Variational Schrödinger Momentum Diffusion

Kevin Rojas[*1], **Yixin Tan**[*2], **Molei Tao**[1], **Yuriy Nevmyvaka**[3], **Wei Deng**[3]
[1]Georgia Institute of Technology   [2]Duke University   [3]Morgan Stanley

Abstract

The Momentum Schrödinger Bridge (mSB) (Chen et al., 2023c) has emerged as a leading method for accelerating generative diffusion processes and reducing transport costs. However, the lack of simulation-free properties inevitably results in high training costs and affects scalability. To obtain a trade-off between transport properties and scalability, we introduce variational Schrödinger momentum diffusion (VSMD), which employs linearized forward score functions (variational scores) to eliminate the dependence on simulated forward trajectories. Our approach leverages a multivariate diffusion process with adaptively transport-optimized variational scores. Additionally, we apply a critical-damping transform to stabilize training by removing the need for score estimations for both velocity and samples. Theoretically, we prove the convergence of samples generated with optimal variational scores and momentum diffusion. Empirical results demonstrate that VSMD efficiently generates anisotropic shapes while maintaining transport efficacy, outperforming overdamped alternatives, and avoiding complex denoising processes. Our approach also scales effectively to real-world data, achieving competitive results in time series and image generation. Code can be found in: https://github.com/KevinRojas1499/Variational-Momentum-SB.

---
[*]Equal contribution. K. Rojas conducted this work during his internship at Morgan Stanley. Correspondence: Wei Deng - weideng056@gmail.com.

1 Introduction

Score-based generative models (SGMs) have become the preferred method for generative modeling, showcasing exceptional capabilities in generating images, videos, and audios (Dhariwal and Nichol, 2022; Ho et al., 2022; Kong et al., 2021; Ramesh et al., 2022). To improve efficiency and simplify the denoising process, critically-damped Langevin diffusion (CLD) (Dockhorn et al., 2022) leverages kinetic (second-order) Langevin dynamics (Dalalyan and Riou-Durand, 2020) by incorporating auxiliary velocity variables, resulting in well-behaved score functions at the boundary. While both SGMs and CLDs offer scalability benefits and simulation-free properties, they lack guaranteed optimal transport (OT) properties (Lavenant and Santambrogio, 2022) and often involve costly evaluations to produce high-quality content (Ho et al., 2020; Salimans and Ho, 2022; Lu et al., 2022).

In contrast, the Momentum Schrödinger Bridge (mSB) (Chen and Georgiou, 2016; Pavon et al., 2021; Caluya and Halder, 2022; De Bortoli et al., 2021; Chen et al., 2023c) focuses on optimizing a stochastic control objective to achieve entropic optimal transport. The extension of forward-backward stochastic differential equations (FB-SDEs) (Chen et al., 2022) with velocity variables not only accelerates the processes but also simplifies the denoising process and lowers tuning costs. However, training the intractable forward score functions for optimal transport relies heavily on simulated trajectories and often requires an additional pipeline using SGMs or CLDs for warm-up training to scale up to real-world data (De Bortoli et al., 2021; Chen et al., 2022). This prompts a critical question: How can we efficiently train momentum diffusion models from scratch while maintaining effective transport?

To address these challenges, we propose the Variational Schrödinger Momentum Diffusion (VSMD) model. Inspired by Deng et al. (2024b), we adopt locally linearized variational scores using variational inference to restore simulation-free properties for training backward scores. Additionally, we introduce a critical-damping transform to simplify and stabilize training by reduc-
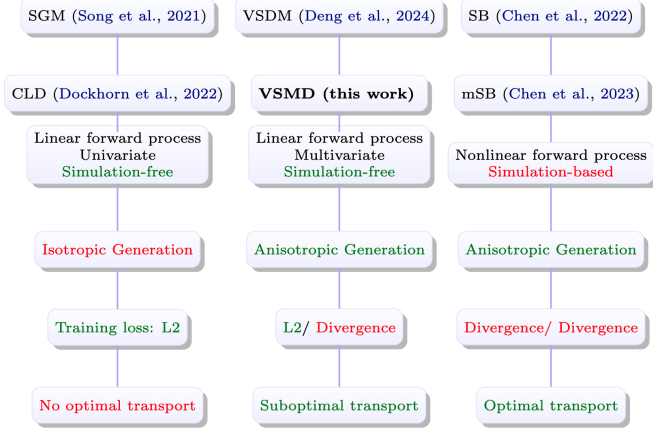
Figure 1: Comparison with existing methodologies and algorithm properties.

