}}} \mid f_{\mathcal{T}} \in \mathcal{T}^{h,s,r}\right\}$;
- *Model output bound:* $\sup_Z \|f_{\mathcal{T}}(Z)\|_2 \leq C_{\mathcal{T}}$;

- *Parameter bound in* $f^{(\mathrm{SA})}$: $\|W_{KQ}\|_{2,\infty} \leq C_{KQ}^{2,\infty}, \|W_{KQ}\|_2 \leq C_{KQ}, \|W_{OV}\|_{2,\infty} \leq C_{OV}^{2,\infty}, \|W_{OV}\|_2 \leq C_{OV}, \left\|E^\top\right\|_{2,\infty} \leq C_E$;
- *Parameter bound in* $f^{(\mathrm{FF})}$: $\max\left\{\|W_1\|_{2,\infty}, \|W_2\|_{2,\infty}\right\} \leq C_F^{2,\infty}, \max\left\{\|W_1\|_2, \|W_2\|_2\right\} \leq C_F$;
- *Lipschitz of* $f_{\mathcal{T}} \in \mathcal{T}^{h,s,r}$: $\|f_{\mathcal{T}}(Z_1) - f_{\mathcal{T}}(Z_2)\|_F \leq L_{\mathcal{T}} \|Z_1 - Z_2\|_F$, *for any* $Z_1, Z_2 \in \mathbb{R}^{d\times L}$.

D. Supplementary Background: Universal Approximation of Transformers

Our analysis of flow matching transformers' ability to achieve arbitrary approximation error on a bounded domain relies on the universal approximation theory of transformers developed in (Hu et al., 2025a; 2024; Kajitsuka & Sato, 2024; Yun et al., 2019).

Of particular importance is the result that, given arbitrary $\epsilon > 0$, the transformer network function class (Section C) can achieve approximation error bounded above by ϵ, provided that the network class satisfies certain parameter norm bounds.

Universal Approximation of Transformers. We now recall a universal approximation result — due to (Hu et al., 2025a; Kajitsuka & Sato, 2024) — showing that a transformer with just a single self-attention layer between two feed-forward layers can approximate any continuous target to arbitrary accuracy on a compact support. In our notation, the realizable functions have the form

$$g(Z) = \mathcal{F}_2^{(\mathrm{FF})} \circ \mathcal{F}^{(\mathrm{SA})} \circ \mathcal{F}_1^{(\mathrm{FF})}(Z) \in \mathcal{T}_R^{h,s,r},$$

where $\mathcal{F}^{(\mathrm{SA})}$ is a single-head softmax self-attention layer and $\mathcal{F}_1^{(\mathrm{FF})}, \mathcal{F}_2^{(\mathrm{FF})}$ are feed-forward layers (see Definition C.3). The approximation error is measured in the L^p-type distance d_p induced by the element-wise ℓ_p norm.

Lemma D.1 (Transformer Universal Approximation, Theorem B.1 of (Hu et al., 2025a) and Proposition 1 of (Kajitsuka & Sato, 2024))**.** *Let* $\epsilon \in (0,1)$ *and* $p \in [1,\infty)$. *Let* $\mathcal{F}_1^{(\mathrm{FF})}, \mathcal{F}_2^{(\mathrm{FF})}$ *be feed-forward layers and let* $\mathcal{F}^{(\mathrm{SA})}$ *be a single-head self-attention layer with softmax (as in Definition C.3). Then, for any continuous function* f *on a compact support and any* ϵ, *there exists*

$$g(Z) = \mathcal{F}_2^{(\mathrm{FF})} \circ \mathcal{F}^{(\mathrm{SA})} \circ \mathcal{F}_1^{(\mathrm{FF})}(Z) \in \mathcal{T}_R^{h,s,r}$$

such that

$$d_p(f(Z), g(Z)) < \epsilon, \qquad \text{where} \quad d_p := \left(\int \|f(Z) - g(Z)\|_p^p \mathrm{d}Z\right)^{1/p},$$

and $\|\cdot\|_p$ *denotes the element-wise* ℓ_p*-norm.*

Remark D.1. *The construction achieving Lemma D.1 uses exactly two feed-forward layers and a single-head, single-layer attention module, i.e., it produces* $g \in \mathcal{T}_R^{1,1,r}$ *with* $r = O(ID)$. *By Definition C.3, the class* $\mathcal{T}_R^{1,1,r}$ *lies within our transformer network class, so the result is obtained within our setup.*

Parameter Norm Bounds for Transformer Approximation. To quantify the universal approximation guarantee (beyond a weaker existential guarantee), we need explicit controls on parameter sizes. The following lemma provides big-O bounds on the spectral and row norms (recall Definition C.3) which give sufficient conditions to realize and ϵ-approximation.

Lemma D.2 (Transformer Matrices Bounds, Modified from Lemma F. 4 and Lemma F. 5 of (Hu et al., 2025b))**.** *Fix* $\epsilon \in (0,1)$ *and* $L \geq 2$. *Denote the input sequence over a compact set by* $Z \in [-I, I]^{d\times L}$ *with* $I > 0$ *an absolute constant. Let* $f : [-I, I]^{d\times L} \to \mathbb{R}^{d\times L}$ *be an function that is Lipschitz continuous (with respect to the matrix 2-norm). Then there exists a choice of parameters within the transformer class* $\mathcal{T}_R^{h,s,r}$ *producing a network* $g \in \mathcal{T}_R^{h,s,r}$ *such that* $d_p(f,g) < \epsilon$ *and, simultaneously, the norm bounds that define the class can be taken as*

$$\text{(attention: keys/queries) } C_{KQ}, C_{KQ}^{2,\infty} = O\left(I^{4d+2}\epsilon^{-4d-2}\right), \tag{37}$$

$$\text{(attention: values/outputs) } C_{OV}, C_{OV}^{2,\infty} = O(\epsilon), \tag{38}$$

$$\text{(feed-forward layers) } C_F, C_F^{2,\infty} = O\left(I\epsilon^{-1} \cdot \max \|f(Z)\|_F\right), \tag{39}$$

$$\text{(positional encodings) } C_E = O(1), \tag{40}$$

where $O(\cdot)$ *hides polynomial and logarithmic factors depending only on* d *and* L.

E. Proof of Theorem 4.1

In this section, we use transformers to approximate velocity under Assumption 2.1 and generic Hölder data assumptions and present an upper bound of the velocity approximation error.

The proof has three ingredients. First, the velocity decomposition reduces the approximation error to a score-approximation error on $\mathbb{R}^{d_0}$ alone. Second, the latent density p_t^h inherits the Hölder regularity and the sub-Gaussian tail of p_1^h, allowing diffused-Taylor approximation on a compact box. Third, we approximate the resulting local polynomial by a transformer on $\mathbb{R}^{d_0}$ via Lemma D.1.

For velocity approximation, we consider the parameterized velocity field

$$u_\theta(x,t) := U[\alpha_t U^\top x + \beta_t \widehat{s}(U^\top x, t)] + \kappa_t(I_{d_x} - UU^\top)x, \tag{41}$$

where $\widehat{s} \in \mathcal{T}_R^{h,s,r}$ is a Transformer network. We start with reducing the velocity approximation theory to latent score approximation.

