
High-Order Flow Matching: Unified Framework and Sharp Statistical Rates

Maojiang Su^{*†} Jerry Yao-Chieh Hu^{*†} Yi-Chen Lee^{*‡} Ning Zhu[§] Jui-Hui Chung[#]
Shang Wu[†] Zhao Song[‡] Minshuo Chen[†] Han Liu[†]

[†]Northwestern University [‡]National Taiwan University [§]University of Glasgow [#]Princeton University
[‡]Simon Institute of Computing, UC Berkeley

{smj, jhu, shangwu2028}@u.northwestern.edu, b10202055@ntu.edu.tw
zhuning0519@gmail.com, juihui@princeton.edu, {minshuo.chen, hanliu}@northwestern.edu

Abstract

Flow matching is an emerging generative modeling framework that learns continuous-time dynamics to map noise into data. To enhance expressiveness and sampling efficiency, recent works have explored incorporating high-order trajectory information. Despite the empirical success, a holistic theoretical foundation is still lacking. We present a unified framework for standard and high-order flow matching that incorporates trajectory derivatives up to an arbitrary order K . Our key innovation is establishing the marginalization technique that converts the intractable K -order loss into a simple conditional regression with exact gradients and identifying the consistency constraint. We establish sharp statistical rates of the K -order flow matching implemented with transformer networks. With n samples, flow matching estimates nonparametric distributions at a rate $\tilde{O}(n^{-\Theta(1/d)})$, matching minimax lower bounds up to logarithmic factors.

1 Introduction

We present a unified theoretical framework and establish sharp statistical rates for standard and variant flow-matching generative models with high-order velocity fields. A rigorous theoretical understanding of such models is crucial in the current era of rapidly advancing generative AI. Flow-based generative models, particularly those employing Flow Matching (FM) principles [Lipman et al., 2022, Liu et al., 2022], have emerged as a powerful class of methods, achieving state-of-the-art performance across diverse domains such as image, speech, and video generation [Esser et al., 2024, Le et al., 2023, Polyak et al., 2024]. Standard flow matching has focused on learning first-order trajectory dynamics by matching the instantaneous velocity field [Lipman et al., 2022, Liu et al., 2022, Lipman et al., 2024, Gat et al., 2024, Chen and Lipman, 2023].

However, there is a growing interest in leveraging richer dynamical information, such as high-order time derivatives of the trajectory, with the intuition that this could lead to more expressive models, smoother generation paths, improved physical plausibility, or more efficient sampling strategies. This trend is evident in recent empirical works. For instance, High-Order Matching for One-Step Shortcut Diffusion (HOMO) [Chen et al., 2025] and Force Matching (ForM) [Cao et al., 2025] have shown that supervising on acceleration and jerk leads to improved smoothness, stability, and precision in generative tasks, particularly in high-curvature regions where first-order methods falter.

^{*}Equal contribution. Version: October 28, 2025. Please see arXiv for the full and latest version.

Despite these promising empirical explorations into high-order dynamics, there lacks a comprehensive theoretical framework that incorporates derivatives up to an arbitrary order K . Rigorous understanding of its statistical properties is also missing. This paper addresses these gaps by introducing High-Order Flow Matching, a generalized theoretical framework for flow-based generative modeling. Specifically, High-Order Flow Matching defines a K -order velocity field f_t . This field is constructed by concatenating K individual d -dimensional column vector fields $u^1, \dots, u^K$. Each u^k component is designed to capture aspects of the flow dynamics, with u^1 representing the primary velocity and u^k (with $k > 1$) capturing higher-order temporal information of an underlying flow. To complete the theoretical foundation of High-Order Flow Matching, we analyze its statistical rates when implemented with transformers [Vaswani et al., 2017] to align with modern developments in practice.

Contributions. Our contributions are two-fold:

- **High-Order Flow Matching: A Unified Theoretical Framework.** We present a unified framework for Flow Matching models. We first introduce the flow ODEs of any order (Definition 3.1) and the mass conservation formula (Theorem 3.2). A key technical innovation is the high-order marginalization technique (Theorem 3.3). This approach, incorporating a consistency constraint, leads to a tractable loss for K -order flow matching (Theorem 3.4). We then prove that High-Order Flow Matching subsumes standard first-order Flow Matching (when $K = 1$, Proposition 3.1) and provides a unified theoretical foundation for understanding emerging high-order flow model approaches. For example, the objective in HOMO [Chen et al., 2025], which target velocity and acceleration, are instantiated by High-Order Flow Matching for $K = 2$.
- **Statistical Rates for High-Order Flow Matching with Transformers.** We provide the first rigorous statistical analysis of the High-Order Flow Matching framework when implemented with transformer architectures. We establish sharp approximation rates for transformers learning the K velocity components $u^1, \dots, u^K$ (Theorem 4.1), derive corresponding estimation error rates (Theorem 4.2), and further provide end-to-end distribution estimation rates under the 2-Wasserstein metric (Theorem 4.3). In addition, we show that these rates are nearly minimax optimal up to logarithmic factors (Theorem 4.4). Importantly, our rates match the established near-minimax optimal rates of standard flow matching [Jiao et al., 2024, Fukumizu et al., 2024].

Organization. Section 2 reviews preliminary concepts about standard flow matching. Section 3 details the High-Order Flow Matching framework, its properties, and its connections to existing methods. Section 4 presents statistical results. Section 5 summarizes our work and discusses the implications of our findings. The appendix includes the supplementary theoretical backgrounds (Appendix B), the detailed proofs of the main text (Appendices C to G), the statistical rates for standard first-order flow matching transformers (Appendix I) and its proof (Appendices J to N).

Notation. We denote the index set $\{1, \dots, I\}$ by $[I]$. Let $x[i]$ denote the i -th component of a vector x . Let $\mathbb{Z}$ denote integers and $\mathbb{Z}_+$ denote positive integers. Given random variables X and Y with marginal densities μ_x and μ_y respectively, we denote the 2-Wasserstein distance between μ_x and μ_y by $W_2(\mu_x, \mu_y)$. Given a matrix $Z \in \mathbb{R}^{d \times L}$, $\|Z\|_2$ and $\|Z\|_F$ denote the 2-norm and the Frobenius norm. Let $u^k \in \mathbb{R}^d$ be column vectors for $k \in [K]$, we denote $\text{col}(u^1, \dots, u^K) \in \mathbb{R}^{kd}$ as the vertical concatenation of $u^1, \dots, u^K$. Let $\text{Div} \cdot$ be the divergence operator.

2 Preliminaries

In this section, we provide a high-level overview of the Flow Model and Flow Matching (FM).

Flow Model. The flow model transforms $X_0 = x_0$ from a source distribution P (e.g., the Gaussian distribution) into samples $X_1 = x_1$ from a target distribution Q . A flow $\psi : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a time-dependent mapping implementing $\psi : (t, x) \mapsto \psi_t(x)$. The flow model is a continuous-time Markov process $(X_t)_{0 \leq t \leq 1}$ defined by applying a flow ψ_t to the random variable $X_0 \sim P$:

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

On the other hand, a time-dependent velocity field $u : [0, 1] \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ implementing $u : (t, x) \mapsto u_t$ defines a unique flow ψ via the following ordinary differential equation (ODE):

$$\frac{d\psi_t}{dt} = u_t(\psi_t(x)) \quad \text{with initial condition} \quad \psi_0(x) = x. \quad (2.1)$$

Given a flow ψ_t , the marginal probability density function (PDF) of flow model $X_t = \psi_t(X_0) \sim p_t$ is a continuous-time probability path $(p_t)_{0 \leq t \leq 1}$. The probability path p_t follows push-forward equation:

$$p_t(x) = [\psi_t]_* p_0(x) := p_0(\psi_t^{-1}(x)) \cdot \left| \det \left[\frac{\partial \psi_t^{-1}}{\partial x} \right] \right|. \quad (2.2)$$

Further, by the equivalence of flows and velocity fields [Lipman et al., 2024], given invertible C^1 diffeomorphism ψ_t , there exists a unique smooth conditional velocity field u_t taking form:

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

For an arbitrary probability path p_t , we define a velocity field u_t that *generates* p_t if its flow ψ_t satisfies (2.2). Continuous Normalizing Flow [Chen et al., 2018] models the velocity field u_t with a neural network u^θ . Once we obtain a well-trained u^θ , we generate samples from solving ODE (2.1).

Flow Matching. Instead of training flow model by maximizing the log-likelihood of training data [Chen et al., 2018], flow matching [Lipman et al., 2022] is a simulation-free framework to train flow generative models without the need of solving ODEs during training. The Flow Matching objective is designed to match the probability path $(p_t)_{0 \leq t \leq 1}$, which allows us to flow from source $p_0 = P$ to target $p_1 = Q$. Suppose u_t generates such probability path p_t , the flow matching loss is

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, X_t \sim p_t} [\|u^\theta(X_t, t) - u_t(X_t)\|_2^2], \quad (2.4)$$

where $t \sim U[0, 1]$, u^θ is a neural network with parameter θ . Flow Matching simplifies the problem of designing a probability path p_t and its corresponding velocity field u_t by adopting a conditional strategy. Formally, conditioning on any arbitrary random vector $Z \in \mathbb{R}^m$ with PDF p_Z , the marginal probability path p_t satisfies

$$p_t(x) = \int p_t(x|z) p_Z(z) dz. \quad (2.5)$$

Suppose conditional velocity field $u_t(x|z)$ generates $p_t(x|z)$, Lipman et al. [2022] show that following marginal velocity field u_t generates marginal probability path p_t under mild assumptions:

$$u_t(x) := \int u_t(x|z) p_{Z|t}(z|x) dz \quad \text{with} \quad p_{Z|t}(z|x) = \frac{p_t(x|z) p_Z(z)}{p_t(x)}, \quad (2.6)$$

where the second equation follows from the Bayes' rule. Combining above, the tractable conditional flow matching loss $\mathcal{L}_{\text{CFM}}$, which satisfies $\nabla_\theta \mathcal{L}_{\text{CFM}}(\theta) = \nabla_\theta \mathcal{L}_{\text{FM}}(\theta)$, is defined as:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, Z \sim p_Z, X_t \sim p_t(\cdot|Z)} [\|u^\theta(X_t, t) - u_t(X_t|Z)\|_2^2]. \quad (2.7)$$

Affine Conditional Flows. The conditional flow matching loss works with any choice of conditional probability path and conditional velocity fields. In this paper, we consider the affine conditional flow with independent data coupling following [Lipman et al., 2022, 2024]:

$$\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x, \quad (2.8)$$

where $\mu_t, \sigma_t : [0, 1] \rightarrow [0, 1]$ are monotone smooth functions satisfying

$$\mu_0 = \sigma_1 = 0, \quad \mu_1 = \sigma_0 = 1, \quad \text{and} \quad \frac{d\mu_t}{dt}, -\frac{d\sigma_t}{dt} > 0 \quad \text{for} \quad t \in (0, 1). \quad (2.9)$$

Setting $Z = X_1 \sim Q$, $X_0 \sim N(0, I)$, the flow ψ_t induces the probability flow $p_t(X_t|X_1) = N(\mu_t X_1, \sigma_t^2 I)$ and velocity field

$$u_t(x|x_1) = \dot{\psi}_t(\psi_t^{-1}(x|x_1)|x_1) = \frac{\dot{\sigma}_t(x - \mu_t x_1)}{\sigma_t} + \dot{\mu}_t x_1. \quad (2.10)$$

Further, using the law of unconscious statistician with $X_t = \psi_t(X_0|X_1)$, the conditional flow matching loss takes the form

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, X_1 \sim q, X_0 \sim N(0, I)} [\|u^\theta(\mu_t X_1 + \sigma_t X_0, t) - (\dot{\mu}_t X_1 + \dot{\sigma}_t X_0)\|_2^2]. \quad (2.11)$$

In practice, for collected i.i.d. data points $\{x_i\}_{i=1}^n$, (2.11) is implemented with Monte-Carlo simulation. To avoid instability, we often clip the interval $[0, 1]$ with t_0 and T . Namely, for any velocity estimator u^θ , we consider the empirical loss function $\hat{\mathcal{L}}_{\text{CFM}}(u^\theta)$:

$$\hat{\mathcal{L}}_{\text{CFM}}(u^\theta) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{X_0 \sim N(0, I)} [\|u^\theta(\mu_t x_i + \sigma_t X_0, t) - (\mu_t x_i + \sigma_t X_0)\|_2^2] dt. \quad (2.12)$$

Transformers. Throughout the paper, we parameterize u^θ by transformers. Due to space limit, we defer formal definition of transformer networks to [Appendix B](#).

3 High-Order Flow Matching

This section extends the flow matching framework in [Section 2](#) to incorporate high-order trajectory information. Recall that these high-order dynamics are proven to be relevant to further improving the performance and stability of flow matching. Specifically, in [Section 3.1](#), we first define a high-order velocity field f_t using an ODE system and subsequently prove its equivalence to the mapping flow ψ_t ([Theorem 3.1](#)). Furthermore, we derive the corresponding Liouville's equation ([Theorem 3.2](#)), which demonstrates mass conservation for this high-order system. Building on this foundation, [Section 3.2](#) addresses the learning objective. We first propose the high-order Flow Matching loss ([Definition 3.2](#)).

However, similar to flow matching [[Lipman et al., 2022](#)], direct optimization is intractable. To address this, we establish the high-order marginalization trick under consistency constraint ([Theorem 3.3](#)). The method allows us to derive a tractable high-order conditional flow matching loss that preserves the original loss's gradients ([Theorem 3.4](#)). [Section 3.3](#) clarify how that High-Order Flow provides a unifying theory. Specifically, we demonstrate that high-order flow matching subsumes existing flow-based generative modeling techniques, with standard Flow Matching serving as a foundational instance within our framework.

3.1 High-Order Flow Model

For $t \in [0, 1]$, let ψ_t and p_t be the time-dependent flow mapping and probability paths follows [Section 2](#). Instead of using velocity field u_t to construct flow ψ_t via the ODE (2.1), we propose using K -order velocity field $f_t : \mathbb{R}^{Kd} \rightarrow \mathbb{R}^{Kd}$ to construct ψ_t :

Definition 3.1 (High-Order Velocity). Let $t \in [0, 1]$, a flow ψ_t can define a K -order velocity field $f_t : \mathbb{R}^{Kd} \rightarrow \mathbb{R}^{Kd}$ via the following ODE:

$$\frac{d}{dt} y_t = \begin{bmatrix} \frac{d^1}{dt^1} \psi_t(x) \\ \frac{d^2}{dt^2} \psi_t(x) \\ \vdots \\ \frac{d^K}{dt^K} \psi_t(x) \end{bmatrix} = \begin{bmatrix} u^1(x_t^{(0)}, t) \\ u^2(x_t^{(0)}, t) \\ \vdots \\ u^K(x_t^{(0)}, t) \end{bmatrix} = f_t(y_t) \quad \text{with} \quad \psi_0(x) = x, \quad (3.1)$$

where $y_t = \text{col}(\psi_t(x), \frac{d}{dt} \psi_t(x), \dots, \frac{d^{K-1}}{dt^{K-1}} \psi_t(x)) := \text{col}(x_t^{(0)}, x_t^{(1)}, \dots, x_t^{(K-1)}) \in \mathbb{R}^{Kd}$ and $u^k : \mathbb{R}^{Kd} \times [0, 1] \rightarrow \mathbb{R}^d$ is k -th order velocity field for all $k \in [K]$. Moreover, notice that $X_t^{(0)} = \psi_t(X_0)$ is random variable since $X_0 \sim p$. Then, the extended state variable of order K is the random vector

$$Y_t = \text{col}(X_t^{(0)}, \dots, X_t^{(K-1)}) \in \mathbb{R}^{Kd} \quad \text{with} \quad X_t^{(k)} := \frac{d^k}{dt^k} \psi_t(x)|_{x=X_0^{(0)}}. \quad (3.2)$$

For $k = 0, \dots, K-1$, define $p_t^k : \mathbb{R}^d \rightarrow \mathbb{R}$ as the probability density function of $X_t^{(k)}$. Denote $\rho_t : \mathbb{R}^{Kd} \rightarrow \mathbb{R}$ as the probability density function of $Y_t = [X_t^{(0)}, \dots, X_t^{(K-1)}]^\top$ at time t . For simplification, we define Y_t satisfy $\frac{d}{dt} Y_t = f_t(Y_t)$ if (3.1) and (3.2) hold.

Remark 3.1 (Total Derivative Constraints). The ODE (3.1) imposes a sequence of total derivative constraints on the velocity fields $u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t)$, for any $k \in [K]$:

$$u^k(x_t^{(0)}, t) = \frac{d^k}{dt^k} \psi_t(x) = \frac{d}{dt} u^{k-1}(x_t^{(0)}, t) = \frac{\partial}{\partial t} u^{k-1}(x_t^{(0)}, t) + \nabla u^{k-1}(x_t^{(0)}, t) \cdot u^1(x_t^{(0)}, t), \quad (3.3)$$

where $u^0(x_t^{(0)}, t) = x_t^{(0)}$. This recursive relation reveals that the velocity fields induced by the flow ψ_t are not independent, but instead coupled through the structure of the ODE via (3.3).

Remark 3.1 guarantees the equivalence between flows ψ_t and K -order velocity field f_t .

Theorem 3.1 (Flow–Velocity Equivalence via ODE). Define the class of structured k -order velocity fields as those of the form:

$$f_t(y_t) = \text{col}(u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t)) \in \mathbb{R}^{Kd}, \quad y_t = \text{col}(x_t^{(0)}, \dots, x_t^{(K-1)}) \in \mathbb{R}^{Kd},$$

where $u^k : \mathbb{R}^{Kd} \times [0, 1] \rightarrow \mathbb{R}^d$ is locally Lipschitz in y_t and continuous in t for any $k \in [K]$. Suppose the velocity fields $u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t)$ satisfy total derivative constraints (3.3). Then, for any initial condition $y_0 \in \mathbb{R}^{Kd}$, the ODE $\frac{d}{dt}y_t = f_t(y_t)$ exists a unique local solution y_t , which defines a K -times differentiable flow $\psi_t(x) := x_t^{(0)}$ and satisfy $\frac{d^k}{dt^k}\psi_t(x) = x_t^{(k)}$ for all $k \in [K]$.

Conversely, any K -times differentiable flow $\psi_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ defines a velocity field f_t via (3.1).

Proof. Please see [Appendix C.1](#) for a detailed proof. $\square$

Recalling from [Section 2](#) and the flow-velocity equivalence established in [Theorem 3.1](#), the K -order velocity field f_t governs the evolution of the probability density ρ_t for the K -order state Y_t . The precise relationship describing this evolution is captured by the mass conservation formula:

Theorem 3.2 (Mass Conservation of High-Order Flow). Let $y_t = (x_t^{(0)}, \dots, x_t^{(K-1)})^\top \in \mathbb{R}^{Kd}$. Let velocity field $f_t(y_t) = (u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t))^\top \in \mathbb{R}^{Kd}$, where $u^k(x_t^{(0)}, t)$ is locally Lipschitz and integrable for all $k \in [K]$. Let $\rho_t : \mathbb{R}^{Kd} \rightarrow \mathbb{R}$ be a time-varying probability density over the extended state $Y_t \in \mathbb{R}^{Kd}$ follows [Definition 3.1](#). Then the following statements are equivalent:

1. The pair (f_t, ρ_t) satisfies the Liouville's equation on the extended space:

$$\frac{\partial}{\partial t}\rho_t(y) + \nabla_y \cdot (\rho_t(y)f_t(y)) = 0, \quad \text{for all } t \in [0, 1].$$

2. Following [Definition 3.1](#), the probability law of Y_t evolves under the flow:

$$\frac{d}{dt}Y_t = f_t(Y_t), \quad \text{with } Y_0 \sim \rho_0, \quad Y_t \sim \rho_t. \quad (3.4)$$

For some arbitrary probability path ρ_t , we define f_t generates ρ_t if (3.4) holds.

Proof. Please see [Appendix C.2](#) for a detailed proof. $\square$

3.2 High-Order Flow Matching

To model the K -order velocity field f_t , we introduce following high-order flow matching loss:

Definition 3.2 (High-Order Flow Matching Loss). Let f_t denote the ground truth K -order velocity field and f_t^θ be its estimator parameterized by a neural network. Let ρ_t be the probability density function of Y_t . Then, the K -order Flow Matching objective minimizes the following regression loss:

$$\mathcal{L}_{\text{FM}}^K(\theta) = \mathbb{E}_{t, Y_t \sim \rho_t} [D(f_t(Y_t), f_t^\theta(Y_t))],$$

where D is a dissimilarity measure between vectors, such as the squared ℓ_2 -norm.

Similar to standard flow matching, the ground truth velocity f_t is intractable. To address this, we adopt the conditional flow matching loss to train our model, leveraging the equivalence between the flow matching loss and its conditional counterpart. As a preliminary step, we introduce the marginalization trick for high-order flow matching.

Theorem 3.3 (Marginalization). Recall that for some arbitrary probability path ρ_t , f_t generates ρ_t if $Y_t \sim \rho_t$ for all $t \in [0, 1]$. Let Z be a random variable, if $f_t(x|z)$ is conditionally integrable and generates the conditional probability path $\rho_t(\cdot|z)$, then the marginal velocity $f_t := \int f_t(y|z)p_t(z|y)dz$ generates the marginal probability path p_t .²

Proof. Please see [Appendix C.3](#) for a detailed proof. $\square$

Now we are ready to prove the higher version of the equivalence between the flow matching loss and conditional flow matching loss. We first define the tractable K -order conditional flow matching loss:

$$\mathcal{L}_{\text{CFM}}^K(\theta) = \mathbb{E}_{t, Z, Y_t \sim \rho_{t|Z}(\cdot|Z)} [D(f_t(Y_t|Z), f_t^\theta(Y_t))]. \quad (3.5)$$

Following [Lipman et al. \[2024\]](#), we specify the dissimilarity metric $D(\cdot, \cdot)$ as a Bregman divergence, which measures the distance between vectors $u, v \in \mathbb{R}^{Kd}$ as $D(u, v) := \Phi(u) - [\Phi(v) + (u - v)^\top \nabla \Phi(v)]$ where $\Phi : \mathbb{R}^{Kd} \rightarrow \mathbb{R}$ is a strictly convex function defined on a convex domain $\Omega \subset \mathbb{R}^{Kd}$. Bregman divergences possess a key property allowing interchanging gradients and expectations [[Holderrieth et al., 2025](#), [Lipman et al., 2024](#)]:

$$\nabla_v D(\mathbb{E}[Y], v) = \mathbb{E}[\nabla_v D(Y, v)] \quad \text{for any random vector } Y \in \mathbb{R}^{Kd}. \quad (3.6)$$

This property implies that the gradients of the flow matching loss and the conditional flow matching loss are identical, making the two objectives equivalent for training.

Theorem 3.4 (Gradient Equivalence of Losses). Let the Flow Matching loss $\mathcal{L}_{\text{FM}}^K$ be defined as in [Definition 3.2](#), and the Conditional Flow Matching loss $\mathcal{L}_{\text{CFM}}^K$ be defined as in (3.5). Then, when $D(\cdot, \cdot)$ is a Bregman divergence, the gradients of the two losses coincide:

$$\nabla \mathcal{L}_{\text{FM}}^K(\theta) = \nabla \mathcal{L}_{\text{CFM}}^K(\theta).$$

Proof. Please see [Appendix C.4](#) for a detailed proof. $\square$

We now consider training the model using the pre-constructed conditional flow $\psi_t(x | x_1)$ as described in [Section 2](#). By the equivalence between flows and high-order velocity fields ([Theorem 3.1](#)), there exists a unique smooth conditional K -order velocity field f_t such that the conditional trajectory y_t satisfies the ODE: $\frac{d}{dt} y_t = f_t(y_t)$, in accordance with (3.1). Following [Definition 3.1](#), we specify $\psi_t(x | x_1) = \mu_t x_1 + \sigma_t x$, which induces a family of k -th order velocity fields u^k . By [Definition 3.1](#), for all $k \in [K]$, we have

$$u^k(x_t^{(0)}, t) = \frac{d^k}{dt^k} x_t^{(0)} = \frac{d^k}{dt^k} \psi_t(x). \quad (\text{By Definition 3.1})$$

Because ψ_t is an invertible diffeomorphism, we define $x' = \psi_t^{-1}(x)$ and obtain

$$u^k(\psi_t(x), t) = u^k(x', t) = \frac{d^k}{dt^k} \psi_t(\psi_t^{-1}(x')).$$

Extending this to the conditional setting, the conditional k -th order velocity field becomes

$$u^k(x, t | X_1^{(0)}) = \frac{d^k}{dt^k} \psi_t(\psi_t^{-1}(x | X_1^{(0)} | X_1^{(0)})). \quad (3.7)$$

Combining the results above, we now revisit the tractable training loss by setting $Z = X_1^{(0)} \sim q$:

$$\mathcal{L}_{\text{CFM}}^K(\theta) = \mathbb{E}_{t, X_1^{(0)} \sim q, Y_t \sim \rho_{t|X_1^{(0)}}(\cdot|X_1^{(0)})} [D(f_t(Y_t | X_1^{(0)}), f_t^\theta(Y_t))]. \quad (\text{By (3.5)})$$

For further simplifications, we adopt the squared ℓ_2 norm as the Bregman divergence. Let u^k denote the k -th order velocity field, and $u^{k,\theta}$ be its estimator parameterized by a neural network. Denoting the distribution of the k -th order state as $X_t^{(k)} \sim p_t^k$, the training objective becomes

$$\begin{aligned} \mathcal{L}_{\text{CFM}}^K(\theta) &= \mathbb{E}_{t, X_1^{(0)} \sim q, Y_t \sim \rho_{t|X_1^{(0)}}(\cdot|X_1^{(0)})} \left[\|f_t(Y_t | X_1^{(0)}) - f_t^\theta(Y_t)\|_2^2 \right] \quad (\text{By (3.5)}) \\ &= \mathbb{E}_{t, X_1^{(0)} \sim q, Y_t \sim \rho_{t|X_1^{(0)}}(\cdot|X_1^{(0)})} \left[\sum_{k=1}^K \|u^k(Y_t, t | X_1^{(0)}) - u^{k,\theta}(Y_t, t)\|_2^2 \right] \quad (\text{By Definition 3.1}) \end{aligned}$$

²The marginal velocity f_t implies a consistency constraint: $u_t^k(y) = \int u_t^k(y|z) \cdot p_t(z|y) dz$ for all $k \in [K]$.

$$\begin{aligned}
&= \mathbb{E}_{t, X_1^{(0)} \sim q} \left[\sum_{k=1}^K \mathbb{E}_{X_0^{(0)} \sim p(\cdot | X_1^{(0)})} \left\| \frac{d^k}{dt^k} \psi_t(X_0^{(0)} | X_1^{(0)}) - u^{k, \theta}(X_t^{(0)}, t) \right\|_2^2 \right] \quad (\text{By (3.7)}) \\
&= \sum_{k=1}^K \mathbb{E}_{t, X_1^{(0)} \sim q, X_0^{(0)} \sim p(\cdot | X_1^{(0)})} \left[\left\| \frac{d^k}{dt^k} \psi_t(X_0^{(0)} | X_1^{(0)}) - u^{k, \theta}(X_t^{(0)}, t) \right\|_2^2 \right]. \quad (3.8)
\end{aligned}$$

The intermediate states $X_t^{(1)}, \dots, X_t^{(k-1)}$ are determined by $X_0^{(0)}$ via the relation $X_t^{(k)} := \frac{d^k}{dt^k} \psi_t(x)|_{x=X_0^{(0)}}$. Therefore, the inside expectation only needs to be taken over $X_0^{(0)}$.

Now, we consider the affine conditional flow $\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$ follows [Section 2](#). Applying (3.8), the high-order conditional flow matching loss takes the form

$$\mathcal{L}_{\text{CFM}}^K(\theta) = \sum_{k=1}^K \mathbb{E}_{t, X_1^{(0)} \sim q, X_0^{(0)} \sim p(\cdot | X_1^{(0)})} \left[\left\| (\mu_t^{(k)} X_1^{(0)} + \sigma_t^{(k)} X_0^{(0)}) - u^{k, \theta}(X_t^{(0)}, t) \right\|_2^2 \right].$$

In practice, we train the general high-order velocity estimator $u^{1, \theta}, \dots, u^{K, \theta}$ with i.i.d samples $\{x_i\}_{i=1}^n$ by optimizing the empirical high-order conditional flow matching loss:

$$\hat{\mathcal{L}}_{\text{CFM}}^K := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim p(\cdot | X_1^{(0)})} \left[\left\| (\mu_t^{(k)} x_i + \sigma_t^{(k)} X_0^{(0)}) - u^{k, \theta}(X_t^{(0)}, t) \right\|_2^2 \right] dt. \quad (3.9)$$

A significant theoretical consequence of learning the complete K -order velocity field f_t is the ability to employ high-order numerical integration schemes for sampling. For instance, to solve the ODE (3.1), we use K -th order Taylor expansion with step size h for the numerical integration:

$$x_{t+h}^{(0)} = x_t + h u^{1, \theta}(x_t^{(0)}, t) + \frac{h^2}{2!} u^{2, \theta}(x_t^{(0)}, t) + \dots + \frac{h^K}{K!} u^{K, \theta}(x_t^{(0)}, t). \quad (3.10)$$

3.3 Unified Perspective on High-Order Flow Dynamics

We show that our K -order flow matching framework offers a significant unification perspective and a theoretical foundation on existing flow-based generative modeling. Firstly, our framework subsumes standard first-order Flow Matching [[Lipman et al., 2022](#)] as a direct special case.

Proposition 3.1 (Reduction to Standard First-Order Flow Matching). When $K = 1$, the entire K -order flow matching framework, including the governing ODE, the probability path definition via the continuity equation, and the K -order flow matching objective, becomes precisely equivalent to the standard first-order Flow Matching framework as detailed in [[Lipman et al., 2022, 2024](#)].

Proof. Please see [Appendix C.5](#) for a detailed proof. □

Proposition 3.1 establishes our K -order framework as a strict generalization of standard first-order Flow Matching. Beyond encompassing established methods, our K -order framework provides a robust theoretical structure for understanding models that leverage high-order trajectory dynamics.

For instance, HOMO framework [[Chen et al., 2025](#)] defines its training objective ([[Chen et al., 2025](#), Definition 4.3]) by matching network predictions against the true velocity $\dot{x}$ and acceleration $\ddot{x}$ of trajectories. Removing the regularization term (aligns with our total derivative constraints [Remark 3.1](#)), their loss is also a direct instantiation of our K -order framework's objective ([Definition 3.2](#)) for $K = 2$. Furthermore, while the Force Matching (ForM) model [[Cao et al., 2025](#)] introduces specific relativistic constraints, its fundamental generative mechanism involves matching a target “force” field ([[Cao et al., 2025](#), Definition 4.1]). Given that force is proportional to acceleration, if separated from its relativistic regularization, aligns with matching the second-order information captured within our $K = 2$ framework.

In summary, the K -order flow matching framework serves as a unifying theoretical structure. It not only subsumes standard flow matching but also provides formal grounding for models that have intuitive or empirical benefits of incorporating richer, high-order dynamical information. The subsequent statistical analysis in [Section 4](#) builds upon this unified perspective.

4 Statistical Rates of High-Order Flow Matching Transformers

This section characterizes sharp statistical rates for K -order flow matching transformers. Building on [Section 2](#) and [Section 3](#), we consider the case of affine conditional flow with independent data coupling. We focus on transformer architectures as Flow matching (FM) with transformers powers today's best generative models, including MovieGen [\[Polyak et al., 2024\]](#) and Voicebox [\[Le et al., 2023\]](#) by Meta, and Rectified Flow [\[Esser et al., 2024\]](#) by Stability AI. [Section 4.1](#) and [Section 4.2](#) establish bounds for the approximation and estimation of the K -order velocity. Based on the K -order velocity estimation rates, [Section 4.3](#) analyzes the distribution estimation rate under the 2-Wasserstein metric. Finally, [Section 4.4](#) presents the nearly minimax optimality of the K -order velocity estimators.

Transformers. We defer standard definition of transformer to [Appendix B](#) due to the page limit.

4.1 High-Order Velocity Approximation

To establish a statistical theory for K -order flow matching transformers, we first investigate an approximation theory for the K -order velocity under sub-Gaussian assumption. In particular, we characterize the regularity of the target density function $q(x_1)$ with Hölder smoothness, defined by:

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

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

Also, we define the Hölder ball of radius B by $\mathcal{H}^\beta(\mathbb{R}^d, B) := \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < B\}$.

With [Definition 4.1](#), we state our assumption on the target density function $q(x_1)$:

Assumption 4.1 (Sub-Gaussian Property and Hölder Smoothness of Target Distribution). The target distribution $q(x_1) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B)$. Further, there exist two positive constants C_1 and C_2 such that $q(x_1) \leq C_1 \exp(-C_2 \|x_1\|_2^2/2)$.

[Assumption 4.1](#) provides a tail bound for the approximation error, and we leverage it to address the error outside the bounded domain where our transformer approximation applies. We now present the approximation theory for high-order flow matching transformers.

Theorem 4.1 (K -order Velocity Approximation with Transformers). Assume [Assumption 4.1](#). Suppose the k -th order velocity field $u^k(x, t)$ is L_k -Lipschitz for all $k \in 0, \dots, K-1$ in ℓ_2 -distance. Let $\epsilon \in (0, 1)$ be the precision parameter satisfying $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$ and smoothness parameter $\beta > 0$. Then, there exists transformers $u^{1,\theta}(x, t), \dots, u^{K,\theta}(x, t) \in \mathcal{T}_R^{h,s,r}$ such that for any $x \in \mathbb{R}^{d_x}$ and $t \in [0, 1]$, it holds:

$$\sum_{k=1}^K \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt = O(N^{-2\beta} \cdot (\log N)^{\frac{d_x}{2}-1}).$$

Further, for all $k \in [K]$, the parameter bounds in transformer network class satisfy

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

where $\lambda^{-1} = O(N^\beta \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof. Please see [Appendix D](#) for a detailed proof. □

4.2 High-Order Velocity Estimation

In this section, we apply the approximation results in [Section 4.1](#) to derive K -order velocity estimation rates ([Theorem 4.2](#)). Given a set of i.i.d samples $\{x_i\}_{i=1}^n$, we train transformer networks

$u^{1,\theta}, \dots, u^{K,\theta}$ by minimizing the high-order empirical conditional flow matching loss (3.9):

$$\hat{\mathcal{L}}_{\text{CFM}}^K = \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim \mathcal{N}(0, I)} \left[\left\| (\mu_t^{(k)} x_i + \sigma_t^{(k)} X_0^{(0)}) - u^{k,\theta}(X_t^{(0)}, t) \right\|_2^2 \right] dt.$$

We evaluate the performance of estimators $u^{1,\theta}, \dots, u^{K,\theta}$ through the K -order flow matching risk:

Definition 4.2 (High-Order Flow Matching Risk). Let $u^{k,\theta}$ be the estimator of the k -th order velocity field u^k . Let Θ be the collection of parameters of $u^{1,\theta}, \dots, u^{K,\theta}$. We define the flow matching risk $\mathcal{R}_K(\Theta)$ as the sum of the expected mean-squared difference between $u^{k,\theta}$ and u^k :

$$\mathcal{R}_K(\Theta) := \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{x \sim p_t^0} [\|u^k(x, t) - u^{k,\theta}(x, t)\|_2^2] dt,$$

where the density function p_t^0 represents the probability density function of $X_t^{(0)}$ (Definition 3.1).

Further, we assume the path coefficients of the affine conditional flow preserve regularity.

Assumption 4.2 (Path Regularity). Consider the affine conditional flow $\psi_t(x|X_1^{(0)}) = \mu_t X_1^{(0)} + \sigma_t x$, the k -th derivative of path coefficients σ_t and μ_t are continuous on $[t_0, T]$, where $t_0, T \in [0, 1]$.

Assuming k -th order velocity Lipschitz continuity and affine path regularity (Assumption 4.2), the following theorem presents the upper bounds on estimation error $\mathcal{R}_K(\Theta)$ with sample size n .

Theorem 4.2 (High-Order Velocity Estimation with Transformer). Assume Assumption 4.1 and Assumption 4.2. Let $\hat{u}^{k,\theta} \in \mathcal{T}_R^{h,s,r}$ be the estimator of the k -th order velocity field u^k trained by minimizing the high-order empirical conditional flow matching loss (3.9). Let $\hat{\Theta}$ be the collection of parameters of $\hat{u}^{k,\theta}$ for $k \in [K]$. Suppose the k -th order velocity field $u^k(x, t)$ is L_k Lipschitz for all $k = 0, \dots, K-1$. Suppose we choose the transformers as in Theorem 4.1, then

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\hat{\Theta})] = O\left(n^{-\frac{1}{16d}} \cdot (\log n)^{10d_x}\right),$$

where d is the feature dimension.

Proof. Please see Appendix E for a detailed proof. $\square$

4.3 High-Order Distribution Estimation

Based on the K -order velocity estimation result in Theorem 4.2, we further analyze the distribution estimation rate for K -order flow matching transformer. The next theorem presents the upper bounds on the expectation of 2-Wasserstein distance between the target and estimated distribution induced by estimators $u^{k,\theta}$ trained by optimizing the empirical conditional loss (3.9).

Theorem 4.3 (High-Order Distribution Estimation under 2-Wasserstein Distance). Assume Assumption 4.1 and Assumption 4.2. Let $\hat{P}_T^K$ be the estimated distribution at time T . Then, it holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T^K, P_T^K)] = O\left(n^{-\frac{1}{18d}} \cdot (\log n)^{6d_x}\right),$$

where d is the feature dimension.

Proof. Please see Appendix F for a detailed proof. $\square$

4.4 High-Order Minimax Optimal Estimation

We show that the K -order flow matching transformers achieves nearly minimax optimal rate:

Theorem 4.4 (Minimax Optimality of High-Order Flow Matching Transformers). Assume that the target density function satisfies $q(x_1) \in \mathcal{H}^\beta([-1, 1]^{d_x}, B)$ and $q(x_1) \geq C$ for some positive constant C . Then, under the setting of $18d(\beta + 1) = d_x + 2\beta$, the distribution estimation rate of flow matching transformers presented in Theorem 4.3 matches the minimax lower bound of Hölder distribution class in 2-Wasserstein distance up to a $\log n$ and Lipschitz constants factors.

Proof. Please see [Appendix G](#) for a detailed proof. $\square$

Remark 4.1 (Comparison with Existing Works). Flow matching with ReLU networks is nearly minimax-optimal on Besov densities in W_2 [Fukumizu et al., 2024], and kernel methods achieve comparable rates in W_1 [Kunkel and Trabs, 2025]. We extend these results to all orders K and to the major powerhouse in practice: transformer architectures. Our analysis proves that flow-matching transformers attain near-minimax rates on Hölder densities in W_2 with assuming Lipschitz velocities, subsuming the first-order case at $K = 1$. Please see [Appendix I](#) for details.

5 Discussion, Limitation, and Open Question

[Section 3](#) and [Section 4](#) establish a unified theoretical framework for High-Order Flow Matching and offer a sharp statistical analysis of High-Order Flow Matching transformers. As discussed in [Section 3.3](#), this framework subsumes the not only original first-order [Lipman et al., 2024, 2022] but also many high-order flow matching models [Chen et al., 2025, Cao et al., 2025]. Furthermore, the established sharp statistical rates provide rigorous support for all models under this unified framework. This broad theoretical guarantee, covering both first-order and high-order approaches, helps explain the empirical success of the high-order flow models.

While our analysis provides foundational statistical guarantees, the compelling empirical evidence and our current theoretical framework present an intriguing open question: it does not elucidate a significant improvement in statistical rates with increasing order K . In addition, while our framework offers a unified perspective for numerous empirical studies, these often assume the validity of the consistency constraint within the marginalization process ([Theorem 3.3](#)). Our research indicates that the general validity of this constraint, or indeed the derivation of similar conclusions under broader conditions, remains an open question. We identify three primary directions for future work stemming from these considerations: (i) Sampling Efficiency: The High-Order Flow Matching framework enables the use of a K -th order Taylor expansion sampler. This sampler achieves a local truncation error of $O(h^{K+1})$ per step, with all K velocity components $u^{k,\theta}$ evaluable in parallel. Future empirical work should investigate whether this high-order accuracy per step translates into practical benefits, such as requiring fewer function evaluations for a target sample quality or faster convergence to high-fidelity samples. (ii) Stable Approximation Error Propagation: In standard flow matching using Runge-Kutta Methods, the sequential nature means approximation errors in u_θ evaluations may propagate and amplify within a single step as they influence subsequent intermediate calculations. However, our K -order flow matching approach solves the ODE without this feedback loop, which might leads to more stable error propagation. (iii) Relaxing the Consistency Constraint: A significant direction for future research involves exploring methods to either remove or relax the consistency constraint highlighted in [Theorem 3.3](#).

6 Concluding Remarks

In this work, we introduce High-Order Flow Matching, a generalized theoretical framework for flow-based generative modeling. Specifically, we characterize the relationship between flow ψ_t , K -order velocity field f_t , probability path ρ_t through governing ODE and mass conservation formula ([Definition 3.1](#) and [Theorem 3.2](#)). Then we purpose the K -order flow matching loss and establish a tractable equivalent conditional K -order flow matching loss ([Theorem 3.4](#)) via high-order marginalization trick ([Theorem 3.3](#)). Further, we prove that High-Order Flow Matching subsumes standard first-order Flow Matching for $K = 1$ ([Proposition 3.1](#)) and providing a unified theoretical foundation for understanding emerging high-order flow model approaches such as HOMO [Chen et al., 2025]. Our second primary contribution is the first rigorous statistical analysis of this High-Order Flow Matching framework when implemented with transformers. We establish sharp approximation, estimation, and distribution learning rates ([Theorems 4.1 to 4.3](#)), and demonstrate their near-minimax optimality up to logarithmic factors ([Theorem 4.4](#)).

Related Work. We defer an extended discussion on related work to [Appendix A](#) due to page limits.

Acknowledgments

JH would like to thank Sitan Chen, Mimi Gallagher, Sara Sanchez, T.Y. Ball, Dino Feng and Andrew Chen for valuable conversations; David Ting-Chun Liu, Sophia Pi and Mingcheng Lu for pointing out typos; Mingcheng Lu, Weimin Wu, Yibo Wen and David Liu for collaborations on related topics; and the Red Maple Family for support. The authors would like to thank the anonymous reviewers and program chairs for constructive comments.

Lastly, JH dedicates this work to the memory of Prof. Bi-Ming Tsai, who passed away during its camera-ready preparation (Oct 2025). Her teachings on health, cooking, good taste in food, and the art of thoughtful living continue to inspire JH.

JH is partially supported by the Walter P. Murphy Fellowship. Han Liu is partially supported by NIH R01LM1372201, NSF AST-2421845, Simons Foundation MPS-AI-00010513, AbbVie, Dolby and Chan Zuckerberg Biohub Chicago Spoke Award. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

References

- Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *The Eleventh International Conference on Learning Representations*, 2022.
- Silas Alberti, Niclas Dern, Laura Thesing, and Gitta Kutyniok. Sumformer: Universal approximation for efficient transformers. In *Topological, Algebraic and Geometric Learning Workshops 2023*, pages 72–86. PMLR, 2023.
- Joe Benton, George Deligiannidis, and Arnaud Doucet. Error bounds for flow matching methods. *Transactions on Machine Learning Research*, 2023.
- Yang Cao, Bo Chen, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Mingda Wan. Force matching with relativistic constraints: A physics-inspired approach to stable and efficient generative modeling. *arXiv preprint arXiv:2502.08150*, 2025.
- Bo Chen, Chengyue Gong, Xiaoyu Li, Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Mingda Wan. High-order matching for one-step shortcut diffusion models. *arXiv preprint arXiv:2502.00688*, 2025.
- Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. *The Twelfth International Conference on Learning Representations*, 2023.
- Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.
- Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.
- Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.
- Hengyu Fu, Zhuoran Yang, Mengdi Wang, and Minshuo Chen. Unveil conditional diffusion models with classifier-free guidance: A sharp statistical theory. *arXiv preprint arXiv:2403.11968*, 2024.
- Kenji Fukumizu, Taiji Suzuki, Noboru Isobe, Kazusato Oko, and Masanori Koyama. Flow matching achieves almost minimax optimal convergence. *arXiv preprint arXiv:2405.20879*, 2024.
- Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.
- Thomas Hakon Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Annals of Mathematics*, 20(4):292–296, 1919.
- Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Heidelberg, second revised edition edition, 1993. ISBN 978-3-540-56670-0. doi: 10.1007/978-3-540-78862-1.

- Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary markov processes. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Jerry Yao-Chieh Hu, Weimin Wu, Zhuoru Li, Sophia Pi, Zhao Song, and Han Liu. On statistical rates and provably efficient criteria of latent diffusion transformers (dits). *Advances in Neural Information Processing Systems*, 38:31562–31628, 2024.
- Jerry Yao-Chieh Hu, Hude Liu, Hong-Yu Chen, Weimin Wu, and Han Liu. Universal approximation with softmax attention. *arXiv preprint arXiv:2504.15956*, 2025a.
- Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025b.
- Jerry Yao-Chieh Hu, Weimin Wu, Yi-Chen Lee, Yu-Chao Huang, Minshuo Chen, and Han Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation, and minimax optimality. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025c.
- Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrod Rector-Brooks, Tara Akhoun-Sadegh, Michael Bronstein, Alexander Tong, and Avishek Joey Bose. Sequence-augmented se(3)-flow matching for conditional protein backbone generation. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- Haotian Jiang and Qianxiao Li. Approximation rate of the transformer architecture for sequence modeling. *Advances in Neural Information Processing Systems*, 37:68926–68955, 2024.
- Yuling Jiao, Yanming Lai, Yang Wang, and Bokai Yan. Convergence analysis of flow matching in latent space with transformers. *arXiv preprint arXiv:2404.02538*, 2024.
- Tokio Kajitsuka and Issei Sato. Are transformers with one layer self-attention using low-rank weight matrices universal approximators? *The Twelfth International Conference on Learning Representations*, 2023.
- Junghwan Kim, Michelle Kim, and Barzan Mozafari. Provable memorization capacity of transformers. In *The Eleventh International Conference on Learning Representations*, 2022.
- Lea Kunkel and Mathias Trabs. On the minimax optimality of flow matching through the connection to kernel density estimation. *arXiv preprint arXiv:2504.13336*, 2025.
- Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.
- Yingyu Liang, Zhizhou Sha, Zhenmei Shi, Zhao Song, and Mingda Wan. Hofar: High-order augmentation of flow autoregressive transformers. *arXiv preprint arXiv:2503.08032*, 2025.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *The Eleventh International Conference on Learning Representations*, 2022.
- Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.
- Hude Liu, Jerry Yao-Chieh Hu, Zhao Song, and Han Liu. Attention mechanism, max-affine partition, and universal approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
- Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *The Eleventh International Conference on Learning Representations*, 2022.

- Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.
- Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.
- Sejun Park, Jaeho Lee, Chulhee Yun, and Jinwoo Shin. Provable memorization via deep neural networks using sub-linear parameters. In *Conference on learning theory*, pages 3627–3661. PMLR, 2021.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.
- Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *CoRR*, 2024.
- Maojiang Su, Mingcheng Lu, Jerry Yao-Chieh Hu, Shang Wu, Zhao Song, Alex Reneau, and Han Liu. A theoretical analysis of discrete flow matching generative models. *arXiv preprint arXiv:2509.22623*, 2025.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- Konstantin Yakovlev and Nikita Puchkin. Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*, 2025.
- Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *International Conference on Learning Representations*, 2019.
- Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Appendix

A	Related Work	16
B	Supplementary Background: Transformer Block	17
B.1	Transformers	17
B.2	Lipschitzness of Transformer Network	18
C	Proofs in Section 3	19
C.1	Proof of Theorem 3.1	19
C.2	Proof of Theorem 3.2	20
C.3	Proof of Theorem 3.3	21
C.4	Proof of Theorem 3.4	21
C.5	Proof of Proposition 3.1	22
D	Proof of Theorem 4.1	23
D.1	Auxiliary Lemmas	23
D.2	Main Proof of Theorem 4.1	26
E	Proof of Theorem 4.2	28
E.1	Preliminaries	28
E.2	Auxiliary Lemmas	29
E.3	Main Proof of Theorem 4.2	34
F	Proof of Theorem 4.3	37
G	Proof of Theorem 4.4	39
H	Preliminaries: Universal Approximation of Transformers	40
I	Statistical Rates of Flow Matching Transformers (FMTs)	54
I.1	Velocity Approximation: Generic Hölder Smooth Data Distributions	54
I.2	Velocity Approximation: Stronger Hölder Smooth Data Distributions	55
I.3	Velocity Estimation and Distribution Estimation	56
I.4	Minimax Optimal Estimation	58
J	Proof of Theorem I.1	59
J.1	Auxiliary Lemmas	59
J.2	Velocity Approximation on Bounded Domain	62
J.3	Main Proof of Theorem I.1	69
K	Proof of Theorem I.2	74
K.1	Auxiliary Lemmas	74
K.2	Velocity Approximation on Bounded Domain	76
K.3	Main Proof of Theorem I.2	80
L	Proof of Theorem I.3	83
L.1	Preliminaries	83
L.2	Auxiliary lemmas	84
L.3	Main Proof of Theorem I.3	97
M	Proof of Theorem I.4	102
M.1	Auxiliary Lemmas	102
M.2	Main Proof of Theorem I.4	102
N	Proof of Theorem I.5	105
O	Experimental Validation	106

O.1	Experimental Setup	106
O.2	Results and Discussion	106

Impact Statement

This theoretical work advances the fundamental understanding of flow matching generative models and presents no foreseeable negative social impacts.