ing the need to estimate two variational scores associated with both velocity and samples. Unlike the single-variate CLD model, VSMD functions as an adaptively transport-optimized multivariate diffusion (Singhal et al., 2023), facilitating efficient training, a simplified denoising process, and effective transport (Chen et al., 2022). Our contributions are highlighted in three key aspects and presented in Figure 1:

- We introduce the Variational Schrödinger Momentum Diffusion (VSMD), an adaptive multivariate diffusion with simulation-free properties. We derive a tailored critical-damping rule to streamline training by avoiding the complexity of estimating additional variational scores.

- Theoretically, we identify the convergence of the adaptively transport-optimized multivariate diffusion using techniques from stochastic approximation (Robbins and Monro, 1951) and stochastic differential equations.

- VSMD surpasses its overdamped counterparts by leveraging momentum accelerations and avoiding complex denoising processes. It demonstrates strong performance in conditional and unconditional generations in both images and time series data, all while eliminating the need for warm-up initializations.

2 Related Works

Schrödinger Bridge (SB) Problems Dynamic SB solvers for high-dimensional problems were initially introduced by De Bortoli et al. (2021); Chen et al. (2022); Vargas et al. (2021); Wang et al. (2021); Chen et al. (2023d); Deng et al. (2024a) to promote smoother

trajectories with optimal transport properties. Subsequent work by Shi et al. (2023); Peluchetti (2023); Chen et al. (2024) enhanced performance by preserving marginal distributions and simplifying objectives inspired by bridge matching (Liu, 2022) and flow matching (Lipman et al., 2023), which were further extended to general cost functions (Neklyudov et al., 2024; Liu et al., 2022). To improve scalability, De Bortoli et al. (2024) proposed an online scheme to avoid caching samples and maintaining two networks. However, the need for simulations still limits scalability, highlighting the ongoing demand for more scalable methods.

Simulation-free Generative Models Lipman et al. (2023) proposed a simulation-free paradigm to train continuous normalizing flows (Chen et al., 2018), inherently connected to the OT displacement map (McCann, 1997). Tong et al. (2024); Pooladian et al. (2023); Eyring et al. (2024) advanced this field by leveraging minibatch OT objectives, approximations from discrete Sinkhorn solvers, non-independent couplings from minibatch data, and unbalanced Monge map estimators. Liu (2022); Liu et al. (2023) introduced methods to rectify non-smooth trajectories and provide theoretical guarantees with convex cost functions. Albergo and Vanden-Eijnden (2023); Albergo et al. (2023) elegantly unified flow and diffusion models in a simulation-free manner. Somnath et al. (2023) addressed data alignment issues, while Kim et al. (2024); Gushchin et al. (2023) employed adversarial objectives to optimize OT losses, although they still do not yield OT maps. Korotin et al. (2024) achieved simulation-free properties on small-scale problems by parameterizing the Schrödinger bridge potentials with Gaussian mixture distributions. Bartosh et al. (2024) supported a broader family of forward diffusion and also introduced extra complexities. However, achieving simulation-free properties often requires sacrifices in OT properties, underscoring the need for more efficient schemes.

3 Preliminaries

SGMs: Score-based generative models (SGMs) (Song et al., 2021) have achieved unprecedented success in generative models. SGMs propose reversing a diffusion process to generate data distributions (Anderson, 1982). However, the simplicity of the forward diffusion process, such as Brownian motion or the Ornstein-Uhlenbeck process, results in a complex denoising process that requires extensive tuning to generalize across different datasets.

