

Speed-Accuracy Relations for Diffusion Models: Wisdom from Nonequilibrium Thermodynamics and Optimal Transport

Kotaro Ikeda^{1,*}, Tomoya Uda², Daisuke Okanohara³, and Sosuke Ito^{4,5,†}

¹*Department of Mathematical Engineering and Information Physics, School of Engineering,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

²*Department of Earth and Planetary Physics, School of Science,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

³*Preferred Networks Inc., 1-6-1, Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan*

⁴*Universal Biology Institute, Graduate School of Science,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*

⁵*Department of Physics, Graduate School of Science,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*



(Received 21 July 2024; accepted 4 June 2025; published 30 July 2025)

We discuss a connection between a generative model, called the diffusion model, and nonequilibrium thermodynamics for the Fokker-Planck equation, called stochastic thermodynamics. Using techniques from stochastic thermodynamics, we derive the speed-accuracy relations for diffusion models, which are inequalities that relate the accuracy of data generation to the entropy production rate. These relations can be interpreted as the relations between accuracy and the speed of the diffusion dynamics in the absence of the nonconservative force. From a stochastic thermodynamic perspective, our results provide quantitative insight into how best to generate data in diffusion models. The optimal learning protocol is introduced by the geodesic of space of the 2-Wasserstein distance in optimal transport theory. We numerically illustrate the validity of the speed-accuracy relations for diffusion models with different noise schedules and different data. We numerically discuss our results for optimal and suboptimal learning protocols. We also demonstrate the applicability of our results to data generation from the real-world image datasets.

DOI: [10.1103/x5vj-8jq9](https://doi.org/10.1103/x5vj-8jq9)

Subject Areas: Statistical Physics

I. INTRODUCTION

Diffusion processes are irreversible phenomena that cause thermodynamic dissipation. Diffusion processes are described by stochastic processes such as Brownian motion [1], and thermodynamic irreversibility is quantified by the entropy production in stochastic thermodynamics [2,3]. In stochastic thermodynamics, there have been various discussions about the relationship between information processing and thermodynamic dissipation for diffusion processes [4–24]. Based on optimal transport theory [25], inevitable thermodynamic dissipation for diffusion processes has been discussed in stochastic thermodynamics [26–29], and the thermodynamic trade-off relations among speed, accuracy, and dissipation for diffusion processes have been discussed as a generalization of the second law of thermodynamics [29–33].

Diffusion processes have recently been discussed in the context of statistical machine-learning models called generative models [34]. The diffusion-based generative models called diffusion models [35,36] were originally inspired by nonequilibrium thermodynamics [35]. By introducing time-reversed dynamics, which are well studied in the context of the fluctuation theorem [5,37,38] and Jarzynski's equality [39], diffusion models generate data with spatial structure from noisy data without spatial structure. Several improvements have been made from different perspectives [40–53], and the improved diffusion models have reached the state of the art in an image-generation task [40,42].

The proposed diffusion models [42,44,51,54,55] are currently understandable without nonequilibrium thermodynamics, and the improved diffusion models are regarded as variants of other statistical machine-learning models. For example, diffusion models [42] are related to the score estimation methods including the score matching [56–58]. By incorporating the existing flow-based methods [59–62], another improved method called the flow matching [63] has also been proposed in the context of diffusion models.

Because diffusion models have been improved outside the context of nonequilibrium thermodynamics, the insights of stochastic thermodynamics are not fully exploited in

*Contact author: kotaro-ikeda@g.ecc.u-tokyo.ac.jp

†Contact author: sosuke.ito@ubi.s.u-tokyo.ac.jp

diffusion models. For example, the discussion of minimum entropy production in stochastic thermodynamics is related to the Wasserstein distance [25] in optimal transport theory [29,30], which is well used in generative models such as the Wasserstein generative adversarial network [64] and diffusion models [65–67]. Discussions on the relationship between optimal transport and dynamics in diffusion models have only recently begun [68–73]. In these discussions, various insights from stochastic thermodynamics, such as the thermodynamic trade-off relations for diffusion processes based on optimal transport theory [29–32], are not well used. There is only one paper that discusses the flow matching method from the perspective of both optimal transport theory and stochastic thermodynamics [74].

In this paper, we reconsider diffusion models in terms of stochastic thermodynamics and derive speed-accuracy relations for diffusion models, which are analogous to the thermodynamic trade-off relations based on optimal transport theory. The speed-accuracy relations explain that the theoretical limits of accurate data generation are generally bounded by the entropy production rate and the temperature, which can be interpreted as the speed of the diffusion dynamics in the absence of the nonconservative force. Furthermore, the speed-accuracy relations explain that the optimal diffusion dynamics for learning is given by the optimal transport. The results provide theoretical support for the trial-and-error protocols of diffusion dynamics, which have been characterized by noise schedules such as the cosine schedule [43] and the conditional optimal transport schedule [63]. We numerically illustrate the speed-accuracy relations for diffusion models using simple

one-dimensional diffusion processes, two-dimensional diffusion processes, and diffusion processes on latent space [44]. We show that accurate data generation can be discussed from a quantitative comparison of the speed-accuracy relations for different diffusion processes. We numerically illustrate that the optimal transport provides the most accurate data generation. We also investigate the situation where the inaccurate data generation is achieved due to a nonconservative force and a nonoptimal protocol as predicted by the speed-accuracy relations. Furthermore, we demonstrate the applicability of the speed-accuracy relations to practical diffusion models trained on real-world image datasets.

This paper is organized as follows (see also Fig. 1). In Sec. II, we discuss the generative models and diffusion models. We first explain the generative models (Sec. II A) and then move on to the basic concepts of diffusion models (Sec. II B). We then present the mathematical details of the diffusion model (Secs. II C and II D). We also explain a practical formulation based on the conditional Gaussian probabilities (Sec. II E) as well as some examples (Sec. II F). In Sec. III, we explain stochastic thermodynamics, especially from the perspective of its relationship to diffusion models and optimal transport theory. In Sec. III A, we introduce the entropy production rate in stochastic thermodynamics. In Sec. III B, we explain applications of optimal transport theory in stochastic thermodynamics such as the thermodynamic trade-off relations. In Sec. IV, we derive the main results, which are the speed-accuracy relations for diffusion models. The main results imply that the entropy production rate introduced in Sec. III A provides a fundamental limit on the accuracy of data

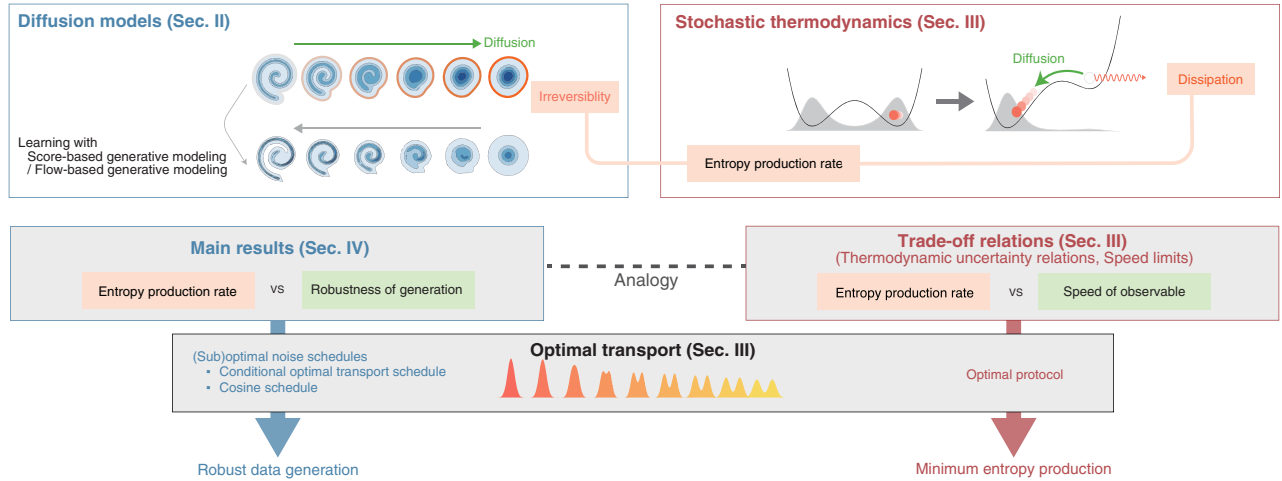


FIG. 1. The organization of this paper. In Sec. II, we explain the basic concepts of diffusion models and discuss their methods, including score-based generative modeling and flow-based generative modeling. In Sec. III, we discuss stochastic thermodynamics and optimal transport theory and introduce the entropy production rate. We also introduce several thermodynamic trade-off relations, such as thermodynamic uncertainty relations and speed limits, which provide the minimum entropy production in a finite time via optimal transport. In Sec. IV, we derive our main results in analogy to thermodynamic trade-off relations. Our main results reveal the essential role of the entropy production rate in diffusion models. Noise schedules based on optimal transport will lead to the most robust generation in diffusion models, and existing noise schedules such as the conditional optimal transport schedule and the cosine schedule are regarded as suboptimal noise schedules.

generation in the diffusion model, and the main results are analogous to the thermodynamic trade-off relations in Sec. III B. Furthermore, we discuss the optimal protocol based on the main results in Sec. IV C and confirm the main results by numerical calculations of diffusion models with three different situations in Sec. IV D. Finally, we conclude with the main results and discuss different perspectives on the results, such as a generalization of the main results, and future studies in Sec. V.

II. GENERATIVE MODELS AND DIFFUSION MODELS

A. Generative models

In this paper, we discuss a class of statistical machine-learning models known as generative models [34]. A generative model is a class of statistical models capable of generating a new dataset that resembles the input dataset. The data generation in generative models can be considered as an estimation of the input data distribution, while the generated dataset is a sample from the estimated distribution.

More specifically, generative models can be described as follows. Let $d_{\text{in}} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N_D)}\}$ be the set of input dataset, where each element is a state of the n_d -dimensional Euclidean space $\mathbf{x}^{(n)} \in \mathbb{R}^{n_d}$. The data $\mathbf{x}^{(n)}$ are assumed to be independent and identically distributed and are sampled from a probability density function $q(\mathbf{x})$ that satisfies $q(\mathbf{x}) > 0$ and $\int d\mathbf{x} q(\mathbf{x}) = 1$. Here, the positivity of the probability density function $q(\mathbf{x}) > 0$ is required for quantities such as $\ln q(\mathbf{x})$ or $1/q(\mathbf{x})$ not to diverge, and we implicitly assume that any probability density functions are positive throughout this paper. We consider the problem of constructing a new probability density function $p(\mathbf{x})$ that satisfies $p(\mathbf{x}) > 0$ and $\int d\mathbf{x} p(\mathbf{x}) = 1$, based on the estimation of the probability density function $q(\mathbf{x})$. Statistical machine-learning models that generate samples from the probability density function $p(\mathbf{x})$ are called generative models.

Several studies have shown that generative models can be applied to practical datasets such as images [75], natural language [76], and audio [77]. In these applications, generative models are constructed by solving optimization problems using machine-learning methods such as deep learning [34,78]. Machine-learning optimization problems are solved by numerically minimizing or maximizing some objective function [34,78]. To judge whether learning was successful, we sometimes discuss whether a distance or a pseudodistance between the distribution of the input data $q(\mathbf{x})$ and the distribution of the generated data $p(\mathbf{x})$ becomes smaller [79–82]. We call the value of this (pseudo)distance the estimation error. Typical examples of the estimation error are the Kullback-Leibler divergence [83], which is a pseudodistance in information theory and information geometry, and the Wasserstein distance [25], which is a distance in optimal transport theory.

B. Diffusion models

Among these machine-learning-based generative models, a series of models called diffusion models [35,36,40,42,48] achieves lower estimation errors than previous generative models for various data [36,47,73,84–90], including image data [36,47,73], video [87,88], and audio [89,90]. Diffusion models generate samples from the probability density function $p(\mathbf{x})$ by estimating the time-reversed diffusion process that generates the data structure, while the corresponding forward process destroys the data structure by adding stochastic noise to the input dataset d_{in} (see Fig. 2) [35,36]. In practice, we sometimes convert the raw data into a low-dimensional latent space and discuss a diffusion process in the latent space to extract the features of the data and reduce the computational cost [44]. In this case, the input dataset d_{in} can be regarded as a dataset of latent variables, which is created by converting the raw dataset into the latent space.

Specifically, the input probability density function $\mathcal{P}_0(\mathbf{x}) = q(\mathbf{x})$ is transformed into a known noise distribution $\mathcal{P}_\tau(\mathbf{x})$ by adding noise to the input data, where $\mathcal{P}_t(\mathbf{x})$ is the probability density function at time $t \in [0, \tau]$. The time evolution of $\mathcal{P}_t(\mathbf{x})$ is given by diffusion processes described by the Fokker-Planck equation [92] or the master equation [93] in diffusion models [35,36]. Sometimes, we consider the stochastic dynamics of the data [36] from the initial state $\mathbf{x}_{t=0}$ to the final state $\mathbf{x}_{t=\tau}$, where \mathbf{x}_t is the state of the data at time t instead of the time evolution of the probability density function $\mathcal{P}_t(\mathbf{x})$. In this case, the stochastic time evolution of \mathbf{x}_t can be described by the Langevin equation [1] or the Markov chain Monte Carlo method [94–96]. The diffusion process from $t = 0$ to $t = \tau$ is called the forward process [35], and the dependence of the noise intensity at time t in the forward process is called

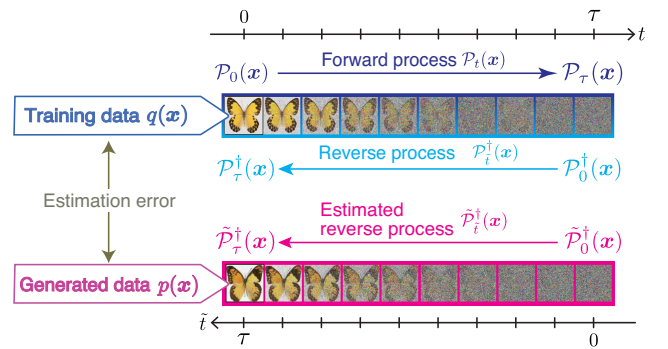


FIG. 2. Illustration of diffusion models in the case of image data. The forward process, denoted by $\mathcal{P}_t(\mathbf{x})$, is a stochastic process which adds noise to the training data sampled from $q(\mathbf{x})$. By inverting this forward process in time, we prepare the reverse process $\mathcal{P}_t^\dagger(\mathbf{x})$. The generated data sampled from $p(\mathbf{x})$ are constructed by an estimated reverse process $\hat{\mathcal{P}}_t^\dagger(\mathbf{x})$ that mimics the reverse process. The image in the figure is taken from the dataset [91].

the noise schedule, which characterizes the time evolution of the diffusion dynamics [35,43,97].

Next, we explain the reverse process, which is the time-reversed dynamics of the forward process. Here, we consider the forward process described by the time evolution of the probability density function $\mathcal{P}_t(\mathbf{x})$. We write the reverse process as $\mathcal{P}_t^\dagger(\mathbf{x})$, defined as $\mathcal{P}_t^\dagger(\mathbf{x}) := \mathcal{P}_{\tau-t}(\mathbf{x})$. In the reverse process, we consider the time evolution from $\tilde{t} = 0$ to $\tilde{t} = \tau$. If τ is large enough, the final probability of the forward process $\mathcal{P}_\tau(\mathbf{x})$ can be considered as a noisy state where the data structure is well broken. Thus, the reverse process means that this noisy state $\mathcal{P}_0^\dagger(\mathbf{x}) = \mathcal{P}_\tau(\mathbf{x})$ is the initial probability density function at time $\tilde{t} = 0$ in the reverse process, and the time evolution in the reverse process allows the reproduction of the input data distribution $\mathcal{P}_\tau^\dagger(\mathbf{x}) = \mathcal{P}_0(\mathbf{x}) = q(\mathbf{x})$ containing the data structure at the final time $\tilde{t} = \tau$.

We introduce the stochastic process for generating the data from $p(\mathbf{x})$, namely, the estimated reverse process, as an imitation of the reverse process. This estimated reverse process may differ from the reverse process, in general, and so we write it as $\tilde{\mathcal{P}}_t^\dagger(\mathbf{x})$ and distinguish it from the reverse process $\mathcal{P}_t^\dagger(\mathbf{x}) [\neq \tilde{\mathcal{P}}_t^\dagger(\mathbf{x})]$. Since the initial probability density function of the estimated reverse process $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ can be different from the initial probability density function of the reverse process $\mathcal{P}_0^\dagger(\mathbf{x}) [\neq \tilde{\mathcal{P}}_0^\dagger(\mathbf{x})]$, the final probability density function of the estimated reverse process $\tilde{\mathcal{P}}_\tau^\dagger(\mathbf{x})$ can be different from the input distribution $q(\mathbf{x}) [\neq \tilde{\mathcal{P}}_\tau^\dagger(\mathbf{x})]$. Here, the idea of diffusion models is to set this probability density function $\tilde{\mathcal{P}}_\tau^\dagger(\mathbf{x})$, which is the output of the estimated reverse process, as the generated data distribution $p(\mathbf{x})$ [35]. In practical diffusion models, we do not necessarily compute the time evolution of the probability density function. For example, there are several methods to generate the data from the estimated reverse process using the Langevin equation [36,42] or a deterministic time evolution via flow [63].

C. Fokker-Planck equation for the forward process

We discuss the time evolution of $\mathcal{P}_t(\mathbf{x})$ for the forward process in diffusion models [36] based on the Fokker-Planck equation [92]. The time evolution of the probability density function $\mathcal{P}_t(\mathbf{x})$ is described by the Fokker-Planck equation as follows:

$$\begin{aligned} \partial_t \mathcal{P}_t(\mathbf{x}) &= -\nabla \cdot (\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x})), \\ \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) &:= \mathbf{f}_t(\mathbf{x}) - T_t \nabla \ln \mathcal{P}_t(\mathbf{x}), \end{aligned} \quad (1)$$

where $\partial_t := \partial/\partial t$ is the partial differential operator and ∇ is the del operator. $\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \in \mathbb{R}^{n_d}$ is the velocity field in a continuity equation. The parameters $T_t \in \mathbb{R}_{\geq 0}$ and $\mathbf{f}_t(\mathbf{x}) \in \mathbb{R}^{n_d}$ are functions of the time t , in general. The

forward process can be regarded as the time evolution of $\mathcal{P}_t(\mathbf{x})$ by the Fokker-Planck equation [Eq. (1)] under the initial condition $\mathcal{P}_0(\mathbf{x}) = q(\mathbf{x})$ with fixed time dependence of the parameters T_t and $\mathbf{f}_t(\mathbf{x})$. The time dependence of T_t characterizes the time variation of the noise intensity in the diffusion process, which is known as the noise schedule in diffusion models.

The Fokker-Planck equation [Eq. (1)] is statistically equivalent to the stochastic differential equation, namely, the Langevin equation [92,93] for the state \mathbf{x}_t :

$$d\mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_t)dt + \sqrt{2T_t}dB_t, \quad (2)$$

where $dB_t = B_{t+dt} - B_t$ is the random white Gaussian noise, mathematically defined as the difference of the Wiener processes B_t . If the Langevin equation is considered as Brownian dynamics for the position of a Brownian particle, the parameters T_t and $\mathbf{f}_t(\mathbf{x}_t)$ correspond to the temperature and the external force, respectively [92]. In the context of diffusion models, the forward process described by the Langevin equation can be viewed as a time-evolving process in which the random noise is sequentially added to the data. Using the Langevin equation, the time evolution of the forward process is computed using the input data $\mathbf{x}^{(n)} \in d_{\text{in}}$ ($n = 1, \dots, N_D$) for fixed time dependence of the parameters T_t and $\mathbf{f}_t(\mathbf{x}_t)$. The time evolution of the probability density function $\mathcal{P}_t(\mathbf{x})$ can be computed numerically by Monte Carlo sampling using Langevin dynamics [96].

D. Reverse and estimated reverse process

Next, we discuss the reverse process by considering the time-reversed dynamics of the Fokker-Planck equation for the forward process [Eq. (1)]. Using the reversed time $\tilde{t} = \tau - t$, the time evolution of $\mathcal{P}_t^\dagger(\mathbf{x}) = \mathcal{P}_{\tau-t}(\mathbf{x}) = \mathcal{P}_t(\mathbf{x})$ is given by the Fokker-Planck equation for the forward process [Eq. (1)] as follows:

$$\begin{aligned} \partial_t \mathcal{P}_t^\dagger(\mathbf{x}) &= -\partial_t \mathcal{P}_t(\mathbf{x}) \\ &= \nabla \cdot (\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x})) \\ &= -\nabla \cdot (\boldsymbol{\nu}_t^\dagger(\mathbf{x}) \mathcal{P}_t^\dagger(\mathbf{x})), \end{aligned} \quad (3)$$

where the velocity field for the reverse process $\boldsymbol{\nu}_t^\dagger(\mathbf{x})$ is defined as $\boldsymbol{\nu}_t^\dagger(\mathbf{x}) := -\boldsymbol{\nu}_{\tau-t}^{\mathcal{P}}(\mathbf{x})$.

Diffusion models are introduced by the estimated reverse process that mimics the reverse process [Eq. (3)]. We denote the probability density function of the data \mathbf{x} in the estimated reverse process by $\tilde{\mathcal{P}}_t^\dagger(\mathbf{x})$. We assume that the time evolution of this probability density function $\tilde{\mathcal{P}}_t^\dagger(\mathbf{x})$ can be written by the continuity equation:

$$\partial_t \tilde{\mathcal{P}}_i^\dagger(\mathbf{x}) = -\nabla \cdot (\tilde{\mathbf{v}}_i^\dagger(\mathbf{x}) \tilde{\mathcal{P}}_i^\dagger(\mathbf{x})). \quad (4)$$

Here, $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x})$ represents the velocity field in the estimated reverse process, which may differ from that of the reverse process $\mathbf{v}_i^\dagger(\mathbf{x})$. The velocity field $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x})$ is estimated numerically from the dynamics of the forward process.

In the following section, we explain two methods for estimating the velocity field in the reverse process and constructing the velocity field of the estimated reverse process $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x})$ for data generation, called the score-based generative modeling [42] and the flow-based generative modeling [63].

1. Example 1: Score-based generative modeling

First, we explain a method called score-based generative modeling [42]. This method consists of two components: score matching [42,56–58] and data generation. Score matching is a method that estimates the score function $\nabla \ln \mathcal{P}_t(\mathbf{x})$ in the forward process [Eq. (1)] in order to arrange the estimated reverse process [36,42]. The score function is typically estimated using neural networks. Specifically, we consider a situation where the score function $\nabla \ln \mathcal{P}_t(\mathbf{x})$ is modeled by a neural network denoted by $s_t^\theta(\mathbf{x})$, where θ represents the set of the network's parameters. We then numerically solve an optimization problem to estimate the score function. The optimized parameters, denoted by θ_{SM}^* , are employed to generate the velocity field of the estimated reverse process $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x})$.

The optimization problem is constructed as follows. The optimization problem to estimate the score function $\nabla \ln \mathcal{P}_t(\mathbf{x})$ is formulated as a minimization problem of the following score matching objective function:

$$L_{\text{SM}}(\theta) = \mathbb{E}_{\mathcal{P}_t, \mathcal{U}}[\|s_t^\theta(\mathbf{x}) - \nabla \ln \mathcal{P}_t(\mathbf{x})\|^2]. \quad (5)$$

Here, the uniform distribution for $t \in [0, \tau]$ is given by $\mathcal{U}(t) = 1/\tau$. The expected value with respect to this uniform distribution and the distribution $\mathcal{P}_t(\mathbf{x})$ is defined as $\mathbb{E}_{\mathcal{P}_t, \mathcal{U}}[\cdot] = \int_0^\tau dt \mathcal{U}(t) \int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \dots$. Then, the optimal set of the parameters θ_{SM}^* is obtained as [36,57]

$$\theta_{\text{SM}}^* \in \operatorname{argmin}_\theta L_{\text{SM}}(\theta). \quad (6)$$

In this optimization problem, $s_t^{\theta_{\text{SM}}^*}(\mathbf{x})$, which reaches the minimum value $L_{\text{SM}}(\theta_{\text{SM}}^*) = 0$, can be the score function $\nabla \ln \mathcal{P}_t(\mathbf{x})$.

2. Example 1-1: Stochastic differential equation

For the data generation by the estimated reverse process, we first explain the method based on the stochastic differential equations [36,40,42]. The estimated score

function $s_t^{\theta_{\text{SM}}^*}(\mathbf{x})$ is employed to reconstruct the velocity field of the forward process $\mathbf{v}_t^\mathcal{P}(\mathbf{x})$ as

$$\hat{\mathbf{v}}_t(\mathbf{x}) := \mathbf{f}_t(\mathbf{x}) - T_t s_t^{\theta_{\text{SM}}^*}(\mathbf{x}). \quad (7)$$

With this estimated velocity field $\hat{\mathbf{v}}_t(\mathbf{x})$, we construct the velocity field of the estimated reverse process $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x})$ as

$$\begin{aligned} \tilde{\mathbf{v}}_i^\dagger(\mathbf{x}) &= [\mathbf{v}_i^\dagger(\mathbf{x})]^\dagger - 2\hat{\mathbf{v}}_{\tau-i}(\mathbf{x}) \\ &= -\mathbf{f}_i^\dagger(\mathbf{x}) + 2T_i^\dagger s_i^{\theta_{\text{SM}}^*}(\mathbf{x}) - T_i^\dagger \nabla \ln \tilde{\mathcal{P}}_i^\dagger(\mathbf{x}), \end{aligned} \quad (8)$$

so that the estimated reverse process [Eq. (4)] mimics the reverse process [Eq. (3)], where $[\mathbf{v}_i^\dagger(\mathbf{x})]^\dagger := \mathbf{f}_i^\dagger(\mathbf{x}) - T_i^\dagger \nabla \ln \tilde{\mathcal{P}}_i^\dagger(\mathbf{x})$, $\mathbf{f}_i^\dagger(\mathbf{x}) := \mathbf{f}_{\tau-i}(\mathbf{x})$, $T_i^\dagger := T_{\tau-i}$, and $s_i^{\theta_{\text{SM}}^*}(\mathbf{x}) := s_{\tau-i}^{\theta_{\text{SM}}^*}(\mathbf{x})$. If $s_i^{\theta_{\text{SM}}^*}(\mathbf{x})$ is exactly equal to $\nabla \ln \mathcal{P}_i(\mathbf{x})$ and the initial condition for the estimated reverse process is equivalent to the initial condition for the reverse process $[\tilde{\mathcal{P}}_0^\dagger(\mathbf{x}) = \mathcal{P}_0^\dagger(\mathbf{x})]$, the equation $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x}) = \mathbf{v}_i^\dagger(\mathbf{x})$ holds and the reverse process [Eq. (3)] is consistent with the estimated reverse process [Eq. (4)].

The continuity equation of the estimated reverse process $\partial_t \tilde{\mathcal{P}}_i^\dagger(\mathbf{x}) = -\nabla \cdot (\tilde{\mathbf{v}}_i^\dagger(\mathbf{x}) \tilde{\mathcal{P}}_i^\dagger(\mathbf{x}))$ using the estimated velocity field $\tilde{\mathbf{v}}_i^\dagger(\mathbf{x})$ can be regarded as the Fokker-Planck equation with the external force $-\mathbf{f}_i^\dagger(\mathbf{x}) + 2T_i^\dagger s_i^{\theta_{\text{SM}}^*}(\mathbf{x})$. The stochastic differential equation (i.e., the Langevin equation) corresponding to this Fokker-Planck equation [36,98] is

$$d\tilde{\mathbf{x}}_i^\dagger = \left(-\mathbf{f}_i^\dagger(\tilde{\mathbf{x}}_i^\dagger) + 2T_i^\dagger s_i^{\theta_{\text{SM}}^*}(\tilde{\mathbf{x}}_i^\dagger) \right) d\tilde{t} + \sqrt{2T_i^\dagger} dB_{\tilde{t}}, \quad (9)$$

where $dB_{\tilde{t}} = B_{\tilde{t}+d\tilde{t}} - B_{\tilde{t}}$ is the difference of the Wiener process $B_{\tilde{t}}$ and $\tilde{\mathbf{x}}_i^\dagger$ is the state of the data in the estimated reverse process. By simulating this stochastic differential equation numerically, we can generate the data in score-based generative modeling [36,40,42].

3. Example 1-2: Probability flow ordinary differential equation

Next, we discuss an alternative method for data generation in score-based generative modeling, namely, the probability flow ordinary differential equation (ODE) [36]. The probability flow ODE is a method in which we use an ODE instead of the stochastic differential equation [Eq. (9)] for data generation under the same objective function [Eq. (5)]. At present, this probability flow ODE is well used instead of the stochastic differential equation, because it is possible to generate data faster and more accurately by performing time evolution based on numerical ODE solvers [99–101].

Specifically, using the velocity field $\hat{\mathbf{v}}_t(\mathbf{x})$ [Eq. (7)] estimated through the objective function [Eq. (5)], the

velocity field of the estimated reverse process $\tilde{\nu}_i^\dagger(\mathbf{x})$ is arranged as follows:

$$\tilde{\nu}_i^\dagger(\mathbf{x}) = -\hat{\nu}_{\tau-i}(\mathbf{x}). \quad (10)$$

This expression is parallel to the expression of the velocity field in the reverse process $\nu_i^\dagger(\mathbf{x}) = -\nu_{\tau-i}^\dagger(\mathbf{x})$. In this case, the continuity equation [Eq. (4)] corresponds to the following ODE [36,61]:

$$\frac{d\tilde{\mathbf{x}}_i^\dagger}{dt} = -\hat{\nu}_{\tau-i}(\tilde{\mathbf{x}}_i^\dagger) = -\mathbf{f}_i^\dagger(\tilde{\mathbf{x}}_i^\dagger) + T_i^\dagger \mathbf{s}_i^{\theta_{\text{SM}}^\dagger}(\tilde{\mathbf{x}}_i^\dagger), \quad (11)$$

where $\tilde{\mathbf{x}}_i^\dagger$ denotes the sample from $\tilde{\mathcal{P}}_i^\dagger(\mathbf{x})$. The probability flow ODE method generates data by numerically simulating this ODE [36].

