We would like to thank all the reviewers for their thoughtful comments. We will respond to each reviewer's questions individually and incorporate the advice on formatting, notations, and references in an updated version of our manuscript.

[R1] Explanation of the transformation. Itô's Lemma shows our model can be used to construct a broad range of Itô diffusion processes with tractable finite-dimensional distributions (FDD). To show the correctness of Eqs. (5-7), it suffices to show the FDD on $\mathbb{R}^n$ of the stochastic process $\boldsymbol{X}_\tau$ defined by Eq.(6) is the same as the distribution obtained by transforming the FDD of the Wiener process with the density of Eq.(7). We present a formal argument based on measure theory and a proof sketch: consider the classical Wiener space (Ω, Σ), where $\Omega = C([0, +\infty), \mathbb{R})$, the set of continuous functions from $[0, +\infty)$ to $\mathbb{R}$, and Σ is the σ-algebra generated by all the cylinder sets of $C([0, +\infty), \mathbb{R})$.[1] We can equip this space with a probability measure (distribution) Q to get a probability space (Ω, Σ, Q) for continuous-time stochastic processes. Given a finite subset $\{\tau_1, \tau_2, ..., \tau_n\} \subset (0, +\infty)$, define the **projection** $\pi_{\{\tau_1,...,\tau_n\}} : (\Omega, \Sigma, Q) \longrightarrow (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ to be $\pi_{\{\tau_1,...,\tau_n\}}(\omega) = (\omega(\tau_1), ..., \omega(\tau_n))$, where $\mathcal{B}(\cdot)$ is the Borel σ-algebra. We will drop the index and simply use π to denote projection from now on. Projection is a measurable mapping. The **finite-dimensional distribution (FDD)** of a process is defined to be the pushforward measure induced by π, that is $Q \circ \pi^{-1}(A) = Q(\{\omega : (\omega(\tau_1), ..., \omega(\tau_2)) \in A\}), A \in \mathcal{B}(\mathbb{R}^n)$. Let P denote the unique measure of Wiener process $\boldsymbol{W}_\tau$ defined on the classical Wiener space. The following proposition and theorem will serve our purpose:

Proposition 1 *The mapping from (Ω, Σ, P) to (Ω, Σ) defined by $\boldsymbol{X}_\tau = F_\theta(\boldsymbol{W}_\tau, \tau)$ is measurable and therefore induces a pushforward measure $P \circ F_\theta^{-1}$.*

Theorem 1 *Given a finite subset $\{\tau_1, \tau_2, ..., \tau_n\} \subset (0, +\infty)$, the FDD of $\boldsymbol{X}_\tau$ is the same as the distribution of $(F_\theta(\boldsymbol{W}_{\tau_1}, \tau_1), ..., F_\theta(\boldsymbol{W}_{\tau_n}, \tau_n))$, where $(\boldsymbol{W}_{\tau_1}, ..., \boldsymbol{W}_{\tau_n})$ is a n-dimensional random variable with FDD of $\boldsymbol{W}_\tau$.*

The distributions, or (pushforward) measures, of $(\boldsymbol{X}_{\tau_1}, ..., \boldsymbol{X}_{\tau_n})$ and $(F_\theta(\boldsymbol{W}_{\tau_1}, \tau_1), ..., F_\theta(\boldsymbol{W}_{\tau_n}, \tau_n))$ are induced by two mappings from (Ω, Σ, P) to $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ respectively: a) $\pi \circ F_\theta$ and b) $F_\theta \circ \pi$, where π is the projection onto $\{\tau_1, \tau_2, ..., \tau_n\}$. To show the two distributions on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ are equal, it suffices to check that they assign the same measure to every Borel set of $\mathbb{R}^n$. This is true because the preimages of every Borel set under the two mappings are identical. The arguments above can be generalized to $\Omega = C([0, +\infty), \mathbb{R}^n)$.