CLD: To address these challenges, critically-damped Langevin diffusion (CLD) has been proposed to accelerate diffusion by augmenting data $\mathbf{x}_t \in \mathbb{R}^d$ with velocity

variables $\mathbf{v}_t \in \mathbb{R}^d$ motivated by Hamiltonian dynamics (Neal, 2012):

$$\begin{pmatrix} \mathrm{d}\mathbf{x}_t \\ \mathrm{d}\mathbf{v}_t \end{pmatrix} = \frac{\beta}{2} \begin{pmatrix} \mathbf{v}_t \\ -\mathbf{x}_t - \gamma\mathbf{v}_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \mathbf{0}_d \\ \sqrt{\beta\gamma}\mathbf{I}_d \end{pmatrix} \mathrm{d}\mathbf{w}_t, \quad (1)$$

where γ is the friction coefficient that controls the randomness, $\mathbf{w}_t$ is the standard Brownian motion in $\mathbb{R}^{2d}$. In the long-time limit, the invariant distribution of continuous-time process (1) is a joint Gaussian distribution $\mathrm{N}(\mathbf{x}; 0, \mathbf{I}_d)\mathrm{N}(\mathbf{v}; 0, \mathbf{I}_d)$.

Damping Regimes: The choices of γ correspond to different damping regimes of Langevin dynamics (McCall, 2011; Dockhorn et al., 2022). For high friction with $\gamma > 2$, it leads to overdamped Langevin dynamics (LD) with straighter trajectories, however, the convergence speed is also impeded. In contrast, lower friction yields more oscillating trajectories and accelerates convergence. The dynamics with $\gamma = 2$ are termed the critical-damped Langevin diffusion (CLD) while $\gamma < 2$ corresponds to underdamped Langevin diffusion (ULD). Theoretically, CLD provides a balance between oscillation and speed, though in practice, different damping may need to be selected for optimal trade-off.

The Effect of Friction on Convergence: The impact of the friction γ on convergence speed is well understood. In particular, LD requires $\Omega(d/\epsilon^2)$ iterations to achieve an ϵ error in 2-Wasserstein (W2) distance for strongly log-concave distributions, whereas ULD requires only $\Omega(\sqrt{d}/\epsilon)$ iterations to achieve the same accuracy (Cheng et al., 2017). Additional literature supporting the speed advantage of employing Hamiltonian dynamics can be found in Mangoubi and Vishnoi (2018); Dalalyan and Riou-Durand (2020); Mangoubi and Smith (2021).

4 Variational Schrödinger Momentum Diffusion

4.1 Momentum Schrödinger Bridge

The momentum Schrödinger bridge (mSB) (Caluya and Halder, 2022; Chen et al., 2023c) can be interpreted as a stochastic optimal control (SOC) objective with optimal transport guarantees (Chen et al., 2021):

$$\inf_{\mathbf{u} \in \mathcal{U}} \mathbb{E}\left\{ \int_0^T \frac{1}{2}\|\mathbf{u}(\overrightarrow{\mathbf{a}}, t)\|_2^2 \mathrm{d}t \right\}$$

$$\text{s.t. } \mathrm{d}\overrightarrow{\mathbf{a}}_t = \left[\mathbf{f}(\overrightarrow{\mathbf{a}}_t) + \mathbf{g}\mathbf{u}(\overrightarrow{\mathbf{a}}_t, t) \right] \mathrm{d}t + \mathbf{g}\mathrm{d}\overrightarrow{\mathbf{w}_t} \quad (2)$$

$$\mathbf{f}(\overrightarrow{\mathbf{a}}) := -\frac{\beta}{2}\left(\begin{pmatrix} 0 & -1 \\ 1 & \gamma \end{pmatrix} \otimes \mathbf{I}_d \right)\overrightarrow{\mathbf{a}}, \ \mathbf{g} := \sqrt{\beta\gamma}\mathbf{J}_{2d},$$

$$\overrightarrow{\mathbf{a}}_0 \sim \rho_0 := p_{\text{data}} \otimes p_{\mathbf{v}_0}, \ \overrightarrow{\mathbf{a}}_T \sim \rho_T := p_{\text{prior}} \otimes p_{\mathbf{v}_T},$$