A Related Work

In the following, we discuss the recent success of the techniques used in our work. We begin with the universal approximation theory of transformers. Then, we discuss the recent theoretical progress in flow matching framework, including approximation, estimation and minimax optimality theories.

Universality of Transformers. The universality of transformers refers to their capability to approximate arbitrary sequence-to-sequence functions with any desired precision. It is a key to connect flow matching Transformer architectures with meaningful statistical estimations in this work. Yun et al. [2019] first prove this capability with deep stacks of self-attention and feed-forward layers through the idea of *contextual mapping* by assuming a minimal separation among all hidden representations. Subsequent work by [Alberti et al., 2023] extend the guarantee to variants that employ sparse attention mechanisms. Building upon these works, Hu et al. [2025b], Kajitsuka and Sato [2023] show that a transformer block with a single self-attention layer is sufficient to achieve universal approximation with refined contextual mapping techniques. Beyond contextual mapping, Jiang and Li [2024] derive explicit Jackson-type approximation rates for single-layer, one-head Transformers, with errors governed by the low-rank structure of the target attention kernel and by head/FFN budgets. Hu et al. [2025a], Liu et al. [2025] show sequence-to-sequence universality of minimal Transformers (attention-only) via interpolation and max-affine partition constructions. In particular, Liu et al. [2025] prove that a single-head self- or cross-attention module achieves universal approximation under L_∞ for continuous targets and extends to L_p .

Remark A.1 (Approximation with High- vs Low-Temperature Softmax Transformers). We remark that, the current statistical rates rely on high-temperature region of softmax attention (i.e., $\text{Softmax}_\lambda(\cdot)$ with small λ , as in Lemma H.2, Theorem H.1 and (H.25)) to cancel the double-exponential factor reported in [Hu et al., 2024, Remark 3.4]. Is it possible to circumvent with above mentioned different approximation results. For example, Hu et al. [2025a] establish universal approximation in the low-temperature regime (large λ). It suggests that analogous statistical guarantees may also hold without resorting to high-temperature softmax scaling. We leave this for future explorations.

Flow Matching and High-Order Flow Matching. Flow Matching generative modeling [Lipman et al., 2024, Gat et al., 2024, Chen and Lipman, 2023, Lipman et al., 2022, Liu et al., 2022] has advanced the state-of-the-art in various fields and applications, including images [Esser et al., 2024], speeches [Le et al., 2023], audios [Polyak et al., 2024] and biomedical data [Hugu et al., 2024]. These standard flow matching frameworks learn first-order trajectory dynamics (velocity field) to smoothly transport a simple source distribution to the target data distribution. However, there is a growing interest for the role of high-order dynamics in generative modeling with improved accuracy and efficiency, which has been applied in various empirical explorations. For instance, Cao et al. [2025] integrate special relativistic mechanics to enhance the stability of generative modeling by supervising on second-order dynamics (acceleration) to ensure sample velocities remain bounded within a safe limit. Similarly, Liang et al. [2025] also augment flow auto-regressive transformers with second-order supervision by capturing complex dependencies through high-order dynamics.

Statistical Rates and Minimax Optimality of Flow Models. Benton et al. [2023], Albergo and Vanden-Eijnden [2022] measure the convergence of flow models by the L_2 -risk of the velocity field but omit explicit convergence rates. Jiao et al. [2024] work in the latent space of an autoencoder and derive explicit convergence rates for flow models; however, they do not consider the smoothness of the target density class. Su et al. [2025] establish statistical rates for discrete flow matching by deriving a model-agnostic, intrinsic error bound. Fukumizu et al. [2024] demonstrate that flow matching achieves nearly minimax-optimal distribution estimation rates in Besov density function spaces under the 2-Wasserstein distance using ReLU network architectures. Kunkel and Trabs [2025] establish similar results under the 1-Wasserstein distance by employing the kernel density estimators. In this work, we provide the first theoretical evidence of the minimax optimality of any order flow matching using transformer architectures, and our results recover the first order case as a special instance. Notably, we show that flow matching transformers (FMTs) achieve nearly minimax optimal rates

in Hölder density function spaces under the 2-Wasserstein distance without imposing the Lipschitz continuity assumption on the velocity field. Please see [Appendix I](#) for a detailed analysis.

B Supplementary Background: Transformer Block

In this section, we introduce the transformer architecture and its Lipschitzness property. [Appendix B.1](#) provides a formal definition of the transformer network class that we use throughout the paper. Further, [Appendix B.2](#) shows that a transformer block is Lipschitz continuous over a compact domain.

B.1 Transformers

Our notation follows [Hu et al., 2025c, 2024]. To begin with, given a matrix $Z \in \mathbb{R}^{d \times L}$, we denote the i -th column and the j -th row by $Z_{:,i}$ and $Z_{j,:}$ respectively.

Transformer Block. Let $\mathcal{F}^{(\text{SA})} : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d \times L}$ denote the self-attention layer. We use h and s to denote the number of heads and hidden dimension in the self-attention layer, and then we have

$$\mathcal{F}^{(\text{SA})}(Z) := Z + \sum_{i=1}^h W_O^i \cdot (W_V^i Z) \text{Softmax}((W_K^i Z)^\top (W_Q^i Z)), \quad (\text{B.1})$$

where $\text{Softmax}(\cdot)$ is the column-wise softmax function, $W_V^i, W_K^i, W_Q^i \in \mathbb{R}^{s \times d}$, and $W_O^i \in \mathbb{R}^{d \times s}$ are the weight matrices. Let r be the MLP dimension. Then, we define the feed-forward layer:

$$\mathcal{F}^{(\text{FF})}(Z) := Z + W_2 \text{ReLU}(W_1 Z + b_1) + b_2, \quad (\text{B.2})$$

where $W_1 \in \mathbb{R}^{r \times d}$ and $W_2 \in \mathbb{R}^{d \times r}$ are weight matrices, and $b_1 \in \mathbb{R}^r$, and $b_2 \in \mathbb{R}^d$ are bias.

Definition B.1 (Transformer Block). We define a transformer block of h -head, s -hidden dimension, r -MLP dimension, and with positional encoding $E \in \mathbb{R}^{d \times L}$ as

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

Now, we define the transformer networks as compositions of transformer blocks.

Definition B.2 (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.,

$$\mathcal{T}^{h,s,r} := \{f_{\mathcal{T}} : \mathbb{R}^{d \times L} \mapsto \mathbb{R}^{d \times L} \mid f_{\mathcal{T}} = \mathcal{F}^{h,s,r} \circ \dots \circ \mathcal{F}^{h,s,r}\}.$$

Flow Matching Transformer. Following architecture of diffusion transformers (DiTs) in [Hu et al., 2025c, 2024, Peebles and Xie, 2023], we adopt the reshape layer R that converts a vector input $x \in \mathbb{R}^{d_x}$ into the sequential matrix input format $Z \in \mathbb{R}^{d \times L}$ for transformer with $d_x = d \cdot L$.

Definition B.3 (Reshape Layer). The reshape layer $R(\cdot) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d \times L}$ transforms d_x -dimensional input into a $d \times L$ matrix. For any $d_x = i \times i$ image input, $R(\cdot)$ converts it into a sequence representation with feature dimension $d := p^2$ ($p \geq 2$) and sequence length $L := (i/p)^2$. Further, We define the reverse reshape layer $R^{-1}(\cdot) : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d_x}$ as the inverse of $R(\cdot)$.

Finally, we define the following transformer network function class with the reshape layer. To simplify, we define $W_{KQ} := (W_K)^\top W_Q$ and $W_{OV} := W_O W_V$.

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

- $\mathcal{T}_R^{h,s,r} := \{R^{-1} \circ f_{\mathcal{T}} \circ R : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_x} \mid f_{\mathcal{T}} \in \mathcal{T}^{h,s,r}\};$
- Transformer network output bound: $\sup_Z \|f_{\mathcal{T}}(Z)\|_2 \leq C_{\mathcal{T}};$

- Parameter bound in $\mathcal{F}^{(\text{FF})}$: $\max\{\|W_1\|_{2,\infty}, \|W_2\|_{2,\infty}\} \leq C_F^{2,\infty}$, $\max\{\|W_1\|_2, \|W_2\|_2\} \leq C_F^2$;
- Parameter bound in $\mathcal{F}^{(\text{SA})}$: $\|W_{KQ}\|_2 \leq C_{KQ}$, $\|W_{OV}\|_2 \leq C_{OV}$, $\|W_{KQ}\|_{2,\infty} \leq C_{KQ}^{2,\infty}$, $\|W_{OV}\|_{2,\infty} \leq C_{OV}^{2,\infty}$, $\|E^\top\|_{2,\infty} \leq C_E$, where 2, ∞ -norm follows $\|\cdot\|_{2,\infty} := \max_{j \in [L]} \|Z_{:j}\|_2$;
- 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 $Z_1, Z_2 \in \mathbb{R}^{d \times L}$.

We remark that these norm bounds are critical to quantify the complexity of the network class.

B.2 Lipschitzness of Transformer Network

In this section, we show the Lipschitzness for our transformer network class [Definition B.2](#). We begin with a helper lemma and the Lipschitzness of softmax function under inverse-temperature scaling:

Lemma B.1 (Lipschitzness from Bounded Jacobian, Lemma A.6 of [\[Edelman et al., 2022\]](#)). Let $\Delta^{d-1} := \{x \in \mathbb{R}^d \mid x \geq 0, \|x\|_1 = 1\}$ and $c_f > 0$ be some constant. Suppose that $f: \mathbb{R}^d \rightarrow \Delta^{d-1}$ is a differentiable function satisfying $\|Jf(x)\|_{1,1} \leq c_f$ for all x . Then, for any $x_1, x_2 \in \mathbb{R}^d$, it holds

$$\|f(x_1) - f(x_2)\|_1 \leq c_f \|x_1 - x_2\|_\infty.$$

We then give the Lipschitz property of the softmax function:

Lemma B.2 (Lipschitzness of Softmax with Inverse Temperature, Modified from Corollary A.7 of [\[Edelman et al., 2022\]](#)). Let $y, z \in \mathbb{R}^d$. Denote the softmax function with inverse temperature $\lambda > 0$ by $\text{Softmax}_\lambda(x)[i] := \exp(\lambda x[i]) / \sum_k \exp(\lambda x[k])$ for $x \in \mathbb{R}^d, k \in [d]$. Then, it holds

$$\|\text{Softmax}_\lambda(y) - \text{Softmax}_\lambda(z)\|_1 \leq 2\lambda \|y - z\|_\infty.$$

Proof. Our proof follows [\[Edelman et al., 2022\]](#) and incorporates inverse temperature scaling $\beta > 0$. Define $\text{Softmax}_\lambda(x)[i] := p_i$ for $i \in [d]$. Jacobian has $J_{i,i} = \lambda p_i(1 - p_i)$ and $J_{i,j} = -\lambda p_i p_j$ for $i \neq j$. Then, this yields $J = \lambda(\text{diga}(p) - pp^\top)$ and $\|J\|_{1,1} \leq 2\lambda$. Then we apply [Lemma B.1](#).

This completes the proof. $\square$

Therefore, [Lemma B.2](#) implies the Lipschitzness of a self-attention layer given bounded weight matrices W_{KQ} and W_{OV} . Furthermore, with bounded W_1, W_2 in the feed-forward layer and $|\text{ReLU}(x) - \text{ReLU}(y)| \leq |x - y|$ for any $x, y \in \mathbb{R}$, we have Lipschitzness of a transformer block.

C Proofs in Section 3

In this section, we formalize the high-order flow matching. [Appendix C.1](#) establishes the flow–velocity equivalence through an ordinary differential equation argument ([Theorem 3.1](#)). [Appendix C.2](#) ensures the mass conservation in high-order flows ([Theorem 3.2](#)). [Appendix C.3](#) derives the marginalization property ([Theorem 3.3](#)). [Appendix C.4](#) shows the gradient equivalence between the flow matching and conditional flow matching objectives ([Theorem 3.4](#)). Finally, [Appendix C.5](#) unifies the framework by proving that K -order flow matching collapses to the standard first-order case ([Proposition 3.1](#)).

C.1 Proof of [Theorem 3.1](#)

In this section, we present the main proof of [Theorem 3.1](#).

Theorem C.1 ([Theorem 3.1](#) Restated: Flow–Velocity Equivalence via ODE). Define the class of structured k -order velocity fields as those of the form:

$$f_t(y_t) = \text{col}(u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t)) \in \mathbb{R}^{Kd}, \quad y_t = \text{col}(x_t^{(0)}, \dots, x_t^{(K-1)}) \in \mathbb{R}^{Kd},$$

where $u^k : \mathbb{R}^{Kd} \times [0, 1] \rightarrow \mathbb{R}^d$ is locally lipschitz in y_t and continues in t for any $k \in [K]$. Suppose the velocity fields $u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t)$ satisfy total derivative constraints (3.3). Then, for any initial condition $y_0 \in \mathbb{R}^{Kd}$, the ODE $\frac{d}{dt}y_t = f_t(y_t)$ exists a unique local solution y_t , which defines a K -times differentiable flow $\psi_t(x) := x_t^{(0)}$ and satisfy $\frac{d^k}{dt^k}\psi_t(x) = x_t^{(k)}$ for all $k \in [K]$. Conversely, any K -times differentiable flow $\psi_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$ defines a velocity field f_t via (3.1).

Proof. We prove both directions:

From velocity field f_t to flow ψ_t : Let $y_0 = (x_0^{(0)}, \dots, x_0^{(K-1)})^\top \in \mathbb{R}^{Kd}$ be any initial condition.

Then, the system (3.1)

$$\frac{d}{dt}y_t = f_t(y_t), \quad \text{with initial condition } y_0,$$

is a standard autonomous first-order ODE on $\mathbb{R}^{Kd}$ with a Lipschitz right-hand side. By the Picard–Lindelöf theorem, there exists a unique local solution y_t . Let us define the flow $\psi_t(x) := x_t^{(0)}$ and since y_t is differentiable, ψ_t is differentiable. By repeatedly applying the total derivative constraint (3.3), we can establish that $\frac{d^k}{dt^k}\psi_t(x) = x_t^{(k)}$ for all $k \in [K]$. Specifically, for any $k \in [K]$, we have:

$$\begin{aligned} x_t^{(k)} &= u^k(x_t^{(0)}, t) && \text{(By definition of the ODE)} \\ &= \frac{d}{dt}u^{k-1}(x_t^{(0)}, t) && \text{(By (3.3))} \\ &= \frac{d}{dt}x_t^{(k-1)} && \text{(By definition of the ODE)} \\ &= \frac{d^k}{dt^k}\psi_t(x). && \text{(By induction)} \end{aligned}$$

This confirms that the k -th order velocity field corresponds exactly to the k -th time derivative of the flow ψ_t .

From flow ψ_t to velocity field f_t : Suppose there is a K -times differentiable flow ψ_t . Define

$$\begin{aligned} y_t &= [\psi_t(x), \frac{d}{dt}\psi_t(x), \dots, \frac{d^{K-1}}{dt^{K-1}}\psi_t(x)]^\top, \\ f_t(y_t) &= \text{col}(\frac{d}{dt}\psi_t(x), \dots, \frac{d^K}{dt^K}\psi_t(x)). \end{aligned}$$

Then, by direct differentiation:

$$\frac{d}{dt}y_t = f_t(y_t).$$

This completes the proof of the bidirectional equivalence. $\square$

C.2 Proof of Theorem 3.2

In this section, we provide the proof of Theorem 3.2.

Theorem C.2 (Theorem 3.2 Restated: Mass Conservation of High-Order Flow). Let $y_t = (x_t^{(0)}, \dots, x_t^{(K-1)})^\top \in \mathbb{R}^{Kd}$. Let velocity field $f_t(y_t) = (u^1(x_t^{(0)}, t), \dots, u^K(x_t^{(0)}, t))^\top \in \mathbb{R}^{Kd}$, where $u^k(x_t^{(0)}, t)$ is locally Lipschitz and integrable for all $k \in [K]$. Let $\rho_t : \mathbb{R}^{Kd} \rightarrow \mathbb{R}$ be a time-varying probability density over the extended state $Y_t \in \mathbb{R}^{Kd}$ follows Definition 3.1. Then the following statements are equivalent:

1. The pair (f_t, ρ_t) satisfies the Liouville's equation on the extended space:

$$\frac{\partial}{\partial t} \rho_t(y) + \nabla_y \cdot (\rho_t(y) f_t(y)) = 0, \quad \text{for all } t \in [0, 1).$$

2. Following Definition 3.1, the probability law of Y_t evolves under the flow:

$$\frac{d}{dt}Y_t = f_t(Y_t), \quad \text{with } Y_0 \sim \rho_0, \quad Y_t \sim \rho_t. \quad (\text{C.1})$$

For some arbitrary probability path ρ_t , we define f_t generates ρ_t if (C.1) holds.

Proof. We prove both directions:

From ODE (C.1) to Liouville's Equation: Let $\phi : \mathbb{R}^{Kd} \rightarrow \mathbb{R}$ be any smooth function with compact support (i.e., a test function). We first compute the time derivative of following quantity

$$\mathbb{E}[\phi(Y_t)] = \int \phi(y) \rho_t(y) dy. \quad (\text{C.2})$$

Since the Y_t satisfy the ODE (C.1), the derivative of the expectation becomes:

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\phi(Y_t)] &= \mathbb{E}\left[\frac{d}{dt} \phi(Y_t)\right] && \text{(By switching the expectation and derivation)} \\ &= \mathbb{E}\left[\nabla_y \phi(Y_t) \cdot \frac{d}{dt} Y_t\right] && \text{(By the chain rule)} \\ &= \mathbb{E}[\nabla_y \phi(Y_t) \cdot f_t(Y_t)] && \text{(By the ODE (C.1))} \\ &= \int \nabla_y \phi(y) \cdot f_t(y) \rho_t(y) dy \\ &= - \int \phi(y) \nabla \cdot (f_t(y) \rho_t(y)) dy. && \text{(By the integration by parts)} \end{aligned}$$

Therefore, for any test function ϕ_t , it holds

$$\int \frac{d}{dt} \phi(y) \rho_t(y) + \phi(y) \nabla \cdot (f_t(y) \rho_t(y)) dy = 0,$$

which leads to Liouville's equation

$$\frac{d}{dt} \rho_t(y) + \nabla_y \cdot (\rho_t(y) f_t(y)) = 0.$$

From Liouville's Equation to ODE (C.1): According to the equivalence between the flow ψ_t and its associated velocity field f_t (Theorem 3.1), the ODE (C.1) admits a unique local solution $\tilde{y}_t$, which defines a unique flow $\tilde{\psi}_t$. By the pushforward formula and the definition in Definition 3.1, this flow induces the distribution $\tilde{Y}_t \sim \tilde{\rho}_t$. Moreover, $\tilde{\rho}_t$ satisfies the Liouville equation associated with the velocity field f_t .

Since the Liouville equation admits a unique solution in the space of probability densities starting from the same initial distribution ρ_0 , and both ρ_t and $\tilde{\rho}_t$ solve the same continuity equation with initial condition ρ_0 , we conclude that $\rho_t = \tilde{\rho}_t$. This completes the proof. $\square$

C.3 Proof of Theorem 3.3

This section presents the proof of Theorem 3.3.

Theorem C.3 (Theorem 3.3 Restated: Marginalization). Recall that for some arbitrary probability path ρ_t , f_t generates ρ_t if $Y_t \sim \rho_t$ for all $t \in [0, 1)$. Let Z be a random variable, if $f_t(x|z)$ is conditionally integrable and generates the conditional probability path $\rho_t(\cdot|z)$, then the marginal velocity $f_t := \int f_t(y|z)p_t(z|y)dz$ generates the marginal probability path p_t .

Proof. Applying the mass conservation follows Theorem 3.2, we only need to verify that the f_t and ρ_t satisfy high-order continuity equation, i.e. Liouville's Equation:

$$\begin{aligned} \frac{d}{dt}\rho_t(y) &= \int \frac{d}{dt}\rho_t(y|z)p_Z(z)dz && \text{(By the law of total probability)} \\ &= \int -\nabla \cdot [f_t(y|z)\rho_t(y|z)]p_Z(z)dz && \text{(By Liouville's equation)} \\ &= -\nabla \cdot \int f_t(y|z)\rho_t(y|z)p_Z(z)dz && \text{(By switching differentiation and integration)} \\ &= -\nabla \cdot \int [f_t(y|z)\rho_t(y|z)p_Z(z)/\rho_t(y)] \cdot \rho_t(y)dz \\ &= -\nabla \cdot [f_t(y)\rho_t(y)]. && \text{(By the definition of } f_t(y) \text{ and the Bayes' rule)} \end{aligned}$$

This completes the proof. $\square$

C.4 Proof of Theorem 3.4

In this section, we prove Theorem 3.4.

Theorem C.4 (Theorem 3.4 Restated: Gradient Equivalence of Losses). Let the Flow Matching loss $\mathcal{L}_{\text{FM}}^K$ be defined as in Definition 3.2, and the Conditional Flow Matching loss $\mathcal{L}_{\text{CFM}}^K$ be defined as in (3.5). Then, when $D(\cdot, \cdot)$ is a Bregman divergence, the gradients of the two losses coincide:

$$\nabla \mathcal{L}_{\text{FM}}^K(\theta) = \nabla \mathcal{L}_{\text{CFM}}^K(\theta).$$

Proof. Similar to the Theorem 4 of [Lipman et al., 2024], the result follows from the Marginalization Trick (Theorem 3.3) and the expectation-swapping property of Bregman divergences (3.6). A direct computation then shows that:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{\text{FM}}^K(\theta) &= \nabla_{\theta} \mathbb{E}_{t, Y_t \sim \rho_t} D(f_t(Y_t), f_t^{\theta}(Y_t)) && \text{(By the definition of Flow Matching Loss)} \\ &= \mathbb{E}_{t, Y_t \sim \rho_t} \nabla_{\theta} D(f_t(Y_t), f_t^{\theta}(Y_t)) && \text{(By swapping the expectation and the gradient computation)} \\ &= \mathbb{E}_{t, Y_t \sim \rho_t} \nabla_v D(f_t(Y_t), f_t^{\theta}(Y_t)) \nabla_{\theta} f_t^{\theta}(Y_t) && \text{(By the chain rule)} \\ &= \mathbb{E}_{t, Y_t \sim \rho_t} \nabla_v D(\mathbb{E}_{Z \sim p_{z|t}(\cdot|y)} [f_t(Y_t|Z)], f_t^{\theta}(Y_t)) \nabla_{\theta} f_t^{\theta}(Y_t) && \text{(By the marginalization trick follows Theorem 3.3)} \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{t, Y_t \sim \rho_t} \mathbb{E}_{Z \sim p_{Z|t}(\cdot|y)} [\nabla_v D([f_t(Y_t|Z)], f_t^\theta(Y_t)) \nabla_\theta f_t^\theta(Y_t)] \\
&\quad \text{(By the property of Bregman divergence follows (3.6))} \\
&= \mathbb{E}_{t, Y_t \sim \rho_t} \mathbb{E}_{Z \sim p_{Z|t}(\cdot|y)} [\nabla_\theta D([f_t(Y_t|Z)], f_t^\theta(Y_t))] \quad \text{(By the chain rule)} \\
&= \nabla_\theta \mathbb{E}_{t, Z, Y_t \sim \rho_{t|Z}(\cdot|Z)} [D(f_t(Y_t|Z), f_t^\theta(Y_t))] \quad \text{(By the Bayes' rule)} \\
&= \nabla_\theta \mathcal{L}_{\text{CFM}}^K(\theta).
\end{aligned}$$

This completes the proof. $\square$

C.5 Proof of Proposition 3.1

This section gives the main proof of Proposition 3.1.

Proposition C.1 (Proposition 3.1 Restated: Reduction to Standard First-Order Flow Matching). When $K = 1$, the entire K -order flow matching framework, including the governing ODE, the probability path definition via the continuity equation, and the K -order flow matching objective, becomes precisely equivalent to the standard first-order Flow Matching framework as detailed in [Lipman et al., 2022, 2024].

Proof. The equivalence follows by setting $K = 1$ in the definitions of our K -order framework.

1. **State Variable and ODE:** From Definition 3.1, when $K = 1$, $Y_t = X_t^{(0)} = X_t$. The ODE system $\frac{d}{dt} Y_t = f_t(Y_t)$ simplifies to $\frac{d}{dt} X_t = u^1(X_t)$, which is the governing ODE for standard flow models ([Lipman et al., 2022, 2024]). The K -order velocity field f_t becomes u^1 .
2. **Probability Path and Continuity Equation:** The K -order mass conservation formula (Theorem 3.2) for $K = 1$ reduces to the standard Mass Conservation Formula (Theorem 2 in [Lipman et al., 2024]).
3. **Loss Objective:** The K -order flow matching loss (Definition 3.2), which targets matching f_t^θ to f_t simplifies to matching only the u^1 component: $\mathbb{E}_{t, X_t \sim p_t} [D(u_t^1(X_t), u_t^{1, \theta}(X_t))]$. This is the standard Flow Matching objective (Eq. (5) in [Lipman et al., 2022]). The conditional formulation via Theorem 3.3 similarly simplifies to the conditional Flow Matching loss used for standard FM.

Thus, all core components of the K -order framework align with standard Flow Matching. $\square$

In this section, we prove [Theorem 4.1](#) following steps similar to the velocity approximation in [Appendix J](#): (i) applying the universal approximation of transformers (ii) leveraging the sub-Gaussian property of the target distribution to bound the approximation error of the K order velocity field.

D.1 Auxiliary Lemmas

We begin with the bounds on $p_t(x)$.

$$\frac{C_4}{\sigma_t^{d_x}} \cdot \exp\left(-\frac{\|x\|_2^2 + 1}{\sigma_t^2}\right) \leq p_t(x) \leq \frac{C_1}{(\mu_t^2 + C_2\sigma_t^2)^{d_x/2}} \cdot \exp\left(-\frac{C_2\|x\|_2^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right).$$

Lemma D.2 (Gaussian Tail Bounds). Consider a random vector $X := (X_1, \dots, X_{d_x})^\top \sim N(0, \sigma_t^2 I)$. Let $\omega_{d_x} := 2\pi^{\frac{d_x}{2}}/\Gamma(\frac{d_x}{2})$. Then, the following two inequalities hold:

$$\begin{aligned} \int_{\|X\|>D} \exp\left(-\frac{\|X\|_2^2}{2\sigma_t^2}\right) dX &\leq \omega_{d_x} \sigma_t^2 D^{d_x-2} \exp\left(-\frac{D^2}{2\sigma_t^2}\right), \\ \int_{\|X\|>D} \|X\|_2^2 \exp\left(-\frac{\|X\|_2^2}{2\sigma_t^2}\right) dX &\leq \omega_{d_x} \cdot (\sigma_t^2 D^{d_x} + d_x \sigma_t^4 D^{d_x-2}) \exp\left(-\frac{D^2}{2\sigma_t^2}\right). \end{aligned}$$

$$\int_{\|X\| \geq D} \exp(-\|X\|_2^2/2\sigma_t^2) dX = \omega_{d_x} \int_D^\infty r^{d_x-1} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr.$$
$$\mathrm{d}u = (d_x - 2)r^{d_x-3}\mathrm{d}r, \quad \text{and} \quad v = -\sigma_t^2 \exp\left(-\frac{r^2}{2\sigma_t^2}\right).$$
$$\begin{aligned}
J(D) &= \left[-r^{d_x-2} \sigma_t^2 \exp\left(-\frac{r^2}{2\sigma_t^2}\right) \right]_{r=D}^{\infty} + (d_x-2) \sigma_t^2 \int_D^{\infty} r^{d_x-3} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr \quad (\text{D.1}) \\
&= \sigma_t^2 D^{d_x-2} \exp\left(-\frac{D^2}{2\sigma_t^2}\right) + (d_x-2) \sigma_t^2 \int_D^{\infty} r^{d_x-3} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr \\
&\hspace{25em} (\text{By integration by parts}) \\
&\leq \sigma_t^2 D^{d_x-2} \exp\left(-\frac{D^2}{2\sigma_t^2}\right). \hspace{10em} (\text{By dropping the second term})
\end{aligned}$$

We obtain the final bound

$$\int_{\|X\|>D} \exp\left(-\frac{\|X\|_2^2}{2\sigma_t^2}\right) dX \leq \omega_{d_x} \sigma_t^2 D^{d_x-2} \exp\left(-\frac{D^2}{2\sigma_t^2}\right).$$

This completes the proof of the first inequality. For the second inequality, we have

$$\begin{aligned} & \int_{\|X\|>D} \|X\|_2^2 \exp\left(-\frac{\|X\|_2^2}{2\sigma_t^2}\right) dX \\ &= \omega_{d_x} \int_D^\infty r^2 r^{d_x-1} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr \\ &= \omega_{d_x} \int_D^\infty r^{d_x+1} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr. \end{aligned}$$

Let $K(D) := \int_D^\infty r^{d_x+1} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr$, $u := r^d$ and $dv := r \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr$. Then,

$$du = d_x r^{d_x-1} dr, \quad \text{and} \quad v = -\sigma_t^2 \exp\left(-\frac{r^2}{2\sigma_t^2}\right).$$

Therefore, the integration by parts gives

$$\begin{aligned} & K(D) \\ &= \left[-r^{d_x} \sigma_t^2 \exp\left(-\frac{r^2}{2\sigma_t^2}\right) \right]_{r=D}^\infty + \int_D^\infty \sigma_t^2 \exp\left(-\frac{r^2}{2\sigma_t^2}\right) d_x r^{d_x-1} dr \\ &= \sigma_t^2 D^{d_x} \exp\left(-\frac{D^2}{2\sigma_t^2}\right) + d_x \sigma_t^2 \int_D^\infty r^{d_x-1} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr. \end{aligned}$$

Recalling (D.1)

$$J_D := \int_D^\infty r^{d_x-1} \exp\left(-\frac{r^2}{2\sigma_t^2}\right) dr, \quad \text{and} \quad J_D \leq \sigma_t^2 D^{d_x-2} \exp\left(-\frac{D^2}{2\sigma_t^2}\right),$$

we have

$$\begin{aligned} & K(D) \\ &= \sigma_t^2 D^{d_x} \exp\left(-\frac{D^2}{2\sigma_t^2}\right) + d_x \sigma_t^2 J_D \\ &\leq \sigma_t^2 D^{d_x} \exp\left(-\frac{D^2}{2\sigma_t^2}\right) + d_x \sigma_t^2 \cdot (\sigma_t^2 D^{d_x-2} \exp\left(-\frac{D^2}{2\sigma_t^2}\right)) \quad (\text{By the bound on } J_D) \\ &= (\sigma_t^2 D^{d_x} + d_x \sigma_t^4 D^{d_x-2}) \exp\left(-\frac{D^2}{2\sigma_t^2}\right). \end{aligned}$$

Then we obtain the final bound

$$\int_{\|X\|>D} \|X\|_2^2 \exp\left(-\frac{\|X\|_2^2}{2\sigma_t^2}\right) dX \leq \omega_{d_x} \cdot (\sigma_t^2 D^{d_x} + d_x \sigma_t^4 D^{d_x-2}) \exp\left(-\frac{D^2}{2\sigma_t^2}\right).$$

This completes the proof of the second inequality. $\square$

Applying the universal approximation of transformers (Theorem H.2), we first approximate the k -th order velocity field u^k over a bounded domain with transformers $u^{k,\theta}$.

Lemma D.3 (Approximate k -th Order Flow with Transformers). Assume [Assumption 4.1](#). Let D be an absolute positive constant. Then, for any $x \in [-I, I]^{d_x}$, $t \in [0, 1]$ and $\epsilon \in (0, 1)$, there exist a transformer $u^{k,\theta}(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_0^1 \int_{[-I, I]^{d_x}} p_t(x) \cdot \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 dx dt \leq \epsilon^2,$$

for all $k \in [K]$. Furthermore, the parameter bounds in the transformer network class satisfy

$$\begin{aligned} C_{KQ}, C_{KQ}^{2,\infty} &= O(\lambda^{-1} I^{4d+2} \epsilon^{-4d-2}); C_{OV}, C_{OV}^{2,\infty} = O(\epsilon); \\ C_F, C_F^{2,\infty} &= O(I \epsilon^{-1} L_{k-1}); C_E = O(I); C_{\mathcal{T}} = O(L_{k-1}) \end{aligned}$$

where $\lambda^{-1} = O(I/\epsilon)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof. By specifying the target function as $f = u^k$ and the transformer-based estimator as $g = u^{k,\theta}$ in [Theorem H.2](#), and applying the bound $p_t(x) \leq 1$, the proof follows [Theorem H.2](#) since the reshape layer ([Definition B.3](#)) does not harm the uniform continuity. Further, by the Lipschitzness of the k -th order flow, we have $\|u^k(x, t)\|_2 \leq L_{k-1}$. Then, the parameter bounds in transformer network follow [Lemma H.5](#), where we set the model output bound $C_{\mathcal{T}} = O(L_{k-1})$. This completes the proof. $\square$

To control the approximation error over an unbounded domain, we introduce tail bounds for the probability flow $p_t(x)$ and the weighted squared norms of the u^k , given by $\|u^k(x, t)\|_2^2 \cdot p_t(x)$.

Lemma D.4 (Truncation of x , Modified from Lemma A.1 of [\[Fu et al., 2024\]](#)). Assume [Assumption 4.1](#). Suppose the k -th order velocity field $u^k(x, t)$ is Lipschitz continuous for all $k = 0, \dots, K-1$. Let L_k denote the Lipschitz constant of u^k , and then the velocity fields are uniformly bounded as $|u^k(x, t)| \leq L_{k-1}$ for any $k \in [K]$. Then, for any $R_1, t > 0$ and $k \in [K]$, the following hold

$$\begin{aligned} \int_{\|x\|_{\infty} > R_1} p_t(x) dx &\lesssim R_1^{d_x-2} \exp\left(-\frac{C_2 R_1^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right), \\ \int_{\|x\|_{\infty} > R_1} \|u^k(x, t)\|_2^2 \cdot p_t(x) dx &\lesssim L_{k-1}^2 R_1^{d_x-2} \exp\left(-\frac{C_2 R_1^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right). \end{aligned}$$

Proof. For the first inequality, it follows

$$\begin{aligned} &\int_{\|x\|_{\infty} > R_1} p_t(x) dx \\ &\leq \int_{\|x\|_{\infty} > R_1} \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx && \text{(By [Lemma D.1](#))} \\ &\leq \int_{\|x\|_2 > R_1} \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx && \text{(By } \|x\|_2 \geq \|x\|_{\infty}) \\ &\lesssim R_1^{d_x-2} \exp\left(-\frac{C_2 R_1^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right). && \text{(By [Lemma D.2](#))} \end{aligned}$$

For the second inequality, it follows

$$\begin{aligned} &\int_{\|x\|_{\infty} \geq R_1} \|u^k(x, t)\|_2^2 \cdot p_t(x) dx \\ &\lesssim \int_{\|x\|_{\infty} \geq R_1} \|u^k(x, t)\|_2^2 \cdot \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx && \text{(By [Lemma D.1](#))} \\ &\lesssim \int_{\|x\|_{\infty} \geq R_1} L_{k-1}^2 \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx && \text{(By the Lipschitzness of the } k\text{-th order flow)} \end{aligned}$$

$$\lesssim L_{k-1}^2 R_1^{d_x-2} \exp\left(-\frac{C_2 R_1^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right). \quad (\text{By Lemma D.2})$$

This completes the proof. $\square$

D.2 Main Proof of Theorem 4.1

We now present the formal proof of Theorem 4.1.

Theorem D.1 (Theorem 4.1 Restated: K -order Velocity Approximation with Transformers). Assume Assumption 4.1. Suppose the k -th order velocity field $u^k(x, t)$ is L_k -Lipschitz for all $k \in 0, \dots, K-1$ in ℓ_2 -distance. Let $\epsilon \in (0, 1)$ be the precision parameter satisfying $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$ and smoothness parameter $\beta > 0$. Then, there exists transformers $u^{1,\theta}(x, t), \dots, u^{K,\theta}(x, t) \in \mathcal{T}_R^{h,s,r}$ such that for any $x \in \mathbb{R}^{d_x}$ and $t \in [0, 1]$, it holds:

$$\sum_{k=1}^K \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt = O(N^{-2\beta} \cdot (\log N)^{\frac{d_x}{2}-1}).$$

Further, for all $k \in [K]$, the parameter bounds in transformer network class satisfy

$$C_{KQ}, C_{KQ}^{2,\infty} = O(\lambda^{-1} N^{2\beta(2d+1)} (\log N)^{2d+1}); \quad C_{OV}, C_{OV}^{2,\infty} = O(N^{-\beta}); \\ C_F, C_F^{2,\infty} = O(N^\beta \sqrt{\log N} L_{k-1}); \quad C_E = O(I); \quad C_T = O(L_{k-1}),$$

where $\lambda^{-1} = O(N^\beta \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof of Theorem 4.1. For $u^{1,\theta}(x, t), \dots, u^{K,\theta}(x, t) \in \mathcal{T}_R^{h,s,r}$, we set the transformer output bound $C_T = O(L_{k-1})$ for the k -th network and let R_3 and ϵ_{low} be two positive numbers to be chosen.

First, we decompose the target into three components and bound each of them

$$\begin{aligned} & \sum_{k=1}^K \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt \\ &= \underbrace{\sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty > R_3} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt}_{(T_1)} \\ & \quad + \underbrace{\sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty \leq R_3} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt}_{(T_2)}. \end{aligned}$$

• **Bound on (T_1) .** It holds

$$\begin{aligned} & (T_1) \\ &= \sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty > R_3} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt \\ &\leq 2 \sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty > R_3} \|u^{k,\theta}(x, t)\|_2^2 \cdot p_t(x) dx dt + 2 \sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty > R_3} \|u^k(x, t)\|_2^2 \cdot p_t(x) dx dt \\ & \quad (\text{By expanding } \ell_2\text{-norm}) \end{aligned}$$

$$\begin{aligned}
&\lesssim \sum_{k=1}^K L_{k-1}^2 \int_{t_0}^T \int_{\|x\|_\infty > R_3} p_t(x) dx dt + \sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty > R_3} \|u^k(x, t)\|_2^2 \cdot p_t(x) dx dt \\
&\hspace{20em} (\text{By } C_{\mathcal{T}} = O(L_{k-1})) \\
&\lesssim \sum_{k=1}^K L_{k-1}^2 \int_{t_0}^T \int_{\|x\|_\infty > R_3} p_t(x) dx dt \hspace{2em} (\text{By the Lipschitzness of the } k\text{-th order flow}) \\
&\lesssim R_3^{d_x-2} \exp\left(-\frac{C_2 R_3^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) \sum_{k=1}^K L_{k-1}^2 \int_{t_0}^T dt. \hspace{2em} (\text{By Lemma D.4}) \\
&\leq R_3^{d_x-2} \exp\left(-\frac{C_2 R_3^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) \sum_{k=1}^K L_{k-1}^2. \hspace{2em} (\text{By } t_0, T \in (0, 1))
\end{aligned}$$

• **Bound on (T₂).** For any $\epsilon \in (0, 1)$, it holds

$$(T_2) = \sum_{k=1}^K \int_{t_0}^T \int_{\|x\|_\infty \leq R_3} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt \leq K\epsilon^2. \quad (\text{By Lemma D.3})$$

By the upper-bound on (T₁) and (T₂), we have

$$\begin{aligned}
&\sum_{k=1}^K \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt \\
&= (T_1) + (T_2) \\
&\lesssim R_3^{d_x-2} \exp\left(-\frac{C_2 R_3^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) \sum_{k=1}^K L_{k-1}^2 + K\epsilon^2 \\
&\lesssim \max\left\{R_3^{d_x-2} \exp\left(-\frac{C_2 R_3^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right), \epsilon^2\right\}.
\end{aligned}$$

Finally, for some $N \in \mathbb{N}$ and $\beta > 0$, we set

$$R_3 := \sqrt{\frac{4\beta(\mu_t^2 + C_2 \sigma_t^2) \log N}{C_2}} \quad \text{and} \quad \epsilon := N^{-\beta}.$$

This gives

$$\sum_{k=1}^K \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2 \cdot p_t(x) dx dt = O(N^{-2\beta} \cdot (\log N)^{\frac{d_x}{2}-1})$$

The transformer parameter bounds follow Lemma D.3 with $I = O(\sqrt{\log N})$ and $\epsilon = N^{-\beta} > 0$:

$$\begin{aligned}
C_{KQ}, C_{KQ}^{2,\infty} &= O(\lambda^{-1} N^{2\beta(2d+1)} (\log N)^{2d+1}); C_{OV}, C_{OV}^{2,\infty} = O(N^{-\beta}); \\
C_F, C_F^{2,\infty} &= O(N^\beta \sqrt{\log N} L_{k-1}); C_E = O(I); C_{\mathcal{T}} = O(L_{k-1}),
\end{aligned} \tag{D.2}$$

where $\lambda^{-1} = O(N^\beta \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function.

This completes the proof. $\square$

E Proof of Theorem 4.2

In this section, we derive the estimation rate of the K order flow matching using transformers. We decompose the proof of Theorem 4.2 into the following three parts due to its complexity.

- **Step 0: Preliminaries.** We introduce several essential definitions, including the K order conditional flow matching loss, K order empirical risk and their domain truncation. These definitions are the extensions from the velocity estimation analysis (see Appendix I.3 and Appendix L).
- **Step 1: Controlling Error from Loss Function outside of the Truncated Domain.** By leveraging the sub-Gaussian tail bound and the Lipschitz continuity of the k -th order velocity field, we derive an upper bound on the loss function outside of the truncated domain in Lemma E.1.
- **Step 2: Upper Bound on the Covering Number.** We present a unified upper bound on the covering number that holds across K transformer networks $u^{1,\theta}, \dots, u^{K,\theta}$ in Lemma E.2.
- **Step 3: Generalization Error.** We apply the covering number technique to bound the deviation between the K order empirical risk and the K order true risk in Lemma E.3.

Organizations. Appendix E.1 includes preliminaries on the framework of estimators' quality evaluation. Appendix E.2 introduces auxiliary lemmas. Appendix E.3 presents the main proof.

E.1 Preliminaries

In this section, we consider affine conditional $\psi_t(x|X_1^{(0)}) = \mu_t X_1^{(0)} + \sigma_t x$ following Section 2. Given k -th order velocity estimator $u^{k,\theta}$, we aim to bound the flow matching risk $\mathcal{R}_K(\Theta)$:

$$\mathcal{R}_K(\Theta) := \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{x \sim p_t^0} [\|u^{k,\theta}(x, t) - u^k(x, t)\|_2^2] dt,$$

where the density function p_t and the k -th order flow are induced by the flow ψ_t (Definition 3.1).

In practice, we use the K order conditional flow matching loss to train $u^{1,\theta}, \dots, u^{K,\theta} \in \mathcal{T}_R^{h,s,r}$.

Definition E.1 (High-Order Conditional Flow Matching Loss). Let q be the ground truth distribution and the normal distribution $N(0, I)$ be the source distribution p . Considering affine conditional flows $\psi_t(x|X_1) = \mu_t X_1 + \sigma_t x$, we define the K order conditional flow matching loss:

$$\mathcal{L}_{\text{CFM}}^K(\Theta) := \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_1^{(0)} \sim q, X_0^{(0)} \sim p} [\|(\mu_t^{(k)} X_1^{(0)} + \sigma_t^{(k)} X_0^{(0)}) - u^{k,\theta}(X_t^{(0)}, t)\|_2^2] dt.$$

Further, we define the K order loss function

$$\ell_K(x; u^{1,\theta}, \dots, u^{K,\theta}) := \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0^{(0)} \sim p} [\|(\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)}) - u^{k,\theta}(X_t^{(0)}, t)\|_2^2] dt.$$

Given a set of i.i.d sample $\{x_i\}_{i=1}^n$, we obtain transformers $u^{1,\theta}, \dots, u^{K,\theta}$ by optimizing the empirical conditional flow matching loss:

$$\hat{\mathcal{L}}_{\text{CFM}}^K := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0, I)} [\|(\mu_t^{(k)} x_i + \sigma_t^{(k)} X_0^{(0)}) - u^{k,\theta}(X_t^{(0)}, t)\|_2^2] dt.$$

Then, we define the K -order empirical risk:

Definition E.2 (High-Order Empirical Risk). Let $u^{k,\theta}$ be the estimator of the k -th order velocity field u^k . Further, consider i.i.d training samples $\{x_i\}_{i=1}^n$ and empirical conditional flow matching loss $\widehat{\mathcal{L}}_{\text{CFM}}^K = \frac{1}{n} \sum_{i=1}^n \ell_K(x_i; \cdot)$. Then, we define the K order empirical risk:

$$\widehat{\mathcal{R}}_K(\Theta) := \frac{1}{n} \sum_{i=1}^n \ell_K(x_i; u^{1,\theta}, \dots, u^{K,\theta}) - \frac{1}{n} \sum_{i=1}^n \ell_K(x_i; u^1, \dots, u^K).$$

Remark E.1. Let $\mathcal{R}_K(f_t)$ be the ground truth inputs of the high-order risk; that is, $u^{k,\theta} = u^k$ for any $k \in [K]$. Then, by the definition of high-order velocity field in [Definition 3.1](#), $\mathcal{R}_K(f_t) = 0$ since $f_t(y_t) = (u^1, \dots, u^K)$ is the collection of K order ground truth velocity fields. Further, the gradient equivalence [Theorem 3.4](#) implies that $\mathcal{R}_K(\Theta) = \mathcal{R}_K(\Theta) - \mathcal{R}_K(f_t) = \mathcal{L}_{\text{CFM}}^K(\Theta) - \mathcal{L}_{\text{CFM}}(f_t)$.

Remark E.2. We use $\widehat{\mathcal{L}}_{\text{CFM}}^{K'}$ and $\widehat{\mathcal{R}}_K'$ to denote the conditional flow matching loss and empirical risk with training samples $\{x'_i\}_{i=1}^n$. Then, by the i.i.d assumption on the training sample, we have $\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{L}}_{\text{CFM}}^{K'}(\Theta)] = \mathcal{L}_{\text{CFM}}(\Theta)$, and therefore $\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K'(\Theta)] = \mathcal{R}_K(\Theta)$.

To obtain finite covering number, we introduce the K truncated loss and truncated risk.

Definition E.3 (Domain Truncation of High-Order Loss and Risk). Let $D > 0$ be constant. Given the K order conditional flow matching loss $\ell_K(x; u^{1,\theta}, \dots, u^{K,\theta})$ defined in [Definition E.1](#), we define its truncated counterparts on a bounded domain $\mathcal{D} := [-D, D]^{d_x}$ by

$$\ell_K^{\text{trunc}}(x; u^{1,\theta}, \dots, u^{K,\theta}) := \ell_K(x; u^{1,\theta}, \dots, u^{K,\theta}) \mathbb{1}_{\{\|x\|_\infty \leq D\}}.$$

Given the K order conditional flow matching risk and the K order empirical risk, we define

$$\mathcal{R}_K^{\text{trunc}}(\Theta) := \mathcal{R}_K(\Theta) \mathbb{1}_{\{\|x\|_\infty \leq D\}}, \quad \widehat{\mathcal{R}}_K^{\text{trunc}}(\Theta) := \widehat{\mathcal{R}}_K(\Theta) \mathbb{1}_{\{\|x\|_\infty \leq D\}}.$$

E.2 Auxiliary Lemmas

We follow the proof of velocity estimation in [Appendix L.2](#) and [Appendix L.3](#) to bound the K order flow matching estimation error. Since direct computation of risk is infeasible, we first decompose the K order flow matching risk $\mathcal{R}_K$ into four terms. Then, we leverage the sub-Gaussian property ([Assumption I.1](#)) and the Lipschitzness of transformer network class ([Definition B.2](#)) to bound each term. Specifically, we introduce three lemmas to bound

1. the error from the domain truncation of loss function class ([Lemma E.1](#)),
2. the log covering number of loss function class ([Lemma E.2](#)), and
3. the generalization error bound ([Lemma E.3](#)).