4. Example 2: Flow-based generative modeling

Next, we explain a method called flow-based generative modeling [63], which consists of flow matching and data generation via the ODE. Flow matching [63] is a method to estimate the velocity field of the forward process without estimating the score function. In flow matching, we model the velocity field of the forward process $\nu_i^\mathcal{P}(\mathbf{x})$ by a neural network $\mathbf{u}_i^\theta \in \mathbb{R}^{n_d}$, where θ is the set of parameters. The objective function

$$L_{\text{FM}}(\theta) = \mathbb{E}_{\mathcal{P}_i, \mathcal{U}}[\|\mathbf{u}_i^\theta(\mathbf{x}) - \nu_i^\mathcal{P}(\mathbf{x})\|^2] \quad (12)$$

is minimized with respect to θ using numerical optimization methods, and θ_{FM}^* is the optimal set of the parameters satisfying

$$\theta_{\text{FM}}^* \in \operatorname{argmin}_\theta L_{\text{FM}}(\theta). \quad (13)$$

The optimal set of the parameters θ_{FM}^* gives the velocity field for the estimated reverse process $\tilde{\nu}_i^\dagger(\mathbf{x})$ as follows:

$$\tilde{\nu}_i^\dagger(\mathbf{x}) = -\mathbf{u}_{\tau-i}^{\theta_{\text{FM}}^*}(\mathbf{x}). \quad (14)$$

This expression corresponds to the velocity field of the reverse process $\nu_i^\dagger(\mathbf{x}) = -\nu_{\tau-i}^\dagger(\mathbf{x})$.

In the flow-based generative modeling, the data generation is done by simulating the ODE [63,102]:

$$\frac{d\tilde{\mathbf{x}}_i^\dagger}{dt} = -\mathbf{u}_{\tau-i}^{\theta_{\text{FM}}^*}(\tilde{\mathbf{x}}_i^\dagger), \quad (15)$$

which is similar to the one used in the probability flow ODE method [Eq. (11)]. This ODE corresponds to the continuity equation for the estimated reverse process [Eq. (4)].

E. Formulations with conditional Gaussian distributions

In a practical implementation of diffusion models, we may consider a process such that the external force $\mathbf{f}_i(\mathbf{x})$ is linear:

$$\mathbf{f}_i(\mathbf{x}) = \mathbf{A}_i \mathbf{x} + \mathbf{b}_i, \quad (16)$$

to reduce the computational complexity [35,36,40,42,63], where $\mathbf{A}_i \in \mathbb{R}^{n_d \times n_d}$ and $\mathbf{b}_i \in \mathbb{R}^{n_d}$ are the matrix and the vector, respectively. When the initial condition is $\mathcal{P}_{t=0}(\mathbf{x}) = q(\mathbf{x})$, the solution $\mathcal{P}_t(\mathbf{x})$ for the process can be given by

$$\mathcal{P}_t(\mathbf{x}) = \int d\mathbf{y} \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y}). \quad (17)$$

Here, the transition probability $\mathcal{P}_t^c(\mathbf{x}|\mathbf{y})$ is a Gaussian distribution $\mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_t(\mathbf{y}), \Sigma_t)$ due to the linearity of the external force [92], where $\boldsymbol{\mu}_t(\mathbf{y})$ is the mean, which depends on \mathbf{y} , and Σ_t is the covariance matrix Σ_t . We assume that Σ_t does not depend on the state \mathbf{y} . Equation (17) at $t = 0$ gives the condition $\mathcal{P}_0^c(\mathbf{x}|\mathbf{y}) = \delta(\mathbf{x} - \mathbf{y})$, where $\delta(\mathbf{x} - \mathbf{y})$ is the delta function. Thus, the covariance matrix at $t = 0$ is $\Sigma_{t=0} = \mathbf{O}$, and the mean is $\boldsymbol{\mu}_{t=0}(\mathbf{y}) = \mathbf{y}$, where \mathbf{O} is the zero matrix.

Under the above condition, the optimization problems [Eqs. (5) and (12)] are easier to solve [42,63]. These training methods using conditional probability in score matching and flow matching are known as denoising score matching [57] and conditional flow matching [63], respectively. To reduce the computational complexity, we consider new objective functions, which correspond to the objective functions [Eqs. (5) and (12)] as follows:

$$L_{\text{SM}}^c(\theta) = \mathbb{E}_{\mathcal{P}_i^c, q, \mathcal{U}}[\|\mathbf{s}_i^\theta(\mathbf{x}) - \nabla \ln \mathcal{P}_i^c(\mathbf{x}|\mathbf{y})\|^2], \quad (18)$$

$$L_{\text{FM}}^c(\theta) = \mathbb{E}_{\mathcal{P}_i^c, q, \mathcal{U}}[\|\mathbf{u}_i^\theta(\mathbf{x}) - \nu_i^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y})\|^2], \quad (19)$$

where $\mathbb{E}_{\mathcal{P}_i^c, q, \mathcal{U}}[\cdots] = \int_0^\tau dt \mathcal{U}(t) \int d\mathbf{x} d\mathbf{y} \mathcal{P}_i^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y}) \cdots$ is the expected value with respect to $\mathcal{P}_i^c(\mathbf{x}|\mathbf{y})$, $q(\mathbf{y})$, $\mathcal{U}(t)$, and $\nu_i^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y}) := \mathbf{f}_i(\mathbf{x}) - T_i \nabla \ln \mathcal{P}_i^c(\mathbf{x}|\mathbf{y})$. As proved in Appendix A, new objective functions satisfy $\nabla_\theta L_{\text{SM}}^c(\theta) = \nabla_\theta L_{\text{SM}}(\theta)$ and $\nabla_\theta L_{\text{FM}}^c(\theta) = \nabla_\theta L_{\text{FM}}(\theta)$, where ∇_θ is the gradient through θ . Since the optimal solutions are given by the conditions $\nabla_\theta L_{\text{SM}}^c(\theta) = \nabla_\theta L_{\text{SM}}(\theta) = 0$ and $\nabla_\theta L_{\text{FM}}^c(\theta) = \nabla_\theta L_{\text{FM}}(\theta) = 0$, we get the optimal solutions

$$\theta_{\text{SM}}^* \in \operatorname{argmin}_\theta L_{\text{SM}}^c(\theta), \quad (20)$$

$$\theta_{\text{FM}}^* \in \operatorname{argmin}_\theta L_{\text{FM}}^c(\theta), \quad (21)$$

which are equivalent to the solutions of the original problems [Eqs. (6) and (13)]. Here, the quantities $\nabla \ln \mathcal{P}_t^c(\mathbf{x}|\mathbf{y})$ and $\nu_t^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y})$ are linear functions of \mathbf{x} because of the Gaussian property, and the conditional probability $\mathcal{P}_t^c(\mathbf{x}|\mathbf{y})$ can be sampled more easily than $\mathcal{P}_t(\mathbf{x})$ due to the reproductive property of Gaussian distributions. Therefore, these optimization problems [Eqs. (18) and (19)] are numerically simpler than the original optimization problems [Eqs. (5) and (12)].

F. Examples of diffusion models

We present some representative examples of diffusion models [40,43,48,63]. Here, we discuss the conditional Gaussian formulations from the previous section. We consider the temperature T_t and the external force $\mathbf{f}_t(\mathbf{x})$ as functions of non-negative parameters $\sigma_t(\geq 0)$ and $m_t(\geq 0)$ such that

$$T_t = m_t \sigma_t \partial_t \left(\frac{\sigma_t}{m_t} \right) = \partial_t \left(\frac{\sigma_t^2}{2} \right) - \sigma_t^2 \partial_t \ln m_t, \quad (22)$$

$$\mathbf{f}_t(\mathbf{x}) = \nabla \left(\frac{\partial_t \ln m_t}{2} \|\mathbf{x}\|^2 \right) = (\partial_t \ln m_t) \mathbf{x}. \quad (23)$$

The external force $\mathbf{f}_t(\mathbf{x})$ is described by the linear function [Eq. (16)] with $\mathbf{A}_t = (\partial_t \ln m_t) \mathbf{I}$, $\mathbf{b}_t = \mathbf{0}$, where \mathbf{I} is the identity matrix and $\mathbf{0}$ is the zero vector. The diffusion dynamics can be determined by the time dependence of m_t and σ_t instead of the time dependence of T_t and $\mathbf{f}_t(\mathbf{x})$. The time dependence of m_t and σ_t is also called the noise schedule [40,43,48,63].

We sometimes make the following assumption for the noise schedules. To satisfy $T_t \geq 0$, the monotonicity condition

$$\partial_t \left(\frac{\sigma_t}{m_t} \right) \geq 0 \quad (24)$$

is assumed. The term m_t/σ_t is called the signal-to-noise ratio [41], which must be monotonically nonincreasing. Furthermore, we assume the initial conditions $m_{t=0} = 1$ and $\sigma_{t=0} = 0$. Under these initial conditions, the solutions of the mean $\mu_t(\mathbf{y})$ and the covariance matrix Σ_t in $\mathcal{P}_t^c(\mathbf{x}|\mathbf{y})$ are given by

$$\mu_t(\mathbf{y}) = m_t \mathbf{y}, \quad \Sigma_t = \sigma_t^2 \mathbf{I} \quad (25)$$

(see Appendix B). For high-dimensional datasets such as images, we often use this assumption [36,40,48,63]. In the following, we show some examples of the noise schedules.

1. Example 1: Variance exploding diffusion (VE diffusion)

We first introduce variance exploding diffusion (VE diffusion) [36]. VE diffusion is a method based on the

noise conditional score networks [42], which is described by the following Langevin equation [36]:

$$d\mathbf{x}_t = \sqrt{2T_t} d\mathbf{B}_t. \quad (26)$$

This equation corresponds to the condition that there is no external force $\mathbf{f}_t(\mathbf{x}) = \mathbf{0}$. From Eqs. (22) and (23), we obtain $\partial_t \ln m_t = 0$ and

$$T_t = \partial_t \left(\frac{\sigma_t^2}{2} \right). \quad (27)$$

The condition $\partial_t \ln m_t = 0$ means that the noise schedule is given by $m_t = 1$ ($0 \leq t \leq \tau$) because the initial condition is given by $m_0 = 1$. Because of Eq. (24), σ_t should be a monotonically nondecreasing function of time. While we explain the VE diffusion in terms of the Langevin equation, this method can also be described in the implementation as the probability flow ODE.

2. Example 2: Conditional optimal transport

Next, we introduce the conditional optimal transport (cond-OT) schedule used in flow matching. The cond-OT schedule is known as a noise schedule that optimizes the transport of the conditional distribution $\mathcal{P}_t^c(\mathbf{x}|\mathbf{y})$ [63], which is described as follows:

$$m_t = 1 - \frac{t}{\tau}, \quad \sigma_t = \frac{t}{\tau}. \quad (28)$$

This noise schedule represents the change of the parameters along the geodesic from $(\sigma_0, m_0) = (0, 1)$ to $(\sigma_\tau, m_\tau) = (1, 0)$ on the two-dimensional Euclidean space of the standard deviation σ_t and the mean m_t . The dynamics along the geodesic can be regarded as optimal transport based on the 2-Wasserstein distance for the conditional probability distributions [63,70].

3. Example 3: Variance preserving diffusion (VP diffusion)

Finally, we introduce variance preserving diffusion (VP diffusion) [36]. VP diffusion is a method based on denoising diffusion probabilistic models (DDPMs) [40,43]. The DDPMs are diffusion models that use the diffusion from the input data distribution to the Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$. VP diffusion is described by the following Langevin equation [36]:

$$d\mathbf{x}_t = -T_t \mathbf{x}_t dt + \sqrt{2T_t} d\mathbf{B}_t. \quad (29)$$

From Eqs. (2) and (23), we get $\mathbf{f}_t(\mathbf{x}) = -T_t \mathbf{x}$ and $T_t = -\partial_t \ln m_t$. By substituting $T_t = -\partial_t \ln m_t$ into Eq. (22), we also obtain the relation

$$\partial_t \ln(m_t^2) = \frac{\partial_t(\sigma_t^2)}{\sigma_t^2 - 1} = \partial_t \ln(1 - \sigma_t^2). \quad (30)$$

Since the initial conditions are given by $\sigma_0 = 0$ and $m_0 = 1$, the condition

$$m_t^2 + \sigma_t^2 = 1 \quad (31)$$

holds. This condition implies that m_t and σ_t are on the unit circle. Thus, we can consider a noise schedule that changes the parameters along the geodesic on the unit circle from $(\sigma_0, m_0) = (0, 1)$ to $(\sigma_\tau, m_\tau) = (1, 0)$ as follows:

$$\begin{aligned} m_t &= \cos\left(\frac{\pi t}{2\tau}\right), \\ \sigma_t &= \sin\left(\frac{\pi t}{2\tau}\right). \end{aligned} \quad (32)$$

This noise schedule is called the cosine schedule [43].

Under the condition of VP diffusion ($\sigma_t^2 + m_t^2 = 1$), we consider the coordinate transformation to treat the diffusion process without the external force. These methods based on the coordinate transformation are called denoising diffusion implicit models (DDIMs) [48]. The coordinate transformation in DDIMs is introduced as follows:

$$\bar{\mathbf{x}} = \frac{\mathbf{x}}{m_t}. \quad (33)$$

Under this coordinate transformation and the conditions Eqs. (22) and (23) with $T_t = -\partial_t \ln m_t$, the Fokker-Planck equation is given by

$$\partial_t \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) = -\nabla_{\bar{\mathbf{x}}} \cdot \left[\left[\frac{\partial_t \ln m_t}{(m_t)^2} \nabla_{\bar{\mathbf{x}}} \ln \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right] \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right] \quad (34)$$

(see Appendix C), where $\bar{\mathcal{P}}_t(\bar{\mathbf{x}}) := \mathcal{P}_t(\mathbf{x}) (|d\bar{\mathbf{x}}/d\mathbf{x}|)^{-1}$, $\nabla_{\bar{\mathbf{x}}} := m_t \nabla$, and $|d\bar{\mathbf{x}}/d\mathbf{x}|$ is the Jacobian. Therefore, the corresponding Langevin equation is

$$d\bar{\mathbf{x}}_t = \sqrt{\frac{2T_t}{(m_t)^2}} d\mathbf{B}_t. \quad (35)$$

This Langevin equation describes a situation where the external force is absent. In DDIMs, we generate data using the probability flow ODE [Eq. (11)] [48].

III. RELATIONSHIPS BETWEEN STOCHASTIC THERMODYNAMICS AND DIFFUSION MODELS

In this section, we review stochastic thermodynamics (Sec. III A) and the relationship between stochastic thermodynamics and optimal transport theory (Sec. III B) from the perspective of diffusion models.

A. Review on stochastic thermodynamics

We introduce nonequilibrium thermodynamics for the overdamped Fokker-Planck equation [Eq. (1)], namely, stochastic thermodynamics [3], and discuss its relation to diffusion models. In stochastic thermodynamics, we mainly consider the quantity called the entropy production rate as a measure of thermodynamic dissipation rate. For the overdamped Fokker-Planck equation [Eq. (1)], the entropy production rate \dot{S}_t^{tot} is defined as

$$\dot{S}_t^{\text{tot}} = \frac{1}{T_t} \int d\mathbf{x} \|\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}). \quad (36)$$

Here, the Boltzmann constant k_B is regularized to $k_B = 1$. This entropy production rate can be decomposed into the entropy change rate in the system \dot{S}_t^{sys} and the entropy change rate in the heat bath \dot{S}_t^{bath} as follows:

$$\dot{S}_t^{\text{tot}} = \dot{S}_t^{\text{sys}} + \dot{S}_t^{\text{bath}}, \quad (37)$$

$$\dot{S}_t^{\text{sys}} = \partial_t \left[- \int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \ln \mathcal{P}_t(\mathbf{x}) \right], \quad (38)$$

$$\dot{S}_t^{\text{bath}} = \frac{\int d\mathbf{x} \mathbf{f}_t(\mathbf{x}) \cdot \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x})}{T_t}, \quad (39)$$

where we use the partial integration and $\int d\mathbf{x} \partial_t \mathcal{P}_t(\mathbf{x}) = 0$ to derive the expression \dot{S}_t^{sys} . Thus, the entropy production rate \dot{S}_t^{tot} is regarded as the entropy change rate in the total system. Its non-negativity $\dot{S}_t^{\text{tot}} \geq 0$ means the second law of thermodynamics. As a measure of thermodynamic dissipation, the entropy production from time $t = 0$ to $t = \tau$ is also defined as

$$S_\tau^{\text{tot}} = \int_0^\tau dt \dot{S}_t^{\text{tot}}. \quad (40)$$

In the original paper of the diffusion model [35], the idea of a reverse process in diffusion models may be inspired by the fluctuation theorem [8,37,38] or the Jarzynski equality [39], which are the relations between the entropy production and the path probabilities of the forward and backward trajectories. Based on mathematical techniques in the fluctuation theorem such as dual dynamics [3], we introduce the following dynamics:

$$d\mathbf{x}_t = \mathbf{f}_t(\mathbf{x}_t) dt - 2\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}_t) dt + \sqrt{2T_t} d\mathbf{B}_t, \quad (41)$$

which corresponds to the estimated reverse process of the score-based generative modeling [Eq. (9)] and is a special case of the interpolated dynamics [32] for the velocity field $-\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}_t)$.

Here, we discuss two path probabilities of the path $\Gamma = \{\mathbf{x}_0, \mathbf{x}_{\Delta t}, \dots, \mathbf{x}_{N_{\Delta t}}\}$. Here, $\Delta t (> 0)$ is the infinitesimal

time interval, and, thus, we consider the limit $\Delta t \rightarrow 0$ with fixed $\tau = N_\tau \Delta t$. For the Langevin equation [Eq. (2)], the transition probability from the state \mathbf{x} at time t to the state \mathbf{y} at time $t + \Delta t$ is given by the expression of the Onsager-Machlup function [92] as follows:

$$\mathcal{T}_t(\mathbf{y}|\mathbf{x}) = \frac{1}{(4\pi T_t \Delta t)^{n_d/2}} e^{-\|\mathbf{y}-\mathbf{x}-\mathbf{f}_t(\mathbf{x})\Delta t\|^2/4T_t \Delta t}. \quad (42)$$

Similarly, the transition probability from the state \mathbf{x} at time t to the state \mathbf{y} at time $t + \Delta t$ for the process [Eq. (41)] is given by

$$\mathcal{T}_t^\dagger(\mathbf{y}|\mathbf{x}) = \frac{1}{(4\pi T_t \Delta t)^{n_d/2}} e^{-\|\mathbf{y}-\mathbf{x}-\mathbf{f}_t(\mathbf{x})\Delta t+2\mathbf{r}_t^\dagger(\mathbf{x})\Delta t\|^2/4T_t \Delta t}. \quad (43)$$

Using the transition probability [Eq. (42)], we define the path probability for the forward process as

$$\mathbb{P}_F(\Gamma) = \mathcal{P}_0(\mathbf{x}_0) \prod_{N=0}^{N_\tau-1} \mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t}). \quad (44)$$

This path probability $\mathbb{P}_F(\Gamma)$ is related to the distribution of the forward process $\mathcal{P}_t(\mathbf{x})$. If we consider the path excluding the state $\{\mathbf{x}_{n\Delta t}\}$ defined as $\Gamma_{t \neq n\Delta t} = \Gamma \setminus \{\mathbf{x}_{n\Delta t}\}$, marginalizing the path probability $\mathbb{P}_F(\Gamma)$ gives $\mathcal{P}_{n\Delta t}(\mathbf{x}_{n\Delta t}) = \int d\Gamma_{t \neq n\Delta t} \mathbb{P}_F(\Gamma)$. Similarly, the path probability for the process in Eq. (41) is given by

$$\mathbb{P}_B(\Gamma) = \mathcal{P}_0(\mathbf{x}_0) \prod_{N=0}^{N_\tau-1} \mathcal{T}_{N\Delta t}^\dagger(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t}). \quad (45)$$

We remark that this path probability \mathbb{P}_B is not related to the reverse process or the estimated reverse process, and marginalizing the path probability $\mathbb{P}_B(\Gamma)$ does not yield $\mathcal{P}_t^\dagger(\mathbf{x})$ or $\tilde{\mathcal{P}}_t^\dagger(\mathbf{x})$, because this path probability \mathbb{P}_B is not introduced as the time-reversed dynamics from the probability distribution function \mathcal{P}_0^\dagger or $\tilde{\mathcal{P}}_0^\dagger$.

The entropy production can be interpreted as the statistical difference between two path probabilities \mathbb{P}_F and \mathbb{P}_B . If we consider the Kullback-Leibler divergence between $\mathbb{P}_F(\Gamma)$ and $\mathbb{P}_B(\Gamma)$ defined as

$$D_{\text{KL}}(\mathbb{P}_F\|\mathbb{P}_B) = \int d\Gamma \mathbb{P}_F(\Gamma) \ln \frac{\mathbb{P}_F(\Gamma)}{\mathbb{P}_B(\Gamma)}, \quad (46)$$

we obtain the following relation between the Kullback-Leibler divergence and the entropy production [3,22,32]:

$$S_\tau^{\text{tot}} = D_{\text{KL}}(\mathbb{P}_F\|\mathbb{P}_B) \quad (47)$$

(see also Appendix D).

We remark that there are several expressions of the entropy production as the Kullback-Leibler divergence. For

example, the entropy production can be formulated as the projection in information geometry, which is a minimization problem of the Kullback-Leibler divergence [32,103]. In the context of the fluctuation theorem, the expression based on the path probability for the backward trajectory is well discussed. The path probability for the backward trajectory is defined as

$$\mathbb{P}'_B(\Gamma) = \mathcal{P}_\tau(\mathbf{x}_{N_\tau \Delta t}) \prod_{N=0}^{N_\tau-1} \mathcal{T}_{N\Delta t}(\mathbf{x}_{N\Delta t}|\mathbf{x}_{(N+1)\Delta t}). \quad (48)$$

We remark that this path probability \mathbb{P}'_B is not related to the reverse process or the estimated reverse process because this path probability \mathbb{P}'_B uses the transition probability \mathcal{T}_t , which is not the transition probability of the reverse process or the estimated reverse process, in the time-reversed dynamics from the initial condition $\mathcal{P}_\tau = \mathcal{P}_0^\dagger$. We can also obtain the similar relation between the Kullback-Leibler divergence and the entropy production [3,32,104]

$$S_\tau^{\text{tot}} = D_{\text{KL}}(\mathbb{P}_F\|\mathbb{P}'_B), \quad (49)$$

(see also Appendix D). We here define the stochastic entropy production for the path Γ as $s_\tau^{\text{tot}}(\Gamma) = \ln[\mathbb{P}_F(\Gamma)/\mathbb{P}'_B(\Gamma)]$. In stochastic thermodynamics, the identity $\exp[s_\tau^{\text{tot}}(\Gamma)] = \mathbb{P}_F(\Gamma)/\mathbb{P}'_B(\Gamma)$ is known as the detailed fluctuation theorem [8,37,38]. If we introduce the expected value with respect to \mathbb{P}_F as $\mathbb{E}_{\mathbb{P}_F}[\dots] = \int d\Gamma \mathbb{P}_F(\Gamma) \dots$, the entropy production satisfies $S_\tau^{\text{tot}} = \mathbb{E}_{\mathbb{P}_F}[s_\tau^{\text{tot}}]$. We also can obtain the following identity:

$$\mathbb{E}_{\mathbb{P}_F}[\exp(-s_\tau^{\text{tot}})] = 1, \quad (50)$$

where we use the normalization $\int d\Gamma \mathbb{P}'_B(\Gamma) = 1$. This formula is known as the integral fluctuation theorem [37,38], and a special case of the integral fluctuation theorem is known as the Jarzynski equality [39].

We remark on the analogous connection between the fluctuation theorem and the diffusion model discussed in the original paper on diffusion models [35]. In Ref. [35], the path probability for the estimated reverse process \mathbb{P}_E is introduced as

$$\mathbb{P}_E(\Gamma) = \tilde{\mathcal{P}}_0^\dagger(\mathbf{x}_{N_\tau \Delta t}) \prod_{N=0}^{N_\tau-1} \mathcal{T}_{N\Delta t}^E(\mathbf{x}_{N\Delta t}|\mathbf{x}_{(N+1)\Delta t}), \quad (51)$$

and $p(\mathbf{x}_0)$ is obtained from the marginalization $p(\mathbf{x}_0) = \int d\Gamma_{t \neq 0} \mathbb{P}_E(\Gamma)$, where \mathcal{T}_t^E is the transition probability for the estimated reverse process. The expression of \mathbb{P}_E is analogous to the expression of \mathbb{P}'_B . While we consider the statistical difference between \mathbb{P}_F and \mathbb{P}'_B in stochastic thermodynamics, we consider the situation where \mathbb{P}_E statistically mimics \mathbb{P}_F in diffusion models. In Ref. [35], the transition probability \mathcal{T}_t^E is estimated from the problem

of maximizing the lower bound K on the model log-likelihood $L(q||p) = \int d\mathbf{x} q(\mathbf{x}) \ln p(\mathbf{x}) (\geq K)$. The lower bound K includes the terms of the Kullback-Leibler divergence and the differential entropy, and the derivation of K is analogous to the derivation of the Jarzynski equality. The maximization of K is solved as a partial minimization of the terms of the Kullback-Leibler divergence.

As described above, there are several analogous connections between stochastic thermodynamics and diffusion models. Although not trivial, it may be possible to apply stochastic thermodynamic techniques to diffusion models due to similarities in mathematical structure. In particular, interesting techniques in stochastic thermodynamics are optimal transport theoretic techniques for the minimum entropy production problem. We next discuss optimal transport theory [25,105].

B. Optimal transport theory and minimum entropy production

In this section, we introduce optimal transport theory and its application to stochastic thermodynamics. We start with the distance in optimal transport theory, namely, the p -Wasserstein distance [25]. The p -Wasserstein distance $\mathcal{W}_p(P, Q)$ between probability density functions $P(\mathbf{x})$ and $Q(\mathbf{y})$ satisfying $P(\mathbf{x}) \geq 0$, $\int d\mathbf{x} P(\mathbf{x}) = 1$, $Q(\mathbf{y}) \geq 0$, and $\int d\mathbf{y} Q(\mathbf{y}) = 1$ is defined as

$$\mathcal{W}_p(P, Q) := \left(\inf_{\pi \in \Pi(P, Q)} \int d\mathbf{x} d\mathbf{y} \|\mathbf{x} - \mathbf{y}\|^p \pi(\mathbf{x}, \mathbf{y}) \right)^{1/p}. \quad (52)$$

Here, $\Pi(P, Q)$ is a set of joint probability density functions defined as

$$\Pi(P, Q) = \left\{ \pi(\mathbf{x}, \mathbf{y}) \mid \pi(\mathbf{x}, \mathbf{y}) \geq 0, \int d\mathbf{y} \pi(\mathbf{x}, \mathbf{y}) = P(\mathbf{x}), \int d\mathbf{x} \pi(\mathbf{x}, \mathbf{y}) = Q(\mathbf{y}) \right\}. \quad (53)$$

If the p th-order moment is finite, the p -Wasserstein distance $\mathcal{W}_p(P, Q)$ remains finite and is well defined. In the space of probability distribution functions which have a finite p th-order moment, the metric space axioms are satisfied [25] as follows:

$$\mathcal{W}_p(P, Q) \geq 0, \quad (54)$$

$$\mathcal{W}_p(P, Q) = 0 \Leftrightarrow P = Q, \quad (55)$$

$$\mathcal{W}_p(P, Q) = \mathcal{W}_p(Q, P), \quad (56)$$

$$\mathcal{W}_p(P, Q) \leq \mathcal{W}_p(P, R) + \mathcal{W}_p(R, Q), \quad (57)$$

where P , Q , and R are probability density functions. It is also known that $\mathcal{W}_p(P, Q) \leq \mathcal{W}_q(P, Q)$ for $p \leq q$ because of Hölder's inequality [25]. The 1-Wasserstein distance and the 2-Wasserstein distance are well used in machine learning and stochastic thermodynamics due to some mathematical properties.

For example, the 1-Wasserstein distance has a computational advantage, because it is computed from expected values. Because of the Kantorovich-Rubinstein duality [25], the 1-Wasserstein distance is given by the following maximization problem:

$$\mathcal{W}_1(P, Q) = \sup_{\psi \in \text{Lip}^1} (\mathbb{E}_P[\psi] - \mathbb{E}_Q[\psi]), \quad (58)$$

where Lip^1 is the set of scalar functions satisfying 1-Lipschitz continuity defined as $\text{Lip}^1 = \{\psi(\mathbf{x}) \mid \|\nabla \psi(\mathbf{x})\| \leq 1\}$ and $\mathbb{E}_P[\psi]$ is the expected value with respect to the probability density function $P(\mathbf{x})$ defined as $\mathbb{E}_P[\psi] = \int d\mathbf{x} P(\mathbf{x}) \psi(\mathbf{x})$. The computational complexity of the 1-Wasserstein distance via the above maximization problem is relatively simple. Therefore, the 1-Wasserstein distance is well used as an objective function in generative models such as the Wasserstein generative adversarial network [64].

In addition, the 2-Wasserstein distance has good differential geometric properties. If P and Q are n_d -dimensional Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_P, \Sigma_P)$ and $\mathcal{N}(\boldsymbol{\mu}_Q, \Sigma_Q)$, respectively, then the 2-Wasserstein distance can be computed as follows [106]:

$$\begin{aligned} \mathcal{W}_2(P, Q)^2 &= \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|^2 \\ &\quad + \text{tr}\{\Sigma_P + \Sigma_Q - 2[(\Sigma_P)^{1/2} \Sigma_Q (\Sigma_P)^{1/2}]^{1/2}\}, \end{aligned} \quad (59)$$

where tr denotes the trace of the matrix. If the variance-covariance matrices are given by $\Sigma_P = \sigma_P^2 \mathbf{I}$ and $\Sigma_Q = \sigma_Q^2 \mathbf{I}$, Eq. (59) is rewritten as

$$\begin{aligned} \mathcal{W}_2(P, Q)^2 &= \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|^2 + \text{tr}[(\sigma_P - \sigma_Q)^2 \mathbf{I}] \\ &= \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|^2 + (\sigma_P - \sigma_Q)^2 n_d \\ &= \|\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q\|^2 + \|\sigma_P \mathbf{1} - \sigma_Q \mathbf{1}\|^2, \end{aligned} \quad (60)$$

where $\mathbf{1}$ is the vector defined as $[\mathbf{1}]_i = 1$ ($i = 1, \dots, n_d$). In this case, the 2-Wasserstein distance is equivalent to the distance between $(\boldsymbol{\mu}_P, \sigma_P \mathbf{1})$ and $(\boldsymbol{\mu}_Q, \sigma_Q \mathbf{1})$ in the $2n_d$ -dimensional Euclidean space of the means and the standard deviations.

Note that the right-hand side of Eq. (59) is called the Fréchet inception distance (FID) [82] when P and Q represent the distributions of the real-world image data and the generated image data in the latent space of the inception model [107]. The right-hand side of Eq. (59)

(i.e., the FID) is always less than $\mathcal{W}_2(P, Q)^2$ for the non-Gaussian distributions P and Q [108], where their means and covariances are given by $(\boldsymbol{\mu}_P, \Sigma_P)$ and $(\boldsymbol{\mu}_Q, \Sigma_Q)$, respectively.