Lemma E.1. *For every $\widehat{s} : \mathbb{R}^{d_0} \times [t_0, T] \to \mathbb{R}^{d_0}$ and every (x,t),*

$$\|u_\theta(x,t) - u_t^\star(x)\|_2^2 = \beta_t^2 \|\widehat{s}(U^\top x, t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})\|_2^2, \tag{42}$$

where u_θ is defined in (41) and $\bar{h} = U^\top x$. Consequently,

$$\int_{\mathbb{R}^{d_x}} \|u_\theta(x,t) - u_t^\star(x)\|_2^2 p_t(x)\mathrm{d}x = \beta_t^2 \int_{\mathbb{R}^{d_0}} \|\widehat{s}(\bar{h},t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})\|_2^2 p_t^h(\bar{h})\mathrm{d}\bar{h}. \tag{43}$$

Proof. By Theorem 3.1, the orthogonal-contraction term of $u_t^\star$ equals the orthogonal-contraction term of u_θ, since both equal $\kappa_t(I_{d_x} - UU^\top)x$. The latent-transport coefficients agree because both equal $U\alpha_t\bar{h}$. Therefore we have

$$u_\theta(x,t) - u_t^\star(x) = U\beta_t[\widehat{s}(\bar{h},t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})].$$

Since U has orthonormal columns, $\|Uv\|_2 = \|v\|_2$ for every $v \in \mathbb{R}^{d_0}$ holds. This proves (42). Integrating on p_t and using the fact that $\bar{h}_t = U^\top X_t$ has marginal density p_t^h yields (43). □

We obtain the latent density p_t^h at time $t \in (0,1)$ by convolving p_1^h with a Gaussian of variance σ_t^2 after scaling by μ_t. The next lemma gathers the standard pointwise bounds.

Lemma E.2 (Density and score bounds, Lemma A.9 and Lemma A.10 of (Fu et al., 2024))**.** *Under Assumption 2.1 and Assumption 4.2, there exist constants $C_4, C_4' > 0$ depending only on d_0, B, C_1, C_2 such that for every $\bar{h} \in \mathbb{R}^{d_0}$ and every $t \in (t_0, T)$,*

$$\frac{C_4}{\sigma_t^{d_0}} \exp\big(-(\|\bar{h}\|_2^2 + 1)/\sigma_t^2\big) \le p_t^h(\bar{h}) \le C_1(\mu_t^2 + C_2\sigma_t^2)^{-d_0/2} \exp\big(-C_2\|\bar{h}\|_2^2/(2(\mu_t^2 + C_2\sigma_t^2))\big)$$

and

$$\|\nabla_{\bar{h}} \log p_t^h(\bar{h})\|_\infty \le \frac{C_4'(\|\bar{h}\|_2 + 1)}{\sigma_t^2}.$$

Proof. See proof of (Fu et al., 2024, Lemma A.9-Lemma A.10) . □

Next, we present proof of score approximation on latent space.

Lemma E.3 (Latent score approximation)**.** *Under Assumption 2.1, Assumption 4.1, and Assumption 4.2, for any $\beta > 0$ and sufficiently large N, there exists a transformer $\widehat{s} \in \mathcal{T}_R^{h,s,r}$ on $\mathbb{R}^{d_0} \times [t_0, T]$ such that*

$$\sup_{t \in [t_0,T]} \int_{\mathbb{R}^{d_0}} \|\widehat{s}(\bar{h},t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})\|_2^2 p_t^h(\bar{h})\mathrm{d}\bar{h} = O(B^2 N^{-\beta}(\log N)^{d_0+\beta/2+1}), \tag{44}$$

with parameter bounds

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

Proof. Let $k_1 = \lfloor \beta \rfloor$ and fix the truncation cube

$$\mathcal{B}_N := [-C_h\sqrt{\log N}, C_h\sqrt{\log N}]^{d_0}$$

with C_h chosen below. Further, by clipping of time interval, we define the uniform-in-t constants

$$\underline{\sigma} := \inf_{t\in[t_0,T]} \sigma_t > 0, \quad \overline{\Sigma} := \sup_{t\in[t_0,T]} (\mu_t^2 + C_2\sigma_t^2) < \infty, \quad \overline{\beta_t^2} := \sup_{t\in[t_0,T]} \beta_t^2 < \infty.$$

Step 1: Local polynomial approximator on $\mathcal{B}_N$. Following the diffused-Taylor construction in Lemma A.4 and Lemma A.6 of (Fu et al., 2024), there exist polynomials $\widehat{\Psi}_1 : \mathcal{B}_N \times (t_0, T) \to \mathbb{R}$ and $\widehat{\Psi}_2 : \mathcal{B}_N \times (t_0, T) \to \mathbb{R}^{d_0}$ of degree at most k_1 in $\bar{h}$ such that for every $t \in (t_0, T)$ and every $\bar{h} \in \mathcal{B}_N$,

$$|\widehat{\Psi}_1(\bar{h}, t) - p_t^h(\bar{h})| \le c_1 B N^{-\beta} (\log N)^{(d_0+k_1)/2}, \tag{45}$$

$$\|\widehat{\Psi}_2(\bar{h}, t) - \nabla_{\bar{h}} p_t^h(\bar{h})\|_\infty \le c_1 B N^{-\beta} (\log N)^{(d_0+k_1+1)/2}, \tag{46}$$

for a constant c_1 depending only on d_0, β, B, C_1, C_2.

Define the truncated denominator $\widehat{\Psi}_1^c := \max\{\widehat{\Psi}_1, \epsilon_{\text{low}}\}$ with $\epsilon_{\text{low}} := 2c_1 B N^{-\beta} (\log N)^{(d_0+k_1)/2}$, and the unclipped polynomial ratio

$$\widehat{v}_0(\bar{h}, t) := \widehat{\Psi}_1^c(\bar{h}, t)^{-1} \widehat{\Psi}_2(\bar{h}, t), \quad (\bar{h}, t) \in \mathcal{B}_N \times [t_0, T].$$

We set the clip radius

$$M_N := C_4'(C_h\sqrt{\log N} + 1)/\underline{\sigma}^2 = O(\sqrt{\log N}), \tag{47}$$

and define the clipped polynomial surrogate score

$$\widehat{v}(\bar{h}, t) := \text{clip}_{M_N}(\widehat{v}_0(\bar{h}, t)), \quad (\bar{h}, t) \in \mathcal{B}_N \times [t_0, T], \tag{48}$$

where the clip acts componentwise via $\text{clip}_M(z)_j := \text{ReLU}(z_j + M) - \text{ReLU}(z_j - M) - M$ for $j = 1, \ldots, d_0$. By Lemma E.2 together with $\sigma_t \ge \underline{\sigma}$ and $\|\bar{h}\|_2 \le C_h\sqrt{\log N}$ on $\mathcal{B}_N$,

$$\|\nabla_{\bar{h}} \log p_t^h(\bar{h})\|_\infty \le M_N \quad \text{for every } (\bar{h}, t) \in \mathcal{B}_N \times [t_0, T], \tag{49}$$

so each coordinate of the true score lies in $[-M_N, M_N]$ on $\mathcal{B}_N$. Further, since component-wise projection onto $[-M_N, M_N]$ is non-expansive with respect to any point already inside that interval, it holds

$$\|\widehat{v}(\bar{h}, t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})\|_2 \le \|\widehat{v}_0(\bar{h}, t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})\|_2 \tag{50}$$

for every $(\bar{h}, t) \in \mathcal{B}_N \times [t_0, T]$.

Step 2: Bound on $\mathcal{B}_N \cap \{p_t^h \ge \epsilon_{\text{low}}\}$. On the set where $p_t^h(\bar{h}) \ge \epsilon_{\text{low}}$, both p_t^h and $\widehat{\Psi}_1^c$ are at least ϵ_{low}. Writing

$$\widehat{v}_0 - \nabla \log p_t^h = \frac{\widehat{\Psi}_2 - \nabla p_t^h}{\widehat{\Psi}_1^c} + \nabla p_t^h \frac{p_t^h - \widehat{\Psi}_1^c}{\widehat{\Psi}_1^c p_t^h},$$

and combining (45), (46) with the score bound $\|\nabla p_t^h\|_\infty \le p_t^h \cdot C_4'(\|\bar{h}\|_2 + 1)/\sigma_t^2$ from Lemma E.2 gives

$$p_t^h(\bar{h}) \|\widehat{v}_0(\bar{h}, t) - \nabla \log p_t^h(\bar{h})\|_2 \le c_2 B N^{-\beta} (\log N)^{(d_0+k_1+1)/2}, \qquad \bar{h} \in \mathcal{B}_N \cap \{p_t^h \ge \epsilon_{\text{low}}\}, \tag{51}$$

for some constant c_2.