Risk Decomposition. For simplicity, we shorthand $\mathcal{R}_K(u^{1,\theta}, \dots, u^{K,\theta})$ with $\mathcal{R}_K$. Let $\{x'_i\}_{i=1}^n$ be a different set of i.i.d samples independent of the training sample $\{x_i\}_{i=1}^n$. Then we decompose:

$$\begin{aligned} \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K] &= \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K' - \widehat{\mathcal{R}}_K^{\text{trunc}}]]}_{\text{(I)}} \\ &+ \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K^{\text{trunc}} - \widehat{\mathcal{R}}_K^{\text{trunc}}]]}_{\text{(II)}} \\ &+ \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K^{\text{trunc}} - \widehat{\mathcal{R}}_K]}_{\text{(III)}} + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K]}_{\text{(IV)}}, \end{aligned}$$

where we use the fact that $\mathbb{E}_{\{x_i\}_{i=1}^n}[\widehat{\mathcal{R}}_K(\Theta)] = \mathcal{R}_K(\Theta)$ (Remark E.1). This decomposition follows standard statistical learning theory technique, formulated in [Hu et al., 2025c, Fu et al., 2024].

High-Order Truncation Loss. We begin with the bounds on term (I) and term (III).

Lemma E.1 (Upper Bound on the High-Order Truncation Error). Let $u^{1,\theta}, \dots, u^{K,\theta} \in \mathcal{T}_R^{h,s,r}$ be transformers in Theorem 4.1. Then, for any $t \in [t_0, T]$ it holds

$$\mathbb{E}_x[|\ell_K(x; u^{1,\theta}, \dots, u^{K,\theta}) - \ell_K^{\text{trunc}}(x; u^{1,\theta}, \dots, u^{K,\theta})|] \lesssim K D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \max_k \{L_k^2\}.$$

Proof. By Theorem 4.1, we have transformers output bounds $C_{\mathcal{T}} = O(L_{k-1})$ for all k .

For all $k \in [K]$, we define

$$\begin{aligned} \ell_k(x; u^{k,\theta}) &:= \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta}(X_t^{(0)}, t) - (\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)})\|_2^2] dt \\ \ell_k^{\text{trunc}}(x; u^{k,\theta}) &:= \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta}(X_t^{(0)}, t) - (\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)})\|_2^2] dt \mathbb{1}\{\|x\|_{\infty} \leq D\}. \end{aligned}$$

Then, it holds

$$\begin{aligned} & \mathbb{E}_x[|\ell_k(x; u^{k,\theta}) - \ell_k^{\text{trunc}}(x; u^{k,\theta})|] \tag{E.1} \\ &= \mathbb{E}_x[|\ell_k(x; u^{k,\theta}) \mathbb{1}\{\|x\| \geq D\}|] \tag{By Definition E.3} \\ &= \frac{1}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta}(X_t^{(0)}, t) - (\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)})\|_2^2] q(x) dx dt \tag{By Definition E.1} \\ &\lesssim \frac{1}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta}(X_t^{(0)}, t)\|_2^2 + \|\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)}\|_2^2] q(x) dx dt \tag{By expanding the \ell_2-norm} \\ &\lesssim \frac{1}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta}(X_t^{(0)}, t)\|_2^2 + \|\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)}\|_2^2] \exp\left(-\frac{1}{2} C_2 \|x\|_2^2\right) dx dt \tag{By Assumption I.1} \\ &\lesssim \frac{1}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{X_0 \sim N(0,I)} [\max_k \{L_k^2\} + \|\mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)}\|_2^2] \exp\left(-\frac{1}{2} C_2 \|x\|_2^2\right) dx dt \tag{By } C_{\mathcal{T}} = O(\max_k \{L_k\}) \\ &\lesssim \frac{1}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} (\max_k \{L_k^2\} + (\sigma_t^{(k)})^2 d_x + (\mu_t^{(k)})^2 \|x\|_2^2) \exp\left(-\frac{1}{2} C_2 \|x\|_2^2\right) dx dt \tag{ } (x_0 \sim N(0, I)) \\ &\lesssim \frac{D^{d_x-2} \exp(-\frac{1}{2} C_2 D^2)}{T-t_0} \int_{t_0}^T (\max_k \{L_k^2\} + (\sigma_t^{(k)})^2 d_x) dt + \frac{D^{d_x} \exp(-\frac{1}{2} C_2 D^2)}{T-t_0} \int_{t_0}^T (\mu_t^{(k)})^2 dt \tag{By Lemma D.2} \\ &\lesssim D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \max_k \{L_k^2\}. \tag{By Assumption I.2} \end{aligned}$$

Therefore,

$$\mathbb{E}_x[|\ell_K(x; u^{1,\theta}, \dots, u^{K,\theta}) - \ell_K^{\text{trunc}}(x; u^{1,\theta}, \dots, u^{K,\theta})|]$$

$$\begin{aligned}
&\leq \sum_{k=1}^K \mathbb{E}_x [|\ell_k(x; u^{k,\theta}) - \ell_k^{\text{trunc}}(x; u^{k,\theta})|] && \text{(By triangle inequality)} \\
&\lesssim K D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \max_k \{L_k^2\}. && \text{(By (E.1))}
\end{aligned}$$

This completes the proof. $\square$

Covering Number of High-Order Loss Function Class with Transformers. The next lemma extends Lemma L.5 to its higher-order counterpart. Please see Definition L.4 for a precise definition.

Lemma E.2 (Covering Number Bounds for $\mathcal{S}(D)$, Lemma K.2 of [Hu et al., 2025c], Theorem A.17 of [Edelman et al., 2022]). Let $\epsilon_c > 0$. We define the loss function class by $\mathcal{S}(D) := \{\ell_K(x; u^{1,\theta}, \dots, u^{K,\theta}) : \mathcal{D} \rightarrow \mathbb{R} \mid u^{1,\theta}, \dots, u^{K,\theta} \in \mathcal{T}_R^{h,s,r}\}$. Further, we define the norm of loss functions by $\|\ell_K\|_{\infty \mathcal{D}} := \max_{x \in [-D, D]^{d_x}} |\ell_K|$. Then, under transformer parameter configuration in Theorem 4.1 the ϵ_c -covering number of $\mathcal{S}(D)$ with respect to $\|\cdot\|_{\infty \mathcal{D}}$ satisfies:

$$\log \mathcal{N}(\epsilon_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}}) \leq O\left(\frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^2 N^{\beta(16d+12)} (\log N)^{8d+8}\right).$$

Proof. We first derive the log covering number of transformers $u^{1,\theta}, \dots, u^{K,\theta}$ in Theorem 4.1. Then, we extend the results to K order loss function class.

• **Log-Covering Number of Transformers Network Class.** From (D.2), for all $k \in [K]$, we have

$$\begin{aligned}
C_{KQ}, C_{KQ}^{2,\infty} &= O(\lambda^{-1} I^{4d+2} \epsilon^{-4d-2}); C_{OV}, C_{OV}^{2,\infty} = O(\epsilon); \\
C_F, C_F^{2,\infty} &= O(I \epsilon^{-1} L_{k-1}); C_E = O(I); C_T = O(L_{k-1}),
\end{aligned}$$

where $\lambda^{-1} = O(N^\beta \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function, $I = O(\sqrt{\log N})$ and $\epsilon = N^{-\beta} > 0$ some $N \in \mathbb{N}$ and $\beta > 0$.

By Lemma L.5, the bounds on log-covering number follow

$$\begin{aligned}
&\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2) \\
&\leq \frac{\alpha^2 \log(nL/\epsilon_c)}{\epsilon_c^2} \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3 \\
&\lesssim \frac{\alpha^2 \log(nL/\epsilon_c)}{\epsilon_c^2} \left(\underbrace{I^{4/3} \epsilon^{-4/3}}_{(C_F^{2,\infty})^{\frac{4}{3}}} + \underbrace{I^{4/3} \epsilon^{-4/3}}_{(C_F)^{4/3}} \underbrace{\epsilon^{2/3}}_{(C_{OV})^{2/3}} \underbrace{I^{(8d+4)/3} \epsilon^{-8d/3-4/3}}_{(\lambda C_{KQ}^{2,\infty})^{2/3}} + \underbrace{I^{4/3} \epsilon^{-4/3}}_{(C_F)^{\frac{4}{3}}} \underbrace{\epsilon^{2/3}}_{(C_{OV}^{2,\infty})^{\frac{2}{3}}} \right)^3 \\
&\lesssim \frac{\alpha^2 \log(nL/\epsilon_c)}{\epsilon_c^2} \cdot (I^{(8d+8)/3} \epsilon^{-8d/3-2})^3 \\
&= \frac{\alpha^2 \log(nL/\epsilon_c)}{\epsilon_c^2} \cdot I^{8d+8} \epsilon^{-8d-6}.
\end{aligned}$$

By Lemma L.5, we have

$$\begin{aligned}
\alpha &\lesssim (C_F)^2 C_{OV} C_{KQ} (D + C_E) \\
&\lesssim \underbrace{I^2 \epsilon^{-2}}_{(C_F)^2} \cdot \underbrace{\epsilon}_{(C_{OV})} \cdot \underbrace{I^{4d+2} \epsilon^{-4d-2}}_{(C_{KQ})} \cdot (D + C_E) && \text{(By the definition of } \alpha) \\
&= D I^{4d+4} \epsilon^{-4d-3}.
\end{aligned}$$

Altogether, for all $u^{k,\theta} \in \mathcal{T}_R^{h,s,r}$, we have

$$\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_2) \lesssim \frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^2 I^{16d+16} \epsilon^{-16d-12}.$$

Further, by $\|\cdot\|_\infty \leq \|\cdot\|_2$, we have

$$\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_\infty) \lesssim \frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^2 I^{16d+16} \epsilon^{-16d-12}. \quad (\text{E.2})$$

for all $u^{k,\theta} \in \mathcal{T}_R^{h,s,r}$.

- **Log-Covering Number of Loss Function Class.** Let $\delta > 0$. Let $u := \{u^{1,\theta}, \dots, u^{K,\theta}\}$ and $\bar{u} := \{\bar{u}^{1,\theta}, \dots, \bar{u}^{K,\theta}\}$ be two sets of transformers network satisfying $\|u^{k,\theta} - \bar{u}^{k,\theta}\|_\infty \leq \delta$ on domain $x \in [-D, D]^{d_x}$ for all $u^{k,\theta} \in u$ and $u^{s,\theta} \in \bar{u}$. Further, let $\psi_{t,k}^*$ denote the ground truth k -th order conditional velocity field ([Definition E.1](#)):

$$\psi_{t,k}^* := \mu_t^{(k)} x + \sigma_t^{(k)} X_0^{(0)}.$$

Then, the distance between two K order conditional loss functions $\ell_{K,1}(x; u^{1,\theta}, \dots, u^{K,\theta})$ and $\ell_{K,2}(x; \bar{u}^{1,\theta}, \dots, \bar{u}^{K,\theta})$ follows:

$$\begin{aligned} & \left| \ell_{K,1}(x; u^{1,\theta}, \dots, u^{K,\theta}) - \ell_{K,2}(x; \bar{u}^{1,\theta}, \dots, \bar{u}^{K,\theta}) \right| \quad (\text{E.3}) \\ &= \frac{1}{T-t_0} \left| \sum_{k=1}^K \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta} - \psi_{t,k}^*\|_2^2] dt - \sum_{s=1}^K \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [\|\bar{u}^{s,\theta} - \psi_{t,k}^*\|_2^2] dt \right| \\ & \quad (\text{By Definition E.1}) \\ &\leq \sum_{k=1}^K \frac{1}{T-t_0} \left| \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [(u^{k,\theta} + \bar{u}^{k,\theta} - 2\psi_{t,k}^*)^\top (u^{k,\theta} - \bar{u}^{k,\theta})] dt \right| \quad (\text{By triangle inequality}) \\ &\leq \sum_{k=1}^K \frac{\delta}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta} + \bar{u}^{k,\theta} - 2\psi_{t,k}^*\|] dt \quad (\text{By } \|u^{k,\theta} - \bar{u}^{k,\theta}\|_\infty \leq \delta) \\ &\leq \sum_{k=1}^K \frac{\delta}{T-t_0} \int_{t_0}^T \sqrt{2 \mathbb{E}_{X_0 \sim N(0,I)} [\|u^{k,\theta} + \bar{u}^{k,\theta}\|_2^2 + 2\|\psi_{t,k}^*\|_2^2]} dt \quad (\text{By Jensen's inequality}) \\ &\lesssim \sum_{k=1}^K \frac{\delta}{T-t_0} \int_{t_0}^T \sqrt{\max_k \{L_k^2\} + 2\|\psi_{t,k}^*\|_2^2} dt \quad (\text{By } C_T = O(\max_k \{L_k\})) \\ &\lesssim \sum_{k=1}^K \frac{\delta \max_k \{L_k\}}{T-t_0} \int_{t_0}^T dt \quad (\text{By the Lipschitzness of } k\text{-th order flow}) \\ &\lesssim \delta \max_k \{L_k\}. \end{aligned}$$

Finally, we extend the log covering number to the loss function class $\mathcal{S}(D)$ by setting

$$\epsilon'_c := \Omega(\epsilon_c \max_k \{L_k\}).$$

This gives

$$\log \mathcal{N}(\epsilon'_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}}) \leq \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_\infty). \quad (\text{By (E.3)})$$

Therefore,

$$\begin{aligned}
& \log \mathcal{N}(\epsilon'_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}}) \\
& \leq \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_{\infty}) \\
& \lesssim \frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^2 I^{16d+16} \epsilon^{-16d-12} \quad (\text{By (E.2)}) \\
& = O\left(\frac{\log(nL/\epsilon_c)}{(\epsilon'_c)^2} D^2 I^{16d+16} \epsilon^{-16d-12} \max_k \{L_k^2\}\right). \quad (\text{By the definition of } \epsilon'_c)
\end{aligned}$$

Finally, we substitute $I = O(\sqrt{\log N})$ and $\epsilon = N^{-\beta} > 0$. This completes the proof. $\square$

Generalization Bound. Based on covering number bounds results in [Lemma E.2](#), we now analyze the upper bound of generalization error $\left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K^{\text{trunc}}(\hat{\Theta}) - \hat{\mathcal{R}}_K^{\text{trunc}}(\hat{\Theta})] \right|$.

Lemma E.3 (Generalization Bound on K Order Flow Matching Risk). For $\epsilon_c > 0$, let $\mathcal{N} := \mathcal{N}(\epsilon_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}})$ be the covering number of function class of loss $\mathcal{S}(D)$ following [Lemma E.2](#). Let $\hat{\Theta}$ be the collection of parameters of transformers trained by optimizing $\mathcal{L}_{\text{CFM}}(\Theta)$ following [Definition E.1](#) with i.i.d training samples $\{x_i\}_{i=1}^n$. Then we bound the generalization error:

$$\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathcal{R}_K^{\text{trunc}}(\hat{\Theta}) - \hat{\mathcal{R}}_K^{\text{trunc}}(\hat{\Theta}) \right] \leq \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{\Theta}) + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right).$$

Proof. Let $\hat{u}^{k,\theta} \in \mathcal{T}_R^{h,s,r}$ be the approximator of the k -th velocity field u^k obtained from minimizing the high-order empirical conditional flow matching loss:

$$\hat{\mathcal{L}}_{\text{CFM}}^K := \frac{1}{n} \sum_{i=1}^n \sum_{k=1}^K \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0,I)} [\|(\mu_t^{(k)} x_i + \sigma_t^{(k)} X_0^{(0)}) - u^{k,\theta}(X_t^{(0)}, t)\|_2^2] dt.$$

Further, we define

$$\mathcal{R}_k^{\text{trunc}}(\hat{u}^{k,\theta}) := \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{x \sim p_t} [\|u^k(x, t) - u^{k,\theta}(x, t)\|_2^2] \mathbb{1}\{\|x\|_{\infty} \leq D\} dt,$$

and

$$\begin{aligned}
& \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{u}^{k,\theta}) \\
& := \frac{1}{n} \sum_{i=1}^n \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0^{(0)} \sim p} [\|(\mu_t^{(k)} x_i + \sigma_t^{(k)} X_0^{(0)}) - u^{k,\theta}(X_t^{(0)}, t)\|_2^2] dt \cdot \mathbb{1}\{\|x_i\|_{\infty} \leq D\} \\
& \quad - \frac{1}{n} \sum_{i=1}^n \frac{1}{T-t_0} \int_{t_0}^T \mathbb{E}_{X_0^{(0)} \sim p} [\|(\mu_t^{(k)} x_i + \sigma_t^{(k)} X_0^{(0)}) - u^k(X_t^{(0)}, t)\|_2^2] dt \cdot \mathbb{1}\{\|x_i\|_{\infty} \leq D\}
\end{aligned}$$

Since every network configurations and log covering number are identical across all K order velocity fields from [Theorem 4.1](#) and [Lemma E.2](#), for any $k \in [K]$, [Lemma L.8](#) extends to

$$\left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{u}^{k,\theta}) - \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{u}^{k,\theta})] \right| \leq \frac{1}{2} \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{u}^{k,\theta})] + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right).$$

Therefore,

$$\left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{\Theta}) - \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{\Theta})] \right|$$

$$\begin{aligned}
&\leq \sum_{k=1}^K \left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{u}^{k,\theta}) - \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{u}^{k,\theta})] \right| && \text{(By the triangle inequality)} \\
&\lesssim \sum_{k=1}^K \frac{1}{2} \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{u}^{k,\theta})] + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) && \text{(By Lemma L.8)} \\
&= \frac{1}{2} \cdot \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K^{\text{trunc}}(\hat{\Theta})] + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right).
\end{aligned}$$

This implies

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{\Theta})] \leq 2 \cdot \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{\Theta}) + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right).$$

Finally, we conclude that

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_k^{\text{trunc}}(\hat{\Theta}) - \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{\Theta})] \leq \hat{\mathcal{R}}_k^{\text{trunc}}(\hat{\Theta}) + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right).$$

This completes the proof. $\square$

E.3 Main Proof of Theorem 4.2

We now present the main proof of Theorem 4.2.

Theorem E.1 (Theorem 4.2 Restated: High-Order Velocity Estimation with Transformer). Assume Assumption 4.1 and Assumption 4.2. Let $\hat{u}^{k,\theta} \in \mathcal{T}_R^{h,s,r}$ be the estimator of the k -th order velocity field u^k trained by minimizing the high-order empirical conditional flow matching loss (3.9). Let $\hat{\Theta}$ be the collection of parameters of $\hat{u}^{k,\theta}$ for $k \in [K]$. Suppose the k -th order velocity field $u^k(x, t)$ is L_k Lipschitz for all $k = 0, \dots, K-1$. Suppose we choose the transformers as in Theorem 4.1, then

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\hat{\Theta})] = O\left(n^{-\frac{1}{10d}} \cdot (\log n)^{10d_x}\right),$$

where d is the feature dimension.

Proof of Theorem 4.2. Let $\{x'_i\}_{i=1}^n$ be a different set of i.i.d samples independent of the training sample $\{x_i\}_{i=1}^n$. Further, we use $\hat{\mathcal{R}}'$ to denote the empirical risk with samples $\{x'_i\}_{i=1}^n$.

Then, we decompose $\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\hat{\Theta})]$ as:

$$\begin{aligned}
\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\hat{\Theta})] &= \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} [\hat{\mathcal{R}}'_K(\hat{\Theta}) - \hat{\mathcal{R}}_K^{\text{trunc}}(\hat{\Theta})] \right]}_{\text{(I)}} \\
&\quad + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} [\hat{\mathcal{R}}_K^{\text{trunc}}(\Theta)] - \hat{\mathcal{R}}_K^{\text{trunc}}(\hat{\Theta}) \right]}_{\text{(II)}} \\
&\quad + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\hat{\mathcal{R}}_K^{\text{trunc}}(\hat{\Theta}) - \hat{\mathcal{R}}_K(\hat{\Theta})]}_{\text{(III)}} + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\hat{\mathcal{R}}_K(\hat{\Theta})]}_{\text{(IV)}},
\end{aligned}$$

Then, we bound each term and incorporate them to obtain the bound on the estimation error.

- **Bound on (I) and III.** By [Lemma E.1](#), (I) and (III) are upper bounded by

$$(I), (III) \lesssim K D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \max_k \{L_k^2\}.$$

- **Bound on (II).** By the generalization error bound ([Lemma E.3](#)), we have

$$\begin{aligned} & (II) \\ &= \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K^{\text{trunc}}(\Theta)] - \widehat{\mathcal{R}}_K^{\text{trunc}}(\Theta)] \\ &= \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K^{\text{trunc}}(\widehat{\Theta}) - \widehat{\mathcal{R}}_K^{\text{trunc}}(\widehat{\Theta})] \quad (\text{By [Remark E.2](#)}) \\ &\leq \mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K^{\text{trunc}}(\widehat{\Theta})] + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) \quad (\text{By [Lemma E.3](#)}) \\ &\lesssim (IV) + D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \max_k \{L_k^2\} + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) \quad (\text{By [Lemma E.1](#)}) \end{aligned}$$

- **Bound on (IV).** Recall [Remark E.1](#), [Remark E.2](#). We have $\widehat{\mathcal{R}}_K(\Theta) := \widehat{\mathcal{L}}_{\text{CFM}}(\Theta) - \widehat{\mathcal{L}}_{\text{CFM}}(f_t)$, where the collection of parameters of K transformers $\widehat{\Theta}$ is trained by optimizing $\widehat{\mathcal{L}}_{\text{CFM}}(\Theta)$.

Therefore, it holds

$$\widehat{\mathcal{R}}_K(\widehat{\Theta}) \leq \widehat{\mathcal{L}}_{\text{CFM}}(\Theta) - \widehat{\mathcal{L}}_{\text{CFM}}(f_t) = \widehat{\mathcal{R}}_K(\Theta).$$

Then, for any velocity estimator Θ , it holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K(\widehat{\Theta})] \leq \mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}_K(\Theta)] = \mathcal{R}_K(\Theta). \quad (\text{E.4})$$

This implies

$$(IV) \leq \mathcal{R}_K(\Theta) \lesssim N^{-2\beta} \cdot (\log N)^{\frac{d_x}{2}-1}. \quad (\text{By [Theorem 4.1](#)})$$

Altogether, the estimation error is upper bounded by

$$\begin{aligned} & \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\widehat{\Theta})] \quad (\text{E.5}) \\ &= (I) + (II) + (III) + (IV) \\ &\lesssim \underbrace{D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right)}_{(T_1)} + \underbrace{O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right)}_{(T_2)} + \underbrace{N^{-2\beta} \cdot (\log N)^{\frac{d_x}{2}-1}}_{(T_3)}, \end{aligned}$$

where

$$\log \mathcal{N} = O\left(\frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^2 N^{\beta(16d+12)} (\log N)^{8d+8}\right). \quad (\text{By [Lemma E.2](#)})$$

Let $\gamma := 16d + 12$. Then, we set $N := n^{\eta_1/(\gamma\beta)}$, $\epsilon_c := n^{-\eta_2}$ and $D := \sqrt{(2\eta_3 \log n)/C_2}$, where $\eta_1, \eta_2, \eta_3 \geq 0$ are constants satisfying $0 \leq \eta_1 + 2\eta_2 < 1$.³ This gives

$$(T_1) = D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \lesssim n^{-\eta_3} (\log n)^{\frac{d_x}{2}}.$$

³The constraint $0 \leq \eta_1 + 2\eta_2 < 1$ is imposed in order to ensure (T_2) converges as $n \rightarrow \infty$.

Further, we have

$$\log \mathcal{N} = O(n^{\eta_1+2\eta_2} (\log n)^{8d_x+11}),$$

implying

$$(T_2) = O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) = O(n^{\eta_1+2\eta_2-1} (\log n)^{8d_x+11} + n^{-\eta_2}).$$

Further,

$$(T_3) = n^{-\frac{2\eta_1}{\gamma}} (\log n)^{\frac{d_x}{2}-1}.$$

Then, (E.5) becomes

$$\begin{aligned} & \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\Theta)] \\ & \lesssim (T_1) + (T_2) + (T_3) \\ & = O\left(n^{-\min\left\{1-(\eta_1+2\eta_2), \eta_2, \frac{2\eta_1}{\gamma}\right\}} \cdot (\log n)^{8d_x+11}\right). \end{aligned}$$

For any η_1 and η_2 satisfying

$$0 < \eta_1 + 2\eta_2 < 1,$$

we consider solving

$$\min \left\{ 1 - (\eta_1 + 2\eta_2), \eta_2, \frac{2\eta_1}{\gamma} \right\}.$$

The linear programming problem has simple solution

$$1 - (\eta_1 + 2\eta_2) = \eta_2 = \frac{2\eta_1}{\gamma}.$$

This gives

$$\eta_1 = \frac{\gamma}{\gamma+6}, \quad \text{and} \quad \eta_2 = \frac{2}{\gamma+6},$$

and $\eta_1 + 2\eta_2 \in (0, 1)$ is satisfied for any $\eta_1, \eta_2 > 0$.

Finally, by $\gamma = 16d + 12$, these free parameters achieves balance and gives

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\hat{\Theta})] \lesssim O\left(n^{-\frac{2}{\gamma+6}} \cdot (\log n)^{8d_x+11}\right) = O\left(n^{-\frac{1}{8d+9}} \cdot (\log n)^{8d_x+11}\right)$$

This completes the proof. □

F Proof of Theorem 4.3

We now present the main proof of Theorem 4.3.

Theorem F.1 (Theorem 4.3 Restated: High-Order Distribution Estimation under 2-Wasserstein Distance). Assume Assumption 4.1 and Assumption 4.2. Let $\hat{P}_T^K$ be the estimated distribution at time T . Then, it holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T^K, P_T^K)] = O\left(n^{-\frac{1}{18d}} \cdot (\log n)^{6d_x}\right),$$

where d is the feature dimension.

Proof of Theorem 4.3. We first consider two general ODE functions that describe the ground truth velocity field and estimated velocity field respectively:

$$\frac{d}{dt}y_t = \begin{bmatrix} u^1(x_t^{(0)}, t) \\ u^2(x_t^{(0)}, t) \\ \vdots \\ u^K(x_t^{(0)}, t) \end{bmatrix} := f(y, t), \quad \frac{d}{dt}y_t = \begin{bmatrix} u^{1,\theta}(x_t^{(0)}, t) \\ u^{2,\theta}(x_t^{(0)}, t) \\ \vdots \\ u^{K,\theta}(x_t^{(0)}, t) \end{bmatrix} := f^\theta(y, t),$$

where the first d rows of $y_t \in \mathbb{R}^{Kd}$ construct $x_t^{(0)} \in \mathbb{R}^d$.

According to the existence uniqueness theorem of ODEs, these two functions can induce another following two corresponding flows $\phi(\cdot) \in \mathbb{R}^{Kd}$ and $\phi^\theta(\cdot) \in \mathbb{R}^{Kd}$ defined for $t \geq s$ that satisfy:

$$\begin{aligned} \frac{d}{dt}\phi(y, s, t) &= f(\phi(y, s, t), t), & \phi(y, s, s) &= y, \\ \frac{d}{dt}\phi^\theta(y, s, t) &= f^\theta(\phi^\theta(y, s, t), t), & \phi^\theta(y, s, s) &= y. \end{aligned}$$

We define the first d rows of $\phi(\cdot)$ construct the flow function $\psi_t(x)$ and the first d rows of $\phi^\theta(\cdot)$ construct the flow function $\psi_t^\theta(x)$. By applying Lemma M.2, it holds that

$$\phi^\theta(y, t_0, T) - \phi(y, t_0, T) = \int_{t_0}^T D\phi^\theta(\phi(y, t_0, s), s, T)(f^\theta(\phi(y, t_0, s), s) - f(\phi(y, t_0, s), s))ds.$$

We extract the first d rows of left-hand-side and it holds:

$$\psi^\theta(x, t_0, T) - \psi(x, t_0, T) = \int_{t_0}^T D\phi^\theta(\phi(y, t_0, s), s, t)[:d](f^\theta(\phi(y, t_0, s), s) - f(\phi(y, t_0, s), s))ds,$$

where $D\phi^\theta(\phi(y, t_0, s), s, t)[:d]$ denotes the first d rows of the Jacobian matrix.

We then bound $\psi^\theta(x, t_0, T) - \psi(x, t_0, T)$ by using similar techniques in proof of Theorem I.4. It shows that

$$\begin{aligned} & \left\| \frac{\partial}{\partial t} D\phi^\theta(\phi(y, t_0, s), s, t)[:d] \right\|_2 \\ & \leq \left\| \frac{\partial}{\partial t} D\phi^\theta(\phi(y, t_0, s), s, t)[:d] \right\|_2 && \text{(By triangle inequality)} \\ & = \|Du^{1,\theta}(\psi^\theta(\psi(x, t_0, s), s, t)D\phi^\theta(\phi(y, t_0, s), s, t))[[:d]]\|_2 && \text{(By chain rule)} \\ & \leq L_{\mathcal{T}} \|D\phi^\theta(\phi(y, t_0, s), s, t)[[:d]]\|_2. && \text{(By Lipschitz constant of transformer)} \end{aligned}$$

Therefore,

$$\|D\phi^\theta(\phi(y, t_0, s), s, t)[\cdot d])\|_2 \lesssim \exp\left\{\int_s^t L_{\mathcal{T}} du\right\} \leq \exp\left\{\int_0^1 L_{\mathcal{T}} du\right\} =: M. \quad (\text{By Lemma M.1})$$

Now we have

$$\begin{aligned} & \|\psi^\theta(x, t_0, T) - \psi(x, t_0, T)\|_2^2 \\ & \leq M^2 \int_{t_0}^T (f^\theta(\phi(y, t_0, s), s) - f(\phi(y, t_0, s), s))^2 ds \quad (\text{By Lemma M.1}) \\ & = M^2 \left(\int_{t_0}^T \left(\sum_{k=1}^K \|u^{k,\theta}(\psi(x, t_0, s), s) - u^k(\psi(x, t_0, s), s)\|_2 \right)^2 ds \right) \quad (\text{By definition of } f^\theta) \\ & \leq M^2 \int_{t_0}^T \left(\sum_{k=1}^K \|u^{k,\theta}(\psi(x, t_0, s), s) - u^k(\psi(x, t_0, s), s)\|_2^2 \right) ds. \quad (\text{By Cauchy Schwarz inequality}) \end{aligned}$$

Then, we take expectation with $x \sim p_{t_0}^0$ on both sides

$$\begin{aligned} & \mathbb{E}_{x \sim p_{t_0}^0} [\|\psi^\theta(x, t_0, T) - \psi(x, t_0, T)\|_2^2] \\ & \leq M^2 \sum_{k=1}^K \mathbb{E}_{x \sim p_{t_0}^0} \left[\int_{t_0}^T \|u^{k,\theta}(\psi(x, t_0, s), s) - u^k(\psi(x, t_0, s), s)\|_2^2 ds \right] \\ & = M^2 (T - t_0) \mathcal{R}_K(\Theta). \quad (\text{By definition of higher order risk in Definition 4.2}) \end{aligned}$$

Finally, we bound the 2-Wasserstein distance between the estimated and true distributions following Appendix M. By using the definition of the 2-Wasserstein metric, it follows that

$$W_2(\hat{P}_T^K, P_T^K) \leq \sqrt{\mathbb{E}_{x \sim p_{t_0}^0} [\|\psi^\theta(x, t_0, T) - \psi(x, t_0, T)\|_2^2]} \lesssim \sqrt{\mathcal{R}_K(\Theta)}$$

Then,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T^K, P_T^K)] \lesssim \sqrt{\mathcal{R}_K(\hat{\Theta})}$$

We apply the high-order velocity estimation results in Theorem 4.2

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}_K(\hat{\Theta})] = O\left(n^{-\frac{1}{8d+9}} \cdot (\log n)^{8d_x+11}\right).$$

This implies

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T^K, P_T^K)] \lesssim \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{\mathcal{R}_K(\hat{\Theta})}] = O\left(n^{-\frac{1}{16d+18}} \cdot (\log n)^{4d_x+6}\right).$$

This completes the proof. □

G Proof of Theorem 4.4

Recall the Hölder density function class and its minimax optimal rate under 2-Wasserstein distance:

Lemma G.1 (Lemma N.2 Restated: Modified from Theorem 3 of [Niles-Weed and Berthet, 2022]). Consider the task of estimating a probability distribution $P(x_1)$ with density function belonging to the space

$$\mathcal{P} := \{q(x_1) \mid q(x_1) \in \mathcal{H}^\beta([-1, 1]^{d_x}, B), q(x_1) \geq C\},$$

Then, for any $r \geq 1$, $\beta > 0$ and $d_x > 2$, we have

$$\inf_{\hat{P}} \sup_{q(x_1) \in \mathcal{P}} \mathbb{E}_{\{x_i\}_{i=1}^n} [W_r(\hat{P}, P)] \gtrsim n^{-\frac{\beta+1}{d_x+2\beta}},$$

where $\{x_i\}_{i=1}^n$ is a set of i.i.d samples drawn from distribution P , and $\hat{P}$ runs over all possible estimators constructed from the data.

We now give the formal proof of Theorem 4.4.

Theorem G.1 (Theorem 4.4 Restated: Minimax Optimality of High-Order Flow Matching Transformers). Assume that the target density function satisfies $q(x_1) \in \mathcal{H}^\beta([-1, 1]^{d_x}, B)$ and $q(x_1) \geq C$ for some constant C . Then, under the setting of $18d(\beta + 1) = d_x + 2\beta$, the distribution estimation rate of flow matching transformers presented in Theorem 4.3 matches the minimax lower bound of Hölder distribution class in 2-Wasserstein distance up to a $\log n$ and Lipschitz constants factors.

Proof of Theorem 4.4. Since the bounded support $[-1, 1]^{d_x}$ guarantees the sub-Gaussian property in Assumption I.1, the distribution estimation Theorem 4.3 holds under $q(x_1) \in \mathcal{H}^\beta([-1, 1]^{d_x}, B)$:

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] \lesssim O\left(n^{-\frac{1}{18d}} \cdot (\log n)^{6d_x}\right).$$

Then, by Lemma N.2, the distribution rates matches the minimax lower bound up to a $\log n$ and Lipschitz constant factors under the setting

$$18d(\beta + 1) = d_x + 2\beta.$$

This completes the proof. □

Remark G.1. Since $d_x = d \cdot L$, the condition $18d(\beta + 1) = d_x + 2\beta$ implies

$$d(18\beta + 18 - L) = 2\beta \quad \text{and} \quad \beta(18d - 2) = d(L - 18),$$

the transformers achieve minimax optimal rate with reshape layer such that $18 \leq L \leq 18\beta + 18$.

H Preliminaries: Universal Approximation of Transformers

Prior works [Hu et al., 2025b, 2024, Kajitsuka and Sato, 2023, Yun et al., 2019] develop the universal approximation of transformers for continuous functions. Here we revisit these methods to establish a foundation for our analysis of k -th order flow matching transformers. Specifically, we revisit (i) the ability of the transformer function class (defined in Appendix B) to approximate any continuous function on a compact domain with arbitrary error, (ii) the parameter norm bounds required to achieve the universal approximation. Notably, controlling the magnitude of these norm bounds is essential for subsequent analysis on the velocity estimation error and distribution error.

Background: Contextual Mapping. Recall the reshape layer Definition B.3. Let $Z \in \mathbb{R}^{d \times L}$ represent input embeddings, where $Z_{:,k} \in \mathbb{R}^d$ denotes the k -th token (column) of each Z sequence. Further, given M embeddings $Z^{(1)}, \dots, Z^{(M)} \in \mathbb{R}^{d \times L}$, we say $Z^{(i)}$ is the i -th sequence for $i \in [M]$. Then, we define the *vocabulary* corresponding to the i -th sequence at the k -th index in Definition H.1.

Definition H.1 (Vocabulary). We define the i -th vocabulary set for $i \in [M]$ by $\mathcal{V}^{(i)} = \bigcup_{k \in [L]} Z_{:,k}^{(i)} \subset \mathbb{R}^d$, and the whole vocabulary set $\mathcal{V}$ is defined by $\mathcal{V} = \bigcup_{i \in [M]} \mathcal{V}^{(i)} \subset \mathbb{R}^d$.

In line with prior works [Hu et al., 2025b, 2024, Kajitsuka and Sato, 2023, Kim et al., 2022, Yun et al., 2019], we assume the embeddings separateness to be $(\gamma_{\min}, \gamma_{\max}, \delta)$ -separated,

Definition H.2 (Tokenwise Separateness). Let $Z^{(1)}, \dots, Z^{(M)} \in \mathbb{R}^{d \times L}$ be embeddings. Then, we say $Z^{(1)}, \dots, Z^{(M)}$ are tokenwise $(\gamma_{\min}, \gamma_{\max}, \delta)$ -separated if the following three conditions hold.

1. For any $i \in [M]$ and $k \in [n]$, $\|Z_{:,k}^{(i)}\| > \gamma_{\min}$ holds.
2. For any $i \in [M]$ and $k \in [n]$, $\|Z_{:,k}^{(i)}\| < \gamma_{\max}$ holds.
3. For any $i, j \in [M]$ and $k, l \in [n]$ if $Z_{:,k}^{(i)} \neq Z_{:,l}^{(j)}$, then $\|Z_{:,k}^{(i)} - Z_{:,l}^{(j)}\| > \delta$ holds.

Further, we say $Z^{(1)}, \dots, Z^{(N)}$ is (γ, δ) -separated when only conditions (ii) and (iii) hold. Also, if only condition (iii) holds, we denote it as (δ) -separateness.

Building on the token separateness, we introduce the *contextual mapping*, that characterizes the ability of transformers' self-attention to capture the relationships among tokens across different sequences. This allows transformers to utilize self-attention for full context representation.

Definition H.3 (Contextual Mapping). Let $Z^{(1)}, \dots, Z^{(M)} \in \mathbb{R}^{d \times L}$ be embeddings. Then, we say a map $T : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d \times L}$ is a (γ, δ) -contextual mapping if the following two conditions hold:

1. For any $i \in [M]$ and $k \in [L]$, it holds

$$\|T(Z^{(i)})_{:,k}\| < \gamma.$$

2. For any $i, j \in [M]$ and $k, l \in [L]$ such that $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$ or $Z_{:,k}^{(i)} \neq Z_{:,l}^{(j)}$, it holds

$$\|T(Z^{(i)})_{:,k} - T(Z^{(j)})_{:,l}\| > \delta.$$

We introduce results from [Hu et al., 2025b] in Theorem H.1, which shows that a one-layer single-head attention mechanism is a contextual mapping.

Helper Lemmas. Before presenting Theorem H.1, we restate several helper lemmas from [Hu et al., 2025b, Kajitsuka and Sato, 2023] to simplify the proof.

Lemma H.1 (Boltz Preserves Distance, Lemma 1 of [Kajitsuka and Sato, 2023]). Given (γ, δ) -tokenwise separated vectors $z^{(1)}, \dots, z^{(N)} \in \mathbb{R}^n$ with no duplicate entries in each vector:

$$z_s^{(i)} \neq z_t^{(i)},$$

where $i \in [N]$ and $s, t \in [L], s \neq t$. Also, let

$$\delta \geq 4 \ln n.$$

Then, the outputs of the Boltzmann operator has the following properties:

$$\begin{aligned} \left| \text{Boltz} \left(z^{(i)} \right) \right| &\leq \gamma, \\ \left| \text{Boltz} \left(z^{(i)} \right) - \text{Boltz} \left(z^{(j)} \right) \right| &> \delta' = \ln^2(n) \cdot e^{-2\gamma} \end{aligned}$$

for all $i, j \in [N], i \neq j$.

We remark that [Lemma H.1](#) gives the key step for constructing a contextual mapping via the self-attention layer [Lemma H.4](#). We extend the results to the case with inverse-temperature scaling.

Lemma H.2 (Boltz Preserves Distance with Finite Inverse-Temperature Scaling in Softmax). Let $z^{(1)}, \dots, z^{(M)} \in \mathbb{R}^n$ be (γ, δ) -tokenwise separated vectors with no duplicate entries in each vector:

$$z_s^{(i)} \neq z_t^{(i)},$$

where $n \geq 2, \delta \geq 4 \ln n / \lambda$ for some $\lambda > 0, i \in [M]$ and $s, t \in [L], s \neq t$. Further, we define

$$\text{Boltz}_\lambda(z^{(i)}) := z^{(i)\top} \text{Softmax}_\lambda(z^{(i)}) \quad \text{for all } i \in [M],$$

where $\text{Softmax}_\lambda(z^{(i)})[k] = \exp(\lambda z^{(i)}[k]) / \sum_j \exp(\lambda z^{(i)}[j])$ is the softmax function with inverse-temperature scaling $\lambda > 0$. Then, the outputs of the Boltzmann operator has the following properties:

$$\left| \text{Boltz}_\lambda \left(z^{(i)} \right) \right| \leq \gamma, \tag{H.1}$$

$$\left| \text{Boltz}_\lambda \left(z^{(i)} \right) - \text{Boltz}_\lambda \left(z^{(j)} \right) \right| > \delta' = \frac{\ln^2(n) \cdot e^{-2\lambda\gamma}}{\lambda} \tag{H.2}$$

for all $i, j \in [M], i \neq j$.

Proof. We first scale the input vectors by $y^{(i)}[j] := z^{(i)}[j] \cdot \lambda$ for all $i \in [M]$ and all $j \in [n]$. This gives a (R, Δ) -tokenwise separated vectors $y^{(1)}, \dots, y^{(M)} \in \mathbb{R}^n$, where $R = \lambda\gamma$ and $\Delta = \lambda\delta$.

Next, notice that $\text{softmax}_\lambda(\cdot) = \text{softmax}(\lambda \cdot)$ and $\text{Boltz}_\lambda(\cdot) = \text{Boltz}(\lambda \cdot) / \lambda$. Therefore, since the condition $\delta \geq 4 \ln n / \lambda$ ensures $\Delta \geq 4 \ln n$, [Lemma H.1](#) implies both [\(H.1\)](#) and [\(H.2\)](#).

This completes the proof. $\square$

Lemma H.3 (Lemma 13 of [\[Park et al., 2021\]](#)). For any finite subset $\mathcal{X} \subset \mathbb{R}^d$, there exists at least one unit vector $u \in \mathbb{R}^d$ such that

$$\frac{1}{|\mathcal{X}|^2} \sqrt{\frac{8}{\pi d}} \|x - x'\| \leq |u^\top (x - x')| \leq \|x - x'\|,$$

for any $x, x' \in \mathcal{X}$.

[Lemma H.3](#) shows the existence of a unit vector $u \in \mathbb{R}^d$ that bounds the inner product of the difference between points in a finite subset $\mathcal{X} \subset \mathbb{R}^d$. Next, we restate the construction of rank- ρ weight matrices in a self-attention layer following [\[Hu et al., 2025b\]](#) in [Lemma H.4](#).

Lemma H.4 (Construction of Weight Matrices, Lemma D.2 of [\[Hu et al., 2025b\]](#)). Given $(\gamma_{\min}, \gamma_{\max}, \epsilon)$ -separated input embeddings $Z^{(1)}, \dots, Z^{(M)} \in \mathbb{R}^{d \times L}$ with finite vocabulary set

$\mathcal{V} \subset \mathbb{R}^d$. There exists rank- ρ weight matrices $W_K, W_Q \in \mathbb{R}^{s \times d}$ such that

$$\left| (W_K v_a)^\top (W_Q v_c) - (W_K v_b)^\top (W_Q v_c) \right| > \delta,$$

for any $\delta > 0$, any $\min(d, s) \geq \rho \geq 1$ and any $v_a, v_b, v_c \in \mathcal{V}$ with $v_a \neq v_b$. In addition, the matrices are constructed as

$$W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s \times d}, \quad W_Q = \sum_{j=1}^{\rho} p'_j q'_j{}^\top \in \mathbb{R}^{s \times d},$$

where $q_i, q'_i \in \mathbb{R}^d$ are unit vectors that satisfy [Lemma H.3](#) for at least one i , and $p_i, p'_i \in \mathbb{R}^s$ satisfies

$$|p_i^\top p'_i| = 5(|\mathcal{V}| + 1)^4 d \frac{\delta}{\epsilon \gamma_{\min}}.$$

Any-Rank Attention is Contextual Mapping. The next lemma shows that the any rank self-attention mechanisms of transformers serve as contextual mappings ([Definition H.3](#)).

Theorem H.1 (Any-Rank Attention is (γ, δ) -Contextual Mapping, Lemma 2.2 of [[Hu et al., 2025b](#)]). Consider $(\gamma_{\min}, \gamma_{\max}, \epsilon)$ -tokenwise separated embeddings $Z^{(1)}, \dots, Z^{(M)} \in \mathbb{R}^{d \times L}$ and vocabulary set $\mathcal{V} = \bigcup_{i \in [N]} \mathcal{V}^{(i)} \subset \mathbb{R}^d$. Let $Z^{(1)}, \dots, Z^{(N)} \in \mathbb{R}^{d \times L}$ be embedding sequences with no duplicate word token in each sequence; that is, $Z_{:,k}^{(i)} \neq Z_{:,l}^{(i)}$, for any $i \in [M]$ and $k, l \in [L]$. Then, there exists a 1-layer single head attention with weight matrices $W_O \in \mathbb{R}^{d \times s}$ and $W_V, W_K, W_Q \in \mathbb{R}^{s \times d}$, that is a (γ, δ) -contextual mapping for embeddings $Z^{(1)}, \dots, Z^{(M)}$ with

$$\gamma = \gamma_{\max} + \epsilon/4, \quad \delta = \exp(-5\lambda\epsilon^{-1}|\mathcal{V}|^4 d \kappa \gamma_{\max} \log L) / \lambda,$$

where $\kappa = \gamma_{\max}/\gamma_{\min}$ and λ is the inverse-temperature scaling in the column-wise Softmax function.

We restate the proof of [Theorem H.1](#) since it is crucial for subsequent analysis.

Proof. For completeness, we restate the proof from Lemma 2.2 of [[Hu et al., 2025c](#)].

The proof consists of two steps:

- **Construct the Softmax Attention.** We ensure that different input tokens are mapped to unique contextual embeddings by configuring the weight matrices in [Lemma H.4](#).
- **Handle Identical Tokens in Different Contexts.** We show that the construction from [Lemma H.4](#) are able to handle identical tokens in different contexts by applying [Lemma H.2](#).

We proceed the proof with these two steps.

Step 1: Attention Construction. We show the construction of matrices: W_K, W_Q, W_O and W_V .

- **Weight Matrices W_K and W_Q .** First, we construct W_K and W_Q by:

$$W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s \times d}, \quad W_Q = \sum_{j=1}^{\rho} p'_j q'_j{}^\top \in \mathbb{R}^{s \times d},$$

where $p_i, p'_j \in \mathbb{R}^s$ and $q_i, q'_j \in \mathbb{R}^d$. In addition, let $\delta = 4 \ln n$ and $p_1, p'_1 \in \mathbb{R}^s$ be an arbitrary vector pair that satisfies

$$|p_1^\top p'_1| = (|\mathcal{V}| + 1)^4 d \frac{\delta}{\epsilon \gamma_{\min}}. \quad (\text{H.3})$$

• **Weight Matrices W_V and W_O .** Next, we construct $W_O \in \mathbb{R}^{d \times s}$ and $W_V \in \mathbb{R}^{s \times d}$ by:

$$W_V = \sum_{i=1}^{\rho} p_i'' q_i''^\top \in \mathbb{R}^{s \times d}, \quad (\text{H.4})$$

where $q'' \in \mathbb{R}^d$, $q_1'' = q_1$ and $p_i'' \in \mathbb{R}^s$ is some nonzero vector that satisfies

$$\|W_O p_i''\| = \frac{\epsilon}{4\rho\gamma_{\max}}. \quad (\text{H.5})$$

This can be accomplished, e.g., $W_O = \sum_{i=1}^{\rho} p_i''' p_i'''^\top$ for any vector p_i''' which satisfies $\|p_i'''\| = \epsilon/(4\rho^2\gamma_{\max}\|p_i''\|^2)$ for any $i \in [\rho]$.

For simplicity, we define $s_{k'}^k := \text{Softmax} \left[\left(W_K Z^{(i)} \right)^\top \left(W_Q Z_{:,k}^{(i)} \right) \right]_{k'}$.

Then, we combine the above weights construction and obtain

$$\begin{aligned} & \|W_O \left(W_V Z^{(i)} \right) \text{Softmax} \left[\left(W_K Z^{(i)} \right)^\top \left(W_Q Z_{:,k}^{(i)} \right) \right]\| \\ &= \left\| \sum_{k'=1}^L s_{k'}^k W_O \left(W_V Z^{(i)} \right) \right\| && (\text{By the definition of } s_{k'}^k) \\ &\leq \sum_{k'=1}^L \|s_{k'}^k W_O \left(W_V Z^{(i)} \right)\| && (\text{By triangle inequality}) \\ &\leq \max_{k' \in [L]} \|W_O \left(W_V Z^{(i)} \right)\| && (\text{By } \sum_{k'=1}^L s_{k'}^k = 1) \\ &\leq \frac{\epsilon}{4\gamma_{\max}} \cdot \max_{k' \in [L]} \left| q^\top Z_{:,k'}^{(i)} \right| && (\text{By (H.4) and (H.5)}) \\ &\leq \frac{\epsilon}{4\gamma_{\max}} \cdot \max_{k' \in [L]} \left| Z_{:,k'}^{(i)} \right| && (\text{By Lemma H.3}) \\ &\leq \frac{\epsilon}{4}. && (\text{By the } (\gamma_{\min}, \gamma_{\max}, \epsilon) \text{ separateness}) \end{aligned} \quad (\text{H.6})$$

for $i \in [M]$ and $k \in [L]$.