Next, we explain that the 2-Wasserstein distance provides a lower bound on the entropy production in stochastic thermodynamics. We start with the following Benamou-Brenier formula [109]:

$$\mathcal{W}_2(\mathcal{P}_\tau, \mathcal{P}_0) = \sqrt{\inf_{\tau} \int_0^\tau dt \int d\mathbf{x} \|\mathbf{v}_t(\mathbf{x})\|^2 q_t(\mathbf{x})}, \quad (61)$$

where the infimum is taken among all paths $[\mathbf{v}_t(\mathbf{x}), q_t(\mathbf{x})]_{0 \leq t \leq \tau}$ satisfying

$$\begin{aligned} \partial_t q_t(\mathbf{x}) &= -\nabla \cdot [\mathbf{v}_t(\mathbf{x}) q_t(\mathbf{x})], \\ q_0(\mathbf{x}) &= \mathcal{P}_0(\mathbf{x}), \quad q_\tau(\mathbf{x}) = \mathcal{P}_\tau(\mathbf{x}). \end{aligned} \quad (62)$$

Because the Fokker-Planck equation [Eq. (1)] is the continuity equation, we obtain

$$[\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)]^2 \leq \tau \int_0^\tau dt \int d\mathbf{x} \|\mathbf{v}_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \quad (63)$$

from Eq. (61). If T_t does not depend on t , this result can be written as

$$S_\tau^{\text{tot}} \geq \frac{\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)^2}{\tau T}, \quad (64)$$

where $T (= T_t)$ is the time-independent temperature. Thus, the 2-Wasserstein distance provides the formula for the minimum entropy production in a finite time τ under the condition that the initial distribution \mathcal{P}_0 and the final distribution \mathcal{P}_τ are fixed [30].

This result [Eq. (64)] can be interpreted from the perspective of the geodesic in the space of the 2-Wasserstein distance [29,32]. If we consider the infinitesimal time evolution from t to $t + \Delta t$, Eq. (61) indicates

$$\begin{aligned} [\mathcal{W}_2(\mathcal{P}_t, \mathcal{P}_{t+\Delta t})]^2 &\leq \Delta t \int_t^{t+\Delta t} dt' \int d\mathbf{x} \|\mathbf{v}_{t'}^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x}) \\ &= (\Delta t)^2 \int d\mathbf{x} \|\mathbf{v}_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) + O[(\Delta t)^3], \end{aligned} \quad (65)$$

where $O[(\Delta t)^3]$ denotes Landau's big O notation. If we define the speed in the space of the 2-Wasserstein distance as

$$v_2(t) = \lim_{\Delta t \rightarrow +0} \frac{\mathcal{W}_2(\mathcal{P}_t, \mathcal{P}_{t+\Delta t})}{\Delta t}, \quad (66)$$

we obtain the lower bound of the entropy production rate [29] from Eq. (65) as

$$\dot{S}_t^{\text{tot}} \geq \frac{[v_2(t)]^2}{T_t}. \quad (67)$$

This lower bound $[v_2(t)]^2/T_t$ is called the excess entropy production rate [31,32,110]. The equality $\dot{S}_t^{\text{tot}} = [v_2(t)]^2/T_t$ holds when the velocity field is given by a gradient of a potential function $\phi_t(\mathbf{x})$ as $\mathbf{v}_t^{\mathcal{P}}(\mathbf{x}) = T_t \nabla \phi_t(\mathbf{x})$ [29,32,111]. The equality condition physically means the situation where the external force $\mathbf{f}_t(\mathbf{x})$ is given by conservative force $\mathbf{f}_t(\mathbf{x}) = -\nabla U_t(\mathbf{x})$ with the potential energy $U_t(\mathbf{x})$. For the case where the external force is given by $\mathbf{f}_t(\mathbf{x}) = \mathbf{A}_t \mathbf{x} + \mathbf{b}_t$, the equality holds if \mathbf{A}_t is a symmetric matrix [112]. Equation (67) can be written as $T_t \dot{S}_t^{\text{tot}} \geq [v_2(t)]^2$. The upper bound $T_t \dot{S}_t^{\text{tot}}$ can be regarded as the instantaneous dissipative work [113] when we consider the transition between equilibrium states.

If $T_t (= T)$ does not depend on t , we obtain the following hierarchy of inequalities for the minimum entropy production known as the thermodynamic speed limits [29]:

$$\begin{aligned} S_\tau^{\text{tot}} = \int_0^\tau dt \dot{S}_t^{\text{tot}} &\geq \frac{\int_0^\tau dt [v_2(t)]^2}{T} \\ &\geq \frac{\mathcal{L}_\tau^2}{\tau T} \\ &\geq \frac{\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)^2}{\tau T}, \end{aligned} \quad (68)$$

where $\mathcal{L}_\tau = \int_0^\tau dt v_2(t)$ is the path length in the space of the 2-Wasserstein distance and we use the Cauchy-Schwarz inequality $[\int_0^\tau dt][\int_0^\tau dt [v_2(t)]^2] \geq [\int_0^\tau dt v_2(t)]^2$ and the triangle inequality $\mathcal{L}_\tau \geq \mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)$. The path length \mathcal{L}_τ can also be calculated as $\mathcal{L}_\tau = \lim_{\Delta t \rightarrow +0} \sum_{n=0}^{N_\tau-1} \mathcal{W}_2(\mathcal{P}_{(n+1)\Delta t}, \mathcal{P}_{n\Delta t})$ with fixed $\tau = N_\tau \Delta t$. If the probability density function \mathcal{P}_t evolves along the geodesic in the space of the 2-Wasserstein distance, the following condition:

$$v_2(t) = \frac{\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)}{\tau} \quad (69)$$

holds. This condition provides the equality condition of Eq. (68):

$$\frac{\int_0^\tau dt [v_2(t)]^2}{T} = \frac{\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)^2}{\tau T}. \quad (70)$$

If the system is driven by a conservative force, $\dot{S}_t^{\text{tot}} = [v_2(t)]^2/T$ holds. Thus, the minimum entropy production $S_\tau^{\text{tot}} = [\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)]^2/(\tau T)$ is achieved when the system is driven by a conservative force and the probability density function \mathcal{P}_t evolves along the geodesic in the space of the 2-Wasserstein distance.

We discuss the thermodynamic uncertainty relations [31,32,110,114] as other lower bounds on the entropy production rate \dot{S}_t^{tot} and the excess entropy production rate

$[v_2(t)]^2/T_t$. Since the equality in Eq. (67) is achieved when $\nu_t^{\mathcal{P}}(\mathbf{x}) = T_t \nabla \phi_t(\mathbf{x})$, the following equality:

$$\frac{[v_2(t)]^2}{T_t} = T_t \int d\mathbf{x} \|\nabla \phi_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \quad (71)$$

holds. This gives us the following inequality:

$$\begin{aligned} & \left[\int d\mathbf{x} \|\nabla \phi_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right] \mathbb{E}_{\mathcal{P}_t}[\|\nabla r\|^2] \\ & \geq \left[\int d\mathbf{x} [\nabla \phi_t(\mathbf{x})] \cdot [\nabla r(\mathbf{x})] \mathcal{P}_t(\mathbf{x}) \right]^2 \\ & = \left[\frac{1}{T_t} \int d\mathbf{x} r(\mathbf{x}) \partial_t \mathcal{P}_t(\mathbf{x}) \right]^2 = \left[\frac{1}{T_t} \partial_t \mathbb{E}_{\mathcal{P}_t}[r] \right]^2 \end{aligned} \quad (72)$$

for any time-independent observable $r(\mathbf{x})$, where the expected value $\mathbb{E}_{\mathcal{P}_t}[r]$ is defined as $\mathbb{E}_{\mathcal{P}_t}[r] = \int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) r(\mathbf{x})$ and we use the Cauchy-Schwarz inequality, the continuity equation $\partial_t \mathcal{P}_t(\mathbf{x}) = -\nabla \cdot \{T_t [\nabla \phi_t(\mathbf{x})] \mathcal{P}_t(\mathbf{x})\}$, and the partial integration. Note that the validity of this continuity equation is due to the Benamou-Brenier formula [109]. From Eqs. (67) and (72), we obtain the thermodynamic uncertainty relation for the excess entropy production rate [31,110]:

$$\dot{S}_t^{\text{tot}} \geq \frac{[v_2(t)]^2}{T_t} \geq \frac{[\partial_t \mathbb{E}_{\mathcal{P}_t}[r]]^2}{T_t \mathbb{E}_{\mathcal{P}_t}[\|\nabla r\|^2]}, \quad (73)$$

which implies a trade-off relation between the speed of the observable $|\partial_t \mathbb{E}_{\mathcal{P}_t}[r]|$ and thermodynamic dissipation rate \dot{S}_t^{tot} . If we define the normalized speed of the observable $r(\mathbf{x})$ as $v_r(t) = |\partial_t \mathbb{E}_{\mathcal{P}_t}[r]| / \sqrt{\mathbb{E}_{\mathcal{P}_t}[\|\nabla r\|^2]}$, this inequality [Eq. (73)] can be written as

$$[v_2(t)]^2 \geq [v_r(t)]^2. \quad (74)$$

Thus, this expression [Eq. (74)] implies that the speed in the space of the 2-Wasserstein distance is the upper bound on the speed of any time-independent observable $r(\mathbf{x})$. We note that this expression [Eq. (74)] can be regarded as the inequality $[\mathcal{W}_2(P_t, P_{t+\Delta t})/\Delta t]^2 \geq [\mathcal{W}_1(P_t, P_{t+\Delta t})/\Delta t]^2$ in the limit $\Delta t \rightarrow +0$ if we substitute the 1-Lipschitz function $\psi \in \text{Lip}^1$, which gives the 1-Wasserstein distance, into r [33].

By considering the following maximization problem of this lower bound with respect to the observable $r(\mathbf{x})$:

$$r^*(\mathbf{x}) \in \text{argmax}_{r(\mathbf{x})} \frac{[\partial_t \mathbb{E}_{\mathcal{P}_t}[r]]^2}{T_t \mathbb{E}_{\mathcal{P}_t}[\|\nabla r\|^2]}, \quad (75)$$

the optimal observable $r^*(\mathbf{x})$ gives the speed

$$v_2(t) = \frac{|\partial_t \mathbb{E}_{\mathcal{P}_t}[r^*]|}{\sqrt{\mathbb{E}_{\mathcal{P}_t}[\|\nabla r^*\|^2]}}, \quad (76)$$

and the gradient of the optimal observable $\nabla r^*(\mathbf{x})$ is proportional to the velocity field $T_t \nabla \phi_t(\mathbf{x}) = c \nabla r^*(\mathbf{x})$ with the proportional coefficient $c = v_2(t) / \sqrt{\mathbb{E}_{\mathcal{P}_t}[\|\nabla r^*\|^2]}$ [31,110]. Even if the external force $\mathbf{f}_t(\mathbf{x})$ is not conservative, there exists a method to estimate the velocity field from a time series data of the Langevin dynamics by considering the maximization problem based on the other thermodynamic uncertainty relation [18,21,23], which may be useful for estimating the velocity field in diffusion models.

IV. MAIN RESULT

In this section, we explain the main results of the paper. Based on optimal transport theory and stochastic thermodynamics, we derive upper bounds on the estimation error in data generation of diffusion models by the entropy production rate, namely, the speed-accuracy relations for diffusion models. This upper bound is derived based on an analogy to the thermodynamic speed limit and the thermodynamic uncertainty relation. Based on the upper bound, we can introduce an optimal protocol for the accuracy of data generation in diffusion models given by the geodesic in the space of the 2-Wasserstein distance. This protocol is analogous to the optimal protocol for minimum entropy production. We discuss the relationship between the optimal protocol and the noise schedules such as the cond-OT schedule and the cosine schedule, and we illustrate the main results numerically.

A. Speed-accuracy relations for diffusion models

We explain the main result of the paper. If the velocity field of the forward process is accurately reconstructed $[\hat{\nu}_t(\mathbf{x}) = \nu_t^{\mathcal{P}}(\mathbf{x})]$ in the probability flow ODE or the flow-based generative modeling, we can derive the following inequality:

$$\frac{1}{\tau} \frac{(\Delta \mathcal{W}_1)^2}{D_0} \leq \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}, \quad (77)$$

which is a relation between the thermodynamic quantity and the estimation error in data generation. Inspired by the notion of energy-speed-accuracy trade-off relation [115], we call this result the speed-accuracy relation for diffusion models, because $\Delta \mathcal{W}_1$ corresponds to the accuracy, and $T_t \dot{S}_t^{\text{tot}}$ can correspond to the speed in the absence of the nonconservative force, as discussed later. The right-hand side of this inequality is given by thermodynamic quantities such as the entropy production rate \dot{S}_t^{tot} and the temperature T_t , and these thermodynamic quantities are computed in the forward process. The quantities on the left-hand side of this inequality are given by the difference between the reverse process and the estimated reverse process. The quantity D_0 is the Pearson χ^2 divergence [116] defined as

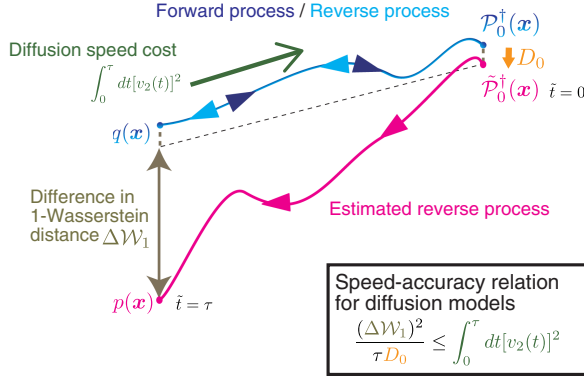


FIG. 3. Illustration of the speed-accuracy relation for diffusion models [Eq. (79)] in the absence of the nonconservative force. We evaluate the difference between the initial conditions of the reverse process \mathcal{P}_0^\dagger and the estimated reverse process $\tilde{\mathcal{P}}_0^\dagger$ by the Pearson χ^2 divergence D_0 and the estimation error of the generated data by the change of the 1-Wasserstein distance $\Delta\mathcal{W}_1$. The speed-accuracy relation for diffusion models implies that the diffusion speed cost $\int_0^\tau dt[v_2(t)]^2$ measured by the 2-Wasserstein distance in the forward process affects the accuracy of the generated data given by the response function $(\Delta\mathcal{W}_1)^2/D_0$.

$$D_0 := \int dx \frac{[\tilde{\mathcal{P}}_0^\dagger(x) - \mathcal{P}_0^\dagger(x)]^2}{\mathcal{P}_0^\dagger(x)}, \quad (78)$$

which quantifies the difference between the initial conditions of the reverse process $\mathcal{P}_0^\dagger(x)$ and the estimated reverse process $\tilde{\mathcal{P}}_0^\dagger(x)$ at $\tilde{t} = 0$. This Pearson's χ^2 divergence is the f divergence, which becomes the Fisher information when the difference $\tilde{\mathcal{P}}_0^\dagger(x) - \mathcal{P}_0^\dagger(x)$ is sufficiently small [83]. The quantity $\Delta\mathcal{W}_1 := \mathcal{W}_1(q, p) - \mathcal{W}_1(\mathcal{P}_0^\dagger, \tilde{\mathcal{P}}_0^\dagger)$ is the change of the 1-Wasserstein distance between reverse and estimated reverse processes from time $\tilde{t} = 0$ to $\tilde{t} = \tau$. In generative models, we sometimes consider the estimation error using the 1-Wasserstein distance $\mathcal{W}_1(q, p)$ for model evaluation [66]. If we consider the initial condition satisfying $\mathcal{W}_1(\mathcal{P}_0^\dagger, \tilde{\mathcal{P}}_0^\dagger) \approx 0$, $\Delta\mathcal{W}_1$ becomes the estimation error $\Delta\mathcal{W}_1 \approx \mathcal{W}_1(q, p)$. Therefore, $(\Delta\mathcal{W}_1)^2/D_0$ is considered as a response function, which quantifies how much the perturbation of the initial distribution D_0 for the estimated reverse process at $\tilde{t} = 0$ affects the estimation error of the generated data $\Delta\mathcal{W}_1$ at $\tilde{t} = \tau$ (see also Fig. 3). If the response function $(\Delta\mathcal{W}_1)^2/D_0$ is smaller, the accuracy of the generated data is better regardless of the initial condition deviations. Thus, a smaller value of $(\Delta\mathcal{W}_1)^2/D_0$ means robustness of diffusion models against the change of the initial distribution $\tilde{\mathcal{P}}_0^\dagger(x)$. In other words, the value of $(\Delta\mathcal{W}_1)^2/D_0$ represents the degree of change in the generated data with respect to the initial perturbation. Therefore, the value of $(\Delta\mathcal{W}_1)^2/D_0$ can also be considered as the sensitivity of the generated data. Our result shows that the thermodynamic quantities

corresponding to the thermodynamic dissipation in the forward process $\int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$ give the upper bound on the sensitivity of diffusion models $(\Delta\mathcal{W}_1)^2/D_0$.

In particular, when the external force is conservative, and, thus, the velocity field of the forward process is given by the gradient of a potential $[\mathbf{v}_t^{\mathcal{P}}(\mathbf{x}) = T_t \nabla \phi_t(\mathbf{x})]$, we also obtain

$$\frac{1}{\tau} \frac{(\Delta\mathcal{W}_1)^2}{D_0} \leq \int_0^\tau dt [v_2(t)]^2 \quad (79)$$

from the equality condition of Eq. (67). This result in the absence of the nonconservative force is also called as the speed-accuracy relation for diffusion models. The right-hand side of this inequality $\int_0^\tau dt [v_2(t)]^2$, which corresponds to the excess entropy production, is given by the diffusion speed measured by the 2-Wasserstein distance $v_2(t)$ in the forward process. This quantity $\int_0^\tau dt [v_2(t)]^2$ is a purely information-theoretic quantity which is considered as the diffusion speed cost in the forward process (see also Fig. 3). According to this result [Eq. (79)], the response function $(\Delta\mathcal{W}_1)^2/D_0$ becomes larger as the diffusion speed cost $\int_0^\tau dt [v_2(t)]^2$ becomes larger. In other words, the accuracy of the generated data can decrease as the diffusion speed increases in the forward process. Since we usually consider a conservative force in the implementations of the flow-based generative modeling and the probability flow ODE, this result without an explicit expression by thermodynamic dissipation may be more useful and intuitive rather than the general result [Eq. (77)].

According to the speed-accuracy relation for diffusion models [Eq. (79)], we can consider minimizing the diffusion speed cost $\int_0^\tau dt [v_2(t)]^2$ to increase the accuracy of data generation, which is quantified by the smallness of the response function $(\Delta\mathcal{W}_1)^2/D_0$. Therefore, minimizing the diffusion speed cost $\int_0^\tau dt [v_2(t)]^2$ can improve the performance of diffusion models. When the initial state \mathcal{P}_0 , the final state \mathcal{P}_τ , and the time duration τ are fixed, the minimization problem can be discussed based on the geodesic in the 2-Wasserstein metric space. We discuss this minimization problem of the diffusion speed cost in Sec. IV C.

As an instantaneous expression of the speed-accuracy relations for diffusion models, the following detailed inequality can be derived:

$$\frac{[\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)]^2}{D_0} \leq T_t \dot{S}_t^{\text{tot}}. \quad (80)$$

This equation is also regarded as $[v_{\text{loss}}(t)]^2 \leq T_t \dot{S}_t^{\text{tot}}$, where $v_{\text{loss}}(t) := |\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)|/\sqrt{D_0}$ is the change rate of the normalized estimation error. This inequality is analogous to the thermodynamic uncertainty relation [Eq. (74)]. We also obtain the hierarchy of inequalities corresponding

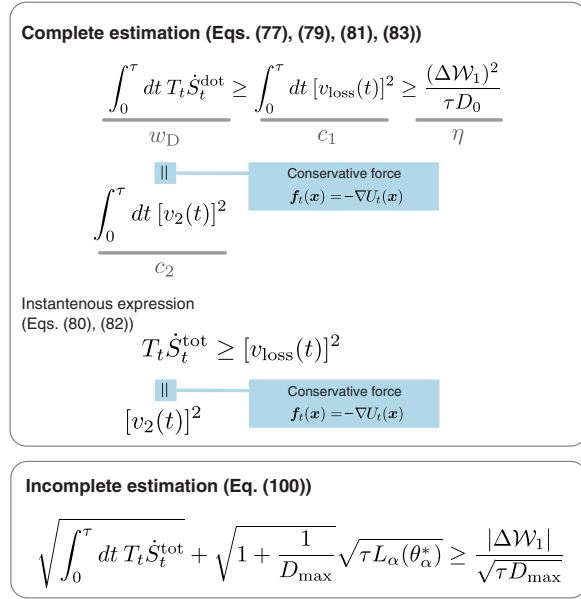


FIG. 4. Summary of the main results. We show a hierarchy between the inequalities and a generalization for incomplete estimation discussed in Sec. VA.

to the speed-accuracy relations for diffusion models as follows:

$$\frac{(\Delta \mathcal{W}_1)^2}{\tau D_0} \leq \int_0^\tau dt [v_{\text{loss}}(t)]^2 \leq \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}, \quad (81)$$

which is analogous to the hierarchy of the thermodynamic speed limits [Eq. (68)].

Especially when the external force is conservative, $[v_2(t)]^2 = T_t \dot{S}_t^{\text{tot}}$ holds. Thus, Eq. (80) indicates that the speed in the space of the 2-Wasserstein distance $v_2(t)$ is the upper bound on $v_{\text{loss}}(t) := |\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)| / \sqrt{D_0}$ as

$$[v_2(t)]^2 \geq [v_{\text{loss}}(t)]^2. \quad (82)$$

Equation (81) is also rewritten as

$$\frac{(\Delta \mathcal{W}_1)^2}{\tau D_0} \leq \int_0^\tau dt [v_{\text{loss}}(t)]^2 \leq \int_0^\tau dt [v_2(t)]^2. \quad (83)$$

In summary, we have illustrated various inequalities in Fig. 4. As shown in Fig. 4, we also note that we later discuss a generalization of the main result for incomplete estimation in Sec. VA. In Sec. IV D, we also use the abbreviated notations $w_D = \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$, $c_2 = \int_0^\tau dt [v_2(t)]^2$, $c_1 = \int_0^\tau dt [v_{\text{loss}}(t)]^2$, and $\eta = (\Delta \mathcal{W}_1)^2 / \tau D_0$. Thus, Eqs. (81) and (83) can be rewritten as $w_D \geq c_1 \geq \eta$ and $c_2 \geq c_1 \geq \eta$, respectively (see also Fig. 4).

B. Proofs of the main result

Here, we show the derivation of the main result. We start with the derivation of Eq. (80). For the sake of simplicity, we introduce a notation for the probability density function of the estimated reverse process $\tilde{\mathcal{P}}_t(\mathbf{x}) := \tilde{\mathcal{P}}_{\tau-t}^\dagger(\mathbf{x})$ according to the time direction of the forward process. Based on the Kantorovich-Rubinstein duality, the 1-Wasserstein distance is given by

$$\mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger) = \mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t) = \mathbb{E}_{\mathcal{P}_t}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi_t^*], \quad (84)$$

where $\psi_t^* = \arg\max_{\psi \in \text{Lip}^1} (\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi])$ is the optimal solution of the maximization problem [Eq. (58)].

Since we assume that the velocity field of the forward process is accurately reconstructed $[\hat{\nu}_t(\mathbf{x}) = \nu_t^{\mathcal{P}}(\mathbf{x})]$, the continuity equations for the reverse process and the estimated reverse process [Eqs. (3) and (4)] are given, respectively, by $\partial_t \mathcal{P}_t(\mathbf{x}) = -\nabla \cdot (\nu_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x}))$ and $\partial_t \tilde{\mathcal{P}}_t(\mathbf{x}) = -\nabla \cdot (\nu_t^{\mathcal{P}}(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x}))$. Therefore, the time evolution of the difference between two probability density functions $\delta \mathcal{P}_t(\mathbf{x}) := \mathcal{P}_t(\mathbf{x}) - \tilde{\mathcal{P}}_t(\mathbf{x})$ is given by

$$\partial_t \delta \mathcal{P}_t(\mathbf{x}) = -\nabla \cdot (\nu_t^{\mathcal{P}}(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x})). \quad (85)$$

Using Eq. (85), we obtain

$$\begin{aligned} \partial_t (\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi]) &= \int d\mathbf{x} \psi(\mathbf{x}) \partial_t \delta \mathcal{P}_t(\mathbf{x}) \\ &= \int d\mathbf{x} \nabla \psi(\mathbf{x}) \cdot \nu_t^{\mathcal{P}}(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x}), \end{aligned} \quad (86)$$

for any time-independent 1-Lipschitz function $\psi \in \text{Lip}^1$, where we use the partial integration and we assume that $\delta \mathcal{P}_t(\mathbf{x})$ disappears at infinity. From the Cauchy-Schwarz inequality and the 1-Lipschitz continuity $\|\nabla \psi(\mathbf{x})\| \leq 1$, we obtain

$$\begin{aligned} &(\partial_t (\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi]))^2 \\ &= \left(\int d\mathbf{x} \nabla \psi(\mathbf{x}) \cdot \nu_t^{\mathcal{P}}(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x}) \right)^2 \\ &\leq \left(\int d\mathbf{x} \|\nu_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right) \left(\int d\mathbf{x} \|\nabla \psi(\mathbf{x})\|^2 \frac{(\delta \mathcal{P}_t(\mathbf{x}))^2}{\mathcal{P}_t(\mathbf{x})} \right) \\ &\leq \left(\int d\mathbf{x} \|\nu_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right) \left(\int d\mathbf{x} \frac{(\delta \mathcal{P}_t(\mathbf{x}))^2}{\mathcal{P}_t(\mathbf{x})} \right) \\ &= T_t \dot{S}_t^{\text{tot}} D_{\tau-t}, \end{aligned} \quad (87)$$

where $D_{\tau-t}$ is the Pearson χ^2 divergence at time $\tilde{t} = \tau - t$ defined as $D_{\tau-t} := \int d\mathbf{x} [\mathcal{P}_{\tau-t}^\dagger(\mathbf{x}) - \tilde{\mathcal{P}}_{\tau-t}^\dagger(\mathbf{x})]^2 / \mathcal{P}_{\tau-t}^\dagger(\mathbf{x})$. We can show that the time derivative of $D_{\tau-t}$ is $\partial_t D_{\tau-t} = 0$ (see Appendix E). Therefore, $D_{\tau-t} = D_0 (= \text{const})$ holds, and we obtain the inequality

$$|\partial_t(\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi])| \leq \sqrt{T_t \dot{S}_t^{\text{tot}} D_0}, \quad (88)$$

for any 1-Lipschitz function $\psi \in \text{Lip}^1$. We now implicitly assume the existence of $\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)$, which can be justified by the smoothness of $\mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)$ as a function of t . In the case of $\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger) \geq 0$, we obtain the lower bound on $\sqrt{T_t \dot{S}_t^{\text{tot}} D_0}$ as follows:

$$\begin{aligned} 0 &\leq \partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger) \\ &= \lim_{\Delta t \rightarrow +0} \frac{\mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t) - \mathcal{W}_1(\mathcal{P}_{t-\Delta t}, \tilde{\mathcal{P}}_{t-\Delta t})}{\Delta t} \\ &\leq \lim_{\Delta t \rightarrow +0} \left[\frac{\mathbb{E}_{\mathcal{P}_t}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi_t^*] - \mathbb{E}_{\mathcal{P}_{t-\Delta t}}[\psi_t^*] + \mathbb{E}_{\tilde{\mathcal{P}}_{t-\Delta t}}[\psi_t^*]}{\Delta t} \right] \\ &= \partial_t(\mathbb{E}_{\mathcal{P}_t}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi_t^*])|_{s=t} \leq \sqrt{T_t \dot{S}_t^{\text{tot}} D_0}, \end{aligned} \quad (89)$$

where we use the fact that $\mathcal{W}_1(\mathcal{P}_{t-\Delta t}, \tilde{\mathcal{P}}_{t-\Delta t}) = \mathbb{E}_{\mathcal{P}_{t-\Delta t}}[\psi_{t-\Delta t}^*] - \mathbb{E}_{\tilde{\mathcal{P}}_{t-\Delta t}}[\psi_{t-\Delta t}^*] \geq \mathbb{E}_{\mathcal{P}_{t-\Delta t}}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_{t-\Delta t}}[\psi_t^*]$ holds because of the expression of the 1-Wasserstein distance [Eq. (58)] and we use the fact that Eq. (87) holds for any t -independent 1-Lipschitz function $\psi_s^* \in \text{Lip}^1$. In the case of $\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger) \leq 0$, we obtain the lower bound on $\sqrt{T_t \dot{S}_t^{\text{tot}} D_0}$ similarly as follows:

$$\begin{aligned} 0 &\leq -\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger) \\ &= \lim_{\Delta t \rightarrow +0} \frac{\mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t) - \mathcal{W}_1(\mathcal{P}_{t+\Delta t}, \tilde{\mathcal{P}}_{t+\Delta t})}{\Delta t} \\ &\leq -\partial_t(\mathbb{E}_{\mathcal{P}_t}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi_t^*])|_{s=t} \leq \sqrt{T_t \dot{S}_t^{\text{tot}} D_0}, \end{aligned} \quad (90)$$

where we use the inequality $\mathcal{W}_1(\mathcal{P}_{t+\Delta t}, \tilde{\mathcal{P}}_{t+\Delta t}) = \mathbb{E}_{\mathcal{P}_{t+\Delta t}}[\psi_{t+\Delta t}^*] - \mathbb{E}_{\tilde{\mathcal{P}}_{t+\Delta t}}[\psi_{t+\Delta t}^*] \geq \mathbb{E}_{\mathcal{P}_{t+\Delta t}}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_{t+\Delta t}}[\psi_t^*]$ and Eq. (87). By squaring both sides of the inequalities (89) and (90), the resulting inequalities are equivalent to Eq. (80). Thus, our main result Eq. (80) is proved regardless of the sign of $\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)$.

We next derive Eq. (77). By integrating Eq. (80) with respect to time and applying the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} \int_0^\tau dt T_t \dot{S}_t^{\text{tot}} &\geq \frac{1}{D_0} \int_0^\tau dt [\partial_t \mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t)]^2 \\ &\geq \frac{1}{\tau D_0} \left(\int_0^\tau dt \partial_t \mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t) \right)^2 \\ &= \frac{(\Delta \mathcal{W}_1)^2}{\tau D_0}, \end{aligned} \quad (91)$$

which is equivalent to Eq. (77). This equation is also equivalent to Eq. (81), because $v_{\text{loss}}(t) = |\partial_t \mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t)| / \sqrt{D_0}$.