Squaring (51) and dividing by $p_t^h \ge \epsilon_{\text{low}}$ derives

$$\|\widehat{v}_0 - \nabla \log p_t^h\|_2^2 p_t^h \le c_2^2 B^2 N^{-2\beta} (\log N)^{d_0+k_1+1} (p_t^h)^{-1}.$$

Integrating over $\mathcal{B}_N \cap \{p_t^h \ge \epsilon_{\text{low}}\}$, using (50),$(p_t^h)^{-1} \le \epsilon_{\text{low}}^{-1}$ and $\text{vol}(\mathcal{B}_N) = (2C_h)^{d_0} (\log N)^{d_0/2}$, it holds

$$\int_{\mathcal{B}_N \cap \{p_t^h \ge \epsilon_{\text{low}}\}} \|\widehat{v} - \nabla \log p_t^h\|_2^2 p_t^h \mathrm{d}\bar{h} \lesssim B N^{-\beta} (\log N)^{d_0 + (k_1+2)/2} \tag{52}$$

$$\le B N^{-\beta} (\log N)^{d_0 + \beta/2 + 1}, \tag{53}$$

where the last line uses $k_1 \le \beta$.

Step 3: Bound on $\mathcal{B}_N \cap \{p_t^h < \epsilon_{\text{low}}\}$. By the clip in Step 1, $\|\widehat{v}(\bar{h},t)\|_\infty \le M_N$ for every $(\bar{h},t) \in \mathcal{B}_N \times [t_0,T]$. Combined with $\|\nabla \log p_t^h(\bar{h})\|_\infty \le M_N$ on $\mathcal{B}_N$ from (49),

$$\|\widehat{v}(\bar{h},t) - \nabla \log p_t^h(\bar{h})\|_2^2 \le d_0(2M_N)^2 = c_3 \log N, \tag{54}$$

with c_3 uniform in t. Multiplying (54) by $p_t^h \le \epsilon_{\text{low}}$ and integrating over $\mathcal{B}_N \cap \{p_t^h < \epsilon_{\text{low}}\}$ using $\text{vol}(\mathcal{B}_N) = O((\log N)^{d_0/2})$ gives

$$\int_{\mathcal{B}_N \cap \{p_t^h < \epsilon_{\text{low}}\}} \|\widehat{v} - \nabla \log p_t^h\|_2^2 p_t^h \mathrm{d}\bar{h} \le c_3 \log N \cdot \epsilon_{\text{low}} \cdot O((\log N)^{d_0/2}).$$

Substituting $\epsilon_{\text{low}} = 2c_1 B N^{-\beta} (\log N)^{(d_0+k_1)/2}$ and using $k_1 \le \beta$,

$$\int_{\mathcal{B}_N \cap \{p_t^h < \epsilon_{\text{low}}\}} \|\widehat{v} - \nabla \log p_t^h\|_2^2 p_t^h \mathrm{d}\bar{h} = O(BN^{-\beta}(\log N)^{d_0+\beta/2+1}). \tag{55}$$

Step 4: Tail outside $\mathcal{B}_N$. Choose C_h such that

$$\frac{C_2 C_h^2}{2\bar{\Sigma}} \ge 2\beta + d_0 \quad \text{for all } t \in [t_0,T].$$

By Lemma E.2, $p_t^h(\bar{h}) \le c_4 \exp\big(-C_2\|\bar{h}\|_2^2/(2(\mu_t^2 + C_2\sigma_t^2))\big)$ for some c_4, hence

$$\int_{\mathbb{R}^{d_0} \setminus \mathcal{B}_N} \|\widehat{v} - \nabla \log p_t^h\|_2^2 p_t^h \mathrm{d}\bar{h} = O(B^2 N^{-2\beta} (\log N)^{d_0/2}), \tag{56}$$

where $\widehat{v}$ is extended by zero outside $\mathcal{B}_N$ and the integrand is controlled by $\|\nabla \log p_t^h\|_2^2 p_t^h \le c(1+\|\bar{h}\|_2^2) p_t^h/\sigma_t^4$, which has integrable Gaussian tail; the rate $N^{-2\beta}$ follows from the choice of C_h.

Step 5: Transformer realization. The surrogate $\widehat{v}$ is continuous on the compact set $\mathcal{B}_N \times [t_0,T] \subset \mathbb{R}^{d_0} \times [t_0,T]$ and bounded by $O(\sqrt{\log N})$ in ℓ_2-norm. By Lemma D.1 applied with target accuracy $\epsilon_{\text{tf}} = N^{-\beta}$ and the parameter scaling of Lemma D.2 specialized to input dimension d_0 and bounded input range $\sqrt{\log N}$, there exists $\widehat{s} \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{t_0}^{T} \int_{\mathcal{B}_N} \|\widehat{s}(\bar{h},t) - \widehat{v}(\bar{h},t)\|_2^2 \mathrm{d}\bar{h}\mathrm{d}t \le \epsilon_{\text{tf}}^2 = N^{-2\beta}.$$

Since p_t^h is uniformly bounded, the transformer approximation error contributes $O(N^{-2\beta})$ to the score risk. The parameter bounds follow Lemma D.1.

Step 6: Combining the four regions. Integrating (52), (55), (56) on $t \in [t_0,T]$ and summing the error in Step 2-5, the dominant rate is $B^2 N^{-\beta}(\log N)^{d_0+\beta/2+1}$ uniformly in $t \in [t_0,T]$. This completes the proof. □

Finally, we present proof of Theorem 4.1.

Theorem E.1 (Theorem 4.1 Restated)**.** *Assume Assumption 2.1, Assumption 4.1 and Assumption 4.2. For any $\beta > 0$ and sufficiently large $N \in \mathbb{N}$, there exists a transformer $\widehat{s} \in \mathcal{T}_R^{h,s,r}$ on $\mathbb{R}^{d_0} \times [t_0,T]$ such that the structured velocity field*

$$u_\theta(x,t) := U[\alpha_t U^\top x + \beta_t \widehat{s}(U^\top x,t)] + \kappa_t(I_{d_x} - UU^\top)x \tag{57}$$

satisfies the integrated approximation error bound

$$\int_{t_0}^{T} \int_{\mathbb{R}^{d_x}} \|u_t^\star(x) - u_\theta(x,t)\|_2^2 p_t(x) \mathrm{d}x\mathrm{d}t = O(B^2 N^{-\beta}(\log N)^{d_0+\beta/2+1}). \tag{58}$$

Writing the input shape as $d_0 + 1 = d \times L$ (patch size times sequence length), the transformer parameter bounds satisfy

$$C_{KQ}, C_{KQ}^{2,\infty} = O((\log N)^{2d+1} N^{(4d+2)\beta}), \quad C_{OV}, C_{OV}^{2,\infty} = O(N^{-\beta}),$$
$$C_F, C_F^{2,\infty} = O((\log N) N^\beta), \quad C_E = O(1), \quad C_T = O(\sqrt{\log N}).$$

The constants in $O(\cdot)$ depend on d_0, β, B, C_1, C_2.