Step 2: The Case of Identical Tokens in Different Contexts. For the second part, we show that with the constructed weight matrices W_O, W_V, W_K, W_Q , the attention layer distinguishes duplicate input tokens with different context, $Z_{:,k}^{(i)} = Z_{:,l}^{(j)}$ with different vocabulary sets $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$.

We define $a^{(i)}, a^{(j)}$ as

$$a^{(i)} = \left(W_K Z^{(i)} \right)^\top \left(W_Q Z_{:,k}^{(i)} \right) \in \mathbb{R}^n, \quad a^{(j)} = \left(W_K Z^{(j)} \right)^\top \left(W_Q Z_{:,l}^{(j)} \right) \in \mathbb{R}^n,$$

where $a^{(i)}$ and $a^{(j)}$ are tokenwise (γ, δ) -separated. Specifically, the following inequality holds

$$|a_{k'}^{(i)}| \leq (|\mathcal{V}| + 1)^4 d \frac{\delta}{\epsilon \gamma_{\min}} \gamma_{\max}^2.$$

Since $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$ and there is no duplicate token in $Z^{(i)}$ and $Z^{(j)}$, we use Lemma H.1 and obtain

$$\begin{aligned} & \left| \text{Boltz} \left(a^{(i)} \right) - \text{Boltz} \left(a^{(j)} \right) \right| \\ &= \left| \left(a^{(i)} \right)^\top \text{Softmax} \left[a^{(i)} \right] - \left(a^{(j)} \right)^\top \text{Softmax} \left[a^{(j)} \right] \right| \end{aligned} \quad (\text{H.7})$$

$$\begin{aligned} &> \delta' \\ &= (\ln n)^2 e^{-2\gamma}. \end{aligned}$$

Additionally, using [Lemma H.4](#) and [\(H.3\)](#), and assuming $Z_{:,k}^{(i)} = Z_{:,l}^{(j)}$, we have

$$\begin{aligned} &\left| \left(a^{(i)} \right)^\top \text{Softmax} \left[a^{(i)} \right] - \left(a^{(j)} \right)^\top \text{Softmax} \left[a^{(j)} \right] \right| \\ &\leq \sum_{i=1}^{\rho} \gamma_{\max} \cdot (|\mathcal{V}| + 1)^4 \frac{\pi d}{8} \frac{\delta}{\epsilon \gamma_{\min}} \cdot \left| \left(q_i^\top Z^{(i)} \right) \text{Softmax} \left[a^{(i)} \right] - \left(q_i^\top Z^{(j)} \right) \text{Softmax} \left[a^{(j)} \right] \right|. \end{aligned} \quad (\text{H.8})$$

By combining [\(H.7\)](#) and [\(H.8\)](#), we have

$$\sum_{i=1}^{\rho} \left| \left(q_i^\top Z^{(i)} \right) \text{Softmax} \left[a^{(i)} \right] - \left(q_i^\top Z^{(j)} \right) \text{Softmax} \left[a^{(j)} \right] \right| > \frac{\delta'}{(|\mathcal{V}| + 1)^4} \frac{\epsilon \gamma_{\min}}{d \delta \gamma_{\max}}. \quad (\text{H.9})$$

Using [\(H.5\)](#) and [\(H.9\)](#), we derive the lower bound of the difference between the self-attention outputs of $Z^{(i)}, Z^{(j)}$ as follows:

$$\left\| \mathcal{F}_S^{(\text{SA})} \left(Z^{(i)} \right)_{:,k} - \mathcal{F}_S^{(\text{SA})} \left(Z^{(j)} \right)_{:,l} \right\| > \frac{\epsilon}{4 \gamma_{\max}} \frac{\delta'}{(|\mathcal{V}| + 1)^4} \frac{\epsilon \gamma_{\min}}{d \delta \gamma_{\max}},$$

where $\delta = 4 \ln L$ and $\delta' = \ln^2(L) e^{-2\gamma}$ with $\gamma = (|\mathcal{V}| + 1)^4 d \delta \gamma_{\max}^2 / (\epsilon \gamma_{\min})$. Finally, we extend the bound to softmax function with inverse-temperature scaling $\lambda > 0$ using [Lemma H.2](#).

This completes the proof. $\square$

Notably, [Theorem H.1](#) shows that, for identical embeddings $Z_{:,k}^{(i)} = Z_{:,l}^{(j)}$ with distinct vocabularies $\mathcal{V}^{(i)} \neq \mathcal{V}^{(j)}$, any-rank self-attention is able to distinguish two identical tokens in distinct contexts.

Universal Approximation of Transformer. We introduce the universal approximation result for transformers with a single self-attention layer from [\[Hu et al., 2025b, Kajitsuka and Sato, 2023\]](#).

Theorem H.2 (Transformer Universal Approximation, Theorem B.1 of [\[Hu et al., 2025b\]](#) and Proposition 1 of [\[Kajitsuka and Sato, 2023\]](#)). Let $\epsilon \in (0, 1)$ and $p \in [1, \infty)$. Let $\mathcal{F}_1^{(\text{FF})}, \mathcal{F}_2^{(\text{FF})}$ be two feed-forward layers and $\mathcal{F}^{(\text{SA})}$ be a single-head self-attention layer with softmax function defined in [\(B.1\)](#) and [\(B.2\)](#). Then, for any permutation equivariant continuous function f on a compact domain and any ϵ , there exists a $g(Z) = \mathcal{F}_2^{(\text{FF})} \circ \mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{(\text{FF})}(Z) \in \mathcal{T}_R^{h,s,r}$ such that $d_p(f(Z), g(Z)) < \epsilon$, where $d_p := (\int \|f(Z) - g(Z)\|_p^p dZ)^{1/p}$ and $\|\cdot\|_p$ is the element-wise ℓ_p -norm.

Proof. Since the universal approximation of transformer over any bounded domain differs only by scaling and shifting the transformer's parameters in $\mathcal{F}_1^{(\text{FF})}$ and $\mathcal{F}_2^{(\text{FF})}$, [Hu et al. \[2025b\]](#), [Kajitsuka and Sato \[2023\]](#) prove [Theorem H.2](#) assuming that the target function f is normalized on domain $[0, 1]^{d \times L}$ for simplicity. To support subsequent derivations of transformer parameter bounds required for ϵ -precision ([Lemma H.5](#)), we provide the proof on a more general bounded domains.

The proof consists of three steps: (i) Quantization by the First Feed-Forward Layer (ii) Contextual Mapping by the Self-Attention Layer (iii) Memorization by the Second Feed-Forward Layer.

Let $\Omega := [-I, I]^{d \times L}$ be the domain of f . Without loss of generality, we consider $I \in \mathbb{N}$.

First Step Quantization. First, we define a grid $\mathbb{G}_D$:

$$\mathbb{G}_D := \left\{ C \in \Omega \mid C_{t,k} = -I + \frac{s_{t,k}}{D}, s_{t,k} = 1, \dots, 2ID \text{ for all } t \in [d], k \in [L] \right\}, \quad (\text{H.10})$$

where $D \in \mathbb{N}$ is the granularity of $\mathbb{G}_D$.

Then, for $Z \in \Omega$, we construct a piece-wise constant function approximator:

$$g_1(Z) := \sum_{C \in \mathbb{G}_D} f(C) \mathbb{1}\{Z \in C + [-1/D, 0)^{d \times L}\} \quad (\text{H.11})$$

By the uniform continuity of f , for any $\epsilon > 0$, there exist a D such that

$$d_p(f(Z), g_1(Z)) < \epsilon/3. \quad (\text{H.12})$$

We use a transformer to approximate $g_1(Z)$ using two feed forward layers and a self-attention layer. First, we introduce the *quantization function* that discretizes the input into $\mathbb{G}_D$:

Quantization Function. We define the quantization function $\text{quant}_D : \mathbb{R} \rightarrow \mathbb{R}$:

$$\text{quant}_D(z) := \begin{cases} -I & z < -I \\ -I + 1/D & -I \leq z < -I + 1/D \\ \vdots & \vdots \\ I & I - 1/D \leq z. \end{cases}$$

Within Ω , $\text{quant}_D^{d \times L}(z)$ outputs regions identical to $C + [-1/D, 0)^{d \times L}$ defined by the indicator function in (H.11). We extend the quantization to $\text{quant}_D^{d \times L}(Z) : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d \times L}$, where $\text{quant}_D(z)$ is applied coordinate-wise. Then, we approximate $\text{quant}_D(z)$ with the following network⁴:

$$f_1(z) := -I + \sum_{t=-ID}^{ID-1} \frac{\text{ReLU}[z/\delta - t/\delta D] - \text{ReLU}[z/\delta - 1 - t/\delta D]}{D} \approx \text{quant}_D(z), \quad (\text{H.13})$$

where $\delta > 0$ determines the steepness of the change from one quantized level to the next. For $z \in \mathbb{R} \setminus [-I, I]$, we add and subtract the first and last step functions scaled by I to obtain zero output:

$$\begin{aligned} f_1^{\text{FF}}(z) &:= f_1(z) - I \cdot \left(\text{ReLU}[z/\delta - I/\delta] - \text{ReLU}[z/\delta - 1 - I/\delta] \right) \\ &\quad + I \cdot \left(\text{ReLU}[-z/\delta - I/\delta] - \text{ReLU}[-z/\delta - 1 - I/\delta] \right), \end{aligned} \quad (\text{H.14})$$

where $f_1^{\text{FF}}(z)$ approximately quantizes $[-I, I]$ into $\{-I + 1/D, \dots, I\}$ and project $\mathbb{R} \setminus [-I, I]$ to 0 by taking sufficiently small δ . Since every element shares identical weights, we realize (H.14) using $W_2' \text{ReLU}[W_1'Z + b_1'] + b_2'$ by constructing $W_1' \in \mathbb{R}^{L \cdot (4ID+4) \times d}$ and $b_1' \in \mathbb{R}^{L \cdot (4ID+4)}$ as:

$$W_1' := \begin{pmatrix} W_1^{(1)} \\ W_1^{(2)} \\ \vdots \\ W_1^{(L)} \end{pmatrix}, \quad b_1' := \begin{pmatrix} b_1^{(1)} \\ b_1^{(2)} \\ \vdots \\ b_1^{(L)} \end{pmatrix}, \quad \text{where} \quad b_1^{(i)} = \begin{pmatrix} \frac{I}{\delta} \\ \frac{I}{\delta} - 1 \\ \vdots \\ -\frac{I}{\delta} + \frac{1}{D} \\ -\frac{I}{\delta} - 1 + \frac{1}{D} \\ -\frac{I}{\delta} \\ -\frac{I}{\delta} - 1 \\ -\frac{I}{\delta} \\ -\frac{I}{\delta} - 1 \end{pmatrix} \quad \text{for all } i \in [L], \quad (\text{H.15})$$

⁴This is by the shifting and stacking step functions from [Yun et al., 2019].

and $W_1^{(1)}, \dots, W_1^{(L)} \in \mathbb{R}^{(4ID+4) \times d}$ take the form:

$$W_1^{(1)} = \begin{pmatrix} 1/\delta & 0 & 0 & \cdots & 0 \\ 1/\delta & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/\delta & 0 & 0 & \cdots & 0 \\ 1/\delta & 0 & 0 & \cdots & 0 \\ 1/\delta & 0 & 0 & \cdots & 0 \\ 1/\delta & 0 & 0 & \cdots & 0 \\ -1/\delta & 0 & 0 & \cdots & 0 \\ -1/\delta & 0 & 0 & \cdots & 0 \end{pmatrix}, W_1^{(2)} = \begin{pmatrix} 0 & 1/\delta & 0 & \cdots & 0 \\ 0 & 1/\delta & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1/\delta & 0 & \cdots & 0 \\ 0 & 1/\delta & 0 & \cdots & 0 \\ 0 & 1/\delta & 0 & \cdots & 0 \\ 0 & 1/\delta & 0 & \cdots & 0 \\ 0 & -1/\delta & 0 & \cdots & 0 \\ 0 & -1/\delta & 0 & \cdots & 0 \end{pmatrix}, \dots, W_1^{(L)} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 1/\delta \\ 0 & 0 & 0 & \cdots & 1/\delta \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1/\delta \\ 0 & 0 & 0 & \cdots & 1/\delta \\ 0 & 0 & 0 & \cdots & 1/\delta \\ 0 & 0 & 0 & \cdots & 1/\delta \\ 0 & 0 & 0 & \cdots & -1/\delta \\ 0 & 0 & 0 & \cdots & -1/\delta \end{pmatrix}.$$

The first $4ID$ rows of W_1' and b_1' implement the ReLU terms in (H.13), and the last four rows implement the ReLU terms in (H.14). Then, we construct $W_2' \in \mathbb{R}^{d \times L \cdot (4ID+4)}$ and $b_2 \in \mathbb{R}^d$ by

$$W_2' = \begin{pmatrix} 1/D & -1/D & \cdots & 1/D & -1/D & -I & I & I & -I \\ 1/D & -1/D & \cdots & 1/D & -1/D & -I & I & I & -I \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1/D & -1/D & \cdots & 1/D & -1/D & -I & I & I & -I \end{pmatrix}, \quad b_2' = \begin{pmatrix} -I \\ -I \\ \vdots \\ -I \end{pmatrix}. \quad (\text{H.16})$$

Besides the quantization, we include an additional penalty function to signify the case where the inputs are not on the target domain $[-I, I]^{d \times L}$.

Penalty Function. We define the penalty function $\text{penalty} : \mathbb{R} \rightarrow \mathbb{R}$

$$\text{penalty}(z) = \begin{cases} -2I & z < -I \\ 0 & z \in [-I, I] \\ -2I & z > I. \end{cases} \quad (\text{H.17})$$

Again, we extend the penalty function to $\text{penalty}^{d \times L}(z) : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d \times L}$ where $\text{penalty}(z)$ is applied coordinate-wise. By taking sufficiently small δ , we approximate $\text{penalty}(z)$ by

$$f_2^{\text{FF}}(z) \approx \text{penalty}(z),$$

where

$$f_2^{\text{FF}} := \underbrace{-2I \left(\text{ReLU}[(z - I)/\delta] + \text{ReLU}[(z - I)/\delta - 1] \right)}_{\text{Approximate step from 0 to } -2I \text{ at } z = I} \underbrace{-2I \left(\text{ReLU}[(-z - I)/\delta] + \text{ReLU}[(-z - I)/\delta - 1] \right)}_{\text{Approximate step from 0 to } -2I \text{ at } z = -I}. \quad (\text{H.18})$$

Let $\mathcal{I} \in \mathbb{R}^{d \times L}$ be a matrix with all entries equal to I . Altogether, we define $g_2(Z) : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d \times L}$

$$g_2(Z) = \underbrace{\text{quant}_D^{d \times L}(Z)}_{(\text{A})} + \underbrace{dL \cdot \text{penalty}^{d \times L}(Z)}_{(\text{B})}. \quad (\text{H.19})$$

Term (A) first quantizes input $[-I, I]^{d \times L}$ into $\mathbb{G}_D$ and scales the grid to $[0, 2I]^{d \times L}$, denoted by $\mathbb{G}_D^\circ$. Term (B) ensures non-positive outputs for any $Z \in \mathbb{R}^{d \times L} \setminus [-I, I]^{d \times L}$.

Together, we approximate (H.19) using the first feed-forward block $W_2 \text{ReLU}[W_1 Z + b_1] + b_2$ with $W_1 \in \mathbb{R}^{L \cdot (4ID+8) \times d}$, $b_1 \in \mathbb{R}^{L \cdot (4ID+8)}$, $W_2 \in \mathbb{R}^{d \times L \cdot (4ID+8)}$ and $b_2 \in \mathbb{R}^d$, where we stack $4L$ rows to W_1' , b_1' to implement the penalty function, scale weights in W_2' , and set b_2 as a zero vector.

Second Step Contextual Mapping. Let $\tilde{\mathbb{G}}_D \subseteq \mathbb{G}_D^\circ$ denote the sub-grid on $[0, 2I]^{d \times L}$:

$$\tilde{\mathbb{G}}_D := \left\{ G \in \mathbb{G}_D^\circ \mid G_{:,k} \neq G_{:,l} \text{ for all } k, l \in [L] \right\}. \quad (\text{H.20})$$

By the construction of $\mathbb{G}_D^\circ$, the sub-grid $\tilde{\mathbb{G}}_D$ is a collection of grids with pairwise distinct tokens, and every $G \in \tilde{\mathbb{G}}_D$ is a token-wise $(1/D, 2I\sqrt{d}, 1/D)$ -separated sequences.

From the construction of $\mathcal{F}^{(\text{SA})}$ in (H.6), it holds:

$$\|\mathcal{F}^{(\text{SA})}(Z)_{:,k} - Z_{:,k}\| < \frac{1}{8I\sqrt{d}D} \max_{k' \in [L]} \|Z_{:,k'}\|,$$

Recall that the magnitude of every entry in $\mathcal{F}_1^{\text{FF}}(Z)$ is at most $2IdL$ by the construction of the penalty function. Therefore, for all $k' \in [L]$, it holds:

$$\max_{Z \in \mathbb{R}^{d \times L}} \|\mathcal{F}_1^{\text{FF}}(Z)_{:,k'}\| \leq 2IdL \times \sqrt{d}.$$

This implies

$$\|\mathcal{F}^{(\text{SA})}(Z)_{:,k} \circ \mathcal{F}_1^{\text{FF}}(Z)_{:,k} - \mathcal{F}_1^{\text{FF}}(Z)_{:,k}\| < \frac{dL}{4D}$$

Taking sufficiently large D , every element of output of $Z \in \mathbb{R}^{d \times L} \setminus [-I, I]^{d \times L}$ is upper-bounded by:

$$\mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{\text{FF}}(Z)_{t,k} < \frac{1}{4D} \quad \text{for all } t \in [d], k \in [L].$$

Also, for input on $Z \in [0, 1]^{d \times L}$, we have lower bound:

$$\mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{\text{FF}}(Z)_{t,k} > \frac{3}{4D} \quad \text{for all } t \in [d], k \in [L].$$

Then, it remains to map sequences on $(3/4D, \infty)^{d \times L}$ to their corresponding target value, and map sequences on $(-\infty, 1/4D)^{d \times L}$ to zero.

Third Step Memorization. For a fixed $\bar{C} \in \tilde{\mathbb{G}}_D$ and a network input Z , we define $u := \mathcal{F}^{(\text{SA})}(\bar{C})$ and $S := \mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{\text{FF}}(Z)$. The goal is to use $\mathcal{F}_2^{\text{FF}}$ to implement $g_3 : \mathbb{R}^{d \times L} \rightarrow \mathbb{R}^{d \times L}$, where

$$g_3(S)_{t,k} = f(\bar{C} - \mathcal{I})_{t,k} \cdot \mathbb{1}[u_{t,k} = S_{t,k}] \quad \text{for all } t \in [d], k \in [L],$$

and f is the target function. We use the bump function with three-piece ReLU to achieve this:

$$\begin{aligned} & \text{bump}_R(S)_{t,k} \\ &= \text{ReLU}[R_{\text{FF}}(S_{t,k} - u_{t,k}) - 1] - 2\text{ReLU}[R_{\text{FF}}(S_{t,k} - u_{t,k})] + \text{ReLU}[R_{\text{FF}}(S_{t,k} - u_{t,k}) + 1], \end{aligned} \quad (\text{H.21})$$

where inputs with a deviation from the correct grid point $u_{t,k}$ greater than $1/R_{\text{FF}}$ are mapped to zero. Therefore, by taking sufficiently large R_{FF} , (H.21) implements $\mathbb{1}[u_{t,k} = S_{t,k}]$ exactly.

Suppose that there exist a feed-forward block $\mathcal{F}_2^{\text{FF}}$ such that (H.21) holds for all $t \in [d], k \in [L]$. By choosing the granularity D sufficiently large such that $|\mathbb{G}_D \setminus \tilde{\mathbb{G}}_D|$ is negligible, it holds

$$d_p(\mathcal{F}_2^{\text{FF}} \circ \mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{\text{FF}}, g_1) < 2\epsilon/3 \quad (\text{H.22})$$

Combining (H.12), we have

$$d_p(\mathcal{F}_2^{\text{FF}} \circ \mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{\text{FF}}, f) < \epsilon.$$

It remains to show that there exists a $\mathcal{F}_2^{\text{FF}}$ that implements (H.21) for all $t \in [d], k \in [L]$. This is possible since Theorem H.1 ensures that $(\gamma_{\min}, \gamma_{\max}, \epsilon)$ -tokenwise separated sequences are at least δ' apart from one another after the self-attention mapping, where $\delta' = \exp(-5\lambda\epsilon^{-1}|\mathcal{V}|^4 d \gamma_{\max}^2 \log L / \gamma_{\min})$, $\lambda > 0$ is the inverse-temperature constant and $|\mathcal{V}|$ is the size of the vocabulary set.

Specifically, for a fixed $\bar{C}$ and $u = \mathcal{F}^{\text{SA}}(\bar{C})$, we construct a $W_1^{(i)} \in \mathbb{R}^{3dL \times L}$ and a $b_1^{(i)} \in \mathbb{R}^{3dL}$:

$$W_1^{(i)} = \begin{pmatrix} R_{\text{FF}} & 0 & 0 & \cdots & 0 \\ R_{\text{FF}} & 0 & 0 & \cdots & 0 \\ R_{\text{FF}} & 0 & 0 & \cdots & 0 \\ 0 & R_{\text{FF}} & 0 & \cdots & 0 \\ 0 & R_{\text{FF}} & 0 & \cdots & 0 \\ 0 & R_{\text{FF}} & 0 & \cdots & 0 \\ 0 & 0 & R_{\text{FF}} & \cdots & 0 \\ 0 & 0 & R_{\text{FF}} & \cdots & 0 \\ 0 & 0 & R_{\text{FF}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & R_{\text{FF}} \\ 0 & 0 & 0 & \cdots & R_{\text{FF}} \\ 0 & 0 & 0 & \cdots & R_{\text{FF}} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_{\text{FF}} & 0 & 0 & \cdots & 0 \\ R_{\text{FF}} & 0 & 0 & \cdots & 0 \\ R_{\text{FF}} & 0 & 0 & \cdots & 0 \\ R_{\text{FF}} & 0 & \cdots & 0 & \\ 0 & R_{\text{FF}} & 0 & \cdots & 0 \\ 0 & R_{\text{FF}} & 0 & \cdots & 0 \\ 0 & 0 & R_{\text{FF}} & \cdots & 0 \\ 0 & 0 & R_{\text{FF}} & \cdots & 0 \\ 0 & 0 & R_{\text{FF}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & R_{\text{FF}} \\ 0 & 0 & 0 & \cdots & R_{\text{FF}} \\ 0 & 0 & 0 & \cdots & R_{\text{FF}} \end{pmatrix}, \quad b_1^{(i)} = \begin{pmatrix} -R_{\text{FF}}u_{1,1} \\ -R_{\text{FF}}u_{1,1} - 1 \\ -R_{\text{FF}}u_{1,1} + 1 \\ -R_{\text{FF}}u_{2,1} \\ -R_{\text{FF}}u_{2,1} - 1 \\ -R_{\text{FF}}u_{2,1} + 1 \\ -R_{\text{FF}}u_{3,1} \\ -R_{\text{FF}}u_{3,1} - 1 \\ -R_{\text{FF}}u_{3,1} + 1 \\ \vdots \\ -R_{\text{FF}}u_{d,1} \\ -R_{\text{FF}}u_{d,1} - 1 \\ -R_{\text{FF}}u_{d,1} + 1 \\ \vdots \\ \vdots \\ -R_{\text{FF}}u_{1,L} \\ -R_{\text{FF}}u_{1,L} - 1 \\ -R_{\text{FF}}u_{1,L} + 1 \\ -R_{\text{FF}}u_{2,L} \\ -R_{\text{FF}}u_{2,L} - 1 \\ -R_{\text{FF}}u_{2,L} + 1 \\ -R_{\text{FF}}u_{3,L} \\ -R_{\text{FF}}u_{3,L} - 1 \\ -R_{\text{FF}}u_{3,L} + 1 \\ \vdots \\ -R_{\text{FF}}u_{d,L} \\ -R_{\text{FF}}u_{d,L} - 1 \\ -R_{\text{FF}}u_{d,L} + 1 \end{pmatrix} \quad (\text{H.23})$$

where the first $3L$ rows of $W_1^{(i)}$ are repeated d times and $S = \mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{\text{FF}}(Z)$ is mapped to $\mathbb{R}^{3dL \times L}$:

$$W_1^{(i)} S + b_1^{(i)}$$

$$= \begin{pmatrix} R_{\text{FF}}(S_{1,1} - u_{1,1}) & R_{\text{FF}}(S_{1,2} - u_{1,1}) & \cdots & R_{\text{FF}}(S_{1,L} - u_{1,1}) \\ R_{\text{FF}}(S_{1,1} - u_{1,1}) - 1 & R_{\text{FF}}(S_{1,2} - u_{1,1}) - 1 & \cdots & R_{\text{FF}}(S_{1,L} - u_{1,1}) - 1 \\ R_{\text{FF}}(S_{1,1} - u_{1,1}) + 1 & R_{\text{FF}}(S_{1,2} - u_{1,1}) + 1 & \cdots & R_{\text{FF}}(S_{1,L} - u_{1,1}) + 1 \\ R_{\text{FF}}(S_{2,1} - u_{2,1}) & R_{\text{FF}}(S_{2,2} - u_{2,1}) & \cdots & R_{\text{FF}}(S_{2,L} - u_{2,1}) \\ R_{\text{FF}}(S_{2,1} - u_{2,1}) - 1 & R_{\text{FF}}(S_{2,2} - u_{2,1}) - 1 & \cdots & R_{\text{FF}}(S_{2,L} - u_{2,1}) - 1 \\ R_{\text{FF}}(S_{2,1} - u_{2,1}) + 1 & R_{\text{FF}}(S_{2,2} - u_{2,1}) + 1 & \cdots & R_{\text{FF}}(S_{2,L} - u_{2,1}) + 1 \\ \vdots & \vdots & \ddots & \vdots \\ R_{\text{FF}}(S_{d,1} - u_{d,1}) & R_{\text{FF}}(S_{d,2} - u_{d,1}) & \cdots & R_{\text{FF}}(S_{d,L} - u_{d,1}) \\ R_{\text{FF}}(S_{d,1} - u_{d,1}) - 1 & R_{\text{FF}}(S_{d,2} - u_{d,1}) - 1 & \cdots & R_{\text{FF}}(S_{d,L} - u_{d,1}) - 1 \\ R_{\text{FF}}(S_{d,1} - u_{d,1}) + 1 & R_{\text{FF}}(S_{d,2} - u_{d,1}) + 1 & \cdots & R_{\text{FF}}(S_{d,L} - u_{d,1}) + 1 \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ R_{\text{FF}}(S_{1,1} - u_{1,L}) & R_{\text{FF}}(S_{1,2} - u_{1,L}) & \cdots & R_{\text{FF}}(S_{1,L} - u_{1,L}) \\ R_{\text{FF}}(S_{1,1} - u_{1,L}) - 1 & R_{\text{FF}}(S_{1,2} - u_{1,L}) - 1 & \cdots & R_{\text{FF}}(S_{1,L} - u_{1,L}) - 1 \\ R_{\text{FF}}(S_{1,1} - u_{1,L}) + 1 & R_{\text{FF}}(S_{1,2} - u_{1,L}) + 1 & \cdots & R_{\text{FF}}(S_{1,L} - u_{1,L}) + 1 \\ R_{\text{FF}}(S_{2,1} - u_{2,L}) & R_{\text{FF}}(S_{2,2} - u_{2,L}) & \cdots & R_{\text{FF}}(S_{2,L} - u_{2,L}) \\ R_{\text{FF}}(S_{2,1} - u_{2,L}) - 1 & R_{\text{FF}}(S_{2,2} - u_{2,L}) - 1 & \cdots & R_{\text{FF}}(S_{2,L} - u_{2,L}) - 1 \\ R_{\text{FF}}(S_{2,1} - u_{2,L}) + 1 & R_{\text{FF}}(S_{2,2} - u_{2,L}) + 1 & \cdots & R_{\text{FF}}(S_{2,L} - u_{2,L}) + 1 \\ \vdots & \vdots & \ddots & \vdots \\ R_{\text{FF}}(S_{d,1} - u_{d,L}) & R_{\text{FF}}(S_{d,2} - u_{d,L}) & \cdots & R_{\text{FF}}(S_{d,L} - u_{d,L}) \\ R_{\text{FF}}(S_{d,1} - u_{d,L}) - 1 & R_{\text{FF}}(S_{d,2} - u_{d,L}) - 1 & \cdots & R_{\text{FF}}(S_{d,L} - u_{d,L}) - 1 \\ R_{\text{FF}}(S_{d,1} - u_{d,L}) + 1 & R_{\text{FF}}(S_{d,2} - u_{d,L}) + 1 & \cdots & R_{\text{FF}}(S_{d,L} - u_{d,L}) + 1 \end{pmatrix}.$$

Then, we construct the second matrix by $W_2^{(i)} := W_2''^{(i)} W_2'^{(i)} \in \mathbb{R}^{d \times 3dL}$ with $W_2'^{(i)} \in \mathbb{R}^{dL \times 3dL}$ and $W_2''^{(i)} \in \mathbb{R}^{d \times dL}$. For $W_2'^{(i)}$, we set

$$W_2'^{(i)} = \begin{pmatrix} -2f(\bar{C} - \mathcal{I})_{1,1} & f(\bar{C} - \mathcal{I})_{1,1} & f(\bar{C} - \mathcal{I})_{1,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & -2f(\bar{C} - \mathcal{I})_{d,L} & f(\bar{C} - \mathcal{I})_{d,L} & f(\bar{C} - \mathcal{I})_{d,L} \end{pmatrix},$$

which maps the output of the first layer after ReLU operation to $\mathbb{R}^{dL \times L}$:

$$W_2'^{(i)} \text{ReLU}[W_1^{(i)} S + b_1^{(i)}] = \begin{pmatrix} f(\bar{C} - \mathcal{I})_{1,1} \text{bump}(S_{1,1} - u_{1,1}) & f(\bar{C} - \mathcal{I})_{1,1} \text{bump}(S_{1,2} - u_{1,1}) & \cdots & f(\bar{C} - \mathcal{I})_{1,1} \text{bump}(S_{1,L} - u_{1,1}) \\ f(\bar{C} - \mathcal{I})_{2,1} \text{bump}(S_{2,1} - u_{2,1}) & f(\bar{C} - \mathcal{I})_{2,1} \text{bump}(S_{2,2} - u_{2,1}) & \cdots & f(\bar{C} - \mathcal{I})_{2,1} \text{bump}(S_{2,L} - u_{2,1}) \\ \vdots & \vdots & \ddots & \vdots \\ f(\bar{C} - \mathcal{I})_{d,1} \text{bump}(S_{d,1} - u_{d,1}) & f(\bar{C} - \mathcal{I})_{d,1} \text{bump}(S_{d,2} - u_{d,1}) & \cdots & f(\bar{C} - \mathcal{I})_{d,1} \text{bump}(S_{d,L} - u_{d,1}) \\ f(\bar{C} - \mathcal{I})_{1,2} \text{bump}(S_{1,1} - u_{1,2}) & f(\bar{C} - \mathcal{I})_{1,2} \text{bump}(S_{1,2} - u_{1,2}) & \cdots & f(\bar{C} - \mathcal{I})_{1,2} \text{bump}(S_{1,L} - u_{1,2}) \\ f(\bar{C} - \mathcal{I})_{2,2} \text{bump}(S_{2,1} - u_{2,2}) & f(\bar{C} - \mathcal{I})_{2,2} \text{bump}(S_{2,2} - u_{2,2}) & \cdots & f(\bar{C} - \mathcal{I})_{2,2} \text{bump}(S_{2,L} - u_{2,2}) \\ \vdots & \vdots & \ddots & \vdots \\ f(\bar{C} - \mathcal{I})_{d,2} \text{bump}(S_{d,1} - u_{d,2}) & f(\bar{C} - \mathcal{I})_{d,2} \text{bump}(S_{d,2} - u_{d,2}) & \cdots & f(\bar{C} - \mathcal{I})_{d,2} \text{bump}(S_{d,L} - u_{d,2}) \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ f(\bar{C} - \mathcal{I})_{1,L} \text{bump}(S_{1,1} - u_{1,L}) & f(\bar{C} - \mathcal{I})_{1,L} \text{bump}(S_{1,2} - u_{1,L}) & \cdots & f(\bar{C} - \mathcal{I})_{1,L} \text{bump}(S_{1,L} - u_{1,L}) \\ f(\bar{C} - \mathcal{I})_{2,L} \text{bump}(S_{2,1} - u_{2,L}) & f(\bar{C} - \mathcal{I})_{2,L} \text{bump}(S_{2,2} - u_{2,L}) & \cdots & f(\bar{C} - \mathcal{I})_{2,L} \text{bump}(S_{2,L} - u_{2,L}) \\ \vdots & \vdots & \ddots & \vdots \\ f(\bar{C} - \mathcal{I})_{d,L} \text{bump}(S_{d,1} - u_{d,L}) & f(\bar{C} - \mathcal{I})_{d,L} \text{bump}(S_{d,2} - u_{d,L}) & \cdots & f(\bar{C} - \mathcal{I})_{d,L} \text{bump}(S_{d,L} - u_{d,L}) \end{pmatrix}. \quad (\text{H.24})$$

Recall that we have tokenwise $(1/D, 2I\sqrt{d}, 1/D)$ -separated sequence after the quantization of the target domain Ω , and $|V| \leq (2ID)^d$. Therefore, it suffices to set the scale of the bump function as

$$R_{\text{FF}} := O(1/\delta') = O(\exp(640\lambda(ID)^{4d+2}d^2 \log L)).$$

(H.24) is a partition of L vertical blocks, each of dimension $d \times L$. Then, if S is a permutation of u , a single column within each block is comprised of correct target output since their bump function all evaluates to one, and all other columns containing zeros. Also, The position of this all-ones column is distinct for every block. Conversely, if S is not a permutation of u , (H.24) becomes the zero matrix.

Lastly, we construct $W_2^{(i)}$ by $(I_d, I_d, \dots, I_d)$ to sum over every column in each block matrix, where the $d \times d$ identity matrix is concatenated L times.

Up to permutation equivariance, there are total $q_2 = (2ID)^{dL}/L!$ possible u , so we stack $W_1^{(i)}, b_1^{(i)}$ and $W_2^{(ii)}$ with weights corresponding to each distinct u using the identical construction.

In sum, there exist a feed-forward block $\mathcal{F}_2^{\text{FF}}$ with $W_1 \in \mathbb{R}^{3dLq_2 \times d}$, a $b_1 \in \mathbb{R}^{3dLq_2}$ and a $W_2 \in \mathbb{R}^{d \times 3dLq_2}$ that implements the bump function (H.21) for all $t \in [d]$ and $k \in [L]$.

This completes the proof. $\square$

Remark H.1. Note that Theorem H.2 uses 2 FFN layers and $g \in \mathcal{T}_R^{1,1,r}$, where hidden dimension $r = O(dL(ID)^{dL}/L!)$. By Definition B.2, $\mathcal{T}_R^{1,1,r}$ belongs to our transformer network class.

To eliminate the permutation equivariance required for the target function, we incorporate the positional encoding to Theorem H.2 to break the symmetry following [Yun et al., 2019].

Corollary H.2.1 (Universal Approximation of Transformers with Positional Encoding, [Kajitsuka and Sato, 2023]). Let $\epsilon \in (0, 1)$ and $p \in [1, \infty)$. Let $\mathcal{F}_1^{(\text{FF})}, \mathcal{F}_2^{(\text{FF})}$ be two feed-forward layers and $\mathcal{F}^{(\text{SA})}$ be a single-head self-attention layer with softmax function defined in (B.1) and (B.2). Then, for any continuous function f on a compact domain and any ϵ , there exists a positional encoding $E \in \mathbb{R}^{d \times L}$ and a $h(Z) = \mathcal{F}_2^{(\text{FF})} \circ \mathcal{F}^{(\text{SA})} \circ \mathcal{F}_1^{(\text{FF})}(Z + E) \in \mathcal{T}_R^{h,s,r}$ such that $d_p(f, h) < \epsilon$, where $d_p := (\int \|f(Z) - g(Z)\|_p^p dZ)^{1/p}$ and $\|\cdot\|_p$ is the element-wise ℓ_p -norm.

Proof. It suffices to show that the universal approximation remains valid with the inclusion of a positional encoding $E \in \mathbb{R}^{d \times L}$ to the weight matrices constructed in Theorem H.2.

Recall the (H.14) and (H.20). We have token-wise $(1/D, 2I\sqrt{d}, 1/D)$ -separated sequences on $[0, 2I]^{d \times L}$ after quantization. Then, we add the positional encoding to the quantized sequences:

$$E := \begin{pmatrix} 2I & 4I & \cdots & 2IL \\ \vdots & \vdots & \ddots & \vdots \\ 2I & 4I & \cdots & 2IL \end{pmatrix},$$

giving token-wise $(2I\sqrt{d}, 2(L+1)I\sqrt{d}, 1/D)$ -separated sequences, where the first column is in $[0, 4I]^d$, the second is in $[0, 6I]^d$ and so on. For each row, entries are monotonically increasing.

The second step contextual mapping remains valid by scaling the penalty function (H.18) to $2I(L+1)$. By Theorem H.1, columns after the self-attention mapping are at least δ' apart (in ℓ_2 -distance) from each other, where $\delta' = \exp(-5\lambda\epsilon^{-1}|\mathcal{V}|^4 d\gamma_{\max}^2 \log L/\gamma_{\min})$ and $|\mathcal{V}| = O((ID)^d)$.

Recall the bump function (H.21) and (H.24). The construction of $W_1^{(i)}, b_1^{(i)}$ and $W_2^{(i)}$ follows by setting the scale parameter $R_{\text{FF}} = O(1/\delta)$. Because of the lack of permutation equivariance here, it necessitates to stack these matrices and biases for all $q_2 = (2ID)^{dL}$ possible values u .

This completes the proof. $\square$

Parameter Norm Bounds for Transformer Approximation. Next lemma provides matrices norm bounds required to achieve the universal approximation of transformer with any error ϵ .

Lemma H.5 (Transformer Matrices Bounds, Modified from Lemma F.4 and Lemma F.5 of [Hu et al., 2025c]). Let $\epsilon \in (0, 1)$. Let $Z \in [-I, I]^{d \times L}$ be an input sequence, where I is an absolute positive constant and $L \geq 2$. Let $f(Z) : [-I, I]^{d \times L} \rightarrow \mathbb{R}^{d \times L}$ be any Lipchitz continuous function with respect to some norm d_Y . Then, for $g \in \mathcal{T}_R^{r,h,s}$ that approximates f within ϵ precision, i.e., $d_Y(f, g) < \epsilon$, the parameter bounds in the transformer network class follow:

$$\begin{aligned} C_{KQ}, C_{KQ}^{2,\infty} &= O(\lambda^{-1} I^{4d+2} \epsilon^{-4d-2}); & C_{OV}, C_{OV}^{2,\infty} &= O(\epsilon); \\ C_F, C_F^{2,\infty} &= O(I \epsilon^{-1} \cdot \max \|f(Z)\|_F); & C_E &= O(I), \end{aligned}$$

where $\lambda^{-1} = O(I/\epsilon)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides polynomial and logarithmic factors depending on d and L . Further, for all feed-forward layers,

$$\max\{\|b_1\|_\infty, \|b_2\|_\infty\} = O(I \epsilon^{-1}).$$

Proof. Hu et al. [2025c] provide parameter bounds for the universal approximation of transformers on domain $[0, 1]^{d \times L}$. We specify these bounds for approximation on domain $[-I, I]^{d \times L}$.

Recall the construction of weight matrices in the proof of Theorem H.2. We achieve the universal approximation by choosing “sufficiently large” granularity D and “sufficiently small” δ (H.13).

To prove Lemma H.5, we first identify the order of δ, D and R_{FF} in terms ϵ . Then, we derive norm bounds on matrices in two feed-forward layers $\mathcal{F}_1^{\text{FF}}, \mathcal{F}_2^{\text{FF}}$, and the self attention layer $\mathcal{F}^{\text{SA}}$.

Bound on δ . Recall the approximation of quantization function in (H.13). In each step function, we have extra partition $(1/D, 1/D + \delta)$. Therefore, it suffices to take $\delta = o(1/D)$.

Bound on the Granularity D . Recall the contextual mapping step in the proof of Theorem H.2. The total omitted duplicated points in the grid $\mathbb{G}_D^\circ$ are $|\mathbb{G}_D^\circ \setminus \tilde{\mathbb{G}}_D| = |D^{-d} \cdot (2ID)^{dL}|$, where $\tilde{\mathbb{G}}_D \subseteq \mathbb{G}_D^\circ$ is the sub-grid consisting of sequences with non-duplicated tokens. Further, by the extreme value theorem, $\|f\|_p^p \leq B_T$ for a constant $B_T > 0$. Then, the difference between the target function f and the piece-wise constant approximator g_1 with granularity D is bounded by

$$\begin{aligned} d_p(f, g_1) &= \left(\int \|f(Z) - g_1(Z)\|_p^p dZ \right)^{\frac{1}{p}} \\ &= O\left((D^{-d}(2ID)^{dL} \cdot B_T(1/D)^{dL})^{\frac{1}{p}}\right) \\ &= O(D^{-d/p} \cdot I^{dL}). \end{aligned}$$

For $p \in [1, \infty)$, we have that $\epsilon = O(D^{-d/p} \cdot I^{dL})$. This implies $D = O(\epsilon^{-p/d} \cdot I^{-L/p})$. Without loss of generality, we drop $I^{-L/p} \in (0, 1)$ and drop the constant p . Then, we have that $D = O(\epsilon^{-1/d})$.

Next, recall the piece-wise constant approximation (H.10), (H.11) and (H.12).

For Lipchitz continuous target function f , there exist a grid $\mathbb{G}_D$ on domain $[-I, I]^{d \times L}$ such that

$$d_p(f(Z), g_1(Z)) < L_f \|Z - Z'\|_2 \leq L_f \|Z - Z'\|_F \leq \sqrt{dL} L_f / D,$$

where $Z' \in \mathbb{G}_D$ and L_f is the Lipchitz constant with respect to the matrix 2-norm. Therefore, it suffices to take $\epsilon = \sqrt{dL} L_f / D$. Altogether, we take $D = O(\epsilon^{-1})$ such that Theorem H.2 holds.

Next, we derive the norm bounds on transformer weight matrices.

- **Bounds on W_Q and W_K in $\mathcal{F}^{\text{SA}}$.** For the self-attention layer, we denote the separatedness of the input tokens by $(\gamma_{\min}, \gamma_{\max}, \epsilon_s)$ and the separatedness of the output tokens by (γ, δ_s) .

Recall **Theorem H.1**. We construct rank ρ matrix W_Q and W_K in the self-attention layer by

$$W_K = \sum_{i=1}^{\rho} p_i q_i^\top \in \mathbb{R}^{s \times d}, \quad W_Q = \sum_{i=1}^{\rho} p'_i q'_i{}^\top \in \mathbb{R}^{s \times d},$$

with the identity $p_i^\top p'_i = (|\mathcal{V}| + 1)^4 d \delta_s / (\epsilon_s \gamma_{\min})$. Then, the bounds on W_{KQ} follows

$$\begin{aligned} \|W_{KQ}\|_2 &\leq \|W_{KQ}\|_F = \|(W_K)^\top W_Q\|_F = O\left(\frac{\delta_s |\mathcal{V}|^4}{\epsilon_s \gamma_{\min}}\right), \\ \|W_{KQ}\|_{2,\infty} &= \|(W_K)^\top W_Q\|_{2,\infty} = O\left(\frac{\delta_s |\mathcal{V}|^4}{\epsilon_s \gamma_{\min}}\right). \end{aligned}$$

We identify the order of each terms. Recall the first step quantization (H.13). We have total $(DI)^{dL}$ input that are token-wise $(1/D, 2I\sqrt{d}, 1/D)$ -separated.

Further, since there are at most DI possible values that each entry can take, we have vocabulary $|\mathcal{V}| = O((DI)^d)$ and $\gamma_{\min}, \epsilon_s = (2DI)^{-1}$. Further, from the proof of the second step contextual mapping in **Theorem H.1**, we construct the self-attention such that $\delta_s \cdot \lambda = 4 \log L$, where we specify the choice of λ in (H.25). Finally, by $D = O(\epsilon^{-1})$ the bounds on W_{KQ} follows

$$\|W_{KQ}\|_2 \leq C_{KQ} = O(\lambda^{-1} \epsilon^{-4d-2} \cdot I^{4d+2}); \quad \|W_{KQ}\|_{2,\infty} \leq C_{KQ}^{2,\infty} = O(\lambda^{-1} \epsilon^{-4d-2} \cdot I^{4d+2}).$$

- **Bounds on W_O and W_V in $\mathcal{F}^{\text{SA}}$.** From the proof of contextual mapping **Theorem H.1**, we have,

$$W_V = \sum_{i=1}^{\rho} p''_i q''_i{}^\top \in \mathbb{R}^{s \times d}; \quad W_O = \sum_{i=1}^{\rho} p'''_i p''_i{}^\top \in \mathbb{R}^{d \times s},$$

with the identity $\|p'''_i\| \lesssim \epsilon_s / (4\rho\gamma_{\max}\|p''_i\|)$ from (H.5), and $p''_i \in \mathbb{R}^s$ is any nonzero vector.

With the $(\gamma_{\min} = 1/D, \gamma_{\max} = \sqrt{d}, \epsilon_s = 1/D)$ separateness and $D = O(\epsilon^{-1})$, we have

$$\begin{aligned} \|W_V\|_2 &= \sup_{\|x\|_2=1} \|W_V x\|_2 \leq C_V = O(\sqrt{\rho}) = O(\sqrt{d}), \\ \|W_V\|_{2,\infty} &= \max_{1 \leq i \leq L} \|(W_V)_{(:,i)}\|_2 \leq C_V^{2,\infty} = O(\rho) = O(d), \\ \|W_O\|_2 &= \sup_{\|x\|_2=1} \|W_O x\|_2 \leq C_O = O(\sqrt{\rho} \cdot \rho^{-1} \cdot \gamma_{\max}^{-1} \cdot \epsilon_s) = O(d^{-1}\epsilon) \\ \|W_O\|_{2,\infty} &= \max_{1 \leq i \leq L} \|(W_O)_{(:,i)}\|_2 \leq C_O^{2,\infty} = O(\rho \cdot \rho^{-1} \cdot \gamma_{\max}^{-1} \cdot \epsilon_s) = O(d^{-1/2}\epsilon). \end{aligned}$$

Therefore,

$$\|W_{OV}\|_2 = \|W_O W_V\|_2 \leq C_{OV} = O(\epsilon); \quad \|W_{OV}\|_{2,\infty} = \|W_O W_V\|_{2,\infty} \leq C_{OV}^{2,\infty} = O(\epsilon).$$

- **Bounds on W_1 and W_2 in $\mathcal{F}_1^{\text{FF}}$** Recall (H.15). By $\delta = o(1/D)$ and $D = O(\epsilon^{-1})$, we have

$$\begin{aligned} \max\{\|W_1\|_2, \|W_2\|_2\} &\leq C_F^2 = O(\epsilon^{-1} \cdot IDL); \\ \max\{\|W_1\|_{2,\infty}, \|W_2\|_{2,\infty}\} &\leq C_F^{2,\infty} = O(\epsilon^{-1} \cdot IDL). \end{aligned}$$

- **Bounds on W_1 and W_2 in $\mathcal{F}_2^{\text{FF}}$.** Recall the construction of bump function (H.24). We take

$$R_{\text{FF}} = O(\exp(640\lambda(ID)^{4d+2}d^2 \log L)),$$

where $(\gamma_{\min}, \gamma_{\max}, \epsilon) = (2I\sqrt{d}, 2(L+1)I\sqrt{d}, 1/D)$. Recall that $D = O(\epsilon^{-1})$. Then, we take

$$\lambda = \frac{(ID)^{-4d-2} \cdot \log \epsilon^{-1}}{640d^2 \cdot \log L}. \quad (\text{H.25})$$

This gives

$$\begin{aligned} \max\{\|W_1\|_2, \|W_2\|_2\} &\leq C_F^2 = O(I\epsilon^{-1} \cdot \max \|f(Z)\|_F), \\ \max\{\|W_1\|_{2,\infty}, \|W_2\|_{2,\infty}\} &\leq C_F^{2,\infty} = O(I\epsilon^{-1} \cdot \max \|f(Z)\|_F), \end{aligned}$$

where $\Omega = [-I, I]^{d \times L}$ is the domain of the target function f .