C. Optimality in the noise schedule

We discuss the diffusion speed cost $\int_0^\tau dt [v_2(t)]^2$, which is the upper bound on the response function in the speed-accuracy relation for diffusion models [Eq. (79)] in the absence of the nonconservative force. Lowering this upper bound in Eq. (79) means making the response function $(\Delta \mathcal{W}_1)^2 / D_0$ small, and this lowering leads to robust data generation against perturbations of the initial distribution in the estimated reverse process. Since the diffusion speed cost depends only on the dynamics of the forward process, it is possible to discuss the optimality of the noise schedules based on the diffusion speed cost.

We explain why we consider only the upper bound in Eq. (79) and not the upper bound in Eq. (77). If we start with the upper bound in Eq. (77), the lowering of the upper bound in Eq. (77), $\int_0^\tau dt [v_2(t)]^2$, is achieved only if the nonconservative force does not exist due to the inequality $T_t \dot{S}_t^{\text{tot}} \geq [v_2(t)]^2$ in Eq. (73). Thus, we may need only to consider the situation where the nonconservative force is absent and Eq. (77) is identical to Eq. (79). Therefore, the absence of the nonconservative force is a necessary condition for an optimal protocol.

We focus on the case where the nonconservative force is absent to discuss an optimal noise schedule. Similar to the discussion of the minimum entropy production [Eq. (68)], the lower bound of the diffusion speed cost is given by

$$\int_0^\tau dt [v_2(t)]^2 \geq \frac{\mathcal{L}_\tau^2}{\tau} \geq \frac{[\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau)]^2}{\tau}. \quad (92)$$

The equality condition is $v_2(t) = \mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau) / \tau$, which causes the time evolution of \mathcal{P}_t to be driven along the geodesic in the space of the 2-Wasserstein distance. This time evolution can be discussed in terms of the optimal transport protocol discussed in Ref. [109]. At least from the perspective of the speed-accuracy relations for diffusion models, an optimal noise schedule is introduced by considering the optimal transport (OT) schedule, which causes the time evolution of \mathcal{P}_t to be driven along the geodesic in the space of the 2-Wasserstein distance. We illustrate the example of an optimal noise schedule given by the OT schedule in comparison to other noise schedules such as the cosine schedule and the cond-OT schedule (see Fig. 5).

While OT schedule can lead to accurate data generation, it is computationally difficult to construct the diffusion process that achieves optimal transport in high-dimensional data. Therefore, the cond-OT schedule, which can be considered as an approximate optimal transport, is proposed. If the input distribution $q(\mathbf{x})$ is normalized such that $\mathbb{E}_q[\mathbf{x}] = \mathbf{0}$ and $\mathbb{E}_q[||\mathbf{x}||^2] = n_d$ and the external force and the temperature are given by Eqs. (22) and (23), the diffusion speed cost is given asymptotically by the conditional kinetic energy \mathcal{E}_c [69] in the limit $N_D / \sqrt{n_d} \rightarrow 0$:

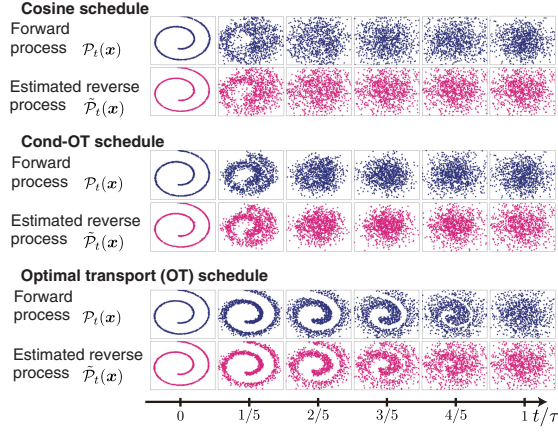


FIG. 5. Examples of the forward process and the estimated reverse process according to three noise schedules: the cosine schedule [Eq. (32)], the cond-OT schedule [Eq. (28)], and the OT schedule which gives the dynamics along the geodesic in the space of the 2-Wasserstein distance. These figures show the time evolution of samples from a two-dimensional Swiss-roll dataset ($n_d = 2$) in the flow-based generative modeling. In all noise schedules, we use the same sample for the initial distribution of the estimated reverse process $\tilde{\mathcal{P}}_\tau(\mathbf{x})$. Compared with the cosine schedule and the cond-OT schedule, the noise in the generated figure is relatively small when we consider the OT schedule. This result is consistent with the optimality based on the speed-accuracy relations for diffusion models.

$$\int_0^\tau dt [v_2(t)]^2 \simeq \tau n_d \mathcal{E}_c, \quad (93)$$

where the dimension of the data n_d is sufficiently large compared to the number of data N_D (Theorem 4.2 in Ref. [69]). Here, the conditional kinetic energy \mathcal{E}_c for the conditional Gaussian probabilities are given by $\mathcal{E}_c = (1/\tau) \int_0^\tau dt [(\partial_t \sigma_t)^2 + (\partial_t m_t)^2]$ [69]. Thus, if the data dimension is sufficiently larger than the number of data, the optimality based on the speed-accuracy relations for diffusion models can be discussed by considering the minimization of this quantity $n_d \int_0^\tau dt [(\partial_t \sigma_t)^2 + (\partial_t m_t)^2]$ instead of the minimization of the diffusion speed cost $\int_0^\tau dt [v_2(t)]^2$.

We explain that minimizing $n_d \int_0^\tau dt [(\partial_t \sigma_t)^2 + (\partial_t m_t)^2]$ leads to the cond-OT schedule [Eq. (28)] and the cosine schedule [Eq. (32)]. When σ_t and m_t are unconstrained, we obtain a lower bound

$$n_d \int_0^\tau dt [(\partial_t \sigma_t)^2 + (\partial_t m_t)^2] \geq n_d \frac{(\sigma_0 - \sigma_\tau)^2 + (m_0 - m_\tau)^2}{\tau}, \quad (94)$$

from the Cauchy-Schwarz inequality. The equality conditions of this inequality are $\partial_t \sigma_t = (\sigma_\tau - \sigma_0)/\tau$ and $\partial_t m_t = (m_\tau - m_0)/\tau$, which can be regarded as the cond-OT schedule [Eq. (28)]. Under the constraint of

VP diffusion, $\sigma_t^2 + m_t^2 = 1$, we can introduce the angle θ_t which satisfies $(m_t, \sigma_t) = (\cos \theta_t, \sin \theta_t)$. From the Cauchy-Schwarz inequality, we also obtain another lower bound:

$$\begin{aligned} n_d \int_0^\tau dt [(\partial_t \sigma_t)^2 + (\partial_t m_t)^2] &= n_d \int_0^\tau dt [(\partial_t \theta_t)^2] \\ &\geq n_d \frac{(\theta_0 - \theta_\tau)^2}{\tau}. \end{aligned} \quad (95)$$

The equality condition of this inequality is $\partial_t \theta_t = (\theta_\tau - \theta_0)/\tau$, which can be regarded as the cosine schedule [Eq. (32)]. The cond-OT schedule or the cosine schedule are considered practical suboptimal protocols in terms of the speed-accuracy relations for diffusion models when the dimension of the data is sufficiently larger than the number of data.

Finally, we mention about the tightness of the bound in Eq. (79), since lowering the diffusion speed cost $\int_0^\tau dt [v_2(t)]^2$ can reduce the response function $(\Delta \mathcal{W}_1)^2/D_0$ if the bound is tight enough. The tightness of the bound in Eq. (79) can be discussed based on the equality condition of the inequalities used in the proof [Eqs. (87), (89), (90), and (91)]. The equality conditions of Eq. (87) are $\nu_t^{\mathcal{P}}(\mathbf{x}) \propto \nabla \psi(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x}) / \mathcal{P}_t(\mathbf{x})$ and $\|\nabla \psi(\mathbf{x})\| = 1$. If we consider $\psi(\mathbf{x}) = \psi_t^*(\mathbf{x})$, the condition $\|\nabla \psi_t^*(\mathbf{x})\| = 1$ may be satisfied almost everywhere, because $\psi_t^*(\mathbf{x})$ is the optimal solution of linear programming with the linear inequality constraint $\|\nabla \psi_t^*(\mathbf{x})\| \leq 1$ [117]. The condition $\nu_t^{\mathcal{P}}(\mathbf{x}) \propto \nabla \psi(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x}) / \mathcal{P}_t(\mathbf{x})$ for $\psi(\mathbf{x}) = \psi_t^*(\mathbf{x})$ is generally not expected to be satisfied. However, the bound is not so loose if $\nu_t^{\mathcal{P}}(\mathbf{x})$ is not orthogonal to $\nabla \psi(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x}) / \mathcal{P}_t(\mathbf{x})$. If $|\partial_t (\mathbb{E}_{\mathcal{P}_t}[\psi_t^*] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi_t^*])|$ has a finite value, the nonorthogonality can be guaranteed and the bound may not be loose. The equality conditions of Eqs. (89) and (90) are $\psi_t^* = \psi_{t-\Delta t}^*$ and $\psi_t^* = \psi_{t+\Delta t}^*$, respectively. Because $\mathcal{P}_t(\mathbf{x}) \simeq \mathcal{P}_{t\pm\Delta t}(\mathbf{x})$ and $\tilde{\mathcal{P}}_t(\mathbf{x}) \simeq \tilde{\mathcal{P}}_{t\pm\Delta t}(\mathbf{x})$ may be satisfied as a consequence of dynamics, $\psi_t^* \simeq \psi_{t\pm\Delta t}^*$ may be expected, and the bound may be tight. The inequalities Eq. (91) become tight when $\partial_t \mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t)$ is constant. While this condition is generally not expected to be satisfied, the numerical result in Sec. IV D supports that this condition may be achievable if we use the OT schedule. In Sec. IV D, we also confirm numerically that both sides of Eq. (79) do not differ by an order of magnitude in simple examples and realistic image generation, and the bound in Eq. (79) can become relatively tight. This fact implies that the discussion in this section may be valid.

D. Numerical experiments

In this section, we first illustrate the validity of our main result [Eq. (79)] in a one-dimensional situation where all variables of Eq. (79) can be computed numerically by density estimation (Sec. IV D 1). Second, we illustrate a two-dimensional situation with a nonconservative force,

where we have to consider the bound in Eq. (77) instead of the bound in Eq. (79). We show the importance of considering the bound given by the entropy production rate in diffusion models (Sec. IV D 2). Finally, we show that our main result [Eq. (79)] provides meaningful bounds in practical situations using the real image dataset and trained flow matching models (Sec. IV D 3).

1. One-dimensional Gaussian mixture distribution

We illustrate the validity of the speed-accuracy relations for diffusion models by simple numerical experiments. We consider the one-dimensional flow-based generative modeling ($\mathbf{x} \in \mathbb{R}$). We now consider the situation where the velocity field of the forward process is accurately reconstructed and $q(\mathbf{x})$ is given by the Gaussian mixture distribution

$$q(\mathbf{x}) = \frac{1}{2}\mathcal{N}(\mathbf{x}|m^a, \sigma^2) + \frac{1}{2}\mathcal{N}(\mathbf{x}|m^b, \sigma^2). \quad (96)$$

Here, $\mathcal{N}(\mathbf{x}|m, \sigma^2)$ is the one-dimensional Gaussian distribution with the mean m and the variance σ^2 . The velocity field $\nu_t^p(\mathbf{x}) (\in \mathbb{R})$ can be calculated analytically as follows (see Appendix G for details of the numerical calculation):

$$\nu_t^p(\mathbf{x}) = \frac{\partial_t \sigma_t}{\sigma_t} \mathbf{x} - \frac{m_t}{\mathcal{P}_t(\mathbf{x})} \left(\frac{\partial_t \sigma_t}{\sigma_t} - \frac{\partial_t m_t}{m_t} \right) \mathbb{E}_{\mathcal{P}_t^c, q}[\mathbf{y}], \quad (97)$$

$$\mathbb{E}_{\mathcal{P}_t^c, q}[\mathbf{y}] = \frac{1}{2} \frac{\sigma^2 m_t \mathbf{x} + \sigma_t^2 m^a \mathbf{1}}{\sigma_t^2 + \sigma^2 m_t^2} \mathcal{N}^a + \frac{1}{2} \frac{\sigma^2 m_t \mathbf{x} + \sigma_t^2 m^b \mathbf{1}}{\sigma_t^2 + \sigma^2 m_t^2} \mathcal{N}^b, \quad (98)$$

where \mathcal{N}^a and \mathcal{N}^b are the Gaussian distributions defined as $\mathcal{N}^x = \mathcal{N}(\mathbf{x}|m_t m^x, \sigma_t^2 + \sigma^2 m_t^2)$ ($x = a, b$). We set the parameters $m^a = -m^b = \sqrt{0.35}$ and $\sigma = 0.3$. We consider various perturbations of the initial condition in the estimated reverse process $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ at $\tilde{t} = 0$ (see also Appendix G for details). We numerically compute the quantities in the speed-accuracy relations for diffusion models [Eqs. (79), (83), and (80)] in Fig. 6.

We discuss the interpretations of the numerical results in Fig. 6. In Fig. 6(a), we show the time evolution of the probability density functions in the forward process $\mathcal{P}_t(\mathbf{x})$ and the time evolution of the probability density functions in the estimated reverse process $\tilde{\mathcal{P}}_t(\mathbf{x})$ with three noise schedules: the cosine schedule, the cond-OT schedule, and the OT schedule. Note that the cosine and cond-OT schedules are not considered the optimal transport because we are considering the situation where $N_D/\sqrt{n_d}$ is not sufficiently small. The initial condition in the estimated reverse process $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ is fixed to the Gaussian distribution $\mathcal{N}(\mathbf{x}|1, 1)$, while the final state in the forward process is given by the Gaussian distribution $\mathcal{N}(\mathbf{x}|0, 1)$.

In Appendix G, we also show the cases with other initial conditions $\mathcal{P}_0^\dagger(\mathbf{x})$ (see also Fig. 9). In Fig. 6(a), the probability density functions in the forward process with the cosine schedule and the cond-OT schedule are significantly changed between time $t = 0$ and $t = (1/3)\tau$ compared to the OT schedule. Thus, the data structure corresponding to the two peaks of the probability density function is not well recovered in the estimated reverse process for the cosine and cond-OT schedules between $t = (1/3)\tau$ and $t = \tau$, while the two peaks can be seen between $t = (1/3)\tau$ and $t = (2/3)\tau$ in the estimated reverse process for the OT schedule.

Next, we numerically confirm the speed-accuracy relations for diffusion models [Eqs. (79) and (83)] and the detailed inequality [Eq. (80)] hold under the various conditions of $\mathcal{P}_0^\dagger(\mathbf{x})$ in Figs. 6(b)–6(e): (b) a Gaussian distribution with a different mean, (c) a Gaussian distribution with a different variance, (d) a Gaussian mixture distribution, and (e) a uniform distribution. Here, we use the notation $\eta \leq c_1 \leq c_2$ for Eq. (83), where $\eta = (\Delta \mathcal{W}_1)^2/D_0$, $c_1 = \int_0^\tau dt [v_{\text{loss}}(t)]^2$, and $c_2 = \int_0^\tau dt [v_2(t)]^2$. We confirm that the speed-accuracy relations for diffusion models are valid under all conditions. We also observe that the bounds in Eq. (80) are tighter in the case of OT than the bounds in the cases of the cosine and cond-OT schedules. There is a tendency for the bound in the detailed inequality [Eq. (80)] to become loose during the time evolution for the cosine schedule and the bound in the detailed inequality [Eq. (80)] to become loose at the beginning and at the end of the time evolution for the cond-OT schedule. The upper bounds of the response function η , namely, c_1 and c_2 , are tighter for the OT schedule compared to the cosine and cond-OT schedules. The value of the response function η for the OT schedule is also the smallest for any schedule (see Table I), which supports our conclusion that the OT schedule provides the most accurate data generation. We also confirm that $[v_{\text{loss}}(t)]^2$ is constant, and, thus, $|\partial_t \mathcal{W}_1(\mathcal{P}_t, \tilde{\mathcal{P}}_t)|$ is also constant in the case of OT schedule. Moreover, our bounds become relatively tight in the case of OT schedule for conditions (b) and (c).

Interestingly, although the values of the upper bounds c_1 and c_2 for the OT schedule are so small compared to the cosine and cond-OT schedules, the value of the response function η for the OT schedule is not so small compared to the cosine and cond-OT schedules. In fact, the probability density functions corresponding to the generated data $p(\mathbf{x})$ in Fig. 6(a) are not significantly different under different conditions. The reason for this fact may be that we are considering only a simple data structure of one-dimensional Gaussian mixtures, so larger differences in η between different noise schedules may occur when the data structure is more complex. Reducing the time duration τ and rapidly changing the time evolution in the forward process may also cause large η differences between different noise schedules.

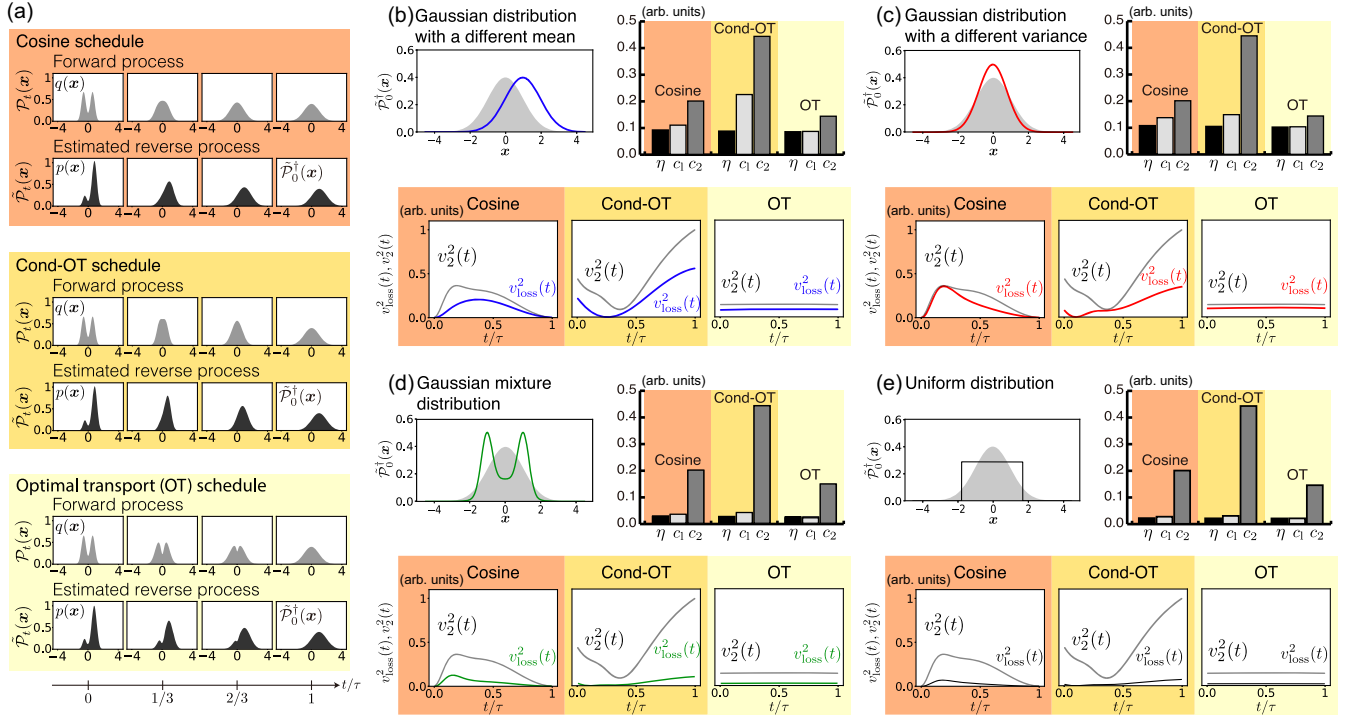


FIG. 6. Numerical calculations of the speed-accuracy relations for diffusion models [Eqs. (79), (80), and (83)]. (a) The time evolution of the probability density functions in the forward process and the estimated reverse process with the cosine schedule, the cond-OT schedule, and the optimal transport (OT) schedule. Here, t denotes time in the forward direction, and we show the probability density functions at $t = 0$, $t = \tau/3$, $t = 2\tau/3$, and $t = \tau$. The final state of the forward process is $\mathcal{P}_\tau(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, 1)$. We use a Gaussian distribution with a different mean $\tilde{\mathcal{P}}_0^+(\mathbf{x}) = \mathcal{N}(\mathbf{x}|1, 1)$ as the initial condition of the estimated reverse process. (b)–(e) We show the situation where the initial condition of the estimated reverse process $\tilde{\mathcal{P}}_0^+(\mathbf{x})$ is (b) a Gaussian distribution with a different mean, (c) a Gaussian distribution with a different variance, (d) a Gaussian mixture distribution, and (e) a uniform distribution. The plot at the top left in each section shows the comparison between $\tilde{\mathcal{P}}_0^+(\mathbf{x})$ and $\mathcal{P}_\tau(\mathbf{x}) = \mathcal{N}(\mathbf{x}|0, 1)$, where $\mathcal{N}(\mathbf{x}|0, 1)$ is shown in gray. In the top right-hand section, we show the bar graph which displays the quantities $\eta = (\Delta\mathcal{W}_1)^2/(\tau D_0)$, $c_1 = \int_0^\tau dt [v_{\text{loss}}(t)]^2$, and $c_2 = \int_0^\tau dt [v_2(t)]^2$, and this figure shows the validity of the inequality $\eta \leq c_1 \leq c_2$ [Eq. (83)]. In the bottom row, we show the time evolution of $[v_2(t)]^2$ and $[v_{\text{loss}}(t)]^2$, and this figure shows the validity of the inequality $[v_2(t)]^2 \geq [v_{\text{loss}}(t)]^2$ [Eq. (80)]. Since $[v_2(t)]^2$ and c_2 depend only on the forward process, these values are consistent within different initial conditions in (b)–(e). See also Table III in Appendix G for more detailed explanations, including the probability density function $\tilde{\mathcal{P}}_0^+(\mathbf{x})$ used in (b)–(e).

TABLE I. The values of the response function $\eta = (\Delta\mathcal{W}_1)^2/(\tau D_0)$ in the numerical calculations (in Fig. 6). Since we compute 3×10^8 samples, we display the value with four significant digits.

Initial conditions	Noise schedules	Values of η
Gaussian distribution with a different mean	Cosine	9.188×10^{-2}
	Cond-OT	8.781×10^{-2}
	OT	8.537×10^{-2}
Gaussian distribution with a different variance	Cosine	1.078×10^{-1}
	Cond-OT	1.047×10^{-1}
	OT	1.022×10^{-1}
Gaussian mixture distribution	Cosine	2.764×10^{-2}
	Cond-OT	2.674×10^{-2}
	OT	2.611×10^{-2}
Uniform distribution	Cosine	2.142×10^{-2}
	Cond-OT	2.068×10^{-2}
	OT	2.033×10^{-2}

Our discussion based on the speed-accuracy relations for diffusion models may not be sufficient to explain why the cosine and cond-OT schedules work well even when $N/\sqrt{n_d}$ is not sufficiently small. Since the value of η is indeed approximately equal to the value of c_1 for the OT schedule, the value of η itself can be considered to be determined by the speed-accuracy relations for diffusion models. To understand why the response function η for the OT schedule is not so small compared to the cosine and cond-OT schedules, it may be necessary to consider not only the upper bound on η derived in this paper, but also the lower bound on η .

2. Two-dimensional Swiss-roll dataset

Second, we illustrate a situation where the nonconservative force is present and we need to consider the bound given by the entropy production rate \dot{S}_t^{tot} [Eq. (77)] instead of the bound given by $v_2(t)$ [Eq. (79)]. Specifically, we consider a two-dimensional Swiss-roll dataset and use a noise schedule with a nonconservative external force as

$$\begin{aligned} \mathbf{f}_t(\mathbf{x}) &= \begin{pmatrix} H_t & -G_t \\ G_t & H_t \end{pmatrix} \mathbf{x}, \\ H_t &= -h \left(1 - \frac{t}{\tau} \right) - T_t, \\ G_t &= g \left(1 - \frac{t}{\tau} \right), \\ T_t &= T_{\text{fin}} \left(\frac{t}{\tau} \right)^2, \end{aligned} \quad (99)$$

where $h \in \mathbb{R}_{>0}$, $T_{\text{fin}} \in \mathbb{R}_{>0}$, and $g \in \mathbb{R}_{\geq 0}$ are parameters to control the noise schedule. In the case of $g = 0$, this external force $\mathbf{f}_t(\mathbf{x})$ is conservative. In the case of $g > 0$, this external force $\mathbf{f}_t(\mathbf{x})$ is nonconservative. Since it is difficult to derive a closed form of the velocity field in this situation, we adopt a denoising score matching method with sufficiently large artificial neural networks and simulate the forward and estimated reverse processes using the estimated score function to evaluate Eq. (77). We set the parameters $h = 1.1$ and $T_{\text{fin}} = 0.5$ and change the value of g in $\{0, 3, 5\}$. We set the initial condition of the estimated reverse process $\tilde{\mathcal{P}}_\tau(\mathbf{x})$ as the Gaussian distribution with different variance $\mathcal{N}(\mathbf{0}, (0.9)^2 \mathbf{I})$, because the probability distribution $\mathcal{P}_\tau(\mathbf{x})$ becomes the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in the limit $\tau \rightarrow \infty$ (see also Appendix G). For comparison, we also numerically estimate the optimal transport noise schedule using a method introduced in Ref. [72] (see also Appendix G). We numerically compute the quantities in the speed-accuracy relations for diffusion models [Eqs. (77) and (80)] in Fig. 7.

Through the illustration in Fig. 7, we can confirm the validity of minimizing the upper bound of our results

[Eqs. (77), (80), and (81)] against improving the robustness of the diffusion model. In Figs. 7(a)–7(d), we can see that the generated data $\tilde{\mathcal{P}}_0(\mathbf{x}) = p(\mathbf{x})$ become significantly different from the training data $\mathcal{P}_0(\mathbf{x}) = q(\mathbf{x})$ in the case of $g = 0, 3, 5$ [Figs. 7(a)–7(c)] compared to the OT schedule [Fig. 7(d)]. We confirm that the bounds [Eqs. (77), (80), and (81)] are valid for each noise schedule in Figs. 7(e) and 7(f). We also confirm that minimizing the thermodynamic cost $w_D = \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$ qualitatively improves the robustness of the generation in Fig. 7(f). The values of η in $g = 0, 3, 5$ differ by orders of magnitude compared to other noise schedules, while the value of η in the OT schedule remains in an acceptable range. We confirm that minimizing the upper bound w_D in Eq. (77) and considering the OT schedule are effective in ensuring that the generated data do not differ from the training data.

We discuss the effects of nonconservative forces and thermodynamic dissipation, which could not be dealt with in the one-dimensional Gaussian mixture distribution. The numerical results also show that the strength of the nonconservative forces g affects the sensitivity of the data generation. The sensitivity $(\Delta \mathcal{W}_1)^2 / (\tau D_0)$ increases with increasing g in the region up to $g = 5$. This trend is similar for the upper bound w_D , and it can be said that the sensitivity of data generation in the diffusion model is higher in situations of strong nonconservative forces, where more thermodynamic dissipation occurs. Nonconservative forces have not been considered in conventional diffusion models using cond-OT and cosine schedules. It can be assumed that nonconservative forces have not been considered to ensure that data generation is robust. On the other hand, if one wants to deliberately create a generated result that is different from the training data, it is necessary to increase the sensitivity $(\Delta \mathcal{W}_1)^2 / (\tau D_0)$. Therefore, one could consider a situation where such nonconservative forces are included in diffusion models and thermodynamic dissipation is dared to occur in the forward processes.

In addition, Figs. 7(a) and 7(d) show a situation where OT schedule provides superior generation results even in a conservative situation as discussed in Fig. 6. As mentioned earlier, the noise schedule in Eq. (99) with $g = 0$ provides a conservative external force $\mathbf{f}_t(\mathbf{x})$, and in this case $[v_2(t)]^2 = T_t \dot{S}_t^{\text{tot}}$, $w_D = \int_0^\tau dt [v_2(t)]^2$ holds. From this perspective, the comparison between the schedule of $g = 0$ and the OT schedule can be seen as a confirmation of Eqs. (79) and (83), similar to the situation we dealt with in Fig. 6. Looking at Fig. 7(f), we can see that the $g = 0$ schedule has a higher η than the OT schedule by a sufficient margin (Table II) compared to the cosine, cond-OT, and OT schedules in the previous setup of a Gaussian mixture in one dimension (Table I). This is because the schedule of $g = 0$ is far from the optimal schedule, while the cosine and cond-OT schedules are suboptimal as discussed in Sec. IV C.