Proof. By Lemma E.1, for the structured velocity u_θ of (41) and the latent score transformer $\widehat{s}$ from Lemma E.3,

$$\int_{t_0}^{T}\int_{\mathbb{R}^{d_x}} \|u_t^\star(x) - u_\theta(x,t)\|_2^2 p_t(x)\mathrm{d}x\mathrm{d}t = \int_{t_0}^{T} \beta_t^2 \int_{\mathbb{R}^{d_0}} \|\widehat{s}(\bar{h},t) - \nabla_{\bar{h}} \log p_t^h(\bar{h})\|_2^2 p_t^h(\bar{h})\mathrm{d}\bar{h}\mathrm{d}t.$$

By Assumption 4.2, $\mu_t, \sigma_t \in (0,1)$ are bounded away from 0 on $[t_0, T]$, so $\sup_{t\in[t_0,T]} \beta_t^2 = O(1)$. Applying Lemma E.3 yields the approximation result. Parameter bounds follow Lemma D.1. □

F. Proof of Theorem 4.2

In this section, we provide proof of Theorem 4.2. The proof of Theorem 4.2 follows the standard approximation–generalization decomposition, adapted to the data-driven structured estimator (41). Concretely, Section F.1 establishes subspace recovery theory, Section F.2 develops the generalization machinery and Section F.3 presents the main proof of Theorem 4.2.

Throughout this section, we let $\{x_i\}_{i=1}^n$ be i.i.d. training samples with $x_i = Uh_i$ and $h_i \sim p_1^h$. We define the L2-loss function as

$$\ell\left(x_i; u_\theta\right) := \frac{1}{T-t_0}\int_{t_0}^{T} \mathop{\mathbb{E}}_{X_0\sim\mathcal{N}(0,I_{d_x})}\left[\left\|u_\theta\left(\mu_t x_i + \sigma_t X_0, t\right) - \left(\dot{\mu}_t x_i + \dot{\sigma}_t X_0\right)\right\|_2^2\right]\mathrm{d}t, \tag{59}$$

where x_i is a sample point drawn from terminal distribution. We let the empirical loss be

$$\widehat{\mathcal{L}}_{\mathrm{CFM}}(u_\theta) := \frac{1}{n(T-t_0)}\sum_{i=1}^{n}\int_{t_0}^{T} \mathop{\mathbb{E}}_{X_0\sim\mathcal{N}(0,I_{d_x})}[\|u_\theta(\mu_t x_i + \sigma_t X_0, t) - \dot{\mu}_t x_i - \dot{\sigma}_t X_0\|_2^2]\mathrm{d}t,$$

and $\widehat{u}_\theta$ is its minimizer over the structured class. We use u^* to stand for true velocity and define the empirical risk as

$$\widehat{\mathcal{R}}(u_\theta) := \frac{1}{n}\sum_{i=1}^{n}(\ell(x_i; u_\theta) - \ell(x_i, u^*)).$$

F.1. Subspace Recovering

Define $\widehat{\Sigma}_x := \sum_{i=1}^n x_i x_i^\top/n \in \mathbb{R}^{d_x\times d_x}$, and let $\widehat{U} \in \mathbb{R}^{d_x\times d_0}$ be a matrix whose columns are orthonormal eigenvectors of $\widehat{\Sigma}_x$ corresponding to the top d_0 eigenvalues. The data-driven structured estimator is

$$\widehat{u}_\theta(x,t) := \widehat{U}[\alpha_t\widehat{U}^\top x + \beta_t\widehat{s}_\theta(\widehat{U}^\top x, t)] + \kappa_t(I_{d_x} - \widehat{U}\widehat{U}^\top)x, \tag{60}$$

with $\widehat{s}_\theta$ a transformer on $\mathbb{R}^{d_0}\times[t_0,T]$ in the class of Lemma D.1. We introduce the following lemma to guarantee that we're capable of recovering the subspace almost surely.

Lemma F.1. *Under Assumption 2.1 and Assumption 4.1, whenever $n \geq d_0$,*

$$\widehat{U}\widehat{U}^\top = UU^\top \quad \textit{almost surely},$$

and there exists a random orthogonal matrix $R \in \mathbb{R}^{d_0\times d_0}$ such that $\widehat{U} = UR$ almost surely.

Proof. By Assumption 2.1, $x_i = Uh_i$, hence

$$\widehat{\Sigma}_x = \sum_{i=1}^{n} Uh_i(Uh_i)^\top/n = U\widehat{\Sigma}_h U^\top, \quad \text{where } \widehat{\Sigma}_h := n^{-1}\sum_{i=1}^{n} h_i h_i^\top.$$

Under Assumption 4.1, p_1^h is absolutely continuous on $\mathbb{R}^{d_0}$. Hence $h_1,\ldots,h_n$ are in general position almost surely; in particular, any d_0 of them are linearly independent. For $n \geq d_0$ this gives $\mathrm{rank}(\widehat{\Sigma}_h) = d_0$ almost surely. Therefore $\widehat{\Sigma}_x = U\widehat{\Sigma}_h U^\top$ has rank exactly d_0 almost surely, with column space equal to $\mathrm{span}(U)$. The matrix $\widehat{U}$ of its top d_0 orthonormal eigenvectors thus spans $\mathrm{span}(U)$ exactly, so $\widehat{U} = UR$ for some orthogonal $R \in \mathbb{R}^{d_0\times d_0}$, and $\widehat{U}\widehat{U}^\top = URR^\top U^\top = UU^\top$. □

Throughout the rest of this section, let $\mathcal{E} := \{\widehat{U}\widehat{U}^\top = UU^\top\}$. By Lemma F.1, $\mathbb{P}(\mathcal{E}) = 1$. On $\mathcal{E}$, we define the rotated latent variable $\widetilde{h} := \widehat{U}^\top x = R^\top U^\top x = R^\top \bar{h}$, with marginal density $\widetilde{p}_t^h(\widetilde{h}) := p_t^h(R\widetilde{h})$. Since R is orthogonal and Hölder

norms and sub-Gaussian tails are rotation-invariant, $\widetilde{p}_1^h \in \mathcal{H}^\beta(\mathbb{R}^{d_0}, B)$ and $\widetilde{p}_1^h(\widetilde{h}) \le C_1 \exp\Big(-C_2\|\widetilde{h}\|_2^2/2\Big)$ with the same constants. Hence Lemma E.2 and Lemma E.3 apply verbatim with p_t^h replaced by $\widetilde{p}_t^h$.

Next, we reduce velocity approximation to score estimation on $\mathcal{E}$.

Lemma F.2. *On $\mathcal{E}$, for every (x,t),*

$$\|\widehat{u}_\theta(x,t) - u_t^\star(x)\|_2^2 = \beta_t^2 \|\widehat{s}_\theta(\widehat{U}^\top x, t) - \widetilde{s}_t(\widehat{U}^\top x)\|_2^2,$$

where $\widetilde{s}_t(\widetilde{h}) := \nabla_{\widetilde{h}} \log \widetilde{p}_t^h(\widetilde{h})$. Consequently

$$\mathcal{R}(\widehat{u}_\theta) = \frac{1}{T-t_0}\int_{t_0}^T \beta_t^2 \mathop{\mathbb{E}}_{\widetilde{h}\sim\widetilde{p}_t^h}[\|\widehat{s}_\theta(\widetilde{h},t) - \widetilde{s}_t(\widetilde{h})\|_2^2]dt.$$

Proof. On $\mathcal{E}$, $I_{d_x} - \widehat{U}\widehat{U}^\top = I_{d_x} - UU^\top$, so the orthogonal-contraction terms of $\widehat{u}_\theta$ and $u_t^\star$ are identical. With $R = U^\top \widehat{U}$ orthogonal and $\widehat{U} = UR$,

$$\widehat{U}\alpha_t \widehat{U}^\top x = URR\alpha_t R^\top U^\top x = U\alpha_t \bar{h},$$

matching the latent-transport term of $u_t^\star$. By the chain rule, $\widetilde{s}_t(\widetilde{h}) = R^\top \nabla_{\bar{h}} \log p_t^h(\bar{h})$, so we have

$$U\beta_t \nabla_{\bar{h}} \log p_t^h(\bar{h}) = URR\beta_t R^\top \nabla_{\bar{h}} \log p_t^h(\bar{h}) = \widehat{U}\beta_t \widetilde{s}_t(\widetilde{h}).$$

Therefore it holds

$$\widehat{u}_\theta(x,t) - u_t^\star(x) = \widehat{U}\beta_t[\widehat{s}_\theta(\widetilde{h},t) - \widetilde{s}_t(\widetilde{h})],$$

and $\|\widehat{U}v\|_2 = \|v\|_2$ since $\widehat{U}$ has orthonormal columns. The second claim follows by integrating the pointwise identity with $p_t(x)\mathrm{d}x \cdot \frac{1}{T-t_0}dt$ and noting that $\widehat{U}^\top X_t = R^\top \bar{h}_t$ has marginal density $\widetilde{p}_t^h$. □

F.2. Generalization Bound

We begin with the definition of covering number.