- **Bounds on Positional Encoding Matrix E .** By [Corollary H.2.1](#), we have:

$$E := \begin{pmatrix} 2I & 4I & \cdots & 2IL \\ \vdots & \vdots & \ddots & \vdots \\ 2I & 4I & \cdots & 2IL \end{pmatrix},$$

Therefore, we have $C_E = O(I)$.

Finally, recall that (i) $u = \mathcal{F}^{(\text{SA})}(\bar{C})$ from [\(H.23\)](#) (ii) $C_{OV} = O(\epsilon)$ and (iii) $R_{\text{FF}} = \epsilon^{-1}$ by [\(H.25\)](#). Then, the bound on the bias holds by the construction in [\(H.15\)](#), [\(H.16\)](#) and [\(H.23\)](#).

This completes the proof. □

I Statistical Rates of Flow Matching Transformers (FMTs)

In this section, we present statistical rates for the first order flow, i.e., the velocity field, $u_t(x)$.

Specifically, we consider the target density function $q_1(x)$ in the Hölder space (Definition I.1) with sub-Gaussian property. Then, we bound the approximation and estimation error for $u_t(x)$. Further, we extend these results to derive distribution estimation rates under the 2-Wasserstein distance. Compared to high-order flow matching statistical rates Section 4, we remove the requirement of Lipschitz continuousness of the velocity field $u_t(x)$.

Organizations. Appendix I.1 presents velocity approximation under a generic Hölder smoothness assumption. Appendix I.2 adopts a stronger Hölder smoothness assumption; this yields tighter approximation error bounds toward minimax optimality in velocity estimation. Appendix I.3 utilizes these approximation results to develop velocity estimation bounds and distribution estimation rates. Finally, Appendix I.4 establishes the nearly minimax optimality of flow matching transformers.

I.1 Velocity Approximation: Generic Hölder Smooth Data Distributions

Establishing our statistical theory begins with approximating the velocity using transformers. We present the corresponding velocity approximation theory under the Hölder smoothness assumption on the initial data [Fu et al., 2024]. This theory ensures our approximation rate adaptive to the initial data's smoothness. First, we restate the definition of Hölder space and Hölder ball.

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

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

Also, we define the Hölder ball of radius B by $\mathcal{H}^\beta(\mathbb{R}^d, B) := \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid \|f\|_{\mathcal{H}^\beta(\mathbb{R}^d)} < B\}$.

Before presenting the main result of velocity approximation, we state our two assumptions: (i) the Generic Hölder Smooth assumption on the target distribution $q(x_1)$. (ii) the regularity assumption on the first derivative of path coefficients. In particular, (i) and (ii) are the counterparts of Assumption 4.1 and Assumption 4.2 in the K order flow matching framework (Section 4) respectively. Notably, we remove the Lipschitzness assumption via a more fine-grained analysis on the velocity field $u_t(x)$.

Assumption I.1 (Generic Hölder Smooth Data). The density function $q(x_1)$ belongs to Hölder ball of radius $B > 0$ with Hölder index $\beta > 0$ (Definition 4.1), denoted by $q(x_1) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B)$. Also, there exist constant $C_1, C_2 > 0$ such that $q(x_1) \leq C_1 \exp(-C_2 \|x_1\|_2^2/2)$.

Assumption I.2 (Path Regularity). Consider the affine conditional flow $\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$. The first-derivative of path coefficients $\dot{\sigma}_t$ and $\dot{\alpha}_t$ are continuous on $[t_0, T]$, where $t_0, T \in (0, 1)$.

Remark I.1. We remark that such path assumption is general and applies to a number of common scenarios. For instance, Lipman et al. [2024] present: (i) the conditional optimal transport schedule: $\psi_t(x|x_1) = tx_1 + (1-t)x$, (ii) the polynomial schedule: $\psi_t(x|x_1) = t^n x_1 + (1-t^n)x$, (iii) the linear variance preserving schedule: $\psi_t(x|x_1) = tx_1 + \sqrt{1-t^2}x$. These cases satisfy Assumption I.2.

We now present the velocity approximation for flow matching transformers.

Theorem I.1 (Velocity Approximation with Transformers under Generic Hölder Smoothness). Assume Assumption I.1 and Assumption I.2. For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$. Then, for all $t \in [t_0, T]$ with $t_0, T \in (0, 1)$,

there exists a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u_t(x) - u_\theta(x, t)\|_2^2 \cdot p_t(x) dx dt = O\left(B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{\beta}{2} + 1}\right).$$

Let d be the feature dimension and L be the sequence length defined by the flow matching reshape layer in [Definition B.3](#). Then, the parameter bounds in transformer network $\mathcal{T}_R^{h,s,r}$ satisfy

$$C_{KQ}, C_{KQ}^{2,\infty} = O(\lambda^{-1} N^{4\beta d + 2\beta} (\log N)^{4d_x + 2}); C_{OV}, C_{OV}^{2,\infty} = O(N^{-\beta});$$

$$C_F, C_F^{2,\infty} = O(N^\beta (\log N)^{\frac{d_x + \beta}{2} + 1}); C_E = O(1); C_T = O(\sqrt{\log N}).$$

where $\lambda^{-1} = O(N^\beta \cdot \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof Sketch. We adopt the following strategy:

- **Step 1: Approximation on a Compact Domain via Transformer Universality.** To reflect the Hölder smoothness of the target density $q(x_1)$, we begin by applying a multivariate Taylor expansion to construct a compactly supported approximation of velocity field $u_t(x)$. We then approximate this function on a compact domain using the universal approximation of transformers.
- **Step 2: Extension to the Full Domain via Sub-Gaussian Tails.** We exploit the sub-Gaussian tail behavior of the target distribution to control the approximation error outside the compact region. Combining the errors from both regions yields the final approximation rate for the velocity field.

Please see [Appendix J](#) for a detailed proof. □

I.2 Velocity Approximation: Stronger Hölder Smooth Data Distributions

We obtain tighter velocity approximation rates than [Appendix I.1](#) by imposing stronger Hölder smoothness assumption on the target distribution $q(x_1)$.

Assumption I.3 (Stronger Hölder Smooth Data). Let C, C_1 and C_2 be positive constants. The density function satisfies $q(x_1) = \exp\left(-C_2 \|x_1\|_2^2 / 2\right) \cdot f(x_1)$, where f belongs to Hölder space $f(x_1) \in \mathcal{H}^\beta(\mathbb{R}^{d_x}, B)$ ([Definition 4.1](#)) and satisfies $C_1 \geq f(x_1) \geq C$ for all x_1 .

The density lower bound prevents $f(x)$ from taking small values, ensuring well-conditioned approximation. Without this bound, small values of $f(x)$ require a chosen threshold to maintain uniform approximation. A positive lower bound eliminates the need for such adjustments, keeping the approximation error controlled across the domain and enabling efficient convergence.

Assuming [Assumption I.3](#), we derive the velocity approximation for flow matching transformers.

Theorem I.2 (Velocity Approximation with Transformers under Stronger Hölder Smoothness). Assume [Assumption I.3](#) and [Assumption I.2](#). For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$. Then, for all $t \in [t_0, T]$ with $t_0, T \in (0, 1)$, there exists a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u_t(x) - u_\theta(x, t)\|_2^2 \cdot p_t(x) dx dt = O\left(B^2 N^{-2\beta} (\log N)^{d_x + \beta}\right),$$

Further, the parameter bounds in the transformer network class follows [Theorem I.1](#).

Proof Sketch. The proof strategy closely follows [Theorem I.1](#):

- **Step 0: Velocity Decomposition.** We invoke [Assumption I.3](#) to decompose the velocity field into a target function that is lower bounded. This step mitigates the influence of low-density regions and enables a more refined approximation analysis, in contrast to the setting under [Assumption I.1](#).
- **Step 1: Approximation with Transformer Universality on Compact Domain.** To capture the Hölder regularity of the target density $q(x_1)$, we construct a compactly supported function as an intermediary to approximate the velocity field $u_t(x)$ using multivariate Taylor expansion. We then apply the universal approximation of transformers to approximate the constructed function.
- **Step 2: Full Domain Approximation.** We extend the approximation to the full space by leveraging the sub-Gaussian tail behavior, ensuring that the error outside the compact region remains controlled. Then, we incorporate all errors terms to achieve the final approximation rates for $u_t(x)$.

Please see [Appendix K](#) for a detailed proof. $\square$

I.3 Velocity Estimation and Distribution Estimation

In this section, we study the statistical estimation problems and develop sample complexity results based on the established approximation results in [Appendix I.1](#) and [Appendix I.2](#). Specifically, we present the estimation error bound of flow matching transformers in [Theorem I.3](#). Applying the velocity estimation rates, we further study the distribution estimation in [Theorem I.4](#).

Velocity Estimation Building on the transformer-based velocity approximation, we evaluate the performance of the velocity estimator u_θ trained with i.i.d. data points $\{x_i\}_{i=1}^n$ by optimizing the empirical loss (2.12). To quantify this, we define flow matching risk:

Definition I.2 (Flow Matching Risk). Let q be the target distribution and $X_1 \sim q$. Given a velocity estimator u_θ , we define the flow matching risk $\mathcal{R}(u_\theta)$ as the expectation of the mean-squared difference between the u_θ and the ground truth u_t :

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

where marginal probability path p_t and marginal velocity field u_t are induced by affine conditional flow $\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$ follows (2.2), (2.3), (2.5) and (2.6).

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

Theorem I.3 (Velocity Estimation with Transformer). Let d be the feature dimension. Suppose we choose the transformers as in [Theorem I.1](#) and [Theorem I.2](#) correspondingly, then we have

- Assume [Assumption I.1](#) and [Assumption I.2](#). Then,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{16d+15}} (\log n)^{20d_x+4\beta+20}).$$

- Assume [Assumption I.2](#) and [Assumption I.3](#). Then,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{8d+9}} (\log n)^{20d_x+4\beta+20}).$$

Proof Sketch. Recall (2.12) from [Section 2](#). We obtain the velocity estimator $\hat{u}_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ by minimizing the empirical conditional flow matching loss:

$$\hat{\mathcal{L}}_{\text{CFM}}(u_\theta) := \frac{1}{n} \sum_{i=1}^n \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{X_0 \sim N(0, I)} [\|u_\theta(\mu_t x_i + \sigma_t X_0, t) - (\mu_t x_i + \sigma_t X_0)\|_2^2] dt.$$

To derive the estimation error, we adopt a standard strategy in empirical process theory. This involves bounding the generalization gap between empirical and true risk using covering number techniques:

- **Step 1: Domain Truncation for Risk Control.** We truncate the domain of the flow matching risk and the flow matching loss to ensure the transformer network has a finite covering number. We then control the error outside of the truncated domain by using the sub-Gaussian tail bound.
- **Step 2: Analysis on the Complexity of the Transformer Network Class via Covering Number.** Using the norm bounds on transformer parameters from [Appendix I.2](#), we derive an upper bound on the covering number of the transformer network function class. This captures the model complexity required to achieve a desired approximation rate on the compact domain.
- **Step 3: Final True Risk Upper Bound.** We apply the covering number bound to control the deviation between the empirical risk and the true risk. Lastly, we incorporate all sources of error from previous steps to derive the final estimation rate for the learned velocity field $\hat{u}_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ via the minimization of the empirical conditional flow matching loss $\hat{\mathcal{L}}_{\text{CFM}}(u_\theta)$ in (2.12).

Please see [Appendix L](#) for a detailed proof. □

Distribution Estimation. Next, we analyze the distribution estimation rate for the velocity estimator $\hat{u}_\theta$ through the 2-Wasserstein distance between estimated and true distributions. Based on the velocity estimation results in [Appendix I.3](#), the next theorem presents upper bounds on the 2-Wasserstein distance between the target distribution and the estimated distribution induced by the velocity estimator $\hat{u}_\theta$ trained from optimizing the empirical conditional loss (2.12).

Theorem I.4 (Distribution Estimation under 2-Wasserstein Distance). Let $\hat{P}_T$ denote the estimated distribution at time T . Let d be the feature dimension.

- Assume [Assumption I.1](#) and [Assumption I.2](#). It holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] = O(n^{-\frac{1}{32d+30}} (\log n)^{10d_x+2\beta+10}).$$

- Assume [Assumption I.2](#) and [Assumption I.3](#). It holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] = O(n^{-\frac{1}{16d+18}} (\log n)^{10d_x+2\beta+10}).$$

Proof Sketch. We derive the distribution estimation rate under the 2-Wasserstein distance by relating it to the velocity estimation error through the flow dynamics. Our proof follows three steps:

- **Step 1: Flow Deviation via Alekseev–Gröbner Lemma.** We apply the Alekseev–Gröbner lemma ([Lemma M.2](#)) to bound the deviation between the learned flow ψ_θ and the true flow ψ in terms of the difference between the estimated velocity $\hat{u}_\theta(x, t)$ and true velocity fields $u_t(x)$.
- **Step 2: Bounding the Jacobian via Grönwall’s Inequality.** The flow deviation bound given by the Alekseev–Gröbner lemma involves the Jacobian matrix $D\psi_\theta$. To ensure the deviation remains controlled over time, we use Grönwall’s inequality ([Lemma M.1](#)) along with the Lipschitz continuity of the network to upper bound the Jacobian norm by an exponential function.
- **Step 3: From Velocity Error to Wasserstein Distance.** We integrate the velocity error over time and apply the definition of the 2-Wasserstein metric to relate the flow deviation to $W_2(\hat{P}_T, P_T)$. Substituting the velocity estimation error from [Theorem I.3](#) then gives the final convergence rate.

Please see [Appendix M](#) for a detailed proof. □

I.4 Minimax Optimal Estimation

In [Theorem I.4](#), we present a fine-grained analysis of distribution estimation. In this section, we further show that the derived estimation rates match the minimax lower bounds in Hölder space under the 2-Wasserstein metric in specific setting. We begin by recalling the minimax optimal rate for distribution estimation over Hölder smooth function classes.

Lemma I.1 (Modified from Theorem 3 of [\[Niles-Weed and Berthet, 2022\]](#)). Consider the task of estimating a probability distribution $P(x_1)$ with density belonging to the space

$$\mathcal{P} := \{q(x_1) | q(x_1) \in \mathcal{H}^\beta([-1, 1]^{d_x}, B), q(x_1) \geq C\},$$

Then, for any $r \geq 1$, $\beta > 0$ and $d_x > 2$, we have

$$\inf_{\hat{P}} \sup_{q(x_1) \in \mathcal{P}} \mathbb{E}_{\{x_i\}_{i=1}^n} [W_r(\hat{P}, P)] \gtrsim n^{-\frac{\beta+1}{d_x+2\beta}},$$

where $\{x_i\}_{i=1}^n$ is a set of i.i.d samples drawn from distribution P , and $\hat{P}$ runs over all possible estimators constructed from the data.

Proof. Please see [Appendix N](#) for a detailed proof. □

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

Theorem I.5 (Minimax Optimality of Flow Matching Transformers). Under the setting of $(16d + 18)(\beta + 1) = d_x + 2\beta$, the distribution estimation rate of flow matching transformers ([Theorem I.4](#)) matches the minimax lower bound of Hölder distribution class in 2-Wasserstein distance up to a $\log n$ and Lipschitz constants factors.

Proof. Please see [Appendix N](#) for a detailed proof. □

J Proof of Theorem I.1

In this section, we use transformers to approximate velocity and give an upper bound of the velocity approximation error. We prove Theorem I.1 following the three steps shown in the proof sketch.

Organizations. Appendix J.1 introduces auxiliary lemmas. Appendix J.2 establishes a bound on the velocity approximation error over a bounded domain by applying the universal approximation of transformers. Appendix J.3 presents the main proof by incorporating the bounded-domain approximation error and controlling the unbounded region using the sub-Gaussian assumption.

J.1 Auxiliary Lemmas

In this section, we introduce auxiliary lemmas for velocity approximation. Specifically, we decompose the velocity field $u_t(x)$ into three components in Lemma J.1 based on the setting of affine conditional flows (Section 2). To approximate each component, we clip the integral domain of x_1 in the integrals defining $\Phi_1(x, t)$, $\Phi_2(x, t)$, and $\Phi_3(x, t)$ to a closed and bounded region in Lemma J.2. This step allows us to perform the approximation on a bounded domain while controlling the error introduced by restricting the integral. Furthermore, we revisit the bounds on the density function $p_t(x)$ in ℓ_∞ -distance, and extend these bounds to the velocity field $u_t(x)$ in Lemma J.3 and Lemma J.4.

Decomposition of Velocity Field. We present the next lemma to decompose the velocity field $u_t(x)$. Constructing an approximator for $u_t(x)$ is difficult due to its complex structure. This decomposition splits the velocity into three functions, each satisfying properties that make approximation feasible. These components allow the use of sub-Gaussian assumptions on the target distribution (Assumption I.1) and provide better control over the approximation error (Lemma J.9).

Lemma J.1 (Decomposition of Velocity Field). Under the flow matching setting (Section 2), the velocity field follows a decomposition:

$$u_t(x) = \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right),$$

where

$$\begin{aligned} \Phi_1(x, t) &:= \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\ \Phi_2(x, t) &:= x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\ \Phi_3(x, t) &:= \int_{\mathbb{R}^{d_x}} \left(\frac{x - \mu_t \cdot x_1}{\sigma_t} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1. \end{aligned}$$

Proof. By (2.5), the density function $p_t(x)$ has the form

$$\begin{aligned} p_t(x) &= \int p_t(x|x_1) \cdot q(x_1) \, dx_1 \\ &= \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \int \exp\left(-\frac{\|\mu_t x_1 - x\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1. \end{aligned}$$

Therefore, we have $p_t(x) = \Phi_1(x, t)$.

Then, we rewrite the velocity field at time t by

$$\begin{aligned} &u_t(x) \\ &= \frac{1}{p_t(x)} \cdot \int_{\mathbb{R}^{d_x}} u_t(x|x_1) p_t(x|x_1) q(x_1) \, dx_1 \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{p_t(x)} \cdot \int_{\mathbb{R}^{d_x}} \left(\frac{\dot{\sigma}_t(x - \mu_t \cdot x_1)}{\sigma_t} + \dot{\mu}_t \cdot x_1 \right) \cdot p_t(x|x_1)q(x_1)dx_1 \quad (\text{By (2.6) and (2.8)}) \\
&= \frac{1}{p_t(x)} \cdot \int_{\mathbb{R}^{d_x}} \left(\frac{\dot{\sigma}_t(x - \mu_t \cdot x_1)}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t}(x - \mu_t \cdot x_1) + \frac{\dot{\mu}_t}{\mu_t} \cdot x \right) \cdot p_t(x|x_1)q(x_1)dx_1 \\
&= \Phi_1(x, t)^{-1} \cdot \left(\dot{\sigma}_t \cdot \Phi_3(x, t) - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \cdot \Phi_3(x, t) + \frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) \right) \\
&\quad (\text{By the definition of } \Phi_1, \Phi_2 \text{ and } \Phi_3) \\
&= \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right).
\end{aligned}$$

This completes the proof. $\square$

Based on decomposition, we construct separate approximators for $\Phi_1(x, t)$, $\Phi_2(x, t)$, and $\Phi_3(x, t)$. Then, we approximate $u_t(x)$ by combining these approximations in [Appendix J.2](#).

Clipping Integral Domain. Next lemma handles unbounded integral domain of $\Phi_1(x, t)$, $\Phi_2(x, t)$, and $\Phi_3(x, t)$. [Lemma J.2](#) ensures that for any small error $\epsilon > 0$ and any fixed $x \in \mathbb{R}^{d_x}$, a bounded domain B_x dependent on ϵ and x exists, where the integral outside B_x remains bounded by ϵ .

Lemma J.2 (Clipping the Multi-Index Gaussian Integral, Lemma A.8 of [\[Fu et al., 2024\]](#) and Lemma F.9 of [\[Oko et al., 2023\]](#)). Assume [Assumption I.1](#). Let d_x be the dimension of the target data x_1 and $n \in \mathbb{N}$. Then, for any $\kappa \in \mathbb{Z}_+^{d_x}$ with $\|\kappa\|_1 \leq n$, $x_1 \in \mathbb{R}^{d_x}$ and $0 < \epsilon \leq 1/e$, there exists a constant $C(n, d_x) \geq 1$ such that

$$\int_{\mathbb{R}^{d_x} \setminus B_x} \left| \left(\frac{\mu_t \cdot x_1 - x}{\sigma_t} \right)^\kappa \right| \cdot \frac{q(x_1)}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \cdot \exp\left(-\frac{\|\mu_t x_1 - x\|^2}{2\sigma_t^2}\right) dx_1 \leq \epsilon,$$

where $(\frac{\mu_t \cdot x_1 - x}{\sigma_t})^\kappa := ((\frac{\mu_t \cdot x_1[1] - x[1]}{\sigma_t})^{\kappa[1]}, \dots, (\frac{\mu_t \cdot x_1[d_x] - x[d_x]}{\sigma_t})^{\kappa[d_x]})$ is a *multi-index vector* and

$$\begin{aligned}
B_x := & \left[\frac{x - \sigma_t C(n, d_x) \sqrt{\log(1/\epsilon)}}{\mu_t}, \frac{x + \sigma_t C(n, d_x) \sqrt{\log(1/\epsilon)}}{\mu_t} \right] \\
& \cap \left[C(n, d_x) \sqrt{\log(1/\epsilon)}, C(n, d_x) \sqrt{\log(1/\epsilon)} \right]^{d_x}. \quad (\text{J.1})
\end{aligned}$$

Remark J.1. The rationale behind this error choice follows from the need to control the clipping error, when we construct a polynomial-like approximator for the components of the decomposed velocity Φ_1 , Φ_2 , and Φ_3 on the bounded domain $B_{x,N}$. Specifically, these approximations capture the smoothness of the density function in Hölder space and leads to an error of order $N^{-\beta}$ up to a logarithmic factor. Therefore, the clipping error is set to match this order.

Bounds on Density Function and Velocity. We introduce two lemmas that provide bounds on the density function $p_t(x)$ and the velocity field $u_t(x)$. These bounds are crucial because the maximum output of the transformer network class plays a key role in analyzing the capacity of the loss function class in estimation error analysis ([Appendix I.3](#)). We start with the bounds on $p_t(x)$ and $\nabla \log p_t(x)$.

Lemma J.3 (Bounds on the Density Function, Lemma A.9 and Lemma A.10 of [\[Fu et al., 2024\]](#)). Recall that $p_t(x) = \int_{\mathbb{R}^{d_x}} p_t(x|x_1)q(x_1)dx_1$ and $p_t(x|x_1) = \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp(-\|x - \mu_t x_1\|_2^2 / 2\sigma_t^2)$. Assume [Assumption I.1](#). There exist a $C_7 > 0$ such that

$$\frac{C_7}{\sigma_t^{d_x}} \cdot \exp\left(-\frac{\|x\|_2^2 + 1}{\sigma_t^2}\right) \leq p_t(x) \leq \frac{C_1}{(\mu_t^2 + C_2 \sigma_t^2)^{d_x/2}} \cdot \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right).$$

Moreover, there exist a positive constant C'_7 such that

$$\|\nabla \log p_t(x)\|_\infty \leq \frac{C'_7}{\sigma_t^2} \cdot (\|x\|_2 + 1).$$

By [Lemma J.1](#), the velocity field $u_t(x)$ follows the decomposition

$$u_t(x) = \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right).$$

With this expression, we apply [Lemma J.3](#) to obtain bound on the velocity $u_t(x)$ in ℓ_∞ -distance.

Lemma J.4 (ℓ_∞ -Bounds on the Velocity Field). Assume [Assumption I.1](#). Then, there exists a positive constant C_5 such that

$$\|u_t(x)\|_\infty \leq \frac{|\dot{\mu}_t|}{\mu_t} \cdot \|x\|_\infty + C_5 \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \cdot (\|x\|_2 + 1).$$

Proof. Recalling from [Lemma J.1](#), we have the velocity decomposition

$$u_t(x) = \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right),$$

where

$$\begin{aligned} \Phi_1(x, t) &= \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\ \Phi_2(x, t) &= x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\ \Phi_3(x, t) &= \int_{\mathbb{R}^{d_x}} \left(\frac{x - \mu_t \cdot x_1}{\sigma_t} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1. \end{aligned}$$

First, we rewrite the expression of $\Phi_2(x, t)$ and $\Phi_3(x, t)$. Then, we derive the bound on $u_t(x)$.

• **Step 1. Rewrite $\Phi_2(x, t)$ and $\Phi_3(x, t)$.** By the definition of $\Phi_2(x, t)$ and $\Phi_3(x, t)$, it holds

$$\Phi_2(x, t) = x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1 = x \cdot \Phi_1(x, t).$$

Therefore, for all $i \in [d_x]$, it holds

$$\left| \frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t)[i] \right| = \left| \frac{\dot{\mu}_t x[i]}{\mu_t} \cdot \Phi_1(x, t) \right|. \quad (\text{J.2})$$

Next, since the gradient of $p_t(x)$ has the expression

$$\nabla p_t(x) = - \int \left(\frac{x - \mu_t \cdot x_1}{\sigma_t^2} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) q(x_1) \, dx_1,$$

we have $\Phi_3(x, t) = -\nabla p_t(x) \cdot \sigma_t$.

Therefore, for all $i \in [d_x]$, it holds

$$\left| \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \cdot \Phi_3(x, t)[i] \right| = \left| \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \sigma_t \cdot \nabla p_t(x)[i] \right|. \quad (\text{J.3})$$

• **Step 2. Bound Velocity Field.** Based on Step 1, the following holds for all $i \in [d_x]$

$$\begin{aligned}
& |u_t[i]| \\
&= \left| \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t)[i] + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t)[i] \right) \right| \\
&\leq \left| \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t)[i] \right) \right| + \left| \Phi_1(x, t)^{-1} \left(\left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \cdot \Phi_3(x, t)[i] \right) \right| \\
&\hspace{15em} \text{(By triangle inequality)} \\
&= \left| \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t x[i]}{\mu_t} \cdot \Phi_1(x, t) \right) \right| + \left| \Phi_1(x, t)^{-1} \left(\left(\frac{\dot{\mu}_t \sigma_t^2}{\mu_t} - \dot{\sigma}_t \sigma_t \right) \cdot \nabla p_t(x)[i] \right) \right| \\
&\hspace{15em} \text{(By (J.2) and (J.3))} \\
&= \left| \frac{\dot{\mu}_t}{\mu_t} \cdot x[i] \right| + \left| \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} - \dot{\sigma}_t \sigma_t \right| \cdot |\nabla \log p_t(x)[i]| \\
&\hspace{15em} \text{(By } \nabla \log p_t = \nabla p_t / p_t \text{)} \\
&\leq \left| \frac{\dot{\mu}_t}{\mu_t} \cdot x[i] \right| + C_5 \left| \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} - \dot{\sigma}_t \sigma_t \right| \cdot \left| \frac{1}{\sigma_t^2} \cdot (\|x\|_2 + 1) \right|. \\
&\hspace{15em} \text{(By Lemma J.3)}
\end{aligned}$$

Therefore, by symmetry,

$$\|u_t(x)\|_\infty \leq \left| \frac{\dot{\mu}_t}{\mu_t} \right| \cdot \|x\|_\infty + C_5 \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \cdot (\|x\|_2 + 1).$$

This completes the proof. $\square$

J.2 Velocity Approximation on Bounded Domain

In this section, we approximate the velocity field $u_t(x)$ on a bounded domain through a two-step approach. Specifically, the first step constructs three compactly supported continuous functions $\Psi_1(x, t)$, $\Psi_2(x, t)$ and $\Psi_3(x, t)$ as approximators for $\Phi_1(x, t)$, $\Phi_2(x, t)$, and $\Phi_3(x, t)$ in Lemma J.5, Lemma J.6, and Lemma J.7 respectively. Then, the second step applies the universal approximation to approximate $\Psi_1(x, t)$, $\Psi_2(x, t)$ and $\Psi_3(x, t)$ with transformers in Lemma J.8. By incorporating these steps, we derive the velocity approximation on a bounded domain in Lemma J.9.

Before proceeding, we reiterate on the velocity expression. By Lemma J.1, $u_t(x)$ has the form

$$u_t(x) = \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right),$$

where

$$\begin{aligned}
\Phi_1(x, t) &= \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\
\Phi_2(x, t) &= x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\
\Phi_3(x, t) &= \int_{\mathbb{R}^{d_x}} \left(\frac{x - \mu_t \cdot x_1}{\sigma_t} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1.
\end{aligned}$$

Approximation of $\Phi_1(x, t)$. This step builds on [Hu et al., 2025c, Fu et al., 2024].

By the expression of $\Phi_1(x, t)$:

$$\Phi_1(x, t) = \int \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|\mu_t x_1 - x\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1,$$

we approximate $q(x_1)$ and $\exp\left(-\frac{\|\mu_t x_1 - x\|^2}{2\sigma_t^2}\right)$ with k_1 -order Taylor polynomial and k_2 -order Taylor polynomial on a bounded domain $B_{x,N}$, introduced in the integral clipping (Lemma J.2). Altogether, we approximate Φ_1 with the local polynomial $\Psi_1(x, t)$ on $B_{x,N}$ with the expression:

$$\Psi_1(x, t) := \sum_{v \in [N]^{d_x}} \sum_{\|n_x\|_1 \leq k_1} \frac{R_B^{\|n_x\|_1}}{n_x!} \frac{\partial^{n_x} \Phi_1}{\partial x^{n_x}} \Big|_{x=R_B(\frac{v}{N}-\frac{1}{2})} g_1(x, n_x, v, t), \quad (\text{J.4})$$

where $n_x \in \mathbb{Z}^{d_x}$ is a multi-index, $R_B > 0$ is a constant depending on the Hölder ball radius B ,

- $g_1(x, n_x, v, t) := \prod_{i=1}^{d_x} \sum_{k_2 < p} g_2(x[i], n_x[i], v[i], k_2)$, and
- $g_2(x[i], n_x[i], v[i], k_2) := \frac{1}{\sigma_t \sqrt{2\pi}} \int \left(\frac{x_1}{R_B} + \frac{1}{2} - \frac{v[i]}{N} \right)^{n_x[i]} \frac{1}{k_2!} \left(-\frac{|x[i] - \mu_t x_1[i]|^2}{2\sigma_t^2} \right)^{k_2} dx_1$.

Hu et al. [2025c], Fu et al. [2024] consider the setting of conditional diffusion transformer with classifier-free guidance. In contrast, we apply (J.4) by removing the condition $y \in \mathbb{R}^{d_y}$.

Since $\Psi_1(x, t)$ is an approximator of $\Phi_1(x, t)$, we need to ensure that it is lower bounded away from zero so that the denominator of velocity $u_t(x)$ in Lemma J.1 does not blow up.

Therefore, we introduce an additional definition.

Definition J.1 (Truncated Density Approximator). Let ϵ_{low} be a positive real number, and let $\Psi_1(x, t)$ be a scalar-valued function defined in (J.4). Then, we define

$$\Psi_1^c(x, t) := \max\{\Psi_1(x, t), \epsilon_{\text{low}}\}.$$

We specify the choice of ϵ_{low} in Lemma J.9. For now, we approximate $\Phi_1(x, t)$ with $\Psi_1(x, t)$.

Lemma J.5 (Local Polynomial Approximation of Φ_1 , Lemma A.4 of [Fu et al., 2024]). Assume Assumption I.1. Let $\Psi_1(x, t)$ be the approximator of $\Phi_1(x, t)$. Then, for any $t \in [0, 1]$ and $x \in \mathbb{R}^{d_x}$, it holds

$$|\Psi_1(x, t) - \Phi_1(x, t)| \lesssim BN^{-\beta} (\log N)^{\frac{d_x + k_1}{2}}.$$

Next, we approximate $\Phi_2(x, t)$.

Approximation of $\Phi_2(x, t)$. By Lemma J.1, the following identity holds

$$\Phi_2(x, t) = x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) dx_1 = x \cdot \Phi_1(x, t). \quad (\text{J.5})$$

Building upon the local polynomial $\Psi_1(x, t)$, we use $x \cdot \Psi_1(x, t)$ as the approximator of $\Phi_2(x, t)$.

Next lemma gives the approximation error rate of $\Phi_2(x, t)$ using $\Psi_2(x, t) := x \cdot \Psi_1(x, t)$

Lemma J.6 (Local Polynomial Approximation of Φ_2). Assume Assumption I.1. Let $\Psi_1(x, t)$ be the local polynomial and $\Psi_2(x, t) := x \Psi_1(x, t)$. Let $C_x(d_x, \beta, C_1, C_2)$ be a positive constant. Then, for any $t \in [0, 1]$ and $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$, it holds for all $i \in [d_x]$

$$|\Psi_2(x, t)[i] - \Phi_2(x, t)[i]|_\infty \lesssim BN^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}.$$

Proof. Since $\Psi_2(x, t) = x \Psi_1(x, t)$ and $\Phi_2(x, t) = x \Phi_1(x, t)$, for all $i \in [d_x]$, it holds

$$\begin{aligned} |\Psi_2[i] - \Phi_2[i]| &= |x \Psi_1[i] - x \Phi_1[i]| \\ &\leq |x[i]| \cdot |\Psi_1 - \Phi_1| && (\text{By (J.5)}) \\ &\lesssim x[i] \cdot BN^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} && (\text{By Lemma J.5}) \end{aligned}$$

$$\lesssim BN^{-\beta} (\log N)^{\frac{d_x+k_1+1}{2}}. \quad (\text{By } x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x})$$

This completes the proof. $\square$

Approximation of $\Phi_3(x, t)$. Similarly, we have approximation results for $\Phi_3(x, t)$.

Lemma J.7 (Local Polynomial Approximation of Φ_3 , Lemma A.6 of [Fu et al., 2024]). Assume **Assumption I.1**. Let $C_x(d_x, \beta, C_1, C_2)$ be a positive constant. There exists local polynomial $\Psi_3(x, t)$ such that for all $t > 0$, $i \in [d_x]$ and $x \in [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x}$, it holds

$$|\Psi_3(x, t)[i] - |\sigma_t \nabla p_t(x)|[i]| \lesssim BN^{-\beta} (\log N)^{\frac{d_x+k_1+1}{2}}.$$

Remark J.2. We clarify that **Lemma J.7** gives the approximation of $\Phi_3(x, t)$ using $\Psi_3(x, t)$. First, the density at time t has the form:

$$p_t(x) = \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) dx_1.$$

Then, the gradient of $p_t(x)$ with respect to x has the form:

$$\nabla p_t(x) = \int_{\mathbb{R}^{d_x}} -\left(\frac{x - \mu_t \cdot x_1}{\sigma_t^2}\right) \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) dx_1.$$

By **Lemma J.1**, we have $\Phi_3(x, t) = |\sigma_t \nabla p_t(x)|$. Therefore,

$$|\Psi_3(x, t)[i] - \Phi_3(x, t)[i]| \lesssim BN^{-\beta} (\log N)^{\frac{d_x+k_1+1}{2}}. \quad (\text{By Lemma J.7})$$

Velocity Approximation with Transformers on Bounded Domain. We first approximate the velocity approximator constructed with $\Psi_1(x, t)$, $\Psi_2(x, t)$ and $\Psi_3(x, t)$. We reiterate that transformers take input $d \times L$ matrices, where $d \times L = d_x$. Then, the next lemma specifies the network configuration for the approximating the velocity approximator with arbitrarily small error.

Lemma J.8 (Approximate Velocity Approximator with Transformers). Assume **Assumption I.1**. Let $C_x(d_x, \beta, C_1, C_2)$ be a positive constant. Further, let $\Psi(x, t) : [-C_x\sqrt{\log N}, C_x\sqrt{\log N}]^{d_x} \times [0, 1] \rightarrow \mathbb{R}^{d_x}$ be the target function:

$$\Psi(x, t) := \frac{\dot{\mu}_t \Psi_2(x, t) / \mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t / \mu_t) \Psi_3(x, t)}{\Psi_1^c(x, t)}.$$

Then, for any $t \in [0, 1]$ and any $\epsilon \in (0, 1)$, there exist a transformer $g(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_0^1 \int_{\|x\|_\infty \leq C_x \sqrt{\log N}} \|g(x, t) - \Psi(x, t)\|_2^2 dx dt \leq \epsilon^2.$$

Furthermore, the parameter bounds in the transformer network class $\mathcal{T}_R^{h,s,r}$ satisfy

$$C_{KQ}, C_{KQ}^{2,\infty} = O(\lambda^{-1} (\log N)^{4d+2} \epsilon^{-4d-2}); C_{OV}, C_{OV}^{2,\infty} = O(\epsilon); \\ C_F, C_F^{2,\infty} = O(\sqrt{\log N} \cdot \epsilon^{-1} \cdot \max \|\Psi\|_2); C_E = O(1),$$

where $\lambda^{-1} = O(\log N / \epsilon)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof. Since the path coefficients are smooth and the first-step approximators $\Psi_1(x, t)$, $\Psi_2(x, t)$, and $\Psi_3(x, t)$ integrate polynomials, the target function is Lipschitz continuous on a compact domain.

Further, the reshape layer [Definition B.4](#) does not harm the continuity of the element-wise ℓ_2 -norm. This continuity ensures that the function satisfies the conditions for applying the universal approximation of transformers. Also, we concatenate t as a additional sequence. Then, we apply [Theorem H.2](#) with $p = 2$ and $Z \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d \times (L+1)}$.⁵ For any $\epsilon \in (0, 1)$, it holds

$$d_2(g, f) = \left(\int \int \|g(x, t) - \Psi(x, t)\|_2^2 dx dt \right)^{1/2} \leq \epsilon. \quad (\text{By Theorem H.2})$$

The parameter bounds in the transformer network class follow [Lemma H.5](#).

This completes the proof. $\square$

Remark J.3. [Lemma J.8](#) modifies [Lemma I.6](#) of [Hu et al., 2025c] by adapting the transformer approximation to decomposed velocity components ([Lemma J.1](#)), whereas their work focuses on approximating $\nabla \log p_t(x)$. Our flow matching framework eliminates the label y and reduces the number of hidden dimensions to one.

Then, by analyzing the error accumulation from both the transformer approximation ([Lemma J.8](#)) and the local polynomial approximations ([Lemma J.5](#), [Lemma J.6](#), and [Lemma J.7](#)), we establish a bound on the velocity approximation error over a bounded domain.

Lemma J.9 (Velocity Approximation with Transformers on Bounded Domain). Assume [Assumption I.1](#) and [Assumption I.3](#). Let $t_0, T \in (0, 1)$. Let $C_x(\beta, C_2)$ and C_3 be two positive constants. Let $\epsilon_{\text{low}} := C_3 N^{-\beta} (\log N)^{(d_x + k_1)/2}$. Then, there exist a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ such that for all $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$, $t \in [t_0, T]$ and $p_t(x) \geq \epsilon_{\text{low}}$, it holds

$$\int_{t_0}^T \int \|u_t(x) - u_\theta(x, t)\|_2^2 (p_t(x))^2 dx dt \lesssim \left(\left| \frac{\dot{\mu}_t}{\mu_t} \right| + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 B^2 N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1},$$

Furthermore, the transformer parameter bounds satisfy

$$\begin{aligned} C_{KQ}, C_{KQ}^{2,\infty} &= O(\lambda^{-1} N^{4\beta d + 2\beta} (\log N)^{4d_x + 2}); & C_{OV}, C_{OV}^{2,\infty} &= O(N^{-\beta}); \\ C_F, C_F^{2,\infty} &= O(N^\beta (\log N)^{\frac{d_x + \beta}{2} + 1}); & C_E &= O(I); & C_T &= O(\sqrt{\log N}). \end{aligned}$$

where $\lambda^{-1} = O(N^\beta \cdot \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof. We use the notation “ $\lesssim$ ” in our derivation when an inequality holds up to a constant factor.

We prove [Lemma J.9](#) with following two steps.

- **Step A: Approximate velocity with constructed function.** We approximate the components $\Phi_1(x, t)$, $\Phi_2(x, t)$, and $\Phi_3(x, t)$ using local polynomials $\Psi_1(x, t)$, $\Psi_2(x, t)$, and $\Psi_3(x, t)$, respectively. Based on the velocity decomposition given in [Lemma J.1](#), we construct an approximation $\Psi(x, t)$ by combining these polynomial components to approximate the full velocity field $u_t(x)$.
- **Step B: Approximate with Transformers.** We leverage the universal approximation of transformers ([Appendix H](#)) to approximate the constructed function Ψ . Based on this approximation, we derive the final velocity approximation rates with the required bounds on model parameters.

By [Lemma J.1](#), the velocity field $u_t(x)$ takes the form

$$u_t(x) = \Phi_1(x, t)^{-1} \cdot \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right),$$

⁵Please see [Appendix H](#) for a detailed proof.

where

$$\begin{aligned}\Phi_1(x, t) &= \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\ \Phi_2(x, t) &= x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1, \\ \Phi_3(x, t) &= \int_{\mathbb{R}^{d_x}} \left(\frac{x - \mu_t \cdot x_1}{\sigma_t}\right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) \, dx_1.\end{aligned}$$

Moreover, by [Lemma J.4](#), the bound on the velocity field in ℓ_∞ -distance follows

$$\begin{aligned}& \|u_t(x)\|_\infty \\ & \leq \frac{|\dot{\mu}_t|}{\mu_t} \cdot \|x\|_\infty + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \cdot (\|x\|_2 + 1) \\ & \lesssim \frac{|\dot{\mu}_t|}{\mu_t} \cdot \sqrt{\log N} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \cdot (\sqrt{\log N} + 1). \quad (\text{By } x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x})\end{aligned} \tag{J.6}$$

Set the transformer network output bound $C_{\mathcal{T}}$ equal to the right-hand side of the expression. Then we are now ready to present the proof of [Lemma J.9](#).

• **Step A: Approximation via Local Polynomial.**

We construct the approximator for $u_t(x)$ based on [Lemma J.5](#), [Lemma J.6](#), and [Lemma J.7](#). Specifically, we define $\Psi(x, t) \in \mathbb{R}^{d_x}$ with each element given by

$$|\Psi(x, t)[i]| := \min \left\{ \frac{\dot{\mu}_t \Psi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Psi_3[i]}{\Psi_1^c}, U \right\}, \tag{J.7}$$

where U is the upper-bound of the ground truth velocity $u_t(x)$ under the sub-Gaussian assumption ([Assumption I.1](#)) and

$$U := \frac{|\dot{\mu}_t|}{\mu_t} \cdot \sqrt{\log N} + C_5 \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \cdot (\sqrt{\log N} + 1). \quad (\text{By } \text{Lemma J.4} \text{ and } x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x})$$

Notice that, for all $i \in [d_x]$, the difference between $\Psi(x, t)[i]$ and $u_t(x)[i]$ follows

$$\begin{aligned}& |u_t(x)[i] - \Psi(x, t)[i]| \\ &= \left| \frac{\dot{\mu}_t \Phi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Phi_3[i]}{\Phi_1} - \frac{\dot{\mu}_t \Psi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Psi_3[i]}{\Psi_1^c} \right| \\ & \quad (\text{By the definition of } u_t \text{ and } \Psi(x, t)) \\ & \leq \underbrace{\left| \frac{\dot{\mu}_t \Phi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Phi_3[i]}{\Psi_1^c} - \frac{\dot{\mu}_t \Psi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Psi_3[i]}{\Psi_1^c} \right|}_{(T_1)} \\ & \quad + \underbrace{\left| \frac{\dot{\mu}_t \Phi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Phi_3[i]}{\Phi_1} - \frac{\dot{\mu}_t \Phi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Phi_3[i]}{\Psi_1^c} \right|}_{(T_2)}. \\ & \quad (\text{By triangle inequality})\end{aligned}$$

Next, we bound (T_1) and (T_2) .

– **Step A.1: Bound term (T₁).** Recall **Definition J.1**. By the definition of ϵ_{low} , we set

$$\Psi_1^c(x, t) := \max \left\{ \Psi_1(x, t), C_3 \cdot N^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} \right\}.$$

By **Lemma J.5**, we have

$$|\Psi_1(x, t) - p_t(x)| \lesssim BN^{-\beta} (\log N)^{\frac{d_x + k_1}{2}}, \quad (\text{J.8})$$

and (J.8) implies

$$p_t(x) - KBN^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} \leq \Psi_1(x, t),$$

for some positive constant K . Next, recall that we consider

$$C_3 N^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} \leq p_t(x).$$

By setting $C_3 = 2KB$, it holds

$$KBN^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} = \frac{C_3}{2} N^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} \leq p_t(x)/2,$$

leading to

$$p_t(x) - p_t(x)/2 \leq p_t(x) - KBN^{-\beta} (\log N)^{\frac{d_x + k_1}{2}} \leq \Psi_1(x, t).$$

As a result, $p_t(x)/2 \leq \Psi_1 \leq \Psi_1^c$ holds.

This allows us to replace the approximator Ψ_1^c with $p_t(x)$ by dropping constant 1/2. Then,

$$\begin{aligned} & (\text{T}_1) \quad \quad \quad (\text{J.9}) \\ &= \left| \frac{\dot{\mu}_t \Phi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Phi_3[i]}{\Psi_1^c} - \frac{\dot{\mu}_t \Psi_2[i]/\mu_t + (\dot{\sigma}_t - \dot{\mu}_t \sigma_t/\mu_t) \Psi_3[i]}{\Psi_1^c} \right| \\ &\leq \frac{2}{p_t} \left| \frac{\dot{\mu}_t}{\mu_t} \cdot (\Phi_2[i] - \Psi_2[i]) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \cdot (\Phi_3[i] - \Psi_3[i]) \right| \quad (\text{By } \Psi_1^c > p_t(x)/2) \\ &\leq \frac{2}{p_t} \frac{|\dot{\mu}_t|}{\mu_t} \cdot |\Phi_2[i] - \Psi_2[i]| + \frac{2}{p_t} \left| \dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right| \cdot |\Phi_3[i] - \Psi_3[i]| \quad (\text{By triangle inequality}) \\ &\lesssim \frac{1}{p_t} \cdot \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right| \right) \cdot BN^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}} \quad (\text{By Lemma J.6 and Lemma J.7}) \\ &\leq \frac{1}{p_t} \cdot \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot BN^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}} \quad (\text{By } \sigma_t \in [0, 1]) \\ &\lesssim \frac{1}{p_t} \cdot BN^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}. \quad (\text{By Assumption I.2}) \end{aligned}$$

Next, we bound (T₂).

– **Step A.2: Bound term (T₂).** By **Lemma J.4** and $\|x\|_2 \lesssim \sqrt{\log N}$, it holds

$$\begin{aligned} & |u_t(x)[i]| \\ &\leq \frac{|\dot{\mu}_t|}{\mu_t} \cdot \sqrt{\log N} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \cdot (\sqrt{\log N} + 1) \\ &= \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot \sqrt{\log N} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right|, \end{aligned}$$

for all $i \in [d_x]$. Next, by the decomposition of velocity in [Lemma J.1](#), it holds

$$\frac{\dot{\mu}_t}{\mu_t} \Phi_2[i] + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3[i] \lesssim \Phi_1 \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot \sqrt{\log N} + \Phi_1 \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right|. \quad (\text{J.10})$$

Therefore,

$$\begin{aligned} & (\text{T}_2) \\ & \leq \left| \frac{\dot{\mu}_t}{\mu_t} \Phi_2[i] + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3[i] \right| \cdot \left| \frac{1}{\Phi_1} - \frac{1}{\Psi_1} \right| \\ & \lesssim \Phi_1 \left(\left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \sqrt{\log N} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot \left| \frac{1}{\Phi_1} - \frac{1}{\Psi_1} \right| \quad (\text{By (J.10)}) \\ & = \frac{1}{\Psi_1} \cdot \left(\left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \sqrt{\log N} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot |\Phi_1 - \Psi_1| \\ & \quad (\text{By factoring out } 1/\Phi_1 \text{ and } 1/\Psi_1) \\ & \leq \frac{1}{p_t} \cdot \left(\left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \sqrt{\log N} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot |\Phi_1 - \Psi_1| \quad (\text{By } \Psi_1^c > p_t(x)/2) \\ & \lesssim \frac{1}{p_t} \cdot \left(\left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \sqrt{\log N} + \left| \frac{\dot{\sigma}_t}{\sigma_t} - \frac{\dot{\mu}_t}{\mu_t} \right| \right) \cdot BN^{-\beta} (\log N)^{\frac{d_x+k_1}{2}} \\ & \quad (\text{By Lemma J.5}) \end{aligned} \quad (\text{J.11})$$

Combining [\(J.9\)](#) and [\(J.11\)](#), we have

$$\begin{aligned} & p_t \cdot |u_t[i] - \Psi[i]| \\ & \leq (\text{T}_1) \cdot p_t + (\text{T}_2) \cdot p_t \\ & \lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right) BN^{-\beta} (\log N)^{\frac{d_x+k_1+1}{2}}, \quad (\text{By (J.9) and (J.11)}) \end{aligned} \quad (\text{J.12})$$

for all $i \in [d_x]$.