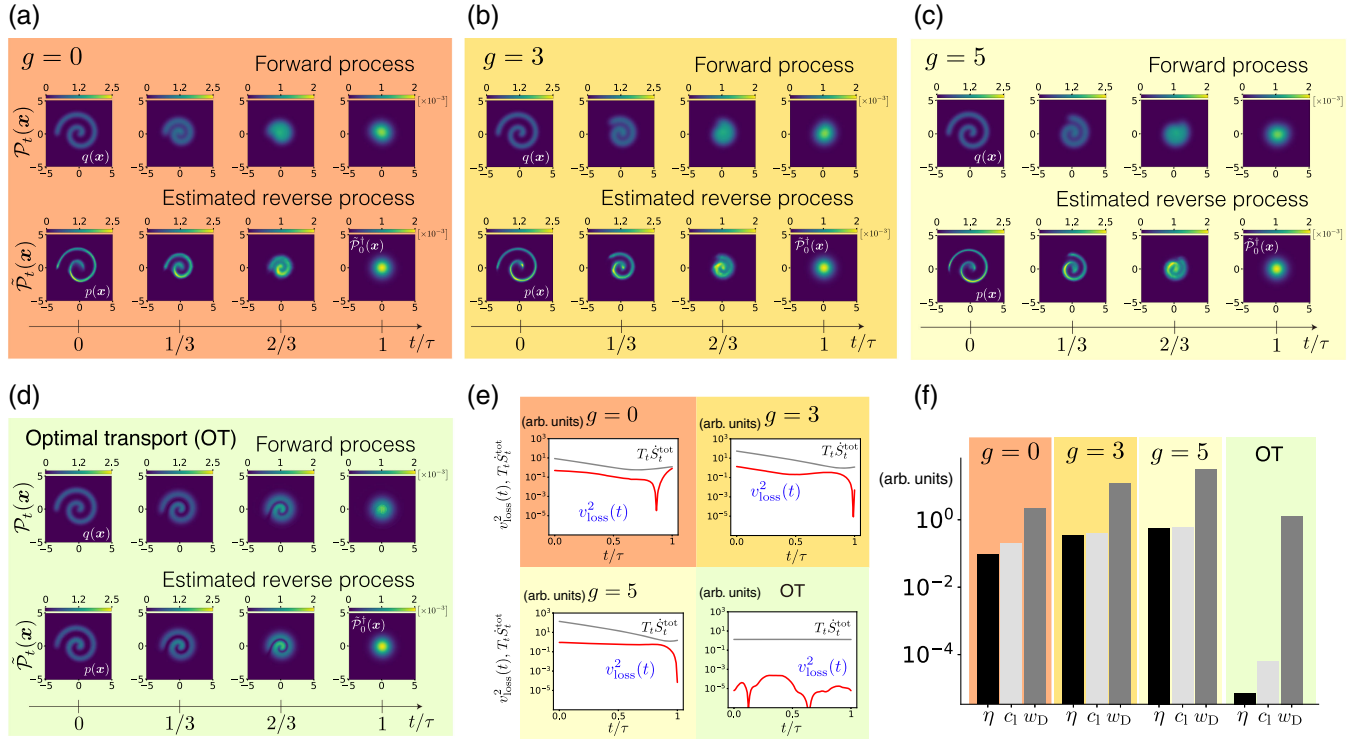


FIG. 7. Numerical experiments of the speed-accuracy relations for diffusion models [Eqs. (77), (80), and (81)] in two-dimensional Swiss-roll situation. (a)–(c) The time evolution of the probability density function in the forward process and the estimated reverse process with the noise schedules using $g = 0, 3, 5$ in Eq. (99). The value g stands for the degree of nonconservativeness, and $g = 0$ denotes a situation with conservative external force $f_t(x)$. (d) The time evolution of the probability density function with OT schedule. In (a)–(d), t denotes time in the forward direction, and we show the probability density functions at $t = 0, t = \tau/3, t = 2\tau/3$, and $t = \tau$. Each axis of the color map denotes the axis of the data space \mathbb{R}^{n_d} . Throughout this experiment, we use a Gaussian distribution with different variance $\hat{P}_0^\dagger(x) = \mathcal{N}(\mathbf{0}, (0.9)^2 \mathbf{I})$ as the initial condition of the estimated reverse process. (e) Temporal evolution of $T_t \dot{S}_t^{\text{tot}}$ and $[v_{\text{loss}}(t)]^2$ proving the validity of $[v_{\text{loss}}(t)]^2 \leq T_t \dot{S}_t^{\text{tot}}$ [Eq. (80)]. (f) A bar graph showing the quantities $\eta = (\Delta \mathcal{W}_1)^2 / (\tau D_0)$, $c_1 = \int_0^\tau dt [v_{\text{loss}}(t)]^2$, and $w_D = \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$. This figure shows the relationship $\eta \leq c_1 \leq w_D$ [Eq. (81)]. For detailed explanations, see Appendix G.

3. Real-world image datasets

We show that the analysis using our main result [Eq. (79)] is feasible in a more realistic situation for the application of diffusion models using the real-world image datasets. We estimate the values of Eq. (79) on the latent space using trained models and benchmark image datasets. Here, we use a trained flow matching model in Ref. [118], which creates a diffusion process on a latent space

TABLE II. The values of the response function $\eta = (\Delta \mathcal{W}_1)^2 / (\tau D_0)$ in the numerical calculations (Fig. 7). Since we compute 1×10^8 samples, we display the value with four significant digits.

Initial conditions	Noise schedules	Values of η
Gaussian distribution with a different variance	$g = 0$	9.301×10^{-2}
	$g = 3$	3.430×10^{-1}
	$g = 5$	5.615×10^{-1}
	OT	7.221×10^{-6}

[44,118]. For the datasets, we use a 256×256 downsampled version of the CelebA HQ dataset [119] and the bedroom subset of the LSUN dataset [120].

In Fig. 8, we have illustrated diffusion models for the real-world image datasets, and the validity of Eq. (79) on the latent space is shown in Fig. 8 (see also the details in Appendix G). These trained models were trained with the cond-OT schedule [Eq. (28)] on the latent space, and we use $\mathcal{N}(x|\mathbf{0}, \mathbf{I})$ for the initial distribution of the estimated reverse process on the latent space. In Figs. 8(a) and 8(b), we have illustrated the samples from the forward process and the estimated reverse process. In this setting, the dimension of the latent space is $n_d = 4096$, and it may be difficult to estimate the time evolution of the probability density functions in the forward and estimated reverse processes with a finite number of samples. However, we may estimate the quantities in our main result [Eq. (79)], as shown in Fig. 8(c), because we need the velocity field in the forward process and the initial and final probability distributions in the forward and estimated reverse processes to estimate the quantities. We find that our result [Eq. (79)] is

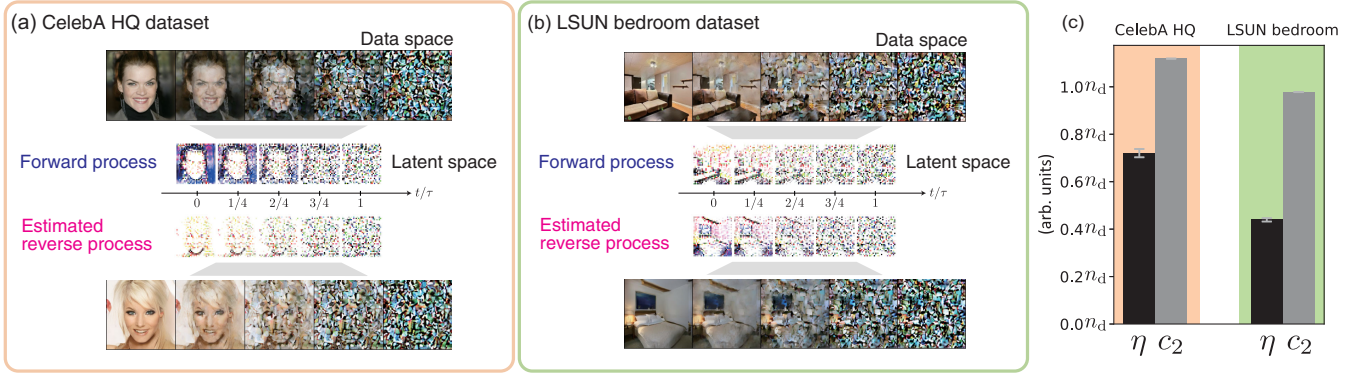


FIG. 8. The application of diffusion models using the real-world image datasets and the validity of Eq. (79) on the latent space. (a) The forward process and the estimated reverse process for the downsampled version of the CelebA HQ dataset [119]. (b) The forward process and the estimated reverse process for the bedrooms subset of the LSUN dataset [120]. (c) The estimated values of $\eta = (\Delta\mathcal{W}_1)^2/(\tau D_0)$ and $c_2 = \int_0^\tau dt [v_2(t)]^2$ computed on the latent space. The error bars show the standard deviation among five calculations using 3×10^4 samples for each calculation. $n_d = 4096$ denotes the dimension of the latent space. The inequality $c_2 \geq \eta$ [Eq. (79)] on the latent space is still valid for realistic applications of diffusion models using the real-world image datasets.

still valid. Moreover, the response function $(\Delta\mathcal{W}_1)^2/(\tau D_0)$ has the same order of magnitude as the upper bound $\int_0^\tau dt [v_2(t)]^2$. Thus, the numerical analysis based on Eq. (79) may be reasonable even if we discuss diffusion models on the latent space for the real-world image datasets with data of large dimension $n_d \gg 1$.

We explain that our treatment based on the response function may be reasonable in realistic situations such as this experiment. In this experiment, the starting point of the estimated reverse process $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ is chosen to be similar to the final state of the forward process $\mathcal{P}_\tau(\mathbf{x})$, so that the estimated reverse process will be close to the genuine reverse process. One of the typical choices of $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ is the standard Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$, since $\mathcal{P}_\tau(\mathbf{x})$ will converge to this distribution for sufficiently large τ in the formulation of Eqs. (22) and (23). However, the discrepancy between $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ and $\mathcal{P}_\tau(\mathbf{x})$ usually exists, since $\mathcal{P}_\tau(\mathbf{x}) = \int d\mathbf{y} \mathcal{P}_\tau^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})$ [Eq. (17)] is not generally equal to $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$ with a finite τ . Because of the large dimension of the data, it is also difficult to estimate $\mathcal{P}_\tau(\mathbf{x})$ from the finite number of samples. Therefore, the difference between $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ and $\mathcal{P}_\tau(\mathbf{x})$ may be unavoidable, and our main result based on the response function [Eq. (79)] may be useful to consider the effect of the difference between $\tilde{\mathcal{P}}_0^\dagger(\mathbf{x})$ and $\mathcal{P}_\tau(\mathbf{x})$.

V. DISCUSSIONS

In this paper, we summarize the relationship between diffusion models and nonequilibrium thermodynamics and derive the speed-accuracy relations for diffusion models based on techniques from stochastic thermodynamics and optimal transport theory. The speed-accuracy relations for diffusion models explain that the robustness of data generation to perturbations is generally limited by the

thermodynamic dissipation in the forward diffusion process. The speed-accuracy relations for diffusion models also quantitatively explain the validity of using the cosine and cond-OT schedules as the noise schedules in the diffusion model and the importance of using optimal transport in the forward diffusion process. We also derive the speed-accuracy relations for diffusion models as analogs of the thermodynamic trade-off relations such as the thermodynamic uncertainty relations [110,114,121] and the thermodynamic speed limits [29,30]. Unlike the conventional thermodynamic trade-off relations, which consider thermodynamic limits on the speed of the observable in a stochastic process, the speed-accuracy relations for diffusion models are conceptually different, because they give limits on how the difference between two different processes—the forward process and the estimated reverse process—changes. Our conclusion for the noise schedule is also justified in light of recent developments in the diffusion model. In our conclusion, the optimal forward process for the data generation should be given by the geodesic in the space of the 2-Wasserstein distance. In recent years, several methods have been proposed to realize the dynamics along the geodesic approximately in the flow-based generative modeling and the probability flow ODE [68–70,72,122], which are becoming the mainstream methods.

Below, we further discuss the generalization of the main result for the incomplete estimation of the velocity field, a comparison with other similar bounds, a comparison with the conventional understanding of noise schedules, the importance of the entropy production rate in diffusion models, and possible future studies.

A. Generalization of Eq. (77) for incomplete estimation

The main result is due to the error-free estimation of the velocity field in the forward process and its use in the

estimated reverse process. If the velocity field estimation is incomplete, then corrections to the speed-accuracy relations for the diffusion model would have to be made due to the incompleteness.

A possible generalization of our main result [Eq. (77)] for incomplete estimation can be obtained as

$$\frac{|\Delta\mathcal{W}_1|}{\sqrt{\tau D_{\max}}} \leq \sqrt{\int_0^\tau dt T_t \dot{S}_t^{\text{tot}}} + \sqrt{1 + \frac{1}{D_{\max}}} \sqrt{\tau L_\alpha(\theta_\alpha^*)}, \quad (100)$$

where $L_\alpha(\theta_\alpha^*)$ ($\alpha \in \{\text{SM}, \text{FM}\}$) is the objective function after training defined in Eqs. (5) and (6) and Eqs. (12) and (13) and $D_{\max} := \max_{t \in [0, \tau]} D_t$ is the maximum value of the χ^2 divergence. The derivation of this bound is given in Appendix F. We can see that Eq. (100) becomes Eq. (77) when the objective function is zero [$L_\alpha(\theta_\alpha^*) = 0$], since $D_{\max} = D_0$ holds in the error-free estimation of the velocity field (see Appendix E).

If the estimation of the velocity field is satisfactorily achieved and $L_\alpha(\theta_\alpha^*) \ll (1 + 1/D_{\max})^{-1} \tau^{-1} \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$ holds, we obtain approximately a similar form of our main result [Eq. (77)] by replacing D_0 with D_{\max} , and our understanding based on Eq. (77) may be useful even for the incomplete estimation of the velocity field. This is due to the fact that the second term of Eq. (100) may be negligible if $L_\alpha(\theta_\alpha^*) \ll (1 + 1/D_{\max})^{-1} \tau^{-1} \int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$. Thus, noise schedules close to the optimal transport may be a better option to use in this case.

If the estimation of the velocity field is not well achieved, the second term $\sqrt{(1 + 1/D_{\max}) \tau L_\alpha(\theta_\alpha^*)}$ is not negligible. If we want to reduce the left-hand side $|\Delta\mathcal{W}_1|/\sqrt{\tau D_{\max}}$ by minimizing the upper bound, the noise schedule in the forward process corresponding to $\sqrt{\int_0^\tau dt T_t \dot{S}_t^{\text{tot}}}$ and the minimization of the objective function corresponding to $\sqrt{(1 + 1/D_{\max}) \tau L_\alpha(\theta_\alpha^*)}$ are both important. Compared to noise schedules that are close to the OT schedule, we may rather use noise schedules which are advantageous for training efficiency at lower values of the objective function. In fact, it is known that one of the viable options for an approximate OT schedule [70] may not achieve the minimum value of the estimation error in a particular situation [123]. Therefore, the OT schedule may not be the best choice for a particular situation where the objective function is not negligible.

B. Comparison with bounds similar to Eq. (77)

We compare our main result [Eq. (77)] with other similar known bounds. As we mentioned in Sec. IV, our bound is inspired by the thermodynamic uncertainty relation and the final form [Eq. (77)] resembles the thermodynamic uncertainty relation. However, our result cannot be derived from the thermodynamic uncertainty relation by changing the observable [Eq. (73)]. For example, a derivation of the

thermodynamic uncertainty relation in Ref. [110] is based on the Cauchy-Schwarz inequality

$$\begin{aligned} & \left(\int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \cdot \nabla r(\mathbf{x}) \right)^2 \\ & \leq \left(\int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \|\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2 \right) \left(\int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \|\nabla r(\mathbf{x})\|^2 \right), \end{aligned} \quad (101)$$

where $r(\mathbf{x})$ denotes an arbitrary time-independent scalar observable [see also Eq. (73)]. On the other hand, the derivation of our main result [Eq. (77)] is based on a different Cauchy-Schwarz inequality:

$$\begin{aligned} & \left(\int d\mathbf{x} \delta \mathcal{P}_t(\mathbf{x}) \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \cdot \nabla \psi(\mathbf{x}) \right)^2 \\ & \leq \left(\int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \|\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2 \right) \left(\int d\mathbf{x} \frac{(\delta \mathcal{P}_t(\mathbf{x}))^2 \|\nabla \psi(\mathbf{x})\|^2}{\mathcal{P}_t(\mathbf{x})} \right) \end{aligned} \quad (102)$$

and the 1-Lipschitz continuity $\|\nabla \psi\| \leq 1$. Therefore, our bound uses the Cauchy-Schwarz inequality in a different way. In addition, our bound requires the discussion of the Kantorovich potential, which is used in Eqs. (89) and (90). Therefore, our result is not regarded as a simple application of the thermodynamic uncertainty relation.

We also compare our main results [Eqs. (77) and (83)] with the trivial inequality in optimal transport theory. We have a trivial relation between the 1-Wasserstein distance and the 2-Wasserstein distance $\mathcal{W}_1(P, Q) \leq \mathcal{W}_2(P, Q)$ for arbitrary probabilities P and Q , and this trivial relation resembles our result [Eq. (77)]. However, our result is the nontrivial relation between the 1-Wasserstein distance and 2-Wasserstein distance, because our result compares the time derivative of the 1-Wasserstein distance $\partial_t \mathcal{W}_1(\mathcal{P}_{\tau-t}^\dagger, \tilde{\mathcal{P}}_{\tau-t}^\dagger)$ for the two processes $\mathcal{P}_{\tau-t}^\dagger(\mathbf{x})$ and $\tilde{\mathcal{P}}_{\tau-t}^\dagger(\mathbf{x})$ with the speed in the space of the 2-Wasserstein distance $v_2(t) = \lim_{\Delta t \rightarrow 0} \mathcal{W}_2(\mathcal{P}_t, \mathcal{P}_{t+\Delta t})/\Delta t$ for the forward process $\mathcal{P}_t(\mathbf{x})$. This inequality is not a straightforward consequence of the trivial relation between the 1-Wasserstein distance and the 2-Wasserstein distance. We also note that the trivial inequality $\mathcal{W}_1(\mathcal{P}_t, \mathcal{P}_{t+\Delta t}) \leq \mathcal{W}_2(\mathcal{P}_t, \mathcal{P}_{t+\Delta t})$ leads to the thermodynamic uncertainty relation (see Sec. III B), and our result [Eq. (77)] differs from the thermodynamic uncertainty relation as discussed above.

In the study of machine learning, there are some studies on a similar problem that derive some bounds in a similar form [65, 66, 123]. However, these bounds are conceptually different from ours. In Eq. (23) in Ref. [65], the authors consider a setup with perturbation on the initial training data and evaluate the effect of the perturbation on the generated data. The authors used the contraction property of the 2-Wasserstein distance, which is

$$\mathcal{W}_2(\mathcal{P}_t, \mathcal{P}_t^p) \leq \exp\left(\int_0^t dt' K_L(t')\right) \mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_0^p), \quad (103)$$

where $K_L(t)$ is the Lipschitz constant of the external force $f_t(\mathbf{x})$ satisfying $\|f_t(\mathbf{x}) - f_t(\mathbf{y})\| \leq K_L(t)\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x} \in \mathbb{R}^{n_d}$ and $\mathbf{y} \in \mathbb{R}^{n_d}$, and $\mathcal{P}_t^p(\mathbf{x})$ is the density function of the forward process with perturbation in the initial data distribution $\mathcal{P}_0(\mathbf{x})$. The author also discussed the upper bound on the 2-Wasserstein distance $\mathcal{W}_2(\tilde{\mathcal{P}}_0, \tilde{\mathcal{P}}_0^p)$ based on the triangle inequality, where $\tilde{\mathcal{P}}_t^p(\mathbf{x})$ is the density function of the estimated reverse process with perturbation in the initial data distribution $\mathcal{P}_0(\mathbf{x})$. This bound differs from ours, because the paper did not focus on the perturbation in the estimated reverse process and the time evolution of the forward process. Indeed, the result in Ref. [65] does not provide insight into the optimality of the noise schedule in the forward process. In Refs. [66,123], they consider the bound on the estimation error when the initial perturbation is negligible $D_{\max} \simeq 0$ and the objective function is not negligible $L_\alpha(\theta_\alpha^*) \neq 0$. In Ref. [66], the authors considered score-based generative modeling and focused on the specific situation where the forward process is given by the Ornstein-Uhlenbeck process and the objective function is optimized with respect to specific parameters. The result is, therefore, not applicable to the general case where, for example, optimal transport is achieved. In Ref. [123], the authors considered flow-based generative modeling and obtained the similar bound obtained in Ref. [66], and, thus, the result is not applicable to the general case either. These analyses are examples of the convergence analysis of generative models [65–67,124–126], which provide bounds on the estimation errors with respect to the objective function. Our generalized result [Eq. (100)] can be considered as the convergence analysis. Compared with the results in Refs. [66,123], our results are generally applicable to diffusion models where the forward process is arbitrarily chosen and the objective function is not well optimized. In fact, optimal transport is generally not achieved with the Ornstein-Uhlenbeck process. Thus, our general inequality allows us to discuss an optimality of data generation via optimal transport, while the results in Refs. [66,123] do not. This fact may be a possible merit of our result based on the universal thermodynamic framework.

Finally, we discuss the possibility of generalizing our result based on other metrics of estimation error instead of the 1-Wasserstein distance. In practical diffusion models, metrics such as the FID [82] and the inception score (IS) [127] are widely used as an index of generation quality. Adapting our methodology to these metrics may not be straightforward, and some other ideas may be needed. First, both metrics are heavily based on the pretrained inception model [107], which is a black-box classification model on images. Therefore, it may be difficult to provide general bounds on these metrics without adding specific assumptions to this black-box model. Although it may be difficult

to apply our methodology, the FID may be more appropriate than the IS for considering a similar upper bound. As discussed in Sec. III B, the FID is a lower bound on the 2-Wasserstein distance $\mathcal{W}_2(p, q)$, and it may be possible to evaluate it by providing an upper bound on $\mathcal{W}_2(p, q)$. Since our bounds are on the 1-Wasserstein distance $\mathcal{W}_1(p, q)$ and the trivial relation $\mathcal{W}_1(p, q) \leq \mathcal{W}_2(p, q)$ does not lead to an upper bound on the FID, a simple modification of our proof may not give an upper bound on the FID. Thus, another idea such as the triangle inequality for the 2-Wasserstein distance may be needed to consider a similar thermodynamic bound on the FID.

C. Comparison with conventional understanding of noise schedules

We discuss our result with the conventional understanding on noise schedules in diffusion models. In studies of diffusion models, there have been many numerical attempts to investigate the effect of noise schedules on the quality of data generation [43,63,70,72,73,128]. For example, the cosine schedule has been shown experimentally to be an efficient protocol for image generation tasks compared to other conventional noise schedules under the $m_t^2 + \sigma_t^2 = 1$ constraint [43], and optimal transport-based flow-matching methods have also been shown experimentally to provide better sampling quality and efficiency [70,72,73,128]. However, there are almost no theoretical studies that explain their efficiency in terms of the universal upper bound on the estimation error. For example, Ref. [70] introduced a method for using an approximation of optimal transport for noise schedules in flow matching, and the authors may not provide any reasons for using optimal transport as noise schedules except for numerical illustration of the generated quality and efficient sampling.

The speed-accuracy relations for diffusion models derived in this study may provide insight into how to determine this noise schedule for accurate data generation. In fact, we have shown that the existing methods, such as the cosine and cond-OT schedules, are suboptimal and that the genuine optimal transport should be the best noise schedules from the perspective of the speed-accuracy relations for diffusion models. Thus, the room for improvement in existing methods can be discussed quantitatively based on the tightness of the inequalities. Our treatment would be a new direction in the theoretical understanding of the optimality of noise schedules for diffusion models.

D. Possible importance of the entropy production rate in diffusion models

We discuss the possible importance of the entropy production rate in diffusion models by comparing our approach with other physically inspired machine learning methods, so-called energy-based methods. In our main

result [Eq. (77)], we use the entropy production rate for the upper bound on the estimation error. As an analogy to the entropy production minimization problem, the reduction of the thermodynamic quantity, $\int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$ for the forward dynamics, is related to a smallness of the response function $(\Delta\mathcal{W}_1)^2/D_0$ against arbitrary initial perturbations.

In machine learning, there are energy-based methods that minimize physical quantities, such as the energy and the free energy, similar to the entropy production. For example, Amari-Hopfield networks [129,130] modeled the brain's memory function using a weighted graph called a neural network, which formed the basis of current machine-learning methods using artificial neural networks. Amari-Hopfield networks are physically viewed as spin-glass systems, and the association and retrieval of given memories is characterized by the energy function in spin-glass systems. Mathematically, the energy function in Amari-Hopfield networks is just the objective function, and minimizing the energy function is understood in terms of optimizing the objective function. Boltzmann machines [131–133] are machine-learning models deeply inspired by Amari-Hopfield networks, which formed the basis of modern machine-learning models. Boltzmann machines treat training data as memories in Amari-Hopfield networks and slightly modify the framework to treat it in a probabilistic way. Specifically, using the same neural network and the energy function as Amari-Hopfield networks, the data in Boltzmann machines are expressed as samples of the Gibbs measure with respect to the energy function. In the training process of Boltzmann machines, we minimize the Kullback-Leibler divergence between the training dataset distribution and the Gibbs measure by changing the weights of the neural network. In this way, Boltzmann machines estimate the probability distribution behind the input data, and Boltzmann machines can be seen as a variant of generative models. In stochastic thermodynamics, the Kullback-Leibler divergence between the current probability and the Gibbs measure is the Lyapunov function for the system with the detailed balance condition and is considered as a nonequilibrium generalization of the free energy [134,135], and, thus, this learning is also translated in terms of the free energy minimization.

From the perspective of stochastic thermodynamics, the application of the energy function and the free energy is limited to the stochastic process that satisfies the detailed balance condition. For example, it is often the case that the nonconservative force exists in Langevin systems or the Fokker-Planck systems, and the detailed balance condition is violated. In this case, the energy may not be physically defined, but the entropy production rate is still well defined. Moreover, if the detailed balance condition holds, the entropy production rate is rewritten in terms of the potential energy or the free energy by the first law of thermodynamics [2]. Thus, the entropy production rate can be a more general thermodynamic quantity than the energy for

stochastic processes such as Langevin dynamics and Fokker-Planck dynamics.

The entropy production rate in diffusion models may be discussed in parallel with the energy and the free energy in Amari-Hopfield networks and Boltzmann machines. Amari-Hopfield networks and Boltzmann machines may be based on systems where the energy function can be defined, and it means that the detailed balance condition may hold. It makes sense that the energy function would be useful in such a system. Since diffusion models are based on the Langevin dynamics or the Fokker-Planck dynamics, our framework based on the entropy production rate may make sense because the entropy production rate is more informative rather than the energy function, especially when the detailed balance condition is violated by the nonconservative force. It is surprising that the idea of minimizing a physically inspired cost $\int_0^\tau dt T_t \dot{S}_t^{\text{tot}}$ is one of the criteria for model selection in diffusion models as in the traditional energy-based methods such as Amari-Hopfield networks and Boltzmann machines. Interestingly, our proposed optimal protocol is related to the minimization of the entropy production when the temperature of the system does not depend on time, analogous to the minimization of the energy and the minimization of the free energy. In fact, the entropy production can also be related to the potential energy $U_t(\mathbf{x})$ when the nonconservative force is absent, i.e., $\mathbf{f}_t(\mathbf{x}) = -\nabla U_t(\mathbf{x})$. Thus, it may be interesting to consider the connection between our framework and the energy-based treatment, where the detailed equilibrium condition holds and the potential energy is defined. In Ref. [136], the symmetry breaking of the phase transition is discussed in diffusion models based on the energy-based spin-glass theory. Such an energy-based treatment of the diffusion model based on statistical physics may also be promising as a complementary approach.

E. Future studies

In this paper, we have demonstrated only one aspect of the usefulness of the analogy between diffusion models and stochastic thermodynamics. We believe that the conventional thermodynamic uncertainty relations [21,24,110,114,121] are also useful, because we can consider the speed of any observable in the forward process, for example, the speed of data structure breakage in the diffusion process. It is also noteworthy that the short-time thermodynamic uncertainty relation is used to estimate the time-varying velocity field [21,23], which is important in the flow-based generative modeling, and it is worth considering whether there are any aspects in which such a thermodynamic-based method is superior to conventional flow-matching methods.

It is also interesting to reconsider the path probability-based method as discussed in the original paper on the diffusion model from a thermodynamic point of view. This is because the method discussed in the original paper can handle not only the simple diffusion processes described by

the Langevin and Fokker-Planck equations, but also the diffusion process on the graph described by the Markov jump processes, which may provide a more scalable method than the current one. In such a case, the thermodynamic trade-off relations and optimal transport for the Markov jump process [137–140] may be useful to consider the optimality of the diffusion model. In such a case, analogies to an information geometric structure of the Kullback-Leibler divergence in stochastic thermodynamics [32,103,140] may also be important, since the objective function is introduced by the Kullback-Leibler divergence and its minimization is mathematically well discussed as the projection theorem in information geometry [83]. In fact, there are several stochastic methods based on the Schrödinger bridge in the diffusion model [50,71,141,142], which is given by the minimization of the Kullback-Leibler divergence. We may be able to obtain some trade-off relations for such a system by considering an analogy to stochastic thermodynamics based on path probability.

It would also be interesting to examine our results from the perspective of increasing the variety of data generated by raising the upper bound of the response function, which is considered to be the sensitivity. Indeed, numerical experiments with thermodynamic nonconservative forces show that, as the upper bound becomes looser, it becomes easier to generate data that are different from the original data. Increasing the strength of the nonconservative force may be a way to generate new data that are difficult to obtain using conventional methods. It is also possible to deliberately account for nonoptimal transport or increase the error in estimating the velocity field to increase such sensitivity, and our main result [Eq. (77)] and the generalized result [Eq. (100)] may also be useful in such cases.

ACKNOWLEDGMENTS

S. I. thanks Sachinori Watanabe for discussions on diffusion models. K. I. and S. I. thank Taiji Suzuki for insightful discussions. S. I. is supported by JSPS KAKENHI Grants No. 21H01560, No. 22H01141, No. 23H00467, and No. 24H00834, JST ERATO Grant No. JPMJER2302, and UTEC-UTokyo FSI Research Grant Program.

DATA AVAILABILITY

The data are not publicly available. The data are available from the authors upon reasonable request.

APPENDIX A: DERIVATION OF THE RELATIONS BETWEEN THE TWO GRADIENTS OF THE OBJECTIVE FUNCTIONS

We show $\nabla_{\theta} L_{\text{SM}}^{\text{c}}(\theta) = \nabla_{\theta} L_{\text{SM}}(\theta)$ for solving the optimization problem by score matching and that $\nabla_{\theta} L_{\text{FM}}^{\text{c}}(\theta) = \nabla_{\theta} L_{\text{FM}}(\theta)$ is satisfied for solving the optimization problem by flow matching.

1. Score matching

We show that two objective functions

$$L_{\text{SM}}(\theta) = \mathbb{E}_{\mathcal{P}_t, \mathcal{U}}[\|s_t^{\theta}(\mathbf{x}) - \nabla \ln \mathcal{P}_t(\mathbf{x})\|^2] \quad (\text{A1})$$

and

$$L_{\text{SM}}^{\text{c}}(\theta) = \mathbb{E}_{\mathcal{P}_t^{\text{c}}, q, \mathcal{U}}[\|s_t^{\theta}(\mathbf{x}) - \nabla \ln \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y})\|^2] \quad (\text{A2})$$

give the same gradient for the parameter θ . The gradient $\nabla_{\theta} L_{\text{SM}}^{\text{c}}(\theta)$ can be calculated as follows:

$$\begin{aligned} \nabla_{\theta} L_{\text{SM}}^{\text{c}}(\theta) &= \frac{2}{\tau} \int_0^{\tau} dt \int d\mathbf{x} \int d\mathbf{y} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) q(\mathbf{y}) [\nabla_{\theta} s_t^{\theta}(\mathbf{x})] \cdot \left(s_t^{\theta}(\mathbf{x}) - \frac{\nabla \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y})}{\mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y})} \right) \\ &= \frac{2}{\tau} \int_0^{\tau} dt \int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) [\nabla_{\theta} s_t^{\theta}(\mathbf{x})] \cdot \left(s_t^{\theta}(\mathbf{x}) - \frac{\nabla \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \right) \\ &= \nabla_{\theta} L_{\text{SM}}(\theta), \end{aligned} \quad (\text{A3})$$

which means that $\nabla_{\theta} L_{\text{SM}}^{\text{c}}(\theta)$ coincides with $\nabla_{\theta} L_{\text{SM}}(\theta)$. Here, we use $\int d\mathbf{y} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) q(\mathbf{y}) = \mathcal{P}_t(\mathbf{x})$ and $q(\mathbf{y}) \nabla \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) = \nabla [\mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) q(\mathbf{y})]$.