Definition F.1 (Covering Number)**.** *For a data distribution P, let $\{x_i\}_{i=1}^n$ be data points sampled from P. Denote $P^n := P \otimes \cdots \otimes P$ as the joint distribution such that $\{x_i\}_{i=1}^n \sim P^n$. Given a function class $\mathcal{F}$ of functions $f : \mathbb{R}^{d_x} \to \mathbb{R}^k$ and $\epsilon_c > 0$, the ϵ_c-covering number $\mathcal{N}(\epsilon_c, \mathcal{F}, \{x_i\}_{i=1}^n, \|\cdot\|)$ with norm $\|\cdot\|$ is the smallest size of a collection $\{f_j\}_{j=1}^N \subset \mathcal{F}$ such that for any $f \in \mathcal{F}$, there exists a $j \in [N]$ satisfying*

$$\max_{i\in[n]} \|f(x_i) - f_j(x_i)\| \le \epsilon_c.$$

Further, we define the covering number with respect to sample size n as

$$\mathcal{N}(\epsilon_c, \mathcal{F}, P^n, \|\cdot\|) := \sup_{\{x_i\}_{i=1}^n \sim P^n} \mathcal{N}(\epsilon_c, \mathcal{F}, \{x_i\}_{i=1}^n, \|\cdot\|).$$

We have the following lemma on upper bound of the covering number for the class of transformer networks.

Lemma F.3 (Covering Number Bounds for $\mathcal{S}(\mathcal{D})$ under Assumption 2.1, Modified from Lemma E.2 of (Su et al., 2025))**.** *Let $\epsilon_c > 0$. We define the loss function class by $\mathcal{S}(\mathcal{D}) := \{\ell(Uh; u_\theta) : h \in \mathcal{D} \to \mathbb{R}\}$, where $\mathcal{D} := [-D, D]^{d_0}$. Further, define the norm of loss functions by $\|\ell(\cdot; u_\theta)\|_{\infty\mathcal{D}} := \max_{h\in\mathcal{D}} |\ell(Uh; u_\theta)|$. under the transformer parameter configuration in Theorem 4.1, the ϵ_c-covering number of $\mathcal{S}(\mathcal{D})$ with respect to $\|\cdot\|_{\infty\mathcal{D}}$ satisfies:*

$$\log \mathcal{N}(\epsilon_c, \mathcal{S}(\mathcal{D}), \|\cdot\|_{\infty\mathcal{D}}) \le O\left(\frac{\log(nL_{\mathcal{T}})}{\epsilon_c^2} D^4 N^{16\beta d + 12\beta} (\log N)^{8d+13}\right).$$

Proof. The proof follows by a straightforward modification of the proof of (Su et al., 2025)[Lemma E.2]. More precisely, replacing the initial parameter bound assumed there by the bound imposed in Theorem 4.1, all subsequent estimates and arguments remain unchanged. With this substitution, the same computational steps yield the desired conclusion, and we omit the repeated details. □

Next, we define the truncated loss function on a bounded domain. For $D > 0$, define the latent cube $\mathcal{D} := [-D, D]^{d_0}$ and

the truncated loss

$$\ell_{\mathrm{trunc}}(x_i; u_\theta) := \ell(x_i; u_\theta) \cdot \mathbf{1}\{h_i \in \mathcal{D}\},$$

where $\ell(x_i; u_\theta)$ is the per-sample CFM loss. The truncated population and empirical risks are R_{trunc} and $\widehat{R}_{\mathrm{trunc}}$ defined as

$$\begin{aligned}\mathcal{R}^{\mathrm{trunc}}(u_\theta) &:= \mathop{\mathbb{E}}_{H_1 \sim p_1^h} \left[\left(\ell(UH_1; u_\theta) - \ell(UH_1; u^\star) \right) \mathbb{1}\{\|H_1\|_\infty \le D\} \right], \\ \widehat{\mathcal{R}}^{\mathrm{trunc}}(u_\theta) &:= \frac{1}{n} \sum_{i=1}^{n} \left(\ell^{\mathrm{trunc}}(x_i; u_\theta) - \ell^{\mathrm{trunc}}(x_i; u^\star) \right). \end{aligned} \tag{61}$$

We have the following lemma on bounding the difference between truncation loss and original loss.

Lemma F.4 (Truncation error)**.** *Under Assumption 4.1 and Assumption 4.2, for the structured estimator with $\|\widehat{s}_\theta\|_\infty = O(\sqrt{\log N})$,*

$$\mathop{\mathbb{E}}_{H_1 \sim p_1^h} [|\ell(UH_1; \widehat{u}_\theta) - \ell_{\mathrm{trunc}}(UH_1; \widehat{u}_\theta)|] \lesssim D \exp\left(-C_2 D^2/2\right) \log N.$$

Proof. By Lemma F.2 and (60), on $\mathcal{E}$ it holds

$$\|\widehat{u}_\theta(X_t, t)\|_2^2 \le 2\|\widehat{U}\alpha_t \widehat{U}^\top X_t\|_2^2 + 2\|\widehat{U}\beta_t \widehat{s}_\theta\|_2^2 + \kappa_t^2 \|(I - \widehat{U}\widehat{U}^\top) X_t\|_2^2.$$

Using $\|\widehat{s}_\theta\|_2 \le \sqrt{d_0}\|\widehat{s}_\theta\|_\infty = O(\sqrt{d_0 \log N})$ and that the projection terms are bounded by $\|X_t\|_2^2$,

$$\|\widehat{u}_\theta(X_t, t)\|_2^2 \lesssim \|X_t\|_2^2 + d_0 \log N. \tag{62}$$

Recall that by (59) the per-sample loss is

$$\ell(UH_1; \widehat{u}_\theta) = (T - t_0)^{-1} \int_{t_0}^{T} \mathop{\mathbb{E}}_{X_0} [\|\widehat{u}_\theta(X_t, t) - \dot{\mu}_t UH_1 - \dot{\sigma}_t X_0\|_2^2] \, dt,$$

with $X_t = \mu_t UH_1 + \sigma_t X_0$ and $X_0 \sim \mathcal{N}(0, I_{d_x})$. The elementary inequality $\|a + b\|_2^2 \le 2\|a\|_2^2 + 2\|b\|_2^2$, applied twice, gives

$$\|\widehat{u}_\theta(X_t, t) - \dot{\mu}_t UH_1 - \dot{\sigma}_t X_0\|_2^2 \le 2\|\widehat{u}_\theta(X_t, t)\|_2^2 + 4\dot{\mu}_t^2 \|H_1\|_2^2 + 4\dot{\sigma}_t^2 \|X_0\|_2^2.$$

Take $\mathbb{E}_{X_0}$ on both sides. The Gaussian moments are $\mathbb{E}_{X_0}[\|X_0\|_2^2] = d_x$ and $\mathbb{E}_{X_0}[\|X_t\|_2^2] = \mu_t^2 \|H_1\|_2^2 + \sigma_t^2 d_x$. Substituting these together with (62) gives

$$\mathop{\mathbb{E}}_{X_0} [\|\widehat{u}_\theta(X_t, t) - \dot{\mu}_t UH_1 - \dot{\sigma}_t X_0\|_2^2] \lesssim \|H_1\|_2^2 + \log N.$$