Therefore,

$$\begin{aligned} & p_t^2 \cdot \|u_t(x) - \Psi(x, t)\|_2^2 \\ & \leq p_t^2 \cdot d_x \|u_t(x) - \Psi(x, t)\|_\infty^2 \quad (\text{By } \|\cdot\|_2 \leq d_x \|\cdot\|_\infty) \\ & \lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 B^2 N^{-2\beta} (\log N)^{d_x+k_1+1}. \quad (\text{By (J.12)}) \end{aligned} \quad (\text{J.13})$$

• Step B: Approximation with Transformer.

By [Lemma J.8](#), there exists a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,r,s}$ such that

$$\int \int \|u_\theta(x, t) - \Psi(x, t)\|_2^2 dx dt \leq \epsilon^2. \quad (\text{J.14})$$

By setting $\epsilon := N^{-\beta}$, it holds

$$\begin{aligned} & \int \int p_t^2 \cdot \|u_t(x) - u_\theta(x, t)\|_2^2 dx dt \\ & \leq \int \int p_t^2 \cdot \|u_t(x) - \Psi(x, t)\|_2^2 dx dt + \int \int p_t^2 \cdot \|\Psi(x, t) - u_\theta(x, t)\|_2^2 dx dt \\ & \quad (\text{By triangle inequality}) \end{aligned}$$

$$\begin{aligned}
&\leq \int \int p_t^2 \cdot \|u_t(x) - \Psi(x, t)\|_2^2 dx dt + \int \int \|\Psi(x, t) - u_\theta(x, t)\|_2^2 dx dt \quad (\text{By } 0 \leq p_t(x) \leq 1) \\
&\lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 B^2 N^{-2\beta} (\log N)^{d_x + k_1 + 1} \int \int dx dt + \int \int \|\Psi(x, t) - u_\theta(x, t)\|_2^2 dx dt \\
&\hspace{15em} (\text{By (J.13)}) \\
&\leq \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 B^2 N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1} + \int \int \|\Psi(x, t) - u_\theta(x, t)\|_2^2 dx dt \\
&\hspace{15em} (\text{By } \|x\|_\infty \leq C_x \sqrt{\log N} \text{ and } t \in [0, 1]) \\
&\lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 B^2 N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1}. \hspace{5em} (\text{By (J.14) and } \epsilon = N^{-\beta})
\end{aligned}$$

By (J.6) and $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$, we have

$$U = O(\sqrt{\log N}),$$

and by (J.12) we have

$$|u_t[i] - \Psi[i]| = O(N^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}).$$

This implies

$$\|\Psi(x, t)\|_2 = O(\sqrt{\log N} + N^{-\beta} (\log N)^{\frac{d_x + k_1 + 1}{2}}).$$

We take a looser bound on $\Psi(x, t)$ such that it holds for all d_x :

$$\|\Psi(x, t)\|_2 \leq d_x \|\Psi(x, t)\|_\infty = O((\log N)^{\frac{d_x + k_1 + 1}{2}}).$$

Then, the parameter bounds follow Lemma J.8 with $\epsilon = N^{-\beta}$.

This completes the proof. □

J.3 Main Proof of Theorem I.1

We establish the velocity approximation with transformers in Lemma J.9. However, it is valid under two settings: (i) the bounded domain $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$ with some constant $C_x(\beta, C_2)$ (ii) the mild and high density region $p_t(x) \geq \epsilon_{\text{low}}$. To obtain general approximation results, we introduce two additional lemmas to tackle the uncontrolled region.

Lemma J.10 (Truncation of x , Modified from Lemma A.1 of [Fu et al., 2024]). Assume Assumption I.1. Then, for any $R_4 > 1$, $t > 0$, the following hold

$$\begin{aligned}
\int_{\|x\|_\infty > R_4} p_t(x) dx &\lesssim R_4^{d_x - 2} \exp\left(-\frac{C_2 R_4^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right), \\
\int_{\|x\|_\infty > R_4} \|u_t(x)\|_2^2 \cdot p_t(x) dx &\lesssim R_4^{d_x} \exp\left(-\frac{C_2 R_4^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right).
\end{aligned}$$

Proof. For the first inequality, it follows

$$\begin{aligned}
&\int_{\|x\|_\infty > R_4} p_t(x) dx \\
&\lesssim \int_{\|x\|_\infty > R_4} \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx \hspace{5em} (\text{By Lemma J.3})
\end{aligned}$$

$$\begin{aligned}
&\leq \int_{\|x\|_2 > R_4} \exp\left(-\frac{C_2\|x\|_2^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) dx && (\text{By } \|x\|_2 \geq \|x\|_\infty) \\
&\lesssim R_4^{d_x-2} \exp\left(-\frac{C_2R_4^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right). && (\text{By Lemma D.2 and dropping constant terms})
\end{aligned}$$

For the second inequality, it follows

$$\begin{aligned}
&\int_{\|x\|_\infty \geq R_4} \|u_t(x)\|_2^2 \cdot p_t(x) dx \\
&\lesssim \int_{\|x\|_\infty \geq R_4} \|u_t(x)\|_2^2 \cdot \exp\left(-\frac{C_2\|x\|_2^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) dx && (\text{By Lemma J.3}) \\
&\lesssim \int_{\|x\|_\infty \geq R_4} \left(\frac{|\dot{\mu}_t|}{\mu_t} \cdot \|x\|_\infty + \left|\frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t}\right| \cdot (\|x\|_2 + 1)\right)^2 \exp\left(-\frac{C_2\|x\|_2^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) dx && (\text{By Lemma J.4}) \\
&\lesssim \int_{\|x\|_\infty \geq R_4} \|x\|_2^2 \exp\left(-\frac{C_2\|x\|_2^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) dx && (\text{By Assumption I.2}) \\
&\leq \int_{\|x\|_2 \geq R_4} \|x\|_2^2 \exp\left(-\frac{C_2\|x\|_2^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) dx && (\text{By } \|x\|_2 \geq \|x\|_\infty) \\
&\lesssim R_4^{d_x} \exp\left(-\frac{C_2R_4^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right). && (\text{By Lemma D.2})
\end{aligned}$$

This completes the proof. $\square$

Lemma J.11 (Bound on Low-Density Region, Modified from Lemma A.2 of [Fu et al., 2024]). Assume **Assumption I.1**. Then, for any $R_5, \epsilon_{\text{low}} > 0$, the following two inequalities hold

$$\begin{aligned}
&\int_{\|x\|_\infty \leq R_5} \mathbb{1}\{|p_t(x)| < \epsilon_{\text{low}}\} \cdot p_t(x) dx \leq R_5^{d_x} \cdot \epsilon_{\text{low}}, \\
&\int_{\|x\|_\infty \leq R_5} \mathbb{1}\{|p_t(x)| < \epsilon_{\text{low}}\} \cdot \|u_t(x)\|_2^2 \cdot p_t(x) dx \lesssim R_5^{d_x+2} \cdot \epsilon_{\text{low}}.
\end{aligned}$$

Proof. The proof for the first inequality is identical to [Fu et al., 2024].

For the second inequality, it follows,

$$\begin{aligned}
&\int_{\|x\|_\infty \leq R_5} \mathbb{1}\{|p_t(x)| < \epsilon_{\text{low}}\} \cdot \|u_t(x)\|_2^2 \cdot p_t(x) dx \\
&\lesssim \int_{\|x\|_\infty \leq R_5} \mathbb{1}\{|p_t(x)| < \epsilon_{\text{low}}\} \cdot \left(\frac{|\dot{\mu}_t|}{\mu_t} \cdot \|x\|_\infty + \left|\frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t}\right| \cdot (\|x\|_2 + 1)\right)^2 \cdot p_t(x) dx && (\text{By Lemma J.4}) \\
&\leq \epsilon_{\text{low}} \int_{\|x\|_\infty \leq R_5} \left(\frac{|\dot{\mu}_t|}{\mu_t} \cdot \|x\|_\infty + \left|\frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t}\right| \cdot (\|x\|_2 + 1)\right)^2 dx \\
&\lesssim R_5^{d_x+2} \cdot \epsilon_{\text{low}}. && (\text{By Assumption I.2})
\end{aligned}$$

This completes the proof. $\square$

Next, we present the formal proof of **Theorem I.1**.

Theorem J.1 (**Theorem I.1** Restated: Velocity Approximation with Transformers under Generic Hölder Smoothness). Assume **Assumption I.1** and **Assumption I.2**. For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$. Then, for all

$t \in [t_0, T]$ with $t_0, T \in (0, 1)$, there exists a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u_t(x) - u_\theta(x, t)\|_2^2 \cdot p_t(x) dx dt = O\left(B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{\beta}{2} + 1}\right).$$

Furthermore, the parameter bounds in transformer network $\mathcal{T}_R^{h,s,r}$ satisfy

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

where $\lambda^{-1} = O(N^\beta \cdot \log N)^{4d+3}$ is the inverse-temperature scaling in the softmax function and $O(\cdot)$ hides all polynomial factors depending on $d_x, d, L, \beta, C_1, C_2$.

Proof of Theorem I.1. Let $R_6 := C_x \sqrt{\log N}$ and $C_x := \sqrt{4\beta(\mu_t^2 + C_2 \sigma_t^2)}/C_2$. Further, we have

$$C_{\mathcal{T}} = O(\sqrt{\log N}); \quad \epsilon_{\text{low}} = C_3 N^{-\beta} (\log N)^{(d_x + k_1)/2}. \quad (\text{By Lemma J.9})$$

First, we decompose the target into three components and bound each of them

$$\begin{aligned} & \int_{t_0}^T \int \|u_\theta - u_t\|_2^2 \cdot p_t(x) dx dt \\ &= \underbrace{\int_{t_0}^T \int_{\|x\|_\infty > R_6} \|u_\theta - u_t\|_2^2 p_t(x) dx dt}_{(T_1)} + \underbrace{\int_{t_0}^T \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \|u_\theta - u_t\|_2^2 p_t(x) dx dt}_{(T_2)} \\ &+ \underbrace{\int_{t_0}^T \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) \geq \epsilon_{\text{low}}\} \|u_\theta - u_t\|_2^2 p_t(x) dx dt}_{(T_3)}. \end{aligned}$$

• **Bound on (T_1) .** It holds

$$\begin{aligned} & \int_{\|x\|_\infty > R_6} \|u_\theta - u_t\|_2^2 \cdot p_t(x) dx \\ &\leq 2 \int_{\|x\|_\infty > R_6} \|u_\theta\|_2^2 \cdot p_t(x) dx + 2 \int_{\|x\|_\infty > R_6} \|u_t\|_2^2 \cdot p_t(x) dx \quad (\text{By expanding } \ell_2\text{-norm}) \\ &\leq 2d_x \int_{\|x\|_\infty > R_6} \|u_\theta\|_\infty^2 \cdot p_t(x) dx + 2 \int_{\|x\|_\infty > R_6} \|u_t\|_2^2 \cdot p_t(x) dx \quad (\text{By } \|\cdot\|_2^2 \leq d_x \|\cdot\|_\infty^2) \\ &\lesssim \underbrace{\int_{\|x\|_\infty > R_6} \log N \cdot p_t(x) dx}_{(T_{1.1})} + \underbrace{\int_{\|x\|_\infty > R_6} \|u_t\|_2^2 \cdot p_t(x) dx}_{(T_{1.2})}. \end{aligned}$$

(By $C_{\mathcal{T}} = O(\sqrt{\log N})$ from Lemma J.9)

We bound $(T_{1.1})$ by

$$\begin{aligned} & (T_{1.1}) \\ &= \log N \cdot \int_{\|x\|_\infty > R_6} p_t(x) dx \\ &\lesssim \log N \cdot R_6^{d_x - 2} \exp\left(-\frac{C_2 R_6^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) \quad (\text{By Lemma J.10}) \end{aligned}$$

$$\lesssim \log N \cdot (\log N)^{\frac{d_x-2}{2}} N^{-\beta}.$$

(By the choice of $R_6 = C_x \sqrt{\log N}$ and $C_x = \sqrt{2\beta(\mu_t^2 + C_2\sigma_t^2)}/C_2$)

We bound (T_{1.2}) by

$$\begin{aligned} & \text{(T}_{1.2}\text{)} \\ &= \int_{\|x\|_\infty > R_6} \|u_t\|_2^2 \cdot p_t(x) \, dx \\ &\lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot R_6^{d_x} \exp\left(-\frac{C_2 R_6^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) \quad (\text{By Lemma J.10}) \\ &\lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot (\log N)^{\frac{d_x}{2}} N^{-\beta}. \end{aligned}$$

(By the choice of $R_6 = C_x \sqrt{\log N}$ and $C_x = \sqrt{2\beta(\mu_t^2 + C_2\sigma_t^2)}/C_2$)

Therefore,

$$(T_1) \lesssim (T_{1.1}) + (T_{1.2}) \lesssim \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot (\log N)^{\frac{d_x}{2}} \cdot N^{-\beta}.$$

• **Bound on (T₂).** It holds

$$\begin{aligned} & \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \cdot \|u_\theta - u_t\|_2^2 \cdot p_t(x) \, dx \\ &\leq 2 \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \cdot (\|u_\theta\|_2^2 + \|u_t\|_2^2) \cdot p_t(x) \, dx \quad (\text{By expanding } \ell_2\text{-norm}) \\ &\leq 2 \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \cdot (d_x \cdot \|u_\theta\|_\infty^2 + \|u_t\|_2^2) \cdot p_t(x) \, dx \quad (\text{By } \|\cdot\|_2^2 \leq d_x \|\cdot\|_\infty^2) \\ &\lesssim \underbrace{\int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t < \epsilon_{\text{low}}\} \cdot \|u_\theta\|_\infty^2 \cdot p_t(x) \, dx}_{(T_{2.1})} + \underbrace{\int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t < \epsilon_{\text{low}}\} \cdot \|u_t\|_2^2 \cdot p_t(x) \, dx}_{(T_{2.2})}. \end{aligned}$$

We bound (T_{2.1}) by

$$\begin{aligned} & \text{(T}_{2.1}\text{)} \\ &= \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \cdot \|u_\theta(x)\|_\infty^2 \cdot p_t(x) \, dx \\ &\lesssim \log N \cdot \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \cdot p_t(x) \, dx \quad (\text{By } C_T = O(\sqrt{\log N}) \text{ from Lemma J.9}) \\ &\lesssim \log N \cdot \epsilon_{\text{low}} R_6^{d_x} \quad (\text{By Lemma J.11}) \\ &\lesssim \log N \cdot (\log N)^{\frac{d_x}{2}} \cdot \epsilon_{\text{low}} \quad (\text{By the choice of } R_6 = C_x \sqrt{\log N} \text{ and } C_x = \sqrt{2\beta(\mu_t^2 + C_2\sigma_t^2)}/C_2) \\ &\lesssim \log N (\log N)^{\frac{d_x}{2}} \cdot N^{-\beta} (\log N)^{\frac{d_x+k_1}{2}} \quad (\text{By the choice of } \epsilon_{\text{low}} = C_3 N^{-\beta} (\log N)^{\frac{d_x+k_1}{2}}) \end{aligned}$$

We bound (T_{2.2}) by

$$\begin{aligned} & \text{(T}_{2.2}\text{)} \\ &= \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) < \epsilon_{\text{low}}\} \cdot \|u_t\|_2^2 \cdot p_t(x) \, dx \end{aligned}$$

$$\begin{aligned}
&\lesssim \epsilon_{\text{low}} R_6^{d_x+2} && \text{(By Lemma J.11)} \\
&\lesssim \epsilon_{\text{low}} (\log N)^{\frac{d_x+2}{2}} && \text{(By the choice of } R_6 = C_x \sqrt{\log N} \text{ and } C_x = \sqrt{2\beta(\mu_t^2 + C_2\sigma_t^2)}/C_2) \\
&\leq N^{-\beta} (\log N)^{\frac{d_x+k_1}{2}} \cdot (\log N)^{\frac{d_x+2}{2}}. && \text{(By the choice of } \epsilon_{\text{low}} = C_3 N^{-\beta} (\log N)^{\frac{d_x+k_1}{2}})
\end{aligned}$$

Therefore,

$$(T_2) \lesssim (T_{2.1}) + (T_{2.2}) \lesssim N^{-\beta} (\log N)^{d_x + \frac{k_1}{2} + 1}.$$

• **Bound on (T_3) .** We bound term (T_3) by

$$\begin{aligned}
&(T_3) \\
&= \int_{t_0}^T \int_{\|x\|_\infty \leq R_6} \mathbb{1}\{p_t(x) \geq \epsilon_{\text{low}}\} \cdot \|u_\theta - u_t\|_2^2 \cdot p_t(x) dx dt \\
&= \int_{t_0}^T \int_{\|x\|_\infty \leq R_6} \frac{1}{p_t} \mathbb{1}\{p_t(x) \geq \epsilon_{\text{low}}\} \cdot d_x \|u_\theta - u_t\|_2^2 \cdot (p_t(x))^2 dx dt && \text{(By multiplying } p_t/p_t) \\
&\leq \int_{t_0}^T \int_{\|x\|_\infty \leq R_6} \frac{1}{\epsilon_{\text{low}}} \mathbb{1}\{p_t(x) \geq \epsilon_{\text{low}}\} \cdot d_x \|u_\theta - u_t\|_2^2 \cdot (p_t(x))^2 dx dt && \text{(By } 1/p_t \leq 1/\epsilon_{\text{low}}) \\
&\leq \frac{d_x}{\epsilon_{\text{low}}} \cdot \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot B^2 N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1} && \text{(By Lemma J.9)} \\
&\lesssim N^\beta (\log N)^{\frac{-(d_x+k_1)}{2}} \cdot \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot B^2 N^{-2\beta} (\log N)^{\frac{3d_x}{2} + k_1 + 1} \\
&\hspace{15em} \text{(By the choice of } \epsilon_{\text{low}}) \\
&= \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{k_1}{2} + 1}.
\end{aligned}$$

By the upper-bound on (T_1) , (T_2) and (T_3) , we have

$$\begin{aligned}
&\int_{t_0}^T \int \|u_t(x) - u_\theta(x, t)\|_2^2 \cdot p_t dx dt \\
&\lesssim (T_1) + (T_2) + (T_3) && \text{(By } t_0, T \in (0, 1)) \\
&= \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot O \left(B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{k_1}{2} + 1} \right) \\
&\leq \left(\frac{|\dot{\mu}_t|}{\mu_t} + \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \right)^2 \cdot O \left(B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{\beta}{2} + 1} \right) && \text{(By } k_1 \leq \beta) \\
&\leq O \left(B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{\beta}{2} + 1} \right). && \text{(By Assumption I.2)}
\end{aligned}$$

Furthermore, the transformer parameter bounds follow Lemma J.9.

This completes the proof. $\square$

K Proof of Theorem I.2

In this section, we derive a tighter error bound for velocity approximation using transformers.

Organizations. Appendix K.1 introduces auxiliary lemmas. Appendix K.2 establishes a bound on the velocity approximation error over a bounded domain by applying the universal approximation of transformers. Appendix K.3 presents the main proof by incorporating the bounded-domain approximation error and controlling the unbounded region using the sub-Gaussian assumption.

K.1 Auxiliary Lemmas

In this section, we introduce auxiliary lemmas for velocity field approximation. Specifically, Lemma K.1 applies a stronger Hölder assumption to decompose the density function $p_t(x)$. Lemma K.2 further decomposes the velocity into two components, differing from the decomposition under a generic Hölder assumption. Then, Lemma K.3 and Lemma K.4 establish upper and lower bounds for the decomposed components and the velocity in ℓ_∞ -distance, respectively.

We begin with the density function decomposition.

Lemma K.1 (Density Function Decomposition, Lemma B.1 of [Fu et al., 2024]). Assume Assumption I.3. Then, the density function $p_t(x)$ and $\nabla \log p_t(x)$ follow the decomposition:

$$p_t(x) = \frac{1}{(\mu_t^2 + C_2 \cdot \sigma_t^2)^{d_x/2}} \exp\left(\frac{-C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) h(x, t),$$

$$\nabla \log p_t(x) = \frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\nabla h(x, t)}{h(x, t)},$$

where $h(x, t) := \int \frac{f(x_1)}{(2\pi)^{d_x/2} \hat{\sigma}_t^{d_x}} \exp\left(-\frac{\|x_1 - \hat{\mu}_t x\|_2^2}{2\hat{\sigma}_t^2}\right) dx_1$, $\hat{\sigma}_t := \frac{\sigma_t}{(\mu_t^2 + C_2 \sigma_t^2)^{1/2}}$ and $\hat{\mu}_t := \frac{\mu_t}{(\mu_t^2 + C_2 \sigma_t^2)^{1/2}}$.

Then, we give the velocity field decomposition.

Lemma K.2 (Velocity Decomposition under Stronger Hölder Smoothness Assumption). Assume Assumption I.3. Then, the velocity field $u_t(x)$ follows the decomposition:

$$u_t(x) = \frac{\dot{\mu}_t}{\mu_t} x - (\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t}) \left(\frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\nabla h(x, t)}{h(x, t)} \right).$$

Remark K.1. The key aspect of Lemma K.2 is the velocity field $u_t(x)$ having a denominator bounded away from zero. Specifically, we apply $f(x_1) \geq C$ to derive the lower bound on $h(x, t)$ (Lemma K.3). This removes the need to impose an additional lower threshold on the density function approximator. In contrast, under Assumption I.1, the approximator is constrained to stay above the threshold ϵ_{low} to prevent explosion, and therefore leads to slower approximation rate.

Proof. Our proof builds on [Fu et al., 2024].

By Lemma J.1, the velocity field $u_t(x)$ has the form

$$u_t(x) = \Phi_1(x, t)^{-1} \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2(x, t) + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3(x, t) \right),$$

where

$$\Phi_1(x, t) = \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) dx_1,$$

$$\Phi_2(x, t) = x \int_{\mathbb{R}^{d_x}} \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp\left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2}\right) \cdot q(x_1) dx_1,$$

$$\Phi_3(x, t) = \int_{\mathbb{R}^{d_x}} \left(\frac{x - \mu_t \cdot x_1}{\sigma_t} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp \left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2} \right) \cdot q(x_1) \, dx_1.$$

Furthermore, we have

$$\begin{aligned} & \sigma_t \nabla p_t(x) \\ &= -\sigma_t \int \left(\frac{x - \mu_t \cdot x_1}{\sigma_t^2} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp \left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2} \right) q(x_1) \, dx_1 \\ &= -\int \left(\frac{x - \mu_t \cdot x_1}{\sigma_t} \right) \cdot \frac{1}{\sigma_t^{d_x} (2\pi)^{d_x/2}} \exp \left(-\frac{\|x - \mu_t \cdot x_1\|^2}{2\sigma_t^2} \right) q(x_1) \, dx_1 \\ &= -\Phi_3(x, t). \end{aligned}$$

Therefore,

$$\begin{aligned} & u_t(x) \\ &= \Phi_1^{-1} \left(\frac{\dot{\mu}_t}{\mu_t} \cdot \Phi_2 + \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \Phi_3 \right) \\ &= \frac{\dot{\mu}_t}{\mu_t} x - \left(\dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right) \sigma_t \nabla \log p_t \quad (\text{By } \Phi_2 = x \Phi_1 \text{ and } \Phi_3 = -\sigma_t \nabla p_t) \\ &= \frac{\dot{\mu}_t}{\mu_t} x - \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\nabla h(x, t)}{h(x, t)} \right). \quad (\text{By Lemma K.1}) \end{aligned}$$

This completes the proof. $\square$

The next lemma bounds $h(x, t)$.

Lemma K.3 (Lemma B.8 of [Fu et al., 2024]). Assume **Assumption I.3**. Then, it holds

$$C_1 \leq h(x, t) \leq B, \quad \left\| \frac{\hat{\sigma}_t}{\hat{\mu}_t} \nabla h(x, t) \right\|_\infty \leq \sqrt{\frac{2}{\pi}} B.$$

Lemma K.3 ensures that $h(x, t)$ remains bounded above and below by a constant. As a result, $u_t(x)$ stays finite for all x . This eliminates the need for an additional threshold ϵ_{low} (**Definition J.1**) in the constructed approximator to prevent divergence, leading to a faster approximation rate.

Bound on Velocity Field. We give the ℓ_∞ -bound on $u_t(x)$ under stronger Hölder assumption.

Lemma K.4 (Bounds on Velocity Field). Assume **Assumption I.3**. Then, there exist a positive constant C_6 such that

$$\|u_t(x)\|_\infty \leq \left| \frac{\dot{\mu}_t}{\mu_t} + \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{C_2}{\mu_t^2 + C_2 \sigma_t^2} \right) \right| \|x\|_\infty + C_6 \left| \dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right|.$$

Proof. Recalling from **Lemma K.2** and **Lemma K.3**, the velocity field has the expression

$$u_t(x) = \frac{\dot{\mu}_t}{\mu_t} x - \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\nabla h(x, t)}{h(x, t)} \right),$$

where $\hat{\sigma}_t = \sigma_t / \sqrt{\mu_t^2 + C_2 \sigma_t^2}$, $\hat{\mu}_t = \mu_t / (\mu_t^2 + C_2 \sigma_t^2)$ and

$$h(x, t) = \int f(x_1) \frac{1}{(2\pi)^{d_x/2} \cdot \hat{\sigma}_t^{d_x}} \cdot \exp \left(-\frac{\|x_1 - \hat{\mu}_t \cdot x\|_2^2}{2\hat{\sigma}_t} \right) \, dx_1.$$

By Lemma K.3 and Assumption I.2, it holds

$$\left\| \frac{\nabla h(x, t)}{h(x, t)} \right\|_{\infty} \leq \frac{\hat{\mu}_t}{\hat{\sigma}_t} \cdot \sqrt{\frac{2}{\pi}} B C_1 = O\left(\frac{1}{\sigma_t}\right). \quad (\text{By Lemma K.3})$$

Therefore,

$$\begin{aligned} & \|u_t(x)\|_{\infty} \\ & \leq \left| \frac{\dot{\mu}_t}{\mu_t} + \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{C_2}{\mu_t^2 + C_2 \sigma_t^2} \right) \right| \cdot \|x\|_{\infty} + \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right| \cdot \left\| \frac{\nabla h(x, t)}{h(x, t)} \right\|_{\infty} \\ & \quad (\text{By triangle inequality}) \\ & \leq \left| \frac{\dot{\mu}_t}{\mu_t} + \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{C_2}{\mu_t^2 + C_2 \sigma_t^2} \right) \right| \cdot \|x\|_{\infty} + C_6 \left| \dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right|, \quad (\text{By ((By Lemma K.3))}) \end{aligned}$$

for some positive constant C_6 .

This completes the proof. $\square$

K.2 Velocity Approximation on Bounded Domain

In this section, we approximate the velocity field $u_t(x)$ using transformers in two steps. The first step constructs two compactly supported continuous functions, $Q_1(x, t)$ and $Q_2(x, t)$, as approximations of $h(x, t)$ and $\nabla h(x, t)$ (Lemma K.5 and Lemma K.6). The second step applies the universal approximation of transformers to approximate $Q_1(x, t)$ and $Q_2(x, t)$ (Lemma K.7). Combining these steps, we present the velocity approximation on a bounded domain in Lemma K.8.

Before proceeding, we reiterate the expression of decomposed velocity under Assumption I.3.

$$u_t(x) = \frac{\dot{\mu}_t}{\mu_t} x - \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\nabla h(x, t)}{h(x, t)} \right).$$

Then, we construct two local polynomials as the approximators for $h(x, t)$, and $\nabla h(x, t)$.

Approximation of $h(x, t)$ and $\nabla h(x, t)$. The differences between

$$h(x, t) = \int f(x_1) \cdot \frac{1}{(2\pi)^{d_x/2} \cdot \hat{\sigma}_t^{d_x}} \cdot \exp\left(-\frac{\|x_1 - \hat{\mu}_t \cdot x\|_2^2}{2\hat{\sigma}_t}\right) dx_1,$$

and

$$p_t(x) = \int q(x_1) \cdot \frac{1}{(2\pi)^{d_x/2} \cdot \sigma_t^{d_x}} \cdot \exp\left(-\frac{\|x_1 - \mu_t \cdot x\|_2^2}{2\sigma_t}\right) dx_1,$$

lie in (i) the target function $f(x_1)$ and $q(x_1)$ (ii) the path coefficients $\hat{\sigma}_t$, $\hat{\mu}_t$ and σ_t , μ_t .

We define local polynomial $\Psi_1(x, t)$ as the approximator for $p_t(x)$ in (J.4). Given the differences between h and p_t , the construction of an approximator for $h(x, t)$ follows the formulation of Ψ_1 .

Formally, we approximate $h(x, t)$ around x with:

$$Q_1(x, t) := \sum_{v \in [N]^{d_x}} \sum_{\|n_x\|_1 \leq k_1} \frac{R_B^{\|n_x\|_1}}{n_x!} \frac{\partial^{n_x} f}{\partial x^{n_x}} \Big|_{x=R_B(\frac{v}{N}-\frac{1}{2})} g_1(x, n_x, v, t), \quad (\text{K.1})$$

where $n_x \in \mathbb{Z}^{d_x}$ is a multi-index, $R_B > 0$ is a constant depending on the Hölder ball radius B ,

- $g_1(x, n_x, v, t) := \prod_{i=1}^{d_x} g_2(x[i], n_x[i], v[i], k_2)$, and

$$\bullet \quad g_2(x[i], n_x[i], v[i], k_2) := \frac{1}{\sigma_t \sqrt{2\pi}} \int \left(\frac{x_1[i]}{R_B} + \frac{1}{2} - \frac{v[i]}{N} \right)^{n_x[i]} \frac{1}{k_2!} \left(-\frac{|x[i] - \hat{\mu}_t x_1[i]|^2}{2\hat{\sigma}_t} \right)^{k_2} dx_1.$$

Remark K.2. Given the differences between $h(x, t)$ and $p_t(x)$, we replace (i) $\partial^{n_x} \Phi_1 / \partial x^{n_x}$ with $\partial^{n_x} f / \partial x^{n_x}$ (ii) σ_t and μ_t with $\hat{\sigma}_t$ and $\hat{\mu}_t$ respectively. Then, the formulation of $Q_1(x, t)$ follows constructions identical to the density function approximator $\Psi_1(x, t)$.

Remark K.3. When the context is clear, we refer to $Q_1(x, t)$ as a local polynomial and distinguish it from $\Psi_1(x, t)$. The generic Hölder assumption ([Assumption I.1](#)) applies to $\Psi_1(x, t)$, while the stronger Hölder assumption ([Assumption I.3](#)) applies to $Q_1(x, t)$.

Then, we approximate $h(x, t)$ using $Q_1(x, t)$.

Lemma K.5 (Approximate of $h(x, t)$, Lemma B.4 of [\[Fu et al., 2024\]](#)). Assume [Assumption I.3](#). Let $Q_1(x, t)$ be the approximator of $h(x, t)$, and $C_x(d_x, \beta, C_1, C_2)$ be a positive constant. Then, for any $t \in [0, 1]$ and $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$, it holds

$$|Q_1(x, t) - h(x, t)| \lesssim BN^{-\beta} (\log N)^{\frac{k_1}{2}}.$$

Based on the approximation of $h(x, t)$ using local polynomial $Q_1(x, t)$, we construct a approximator of $\nabla h(x, t)$ following similar formulation

Definition K.1 (Approximator of $\nabla h(x, t)$). We define $Q_2(x, t)$ as the approximator of $\nabla h(x, t)$, with each component $Q_2[i]$ following the form of local polynomial presented in ([K.1](#)).

Then, we approximate $h'(x, t)$ and $\nabla h(x, t)$ with $Q_2(x, t)$.

Lemma K.6 (Approximate $\nabla h(x, t)$, Lemma B.6 of [\[Fu et al., 2024\]](#)). Assume [Assumption I.3](#). Let $C_x(d_x, \beta, C_1, C_2)$ be a positive constant. Then, for all $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$, $i \in [d_x]$ and $t > 0$, it holds

$$\left| Q_2(x, t)[i] - \frac{\hat{\sigma}_t}{\hat{\mu}_t} \cdot \nabla h(x, t)[i] \right| \lesssim BN^{-\beta} (\log N)^{\frac{k_1+1}{2}}.$$

Approximate Velocity Approximator with Transformers Before deriving the velocity approximation with transformers on a bounded domain, we first approximate the velocity approximator constructed with $Q_1(x, t)$ and $Q_2(x, t)$ using transformers.

Lemma K.7 (Approximate Velocity Approximators with Transformers Network). Assume [Assumption I.3](#). Let C_x be a positive constant dependent on d_x, β, C_1 and C_2 . Then, for any $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$ and $t \in [0, 1]$, there exist a transformer $\mathcal{T} \in \mathcal{T}_R^{h,s,r}$ such that,

$$\int_0^1 \int \left\| \mathcal{T}(x, t) - \frac{\dot{\mu}_t}{\mu_t} x + \left(\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right) \left(\frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\hat{\mu}_t \nabla Q_2[i]}{\hat{\sigma}_t Q_1} \right) \right\|_2^2 dx dt \leq \epsilon^2.$$

Further, the parameter bounds in the transformer network class follows [Lemma J.8](#).

Proof. The proof closely follows [Lemma J.8](#). □

Approximate Velocity with Transformers on Bounded Domain. We incorporate the approximations with Q_1 , Q_2 and $\mathcal{T}(x, t)$ to derive the velocity approximation on a bounded domain.

Lemma K.8 (Velocity Approximation with Transformers on Bounded Domain). Assume [Assumption I.3](#). Then, for any $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$ and $t \in [t_0, T]$ with a positive constant $C_x(d_x, \beta, C_1, C_2)$ and $t_0, T \in (0, 1)$, there exist a $u_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{t_0}^T \int_{\|x\|_\infty \leq C_x \sqrt{\log N}} \|u_t(x) - u_\theta(x, t)\|_2^2 p_t(x) dx dt \lesssim B^2 N^{-2\beta} (\log N)^{k_1+d_x}.$$

Further, the parameter bounds in the transformer network class follows [Lemma J.9](#).

Proof. Building upon [\[Hu et al., 2025c, Fu et al., 2024\]](#), we prove [Lemma K.8](#) with two steps.

- **Step 1: Approximate velocity with constructed function.** We approximate the decomposed velocity field ([Lemma K.2](#)) and its components with approximator $Q_1(x, t)$ and $Q_2(x, t)$.
- **Step 2: Approximate with transformers.** We apply the universal approximation of transformers presented in [Appendix H](#) to approximate the constructed function in Step 1.

Before proceeding, we recall some previous lemmas to prepare our proof.

By [Lemma K.2](#), the velocity follows the decomposition under [Assumption I.3](#):

$$u_t(x) = \frac{\dot{\mu}_t}{\mu_t}x - (\dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t}) \left(\frac{-C_2x}{\mu_t^2 + C_2\sigma_t^2} + \frac{\nabla h(x, t)}{h(x, t)} \right),$$

where $\hat{\sigma}_t = \sigma_t/(\mu_t^2 + C_2\sigma_t^2)^{1/2}$, $\hat{\mu}_t = \mu_t/(\mu_t^2 + C_2\sigma_t^2)$ and

$$h(x, t) = \int f(x_1) \frac{1}{(2\pi)^{d_x/2} \cdot \hat{\sigma}_t^{d_x}} \cdot \exp\left(-\frac{\|x_1 - \hat{\mu}_t \cdot x\|_2^2}{2\hat{\sigma}_t}\right) dx_1.$$

Furthermore, by [Lemma K.4](#), the bound on $u_t(x)$ in ℓ_∞ -distance follows

$$\|u_t(x)\|_\infty \leq \left| \frac{\dot{\mu}_t}{\mu_t} + (\dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t}) \left(\frac{C_2}{\mu_t^2 + C_2\sigma_t^2} \right) \right| \cdot \|x\|_\infty + C_6 \left| \dot{\sigma}_t - \frac{\dot{\mu}_t\sigma_t}{\mu_t} \right|.$$

First, we apply $\|x\|_2 \lesssim \sqrt{\log N}$ to [Lemma K.4](#). Next, we apply [Lemma K.4](#) and [Lemma K.6](#) to construct the first-step approximator $Q(x, t) \in \mathbb{R}^{d_x}$, with each element defined by:

$$\|Q[i]\| := \min \left\{ \frac{\dot{\mu}_t}{\mu_t}x - (\dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t}) \left(\frac{-C_2x}{\mu_t^2 + C_2\sigma_t^2} + \frac{\hat{\mu}_t \nabla Q_2[i]}{\hat{\sigma}_t Q_1} \right), \|u_t(x)\|_\infty \right\}. \quad (\text{K.2})$$

The first element consists of approximators for $h(x, t)$ and $\nabla h(x, t)$. The second element ensures that $\Psi(x, t)$ does not output value larger than the maximum of $u_t(x)$ in ℓ_∞ .

- **Step A: Approximation via Local Polynomial.**

By symmetry, for all $i \in [d_x]$, the difference between $Q(x, t)[i]$ and $u_t(x)[i]$ follows

$$\begin{aligned} & |u_t[i] - Q[i]| \\ &= \left| \left(\dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t} \right) \left(\frac{\nabla h[i]}{h} - \frac{\hat{\mu}_t Q_2[i]}{\hat{\sigma}_t Q_1} \right) \right| \\ &\leq \left| \dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t} \right| \cdot \left| \left(\frac{\nabla h[i]}{h} - \frac{\hat{\mu}_t Q_2[i]}{\hat{\sigma}_t Q_1} \right) \right| \\ &= \left| \dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t} \right| \cdot \underbrace{\left| \frac{\nabla h[i]}{h} - \frac{\nabla h[i]}{Q_1} \right|}_{(T_1)} + \left| \dot{\sigma}_t\sigma_t - \frac{\dot{\mu}_t\sigma_t^2}{\mu_t} \right| \cdot \underbrace{\left| \frac{\nabla h[i]}{Q_1} - \frac{\hat{\mu}_t Q_2[i]}{\hat{\sigma}_t Q_1} \right|}_{(T_2)}. \end{aligned}$$

(By triangle inequality)

Next, we bound (T_1) and (T_2) .

Step A.1: Bound (T₁). By Lemma K.3, we have $C_1 \leq h \leq B$ and

$$\left\| \frac{\widehat{\sigma}_t}{\widehat{\mu}_t} \nabla h(x, t) \right\|_{\infty} \leq \sqrt{\frac{2}{\pi}} B.$$

Moreover, by Lemma K.5, it holds

$$|Q_1(x, t) - h(x, t)| \lesssim BN^{-\beta} (\log N)^{\frac{k_1}{2}}.$$

It implies that

$$h(x, t) - K'BN^{-\beta} (\log N)^{\frac{k_1}{2}} \leq Q_1(x, t),$$

for some positive constant K' . This gives

$$|Q_1(x, t)| \lesssim BN^{-\beta} (\log N)^{\frac{k_1}{2}}. \quad (\text{K.3})$$

Therefore,

$$\begin{aligned} (\text{T}_1) &= \left| \frac{\nabla h[i]}{h} - \frac{\nabla h[i]}{Q_1} \right| \\ &\leq |\nabla h[i]| \cdot \left| \frac{h - Q_1}{hQ_1} \right| \\ &\leq \sqrt{\frac{2}{\pi}} \frac{\widehat{\mu}_t}{\widehat{\sigma}_t} B \left| \frac{h - Q_1}{hQ_1} \right| \quad (\text{By Lemma K.3}) \\ &\lesssim \frac{B}{\sigma_t} N^{-\beta} (\log N)^{\frac{k_1}{2}}. \end{aligned}$$

Step 1.2: Bound (T₂). It holds

$$\begin{aligned} (\text{T}_2) &= \left| \frac{\nabla h[i]}{Q_1} - \frac{\widehat{\mu}_t Q_2[i]}{\widehat{\sigma}_t Q_1} \right| \\ &\leq \frac{\widehat{\mu}_t}{\widehat{\sigma}_t} \left| \frac{Q_2[i] - \frac{\widehat{\sigma}_t}{\widehat{\mu}_t} \nabla h[i]}{Q_1} \right| \quad (\text{By factoring out } \widehat{\mu}_t / \widehat{\sigma}_t) \\ &\lesssim \frac{B}{\sigma_t} N^{-\beta} (\log N)^{\frac{k_1+1}{2}}. \quad (\text{By (K.3) and Lemma K.6}) \end{aligned}$$

Combining bounds on (T₁) and (T₂), it holds

$$\begin{aligned} &|u_t[i] - Q[i]| \quad (\text{K.4}) \\ &\leq \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right| \cdot ((\text{T}_1) + (\text{T}_2)) \\ &\lesssim \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right| BN^{-\beta} (\log N)^{\frac{k_1+1}{2}}, \end{aligned}$$

for all $i \in [d_x]$.

Therefore, by symmetry, it holds

$$\begin{aligned} &\|u_t(x) - Q(x, t)\|_2^2 \quad (\text{K.5}) \\ &\leq d_x \|u_t(x) - Q(x, t)\|_{\infty}^2 \quad (\text{By } \|\cdot\|_2 \leq d_x \|\cdot\|_{\infty}) \end{aligned}$$

$$\lesssim \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right|^2 B^2 N^{-2\beta} (\log N)^{k_1}. \quad (\text{By (K.4)})$$

• **Step B: Approximation with Transformers.**

By Lemma K.7, there exists a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,r,s}$ such that

$$\int \int \|u_\theta(x, t) - \frac{\dot{\mu}_t}{\mu_t} x + (\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t}) (\frac{-C_2 x}{\mu_t^2 + C_2 \sigma_t^2} + \frac{\hat{\mu}_t \nabla Q_2[i]}{\hat{\sigma}_t Q_1})\|_2^2 dx dt \leq \epsilon^2.$$

By setting $\epsilon := N^{-\beta}$, the velocity approximation using transformers follows

$$\begin{aligned} & \int \int p_t \|u_t(x) - u_\theta(x, t)\|_2^2 dx dt \\ & \leq \int \int p_t \|u_t(x) - Q(x, t)\|_2^2 dx dt + \int \int p_t \|Q(x, t) - u_\theta(x, t)\|_2^2 dx dt \quad (\text{By triangle inequality}) \\ & \leq \int \int \|u_t(x) - Q(x, t)\|_2^2 dx dt + \int \int \|Q(x, t) - u_\theta(x, t)\|_2^2 dx dt \quad (\text{By } 0 \leq p_t(x) \leq 1) \\ & \lesssim \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right|^2 B^2 N^{-2\beta} (\log N)^{k_1} \int \int dx dt + \int \|Q(x, t) - u_\theta(x, t)\|_2^2 dx dt \quad (\text{By (K.5)}) \\ & \leq \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right|^2 B^2 N^{-2\beta} (\log N)^{k_1+d_x} + \int \|Q(x, t) - u_\theta(x, t)\|_2^2 dx dt \\ & \quad (\text{By } t \in (0, 1) \text{ and } \|x\|_\infty \leq C_x \sqrt{\log N}) \\ & \lesssim \left| \dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t} \right|^2 B^2 N^{-2\beta} (\log N)^{k_1+d_x}, \quad (\text{By Lemma K.7}) \\ & \lesssim B^2 N^{-2\beta} (\log N)^{k_1+d_x}. \quad (\text{By Assumption I.2}) \end{aligned}$$

By Lemma K.4, it holds

$$\begin{aligned} & \|u_t(x)\|_\infty \\ & \leq \left| \frac{\dot{\mu}_t}{\mu_t} + (\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t}) (\frac{C_2}{\mu_t^2 + C_2 \sigma_t^2}) \right| \cdot \|x\|_\infty + C_6 \left| \dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right|. \quad (\text{By Lemma K.4}) \\ & \lesssim O(\sqrt{\log N}). \quad (\text{By Assumption I.2}) \end{aligned}$$

Therefore, we set transformer output bound $C_{\mathcal{T}} := O(\|u_t(x)\|_\infty)$. Then, the parameter bounds in the transformer network follow Lemma J.9.

This completes the proof. $\square$

K.3 Main Proof of Theorem I.2

In Lemma J.9, we give the velocity field approximation using transformer on a bounded domain $x \in [-C_x \sqrt{\log N}, C_x \sqrt{\log N}]^{d_x}$ under stronger Hölder assumption. To obtain general approximation result, we introduce the next lemma that bounds the uncontrolled region.

Lemma K.9 (Truncation of x , Modified from Lemma B.2 of [Fu et al., 2024]). Assume Assumption I.3. Then, for any $R_7 > 1$ and $t > 0$, the following hold

$$\int_{\|x\|_\infty \geq R_7} p_t(x) dx \lesssim R_7^{d_x-2} \exp\left(-\frac{C_2 R_7^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right),$$

$$\int_{\|x\|_\infty \geq R_7} \|u_t(x)\|_2^2 \cdot p_t(x) dx \lesssim R_7^{d_x} \exp\left(-\frac{C_2 R_7^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right).$$

Proof. The first part of the proof is identical to [Lemma J.10](#)

Recall [Lemma J.3](#). The density function at time t is upper bounded by

$$\begin{aligned} p_t &\leq \frac{C_1}{(\mu_t^2 + C_2 \sigma_t^2)^{d_x/2}} \cdot \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) && \text{(By dropping constant term)} \\ &\lesssim \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right). \end{aligned}$$

Furthermore, by [Lemma K.4](#) we have

$$\begin{aligned} &\|u_t(x)\|_\infty && \text{(K.6)} \\ &\leq \left| \frac{\dot{\mu}_t}{\mu_t} + (\dot{\sigma}_t \sigma_t - \frac{\dot{\mu}_t \sigma_t^2}{\mu_t}) \left(\frac{C_2}{\mu_t^2 + C_2 \sigma_t^2} \right) \right| \cdot \|x\|_\infty + C_6 \left| \dot{\sigma}_t - \frac{\dot{\mu}_t \sigma_t}{\mu_t} \right| \\ &\lesssim \|x\|_\infty && \text{(By Assumption I.2)} \\ &\leq \|x\|_2. && \text{(By } \|\cdot\|_\infty \leq \|\cdot\|_2) \end{aligned}$$

Therefore, the second inequality follows

$$\begin{aligned} &\int_{\|x\|_\infty \geq R_7} \|u_t(x)\|_2^2 p_t(x) dx \\ &\leq d_x \int_{\|x\|_\infty \geq R_7} \|u_t(x)\|_\infty^2 p_t(x) dx && \text{(By } \|\cdot\|_2 \leq d_x \|\cdot\|_\infty) \\ &\lesssim \int_{\|x\|_\infty \geq R_7} \|x\|_2^2 \cdot p_t(x) dx && \text{(By (K.6))} \\ &\lesssim \int_{\|x\|_\infty \geq R_7} \|x\|_2^2 \cdot \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx && \text{(By Lemma J.3)} \\ &\lesssim \int_{\|x\|_2 \geq R_7} \|x\|_2^2 \cdot \exp\left(-\frac{C_2 \|x\|_2^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right) dx && \text{(By } \|x\|_2 \geq \|x\|_\infty) \\ &\lesssim R_7^{d_x} \exp\left(-\frac{C_2 R_7^2}{2(\mu_t^2 + C_2 \sigma_t^2)}\right). && \text{(By Lemma D.2)} \end{aligned}$$

This completes the proof. $\square$

Next, we present the main proof of [Theorem I.2](#)

Theorem K.1 ([Theorem I.2](#) Restated: Velocity Approximation with Transformers under Stronger Hölder Smoothness). Assume [Assumption I.3](#) and [Assumption I.2](#). For any precision parameter $0 < \epsilon < 1$ and smoothness parameter $\beta > 0$, let $\epsilon \leq O(N^{-\beta})$ for some $N \in \mathbb{N}$. Then, for all $t \in [t_0, T]$ with $t_0, T \in (0, 1)$, there exists a transformer $u_\theta(x, t) \in \mathcal{T}_R^{h,s,r}$ such that

$$\int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u_t(x) - u_\theta(x, t)\|_2^2 \cdot p_t(x) dx dt = O\left(B^2 N^{-2\beta} (\log N)^{d_x + \beta}\right),$$

Further, the parameter bounds in the transformer network class follows [Theorem I.1](#).