2. Flow matching

Similarly, in flow matching, we show that two objective functions

$$L_{\text{FM}}(\theta) = \mathbb{E}_{\mathcal{P}_t, \mathcal{U}}[\|\mathbf{u}_t^{\theta}(\mathbf{x}) - \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2] \quad (\text{A4})$$

and

$$L_{\text{FM}}^{\text{c}}(\theta) = \mathbb{E}_{\mathcal{P}_t^{\text{c}}, q, \mathcal{U}}[\|\mathbf{u}_t^{\theta}(\mathbf{x}) - \boldsymbol{\nu}_t^{\mathcal{P}^{\text{c}}}(\mathbf{x}|\mathbf{y})\|^2] \quad (\text{A5})$$

give the same gradient for the parameter θ . The gradient $\nabla_{\theta} L_{\text{FM}}^{\text{c}}(\theta)$ can be calculated as follows:

$$\begin{aligned}\nabla_{\theta} L_{\text{FM}}^{\text{c}}(\theta) &= \frac{2}{\tau} \int dt \int dy \int dx \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) q(\mathbf{y}) [\nabla_{\theta} \mathbf{u}_t^{\theta}(\mathbf{x})] \cdot [\mathbf{u}_t^{\theta}(\mathbf{x}) - \boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y})] \\ &= \frac{2}{\tau} \int dt \int dx \mathcal{P}_t(\mathbf{x}) [\nabla_{\theta} \mathbf{u}_t^{\theta}(\mathbf{x})] \cdot [\mathbf{u}_t^{\theta}(\mathbf{x}) - \boldsymbol{\nu}_t^{\text{P}}(\mathbf{x})] \\ &= \nabla_{\theta} L_{\text{FM}}(\theta),\end{aligned}\tag{A6}$$

which means that $\nabla_{\theta} L_{\text{FM}}^{\text{c}}(\theta)$ coincides with $\nabla_{\theta} L_{\text{FM}}(\theta)$. Here, we use $\int dy \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) q(\mathbf{y}) = \mathcal{P}_t(\mathbf{x})$ and

$$\boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x}) = \int dy \boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) q(\mathbf{y}).\tag{A7}$$

APPENDIX B: PARAMETERS OF THE CONDITIONAL PROBABILITY DENSITY FUNCTION

Here, we explain the specific expressions for the parameters $\boldsymbol{\mu}_t(\mathbf{y})$ and Σ_t of the conditional probability density function $\mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_t(\mathbf{y}), \Sigma_t)$. The Fokker-Planck equation for the conditional probability density function $\mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y})$ is given as follows:

$$\partial_t \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) = -\nabla \cdot \left(\boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \right),\tag{B1}$$

$$\begin{aligned}\boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y}) &:= \mathbf{A}_t \mathbf{x} + \mathbf{b}_t - T_t \nabla \ln \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \\ &= (\mathbf{A}_t + T_t \Sigma_t^{-1})[\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})] + \mathbf{A}_t \boldsymbol{\mu}_t(\mathbf{y}) + \mathbf{b}_t.\end{aligned}\tag{B2}$$

$$\begin{aligned}\partial_t \Sigma_t &= \partial_t \int d\mathbf{x} [\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})][\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})]^{\text{T}} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \\ &= \int d\mathbf{x} [\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})][\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})]^{\text{T}} \left(-\nabla \cdot \left(\boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \right) \right) \\ &= \int d\mathbf{x} (\mathbf{A}_t + T_t \Sigma_t^{-1})[\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})][\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})]^{\text{T}} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) + \int d\mathbf{x} [\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})][(\mathbf{A}_t + T_t \Sigma_t^{-1})[\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})]]^{\text{T}} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \\ &= \mathbf{A}_t \Sigma_t + \Sigma_t \mathbf{A}_t^{\text{T}} + 2T_t \mathbf{I},\end{aligned}\tag{B4}$$

where T stands for the transpose and we use $\Sigma_t = \int d\mathbf{x} [\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})][\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})]^{\text{T}} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y})$, $\int d\mathbf{x} [\mathbf{x} - \boldsymbol{\mu}_t(\mathbf{y})] \times \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) = \mathbf{0}$, and $(\Sigma_t^{-1})^{\text{T}} = \Sigma_t^{-1}$.

Now, we consider a situation where \mathbf{A}_t , \mathbf{b}_t , and T_t are expressed, respectively, as $\mathbf{A}_t = (\partial_t \ln m_t) \mathbf{I}$, $\mathbf{b}_t = \mathbf{0}$, and $T_t = \partial_t (\sigma_t^2/2) - \sigma_t^2 \partial_t \ln m_t$ with non-negative parameters $\sigma_t (\geq 0)$ and $m_t (\geq 0)$, and the initial conditions of the parameters are given by $m_{t=0} = 1$ and $\sigma_{t=0} = 0$. In this situation, Eqs. (B3) and (B4) are given as follows:

Thus, the time evolution of the parameter $\boldsymbol{\mu}_t(\mathbf{y})$ can be calculated as follows:

$$\begin{aligned}\partial_t \boldsymbol{\mu}_t(\mathbf{y}) &= \partial_t \int d\mathbf{x} \mathbf{x} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \\ &= \int d\mathbf{x} \mathbf{x} \left(-\nabla \cdot \left(\boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \right) \right) \\ &= \int d\mathbf{x} \boldsymbol{\nu}_t^{\text{P}}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \\ &= \int d\mathbf{x} [\mathbf{A}_t \mathbf{x} + \mathbf{b}_t] \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) - \int d\mathbf{x} T_t \nabla \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) \\ &= \mathbf{A}_t \boldsymbol{\mu}_t(\mathbf{y}) + \mathbf{b}_t,\end{aligned}\tag{B3}$$

where we assume that the contribution vanishes at infinity in the partial integration and we use $\boldsymbol{\mu}_t(\mathbf{y}) = \int d\mathbf{x} \mathbf{x} \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y})$ and $\int d\mathbf{x} \nabla \mathcal{P}_t^{\text{c}}(\mathbf{x}|\mathbf{y}) = 0$. The time evolution of the parameter Σ_t is also calculated as follows:

$$\partial_t \boldsymbol{\mu}_t(\mathbf{y}) = (\partial_t \ln m_t) \boldsymbol{\mu}_t(\mathbf{y}),\tag{B5}$$

$$\partial_t \Sigma_t(\mathbf{y}) = (2\partial_t \ln m_t) \Sigma_t(\mathbf{y}) + [\partial_t (\sigma_t^2) - 2\sigma_t^2 \partial_t \ln m_t] \mathbf{I}.\tag{B6}$$

Therefore, the k th element of the vector $[\boldsymbol{\mu}_t(\mathbf{y})]_k$ and the (k, l) element of the matrix $[\Sigma_t(\mathbf{y})]_{kl}$ satisfy the following equations:

$$\partial_t[\boldsymbol{\mu}_t(\mathbf{y})]_k = (\partial_t \ln m_t)[\boldsymbol{\mu}_t(\mathbf{y})]_k, \quad (\text{B7})$$

$$\partial_t\{[\boldsymbol{\Sigma}_t(\mathbf{y})]_{kl} - \sigma_t^2 \delta_{kl}\} = (2\partial_t \ln m_t)\{[\boldsymbol{\Sigma}_t(\mathbf{y})]_{kl} - \sigma_t^2 \delta_{kl}\}, \quad (\text{B8})$$

where $\delta_{kl} = [\mathbb{I}]_{kl}$ is the Kronecker delta. Since $\boldsymbol{\mu}_{t=0}(\mathbf{y}) = \mathbf{y}$ and $\boldsymbol{\Sigma}_{t=0} = \mathbf{O}$, performing a time integral from $t = 0$ to t yields

$$[\boldsymbol{\mu}_t(\mathbf{y})]_k = m_t[\mathbf{y}]_k, \quad (\text{B9})$$

$$[\boldsymbol{\Sigma}_t(\mathbf{y})]_{kl} - \sigma_t^2 \delta_{kl} = 0, \quad (\text{B10})$$

and, thus, the following specific expressions are obtained:

$$\boldsymbol{\mu}_t(\mathbf{y}) = m_t \mathbf{y}, \quad \boldsymbol{\Sigma}_t = \sigma_t^2 \mathbb{I}. \quad (\text{B11})$$

APPENDIX C: COORDINATE TRANSFORMATION IN DDIMS

We show the detailed calculation of the coordinate transformation in DDIMS. The Fokker-Planck equation [Eq. (1)] with Eq. (23) is given as follows:

$$\partial_t \mathcal{P}_t(\mathbf{x}) = -\nabla \cdot [(\partial_t \ln m_t) \mathbf{x} - T_t \nabla \ln \mathcal{P}_t(\mathbf{x})] \mathcal{P}_t(\mathbf{x}). \quad (\text{C1})$$

From the coordinate transform

$$\bar{\mathbf{x}} = \frac{\mathbf{x}}{m_t}, \quad (\text{C2})$$

we obtain $\mathcal{P}_t(\mathbf{x}) = \bar{\mathcal{P}}_t(\bar{\mathbf{x}})(|d\bar{\mathbf{x}}/d\mathbf{x}|) = \bar{\mathcal{P}}_t(\bar{\mathbf{x}})(m_t)^{-n_d}$. Substituting this into Eq. (C1), the left-hand side is calculated as

$$\partial_t \mathcal{P}_t(\mathbf{x}) = (m_t)^{-n_d} \partial_t \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) - n_d (\partial_t \ln m_t) (m_t)^{-n_d} \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) - (\partial_t \ln m_t) (m_t)^{-(n_d+1)} \mathbf{x} \cdot \nabla_{\bar{\mathbf{x}}} \bar{\mathcal{P}}_t(\bar{\mathbf{x}}), \quad (\text{C3})$$

where $\nabla_{\bar{\mathbf{x}}}$ is defined as $\nabla_{\bar{\mathbf{x}}} = m_t \nabla$. The right-hand side of Eq. (C1) is calculated as

$$\begin{aligned} & -\nabla \cdot [(\partial_t \ln m_t) \mathbf{x} - T_t \nabla \ln \mathcal{P}_t(\mathbf{x})] \mathcal{P}_t(\mathbf{x}) \\ &= -n_d (\partial_t \ln m_t) (m_t)^{-n_d} \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) - (\partial_t \ln m_t) (m_t)^{-(n_d+1)} \mathbf{x} \cdot \nabla_{\bar{\mathbf{x}}} \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) + \nabla_{\bar{\mathbf{x}}} \cdot \left[T_t \nabla_{\bar{\mathbf{x}}} \ln \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right] \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) (m_t)^{-(n_d+2)}. \end{aligned} \quad (\text{C4})$$

From Eqs. (C3) and (C4), we obtain

$$\partial_t \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) = -\nabla_{\bar{\mathbf{x}}} \cdot \left[\left[-\frac{T_t}{(m_t)^2} \nabla_{\bar{\mathbf{x}}} \ln \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right] \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right]. \quad (\text{C5})$$

By substituting $T_t = -\partial_t \ln m_t$ into this equation,

$$\partial_t \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) = -\nabla_{\bar{\mathbf{x}}} \cdot \left[\left[\frac{\partial_t \ln m_t}{(m_t)^2} \nabla_{\bar{\mathbf{x}}} \ln \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right] \bar{\mathcal{P}}_t(\bar{\mathbf{x}}) \right] \quad (\text{C6})$$

is derived.

APPENDIX D: THE RELATIONS BETWEEN THE KULLBACK-LEIBLER DIVERGENCE AND THE ENTROPY PRODUCTION

First, we prove that the Kullback-Leibler divergence

$$D_{\text{KL}}(\mathbb{P}_F \| \mathbb{P}_B) = \sum_{N=0}^{N_\tau-1} \int d\Gamma \mathbb{P}_F(\Gamma) \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t} | \mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}^\dagger(\mathbf{x}_{(N+1)\Delta t} | \mathbf{x}_{N\Delta t})} \quad (\text{D1})$$

coincides with the entropy production

$$S_\tau^{\text{tot}} = \int_0^\tau dt \left[\frac{1}{T_t} \int d\mathbf{x} \|\nu_t^\mathcal{P}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right], \quad (\text{D2})$$

in the limit $\Delta t \rightarrow 0$ with fixed τ . By calculating $\ln[\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t} | \mathbf{x}_{N\Delta t}) / \mathcal{T}_{N\Delta t}^\dagger(\mathbf{x}_{(N+1)\Delta t} | \mathbf{x}_{N\Delta t})]$, we obtain

$$\begin{aligned} \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}^\dagger(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})} &= \frac{\|\mathbf{x}_{(N+1)\Delta t} - \mathbf{x}_{N\Delta t} - \mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t})\Delta t + 2\boldsymbol{\nu}_{N\Delta t}^{\mathcal{P}}(\mathbf{x}_{N\Delta t})\Delta t\|^2}{4T_{N\Delta t}\Delta t} - \frac{\|\mathbf{x}_{(N+1)\Delta t} - \mathbf{x}_{N\Delta t} - \mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t})\Delta t\|^2}{4T_{N\Delta t}\Delta t} \\ &= \frac{\|\boldsymbol{\nu}_{N\Delta t}^{\mathcal{P}}(\mathbf{x}_{N\Delta t})\|^2}{T_{N\Delta t}}\Delta t + \frac{[\mathbf{x}_{(N+1)\Delta t} - \mathbf{x}_{N\Delta t} - \mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t})\Delta t] \cdot \boldsymbol{\nu}_{N\Delta t}^{\mathcal{P}}(\mathbf{x}_{N\Delta t})}{T_{N\Delta t}}. \end{aligned} \quad (\text{D3})$$

By using $\int d\mathbf{x}_{(N+1)\Delta t} \mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})[\mathbf{x}_{(N+1)\Delta t} - \mathbf{x}_{N\Delta t} - \mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t})\Delta t] = \mathbf{0}$ and the normalization of the probability distributions, we obtain

$$\begin{aligned} \int d\Gamma \mathbb{P}^{\text{F}}(\Gamma) \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}^\dagger(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})} &= \int d\mathbf{x}_{N\Delta t} \int d\mathbf{x}_{(N+1)\Delta t} \mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t}) \mathcal{P}_{N\Delta t}(\mathbf{x}_{N\Delta t}) \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}^\dagger(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})} \\ &= \frac{\Delta t}{T_{N\Delta t}} \int d\mathbf{x}_{N\Delta t} \|\boldsymbol{\nu}_{N\Delta t}^{\mathcal{P}}(\mathbf{x}_{N\Delta t})\|^2 \mathcal{P}_{N\Delta t}(\mathbf{x}_{N\Delta t}). \end{aligned} \quad (\text{D4})$$

By taking the limit $\Delta t \rightarrow 0$ with fixed τ , we obtain

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_{\text{F}}\|\mathbb{P}_{\text{B}}) &= \sum_{N=0}^{N_\tau-1} \frac{\Delta t}{T_{N\Delta t}} \int d\mathbf{x} \|\boldsymbol{\nu}_{N\Delta t}^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_{N\Delta t}(\mathbf{x}) \\ &\rightarrow \int_0^\tau dt \left[\frac{1}{T_t} \int d\mathbf{x} \|\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right] = S_\tau^{\text{tot}}. \end{aligned} \quad (\text{D5})$$

Next, we prove that Kullback-Leibler divergence $D_{\text{KL}}(\mathbb{P}_{\text{F}}\|\mathbb{P}'_{\text{B}}) = \mathbb{E}_{\mathbb{P}_{\text{F}}}[s_\tau^{\text{tot}}]$ coincides with $S_\tau^{\text{tot}} = \int_0^\tau dt \dot{S}_t^{\text{sys}} + \int_0^\tau dt \dot{S}_t^{\text{bath}}$ in the limit $\Delta t \rightarrow 0$ with fixed τ . Since the stochastic entropy production is given by

$$s_\tau^{\text{tot}} = \ln \frac{\mathcal{P}_0(\mathbf{x}_0)}{\mathcal{P}_{N_\tau\Delta t}(\mathbf{x}_{N_\tau\Delta t})} + \sum_{N=0}^{N_\tau-1} \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}(\mathbf{x}_{N\Delta t}|\mathbf{x}_{(N+1)\Delta t})}, \quad (\text{D6})$$

$D_{\text{KL}}(\mathbb{P}_{\text{F}}\|\mathbb{P}'_{\text{B}})$ is calculated as follows:

$$\begin{aligned} D_{\text{KL}}(\mathbb{P}_{\text{F}}\|\mathbb{P}'_{\text{B}}) &= \mathbb{E}_{\mathbb{P}_{\text{F}}}[s_\tau^{\text{tot}}] \\ &= \int d\Gamma \mathbb{P}_{\text{F}}(\Gamma) \ln \frac{\mathcal{P}_0(\mathbf{x}_0)}{\mathcal{P}_{N_\tau\Delta t}(\mathbf{x}_{N_\tau\Delta t})} + \sum_{N=0}^{N_\tau-1} \int d\Gamma \mathbb{P}_{\text{F}}(\Gamma) \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t}|\mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}(\mathbf{x}_{N\Delta t}|\mathbf{x}_{(N+1)\Delta t})}. \end{aligned} \quad (\text{D7})$$

From the marginalization of the path probability, the first term of Eq. (D7) is rewritten as follows:

$$\begin{aligned} \int d\Gamma \mathbb{P}_{\text{F}}(\Gamma) \ln \frac{\mathcal{P}_0(\mathbf{x}_0)}{\mathcal{P}_{N_\tau\Delta t}(\mathbf{x}_{N_\tau\Delta t})} &= \int d\mathbf{x}_0 \mathcal{P}_0(\mathbf{x}_0) \ln \mathcal{P}_0(\mathbf{x}_0) - \int d\mathbf{x}_{N_\tau\Delta t} \mathcal{P}_{N_\tau\Delta t}(\mathbf{x}_{N_\tau\Delta t}) \ln \mathcal{P}_{N_\tau\Delta t}(\mathbf{x}_{N_\tau\Delta t}) \\ &= \int_0^\tau dt \partial_t \left[- \int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \ln \mathcal{P}_t(\mathbf{x}) \right] \\ &= \int_0^\tau dt \dot{S}_t^{\text{sys}}. \end{aligned} \quad (\text{D8})$$

The second term of Eq. (D7) is also calculated as

$$\begin{aligned}
& \int d\Gamma \mathbb{P}_F(\Gamma) \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t} | \mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}(\mathbf{x}_{N\Delta t} | \mathbf{x}_{(N+1)\Delta t})} \\
&= \int d\Gamma \mathbb{P}_F(\Gamma) \left[-\frac{\|\mathbf{x}_{(N+1)\Delta t} - \mathbf{x}_{N\Delta t} - \mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t})\Delta t\|^2}{4T_{N\Delta t}\Delta t} + \frac{\|\mathbf{x}_{N\Delta t} - \mathbf{x}_{(N+1)\Delta t} - \mathbf{f}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t})\Delta t\|^2}{4T_{N\Delta t}\Delta t} \right] \\
&= \int d\Gamma \mathbb{P}_F(\Gamma) \left[\frac{(\mathbf{x}_{(N+1)\Delta t} - \mathbf{x}_{N\Delta t}) \cdot \frac{\mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t}) + \mathbf{f}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t})}{2}}{T_{N\Delta t}} + \frac{\|\mathbf{f}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t})\|^2 - \|\mathbf{f}_{N\Delta t}(\mathbf{x}_{N\Delta t})\|^2}{4T_{N\Delta t}} \Delta t \right] \\
&= \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \int d\mathbf{y} \mathcal{T}_{N\Delta t}(\mathbf{y} | \mathbf{x}) \left[\frac{(\mathbf{y} - \mathbf{x}) \cdot \frac{\mathbf{f}_{N\Delta t}(\mathbf{x}) + \mathbf{f}_{N\Delta t}(\mathbf{y})}{2}}{T_{N\Delta t}} \right] \\
&\quad + \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \int d\mathbf{y} \mathcal{T}_{N\Delta t}(\mathbf{y} | \mathbf{x}) \left[\frac{\|\mathbf{f}_{N\Delta t}(\mathbf{y})\|^2 - \|\mathbf{f}_{N\Delta t}(\mathbf{x})\|^2}{4T_{N\Delta t}} \Delta t \right] \\
&= \frac{1}{T_{N\Delta t}} \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \int d\mathbf{y} \mathcal{T}_{N\Delta t}(\mathbf{y} | \mathbf{x}) [(\mathbf{y} - \mathbf{x}) \cdot \mathbf{f}_{N\Delta t}(\mathbf{x})] \\
&\quad + \frac{1}{T_{N\Delta t}} \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \int d\mathbf{y} \mathcal{T}_{N\Delta t}(\mathbf{y} | \mathbf{x}) \left[\sum_{i,j} \frac{[y_i - x_i][y_j - x_j]}{2} \nabla_{x_j} [\mathbf{f}_{N\Delta t}(\mathbf{x})]_i \right] + O[(\Delta t)^2] \\
&= \frac{\Delta t}{T_{N\Delta t}} \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \|\mathbf{f}_{N\Delta t}(\mathbf{x})\|^2 + \Delta t \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \nabla \cdot \mathbf{f}_{N\Delta t}(\mathbf{x}) + O[(\Delta t)^2] \\
&= \frac{\Delta t}{T_{N\Delta t}} \int d\mathbf{x} \mathcal{P}_{N\Delta t}(\mathbf{x}) \mathbf{f}_{N\Delta t}(\mathbf{x}) \cdot \boldsymbol{\nu}_{N\Delta t}^{\mathcal{P}}(\mathbf{x}) + O[(\Delta t)^2]. \tag{D9}
\end{aligned}$$

Therefore, in the limit $\Delta t \rightarrow 0$ with fixed τ , we obtain

$$\sum_{N=0}^{N_\tau-1} \int d\Gamma \mathbb{P}_F(\Gamma) \ln \frac{\mathcal{T}_{N\Delta t}(\mathbf{x}_{(N+1)\Delta t} | \mathbf{x}_{N\Delta t})}{\mathcal{T}_{N\Delta t}(\mathbf{x}_{N\Delta t} | \mathbf{x}_{(N+1)\Delta t})} = \int_0^\tau dt \frac{\int d\mathbf{x} \mathcal{P}_t(\mathbf{x}) \mathbf{f}_t(\mathbf{x}) \cdot \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})}{T_t} = \int_0^\tau dt \dot{S}_t^{\text{bath}}. \tag{D10}$$

From Eqs. (D7), (D8), and (D10), we obtain

$$D_{\text{KL}}(\mathbb{P}_F \| \mathbb{P}'_B) = \int_0^\tau dt \dot{S}_t^{\text{sys}} + \int_0^\tau dt \dot{S}_t^{\text{bath}} = S_\tau^{\text{tot}}. \tag{D11}$$

We note that $\mathbb{E}_{\mathbb{P}_F}[\ln \mathbb{P}_B] = \mathbb{E}_{\mathbb{P}_F}[\ln \mathbb{P}'_B]$ is satisfied, because $S_\tau^{\text{tot}} = D_{\text{KL}}(\mathbb{P}_F \| \mathbb{P}_B) = D_{\text{KL}}(\mathbb{P}_F \| \mathbb{P}'_B)$, while $\mathbb{P}_B(\Gamma) \neq \mathbb{P}'_B(\Gamma)$.

APPENDIX E: TIME INDEPENDENCE OF THE PEARSON χ^2 DIVERGENCE

We show that $D_{\tau-t}$ is independent of time if the velocity field of the forward process is accurately reconstructed $[\hat{\nu}_t(\mathbf{x}) = \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})]$ in the flow-based generative modeling or the probability flow ODE. From Eq. (85) and the continuity equation $\partial_t \mathcal{P}_t(\mathbf{x}) = -\nabla \cdot (\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x}))$, the time derivative of $D_{\tau-t}$ is calculated as

$$\begin{aligned}
\partial_t D_{\tau-t} &= \int d\mathbf{x} \partial_t \frac{[\delta \mathcal{P}_t(\mathbf{x})]^2}{\mathcal{P}_t(\mathbf{x})} \\
&= \int d\mathbf{x} \left[\frac{2\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \partial_t \delta \mathcal{P}_t(\mathbf{x}) - \left(\frac{\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \right)^2 \partial_t \mathcal{P}_t(\mathbf{x}) \right] \\
&= - \int d\mathbf{x} \left[\frac{2\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \{ \nabla \cdot [\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x})] \} \right] + \int d\mathbf{x} \left[\left(\frac{\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \right)^2 \{ \nabla \cdot [\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x})] \} \right] \\
&= \int d\mathbf{x} \left[2 \nabla \left(\frac{\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \right) \cdot \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \delta \mathcal{P}_t(\mathbf{x}) \right] - \int d\mathbf{x} \left[2 \nabla \left(\frac{\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \right) \cdot \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \frac{\delta \mathcal{P}_t(\mathbf{x})}{\mathcal{P}_t(\mathbf{x})} \mathcal{P}_t(\mathbf{x}) \right] \\
&= 0, \tag{E1}
\end{aligned}$$

where we use the partial integration and we assume that $\mathcal{P}_t(\mathbf{x})$ and $\delta\mathcal{P}_t(\mathbf{x})$ vanish at infinity. Therefore, we obtain $\partial_t D_{\tau-t} = 0$, which means that $D_{\tau-t}$ is independent of time.

APPENDIX F: THE DERIVATION OF EQ. (100)

Here, we derive Eq. (100), which is a generalization of the main result for incomplete estimation. In the case of flow-based generative modeling or the probability flow ODE, the time evolution of the probability densities for the forward process is given by $\partial_t \mathcal{P}_t(\mathbf{x}) = -\nabla \cdot (\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x}))$. For the estimated reverse process, the time evolution is also given by

$\partial_t \tilde{\mathcal{P}}_t = -\nabla \cdot (\tilde{\boldsymbol{\nu}}_t(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x}))$ with the incomplete estimated velocity field $\tilde{\boldsymbol{\nu}}_t(\mathbf{x}) [\neq \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})]$. Analogous to Eq. (85), the time evolution of $\delta\mathcal{P}_t(\mathbf{x}) = \mathcal{P}_t(\mathbf{x}) - \tilde{\mathcal{P}}_t(\mathbf{x})$ is calculated as

$$\begin{aligned} \partial_t \delta\mathcal{P}_t(\mathbf{x}) &= -\nabla \cdot \left(\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \mathcal{P}_t(\mathbf{x}) - \tilde{\boldsymbol{\nu}}_t(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x}) \right) \\ &= -\nabla \cdot \left(\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \delta\mathcal{P}_t(\mathbf{x}) + \delta\boldsymbol{\nu}_t(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x}) \right), \end{aligned} \quad (\text{F1})$$

where $\delta\boldsymbol{\nu}_t(\mathbf{x})$ is the difference of the velocity fields defined as $\delta\boldsymbol{\nu}_t(\mathbf{x}) := \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) - \tilde{\boldsymbol{\nu}}_t(\mathbf{x})$. From Eq. (F1), we obtain

$$\begin{aligned} \partial_t (\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi]) &= \int d\mathbf{x} \psi(\mathbf{x}) \partial_t \delta\mathcal{P}_t(\mathbf{x}) \\ &= \int d\mathbf{x} \nabla \psi(\mathbf{x}) \cdot \left(\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \delta\mathcal{P}_t(\mathbf{x}) + \delta\boldsymbol{\nu}_t(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x}) \right) \end{aligned} \quad (\text{F2})$$

for any time-independent 1-Lipschitz function $\psi \in \text{Lip}^1$, using partial integration under the assumption that the surface term is negligible, similar to the proof of our main result [Eqs. (86) and (87)]. From the Cauchy-Schwarz inequality, the 1-Lipschitz continuity $\|\nabla \psi(\mathbf{x})\| \leq 1$, the triangle inequality and Eq. (F2), and $\int d\mathbf{x} \delta\mathcal{P}_t(\mathbf{x}) = 0$, we obtain

$$\begin{aligned} &|\partial_t (\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi])| \\ &= \left| \int d\mathbf{x} \nabla \psi(\mathbf{x}) \cdot (\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \delta\mathcal{P}_t(\mathbf{x}) + \delta\boldsymbol{\nu}_t(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x})) \right| \\ &\leq \left| \int d\mathbf{x} \nabla \psi(\mathbf{x}) \cdot \boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x}) \delta\mathcal{P}_t(\mathbf{x}) \right| + \left| \int d\mathbf{x} \nabla \psi(\mathbf{x}) \cdot \delta\boldsymbol{\nu}_t(\mathbf{x}) \tilde{\mathcal{P}}_t(\mathbf{x}) \right| \\ &\leq \sqrt{\left(\int d\mathbf{x} \|\nabla \psi(\mathbf{x})\|^2 \frac{[\delta\mathcal{P}_t(\mathbf{x})]^2}{\mathcal{P}_t(\mathbf{x})} \right) \left(\int d\mathbf{x} \|\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right)} + \sqrt{\left(\int d\mathbf{x} \|\nabla \psi(\mathbf{x})\|^2 \frac{\tilde{\mathcal{P}}_t(\mathbf{x})^2}{\mathcal{P}_t(\mathbf{x})} \right) \left(\int d\mathbf{x} \|\delta\boldsymbol{\nu}_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right)} \\ &\leq \sqrt{\left(\int d\mathbf{x} \frac{[\delta\mathcal{P}_t(\mathbf{x})]^2}{\mathcal{P}_t(\mathbf{x})} \right) \left(\int d\mathbf{x} \|\boldsymbol{\nu}_t^{\mathcal{P}}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right)} + \sqrt{\left(\int d\mathbf{x} \frac{[\mathcal{P}_t(\mathbf{x}) - \delta\mathcal{P}_t(\mathbf{x})]^2}{\mathcal{P}_t(\mathbf{x})} \right) \left(\int d\mathbf{x} \|\delta\boldsymbol{\nu}_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) \right)} \\ &\leq \sqrt{T_t \dot{S}_t^{\text{tot}} D_{\tau-t}} + \sqrt{1 + D_{\tau-t}} \sqrt{\int d\mathbf{x} \|\delta\boldsymbol{\nu}_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x})}. \end{aligned} \quad (\text{F3})$$

From $|\partial_t \mathcal{W}_1(\mathcal{P}_t(\mathbf{x}), \tilde{\mathcal{P}}_t(\mathbf{x}))| \leq |\partial_t (\mathbb{E}_{\mathcal{P}_t}[\psi] - \mathbb{E}_{\tilde{\mathcal{P}}_t}[\psi])|$ in Eqs. (89) and (90), we obtain a generalization of the instantaneous bound for incomplete estimation as

$$|\partial_t \mathcal{W}_1(\mathcal{P}_t(\mathbf{x}), \tilde{\mathcal{P}}_t(\mathbf{x}))| \leq \sqrt{D_{\tau-t}} \sqrt{T_t \dot{S}_t^{\text{tot}}} + \sqrt{1 + D_{\tau-t}} \sqrt{\int d\mathbf{x} \|\delta\boldsymbol{\nu}_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x})}. \quad (\text{F4})$$

Compared to Eq. (80), this result contains an additional term $\sqrt{1 + D_{\tau-t}} \sqrt{\int d\mathbf{x} \|\delta\boldsymbol{\nu}_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x})}$.