Integrating over $t \in [t_0, T]$ and dividing by $T - t_0$ yields

$$\ell(UH_1; \widehat{u}_\theta) \lesssim \|H_1\|_2^2 + \log N. \tag{63}$$

The sub-Gaussian density bound of Assumption 4.1 and the Gaussian-tail integration over $\mathbb{R}^{d_0} \setminus \mathcal{D}$, specialized to the latent density in dimension d_0 as in (Fu et al., 2024, Theorem 4.1), gives

$$\mathbb{E}[\|H_1\|_2^2 \mathbf{1}\{H_1 \notin \mathcal{D}\}] \lesssim D \exp\left(-C_2 D^2/2\right), \quad \mathbb{P}\{H_1 \notin \mathcal{D}\} \lesssim D \exp\left(-C_2 D^2/2\right). \tag{64}$$

Since $\ell_{\mathrm{trunc}}(UH_1; \widehat{u}_\theta) = \ell(UH_1; \widehat{u}_\theta) \mathbf{1}\{H_1 \in \mathcal{D}\}$, the truncation gap equals $\ell \cdot \mathbf{1}\{H_1 \notin \mathcal{D}\}$. Combining (63) with (64),

$$\begin{aligned}\mathop{\mathbb{E}}_{H_1} [|\ell(UH_1; \widehat{u}_\theta) - \ell_{\mathrm{trunc}}(UH_1; \widehat{u}_\theta)|] &\lesssim \mathbb{E}[\|H_1\|_2^2 \mathbf{1}\{H_1 \notin \mathcal{D}\}] + \log N \cdot \mathbb{P}\{H_1 \notin \mathcal{D}\} \\ &\lesssim D \exp\left(-C_2 D^2/2\right) \log N.\end{aligned}$$

This completes the proof. □

We now bound the generalization error of truncated loss using standard symmetrization technique and covering number bound.

Lemma F.5 (Generalization bound under Assumption 2.1)**.** *Let $\widehat{u}_\theta$ be the minimiser of the empirical loss $\widehat{\mathcal{L}}_{\mathrm{CFM}}(u_\theta)$ built from i.i.d. latent samples $\{h_i\}_{i=1}^n$ (and $x_i = Uh_i$). For $\epsilon_c > 0$, set $\mathcal{N} := \mathcal{N}(\epsilon_c, \mathcal{S}(\mathcal{D}), P_h^n, \|\cdot\|_{\infty\mathcal{D}})$ as in Lemma F.3.*

Define the uniform bound on the truncated loss:

$$\Upsilon := \sup_{u_\theta \in \mathcal{T}_R^{h,s,r}, h \in \mathcal{D}} |\ell^{\text{trunc}}(Uh; u_\theta)| = O\big(\log N + D^2\big).$$

Then, we have

$$\Big| \mathop{\mathbb{E}}_{\{h_i\}} \big[\mathcal{R}^{\text{trunc}}(\widehat{u}_\theta) - \widehat{\mathcal{R}}^{\text{trunc}}(\widehat{u}_\theta)\big] \Big| \le \mathop{\mathbb{E}}_{\{h_i\}} \big[\widehat{\mathcal{R}}^{\text{trunc}}(\widehat{u}_\theta)\big] + O\Big(\frac{\Upsilon}{n} \log \mathcal{N} + \epsilon_c\Big).$$

Proof. The proof is standard and follows from the classical symmetrization argument. For completeness, we refer the reader to proof of (Fu et al., 2024)[Theorem 4.1] (paragraph bounding on **B**), where authors carry out same method in detail in the same setting. □

F.3. Main proof of Theorem 4.2

Theorem F.1 (Theorem 4.2 Restated)**.** *Assume Assumption 2.1, Assumption 4.1 and Assumption 4.2. Let $d \times L = d_0 + 1$ be the input shape used by the transformer. Suppose we choose the transformer as in Theorem 4.1. Then, it holds*

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\widehat{u}_\theta)] = O\left(n^{-\frac{1}{16d+15}} (\log n)^{\max\{d_0 + \frac{1}{2}\beta + 1, (8d+16)/3\}}\right).$$

Proof. By Lemma F.1, the event $\mathcal{E} = \{\widehat{U}\widehat{U}^\top = UU^\top\}$ has probability one, so every expectation below may be taken conditional on $\mathcal{E}$. We fix a truncation radius $D > 0$ and a covering scale $\epsilon_c \in (0, 1)$ to be chosen at the end of the proof. Let $u_\theta^\star$ denote the structured velocity field of Theorem 4.1. On $\mathcal{E}$ the matrices $\widehat{U}$ and U have the same column space, so $u_\theta^\star$ belongs to the data-driven structured class (60), and the empirical risk minimizer $\widehat{u}_\theta$ satisfies $\widehat{\mathcal{R}}(\widehat{u}_\theta) \le \widehat{\mathcal{R}}(u_\theta^\star)$, where $\widehat{\mathcal{R}}$ is the untruncated empirical risk and R is the population risk of Definition 4.2. Writing R_{trunc} and $\widehat{\mathcal{R}}_{\text{trunc}}$ cooresponding to ℓ_{trunc} as in (61), we decompose the risk of $\widehat{u}_\theta$ as

$$\mathbb{E}[\mathcal{R}(\widehat{u}_\theta)] = \underbrace{\mathbb{E}[\mathcal{R}(\widehat{u}_\theta) - \mathcal{R}_{\text{trunc}}(\widehat{u}_\theta)]}_{\text{(I)}} + \underbrace{\mathbb{E}[\mathcal{R}_{\text{trunc}}(\widehat{u}_\theta) - \widehat{\mathcal{R}}_{\text{trunc}}(\widehat{u}_\theta)]}_{\text{(II)}} + \underbrace{\mathbb{E}[\widehat{\mathcal{R}}_{\text{trunc}}(\widehat{u}_\theta) - \widehat{\mathcal{R}}(\widehat{u}_\theta)]}_{\text{(III)}} + \underbrace{\mathbb{E}[\widehat{\mathcal{R}}(\widehat{u}_\theta)]}_{\text{(IV)}}. \tag{65}$$

Bounding the truncation gaps (I) and (III). Since every element of the structured class (60) has uniformly bounded latent component $\|\widehat{s}_\theta\|_\infty = O(\sqrt{\log N})$, Lemma F.4 applies and gives

$$|\text{(I)}| \lesssim D \exp\big(-C_2 D^2/2\big) \log N, \quad |\text{(III)}| \lesssim D \exp\big(-C_2 D^2/2\big) \log N.$$

We choose $D^2 = 4\beta d_0 \log N / C_2$ and $D = O(\sqrt{\log N})$, so $|\text{(I)}| + |\text{(III)}| = O((\log N)^{3/2} N^{-2\beta})$.

Bounding the generalization gap (II). Applying Lemma F.5 with these inputs gives

$$\text{(II)} \lesssim \mathbb{E}[\widehat{\mathcal{R}}_{\text{trunc}}(\widehat{u}_\theta)] + [\Upsilon \log \mathcal{N}/n + \epsilon_c].$$

Substituting Lemma F.3 and $D^4 = O((\log N)^2)$ gives

$$\Upsilon(\log \mathcal{N})/n \lesssim N^{16\beta d + 12\beta} (\log N)^{8d+16} / (n\epsilon_c^2),$$

so that

$$\text{(II)} \lesssim \text{(III)} + \text{(IV)} + [N^{16\beta d + 12\beta} (\log N)^{8d+16} / (n\epsilon_c^2) + \epsilon_c]. \tag{66}$$

Bounding the approximation error (IV). Since $\widehat{\mathcal{R}}(\widehat{u}_\theta) \le \widehat{\mathcal{R}}(u_\theta^\star)$ and $u_\theta^\star$ is nonrandom, taking expectations gives $\text{(IV)} \le \mathbb{E}[\widehat{\mathcal{R}}(u_\theta^\star)] = R(u_\theta^\star)$. Hence Theorem 4.1 gives

$$\text{(IV)} \le \mathcal{R}(u_\theta^\star) = \frac{1}{T - t_0} \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u_t^\star - u_\theta^\star\|_2^2 \, p_t \, dx \, dt = O(B^2 N^{-\beta} (\log N)^{d_0 + \beta/2 + 1}). \tag{67}$$