Proof of Theorem I.2. Recall Lemma K.8, Lemma K.9. We have $C_{\mathcal{T}} = O(\sqrt{\log N})$ and we set

$$R_3 := \sqrt{\frac{4\beta(\mu_t^2 + C_2\sigma_t^2) \log N}{C_2}}. \quad (\text{K.7})$$

Then, it holds

$$\begin{aligned} & \int_{t_0}^T \int_{\mathbb{R}^{d_x}} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt \\ &= \int_{t_0}^T \int_{\|x\|_{\infty} > R_3} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt + \int_{t_0}^T \int_{\|x\|_{\infty} \leq R_3} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt \\ &\leq 2 \int_{t_0}^T \int_{\|x\|_{\infty} > R_3} (\|u_{\theta}(x)\|_2^2 + \|u_t(x)\|_2^2) p_t(x) dx dt + \int_{t_0}^T \int_{\|x\|_{\infty} \leq R_3} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt \\ &\quad (\text{By expanding } \|\cdot\|_2^2) \\ &\lesssim \int_{t_0}^T \int_{\|x\|_{\infty} > R_3} (\log N + \|u_t(x)\|_2^2) \cdot p_t(x) dx dt + \int_{t_0}^T \int_{\|x\|_{\infty} \leq R_3} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt \\ &\quad (\text{By } C_{\mathcal{T}} = O(\sqrt{\log N})) \\ &\lesssim (\log N \cdot R_3^{d_x-2} + R_3^{d_x}) \exp\left(-\frac{C_2 R_3^2}{2(\mu_t^2 + C_2\sigma_t^2)}\right) + \int_{t_0}^T \int_{\|x\|_{\infty} \leq R_3} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt \\ &\quad (\text{By Lemma K.9}) \\ &\lesssim (\log N)^{\frac{d_x}{2}} N^{-2\beta} + \int_{t_0}^T \int_{\|x\|_{\infty} \leq R_3} \|u_{\theta}(x) - u_t(x)\|_2^2 p_t(x) dx dt \quad (\text{By (K.7)}) \\ &\leq (\log N)^{\frac{d_x}{2}} N^{-2\beta} + B^2 N^{-2\beta} (\log N)^{k_1+d_x} \quad (\text{By Lemma K.8}) \\ &= O\left(B^2 N^{-2\beta} (\log N)^{k_1+d_x}\right). \quad (\text{By } k_1 \leq \beta) \end{aligned}$$

Furthermore, the parameter bounds in transformer network follow Lemma K.8.

This completes the proof. $\square$

L Proof of Theorem I.3

In this section, we prove Theorem I.3 following the three steps presented in the proof sketch.

Organizations. Appendix L.1 provides fundamental definitions of flow matching and discusses key properties of the flow matching loss. Appendix L.2 introduces several auxiliary lemmas that support our proof. Finally, Appendix L.3 presents the main proof of Theorem I.3.

L.1 Preliminaries

In this section, we consider affine conditional flows $\psi_t(x|x_1) = \mu_t x_1 + \sigma_t x$ follows Section 2. Given a velocity approximator u_θ , we aim to bound the following flow matching risk $\mathcal{R}(u_\theta)$:

$$\mathcal{R}(u_\theta) := \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{x_t \sim p_t} [\|u_\theta(X_t, t) - u_t(X_t)\|_2^2] dt, \quad (\text{L.1})$$

where marginal probability path p_t and marginal velocity field u_t are induced by affine conditional flow $\psi_t(x|X_1) = \mu_t X_1 + \sigma_t x$ follows (2.2), (2.3), (2.5) and (2.6).

In practice, we use conditional flow matching loss to train velocity estimator u_θ :

Definition L.1 (Conditional Flow Matching). Let q be the ground truth distribution and the normal distribution $N(0, I)$ be the source distribution p . Considering affine conditional flows $\psi_t(x|x_1) = \mu_t X_1 + \sigma_t x$, we define the loss function and the conditional flow matching loss:

$$\begin{aligned} \ell(x; u_\theta) &:= \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0, I)} [\|u_\theta(\mu_t x + \sigma_t X_0, t) - (\dot{\mu}_t x + \dot{\sigma}_t X_0)\|_2^2] dt, \\ \mathcal{L}_{\text{CFM}}(u_\theta) &:= \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{X_1 \sim q, X_0 \sim N(0, I)} [\|u_\theta(\mu_t x + \sigma_t X_0, t) - (\dot{\mu}_t X_1 + \dot{\sigma}_t X_0)\|_2^2] dt. \end{aligned}$$

Remark L.1. Holderrieth et al. [2025] prove that the gradients of the flow matching loss (risk) and the conditional flow matching loss coincide. Therefore, minimizing the flow matching loss (risk) $\mathcal{R}(u_\theta)$ is equivalent to minimizing the conditional flow matching loss $\mathcal{L}_{\text{CFM}}(u_\theta)$.

To better evaluate the estimator u_θ , now we introduce the empirical flow matching risk $\hat{\mathcal{R}}(u_\theta)$.

Definition L.2 (Empirical Risk). Consider a velocity estimator $u_\theta \in \mathcal{T}_R^{h,s,r}$ and i.i.d training samples $\{x_i\}_{i=1}^n$, the empirical conditional flow matching loss $\hat{\mathcal{L}}_{\text{CFM}}(u_\theta) := \frac{1}{n} \sum_{i=1}^n \ell(x_i; u_\theta)$. Let $u^* := u_t$ be the ground truth velocity field, we define empirical flow matching risk:

$$\hat{\mathcal{R}}(u_\theta) := \hat{\mathcal{L}}_{\text{CFM}}(u_\theta) - \hat{\mathcal{L}}_{\text{CFM}}(u^*) = \frac{1}{n} \sum_{i=1}^n \ell(x_i; u_\theta) - \frac{1}{n} \sum_{i=1}^n \ell(x_i; u^*).$$

Remark L.2. Notice that $R(u^*) = 0$ since u^* is the ground truth velocity field. Furthermore, the fact that the gradients of the flow matching loss (risk) and the conditional flow matching loss coincide implies that $R(u_\theta) = R(u_\theta) - R(u^*) = \mathcal{L}_{\text{CFM}}(u_\theta) - \mathcal{L}_{\text{CFM}}(u^*)$.

Remark L.3 (Unbiased Property). We use $\hat{\mathcal{L}}'_{\text{CFM}}$ and $\hat{\mathcal{R}}'$ to denote the conditional flow matching loss and empirical risk with training samples $\{x'_i\}_{i=1}^n$. Then for any velocity estimator u_θ , the i.i.d. assumption implies that $\mathbb{E}_{\{x'_i\}_{i=1}^n} [\hat{\mathcal{L}}'_{\text{CFM}}(u_\theta)] = \mathcal{L}_{\text{CFM}}(u_\theta)$, leading to $\mathbb{E}_{\{x'_i\}_{i=1}^n} [\hat{\mathcal{R}}'(u_\theta)] = \mathcal{R}(u_\theta)$.

Next, we introduce the truncated version of (i) loss function $\ell(x; u_\theta)$, (ii) conditional flow matching loss $\mathcal{L}_{\text{CFM}}(u_\theta)$, (iii) the conditional flow matching risk, $\mathcal{R}(u_\theta)$ (iv) the empirical risk $\hat{\mathcal{R}}(u_\theta)$.

Definition L.3 (Domain Truncation of Loss and Risk). Given $\ell(x; u_\theta)$, $\mathcal{L}_{\text{CFM}}(u_\theta)$, $\mathcal{R}(u_\theta)$ and $\widehat{\mathcal{R}}(u_\theta)$, we define their truncated counterparts on a bounded domain $\mathcal{D} := [-D, D]^{d_x}$ by

$$\begin{aligned}\ell^{\text{trunc}}(x; u_\theta) &:= \ell(x; u_\theta) \mathbb{1}\{\|x\|_\infty \leq D\}, & \mathcal{L}_{\text{CFM}}^{\text{trunc}}(u_\theta) &:= \mathcal{L}(u_\theta) \mathbb{1}\{\|x\|_\infty \leq D\}, \\ \mathcal{R}^{\text{trunc}}(u_\theta) &:= \mathcal{R}(x; u_\theta) \mathbb{1}\{\|x\|_\infty \leq D\}, & \widehat{\mathcal{R}}^{\text{trunc}}(u_\theta) &:= \widehat{\mathcal{R}}(u_\theta) \mathbb{1}\{\|x\|_\infty \leq D\},\end{aligned}$$

where $D > 0$ is a constant.

With **Definition L.3**, we refer to $\ell^{\text{trunc}}(x; u_\theta)$, $\mathcal{L}_{\text{CFM}}^{\text{trunc}}(u_\theta)$, $\mathcal{R}^{\text{trunc}}(u_\theta)$ and $\widehat{\mathcal{R}}^{\text{trunc}}(u_\theta)$ as truncated loss, truncated CFM loss, truncated risk and truncated empirical risk respectively.

L.2 Auxiliary lemmas

Since the target distribution $q(x_1)$ is unknown, direct computation of the risk is infeasible. Therefore, we first decompose the estimation error into four components and present supporting lemmas to bound each of them. Then, we incorporate these results in the main proof in **Appendix L.3**.

Estimation Error Decomposition. Let $\widehat{u}_\theta$ be the optimizer of the empirical conditional flow matching loss $\widehat{\mathcal{L}}_{\text{CFM}}(u_\theta)$ using i.i.d samples $\{x_i\}_{i=1}^n$. Next, we introduce a different set of i.i.d samples $\{x'_i\}_{i=1}^n$ independent of the training sample $\{x_i\}_{i=1}^n$. Then, we decompose $\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\widehat{u}_\theta)]$:

$$\begin{aligned}\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\widehat{u}_\theta)] &= \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{R}}'(\widehat{u}_\theta) - \widehat{\mathcal{R}}'^{\text{trunc}}(\widehat{u}_\theta)] \right]}_{\text{(I)}} \\ &+ \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} [\widehat{\mathcal{R}}'^{\text{trunc}}(\widehat{u}_\theta) - \widehat{\mathcal{R}}^{\text{trunc}}(\widehat{u}_\theta)] \right]}_{\text{(II)}} \\ &+ \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}^{\text{trunc}}(\widehat{u}_\theta) - \widehat{\mathcal{R}}(\widehat{u}_\theta)]}_{\text{(III)}} + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}(\widehat{u}_\theta)]}_{\text{(IV)}}. \quad (\text{L.2})\end{aligned}$$

We refer to terms (I) and (III) as *truncation error*, and we control these errors by leveraging the sub-Gaussian assumption in **Lemma L.1**. Then, we derive the generalization bound to control term (II) using covering number in **Lemma L.5** and **Lemma L.6**. Finally, we apply the approximation error using transformers to bound term (IV) in **Lemma L.8**.

Truncation Error. We apply the sub-Gaussian assumption to bound the truncation error.

Lemma L.1 (Upper Bound on the Truncation Error). Assume **Assumption I.1**. Let $t_0, T \in (0, 1)$ and $u_\theta(x, t)$ be the velocity approximators in **Theorem I.1** and **Theorem I.2**. Then, it holds

$$\mathbb{E}_x [|\ell(x; u_\theta) - \ell^{\text{trunc}}(x; u_\theta)|] \lesssim D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \log N \quad \text{for any } t \in [t_0, T].$$

Proof. Our proof builds on Section D.2 of [Fu et al., 2024]. By **Theorem I.1** and **Theorem I.2**, the transformers output bound $C_T = O(\sqrt{\log N})$. Then, for all approximator $u_\theta \in \mathcal{T}_R^{h,s,r}$, it holds

$$\begin{aligned}& \mathbb{E}_x [|\ell(x; u_\theta) - \ell^{\text{trunc}}(x; u_\theta)|] \\ &= \mathbb{E}_x [|\ell(x; u_\theta) \mathbb{1}\{\|x\| \geq D\}|] \quad (\text{By Definition L.3}) \\ &= \frac{1}{T - t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{x_0 \sim N(0, I)} [\|u_\theta - (\dot{\mu}_t x + \dot{\sigma}_t x_0)\|_2^2] q(x) dx dt \quad (\text{By Definition L.1}) \\ &\leq \frac{2}{T - t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{x_0 \sim N(0, I)} [\|u_\theta\|_2^2 + \|\dot{\mu}_t x + \dot{\sigma}_t x_0\|_2^2] q(x) dx dt \quad (\text{By expanding the } \ell_2\text{-norm})\end{aligned}$$

$$\begin{aligned}
&\lesssim \frac{2}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{x_0 \sim N(0, I)} [\|u_\theta\|_2^2 + \|\dot{\mu}_t x + \dot{\sigma}_t x_0\|_2^2] \exp\left(-\frac{1}{2} C_2 \|x\|_2^2\right) dx dt \\
&\quad \text{(By Assumption I.1)} \\
&\lesssim \frac{2}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} \mathbb{E}_{x_0 \sim N(0, I)} [\log N + \|\dot{\mu}_t x + \dot{\sigma}_t x_0\|_2^2] \exp\left(-\frac{1}{2} C_2 \|x\|_2^2\right) dx dt \\
&\quad \text{(By } C_T = O(\sqrt{\log N})\text{)} \\
&\lesssim \frac{1}{T-t_0} \int_{t_0}^T \int_{\|x\| \geq D} (\log N + \dot{\sigma}_t^2 d + \dot{\mu}_t^2 \|x\|_2^2) \exp\left(-\frac{1}{2} C_2 \|x\|_2^2\right) dx dt \quad \text{(By } x_0 \sim N(0, I)\text{)} \\
&\lesssim \frac{D^{d_x-2} \exp(-\frac{1}{2} C_2 D^2)}{T-t_0} \int_{t_0}^T (\log N + \dot{\sigma}_t^2 d) dt + \frac{D^{d_x} \exp(-\frac{1}{2} C_2 D^2)}{T-t_0} \int_{t_0}^T \dot{\mu}_t^2 dt \\
&\quad \text{(By Lemma D.2)} \\
&\lesssim D^{d_x} \exp\left(-\frac{1}{2} C_2 D^2\right) \log N. \quad \text{(By Assumption I.2)}
\end{aligned}$$

This completes the proof. $\square$

Covering Number of Loss Function Class with Transformer Estimator. Recall (II) in (L.2):

$$(II) = \mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} \left[\widehat{\mathcal{R}}'^{\text{trunc}}(\widehat{u}_\theta) - \widehat{\mathcal{R}}^{\text{trunc}}(\widehat{u}_\theta) \right] \right].$$

To derive an upper bound on (II), we introduce (i) the covering number technique in Lemma L.5 and Lemma L.6 (ii) the generalization error bound to bound in Lemma L.8.

We begin with the definition of covering number.

Definition L.4 (Covering Number). Let Ω be a compact domain and $\{x_i\}_{i=1}^n$ be data drawn from distribution P . Denote the joint distribution $\{x_i\}_{i=1}^n \sim P^n := P \otimes \dots \otimes P$. Given a function class $\mathcal{F}$, a $t \in \Omega$, a norm $\|\cdot\|$ and a $\epsilon_c > 0$, the ϵ_c -covering number $\mathcal{N}(\epsilon, \mathcal{F}, \{x_i\}_{i=1}^n \times \Omega, \|\cdot\|)$ 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

$$\sup_t \max_i \|f(x_i, t) - f_j(x_i, t)\| \leq \epsilon. \quad (\text{L.3})$$

Also, we define the covering number with respect to the data distribution P and size n as

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

Further, for $\Omega = \emptyset$, we denote the covering number by $\mathcal{N}(\epsilon, \mathcal{F}, \{x_i\}_{i=1}^n, \|\cdot\|)$ and $\mathcal{N}(\epsilon, \mathcal{F}, P^n, \|\cdot\|)$.

Remark L.4 (Covering Number of Transformer Network Class). We define the covering number over domain $\{x_i\}_{i=1}^n \times \Omega$ to align with the flow matching loss formulation in Definition L.1, where temporal dependence in transformers introduces no additional statistical error. Specifically, the loss averages over the time component, unlike the n i.i.d. data points sampled from target distribution q^n .

Next, we derive an upper bound on the covering number of transformer networks. Our proof builds on [Edelman et al., 2022] and studies the class in Definition B.4, with a self-attention layer that applies softmax under inverse-temperature scaling and a feed-forward layer with ReLU activation.

Covering Number of Linear Function Class. Our norm-based upper bound on the covering number of the transformer network class extends the classical norm-based bound for the linear function class:

Lemma L.2 (Covering Number of Linear Function Class, Lemma 4.6 of [Edelman et al., 2022] and [Zhang, 2002]). Let $z_1, \dots, z_n \in \mathbb{R}^d$ be sample points satisfying $\|z_i\| \leq B_X$ for all $i \in [n]$. Then,

for linear function class $\mathcal{F} := \{f : z \rightarrow Wz \mid W \in \mathbb{R}^{d' \times d}, \|W\|_{2,1} \leq B_W, B_W > 0\}$, it holds:

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \{z_i\}_{i=1}^n; \|\cdot\|_\infty) \lesssim \frac{B_X^2 B_W^2}{\epsilon^2} \log(d'n) \quad \text{for any } \epsilon > 0.$$

Remark L.5 (Covering Number Equivalence). We remark that [Lemma L.2](#) applies to our feed-forward layer for two reasons. First, the bias terms b_1 and b_2 admit an augmented form. By appending a bottom row of ones to Z and set $\widetilde{W}_1 := (W_1, b_1)$, it holds that $W_1 Z + b_1 = \widetilde{W}_1 \widetilde{Z}$, where $W_1 \in \mathbb{R}^{r \times d}$, $b_1 \in \mathbb{R}^r$, and $\widetilde{W}_1 \in \mathbb{R}^{r \times (d+1)}$, $\widetilde{Z} \in \mathbb{R}^{(d+1) \times L}$. This two forms define the same function class. Second, the norm bound in [Lemma H.5](#) keeps the biases on the same order as the matrices operator. Since our $\lesssim$ and $O(\cdot)$ notation hides polynomial and logarithmic factors in d and L , [Lemma L.2](#) gives the covering number for the linear class in our feed-forward layer with biases.

Following [\[Edelman et al., 2022\]](#), we extend [Lemma L.2](#) to a transformer block. We view the block as a composition of linear function classes. We construct covers for each linear function class, balance the errors across components, and minimize the size of the concatenated cover using the next lemmas:

Lemma L.3 (Lemma A.8 of [\[Edelman et al., 2022\]](#)). Consider the following optimization problem:

$$\min_{x_1, \dots, x_n} \sum_{i=1}^n \frac{\alpha_i}{x_i^2} \quad \text{subject to} \quad \sum_{i=1}^n \omega_i x_i = C \quad (\text{L.4})$$

for some α_i, ω_i and a constant C . Then, (L.4) has solution

$$\frac{\left(\sum_{i=1}^n \alpha_i^{\frac{1}{3}} \omega_i^{\frac{2}{3}}\right)^3}{C^2} \quad \text{when} \quad x_i = \frac{C}{\sum_{i=1}^n \alpha_i^{\frac{1}{3}} \omega_i^{\frac{2}{3}}} \cdot \left(\frac{\alpha_i}{\omega_i}\right)^{\frac{2}{3}}.$$

Equipped with [Lemma L.3](#), we have the upper bound for the covering number of transformer block:

Lemma L.4 (Covering Number of Transformer Block, Modified from Corollary 4.5 of [\[Edelman et al., 2022\]](#)). Let $\{x_i\}_{i=1}^n$ be sample points satisfying $\max_i \|x_i\|_\infty \leq D$ for some constant $D > 0$ and $R(\cdot) : \mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d \times L}$ be the reshape layer ([Definition B.3](#)). Let $\{Z_i\}_{i=1}^n := \{R(x_i)\}_{i=1}^n$, $E \in \mathbb{R}^{d \times L}$ be the positional encoding and $\mathcal{T}_R^{1,s,r}(Z)$ denote a two-layer transformer class ([Definition B.1](#)) with single-head self-attention and s -hidden dimension and r -MLP dimension. Then, it holds:

$$\begin{aligned} & \log \mathcal{N}(\epsilon, \mathcal{T}_R^{1,s,r}, \{Z_i\}_{i=1}^n; \|\cdot\|_\infty) \\ & \lesssim \frac{\log(nL)}{\epsilon_c^2} \alpha^2 \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3, \end{aligned}$$

where $\alpha := O(C_F^2 C_{OV} C_{KQ} (D + C_E))$.

Proof. Our proof builds on [\[Edelman et al., 2022\]](#) by incorporating the scaling $\lambda > 0$ for the column-wise softmax function. For $Z \in [-D, D]^{d \times L}$ and $W_K, W_Q \in \mathbb{R}^{s \times d}$, we define

$$\mathcal{F}_{KQ} := \{f : Z \rightarrow (W_K Z)^\top (W_Q Z) \mid W_{KQ} = W_K^\top W_Q, \|W_{KQ}\|_2 \leq C_{KQ}\},$$

for some constant $D > 0$. Similarly, for $W_V \in \mathbb{R}^{s \times d}$ and $W_O \in \mathbb{R}^{d \times s}$, we define

$$\mathcal{F}_{OV} := \{f : Z \rightarrow W_O \cdot W_V Z \mid W_{OV} = W_O W_V, \|W_{OV}\|_2 \leq C_{OV}\}.$$

Recall [Remark L.5](#). For feed-forward layer, we define:

$$\mathcal{F}^{(\text{FF})} := \{f : Z \rightarrow W_2 \text{ReLU}[W_1 Z] \mid \|W_1\|_2, \|W_2\|_2 \leq C_F\}.$$

For simplicity, we denote the k -th column of Z_i by $z_i^{(k)}$ for all $i \in [n]$. First observe that

$$\max_{i \in [n]} \|f(Z_i) - \hat{f}(Z_i)\|_{2,\infty} = \max_{i \in [n], k \in [L]} \|f(z_i^{(k)}) - \hat{f}(z_i^{(k)})\|_2 \quad (\text{L.5})$$

for any distinct $f, \hat{f}$ in $\mathcal{F}_{KQ}, \mathcal{F}_{OV}$ or $\mathcal{F}^{(\text{FF})}$. With this, we consider $\{Z_i\}_{i=1}^n$ as nL samples in $[-D, D]^d$ and apply [Lemma L.2](#) to construct covers $\mathcal{C}_{KQ}, \mathcal{C}_{OV}$ and $\mathcal{C}_{\text{FF}}$ for $\mathcal{F}_{KQ}, \mathcal{F}_{OV}$ and $\mathcal{F}^{(\text{FF})}$.

Covering Number for $\mathcal{F}_{KQ}$ and $\mathcal{F}_{OV}$. Since $Z_i \in [-D, D]^{d \times L}$, for any $i, j \in [n]$ and $k \in [L]$, we have that $\|f(z_i^{(k)}) - f(z_j^{(k)})\|_2^2 \leq dD^2 \cdot \|(W_K^\top W_Q)Z\|_2^2$. Thus, by [Lemma L.2](#) and [\(L.5\)](#), it holds

$$\log \mathcal{N}(\epsilon_{KQ}, \mathcal{F}_{KQ}, \{Z_i\}_{i=1}^n; \|\cdot\|_\infty) \lesssim \frac{D^3 C_{KQ}^2}{\epsilon_{KQ}^2} \log(nL), \quad (\text{L.6})$$

where $\lesssim$ hides polynomial factors dependent on d and L . Similarly, for $\mathcal{F}_{OV}$, it holds

$$\log \mathcal{N}(\epsilon_{OV}, \mathcal{F}_{OV}, \{Z_i\}_{i=1}^n; \|\cdot\|_\infty) \lesssim \frac{D^2 C_{OV}^2}{\epsilon_{OV}^2} \log(nL), \quad (\text{L.7})$$

That is, we have cover $\mathcal{C}_{KQ}, \mathcal{C}_{OV}$ for $\mathcal{F}_{KQ}, \mathcal{F}_{OV}$ whose sizes are upper-bounded by [\(L.6\)](#) and [\(L.7\)](#).

By triangle inequality and by [Lemma B.2](#), for any $W_K, W_Q \in \mathcal{F}_{KQ}$ and any $W_O, W_V \in \mathcal{F}_{OV}$, there exist some $\widehat{W}_K, \widehat{W}_Q \in \mathcal{C}_{KQ}$ and $\widehat{W}_O, \widehat{W}_V \in \mathcal{C}_{OV}$ such that for all $i \in [n]$ and all $k \in [L]$, it holds

$$\begin{aligned} & \|\widehat{W}_O(\widehat{W}_V z_i^{(k)}) \text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})] - W_O(W_V z_i^{(k)}) \text{Softmax}[(W_K z_i^{(k)})^\top (W_Q z_i^{(k)})]\| \\ & \hspace{25em} (\text{L.8}) \\ & \leq \underbrace{\|\widehat{W}_O(\widehat{W}_V z_i^{(k)}) \text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})] - W_O(W_V z_i^{(k)}) \text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})]\|}_{(\text{I})} \\ & \quad + \underbrace{\|W_O(W_V z_i^{(k)}) \text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})] - W_O(W_V z_i^{(k)}) \text{Softmax}[(W_K z_i^{(k)})^\top (W_Q z_i^{(k)})]\|}_{(\text{II})} \\ & \hspace{15em} (\text{By triangle inequality}) \\ & \lesssim \underbrace{\|\widehat{W}_O \widehat{W}_V - W_O W_V\| z_i^{(k)}}_{(\text{I})} + \underbrace{\|W_O W_V z_i^{(k)}\| \|\text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})] - \text{Softmax}[(W_K z_i^{(k)})^\top (W_Q z_i^{(k)})]\|}_{(\text{II})} \\ & \hspace{15em} (\text{By } \|\text{Softmax}[\cdot]\|_F \leq dL) \\ & \lesssim \underbrace{\|\widehat{W}_O \widehat{W}_V - W_O W_V\| z_i^{(k)}}_{(\text{I})} + \underbrace{C_{OV} D \|\text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})] - \text{Softmax}[(W_K z_i^{(k)})^\top (W_Q z_i^{(k)})]\|}_{(\text{II})} \\ & \hspace{15em} (\text{By } \|W_{OV}\| \leq C_{OV} \text{ and } \|z_i^{(k)}\|_\infty \leq D) \\ & \lesssim \epsilon_{OV} + \underbrace{C_{OV} D \|\text{Softmax}[(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)})] - \text{Softmax}[(W_K z_i^{(k)})^\top (W_Q z_i^{(k)})]\|}_{(\text{II})} \\ & \hspace{25em} (\text{By (L.6)}) \\ & \lesssim \epsilon_{OV} + \underbrace{2\lambda C_{OV} D \|(\widehat{W}_K z_i^{(k)})^\top (\widehat{W}_Q z_i^{(k)}) - (W_K z_i^{(k)})^\top (W_Q z_i^{(k)})\|}_{(\text{II})} \quad (\text{By Lemma B.2}) \\ & \leq \epsilon_{OV} + 2\lambda C_{OV} D \epsilon_{KQ}. \quad (\text{By (L.7)}) \end{aligned}$$

Then, we have a cover for the self-attention layer

$$\mathcal{C}_{\text{SA}} := \{f : Z \rightarrow \widehat{W}_O(\widehat{W}_V Z) \text{Softmax}[(\widehat{W}_K Z)^\top (\widehat{W}_Q Z)] \mid \widehat{W}_O \widehat{W}_V \in \mathcal{C}_{OV}, \widehat{W}_K^\top \widehat{W}_Q \in \mathcal{C}_{KQ}\}.$$

Covering Number for $\mathcal{F}^{(\text{FF})}$. Similarly, we apply [Lemma L.2](#) twice for the inner and outer linear function class operated by W_1 and W_2 . Specifically, we construct cover $\mathcal{C}_1$ and $\mathcal{C}_2$ for

$$\mathcal{F}_1^{(\text{FF})} := \{f : Z \rightarrow W_1 Z \mid \|W_1\|_2 \leq C_F\} \quad \text{and} \quad \mathcal{F}_2^{(\text{FF})} := \{f : Z \rightarrow W_2 Z \mid \|W_2\|_2 \leq C_F\}$$

respectively. Then, by [Lemma L.2](#), we have cover $\mathcal{C}_1$ and $\mathcal{C}_2$ for

$$\log \mathcal{N}(\epsilon_{F,1}, \mathcal{F}_1^{(\text{FF})}, \{Z_i\}_{i=1}^n; \|\cdot\|_\infty) \lesssim \frac{D^2 C_F^2}{\epsilon_{F,1}^2} \log(nL). \quad (\text{L.9})$$

and

$$\log \mathcal{N}(\epsilon_{F,2}, \mathcal{F}_2^{(\text{FF})}, \{Z_i\}_{i=1}^n; \|\cdot\|_\infty) \lesssim \frac{D^2 C_F^2}{\epsilon_{F,2}^2} \log(nL). \quad (\text{L.10})$$

That is, we have cover $\mathcal{C}_1, \mathcal{C}_2$ for $\mathcal{F}_1^{(\text{FF})}, \mathcal{F}_2^{(\text{FF})}$ whose sizes are upper-bounded by [\(L.9\)](#) and [\(L.10\)](#).

Let $\widehat{\mathcal{F}}^{(\text{SA})}(Z)$ denote the self-attention layer with weights chosen from cover $\mathcal{C}_{\text{SA}}$. For any W_1, W_K, W_Q, W_O, W_V , there exist some $\widehat{W}_1 \in \mathcal{C}_1$ and $\widehat{\mathcal{F}}^{(\text{SA})}$ such that for all $i \in [n], k \in [L]$, it holds

$$\begin{aligned} & \|\widehat{W}_1 \circ \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - W_1 \circ \mathcal{F}^{(\text{SA})}(z_i^{(k)})\| \\ & \leq \|\widehat{W}_1 \circ \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - \widehat{W}_1 \circ \mathcal{F}^{(\text{SA})}(z_i^{(k)})\| + \|\widehat{W}_1 \circ \mathcal{F}^{(\text{SA})}(z_i^{(k)}) - W_1 \circ \mathcal{F}^{(\text{SA})}(z_i^{(k)})\| \\ & \quad \text{(By triangle inequality)} \\ & \lesssim C_F \|\widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - \mathcal{F}^{(\text{SA})}(z_i^{(k)})\| + \|\widehat{W}_1 \circ \mathcal{F}^{(\text{SA})}(z_i^{(k)}) - W_1 \circ \mathcal{F}^{(\text{SA})}(z_i^{(k)})\| \\ & \quad \text{(By norm bound on } W_1) \\ & \lesssim C_F \|\widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - \mathcal{F}^{(\text{SA})}(z_i^{(k)})\| + C_{OV} \|\widehat{W}_1 - W_1\|(z_i^{(k)}) \\ & \quad \text{(By the norm bound on } W_O \cdot W_V) \\ & \leq C_F(\epsilon_{OV} + 2\lambda C_{OV} D \epsilon_{KQ}) + C_{OV} \epsilon_{F,1}, \quad \text{(By (L.8) and (L.9))} \end{aligned} \quad (\text{L.11})$$

Building on [\(L.11\)](#), we have the cover for $\mathcal{F}^{(\text{FF})} \circ \mathcal{F}^{(\text{SA})}$:

$$\begin{aligned} & \|\widehat{W}_2 \text{ReLU}[\widehat{W}_1 \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)})] - W_2 \text{ReLU}[W_1 \mathcal{F}^{(\text{SA})}(z_i^{(k)})]\| \\ & \lesssim \underbrace{\|\widehat{W}_2 \widehat{W}_1 \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - W_2 \widehat{W}_1 \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)})\|}_{\text{(I)}} + \underbrace{\|W_2 \widehat{W}_1 \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - W_2 W_1 \mathcal{F}^{(\text{SA})}(z_i^{(k)})\|}_{\text{(II)}} \\ & \quad \text{(By triangle inequality and Lipschitzness of ReLU)} \\ & \lesssim \underbrace{C_F C_{OV} \|\widehat{W}_2 - W_2\| z_i^{(k)}}_{\text{(I)}} + \underbrace{\|W_2 \widehat{W}_1 \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - W_2 W_1 \mathcal{F}^{(\text{SA})}(z_i^{(k)})\|}_{\text{(II)}} \\ & \quad \text{(By norm bound on } W_1 \text{ and } W_{OV}) \\ & \lesssim \underbrace{C_F C_{OV} \|\widehat{W}_2 - W_2\| z_i^{(k)}}_{\text{(I)}} + C_F \underbrace{\|\widehat{W}_1 \widehat{\mathcal{F}}^{(\text{SA})}(z_i^{(k)}) - W_1 \mathcal{F}^{(\text{SA})}(z_i^{(k)})\|}_{\text{(II)}} \quad \text{(By norm bound on } W_2) \\ & \lesssim C_F C_{OV} \epsilon_{F,2} + C_F^2 (\epsilon_{OV} + 2\lambda C_{OV} D \epsilon_{KQ}) + C_F C_{OV} \epsilon_{F,1}. \quad \text{(By (L.10) and (L.11))} \end{aligned} \quad (\text{L.12})$$

The overall size of the ϵ -cover of a transformer block $\mathcal{F}^{1,s,r} = \mathcal{F}^{(\text{FF})} \circ \mathcal{F}^{(\text{SA})}$ is

$$\mathcal{C}_{\text{trans}} := \{\mathcal{F}^{(\text{FF})} \circ \mathcal{F}^{(\text{SA})} \mid \widehat{W}_O \widehat{W}_V \in \mathcal{C}_{OV}, \widehat{W}_K^\top \widehat{W}_Q \in \mathcal{C}_{KQ}, \widehat{W}_1, \widehat{W}_2 \in \mathcal{C}_{\text{FF}}\}.$$

This gives

$$\log |\mathcal{C}_{\text{trans}}| = \log |\mathcal{C}_{KQ}| + \log |\mathcal{C}_{OV}| + \log |\mathcal{C}_{\text{FF}}|,$$

satisfying $C_F C_{OV}(\epsilon_{F,1} + \epsilon_{F,2}) + C_F^2(\epsilon_{OV} + 2\lambda C_{OV} D \epsilon_{KQ}) \leq \epsilon$. Let

$$\begin{aligned}\alpha_1 &:= D^3 C_{KQ}^2 \log nL; & \alpha_2 &:= D^2 C_{OV}^2 \log(nL); & \alpha_3 &:= D^2 C_F^2 \log(nL); \\ \omega_1 &:= 2\lambda C_F^2 C_{OV} D; & \omega_2 &:= C_F^2 C_{OV} D; & \omega_3 &:= \lambda C_F C_{OV}.\end{aligned}$$

Finally, we apply [Lemma L.3](#) and obtain the optimal size of the the cover $\mathcal{C}_{\text{trans}}$:

$$\frac{D^2 \log(nL) (C_{KQ}^{2/3} \lambda^{2/3} C_F^{4/3} C_{OV}^{2/3} D^{1/3} + C_{OV}^{4/3} C_F^{4/3} D^{2/3} + \lambda^{2/3} C_{OV}^{2/3} C_F^{4/3})^3}{\epsilon^2}.$$

We extend the argument to two blocks by invoking the composition step in [\[Edelman et al., 2022\]](#). For inputs with positional encoding E , we replace the bound D with $D + C_E$, where $\|E\| \leq C_E$.

This completes the proof. $\square$

Remark L.6 (Looseness of the Covering Number Bound). The bound in [Lemma L.4](#) is loose yet sufficient for our two layer transformer class with inverse-temperature scaling and one attention head (recall [Appendix H](#)). The extension from one block to two follows from the composition step in Theorem A.17 of [\[Edelman et al., 2022\]](#), where it provides the complete induction argument.

Lemma L.5 (Covering Number Bounds for Transformer Network Class, Modified from Theorem A.17 of [\[Edelman et al., 2022\]](#)). Let $\mathcal{T}_R^{h,s,r}(C_{\mathcal{T}}, C_{KQ}^{2,\infty}, C_{KQ}, C_{OV}^{2,\infty}, C_{OV}, C_E, C_F^{2,\infty}, C_F, L_{\mathcal{T}})$ be the class of functions of one transformer block satisfying the norm bound for matrix and the Lipschitz property for feed-forward layers. Then, for all data points satisfying $\|x_i\|_{2,\infty} \leq D$, it holds

$$\begin{aligned}& \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, P^n \times [0, 1], \|\cdot\|_2) \\ & \lesssim \frac{\log(nL/\epsilon_c)}{\epsilon_c^2} \alpha^2 \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3,\end{aligned}\tag{L.13}$$

where $\alpha := O(C_F^2 C_{OV} C_{KQ} (D + C_E))$.

Proof. [Lemma L.4](#) shows [Lemma L.5](#) with the absence of domain $[0, 1]$. That is,

$$\begin{aligned}& \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, P^n, \|\cdot\|_2) \\ & \lesssim \frac{\log(nL)}{\epsilon_c^2} \alpha^2 \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3,\end{aligned}\tag{L.14}$$

holds with data points drawn from P^n . To extend it to $P^n \times [0, 1]$, we discretize $[0, 1]$ into a δ -grid $\mathcal{G} := \{t_k = k \cdot \delta \mid k = 0, 1, \dots, \lfloor 1/\delta \rfloor, \delta \in (0, 1)\}$. For simplicity, we denote $|\mathcal{G}| := m = O(1/\delta)$.

Let $\mathcal{T}_{R,\mathcal{G}}^{h,s,r}$ be the transformer network class on domain $[-D, D]^{d_x} \times \mathcal{G}$. We first suppose that

$$\mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, P^n \times [0, 1], \|\cdot\|_2) \leq \mathcal{N}(\epsilon_c/2, \mathcal{T}_{R,\mathcal{G}}^{h,s,r}, P^n \times \mathcal{G}, \|\cdot\|_2)\tag{L.15}$$

holds when $\delta := \epsilon_c/(4L_{\mathcal{T}})$, where $L_{\mathcal{T}}$ is the Lipschitz constant of the transformer block.

Then, since domain $P^n \times \mathcal{G}$ is a set of sample points with size at most nm , [\(L.14\)](#) gives

$$\begin{aligned}& \mathcal{N}(\epsilon_c/2, \mathcal{T}_{R,\mathcal{G}}^{h,s,r}, P^n \times \mathcal{G}, \|\cdot\|_2) \\ & \lesssim \frac{\log(nmL)}{(\epsilon_c/2)^2} \alpha^2 \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3\end{aligned}\tag{L.16}$$

Therefore, [\(L.15\)](#) and [\(L.16\)](#) implies [\(L.13\)](#):

$$\mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, P^n \times [0, 1], \|\cdot\|_2)$$

$$\begin{aligned}
&\leq \mathcal{N}(\epsilon_c/2, \mathcal{T}_{R,\mathcal{G}}^{h,s,r}, P^n \times \mathcal{G}, \|\cdot\|_2) && \text{(By (L.15))} \\
&\leq \frac{\log(nmL)}{(\epsilon_c/2)^2} \alpha^2 \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3 && \text{(By (L.16))} \\
&\lesssim \frac{\log(nL/\epsilon_c)}{\epsilon_c^2} \alpha^2 \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3. \\
&\hspace{15em} \text{(By } \delta = \epsilon_c/L_{\mathcal{T}} \text{ and dropping lower order terms)}
\end{aligned}$$

Lastly, it suffices to prove that (L.15) holds when $\delta := \epsilon_c/(4L_{\mathcal{T}})$. We show this by utilizing the Lipchitz property of the transformer networks (Definition B.4.) Specifically, let $\{f_j\}_{j=1}^{\mathcal{N}}$ be a $\epsilon_c/2$ cover of $\mathcal{T}_{R,\mathcal{G}}^{h,s,r}$. Then, for any $f \in \mathcal{T}_{R,\mathcal{G}}^{h,s,r}$ and $t \in [0, 1]$, it holds:

$$\begin{aligned}
&\|f(x_i, t) - f_j(x_i, t)\|_2 \\
&\leq \underbrace{\|f(x_i, t) - f(x_i, t_k)\|_2}_{(A)} + \underbrace{\|f(x_i, t_k) - f_j(x_i, t_k)\|_2}_{(B)} + \underbrace{\|f_j(x_i, t_k) - f_j(x_i, t)\|_2}_{(C)}. \\
&\hspace{15em} \text{(By triangle inequality)}
\end{aligned}$$

We then show that the RHS is bounded by ϵ_c and this implies (L.15). For (A) and (C), it holds:

$$(A), (C) \leq L_{\mathcal{T}} \cdot |t - t_k| \leq L_{\mathcal{T}} \cdot \delta \leq \frac{\epsilon_c}{4},$$

where the first inequality is by the Lipchitzness of the transformer network, the second inequality is by the definition of the δ -grid $\mathcal{G}$, and the last inequality is by taking $\delta := \epsilon_c/(4L_{\mathcal{T}})$.

Further, (B) is bounded by $\epsilon_c/2$ by the definition of $\{f_j\}_{j=1}^{\mathcal{N}}$. Altogether, we have that

$$\|f(x_i, t) - f_j(x_i, t)\|_2 \leq \frac{\epsilon_c}{4} + \frac{\epsilon_c}{2} + \frac{\epsilon_c}{4} = \epsilon_c. \quad (\text{L.17})$$

Since (L.17) holds for all t , (L.3) in Definition L.4 holds after taking the supremum over t .

This completes the proof. $\square$

Equipped with Lemma L.5, we now derive the the covering number bounds of loss function class under transformer weights configuration in Theorem I.1 and Theorem I.2.

Lemma L.6 (Covering Number Bounds for $\mathcal{S}(D)$). Let $\epsilon_c > 0$. We define the loss function class by $\mathcal{S}(D) := \{\ell(x; u_{\theta}) : \mathcal{D} \rightarrow \mathbb{R} | u_{\theta} \in \mathcal{T}_R^{h,s,r}\}$, where $\mathcal{D} := [-D, D]^{d_x}$ for some $D > 0$. Given a fixed set of i.i.d. sample $\{x_i\}_{i=1}^n$ drawn from the target distribution q , we define the norm of loss functions by $\|\ell(x; u_{\theta})\|_{q^n} := \sup_{x_i} |\ell^{\text{trunc}}(x_i; u_{\theta}(x_i))|$. Then, under parameter configuration in Theorem I.1 and Theorem I.2, the ϵ_c -covering number of $\mathcal{S}(D)$ with respect to $\|\cdot\|_{\infty \mathcal{D}}$ satisfies:

$$\log \mathcal{N}(\epsilon_c, \mathcal{S}(D), \{x_i\}_{i=1}^n, \|\cdot\|_{q^n}) \leq O\left(\frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^4 N^{16\beta d + 12\beta} (\log N)^{20d_x + 4\beta + 17}\right).$$

For all $f(x, t) \in \mathcal{T}_R^{h,s,r}$, $t \in [0, 1]$ and $\{x_i\}_{i=1}^n \sim q^n$, we equip the transformers with the norm:

$$\|f(x, t)\|_{q^n, \mathcal{D}} := \|f(x_i, t) \mathbb{1}\{\|x_i\|_{2,\infty} \leq D\}\|_2.$$

Then, the ϵ_c -covering number of the transformer network class satisfies:

$$\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, q^n \times [0, 1], \|\cdot\|_{q^n, \mathcal{D}}) \leq O\left(\frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^2 N^{16\beta d + 12\beta} (\log N)^{20d_x + 4\beta + 16}\right).$$

Proof. First, we apply transformers parameter bounds in Theorem I.1 and Theorem I.2. Then, we extend the covering number bound to loss function calss $\mathcal{S}(D)$.

- **Log-Covering Number of Transformers Network Class.** From [Theorem I.1](#), we have

$$C_{KQ}, C_{KQ}^{2,\infty} = O(\lambda^{-1} N^{4\beta d+2\beta} (\log N)^{4d_x+2}); C_{OV}, C_{OV}^{2,\infty} = O(N^{-\beta});$$

$$C_F, C_F^{2,\infty} = O(N^\beta (\log N)^{\frac{d_x+\beta}{2}+1}); C_E = O(I); C_{\mathcal{T}} = O(\sqrt{\log N}).$$

By [Lemma L.5](#), the bounds on log-covering number follow

$$\begin{aligned} & \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, q^n \times [0, 1], \|\cdot\|_2) \\ & \leq \frac{\alpha^2 \log nL/\epsilon_c}{\epsilon_c^2} \left((C_F^{2,\infty})^{\frac{4}{3}} + (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^{\frac{2}{3}} + ((C_F)^2 C_{OV}^{2,\infty})^{\frac{2}{3}} \right)^3 \\ & \lesssim \frac{\alpha^2 \log(nL/\epsilon_c)}{\epsilon_c^2} (\lambda(C_F)^2 C_{OV} C_{KQ}^{2,\infty})^2, \quad (\text{By dropping lower order terms}) \end{aligned}$$

where

$$\begin{aligned} & (C_F)^2 C_{OV} C_{KQ}^{2,\infty} \\ & = O(\underbrace{N^{4\beta} (\log N)^{2d_x+2\beta+4}}_{(C_F)^4} \underbrace{N^{-2\beta}}_{(C_{OV})^2} \underbrace{N^{8\beta d+4\beta} (\log N)^{8d_x+4}}_{(\lambda C_{KQ}^{2,\infty})^2}) \\ & = O(N^{8\beta d+6\beta} (\log N)^{10d_x+2\beta+8}). \end{aligned}$$

Therefore,

$$\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, q^n \times [0, 1], \|\cdot\|_2) \lesssim \frac{\alpha^2 \log(nL_{\mathcal{T}})}{\epsilon_c^2} (N^{8\beta d+6\beta} (\log N)^{10d_x+2\beta+8}).$$

By [Lemma L.5](#), we have

$$\begin{aligned} \alpha & \lesssim (C_F)^2 C_{OV} C_{KQ} (D + C_E) \\ & \lesssim \underbrace{N^{2\beta} (\log N)^{d_x+\beta+2}}_{(C_F)^2} \underbrace{N^{-\beta}}_{(C_{OV})} \underbrace{N^{4\beta d+2\beta} (\log N)^{4d_x+2}}_{(\lambda C_{KQ})} (D + C_E) \quad (\text{By the definition of } \alpha) \\ & = O(D N^{4\beta d+3\beta} (\log N)^{5d_x+\beta+4}). \end{aligned}$$

Altogether, we have

$$\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, q^n \times [0, 1], \|\cdot\|_2) \lesssim \frac{\log(nL_{\mathcal{T}})}{\epsilon_c^2} D^2 N^{16\beta d+12\beta} (\log N)^{20d_x+4\beta+16}.$$

Further, by $\|\cdot\|_\infty \leq \|\cdot\|_2$, we have that

$$\log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, q^n \times [0, 1], \|\cdot\|_\infty) \lesssim \frac{\log(nL_{\mathcal{T}})}{\epsilon_c^2} D^2 N^{16\beta d+12\beta} (\log N)^{20d_x+4\beta+16}. \quad (\text{L.18})$$

- **Log-Covering Number of Loss Function Class.** Recall the of loss function [Definition L.1](#) and its truncated counterpart [Definition L.3](#). Let $\delta > 0$ and $u_1(x, t), u_2(x, t) \in \mathcal{T}_R^{h,r,s}$ be any two transformers satisfying $\max_i \|u_1(x_i, t) - u_2(x_i, t)\|_\infty \leq \delta$ for all $i \in [n]$.

First, we derive the upper bound on the expectation of $\|u_t(x|x_1)\|$:

$$\begin{aligned} & \mathbb{E}_{X_0 \sim N(0, I)} [\|u_t(x|x_1)\|_2] \\ & = \mathbb{E}_{X_0 \sim N(0, I)} [\|\dot{\mu}_t x + \dot{\sigma}_t X_0\|] \quad (\text{By Definition L.1}) \end{aligned} \quad (\text{L.19})$$

$$\begin{aligned}
&\leq \sqrt{\mathbb{E}_{X_0 \sim N(0, I)} [\|\dot{\mu}_t x + \dot{\sigma}_t X_0\|_2^2]} && \text{(By Jensen's inequality)} \\
&\leq \sqrt{\mathbb{E}_{X_0 \sim N(0, I)} [\dot{\mu}_t^2 \|x\|_2^2 + \dot{\sigma}_t^2 \|X_0\|_2^2]} && \text{(By expanding the } \ell_2 \text{ norm)} \\
&= \sqrt{\mathbb{E}_{X_0 \sim N(0, I)} [\dot{\mu}_t^2 \|x\|_2^2] + \dot{\sigma}_t^2} && \text{(By } X_0 \sim N(0, I)\text{)} \\
&\leq \sqrt{\dot{\mu}_t^2 D^2 + \dot{\sigma}_t^2}. && \text{(By } x \in [-D, D]^{d_x}\text{)}
\end{aligned}$$

Then, the distance between loss function $\ell_1(x; u_1)$ and $\ell_2(x; u_2)$ follows:

$$\begin{aligned}
&|\ell_1(x; u_1) - \ell_2(x; u_2)| && \text{(L.20)} \\
&= \frac{1}{T - t_0} \left| \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0, I)} [\|u_1(x, t) - u_t(x|x_1)\|_2^2 - \|u_2(x, t) - u_t(x|x_1)\|_2^2] dt \right| && \text{(By Definition L.1)} \\
&= \frac{1}{T - t_0} \left| \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0, I)} [(u_1(x, t) + u_2(x, t) - 2u_t(x|x_1))^\top (u_1(x, t) - u_2(x, t))] dt \right| \\
&\leq \frac{\delta}{T - t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0, I)} [\|u_1(x, t) + u_2(x, t) - 2u_t(x|x_1)\|] dt && \text{(By } \|u_1 - u_2\| \leq \delta\text{)} \\
&\leq \frac{\delta}{T - t_0} \int_{t_0}^T \sqrt{\mathbb{E}_{X_0 \sim N(0, I)} [\|u_1(x, t) + u_2(x, t) - 2u_t(x|x_1)\|_2^2]} dt && \text{(By Jensen's inequality)} \\
&\leq \frac{\delta}{T - t_0} \int_{t_0}^T \sqrt{2 \mathbb{E}_{X_0 \sim N(0, I)} [\|u_1(x, t) + u_2(x, t)\|_2^2 + 2\|u_t(x|x_1)\|_2^2]} dt && \text{(By expanding the } \ell_2 \text{ norm)} \\
&\lesssim \frac{\delta}{T - t_0} \int_{t_0}^T \sqrt{\mathbb{E}_{X_0 \sim N(0, I)} [\log N + 2\|u_t(x|x_1)\|_2^2]} dt && \text{(By } C_T = O(\sqrt{\log N})\text{)} \\
&\lesssim \frac{\delta}{T - t_0} \int_{t_0}^T \sqrt{\log N + \dot{\mu}_t^2 D^2 + 4\dot{\sigma}_t^2} dt && \text{(By (L.19))} \\
&\lesssim \delta \sqrt{\log N + D^2}. && \text{(By Assumption L.2)}
\end{aligned}$$

Finally, we extend the log covering number to the loss function class $\mathcal{S}(D)$ by setting

$$\epsilon'_c := \Omega(\epsilon_c \sqrt{\log N + D^2}).$$

This gives

$$\log \mathcal{N}(\epsilon'_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}}) \leq \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_{\infty}). \quad \text{(By (L.20))}$$

Therefore,

$$\begin{aligned}
&\log \mathcal{N}(\epsilon'_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}}) \\
&\leq \log \mathcal{N}(\epsilon_c, \mathcal{T}_R^{h,s,r}, \|\cdot\|_{\infty}) \\
&\lesssim \frac{\log(nL)}{\epsilon_c^2} \cdot D^2 N^{16\beta d + 12\beta} (\log N)^{20d_x + 4\beta + 16} && \text{(By (L.18))} \\
&= O\left(\frac{\log(nL)}{(\epsilon'_c)^2} D^4 N^{16\beta d + 12\beta} (\log N)^{20d_x + 4\beta + 17}\right). && \text{(By the definition of } \epsilon'_c\text{)}
\end{aligned}$$

This completes the proof. $\square$

Generalization Bound. Based on covering number bounds results in [Lemma L.6](#), we analyze the upper bound of generalization error $\left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \right|$. Note that the following distinction separates generalization bound for the flow-matching loss from classical learning theory.