From Eq. (F4), the triangle inequality, Cauchy-Schwarz inequality, $\Delta \mathcal{W}_1 = \mathcal{W}_1(\mathcal{P}_\tau, \tilde{\mathcal{P}}_\tau) - \mathcal{W}_1(\mathcal{P}_0, \tilde{\mathcal{P}}_0)$, and $D_{\max} := \max_{t \in [0, \tau]} D_t$, we can also obtain an upper bound on $|\Delta \mathcal{W}_1|$ as

$$\begin{aligned}
|\Delta \mathcal{W}_1| &\leq \int_0^\tau dt' |\partial_{t'} \mathcal{W}_1(\mathcal{P}_{t'}(\mathbf{x}), \tilde{\mathcal{P}}_{t'}(\mathbf{x}))| \\
&\leq \int_0^\tau dt' \sqrt{D_{\tau-t'}} \sqrt{T_{t'} \dot{S}_{t'}^{\text{tot}}} + \int_0^\tau dt' \sqrt{1 + D_{\tau-t'}} \sqrt{\int d\mathbf{x} \|\delta \nu_{t'}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x})} \\
&\leq \sqrt{D_{\max}} \int_0^\tau dt' \sqrt{T_{t'} \dot{S}_{t'}^{\text{tot}}} + \sqrt{1 + D_{\max}} \int_0^\tau dt' \sqrt{\int d\mathbf{x} \|\delta \nu_{t'}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x})} \\
&\leq \sqrt{D_{\max}} \sqrt{\tau \int_0^\tau dt' T_{t'} \dot{S}_{t'}^{\text{tot}}} + \sqrt{1 + D_{\max}} \sqrt{\tau \int_0^\tau dt' \int d\mathbf{x} \|\delta \nu_{t'}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x})}. \tag{F5}
\end{aligned}$$

The term $\int_0^\tau dt' \int d\mathbf{x} \|\delta \nu_{t'}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x})$ is considered as the objective function, since $L_{\text{SM}}(\theta_{\text{SM}}^*) = (1/\tau) \int_0^\tau dt' \int d\mathbf{x} \|\delta \nu_{t'}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x})$ and $L_{\text{FM}}(\theta_{\text{FM}}^*) = (1/\tau) \int_0^\tau dt' \int d\mathbf{x} \|\delta \nu_{t'}(\mathbf{x})\|^2 \mathcal{P}_{t'}(\mathbf{x})$ hold in the probability flow ODE and in the flow matching, respectively. Taking this into account, we can rewrite this bound as

$$\frac{|\Delta \mathcal{W}_1|}{\sqrt{\tau D_{\max}}} \leq \sqrt{\int_0^\tau dt' T_{t'} \dot{S}_{t'}^{\text{tot}}} + \sqrt{1 + \frac{1}{D_{\max}}} \sqrt{\tau L_\alpha(\theta_\alpha^*)} \quad (\alpha = \text{SM, FM}). \tag{F6}$$

APPENDIX G: DETAILS OF THE NUMERICAL EXPERIMENTS

1. One-dimensional Gaussian mixture distribution

We explain the details of the numerical calculations in Fig. 6. In the numerical calculation, the input distribution $q(\mathbf{x})$ is given by the one-dimensional situation Gaussian mixture distribution

$$q(\mathbf{x}) = \frac{1}{2} \mathcal{N}(\mathbf{x}|m^a, \sigma^2) + \frac{1}{2} \mathcal{N}(\mathbf{x}|m^b, \sigma^2), \tag{G1}$$

with $\sigma = 0.3$ and $m^a = -m^b = \sqrt{0.35}$. Here, we can analytically obtain $\nu_t^{\mathcal{P}}(\mathbf{x})$ as follows. By considering the Fokker-Planck equation for the conditional distribution,

$\partial_t \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) = -\nabla \cdot (\nu_t^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}))$, the time evolution of $\mathcal{P}_t(\mathbf{x}) = \int d\mathbf{y} \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})$ is given by

$$\begin{aligned}
\partial_t \mathcal{P}_t(\mathbf{x}) &= \partial_t \left[\int d\mathbf{y} \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y}) \right] \\
&= -\nabla \cdot \left(\frac{\int d\mathbf{y} \nu_t^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})}{\mathcal{P}_t(\mathbf{x})} \mathcal{P}_t(\mathbf{x}) \right). \tag{G2}
\end{aligned}$$

Comparing this equation with Eq. (1), we obtain

$$\nu_t^{\mathcal{P}}(\mathbf{x}) = \int d\mathbf{y} \frac{\nu_t^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y}) \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})}{\mathcal{P}_t(\mathbf{x})}. \tag{G3}$$

From $\nu_t^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y}) = \mathbf{f}_t(\mathbf{x}) - T_t \nabla \ln \mathcal{P}_t^c(\mathbf{x}|\mathbf{y})$ and Eqs. (22) and (23), we obtain

$$\begin{aligned}
\nu_t^{\mathcal{P}^c}(\mathbf{x}|\mathbf{y}) &= (\partial_t \ln m_t) \mathbf{x} - \left[\partial_t \left(\frac{\sigma_t^2}{2} \right) - \sigma_t^2 \partial_t \ln m_t \right] \nabla \ln \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) \\
&= \frac{\partial_t m_t}{m_t} \mathbf{x} - \sigma_t^2 \left(\frac{\partial_t \sigma_t}{\sigma_t} - \frac{\partial_t m_t}{m_t} \right) \nabla \ln \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}). \tag{G4}
\end{aligned}$$

Therefore, $\nu_t^{\mathcal{P}}(\mathbf{x})$ is calculated as

$$\begin{aligned}
\nu_t^{\mathcal{P}}(\mathbf{x}) &= \int d\mathbf{y} \left[\frac{\partial_t m_t}{m_t} \mathbf{x} - \sigma_t^2 \left(\frac{\partial_t \sigma_t}{\sigma_t} - \frac{\partial_t m_t}{m_t} \right) \nabla \ln \mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) \right] \frac{\mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})}{\mathcal{P}_t(\mathbf{x})} \\
&= \int d\mathbf{y} \left[\frac{\partial_t m_t}{m_t} \mathbf{x} + \left(\frac{\partial_t \sigma_t}{\sigma_t} - \frac{\partial_t m_t}{m_t} \right) (\mathbf{x} - m_t \mathbf{y}) \right] \frac{\mathcal{P}_t^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})}{\mathcal{P}_t(\mathbf{x})} \\
&= \frac{\partial_t \sigma_t}{\sigma_t} \mathbf{x} - \frac{m_t}{\mathcal{P}_t(\mathbf{x})} \left(\frac{\partial_t \sigma_t}{\sigma_t} - \frac{\partial_t m_t}{m_t} \right) \mathbb{E}_{\mathcal{P}_t^c, q}[\mathbf{y}], \tag{G5}
\end{aligned}$$

where we use $\mathcal{P}_i^c(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\mathbf{m}_i\mathbf{y}, \sigma_i^2)$ and the expected value $\mathbb{E}_{\mathcal{P}_i^c, q}[\mathbf{y}]$ is defined as $\mathbb{E}_{\mathcal{P}_i^c, q}[\mathbf{y}] = \int d\mathbf{y} \mathbf{y} \mathcal{P}_i^c(\mathbf{x}|\mathbf{y}) q(\mathbf{y})$. Here, $\mathbb{E}_{\mathcal{P}_i^c, q}[\mathbf{y}]$ is calculated as

$$\begin{aligned} \mathbb{E}_{\mathcal{P}_i^c, q}[\mathbf{y}] &= \frac{1}{2} \int d\mathbf{y} \mathbf{y} \mathcal{N}(\mathbf{x}|\mathbf{m}_i\mathbf{y}, \sigma_i^2) \mathcal{N}(\mathbf{y}|\mathbf{m}^a, \sigma^2) + \frac{1}{2} \int d\mathbf{y} \mathbf{y} \mathcal{N}(\mathbf{x}|\mathbf{m}_i\mathbf{y}, \sigma_i^2) \mathcal{N}(\mathbf{y}|\mathbf{m}^b, \sigma^2) \\ &= \frac{1}{2} \frac{\sigma^2 \mathbf{m}_i \mathbf{x} + \sigma_i^2 \mathbf{m}^a \mathbf{1}}{\sigma_i^2 + \sigma^2 \mathbf{m}_i^2} \mathcal{N}(\mathbf{x}|\mathbf{m}_i \mathbf{m}^a, \sigma_i^2 + \sigma^2 \mathbf{m}_i^2) + \frac{1}{2} \frac{\sigma^2 \mathbf{m}_i \mathbf{x} + \sigma_i^2 \mathbf{m}^b \mathbf{1}}{\sigma_i^2 + \sigma^2 \mathbf{m}_i^2} \mathcal{N}(\mathbf{x}|\mathbf{m}_i \mathbf{m}^b, \sigma_i^2 + \sigma^2 \mathbf{m}_i^2) \\ &= \frac{1}{2} \frac{\sigma^2 \mathbf{m}_i \mathbf{x} + \sigma_i^2 \mathbf{m}^a \mathbf{1}}{\sigma_i^2 + \sigma^2 \mathbf{m}_i^2} \mathcal{N}^a + \frac{1}{2} \frac{\sigma^2 \mathbf{m}_i \mathbf{x} + \sigma_i^2 \mathbf{m}^b \mathbf{1}}{\sigma_i^2 + \sigma^2 \mathbf{m}_i^2} \mathcal{N}^b. \end{aligned} \quad (\text{G6})$$

We use this analytical expression of the velocity field $\boldsymbol{\nu}_i^p(\mathbf{x})$ [Eq. (98)] for the cosine and cond-OT schedules.

To compute the velocity field for the optimal transport, we adopted the optimal transport conditional flow matching method [70]. This method is a conditional flow-matching method that estimates the velocity field for the optimal transport by using samples generated from the optimal transport plan. To compute the optimal transport plan, we use the PYTHON optimal transport (POT) library [143]. As an estimation architecture, we use a fully connected multi-layer perceptron with five layers consisting of 1024 neurons in each layer. The hyperbolic tangent function was used as the activation function. For training, we use the Adam optimizer [144] with a learning rate of 3×10^{-4} for learning 10^7 data points with a batch size of 4056.

To obtain the probability density function for the estimated reverse process, we numerically compute the ODE $d\tilde{\mathbf{x}}_i^\dagger/d\tilde{t} = \tilde{\boldsymbol{\nu}}_i^\dagger(\tilde{\mathbf{x}}_i^\dagger) = -\boldsymbol{\nu}_{\tau-\tilde{t}}^p(\tilde{\mathbf{x}}_i^\dagger)$ for the velocity fields of the cosine schedule, the cond-OT schedule and the OT schedule with various initial conditions, using the fourth-order explicit Runge-Kutta method. Specifically, we compute 3×10^8 samples and use a cubic spline interpolation for the histogram of the samples to approximate the probability density function. To numerically compute the each term in the inequality [Eq. (80)], we use numerical integration methods for $D_{\tau-t} = \int d\mathbf{x} [\delta \mathcal{P}_t(\mathbf{x})]^2 / \mathcal{P}_t(\mathbf{x})$ and $[v_2(t)]^2 = \int d\mathbf{x} \|\boldsymbol{\nu}_t(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x})$, and we use the POT library for the 1-Wasserstein distance. We use $\tau/\Delta t = 100$ for the time step Δt and create a uniform grid for the spatial step

Δx as $[4.5 - (-4.5)]/\Delta x = 5 \times 10^2$ in $x \in [-4.5, 4.5]$. When the initial condition is given by the uniform distribution, we set the spatial step to $[4.5 - (-4.5)]/\Delta x = 5 \times 10^3$ because the computation can be numerically unstable for a large spatial step due to the discontinuity of the distribution. We show the initial conditions in Table III. We also show the time evolution of the probability density functions in the forward and estimated reverse process for each condition in Fig. 9.

2. Two-dimensional Swiss-roll dataset

We explain the details of the numerical calculations in Fig. 7. First, we provide the details of the score estimation method under Eq. (99). We adopt a slightly modified version of the denoising score matching method introduced in Sec. II E due to the nonconservativeness of $\mathbf{f}_t(\mathbf{x})$. Similar to the derivation in Appendix B, we solve the differential equations (B3) and (B4) to obtain a closed-form solution of $\mathcal{P}_i^c(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i(\mathbf{y}), \boldsymbol{\Sigma}_i)$ under the initial conditions $\boldsymbol{\mu}_0(\mathbf{y}) = \mathbf{y}$ and $\boldsymbol{\Sigma}_0 = \mathbf{O}$. Substituting Eq. (99) into Eq. (B3), we obtain

$$\partial_t \boldsymbol{\mu}_t(\mathbf{y}) = \begin{pmatrix} H_t & -G_t \\ G_t & H_t \end{pmatrix} \boldsymbol{\mu}_t(\mathbf{y}). \quad (\text{G7})$$

This differential equation is solvable, and the solution is given by

TABLE III. The initial conditions used in the numerical calculations (Fig. 6). Here, $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \sigma^2)$ denotes the one-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and variance σ^2 .

Initial conditions	Probability density functions
Gaussian distribution with a different mean	$q(\mathbf{x}) = \mathcal{N}(\mathbf{x} 1.0, 1.0)$
Gaussian distribution with a different variance	$q(\mathbf{x}) = \mathcal{N}(\mathbf{x} 0.0, 0.8^2)$
Gaussian mixture distribution	$q(\mathbf{x}) = \frac{1}{3} [\mathcal{N}(\mathbf{x} \frac{-1}{0.94}, (\frac{0.3}{0.94})^2) + \mathcal{N}(\mathbf{x} 0, (\frac{0.8}{0.94})^2) + \mathcal{N}(\mathbf{x} \frac{1}{0.94}, (\frac{0.3}{0.94})^2)]$
Uniform distribution	$q(\mathbf{x}) = \frac{1}{2\sqrt{3}} (\mathbf{x} \in [-\sqrt{3}, \sqrt{3}])$

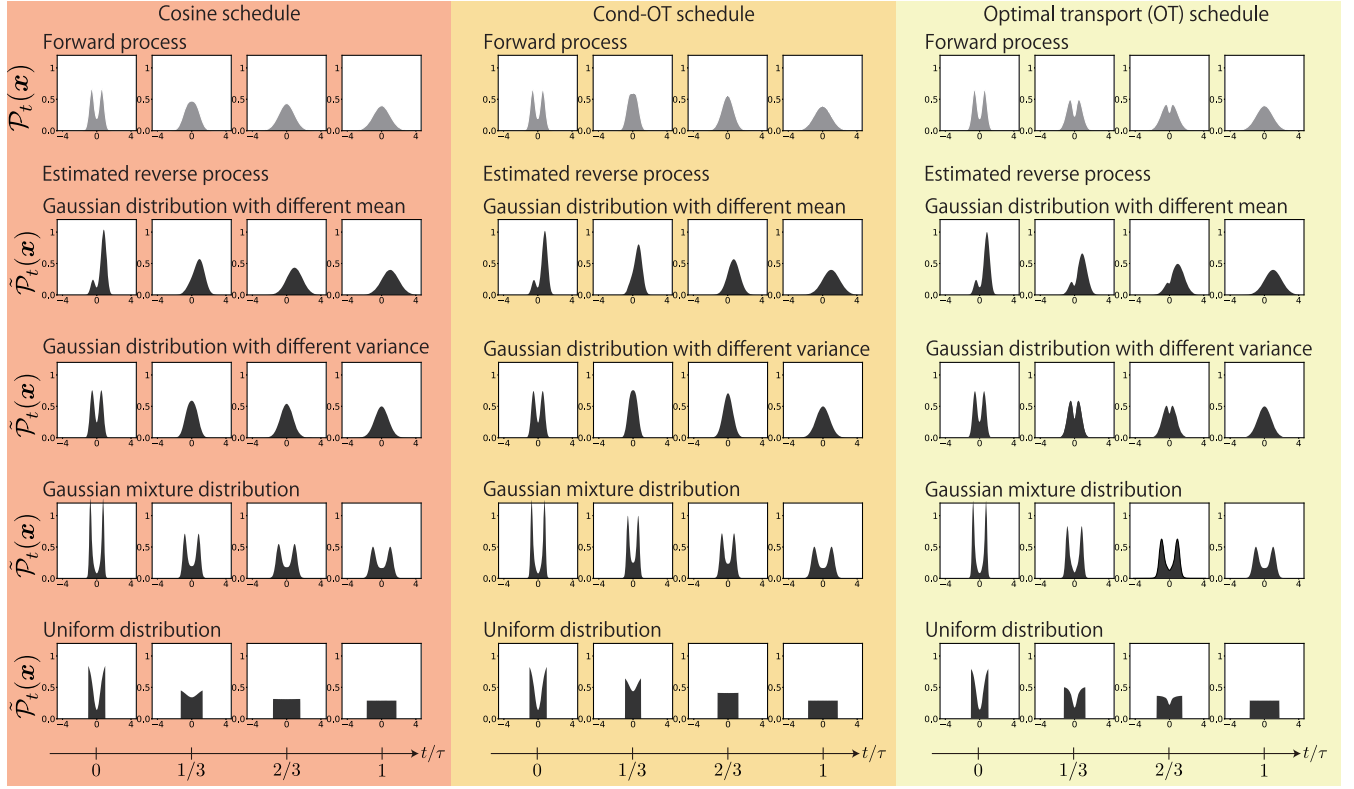


FIG. 9. The time evolution of the probability density functions in the forward and estimated reverse processes, for each condition used in the numerical calculations (Fig. 6).

$$\begin{aligned}
 \mu_t(\mathbf{y}) &= \mathbf{P} \exp \left(\int_0^t dt' \begin{pmatrix} H_{t'} + iG_{t'} & 0 \\ 0 & H_{t'} - iG_{t'} \end{pmatrix} \right) \mathbf{P}^* \mu_0(\mathbf{y}), \\
 &= \exp \left(\int_0^t dt' H_{t'} \right) \begin{pmatrix} \cos \int_0^t dt' G_{t'} & -\sin \int_0^t dt' G_{t'} \\ \sin \int_0^t dt' G_{t'} & \cos \int_0^t dt' G_{t'} \end{pmatrix} \mu_0(\mathbf{y}),
 \end{aligned} \tag{G8}$$

where $i = \sqrt{-1}$ is an imaginary unit, \mathbf{P} is a unitary matrix defined as

$$\mathbf{P} = \frac{1}{\sqrt{2}} \begin{pmatrix} i & -i \\ 1 & 1 \end{pmatrix}, \tag{G9}$$

and \mathbf{P}^* denotes an adjoint matrix of \mathbf{P} .

Next, we solve the differential equation for the covariance matrix [Eq. (B4)]. First, we define the matrix \mathbf{A}_t as

$$\mathbf{A}_t = \begin{pmatrix} H_t & -G_t \\ G_t & H_t \end{pmatrix} = \begin{pmatrix} -h\left(1 - \frac{t}{\tau}\right) - T_{\text{fin}}\left(\frac{t}{\tau}\right)^2 & -g\left(1 - \frac{t}{\tau}\right) \\ g\left(1 - \frac{t}{\tau}\right) & -h\left(1 - \frac{t}{\tau}\right) - T_{\text{fin}}\left(\frac{t}{\tau}\right)^2 \end{pmatrix}. \tag{G10}$$

Here, Eq. (B4) is

$$\mathbf{A}_t \Sigma_t + \Sigma_t \mathbf{A}_t^* + 2T_t \mathbf{I} - \partial_t \Sigma_t = 0. \tag{G11}$$

The real part of the eigenvalue of \mathbf{A}_t , that is, H_t , is negative due to $h > 0$ and $T_{\text{fin}} > 0$. Thus, an analytic expression of Σ_t is formally obtained as

$$\begin{aligned}
\Sigma_t &= \int_0^\infty ds e^{A_t s} (2T_t \mathbf{I} - \partial_t \Sigma_t) e^{A_t^* s} \\
&= 2T_t \int_0^\infty ds e^{(A_t + A_t^*) s} - \int_0^\infty ds e^{A_t s} \partial_t \Sigma_t e^{A_t^* s} \\
&= 2T_t \int_0^\infty ds e^{2H_t s} \mathbf{I} - \int_0^\infty ds e^{A_t s} \partial_t \Sigma_t e^{A_t^* s} \\
&= -\frac{T_t}{H_t} \mathbf{I} - \int_0^\infty ds e^{A_t s} \partial_t \Sigma_t e^{A_t^* s},
\end{aligned} \tag{G12}$$

because we can obtain $A_t \Sigma_t + \Sigma_t A_t^* = \int_0^\infty ds (d/ds) [e^{A_t s} (2T_t \mathbf{I} - \partial_t \Sigma_t) e^{A_t^* s}] = -2T_t \mathbf{I} + \partial_t \Sigma_t$ if the real part of the eigenvalue of A_t is negative. In this case, it is possible to derive a closed-form solution. Let \tilde{A}_t and $\tilde{\Sigma}_t$ be $\tilde{A}_t = \mathbf{P}^* A_t \mathbf{P}$ and $\tilde{\Sigma}_t = \mathbf{P}^* \Sigma_t \mathbf{P}$, respectively. Equation (G12) can be rewritten as

$$\begin{aligned}
\tilde{\Sigma}_t &= -\frac{T_t}{H_t} \mathbf{I} - \int_0^\infty ds e^{\tilde{A}_t s} \partial_t (\mathbf{P}^* \Sigma_t \mathbf{P}) e^{\tilde{A}_t^* s} \\
&= -\frac{T_t}{H_t} \mathbf{I} - \int_0^\infty ds \begin{pmatrix} e^{(H_t + iG_t)s} & 0 \\ 0 & e^{(H_t - iG_t)s} \end{pmatrix} \partial_t \tilde{\Sigma}_t \begin{pmatrix} e^{(H_t - iG_t)s} & 0 \\ 0 & e^{(H_t + iG_t)s} \end{pmatrix} \\
&= -\frac{T_t}{H_t} \mathbf{I} - \int_0^\infty ds \begin{pmatrix} e^{2H_t s} [\partial_t \tilde{\Sigma}_t]_{11} & e^{2(H_t + iG_t)s} [\partial_t \tilde{\Sigma}_t]_{12} \\ e^{2(H_t - iG_t)s} [\partial_t \tilde{\Sigma}_t]_{21} & e^{2H_t s} [\partial_t \tilde{\Sigma}_t]_{22} \end{pmatrix} \\
&= -\frac{T_t}{H_t} \mathbf{I} + \frac{1}{2} \begin{pmatrix} [\partial_t \tilde{\Sigma}_t]_{11}/H_t & [\partial_t \tilde{\Sigma}_t]_{12}/(H_t + iG_t) \\ [\partial_t \tilde{\Sigma}_t]_{21}/(H_t - iG_t) & [\partial_t \tilde{\Sigma}_t]_{22}/H_t \end{pmatrix}.
\end{aligned} \tag{G13}$$

Considering this equation for each component of the matrix $\partial_t \tilde{\Sigma}_t$, we obtain

$$\partial_t [\tilde{\Sigma}_t]_{ii} = 2T_t + 2H_t [\tilde{\Sigma}_t]_{ii} \quad (i = 1, 2), \tag{G14}$$

$$\partial_t [\tilde{\Sigma}_t]_{12} = 2(H_t + iG_t) [\tilde{\Sigma}_t]_{12}, \tag{G15}$$

$$\partial_t [\tilde{\Sigma}_t]_{21} = 2(H_t - iG_t) [\tilde{\Sigma}_t]_{21}, \tag{G16}$$

which is solvable under the initial condition $\Sigma_0 = \tilde{\Sigma}_0 = \mathbf{O}$. The solution is given by

$$\tilde{\Sigma}_t = \left[e^{2 \int_0^t dt' H_t'} \int_0^t ds e^{-2 \int_0^s dt' H_t'} T_s \right] \mathbf{I}. \tag{G17}$$

Since this is a scaled identity matrix \mathbf{I} , $\Sigma_t = \mathbf{P} \tilde{\Sigma}_t \mathbf{P}^*$ is also obtained as

$$\Sigma_t = \left[e^{2 \int_0^t dt' H_t'} \int_0^t ds e^{-2 \int_0^s dt' H_t'} T_s \right] \mathbf{I}. \tag{G18}$$

Using the solutions [Eqs. (G8) and (G18)], we empirically minimize the conditional objective function Eq. (18) and estimate the score function $s_t^\theta(\mathbf{x})$. We use a fully connected multilayer perceptron with eight layers consisting of 1024 neurons in each layer for $s_t^\theta(\mathbf{x})$. The hyperbolic tangent function was used as the activation function. For training, we use the Adam optimizer [144] with a learning

rate of 3×10^{-4} to learn 10^7 data points with a batch size of 1024. The number of samples for stochastic processes was set to 1×10^8 . For the sampling procedure, we use the Euler-Maruyama method for the SDE in the forward process and the fourth-order Runge-Kutta method for the ODE in the estimated reverse process.

To estimate the optimal transport, we use the Nadaraya-Watson style model described in Ref. [72]. Here, we consider the 2-Wasserstein distance between $\mathcal{P}_0(\mathbf{x}_0)$ and $\mathcal{P}_\tau(\mathbf{x}_\tau)$ and obtain the optimal transport plan $\pi^*(\mathbf{x}_0, \mathbf{x}_\tau)$ for the 2-Wasserstein distance $\mathcal{W}_2(\mathcal{P}_0, \mathcal{P}_\tau) = \sqrt{\mathbb{E}_{\pi^*}[\|\mathbf{x}_0 - \mathbf{x}_\tau\|^2]}$, where $\mathbb{E}_{\pi^*}[\cdot]$ is defined as $\mathbb{E}_{\pi^*}[\cdot] = \int d\mathbf{x}_0 \int d\mathbf{x}_\tau \pi^*(\mathbf{x}_0, \mathbf{x}_\tau) \cdot$. We consider the interpolation $\mathbf{x}_t = (1 - t/\tau)\mathbf{x}_0 + (t/\tau)\mathbf{x}_\tau$ and velocity $(\mathbf{x}_\tau - \mathbf{x}_0)/\tau = (\mathbf{x}_\tau - \mathbf{x}_t)/(\tau - t)$. We consider the expression of the optimal velocity field $\mathbf{v}_t^{\mathcal{P}}(\mathbf{x}) = \mathbb{E}[(\mathbf{x}_\tau - \mathbf{x}_0)/\tau | \mathbf{x}_t = \mathbf{x}]$, which is the conditional expected value of $(\mathbf{x}_\tau - \mathbf{x}_0)/\tau$ under the condition $\mathbf{x}_t = \mathbf{x}$. We introduce the normalized Gaussian kernel $\kappa(\mathbf{x}_t, \mathbf{x}) = \exp[-\|\mathbf{x}_t - \mathbf{x}\|^2 / (2\sigma_K^2)] / \mathbb{E}_{\pi^*}[\exp[-\|\mathbf{x}_t - \mathbf{x}\|^2 / (2\sigma_K^2)]]$. If σ_K is relatively small, $\pi^*(\mathbf{x}_0, \mathbf{x}_\tau) \kappa(\mathbf{x}_t, \mathbf{x})$ can be regarded as the conditional probability of $(\mathbf{x}_0, \mathbf{x}_\tau)$ under the condition $\mathbf{x}_t = \mathbf{x}$. Therefore, the optimal velocity field may be obtained from the expression

$$\mathbb{E}_{\pi^*} \left[\frac{\mathbf{x}_\tau - \mathbf{x}}{\tau - t} \kappa(\mathbf{x}_t, \mathbf{x}) \right], \tag{G19}$$

and we use 1×10^7 samples and $\sigma_K = 0.08$ for this expression to estimate the optimal velocity field.

To numerically compute each term in the inequality [Eq. (80)], we use numerical integration methods for $D_{\tau-t} = \int d\mathbf{x} [\delta \mathcal{P}_t(\mathbf{x})]^2 / \mathcal{P}_t(\mathbf{x})$ and $T_i \dot{S}_i^{\text{tot}} = \int d\mathbf{x} \|\nu_i(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x})$, and we use the POT library for the 1-Wasserstein distance. We use $\tau/\Delta t = 100$ for the time step Δt and create a uniform two-dimensional grid for the spatial step $\Delta x = 2 \times 10^{-2}$ in the region $x_1 \in [-5, 5]$ and $x_2 \in [-5, 5]$ for $\mathbf{x} = (x_1, x_2)$.

3. Real-world image datasets

Here, we describe the details of the experiment in Fig. 8. In this experiment, we use the trained flow matching model for the 256×256 downsampled CelebA HQ dataset [119] and the LSUN bedrooms subset [120] provided in Ref. [118]. From each dataset, we sample 3×10^4 points in the latent space and compute the forward and the estimated reverse process. We assume that the estimation of the velocity field in the trained model is accurate. Let $\mathbf{u}_t^{\theta^*}(\mathbf{x})$ be the velocity field for the forward process in the trained model, and we compute $\partial_t \mathbf{x}_t = \mathbf{u}_t^{\theta^*}(\mathbf{x}_t)$ and $\partial_t \tilde{\mathbf{x}}_t^{\dagger} = -\mathbf{u}_{\tau-t}^{\theta^*}(\tilde{\mathbf{x}}_t^{\dagger})$ for forward and estimated reverse processes, respectively. For the initial state of the estimated reverse process $\tilde{\mathbf{x}}_0^{\dagger}$, we use 3×10^4 samples from the standard Gaussian distribution $\mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{I})$.