[R1] Efficiency of neural ODE. We would like to clarify that our main contribution is a continuous-time stochastic process model, of which the normalizing flow (NF) is just one component. We use a continuous normalizing flow (neural ODE) primarily because of its free-form Jacobian matrix property, flexibility w.r.t. transformations, and model architecture. It also shows competitive performance on low-dimensional data compared with GLOW and autoregressive flows. Since our experiments focus on low-dimensional data, the time cost is not a major bottleneck. Other types of data may require a different choice of flow, which is possible and within the specifications of the proposed framework.

[R1] Ablation study of sampling intervals. It is worth noting that the observation time intervals are different samples from the same Poisson process for each sequence. We would like to make a minor correction on data-preprocessing details: for real-world data we use $\lambda = 2$ rather than $\lambda = 0.5$. We present the requested ablation study results below:

Table 1: Ablation Study on Time Interval for Real-World Data

Model	Mujoco-Hopper		BAQD		PTBDB	
	$\lambda_{\text{test}} = 1$	$\lambda_{\text{test}} = 5$	$\lambda_{\text{test}} = 1$	$\lambda_{\text{test}} = 5$	$\lambda_{\text{test}} = 1$	$\lambda_{\text{test}} = 5$
Latent ODE	25.082 ± 0.011	24.599 ± 0.004	2.948 ± 0.006	2.686 ± 0.006	-0.633 ± 0.006	-0.892 ± 0.009
VRNN	10.553 ± 0.010	8.543 ± 0.008	0.044 ± 0.007	-1.016 ± 0.001	$\mathbf{-1.552 \pm 0.011}$	$\mathbf{-2.545 \pm 0.005}$
CTFP	-5.860 ± 0.013	-20.530 ± 0.003	-0.890 ± 0.001	-3.595 ± 0.001	-0.982 ± 0.041	-1.793 ± 0.015
Latent CTFP	$\mathbf{-28.272 \pm 0.043}$	$\mathbf{-32.388 \pm 0.057}$	$\mathbf{-7.212 \pm 0.064}$	$\mathbf{-6.157 \pm 0.035}$	$\mathbf{-1.549 \pm 0.009}$	$\mathbf{-2.525 \pm 0.007}$

[R2] Independent normalizing flow transformation. While there is similarity between the graphical model representations of our approach and state-space models, we would like to stress the fundamental differences between them: CTFP directly models a distribution of continuous functions from the time axis to an observation space, or equivalently a stochastic process. It takes the evaluations of the functions at an arbitrary given time-grid to be the distribution of observations and can directly compute its density. The CTFP mapping is injective and does not rely on an emission process with observation noise. We condition each transformation only on the time stamp τ rather than previous observations to enforce marginalization consistency of the stochastic process: given the finite-dimensional distribution (FDD) of $(\boldsymbol{X}_{\tau_1}, ..., \boldsymbol{X}_{\tau_i}, ..., \boldsymbol{X}_{\tau_n})$ for a stochastic process $\boldsymbol{X}_\tau$, the distribution obtained by marginalizing over one of the dimensions, $\boldsymbol{X}_{\tau_i}$, must be the same as the FDD of $(\boldsymbol{X}_{\tau_1}, ..., \boldsymbol{X}_{\tau_{i-1}}, \boldsymbol{X}_{\tau_{i+1}}, ..., \boldsymbol{X}_{\tau_n})$. Our experiments show that our models outperform VRNN, which conditions emission and transition on all previous observations and is arguably a more powerful filtering-based model than the extended Kalman filter. We agree with the reviewer's comment on Eq.(12): the latent variable z could be interpreted as containing history information to relax the Markov property of CTFP.

[R3] Importance of continuity. We thank the reviewer for recognizing continuity as a unique property of CTFP. Its importance is shown from two aspects in our work: our models show better performance than other models when evaluated using denser observation intervals (larger λ) on most of the datasets. This is partially due to our model's continuity as the dependence between neighboring observations is stronger with denser observations. Moreover, the qualitative examples in Fig. 3(c) show that our models, which can generate continuous trajectories, are better for interpolation than non-continuous models.

[1] We refer the reviewer to Chapter 2 of *Brownian Motion, Martingales, and Stochastic Calculus* by Jean François Le Gall for more details.