Balancing the free parameters. Combining (65) with the bounds on (I), (III), (66), and (67), and discarding the negligible $O((\log N)^{d_0/2} N^{-2\beta})$ contribution of the truncation gaps derives

$$\mathbb{E}[\mathcal{R}(\widehat{u}_\theta)] \lesssim N^{-\beta} (\log N)^{d_0 + \beta/2 + 1} + N^{16\beta d + 12\beta} (\log N)^{8d+16} / (n\epsilon_c^2) + \epsilon_c.$$

Setting $N = n^{1/[\beta(16d+15)]}, \epsilon_c = n^{-1/(16d+15)}(\log n)^{(8d+16)/3}$ derives

$$\mathbb{E}[\mathcal{R}(\widehat{u}_\theta)] \lesssim n^{-1/(16d+15)} \, (\log n)^{\max(d_0+\beta/2+1,\, (8d+16)/3)}$$

This completes the proof. □

G. Proof of Theorem 4.3

We now give the formal proof of Theorem 4.3:

Theorem G.1 (Distribution Estimation With Wasserstein Distance under Assumption 2.1, Theorem 4.3 Restated)**.** *Let $\widehat{P}_T$ be the distribution obtained at clipped time $[t_0, T]$ by running the flow driven by the learned velocity field $\widehat{u}_\theta$. Assume Assumption 2.1, Assumption 4.1 and Assumption 4.2. Then, for sample size $n \geq d_0$,*

$$\begin{aligned}&\mathop{\mathbb{E}}_{\{x_i\}_{i=1}^n} \left[W_2\big(\widehat{P}_T, P_T\big)\right] \\ &= O\left(n^{-\frac{1}{32d+30}} (\log n)^{\frac{1}{2}\max\{d_0+\beta/2+1,\, (8d+16)/3\}}\right).\end{aligned}$$

Proof of Theorem 4.3. We bound the 2-Wasserstein distance between the estimated and true distributions with the ℓ_2 difference of the velocity field transformer network and the true velocity field. Following the proofs of (Fukumizu et al., 2025, Theorem 3) and (Benton et al., 2024, Theorem 1), we have:

$$\mathbb{E}_{\{h_i\}_{i=1}^n}\left[W_2\left(\widehat{P}_T, P_T\right)\right] \lesssim \sqrt{\mathbb{E}_{\{h_i\}_{i=1}^n}[\mathcal{R}(\widehat{u}_\theta)]}. \tag{68}$$

Now, we plug in the velocity estimation rate from Theorem 4.2. Under Assumption 4.1 (for latent density p_1^h):

$$\mathbb{E}_{\{h_i\}_{i=1}^n}[\mathcal{R}(\widehat{u}_\theta)] = O\left(n^{-\frac{1}{16d+15}} (\log n)^{\max\{d_0+\beta/2+1,\, (8d+16)/3\}}\right).$$

Taking the square root gives the rate for W_2:

$$\mathbb{E}_{\{h_i\}_{i=1}^n}\left[W_2\left(\widehat{P}_T, P_T\right)\right] = O\left(n^{-\frac{1}{32d+30}} (\log n)^{\frac{1}{2}\max\{d_0+\beta/2+1,\, (8d+16)/3\}}\right).$$

This completes the proof. □

H. Additional Experimental Results

In this appendix, we provide a small synthetic study validating the intrinsic-dimension scaling behavior predicted by our theory. Our main theoretical claim is that, under the low-dimensional linear latent subspace assumption, the statistically difficult component of the velocity field is the tangent component, whose complexity depends primarily on the intrinsic dimension d_0 rather than the ambient dimension d_x. The experiments in this appendix test this claim in a controlled setting.

H.1. Synthetic Setup

We consider the synthetic latent-subspace model

$$h \sim \sum_{m=1}^{M} \pi_m \mathcal{N}(\mu_m, \Sigma_m), \qquad x = Uh, \tag{69}$$

where $U \in \mathbb{R}^{d_x \times d_0}$ has orthonormal columns and $h \in \mathbb{R}^{d_0}$ follows a Gaussian mixture model with $M = 4$ components. We generate U by orthogonalizing a Gaussian random matrix and fix one canonical GMM for each value of d_0.

For the forward interpolation path, we use the standard linear flow-matching path

$$X_t = tX_1 + (1-t)X_0, \qquad X_0 \sim \mathcal{N}(0, I_{d_x}), \tag{70}$$

with $t \sim \mathrm{Unif}[\varepsilon, 1-\varepsilon]$ and $\varepsilon = 0.01$.

Our latent model receives the projected latent coordinate

$$\bar{h} = U^\top x_t \in \mathbb{R}^{d_0} \tag{71}$$

together with a sinusoidal time embedding and predicts the tangent velocity in latent space. The analytically known normal component is added back exactly:

$$\widehat{u}(x_t, t) = U\widehat{v}(\bar{h}, t) - \frac{1}{1-t}(I - UU^\top)x_t. \tag{72}$$

The model architecture is fixed across all settings: 2-layer Transformer with hidden width 256, SiLU activations, AdamW optimizer, learning rate 3×10^{-4}, batch size 256, and 250 epochs. We report averages over five random seeds.

H.2. Evaluation Metrics

Our primary metric is the per-dimension tangent oracle MSE:

$$\mathrm{MSE}_{\mathrm{tan}}^{\text{per-dim}} = \frac{1}{d_0}\mathbb{E}\left[\left\|\widehat{v}(\bar{h}, t) - v^\star(\bar{h}, t)\right\|_2^2\right], \tag{73}$$

where the oracle tangent velocity is

$$v^\star(\bar{h}, t) = \frac{1}{t}\bar{h} + \frac{1-t}{t}\nabla_{\bar{h}} \log p_t^h(\bar{h}). \tag{74}$$

We divide by d_0 so that changes in the metric reflect statistical difficulty rather than the number of output coordinates. We also report the total tangent oracle MSE

$$\mathrm{MSE}_{\mathrm{tan}}^{\mathrm{total}} = \mathbb{E}\left[\left\|\widehat{v}(\bar{h}, t) - v^\star(\bar{h}, t)\right\|_2^2\right] \tag{75}$$

as a secondary metric. We additionally verify that the normal MSE—measuring the discrepancy between the analytically reconstructed normal component and its theoretical expression—is numerically negligible across all runs ($\leq 3 \times 10^{-10}$), confirming that the analytical reconstruction is implemented correctly.

H.3. Intrinsic versus Ambient Dimension Scaling

Figure 2 shows per-dimension tangent oracle MSE as a function of training set size for two experimental panels.