The empirical risk ([Definition L.2](#)) takes the form $\ell(x; u_\theta) - \ell(x; u^*)$, where u^* denotes the ground truth velocity. While most standard losses stay non-negative almost everywhere, the flow matching loss may take negative values. We use the next lemma in ([L.25](#)), which bounds the second moment of the flow matching loss in terms of its first moment. Without it, the sign issue breaks the derivation.

Lemma L.7 (Bounds on Second Moment of Flow Matching Loss, Modified from Lemma C.1 of [[Yakovlev and Puchkin, 2025](#)]). Assume [Assumption I.1](#) and [Assumption I.3](#). Then, it holds

$$\mathbb{E}_{x \sim q} \left[\left| \ell^{\text{trunc}}(x; u_\theta) - \ell^{\text{trunc}}(x; u^*) \right|^2 \right] \lesssim \kappa \cdot \mathbb{E}_{x \sim q} \left[\ell^{\text{trunc}}(x; u_\theta) - \ell^{\text{trunc}}(x; u^*) \right],$$

where $\kappa := D^2 + \sqrt{\log N}$.

Proof. Recall [Definition L.1](#) and [Definition L.2](#). We have

$$\ell^{\text{trunc}}(x; u_\theta) := \ell(x; u_\theta) \mathbb{1}_{\{\|x\|_\infty \leq D\}} \quad \text{and} \quad \hat{\mathcal{R}}(u_\theta) = \frac{1}{n} \sum_{i=1}^n \ell(x_i; u_\theta) - \frac{1}{n} \sum_{i=1}^n \ell(x_i; u^*),$$

where $u^*(x, t) = \frac{1}{p_t(x)} \cdot \int_{\mathbb{R}^{d_x}} u_t(x|x_1) p_t(x|x_1) q(x_1) dx_1$ is the ground truth velocity and

$$\ell(x; u_\theta) := \frac{1}{T - t_0} \int_{t_0}^T \mathbb{E}_{X_0 \sim N(0, I)} [\|u_\theta(\mu_t x + \sigma_t X_0, t) - (\mu_t x + \sigma_t X_0)\|_2^2] dt.$$

For any x_i , the flow matching loss takes the form $\ell(x_i; u_\theta) - \ell(x_i; u^*)$. To simplify notation, we omit the indicator $\mathbb{1}_{\{\|x\|_\infty \leq D\}}$ when expanding ℓ^{trunc} , with the understanding that we focus only on the bounded domain where the flow matching loss is defined. Then, we compute

$$\begin{aligned} & \left| \ell^{\text{trunc}}(x; u_\theta) - \ell^{\text{trunc}}(x; u^*) \right| \\ &= \left| \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{X_0 \sim N(0, I)} [\|u_\theta - (\mu_t x + \sigma_t X_0)\|_2^2 - \|u^* - (\mu_t x + \sigma_t X_0)\|_2^2] dt \right| \\ &= \left| \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{X_0 \sim N(0, I)} [(u_\theta - u^*)^\top (u_\theta + u^* - 2 \cdot (\mu_t x + \sigma_t X_0))] dt \right| \\ &\leq \underbrace{\left(\int_{t_0}^T \frac{1}{T - t_0} \mathbb{E} [\|u_\theta - u^*\|_2^2] dt \right)^{\frac{1}{2}} \cdot \left(\int_{t_0}^T \frac{1}{T - t_0} \mathbb{E} [\|u_\theta + u^* - 2 \cdot (\mu_t x + \sigma_t X_0)\|_2^2] dt \right)^{\frac{1}{2}}}_{(A)}, \end{aligned} \tag{L.21}$$

where we apply the Cauchy-Schwarz inequality for the last inequality. Next, we bound (A) using previous results for the bounds on the true velocity, conditional velocity and transformer network.

Recall [Lemma J.4](#). It holds

$$\|u^*\|_\infty \leq \frac{|\dot{\mu}_t|}{\mu_t} \cdot \|x\|_\infty + C_5 \left| \frac{\dot{\mu}_t}{\mu_t} - \frac{\dot{\sigma}_t}{\sigma_t} \right| \cdot (\|x\|_2 + 1),$$

and by [Assumption I.3](#) we have $\|u^*\|_\infty^2 \lesssim \|x\|_2^2$ and here we consider bounded domain $\|x\|_\infty \leq D$.

Further, under the transformer network configuration in either [Theorem I.1](#) or [Theorem I.2](#), we have the transformer output bounds $C_{\mathcal{T}} = O(\sqrt{\log N})$. Lastly, for $\dot{\mu}_t x + \dot{\sigma}_t X_0$, it holds:

$$\mathbb{E}_{X_0 \sim N(0, I)} [\|\dot{\mu}_t x + \dot{\sigma}_t X_0\|_2^2] \leq \mathbb{E}_{X_0 \sim N(0, I)} [\|\dot{\mu}_t x\|_2^2 + \|\dot{\sigma}_t X_0\|_2^2] \lesssim D^2,$$

where we invoke [Assumption I.3](#) and $\|x\|_2^2 \leq d_x D^2$ in the last inequality.

Altogether, we have

$$(A) \leq \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E} [\|u_\theta\|_2^2 + \|u^*\|_2^2 + \|2 \cdot (\dot{\mu}_t x + \dot{\sigma}_t X_0)\|_2^2] dt \lesssim D^2 + \sqrt{\log N}.$$

Therefore, [\(L.21\)](#) becomes:

$$|\ell^{\text{trunc}}(x; u_\theta) - \ell^{\text{trunc}}(x; u^*)|^2 \lesssim \left(\int_{t_0}^T \frac{1}{T - t_0} \mathbb{E} [\|u_\theta - u^*\|_2^2] dt \right) \cdot (D^2 + \sqrt{\log N}).$$

Then, we conclude that

$$\begin{aligned} & \mathbb{E}_{x \sim q} [\ell^{\text{trunc}}(x; u_\theta) - \ell^{\text{trunc}}(x; u^*)]^2 \\ & \lesssim (D^2 + \sqrt{\log N}) \cdot \int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{x \sim q} \left[\mathbb{E}_{X_0 \sim N(0, I)} [\|u_\theta - u^*\|_2^2] dt \right] \\ & = (D^2 + \sqrt{\log N}) \cdot \underbrace{\int_{t_0}^T \frac{1}{T - t_0} \mathbb{E}_{x_t \sim p_t} [\|u_\theta - u^*\|_2^2] dt}_{(B)} \quad (\text{By tower property}) \\ & = (D^2 + \sqrt{\log N}) \cdot \mathbb{E}_{x \sim q} [\ell^{\text{trunc}}(x; u_\theta) - \ell^{\text{trunc}}(x; u^*)]. \quad (\text{By Remark L.2}) \end{aligned}$$

We remark that (B) is the conditional flow matching risk $\mathcal{R}(u_\theta)$ defined in [\(L.1\)](#).

This completes the proof. $\square$

Lemma L.8 (Generalization Bound, Modified from the Theorem C.4 of [\[Oko et al., 2023\]](#)). Let $\hat{u}_\theta$ be the velocity estimator trained by optimizing $\mathcal{L}_{\text{CFM}}(u_\theta)$ following [Definition L.1](#) with i.i.d training samples $\{x_i\}_{i=1}^n$. For $\epsilon_c > 0$, let $\mathcal{N} := \mathcal{N}(\epsilon_c, \mathcal{S}(D), q^n, \|\cdot\|_\infty)$ be the covering number of function class of loss $\mathcal{S}(D)$ following [Lemma L.6](#). Then we bound the generalization error:

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \lesssim \mathbb{E}_{\{x_i\}_{i=1}^n} [\hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right).$$

Proof. We use $\hat{\mathcal{L}}'_{\text{CFM}}$ and $\hat{\mathcal{R}}'$ to denote the conditional flow matching loss and empirical risk with ghost training samples $\{x'_i\}_{i=1}^n$. Further, let u^* denote the ground truth velocity field.

Then, following [Remark L.3](#), we rewrite the generalization error:

$$\begin{aligned} & \left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \right| \quad (\text{L.22}) \\ & = \left| \mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} [\hat{\mathcal{R}}'^{\text{trunc}}(\hat{u}_\theta)] - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) \right] \right| \quad (\text{By Remark L.3}) \\ & = \left| \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\hat{\mathcal{R}}'^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \right| \quad (\text{By the independence between } x'_i \text{ and } \hat{\mathcal{R}}(\hat{u}_\theta)) \end{aligned}$$

$$= \left| \frac{1}{n} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\left(\sum_{i=1}^n \ell^{\text{trunc}}(x'_i; \hat{u}_\theta) - \sum_{i=1}^n \ell^{\text{trunc}}(x'_i; u^*) \right) - \left(\sum_{i=1}^n \ell^{\text{trunc}}(x_i; \hat{u}_\theta) - \sum_{i=1}^n \ell^{\text{trunc}}(x_i; u^*) \right) \right] \right|. \quad (\text{By Definition L.2})$$

For $\epsilon_c > 0$ to be chosen later, let $\mathcal{J} := \{\ell_1, \ell_2, \dots, \ell_N\}$ be a ϵ_c -covering of the loss function class $\mathcal{S}(\mathcal{D})$ with the minimum cardinality in the L_∞ metric. Note that $\ell_1, \dots, \ell_N$ have domain $\mathcal{D} = [-D, D]^{d_x}$ by Definition L.3 and Definition L.4. Further, let J be the random variable such that $\|\ell(\cdot, \hat{u}_\theta) - \ell_J(\cdot, u_J)\|_\infty \leq \epsilon_c$. Moreover, we introduce following definitions for simplicity:

$$\begin{aligned} \omega(x) &:= \ell^{\text{trunc}}(x; \hat{u}_\theta) - \ell^{\text{trunc}}(x; u^*), \\ \omega_j(x) &:= \ell_j(x; u_j) - \ell^{\text{trunc}}(x; u^*), \\ h_j &:= \max \left\{ A, \sqrt{\mathbb{E}_z[\ell_j(z; u_j) - \ell^{\text{trunc}}(z; u^*)]} \right\}, \\ \Omega &:= \max_{1 \leq j \leq N} \left| \sum_{i=1}^n \frac{\omega_j(x'_i) - \omega_j(x_i)}{h_j} \right|, \end{aligned}$$

where $z \sim q$ is independent of $\{x_i, x'_i\}_{i=1}^n$. Then we can further bound (L.22) as follows:

$$\begin{aligned} & \left| \frac{1}{n} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\left(\sum_{i=1}^n \ell^{\text{trunc}}(x'_i; \hat{u}_\theta) - \sum_{i=1}^n \ell^{\text{trunc}}(x'_i; u^*) \right) - \left(\sum_{i=1}^n \ell^{\text{trunc}}(x_i; \hat{u}_\theta) - \sum_{i=1}^n \ell^{\text{trunc}}(x_i; u^*) \right) \right] \right| \\ & \quad \quad \quad (\text{L.23}) \\ & \leq \left| \frac{1}{n} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\left(\sum_{i=1}^n (\omega_J(x'_i) - \omega_J(x_i)) \right) \right] \right| + 2\epsilon_c \quad (\text{By the definitions of } \omega_J \text{ and covering number}) \\ & \leq \frac{1}{n} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\left| \left(\sum_{i=1}^n (\omega_J(x'_i) - \omega_J(x_i)) \right) \right| \right] + 2\epsilon_c \quad (\text{By the property of expectation}) \\ & \leq \frac{1}{n} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J \Omega] + 2\epsilon_c \quad (\text{By the definitions of } h_j \text{ and } \Omega) \\ & \leq \frac{1}{n} \sqrt{\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2] \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\Omega^2]} + 2\epsilon_c \quad (\text{By Cauchy-Schwarz inequality}) \\ & \leq \frac{1}{n} \left(\frac{n}{2} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2] + \frac{1}{2n} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\Omega^2] \right) + 2\epsilon_c \quad (\text{By AM-GM Inequality}) \\ & = \frac{1}{2} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2] + \frac{1}{2n^2} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\Omega^2] + 2\epsilon_c. \end{aligned}$$

Now we bound $\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2]$ and $\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\Omega^2]$ separately. For $\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2]$, we have

$$\begin{aligned} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2] &\leq A^2 + \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\mathbb{E}_z[\ell_j(z; u_J) - \ell^{\text{trunc}}(z; u^*)]] \quad (\text{By the definition of } h_j) \\ &\leq A^2 + \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\mathbb{E}_z[\ell^{\text{trunc}}(z; \hat{u}_\theta) - \ell^{\text{trunc}}(z; u^*)]] + 2\epsilon_c \quad (\text{By the definition of } \epsilon_c) \\ &\leq A^2 + \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta)] + 2\epsilon_c. \quad (\text{By Remark L.3}) \end{aligned}$$

Then we start to bound $\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\Omega^2]$. By the definition of $\omega_j(x)$ and the independence between $\{x_i\}_{i=1}^n$ and $\{x'_i\}_{i=1}^n$, we have

$$\begin{aligned} & \mathbb{E}_{x_i, x'_i} \left[\frac{\omega_j(x_i) \omega_j(x'_i)}{h_j^2} \right] \\ &= \frac{1}{h_j^2} \mathbb{E}_{x_i} [\omega_j(x_i)] \cdot \mathbb{E}_{x'_i} [\omega_j(x'_i)] \quad (\text{By the independence between } h_j \text{ and } \{x_i, x'_i\}_{i=1}^n) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{h_j^2} \left(\mathbb{E}_{x_i} [\omega_j(x_i)] \right)^2 && \text{(By the independence between } w_j \text{ and } \{x_i, x'_i\}_{i=1}^n) \\
&\geq 0. && \text{(L.24)}
\end{aligned}$$

To use Bernstein's Inequality, for any j , we bound the following expectation as

$$\begin{aligned}
&\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\sum_{i=1}^n \left(\frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \right)^2 \right] \\
&= \sum_{i=1}^n \left(\mathbb{E}_{x_i, x'_i} \left[\left(\frac{\omega_j(x_i)}{h_j} \right)^2 + \left(\frac{\omega_j(x'_i)}{h_j} \right)^2 \right] - 2 \mathbb{E}_{x_i, x'_i} \left[\frac{\omega_j(x_i) \omega_j(x'_i)}{h_j^2} \right] \right) \\
&\leq \sum_{i=1}^n \mathbb{E}_{x_i, x'_i} \left[\left(\frac{\omega_j(x)}{h_j} \right)^2 + \left(\frac{\omega_j(x')}{h_j} \right)^2 \right]. && \text{(By (L.24))}
\end{aligned}$$

Recall that for any $j \in [\mathcal{N}]$, $\omega_j(x) := \ell_j(x; u_j) - \ell^{\text{trunc}}(x; u^*)$. For any $\ell \in \mathcal{S}(D)$, assume $|\ell^{\text{trunc}}(\cdot; u_\theta)| \leq \kappa$, then for any $i \in [n], j \in [\mathcal{N}]$, we have $\mathbb{E}_{x_i, x'_i} [\omega_j(x_i)] = \mathbb{E}_{x_i, x'_i} [\omega_j(x'_i)]$, which leads to

$$\begin{aligned}
\mathbb{E}_{x_i, x'_i} [\omega_j(x_i)] &= \mathbb{E}_{x_i, x'_i} [\omega_j(x'_i)] \\
&= \mathbb{E}_{x_i, x'_i} [\ell_j(x'_i; u_j) - \ell^{\text{trunc}}(x'_i; u^*)] && \text{(By the definition of } \omega_j(x)) \\
&= \mathbb{E}_z [\ell_j(z; u_j) - \ell^{\text{trunc}}(z; u^*)] \\
&\leq \mathbb{E}_{x_i, x'_i} [h_j^2]. && \text{(By the definition of } h_j)
\end{aligned}$$

Then, it holds

$$\begin{aligned}
&\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\sum_{i=1}^n \left(\frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \right)^2 \right] \\
&\leq \sum_{i=1}^n \mathbb{E}_{x_i, x'_i} \left[\left(\frac{\omega_j(x_i)}{h_j} \right)^2 + \left(\frac{\omega_j(x'_i)}{h_j} \right)^2 \right] \\
&\leq 2\kappa \sum_{i=1}^n \mathbb{E}_{x_i, x'_i} \left[\left(\frac{\omega_j(x_i)}{h_j^2} \right) + \left(\frac{\omega_j(x'_i)}{h_j^2} \right) \right] && \text{(By Lemma L.7)} \\
&\leq 4n\kappa. && \text{(L.25)}
\end{aligned}$$

Since $\left| \frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \right| \leq \frac{\kappa}{A}$ and $\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} \left[\frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \right] = 0$, by Bernstein's Inequality, we have for any $j \in [\mathcal{N}]$, $h > 0$,

$$\begin{aligned}
\Pr \left[\left(\sum_{i=1}^n \frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \right)^2 \geq h \right] &= 2 \Pr \left[\sum_{i=1}^n \frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \geq \sqrt{h} \right] \\
&\leq 2 \exp \left(- \frac{h/2}{\kappa(4n + \frac{\sqrt{h}}{3A})} \right).
\end{aligned}$$

Thus, we have

$$\Pr [\Omega^2 \geq h] \leq \sum_{j=1}^{\mathcal{N}} \Pr \left[\left(\sum_{i=1}^n \frac{\omega_j(x_i) - \omega_j(x'_i)}{h_j} \right)^2 \geq h \right] \quad \text{(By union bound.)}$$

$$\leq 2\mathcal{N} \exp\left(-\frac{h/2}{\kappa(4n + \frac{\sqrt{h}}{3A})}\right).$$

Thus, for any $h_0 > 0$, we bound $\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n}[\Omega^2]$ as

$$\begin{aligned} & \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n}[\Omega^2] \\ &= \int_0^{h_0} \Pr[\Omega^2 \geq h] dh + \int_{h_0}^{\infty} \Pr[\Omega^2 \geq h] dh \\ &\leq h_0 + \int_{h_0}^{\infty} 2\mathcal{N} \exp\left(-\frac{h/2}{\kappa(4n + \frac{\sqrt{h}}{3A})}\right) dh && \text{(By tail-sum formula)} \\ &\leq h_0 + 2\mathcal{N} \int_{h_0}^{\infty} \left[\exp\left(-\frac{h}{16\kappa n}\right) + \exp\left(-\frac{3A\sqrt{h}}{4\kappa}\right)\right] dh \\ &\leq h_0 + 2\mathcal{N} \left[16\kappa n \exp\left(-\frac{h_0}{16\kappa n}\right) + \left(\frac{8\kappa\sqrt{h_0}}{3A} + \frac{32\kappa}{9A^2}\right) \exp\left(-\frac{3A\sqrt{h_0}}{4\kappa}\right)\right]. \end{aligned}$$

Taking $A = \frac{\sqrt{h_0}}{12n}$ and $h_0 = 16\kappa n \log \mathcal{N}$, we have

$$\mathbb{E}_{\{x_i, x'_i\}_{i=1}^n}[\Omega^2] \lesssim n\kappa \log \mathcal{N}.$$

Combining above, we bound the generalization error as

$$\begin{aligned} & \left| \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \right| \\ &\leq \frac{1}{2} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [h_J^2] + \frac{1}{2n^2} \mathbb{E}_{\{x_i, x'_i\}_{i=1}^n} [\Omega^2] + 2\epsilon_c && \text{(By (L.22))} \\ &\leq \frac{1}{2} \left(A^2 + \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta)] + 2\epsilon_c \right) + \frac{1}{2n^2} O(n\kappa \log \mathcal{N}) \\ &\lesssim \frac{1}{2} \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta)] + O\left(\frac{\kappa}{n} \log \mathcal{N} + \epsilon_c\right). \end{aligned}$$

This implies

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta)] \lesssim 2 \cdot \mathbb{E}_{\{x_i\}_{i=1}^n} [\hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] + O\left(\frac{\kappa}{n} \log \mathcal{N} + \epsilon_c\right).$$

Therefore,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \lesssim \mathbb{E}_{\{x_i\}_{i=1}^n} [\hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] + O\left(\frac{\kappa}{n} \log \mathcal{N} + \epsilon_c\right).$$

This completes the proof. $\square$

L.3 Main Proof of Theorem I.3

We now give the formal proof of Theorem I.3.

Theorem L.1 (Theorem I.3 Restated: Velocity Estimation with Transformer). Let d be the feature dimension. Suppose we choose the transformers as in Theorem I.1 and Theorem I.2 correspondingly, then we have

- Assume **Assumption I.1** and **Assumption I.2**. Then,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{16d+15}} (\log n)^{20d_x+4\beta+20}).$$

- Assume **Assumption I.2** and **Assumption I.3**. Then,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{8d+9}} (\log n)^{20d_x+4\beta+20}).$$

Proof of Theorem I.3. Let $\{x'_i\}_{i=1}^n$ be a different set of i.i.d samples independent of the training sample $\{x_i\}_{i=1}^n$. Further, we use $\hat{\mathcal{R}}'$ to denote the empirical risk with samples $\{x'_i\}_{i=1}^n$.

Then, following (L.2), we decompose $\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)]$ as:

$$\begin{aligned} \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] &= \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} \left[\hat{\mathcal{R}}'(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) \right] \right]}_{\text{(I)}} \\ &\quad + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} \left[\hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) \right] - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) \right]}_{\text{(II)}} \\ &\quad + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}(\hat{u}_\theta) \right]}_{\text{(III)}} + \underbrace{\mathbb{E}_{\{x_i\}_{i=1}^n} \left[\hat{\mathcal{R}}(\hat{u}_\theta) \right]}_{\text{(IV)}}. \end{aligned}$$

Then, we bound each term and incorporate them to obtain the upper bound on the estimation error.

- **Bound (I) and (III).** By **Lemma L.1**, term (I) and term (III) are upper bounded by

$$\text{(I), (III)} \lesssim D^{d_x} \exp\left(-\frac{1}{2}C_2 D^2\right) \log N.$$

- **Bound (II).** By the generalization error bound (**Lemma L.8**), we have

$$\begin{aligned} \text{(II)} &= \mathbb{E}_{\{x_i\}_{i=1}^n} \left[\mathbb{E}_{\{x'_i\}_{i=1}^n} \left[\hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) \right] - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta) \right] \tag{L.26} \\ &= \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta) - \hat{\mathcal{R}}^{\text{trunc}}(\hat{u}_\theta)] \quad (\text{By } \mathbb{E}_{\{x'_i\}_{i=1}^n} [\hat{\mathcal{R}}^{\text{trunc}}] = \mathcal{R}^{\text{trunc}}) \\ &\lesssim \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}^{\text{trunc}}(\hat{u}_\theta)] + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) \quad (\text{By Lemma L.8}) \\ &\lesssim \text{(IV)} + D^{d_x} \exp\left(-\frac{1}{2}C_2 D^2\right) \log N + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right). \quad (\text{By Lemma L.1}) \end{aligned}$$

where $\mathcal{N}(\epsilon_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}})$ is the covering number (**Definition L.4**) of loss function class.

- **Bound (IV).** Recall that $\hat{\mathcal{R}}(\hat{u}_\theta) := \hat{\mathcal{L}}_{\text{CFM}}(\hat{u}_\theta) - \hat{\mathcal{L}}_{\text{CFM}}(u^*)$ and $\hat{u}_\theta$ is trained by optimizing $\hat{\mathcal{L}}_{\text{CFM}}(u_\theta)$ following **Definition L.2**. Therefore, for any velocity estimator u_θ , it holds

$$\hat{\mathcal{R}}(\hat{u}_\theta) \leq \hat{\mathcal{L}}_{\text{CFM}}(u_\theta) - \hat{\mathcal{L}}_{\text{CFM}}(u^*) = \hat{\mathcal{R}}(u_\theta).$$

Then, for any velocity estimator u_θ , it holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}(\widehat{u}_\theta)] \leq \mathbb{E}_{\{x_i\}_{i=1}^n} [\widehat{\mathcal{R}}(u_\theta)] = \mathcal{R}(u_\theta). \quad (\text{L.27})$$

Altogether, the estimation error is upper bounded by

$$\begin{aligned} & \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\widehat{u}_\theta)] \\ &= (\text{I}) + (\text{II}) + (\text{III}) + (\text{IV}) \\ &\lesssim D^{d_x} \exp(-C_2 D^2) \log N + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) + 2(\text{IV}) \\ &\leq O(N^{-2\beta} (\log N)^{d_x/2+1}) + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) + 2(\text{IV}). \quad (\text{By setting } D := \sqrt{2\beta \log N / C_2}) \end{aligned} \quad (\text{L.28})$$

Furthermore, the log covering number is upper bounded by

$$\begin{aligned} & \log \mathcal{N}(\epsilon_c, \mathcal{S}(D), \|\cdot\|_{\infty \mathcal{D}}) \\ &\leq O\left(\frac{\log(nL/\epsilon_c)}{\epsilon_c^2} D^4 N^{16\beta d + 12\beta} (\log N)^{20d_x + 4\beta + 17}\right) \quad (\text{By Lemma L.6}) \\ &\leq O\left(\frac{\log(nL/\epsilon_c)}{\epsilon_c^2} N^{16\beta d + 12\beta} (\log N)^{20d_x + 4\beta + 19}\right). \quad (\text{By } D := \sqrt{2\beta \log N / C_2}) \end{aligned} \quad (\text{L.29})$$

Next, we bound the velocity field estimation error.

- **Estimation Rates under Generic Hölder Smoothness.** By Theorem I.1, it holds

$$\begin{aligned} (\text{IV}) &\leq \mathcal{R}(u_\theta(x, t)) \quad (\text{By (L.27)}) \\ &= \int \frac{1}{T - t_0} \int_{\mathbb{R}^{d_x}} \|u_t(x) - u_\theta(x, t)\|_2^2 p_t(x) dx dt \\ &= O(B^2 N^{-\beta} \cdot (\log N)^{d_x + \frac{\beta}{2} + 1}). \quad (\text{By Theorem I.1}) \end{aligned}$$

Then, (L.28) becomes

$$\begin{aligned} & \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\widehat{u}_\theta)] \\ &\leq O(N^{-2\beta} (\log N)^{d_x/2+1}) + O\left(\frac{1}{n} \log \mathcal{N} + \epsilon_c\right) + O(B^2 N^{-\beta} (\log N)^{d_x + \frac{\beta}{2} + 1}) \\ &\leq O(N^{-2\beta} (\log N)^{d_x/2+1}) + O\left(\frac{\log(nL/\epsilon_c)}{n\epsilon_c^2} N^\nu (\log N)^{20d_x + 4\beta + 19} + \epsilon_c\right) + O(B^2 N^{-\beta} (\log N)^{d_x + \frac{\beta}{2} + 1}), \end{aligned}$$

(By (L.29))

where $\nu := 16\beta d + 12\beta$.

Let $\gamma_1, \gamma_2 \in (0, 1)$ be two arbitrary numbers. We take $N = n^{\gamma_1/\nu}$ and $\epsilon_c = n^{-\gamma_2}$. Then,

$$\begin{aligned} & \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\widehat{u}_\theta)] \\ &\leq O(n^{-\frac{2\beta\gamma_1}{\nu}} (\log n)^{\frac{d_x}{2}+1}) + O(n^{-1+\gamma_1+2\gamma_2} (\log N)^{20d_x + 4\beta + 20} \log n + n^{-\gamma_2}) + O(B^2 n^{-\frac{\beta\gamma_1}{\nu}} (\log n)^{d_x + \frac{\beta}{2} + 1}) \\ &\leq O(n^{-\min\{\frac{\beta\gamma_1}{\nu}, 1-\gamma_1-2\gamma_2, \gamma_2\}} (\log n)^{20d_x + 4\beta + 21}). \end{aligned}$$

For any $\gamma_1, \gamma_2 \in (0, 1)$ satisfying

$$\gamma_1 + 2\gamma_2 < 1,$$

we consider

$$\min\{\frac{\beta\gamma_1}{\nu}, 1 - \gamma_1 - 2\gamma_2, \gamma_2\}.$$

To simplify, we set

$$\frac{\beta\gamma_1}{\nu} = 1 - \gamma_1 - 2\gamma_2 = \gamma_2,$$

giving

$$\gamma_1 = \frac{\nu}{\nu + 3\beta}, \quad \gamma_2 = \frac{\beta}{\nu + 3\beta}.$$

Therefore,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{16d+15}} (\log n)^{20d_x+4\beta+21}).$$

- **Estimation Rates under Stronger Hölder Smoothness.** By [Theorem I.2](#), it holds

$$\begin{aligned} \text{(IV)} &\leq \mathcal{R}(u_\theta(x, t)) && \text{(By (L.27))} \\ &= \int \int \|u_t(x) - u_\theta(x, t)\|_2^2 \cdot p_t(x) dx dt \\ &= O(B^2 N^{-2\beta} (\log N)^{d_x+\beta}). \end{aligned}$$

Then, [\(L.28\)](#) becomes

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] \\ &\leq O(N^{-2\beta} (\log N)^{\frac{d_x}{2}+1}) + O(\frac{1}{n} \log \mathcal{N} + \epsilon_c) + O(B^2 N^{-2\beta} (\log N)^{d_x+\beta}) \\ &\leq O(N^{-2\beta} (\log N)^{\frac{d_x}{2}+1}) + O(\frac{\log n}{n\epsilon_c^2} N^\nu (\log N)^{20d_x+4\beta+19} + \epsilon_c) + O(B^2 N^{-2\beta} (\log N)^{d_x+\beta}), \end{aligned}$$

(By [\(L.29\)](#))

where $\nu := 16\beta d + 12\beta$.

Let $\gamma_3, \gamma_4 \in (0, 1)$ be two arbitrary numbers. We take $N = n^{\gamma_3/\nu}$ and $\epsilon_c = n^{-\gamma_4}$. Then,

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] \\ &\leq O(n^{-\frac{2\beta\gamma_3}{\nu}} (\log n)^{\frac{d_x}{2}+1}) + O(n^{-1} n^{2\gamma_4} n^{\gamma_3} (\log n)^{20d_x+4\beta+20} + n^{-\gamma_4}) + O(B^2 n^{-\frac{2\beta\gamma_3}{\nu}} (\log n)^{d_x+\beta}) \\ &\leq O(n^{-\min\{\frac{2\beta\gamma_3}{\nu}, 1-\gamma_3-2\gamma_4, \gamma_4\}} (\log n)^{20d_x+4\beta+20}). \end{aligned}$$

For any $\gamma_3, \gamma_4 \in (0, 1)$ satisfying

$$\gamma_3 + 2\gamma_4 < 1,$$

we consider

$$\min\{\frac{2\beta\gamma_3}{\nu}, 1 - \gamma_3 - 2\gamma_4, \gamma_4\}.$$

To simplify, we set

$$\frac{2\beta\gamma_3}{\nu} = 1 - \gamma_3 - 2\gamma_4 = \gamma_4,$$

giving

$$\gamma_3 = \frac{\nu}{\nu + 6\beta}, \quad \gamma_4 = \frac{2\beta}{\nu + 6\beta}.$$

Therefore,

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] = O(n^{-\frac{1}{8d+9}} (\log n)^{20d_x+4\beta+20}).$$

This completes the proof. □

M Proof of Theorem I.4

In this section, we apply the Grönwall's inequality and the Alekseev–Gröbner lemma to extend the velocity estimation to distribution estimation under 2-Wasserstein distance.

Organizations. Appendix M.1 introduces auxiliary lemmas. Appendix M.2 presents the main proof.

M.1 Auxiliary Lemmas

In this section, we introduce auxiliary lemmas for extending the velocity estimation to distribution estimation in 2-Wasserstein distance. Specifically, we state the Grönwall's inequality in Lemma M.1. Furthermore, we introduce the Alekseev–Gröbner lemma that quantifies the deviation between solutions of two distinct ODEs in terms of the discrepancy between their velocity in Lemma M.2.

We begin with the Grönwall's inequality.

Lemma M.1 (Grönwall's Inequality, [Gronwall, 1919]). Let $a, b \in \mathbb{R}$ with $a < b$. Let $g(t)$ and $y(t)$ be two real-valued continuous functions defined on $[a, b]$. Then, if $y(t)$ is differentiable on $[a, b]$ and satisfies:

$$\frac{d}{dt}y(t) \leq y(t)g(t), \quad t \in [a, b],$$

it holds

$$y(t) \leq y(a) \exp\left(\int_a^b g(s)ds\right).$$

Next, we introduce the Alekseev–Gröbner lemma.

Lemma M.2 (Alekseev–Gröbner Lemma, Lemma 16 of [Fukumizu et al., 2024], Proposition 2 of [Benton et al., 2023], Theorem 14.5 of [Hairer et al., 1993]). Let $u(x, t)$ and $u_\theta(x, t)$ be smooth vector fields and $\psi(x, s, t)$ and $\psi_\theta(x, s, t)$ be the respective flows defined for $t \geq s$ that satisfy

$$\begin{aligned} \frac{d}{dt}\psi(x, s, t) &= u(\psi(x, s, t), t), & \psi(x, s, s) &= x \\ \frac{d}{dt}\psi_\theta(x, s, t) &= u_\theta(\psi_\theta(x, s, t), t), & \psi_\theta(x, s, s) &= x. \end{aligned}$$

Then,

$$\psi_\theta(x, t_0, T) - \psi(x, t_0, T) = \int_{t_0}^T D\psi_\theta(\psi(x, t_0, s), s, T)(u_\theta(\psi(x, t_0, s), s) - u(\psi(x, t_0, s), s))ds,$$

where the partial derivatives in the Jacobian matrix $D\psi_\theta(\psi(x, t_0, s), s, T)$ is with respect to its first argument.

M.2 Main Proof of Theorem I.4

We now present the main proof of Theorem I.4.

Theorem M.1 (Theorem I.4 Restated: Distribution Estimation under Wasserstein Distance). Let $\hat{P}_T$ denote the estimated distribution at time T . Further, we define a constant $\nu := 16(L + 1) + 12/d$.

- Assume Assumption I.1 and Assumption I.2. It holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] = O(n^{-\frac{1}{32d+30}} (\log n)^{10d_x+2\beta+10}).$$

- Assume [Assumption I.2](#) and [Assumption I.3](#). It holds

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] = O(n^{-\frac{1}{16d+18}} (\log n)^{10d_x+2\beta+10}).$$

Proof of [Theorem I.4](#). We bound the 2-Wasserstein distance between the estimated and true distributions with the ℓ_2 difference of the velocity field network and the true velocity field. Our proof structure follows [\[Fukumizu et al., 2024, Theorem 3\]](#) and [\[Benton et al., 2023, Theorem 1\]](#).

The distributions $\hat{P}_T$ and P_T are the pushforwards of P_{t_0} by $\psi_\theta(\cdot, t_0, T)$ and $\psi(\cdot, t_0, T)$. Thus, using the definition of the 2-Wasserstein metric, it follows that

$$W_2(\hat{P}_T, P_T) \leq \sqrt{\mathbb{E}_{x \sim p_{t_0}} [\|\psi_\theta(x, t_0, T) - \psi(x, t_0, T)\|_2^2]}.$$

We use [Lemma M.2](#) to bound the ℓ_2 difference of the flows. To that end, let us first bound the Jacobian matrix $D\psi_\theta(\psi(x, t_0, s), s, t)$. We have

$$\begin{aligned} & \frac{\partial}{\partial t} \|D\psi_\theta(\psi(x, t_0, s), s, t)\|_2 \\ & \leq \left\| \frac{\partial}{\partial t} D\psi_\theta(\psi(x, t_0, s), s, t) \right\|_2 \\ & = \|Du_\theta(\psi_\theta(\psi(x, t_0, s), t), s, t) D\psi_\theta(\psi(x, t_0, s), s, t)\|_2 \\ & \leq L_{\mathcal{T}} \|D\psi_\theta(\psi(x, t_0, s), s, t)\|_2, \end{aligned}$$

where the first inequality follows from triangle inequality of the $\|\cdot\|_2$ -norm, and the second equality follows from the flow ODE in the assumption of [Lemma M.2](#), and the third inequality follows from the Lipschitzness of transformer network ([Definition B.2](#)). Therefore,

$$\|D\psi_\theta(\psi(x, t_0, s), s, t)\|_2 \lesssim \exp\left\{\int_s^t L_{\mathcal{T}} du\right\} \leq \exp\left\{\int_0^1 L_{\mathcal{T}} du\right\} =: M. \quad (\text{By [Lemma M.1](#)})$$

Now we have

$$\begin{aligned} & \|\psi_\theta(x, t_0, T) - \psi(x, t_0, T)\|_2^2 \\ & \leq M^2 \cdot \left(\int_{t_0}^T \|u_\theta(\psi(x, t_0, s), s) - u(\psi(x, t_0, s), s)\|_2 ds\right)^2 \\ & \leq M^2 \int_{t_0}^T \|u_\theta(\psi(x, t_0, s), s) - u(\psi(x, t_0, s), s)\|_2^2 ds, \end{aligned}$$

where in the first line we apply [Lemma M.2](#) and in the second line we apply the Hölder's inequality. Then, we take expectation with respect to $x \sim p_{t_0}$ on both sides of the above inequality

$$\begin{aligned} \mathbb{E}_{x \sim p_{t_0}} [\|\psi_\theta(x, t_0, T) - \psi(x, t_0, T)\|_2^2] & \leq M^2 \mathbb{E}_{x \sim p_{t_0}} \left[\int_{t_0}^T \|u_\theta(\psi(x, t_0, s), s) - u(\psi(x, t_0, s), s)\|_2^2 ds \right] \\ & = M^2 \int_{t_0}^T \mathbb{E}_{x \sim p_s} [\|u_\theta(x, s) - u(x, s)\|_2^2] ds, \end{aligned}$$

where the last equality follows since the samples $\psi(x, t_0, s)$ with $x \sim p_{t_0}$ are the same as the samples $x \sim p_s$ by construction of the flow.

Therefore, we have

$$W_2(\hat{P}_T, P_T) \leq M \cdot \left(\int_{t_0}^T \mathbb{E}_{x \sim p_s} [\|u_\theta(x, s) - u(x, s)\|_2^2] ds \right)^{\frac{1}{2}},$$

where

$$\int_{t_0}^T \mathbb{E}_{x \sim p_s} [\|u_\theta(x, s) - u(x, s)\|_2^2] ds = (T - t_0) \mathcal{R}(u_\theta). \quad (\text{By Definition I.2})$$

Then, by Assumption I.2, we have

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] \leq M \cdot (T - t_0) \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{\mathcal{R}(\hat{u}_\theta)}] \lesssim M \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{\mathcal{R}(\hat{u}_\theta)}].$$

Finally, we apply the flow estimation results in Theorem I.3 and get

$$\begin{aligned} \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] &= O(n^{-\frac{1}{16d+15}} (\log n)^{20d_x+4\beta+20}), \\ \mathbb{E}_{\{x_i\}_{i=1}^n} [\mathcal{R}(\hat{u}_\theta)] &= O(n^{-\frac{1}{8d+9}} (\log n)^{20d_x+4\beta+20}), \end{aligned}$$

under Assumption I.1 and Assumption I.3 respectively. These imply

$$\begin{aligned} \mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] &\lesssim M \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{\mathcal{R}(\hat{u}_\theta)}] = O(n^{-\frac{1}{32d+30}} (\log n)^{10d_x+2\beta+10}), \\ \mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] &\lesssim M \mathbb{E}_{\{x_i\}_{i=1}^n} [\sqrt{\mathcal{R}(\hat{u}_\theta)}] = O(n^{-\frac{1}{16d+18}} (\log n)^{10d_x+2\beta+10}). \end{aligned}$$

This completes the proof. □

N Proof of Theorem I.5

In this section, we prove the nearly minimax optimality results of flow matching transformers under specified settings (Theorem I.5).

We begin with the definition of modulus of smoothness following [Oko et al., 2023].

Definition N.1 (Modulus of Smoothness). Let Ω be a domain in $\mathbb{R}^{d_x}$ and $f \in L^{p'}(\Omega)$ be a function for some $p' \in (0, \infty]$. We define the r -th modulus of smoothness of f by:

$$\omega_{r,p'}(f, t) := \sup_{\|h\|_2 \leq t} \|\Delta_h^r(f)\|_{p'},$$

where $\Delta_h^r(\Omega)$ is the difference operator defined by

$$\Delta_h^r(f)(x) := \begin{cases} \sum_{j=0}^r \binom{r}{j} (-1)^{r-j} f(x + jh), & \text{if } x + jh \in \Omega \text{ for all } j, \\ 0, & \text{otherwise.} \end{cases}$$

Next, we define the Besov space.

Definition N.2 (Besov Space $B_{p',q'}^s$). Let $0 < p', q' \leq \infty$, $s > 0$ and $r := \lfloor s \rfloor + 1$. The Besov norm of a function $f \in L^{p'}(\Omega)$ is defined by $\|f\|_{B_{p',q'}^s} := \|f\|_{p'} + |f|_{B_{p',q'}^s}$, where

$$|f|_{B_{p',q'}^s} := \begin{cases} \left(\int_0^\infty \left((t^{-s} \omega_{r,p'}(f, t))^{q'} \frac{dt}{t} \right)^{\frac{1}{q'}} dt, & q' < \infty, \\ 0, & q' = \infty. \end{cases}$$

Given $m, L > 0$ we have the Besov space $B_{p',q'}^s(L, m) := \{f \in L^{p'}(\Omega) \mid \|f\|_{B_{p',q'}^s} < L, f \geq m\}$.

The next lemma provides the minimax optimal rate for density in the Besov space $B_{p',q'}^s$.

Lemma N.1 (Theorem 3 of [Niles-Weed and Berthet, 2022]). Let $\Omega := [-1, 1]^{d_x}$ be the domain of density $q(x_1)$ in Besov space $B_{p',q'}^s(L, m)$. Then, for any $r, p', q' \geq 1$ and $s > 0$,

$$\inf_{\hat{P}} \sup_{q \in B_{p',q'}^s(L, m)} \mathbb{E} [W_r(\hat{P}, P)] \gtrsim n^{-\frac{s+1}{d_x+2s}},$$

where $\{x_i\}_{i=1}^n$ is a set of i.i.d samples drawn from distribution P , and $\hat{P}$ runs over all possible estimators constructed from the data.

Then, we revisit the definition of Wasserstein distance:

Definition N.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:

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

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

We then give the minimax optimal rate in the Hölder density function spaces.

Lemma N.2 (Modified from Theorem 3 of [Niles-Weed and Berthet, 2022]). Consider the task of estimating a probability distribution $P(x_1)$ with density function belonging to the space

$$\mathcal{P} := \{q(x_1) \mid q(x_1) \in \mathcal{H}^\beta([-1, 1]^{d_x}, B), q(x_1) \geq C\},$$

Then, for any $r \geq 1$, $\beta > 0$ and $d_x > 2$, we have

$$\inf_{\hat{P}} \sup_{q(x_1) \in \mathcal{P}} \mathbb{E}_{\{x_i\}_{i=1}^n} [W_r(\hat{P}, P)] \gtrsim n^{-\frac{\beta+1}{d_x+2\beta}},$$

where $\{x_i\}_{i=1}^n$ is a set of i.i.d samples drawn from distribution P , and $\hat{P}$ runs over all possible estimators constructed from the data.

Proof. Let Ω be some domains. Since $B_{\infty,\infty}^s(\Omega) = \mathcal{H}^s(\Omega)$ for any $s \in \mathbb{R}_+ \setminus \mathbb{Z}_+$, [Lemma N.1](#) directly implies [Lemma N.2](#). This completes the proof. $\square$

Next, we present the proof of [Theorem I.5](#).

Theorem N.1 ([Theorem I.5](#) Restated: Minimax Optimality of Flow Matching Transformers). Under the setting of $(16d + 18)(\beta + 1) = d_x + 2\beta$, the distribution estimation rate of flow matching transformers ([Theorem I.4](#)) matches the minimax lower bound of Hölder distribution class in 2-Wasserstein distance up to a log n and Lipschitz constants factors.

Proof of Theorem I.5. By [Theorem I.4](#), we have the distribution estimation rate in 2-Wasserstein distance under [Assumption I.2](#) and [Assumption I.3](#):

$$\mathbb{E}_{\{x_i\}_{i=1}^n} [W_2(\hat{P}_T, P_T)] = O(n^{-\frac{1}{16d+18}} (\log n)^{10d_x+2\beta+10}).$$

Then, by [Lemma N.2](#), the distribution rates matches the minimax lower bound up to a log n and Lipschitz constant factors under the setting

$$(16d + 18)(\beta + 1) = d_x + 2\beta.$$

This completes the proof. $\square$

O Experimental Validation

To provide empirical support for the proposed High-Order Flow Matching (HOFM) framework, we conduct a series of synthetic experiments designed to evaluate the practical benefits of incorporating higher-order dynamics. We compare the performance of standard first-order flow matching (equivalent to our framework with $K = 1$) against second-order flow matching ($K = 2$).

O.1 Experimental Setup

Task and Datasets. We evaluate the models on 2D density matching tasks, transitioning a standard multivariate Gaussian distribution, π_0 , to three complex target distributions, π_1 . Following the experimental setting in [\[Chen et al., 2025\]](#), we use target distributions shaped as: (1) a square, (2) two intertwined spirals, and (3) three intertwined spirals. These datasets are chosen to test the models' ability to learn distributions with sharp corners and high-curvature manifolds.

Evaluation Metric. To quantify the quality of the generated samples, we measure the 2-Wasserstein distance between the generated distribution and the target distribution. A lower Wasserstein distance indicates a better match and, therefore, superior performance.

O.2 Results and Discussion

The results of our comparison are summarized in [Appendix O.2](#). The findings demonstrate the advantages of using second-order dynamics.

Across all three target distributions and for every sampling step count (10, 50, and 100), the **second-order model achieves a lower Wasserstein distance** than the first-order model. This suggests that incorporating higher-order information allows the model to learn more accurate and stable generation paths, which aligns with the motivations discussed in [Section 1](#).

Distribution	Sampling Steps	First Order ($K = 1$)	Second Order ($K = 2$)
Square	10	8.51	7.09
	50	6.45	6.08
	100	5.48	2.82
Two Spirals	10	114.39	74.57
	50	73.37	68.47
	100	66.15	46.71
Three Spirals	10	192.19	109.93
	50	123.53	87.70
	100	93.26	68.81

Table 1: Comparison of first-order and second-order flow matching on synthetic 2D datasets.

Furthermore, these results highlight a notable improvement in **sampling efficiency**. For instance, in the Three Spirals task, the second-order model with only 50 sampling steps (Wasserstein distance of 87.70) outperforms the first-order model with 100 steps (93.26). This empirical evidence supports the theoretical premise that HOFM lead to more efficient sampling strategies (Section 5).

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Section 3 presents the K -order flow matching framework.

Section 4 presents the statistical rates of K -order flow matching using transformers. Appendix I provides the statistical rates of the first order flow matching using transformers.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitation and potential future works in Section 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Assumptions are stated before every results, and we either give the complete proof directly or refer it to the corresponding section in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not includes experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).

- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This paper does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: All aspects of this work comply with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This work is dedicated to advancing the fundamental understanding of flow-matching generative models; we do not anticipate any significant positive or negative societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: This paper does not release assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper involves neither crowdsourcing nor research with human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper involves neither crowdsourcing nor research with human participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The primary methodological contributions of this study do not depend on large language models in any meaningful, novel, or non-standard capacity.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.