Using the generated trajectories of the processes \mathbf{x}_t and $\tilde{\mathbf{x}}_t^{\dagger}$, we estimate $c_2 = \int_0^{\tau} dt [v_2(t)]^2$ by empirically calculating the expectation as follows:

$$c_2 = \int_0^{\tau} dt \int d\mathbf{x} \|\mathbf{u}_t^{\theta^*}(\mathbf{x})\|^2 \mathcal{P}_t(\mathbf{x}) = \tau \mathbb{E}_{\mathcal{U}, \mathcal{P}_t} [\|\mathbf{u}_t^{\theta^*}(\mathbf{x})\|^2]. \quad (\text{G20})$$

For the response function $\eta = (\Delta \mathcal{W}_1)^2 / D_0$, we estimate $D_0 = \int d\mathbf{x} [\mathcal{P}_0^{\dagger}(\mathbf{x}) - \tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x})]^2 / \mathcal{P}_0^{\dagger}(\mathbf{x})$ by assuming that $\mathcal{P}_0^{\dagger}(\mathbf{x})$ and $\tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x})$ are Gaussian distributions, $\mathcal{P}_0^{\dagger}(\mathbf{x})$ is close to $\tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x})$, and the higher order of the difference $\delta \mathcal{P}_0^{\dagger}(\mathbf{x}) = \tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x}) - \mathcal{P}_0^{\dagger}(\mathbf{x})$ is negligible in this experiment. In this situation, D_0 may be estimated using the Kullback-Leibler divergence $D_0/2 \approx D_{\text{KL}}(\mathcal{P}_0^{\dagger} \parallel \tilde{\mathcal{P}}_0^{\dagger}) = \int d\mathbf{x} \mathcal{P}_0^{\dagger}(\mathbf{x}) \ln [\mathcal{P}_0^{\dagger}(\mathbf{x}) / \tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x})]$ because $D_{\text{KL}}(\mathcal{P}_0^{\dagger} \parallel \tilde{\mathcal{P}}_0^{\dagger})$ is given by

$$\begin{aligned} D_{\text{KL}}(\mathcal{P}_0^{\dagger} \parallel \tilde{\mathcal{P}}_0^{\dagger}) &= \int d\mathbf{x} \mathcal{P}_0^{\dagger}(\mathbf{x}) \ln \frac{\mathcal{P}_0^{\dagger}(\mathbf{x})}{\tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x})} \\ &= \frac{1}{2} D_0 + O((\delta \mathcal{P}_0^{\dagger}(\mathbf{x}))^3). \end{aligned} \quad (\text{G21})$$

The advantage of using the Kullback-Leibler divergence instead of the χ^2 divergence is that the Kullback-Leibler divergence between two Gaussian distributions can be computed using only the mean and covariance matrix. In a realistic situation of the diffusion model for the real-world

image dataset, such as the one we are dealing with here, we can assume that $\mathcal{P}_0^{\dagger}(\mathbf{x})$ and $\tilde{\mathcal{P}}_0^{\dagger}(\mathbf{x})$ are given by high-dimensional Gaussian distributions and the difference between \mathcal{P}_0^{\dagger} and $\tilde{\mathcal{P}}_0^{\dagger}$ is relatively small. In this case, it is numerically more stable and feasible to compute D_0 using the Kullback-Leibler divergence than using the conventional definition of the χ^2 divergence. We also assume that the covariance matrices of the Gaussian distributions are proportional to the identity matrix to ensure that the estimated value of the Kullback-Leibler divergence is finite for the high-dimensional systems. For $\Delta \mathcal{W}_1 = \mathcal{W}_1(p, q) - \mathcal{W}_1(\mathcal{P}_0^{\dagger}, \tilde{\mathcal{P}}_0^{\dagger})$, we compute the 1-Wasserstein distances empirically using the POT library [143].

-
- [1] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, New York, 1992), Vol. 1.
 - [2] K. Sekimoto, *Stochastic Energetics*, Lecture Notes in Physics Vol. 799 (Springer, Berlin, Heidelberg, 2010), [10.1007/978-3-642-05411-2](https://doi.org/10.1007/978-3-642-05411-2).
 - [3] U. Seifert, *Stochastic thermodynamics, fluctuation theorems and molecular machines*, *Rep. Prog. Phys.* **75**, 126001 (2012).
 - [4] K. Sekimoto, *Langevin equation and thermodynamics*, *Prog. Theor. Phys. Suppl.* **130**, 17 (1998).
 - [5] J. Kurchan, *Fluctuation theorem for stochastic dynamics*, *J. Phys. A* **31**, 3719 (1998).
 - [6] T. Hatano and S. I. Sasa, *Steady-state thermodynamics of Langevin systems*, *Phys. Rev. Lett.* **86**, 3463 (2001).
 - [7] U. Seifert, *Entropy production along a stochastic trajectory and an integral fluctuation theorem*, *Phys. Rev. Lett.* **95**, 040602 (2005).
 - [8] V. Y. Chernyak, M. Chertkov, and C. Jarzynski, *Path-integral analysis of fluctuation theorems for general Langevin processes*, *J. Stat. Mech.* (2006) P08001.
 - [9] T. Schmiedl and U. Seifert, *Efficiency at maximum power: An analytically solvable model for stochastic heat engines*, *Europhys. Lett.* **81**, 20003 (2007).
 - [10] A. E. Allahverdyan, D. Janzing, and G. Mahler, *Thermodynamic efficiency of information and heat flow*, *J. Stat. Mech.* (2009) P09011.
 - [11] C. Van den Broeck and M. Esposito, *Three faces of the second law. II. Fokker-Planck formulation*, *Phys. Rev. E* **82**, 011144 (2010).
 - [12] T. Sagawa and M. Ueda, *Nonequilibrium thermodynamics of feedback control*, *Phys. Rev. E* **85**, 021104 (2012).
 - [13] S. Ito and T. Sagawa, *Information thermodynamics on causal networks*, *Phys. Rev. Lett.* **111**, 180603 (2013).
 - [14] J. M. Horowitz and H. Sandberg, *Second-law-like inequalities with information and their interpretations*, *New J. Phys.* **16**, 125007 (2014).
 - [15] S. Ito and T. Sagawa, *Maxwell's demon in biochemical signal transduction with feedback loop*, *Nat. Commun.* **6**, 1 (2015).
 - [16] T. R. Gingrich, G. M. Rotskoff, and J. M. Horowitz, *Inferring dissipation from current fluctuations*, *J. Phys. A* **50**, 184004 (2017).

- [17] A. Dechant and S. I. Sasa, *Entropic bounds on currents in Langevin systems*, *Phys. Rev. E* **97**, 062101 (2018).
- [18] J. Li, J. M. Horowitz, T. R. Gingrich, and N. Fakhri, *Quantifying dissipation using fluctuating currents*, *Nat. Commun.* **10**, 1666 (2019).
- [19] Y. Hasegawa and T. Van Vu, *Uncertainty relations in stochastic processes: An information inequality approach*, *Phys. Rev. E* **99**, 062126 (2019).
- [20] S. Ito and A. Dechant, *Stochastic time evolution, information geometry, and the Cramér-Rao bound*, *Phys. Rev. X* **10**, 021056 (2020).
- [21] S. Otsubo, S. Ito, A. Dechant, and T. Sagawa, *Estimating entropy production by machine learning of short-time fluctuating currents*, *Phys. Rev. E* **101**, 062106 (2020).
- [22] A. Dechant and S. I. Sasa, *Continuous time reversal and equality in the thermodynamic uncertainty relation*, *Phys. Rev. Res.* **3**, L042012 (2021).
- [23] S. Otsubo, S. K. Manikandan, T. Sagawa, and S. Krishnamurthy, *Estimating time-dependent entropy production from non-equilibrium trajectories*, *Commun. Phys.* **5**, 11 (2022).
- [24] T. Koyuk and U. Seifert, *Thermodynamic uncertainty relation for time-dependent driving*, *Phys. Rev. Lett.* **125**, 260604 (2020).
- [25] C. Villani et al., *Optimal Transport: Old and New* (Springer, New York, 2009), Vol. 338.
- [26] R. Jordan, D. Kinderlehrer, and F. Otto, *The variational formulation of the Fokker–Planck equation*, *SIAM J. Math. Anal.* **29**, 1 (1998).
- [27] E. Aurell, C. Mejía-Monasterio, and P. Muratore-Ginanneschi, *Optimal protocols and optimal transport in stochastic thermodynamics*, *Phys. Rev. Lett.* **106**, 250601 (2011).
- [28] Y. Chen, T. T. Georgiou, and A. Tannenbaum, *Stochastic control and nonequilibrium thermodynamics: Fundamental limits*, *IEEE Trans. Autom. Control* **65**, 2979 (2019).
- [29] M. Nakazato and S. Ito, *Geometrical aspects of entropy production in stochastic thermodynamics based on Wasserstein distance*, *Phys. Rev. Res.* **3**, 043093 (2021).
- [30] E. Aurell, K. Gawędzki, C. Mejía-Monasterio, R. Mohayaee, and P. Muratore-Ginanneschi, *Refined second law of thermodynamics for fast random processes*, *J. Stat. Phys.* **147**, 487 (2012).
- [31] A. Dechant, S. I. Sasa, and S. Ito, *Geometric decomposition of entropy production into excess, housekeeping, and coupling parts*, *Phys. Rev. E* **106**, 024125 (2022).
- [32] S. Ito, *Geometric thermodynamics for the Fokker–Planck equation: Stochastic thermodynamic links between information geometry and optimal transport*, *Inf. Geom.* **7**, 441 (2024).
- [33] R. Nagayama, K. Yoshimura, A. Kolchinsky, and S. Ito, *Geometric thermodynamics of reaction-diffusion systems: Thermodynamic trade-off relations and optimal transport for pattern formation*, *Phys. Rev. Res.* **7**, 033011 (2025).
- [34] J. M. Tomczak, *Deep Generative Modeling* (Springer, New York, 2022).
- [35] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, *Deep unsupervised learning using nonequilibrium thermodynamics*, in *International Conference on Machine Learning* (PMLR, 2015), pp. 2256–2265, <https://proceedings.mlr.press/v37/sohl-dickstein15.html>.
- [36] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, *Score-based generative modeling through stochastic differential equations*, in *International Conference on Learning Representations* (2020), <https://iclr.cc/virtual/2021/poster/3177>.
- [37] D. J. Evans and D. J. Searles, *The fluctuation theorem*, *Adv. Phys.* **51**, 1529 (2002).
- [38] G. E. Crooks, *Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences*, *Phys. Rev. E* **60**, 2721 (1999).
- [39] C. Jarzynski, *Nonequilibrium equality for free energy differences*, *Phys. Rev. Lett.* **78**, 2690 (1997).
- [40] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, in *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada (2020), <https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf>.
- [41] D. Kingma, T. Salimans, B. Poole, and J. Ho, *Variational diffusion models*, in *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS 2021)* (2021), <https://proceedings.neurips.cc/paper/2021/hash/b578f2a52a0229873fec2a4b06377fa-Abstract.html>.
- [42] Y. Song and S. Ermon, *Generative modeling by estimating gradients of the data distribution*, in *Proceedings of the 32th Conference on Advances in Neural Information Processing Systems (NeurIPS 2019)* (2019), <https://proceedings.neurips.cc/paper/2019/hash/3001ef257407d5a371a96dcd947c7d93-Abstract.html>.
- [43] A. Q. Nichol and P. Dhariwal, *Improved denoising diffusion probabilistic models*, in *International Conference on Machine Learning* (PMLR, 2021) pp. 8162–8171, <https://proceedings.mlr.press/v139/nichol21a.html>.
- [44] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-resolution image synthesis with latent diffusion models*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 10684–10695, https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.
- [45] Y. Song and S. Ermon, *Improved techniques for training score-based generative models*, in *Proceedings of the 33th Conference on Advances in Neural Information Processing Systems (NeurIPS 2020)* (2020), <https://proceedings.neurips.cc/paper/2020/hash/92c3b916311a5517d9290576e3ea37ad-Abstract.html>.
- [46] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, *Consistency models*, in *International Conference on Machine Learning* (PMLR, 2023), pp. 32211–32252, <https://proceedings.mlr.press/v202/song23a>.
- [47] P. Dhariwal and A. Nichol, *Diffusion models beat GANs on image synthesis*, in *Proceedings of the 34th Conference on Advances in Neural Information Processing Systems (NeurIPS 2021)* (2021), https://proceedings.neurips.cc/paper_files/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

- [48] J. Song, C. Meng, and S. Ermon, *Denoising diffusion implicit models*, in *International Conference on Learning Representations* (2020), <https://iclr.cc/virtual/2021/poster/2804>.
- [49] T. Karras, M. Aittala, T. Aila, and S. Laine, *Elucidating the design space of diffusion-based generative models*, in *Proceedings of the 35th Conference on Advances in Neural Information Processing Systems (NeurIPS 2022)* (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.
- [50] T. Chen, G.-H. Liu, and E. Theodorou, *Likelihood training of Schrödinger bridge using forward-backward SDEs theory*, in *International Conference on Learning Representations* (2021), <https://iclr.cc/virtual/2022/poster/6506>.
- [51] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, *Hierarchical text-conditional image generation with clip latents*, [arXiv:2204.06125](https://arxiv.org/abs/2204.06125).
- [52] J. Ho and T. Salimans, *Classifier-free diffusion guidance*, in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications* (2021), <https://openreview.net/forum?id=qw8AKxfYbI>.
- [53] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, *Glide: Towards photorealistic image generation and editing with text-guided diffusion models*, in *International Conference on Machine Learning* (PMLR, 2022), pp. 16784–16804, <https://proceedings.mlr.press/v162/nichol22a.html>.
- [54] A. Sauer, F. Boesel, T. Dockhorn, A. Blattmann, P. Esser, and R. Rombach, *Fast high-resolution image synthesis with latent adversarial diffusion distillation*, in *SIGGRAPH Asia 2024 Conference Papers* (2024), pp. 1–11, [arXiv:2403.12015](https://arxiv.org/abs/2403.12015).
- [55] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, *SDXL: Improving latent diffusion models for high-resolution image synthesis*, in *The Twelfth International Conference on Learning Representations* (2024), <https://iclr.cc/virtual/2024/poster/18250>.
- [56] A. Hyvärinen and P. Dayan, *Estimation of non-normalized statistical models by score matching*, *J. Mach. Learn. Res.* **6** (2005), <https://jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf>.
- [57] P. Vincent, *A connection between score matching and denoising autoencoders*, *Neural Comput.* **23**, 1661 (2011).
- [58] D. P. Kingma and Y. Cun, *Regularized estimation of image statistics by score matching*, *Adv. Neural Inf. Process. Syst.* **23**, 1126 (2010), https://papers.nips.cc/paper_files/paper/2010/hash/6f3e29a35278d71c7f65495871231324-Abstract.html.
- [59] L. Dinh, D. Krueger, and Y. Bengio, *Nice: Non-linear independent components estimation*, in *ICLR Workshop* (2015), [arXiv:1410.8516](https://arxiv.org/abs/1410.8516).
- [60] D. Rezende and S. Mohamed, *Variational inference with normalizing flows*, in *International Conference on Machine Learning* (PMLR, 2015), pp. 1530–1538, <http://proceedings.mlr.press/v37/rezende15.pdf>.
- [61] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, in *Proceedings of the Advances in Neural Information Processing Systems 31 (NeurIPS 2018)* (2018), <https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html>.
- [62] K.-U. Song, *Applying regularized Schrödinger-bridge-based stochastic process in generative modeling*, [arXiv:2208.07131](https://arxiv.org/abs/2208.07131).
- [63] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, *Flow matching for generative modeling*, in *The Eleventh International Conference on Learning Representations* (2022), <https://iclr.cc/virtual/2023/poster/11309>.
- [64] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein generative adversarial networks*, in *International Conference on Machine Learning* (PMLR, 2017) pp. 214–223, <https://proceedings.mlr.press/v70/arjovsky17a.html>.
- [65] D. Kwon, Y. Fan, and K. Lee, *Score-based generative modeling secretly minimizes the Wasserstein distance*, in *Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/7f52f6b8f107931127eefe15429ee278-Abstract-Conference.html.
- [66] K. Oko, S. Akiyama, and T. Suzuki, *Diffusion models are minimax optimal distribution estimators*, in *International Conference on Machine Learning* (PMLR, 2023), pp. 26517–26582, <https://proceedings.mlr.press/v202/oko23a.html>.
- [67] V. D. Bortoli, *Convergence of denoising diffusion models under the manifold hypothesis*, *Trans. Mach. Learn. Res.* (2022), <https://openreview.net/forum?id=MhK5aXo3gB>.
- [68] N. M. Kornilov, P. Mokrov, A. Gasnikov, and A. Korotin, *Optimal flow matching: Learning straight trajectories in just one step*, in *The Thirty-Eighth Annual Conference on Neural Information Processing Systems* (2024), https://proceedings.neurips.cc/paper_files/paper/2024/hash/bc8f76d9caadd48f77025b1c889d2e2d-Abstract-Conference.html.
- [69] N. Shaul, R. T. Chen, M. Nickel, M. Le, and Y. Lipman, *On kinetic optimal probability paths for generative models*, in *International Conference on Machine Learning* (PMLR, 2023), pp. 30883–30907, <https://proceedings.mlr.press/v202/shaul23a.html>.
- [70] A. Tong, N. Malkin, G. Hugué, Y. Zhang, J. Rector-Brooks, K. Fatras, G. Wolf, and Y. Bengio, *Improving and generalizing flow-based generative models with minibatch optimal transport*, in *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems* (2023), <https://openreview.net/forum?id=CD9Snc73AW>.
- [71] Y. Shi, V. De Bortoli, G. Deligiannidis, and A. Doucet, *Conditional simulation using diffusion Schrödinger bridges*, in *Uncertainty in Artificial Intelligence* (PMLR, 2022), pp. 1792–1802, <https://proceedings.mlr.press/v180/shi22a.html>.
- [72] X. Liu, C. Gong, and qiang liu, *Flow straight and fast: Learning to generate and transfer data with rectified flow*, in *The Eleventh International Conference on Learning Representations* (2023), <https://iclr.cc/virtual/2023/poster/11266>.
- [73] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, *Scaling rectified flow transformers for high-resolution image synthesis*, in *Forty-First International Conference on Machine Learning*

- ing (2024), <https://proceedings.mlr.press/v235/esser24a.html>.
- [74] J. Klinger and G. M. Rotskoff, *Universal energy-speed-accuracy trade-offs in driven nonequilibrium systems*, *Phys. Rev. E* **111**, 014114 (2025).
- [75] A. Brock, J. Donahue, and K. Simonyan, *Large scale GAN training for high fidelity natural image synthesis*, in *International Conference on Learning Representations* (2019), <https://iclr.cc/Conferences/2019/Schedule?showEvent=1152>.
- [76] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, *Language models are few-shot learners*, in *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020)* (2020), <https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html>.
- [77] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, *Wavenet: A generative model for raw audio*, [arXiv:1609.03499](https://arxiv.org/abs/1609.03499).
- [78] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).
- [79] L. Theis, A. v. d. Oord, and M. Bethge, *A note on the evaluation of generative models*, [arXiv:1511.01844](https://arxiv.org/abs/1511.01844).
- [80] E. Betzalel, C. Penso, A. Navon, and E. Fetaya, *A study on the evaluation of generative models*, [arXiv:2206.10935](https://arxiv.org/abs/2206.10935).
- [81] S. Bischoff, A. Darcher, M. Deistler, R. Gao, F. Gerken, M. Gloeckler, L. Haxel, J. Kapoor, J. K. Lappalainen, J. H. Macke *et al.*, *A practical guide to statistical distances for evaluating generative models in science*, *Trans. Mach. Learn. Res.* (2024), <https://openreview.net/forum?id=isEFziui9p>.
- [82] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, *GANs trained by a two time-scale update rule converge to a local nash equilibrium*, *Adv. Neural Inf. Process. Syst.* **30** (2017), [arXiv:1706.08500](https://arxiv.org/abs/1706.08500).
- [83] S.-i. Amari, *Information Geometry and its Applications* (Springer, New York, 2016), Vol. 194.
- [84] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, *Dreamfusion: Text-to-3D using 2D diffusion*, in *Proceedings of the Eleventh International Conference on Learning Representations* (2023), <https://iclr.cc/virtual/2023/poster/10961>.
- [85] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, and J. Tang, *GeoDiff: A geometric diffusion model for molecular conformation generation*, in *Proceedings of the International Conference on Learning Representations* (2022), <https://iclr.cc/virtual/2022/poster/7028>.
- [86] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick *et al.*, *Accurate structure prediction of biomolecular interactions with ALPHAFOLD3*, *Nature (London)* **630**, 493 (2024).
- [87] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, *Imagen video: High definition video generation with diffusion models*, [arXiv:2210.02303](https://arxiv.org/abs/2210.02303).
- [88] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, *Video generation models as world simulators* (2024), <https://openai.com/research/video-generation-models-as-world-simulators>.
- [89] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, *Wavegrad: Estimating gradients for waveform generation*, in *International Conference on Learning Representations* (2020), <https://iclr.cc/virtual/2021/poster/3220>.
- [90] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, *Diffwave: A versatile diffusion model for audio synthesis*, in *International Conference on Learning Representations* (2020), <https://iclr.cc/virtual/2021/poster/2979>.
- [91] *Smithsonian Butterflies Subset*, https://huggingface.co/datasets/huggan/smithsonian_butterflies_subset (accessed: 2024-02-09).
- [92] H. Risken and H. Risken, *Fokker-Planck Equation* (Springer, New York, 1996).
- [93] N. G. Van Kampen, *Stochastic differential equations*, *Phys. Rep.* **24**, 171 (1976).
- [94] S. Brooks, *Markov chain Monte Carlo method and its application*, *J. R. Stat. Soc.* **47**, 69 (1998).
- [95] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, *An introduction to MCMC for machine learning*, *Mach. Learn.* **50**, 5 (2003).
- [96] M. Welling and Y. W. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*, in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (Citeseer, 2011), pp. 681–688.
- [97] T. Chen, *On the importance of noise scheduling for diffusion models*, [arXiv:2301.10972](https://arxiv.org/abs/2301.10972).
- [98] B. D. Anderson, *Reverse-time diffusion equation models*, *Stoch. Proc. Appl.* **12**, 313 (1982).
- [99] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, *DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps*, in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/260a14acce2a89dad36adc8eefc7c59e-Abstract-Conference.html.
- [100] C. Lu, K. Zheng, F. Bao, J. Chen, C. Li, and J. Zhu, *Maximum likelihood training for score-based diffusion ODEs by high order denoising score matching*, in *International Conference on Machine Learning (PMLR, 2022)*, pp. 14429–14460, <https://proceedings.mlr.press/v162/lu22f.html?ref=https://githubhelp.com>.
- [101] Q. Zhang and Y. Chen, *Fast sampling of diffusion models with exponential integrator*, in *The Eleventh International Conference on Learning Representations* (2023), <https://iclr.cc/virtual/2023/poster/10904>.
- [102] I. Kobyzev, S. J. Prince, and M. A. Brubaker, *Normalizing flows: An introduction and review of current methods*, *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 3964 (2020).
- [103] S. Ito, M. Oizumi, and S. I. Amari, *Unified framework for the entropy production and the stochastic interaction based on information geometry*, *Phys. Rev. Res.* **2**, 033048 (2020).
- [104] R. Kawai, J. M. R. Parrondo, and C. Van den Broeck, *Dissipation: The phase-space perspective*, *Phys. Rev. Lett.* **98**, 080602 (2007).
- [105] C. Villani, *Topics in Optimal Transportation* (American Mathematical Society, Providence, 2021), Vol. 58.

- [106] C. R. Givens and R. M. Shortt, *A class of Wasserstein metrics for probability distributions*, *Mich. Math. J.* **31**, 231 (1984).
- [107] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 1–9, https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html.
- [108] M. Gelbrich, *On a formula for the l2 Wasserstein metric between measures on Euclidean and Hilbert spaces*, *Mathematische Nachrichten* **147**, 185 (1990).
- [109] J.-D. Benamou and Y. Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, *Num. Math.* **84**, 375 (2000).
- [110] A. Dechant, S. I. Sasa, and S. Ito, *Geometric decomposition of entropy production in out-of-equilibrium systems*, *Phys. Rev. Res.* **4**, L012034 (2022).
- [111] J.-D. Benamou and Y. Brenier, *A numerical method for the optimal time-continuous mass transport problem and related problems*, *Contemp. Math.* **226**, 1 (1999).
- [112] D. Sekizawa, S. Ito, and M. Oizumi, *Decomposing thermodynamic dissipation of neural dynamics via spatio-temporal oscillatory modes*, *Phys. Rev. X* **14**, 041003 (2024).
- [113] G. E. Crooks, *Path-ensemble averages in systems driven far from equilibrium*, *Phys. Rev. E* **61**, 2361 (2000).
- [114] J. M. Horowitz and T. R. Gingrich, *Thermodynamic uncertainty relations constrain non-equilibrium fluctuations*, *Nat. Phys.* **16**, 15 (2020).
- [115] G. Lan, P. Sartori, S. Neumann, V. Sourjik, and Y. Tu, *The energy–speed–accuracy trade-off in sensory adaptation*, *Nat. Phys.* **8**, 422 (2012).
- [116] K. Pearson, *On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling*, *London Edinburgh Dublin Phil. Mag. J. Sci.* **50**, 157 (1900).
- [117] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, *Improved training of Wasserstein GANs*, in *Proceedings of the Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), <https://proceedings.neurips.cc/paper/2017/hash/892c3b1c6dccc52936e27cbd0ff683d6-Abstract.html>.
- [118] Q. Dao, H. Phung, B. Nguyen, and A. Tran, *Flow matching in latent space*, *arXiv:2307.08698*.
- [119] T. Karras, T. Aila, S. Laine, and J. Lehtinen, *Progressive growing of GANs for improved quality, stability, and variation*, in *International Conference on Learning Representations* (2018), <https://openreview.net/forum?id=Hk99zCeAb>.
- [120] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, *Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop*, *arXiv:1506.03365*.
- [121] A. C. Barato and U. Seifert, *Thermodynamic uncertainty relation for biomolecular processes*, *Phys. Rev. Lett.* **114**, 158101 (2015).
- [122] A. A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Chen, *Multisample flow matching: Straightening flows with minibatch couplings*, *Proc. Mach. Learn. Res.* **202**, 28100 (2023), <https://proceedings.mlr.press/v202/pooladian23a.html>.
- [123] K. Fukumizu, T. Suzuki, N. Isobe, K. Oko, and M. Koyama, *Flow matching achieves almost minimax optimal convergence*, in *The Thirteenth International Conference on Learning Representations* (2025), <https://openreview.net/forum?id=2OMyAFjiJJ>.
- [124] Y. Song, C. Durkan, I. Murray, and S. Ermon, *Maximum likelihood training of score-based diffusion models*, in *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (2021), <https://proceedings.neurips.cc/paper/2021/hash/0a9fdbb17feb6ccb7ec405cfb85222c4-Abstract.html>.
- [125] H. Lee, J. Lu, and Y. Tan, *Convergence for score-based generative modeling with polynomial complexity*, in *Proceedings of the Advances in Neural Information Processing Systems 35 (NeurIPS 2022)* (2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/8ff87c96935244b63503f542472462b3-Abstract-Conference.html.
- [126] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang, *Sampling is as easy as learning the score: Theory for diffusion models with minimal data assumptions*, in *International Conference on Learning Representations* (2023), <https://iclr.cc/virtual/2023/poster/10972>.
- [127] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, *Improved techniques for training GANs*, in *Proceedings of the Advances in Neural Information Processing Systems 29 (NIPS 2016)* (2016), https://papers.nips.cc/paper_files/paper/2016/hash/8a3363abe792db2d8761d6403605aeb7-Abstract.html.
- [128] M. S. Albergo and E. Vanden-Eijnden, *Building normalizing flows with stochastic interpolants*, in *The Eleventh International Conference on Learning Representations* (2023), <https://iclr.cc/virtual/2023/poster/10866>.
- [129] S.-I. Amari, *Learning patterns and pattern sequences by self-organizing nets of threshold elements*, *IEEE Trans. Comput.* **100**, 1197 (1972).
- [130] J. J. Hopfield, *Neural networks and physical systems with emergent collective computational abilities.*, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554 (1982).
- [131] G. E. Hinton, T. J. Sejnowski, and D. H. Ackley, *Boltzmann Machines: Constraint Satisfaction Networks That Learn* (Carnegie-Mellon University, Department of Computer Science, Pittsburgh, PA, 1984).
- [132] R. Salakhutdinov and G. Hinton, *Deep Boltzmann machines*, in *Artificial Intelligence and Statistics (PMLR, 2009)*, pp. 448–455, <http://proceedings.mlr.press/v5/salakhutdinov09a>.
- [133] G. Hinton, *A practical guide to training restricted Boltzmann machines*, *Momentum* **9**, 926 (2010).
- [134] J. Schnakenberg, *Network theory of microscopic and macroscopic behavior of master equation systems*, *Rev. Mod. Phys.* **48**, 571 (1976).
- [135] H. Qian, *Relative entropy: Free energy associated with equilibrium fluctuations and nonequilibrium deviations*, *Phys. Rev. E* **63**, 042103 (2001).
- [136] G. Biroli, T. Bonnaire, V. De Bortoli, and M. Mézard, *Dynamical regimes of diffusion models*, *Nat. Commun.* **15**, 9957 (2024).

- [137] A. Dechant, *Minimum entropy production, detailed balance and Wasserstein distance for continuous-time Markov processes*, *J. Phys. A* **55**, 094001 (2022).
- [138] K. Yoshimura, A. Kolchinsky, A. Dechant, and S. Ito, *Housekeeping and excess entropy production for general nonlinear dynamics*, *Phys. Rev. Res.* **5**, 013017 (2023).
- [139] T. Van Vu and K. Saito, *Thermodynamic unification of optimal transport: Thermodynamic uncertainty relation, minimum dissipation, and thermodynamic speed limits*, *Phys. Rev. X* **13**, 011013 (2023).
- [140] A. Kolchinsky, A. Dechant, K. Yoshimura, and S. Ito, *Information geometry of excess and housekeeping entropy production*, [arXiv:2206.14599](https://arxiv.org/abs/2206.14599).
- [141] G. Wang, Y. Jiao, Q. Xu, Y. Wang, and C. Yang, *Deep generative learning via Schrödinger bridge*, in *International Conference on Machine Learning* (PMLR, 2021), pp. 10794–10804, <https://proceedings.mlr.press/v139/wang211.html>.
- [142] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet, *Diffusion Schrödinger bridge with applications to score-based generative modeling*, in *Proceedings of the Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (2021), https://proceedings.neurips.cc/paper_files/paper/2021/hash/940392f5f32a7ade1cc201767cf83e31-Abstract.html.
- [143] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier *et al.*, *POT: PYTHON optimal transport*, *J. Mach. Learn. Res.* **22**, 1 (2021), <http://jmlr.org/papers/v22/20-451.html>.
- [144] D. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in *International Conference on Learning Representations* (2015), [arXiv:1412.6980](https://arxiv.org/abs/1412.6980).