Panel A: varying ambient dimension. We fix $d_0 = 4$ and vary $d_x \in \{32, 128, 512\}$. To make the comparison as clean as possible, all three values of d_x use the same latent GMM, the same latent training samples, and the same held-out evaluation

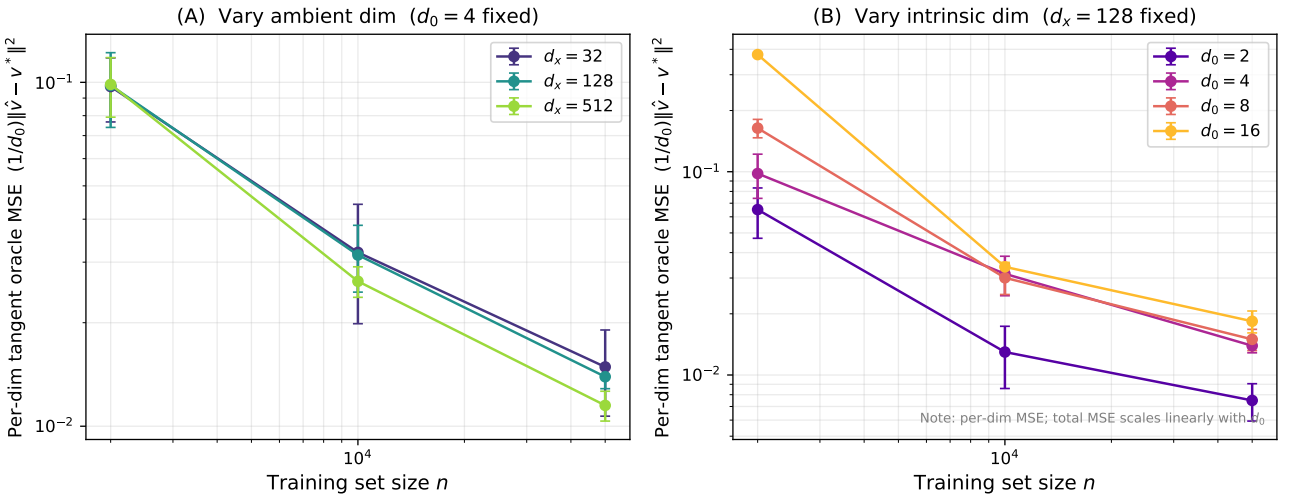

Figure 2. Per-dimension tangent oracle MSE as a function of training set size. *Left (Panel A):* fixing $d_0 = 4$ and varying $d_x \in \{32, 128, 512\}$, the three curves are nearly identical at all training set sizes, confirming that ambient dimension has little effect on tangent estimation accuracy once the estimator operates only on latent coordinates. *Right (Panel B):* fixing $d_x = 128$ and varying $d_0 \in \{2, 4, 8, 16\}$, larger intrinsic dimension produces higher MSE and slower improvement with training set size; $d_0 = 2$ is consistently the easiest and $d_0 = 16$ consistently the hardest. Error bars denote standard deviation across five random seeds.

set; only the ambient embedding matrix U changes across conditions.

The three curves are nearly identical at all training set sizes. At $n_{\text{train}} = 2000$ the spread across d_x values is negligible (ratio of largest to smallest mean MSE: $1.01\times$), and at $n_{\text{train}} = 50000$ it remains small ($1.29\times$), with all three curves lying in the same order of magnitude and overlapping within error bars. There is no systematic trend with d_x: the curves do not separate or diverge as ambient dimension increases. This confirms that, once the estimator operates only on latent coordinates with oracle access to U, changing the ambient embedding dimension has little effect on tangent estimation accuracy.

Panel B: varying intrinsic dimension. We fix $d_x = 128$ and vary $d_0 \in \{2, 4, 8, 16\}$, using one independently sampled canonical GMM per value of d_0.

The curves separate clearly according to intrinsic dimension. At $n_{\text{train}} = 2000$ the ordering is perfectly monotone, with the largest gap occurring between $d_0 = 2$ and $d_0 = 16$. At larger training set sizes, $d_0 = 2$ remains consistently the easiest and $d_0 = 16$ consistently the hardest across all seeds; the middle values $d_0 = 4$ and $d_0 = 8$ are close to each other at $n_{\text{train}} = 50000$ (means 0.014 and 0.015 respectively, within one standard deviation), which we attribute to the different canonical GMMs drawn for each d_0 producing problems of similar difficulty at that sample size. All curves decrease with n_{train}, with higher-d_0 curves decreasing more slowly.

H.4. Tangent Velocity Diagnostics

Figure 3 reports three diagnostics for the representative configuration $(d_x, d_0, n_{\text{train}}) = (128, 4, 50000)$.

Panel A: learned versus oracle tangent velocity. We plot learned against oracle tangent velocity coordinates for 5000 held-out points $(\bar{h}, t)$, coloring each point by latent coordinate index. The scatter lies tightly along the diagonal for all four coordinates, with per-coordinate correlations above 0.99 and per-coordinate MSE ranging from 0.002 to 0.003. There is no visible coordinate-specific failure mode.

Panel B: tangent oracle MSE during training. Per-dimension tangent oracle MSE decreases from 0.19365 at epoch 10 to 0.00240 at epoch 250, a reduction of approximately $80.7\times$. The training loss decreases in parallel. Both quantities improve steadily throughout training with no sign of stagnation or divergence.

Panel C: latent endpoint distribution. We generate 2000 latent samples by integrating the learned ODE $dh/dt = \widehat{v}(h, t)$ from Gaussian noise using Euler integration (100 steps), and compare against 2000 draws from the true latent GMM. The

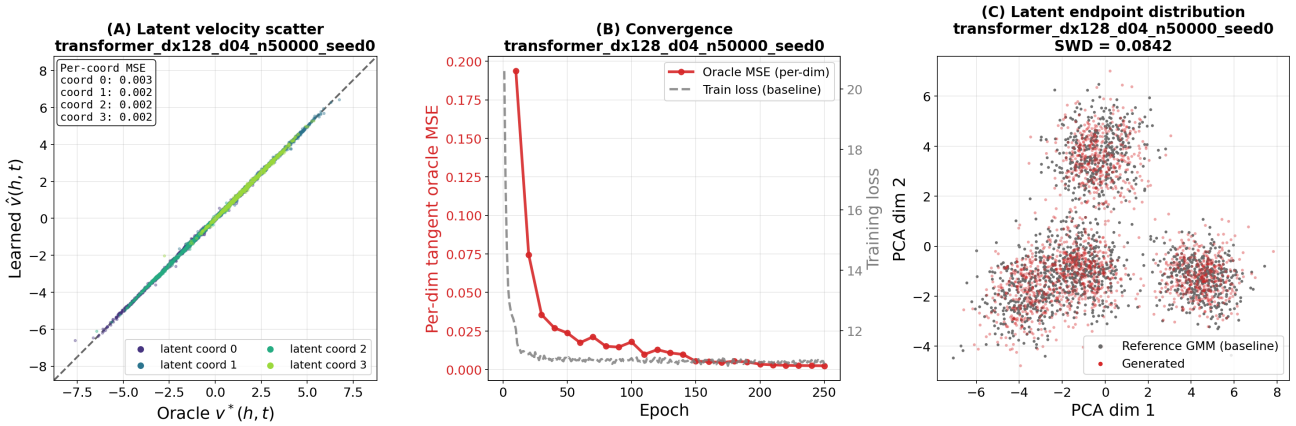

Figure 3. Tangent-component diagnostics for the representative setting $(d_x, d_0, n_{\text{train}}) = (128, 4, 50000)$. *Left (Panel A):* learned tangent velocities align closely with oracle tangent velocities across all four latent coordinates, with per-coordinate correlations above 0.99. *Middle (Panel B):* per-dimension tangent oracle MSE decreases $80.7\times$ over 250 epochs, alongside the flow-matching training loss. *Right (Panel C):* generated latent samples match the reference GMM distribution closely; sliced Wasserstein distance is 0.084.

per-coordinate means and standard deviations of the generated and reference distributions match closely. Sliced Wasserstein distance (500 projections) between the two distributions is 0.084. The joint-PCA scatter shows good overlap between the generated and reference clouds.

H.5. Summary

In Figure 2, Panel A shows that once subspace geometry is provided and the estimator operates in latent coordinates, changing d_x over a $16\times$ range produces negligible variation in per-dimension tangent oracle MSE. Panel B shows that increasing d_0 makes the problem harder, with a clear separation between $d_0 = 2$ and $d_0 = 16$ at all training set sizes and all curves decreasing with n_{train}. The diagnostic panels in Figure 3 confirm that the learned tangent velocity field closely matches the oracle and that the latent ODE generates a distribution consistent with the true GMM. Taken together, these results support the tangent/normal decomposition and the intrinsic-dimension scaling behavior predicted by the theory.