where $\mathbf{J}_{2d} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \otimes \mathbf{I}_d$, $\overrightarrow{\mathbf{a}} = \begin{pmatrix} \overrightarrow{\mathbf{x}} \\ \overrightarrow{\mathbf{v}} \end{pmatrix} \in \mathbb{R}^{2d}$ is the augmented variable; $\mathbf{u} : \mathbb{R}^{2d} \times [0, T] \to \mathbb{R}^{2d}$ is a control; $\mathbf{f} := \mathbb{R}^{2d} \times [0, T] \to \mathbb{R}^{2d}$ is a vector field. The probability density function (PDF) for the process (2) is denoted by $\overrightarrow{\rho}(\cdot, t)$. We fix $p_{\mathbf{v}_0}$ and $p_{\mathbf{v}_T}$ as the standard Gaussian distribution $\mathrm{N}(0, \mathbf{I})$.

The Lagrangian of Eq.(2) leads to the Hamilton–Jacobi–Bellman (HJB) equation (Caluya and Halder, 2022; Chen et al., 2023c); applying the Hopf-Cole transform, we can solve the *Schrödinger system* via the backward-forward Kolmogorov equations

$$\begin{cases} \frac{\partial \overrightarrow{\psi}}{\partial t} + \langle \nabla \overrightarrow{\psi}, \mathbf{f} \rangle + \frac{1}{2}\mathbf{g}\mathbf{g}^\mathsf{T}\Delta \overrightarrow{\psi} = 0 \\ \frac{\partial \overleftarrow{\varphi}}{\partial t} + \nabla \cdot (\overleftarrow{\varphi}\mathbf{f}) - \frac{1}{2}\mathbf{g}\mathbf{g}^\mathsf{T}\Delta \overleftarrow{\varphi} = 0, \end{cases}$$

$$\text{s.t. } \overrightarrow{\psi}(\mathbf{x}, 0)\overleftarrow{\varphi}(\mathbf{x}, 0) = \rho_0, \quad \overrightarrow{\psi}(\mathbf{y}, T)\overleftarrow{\varphi}(\mathbf{y}, T) = \rho_T.$$

Considering the stochastic representation for the forward Kolmogorov equation and the time reversal (Anderson, 1982), we have the forward-backward stochastic differential equation (FB-SDE) Chen et al. (2023c):

$$\mathrm{d}\overrightarrow{\mathbf{a}}_t = \left[\mathbf{f}(\overrightarrow{\mathbf{a}}_t, t) + \mathbf{g}\mathbf{g}^\mathsf{T} \begin{pmatrix} \mathbf{0} \\ \nabla_\mathbf{v} \log \overrightarrow{\psi}(\overrightarrow{\mathbf{a}}_t, t) \end{pmatrix} \right] \mathrm{d}t$$
$$+ \mathbf{g}\mathrm{d}\overrightarrow{\mathbf{w}_t}, \quad \overrightarrow{\mathbf{a}}_0 \sim \rho_0, \quad (3a)$$

$$\mathrm{d}\overleftarrow{\mathbf{a}}_t = \left[\mathbf{f}(\overleftarrow{\mathbf{a}}_t, t) - \mathbf{g}\mathbf{g}^\mathsf{T} \begin{pmatrix} \mathbf{0} \\ \nabla_\mathbf{v} \log \overleftarrow{\varphi}(\overleftarrow{\mathbf{a}}_t, t) \end{pmatrix} \right] \mathrm{d}t$$
$$+ \mathbf{g}\mathrm{d}\overleftarrow{\mathbf{w}_t}, \quad \overleftarrow{\mathbf{a}}_T \sim \rho_T. \quad (3b)$$

We can next solve $(\overrightarrow{\psi}, \overleftarrow{\varphi})$ for the augmented variable $\mathbf{a} = \begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix}$ to the *Schrödinger system* by the nonlinear Feynman-Kac formula (Ma and Yong, 2007; Chen et al., 2022, 2023c):

Proposition 1 (Feynman-Kac formula). *Given $\beta, \gamma > 0$, the stochastic representation of the solution follows*

$$\overleftarrow{y}_s = \mathbb{E}\left[\overleftarrow{y}_T - \int_s^T \Gamma_\zeta(\overleftarrow{\mathbf{z}}_t; \overrightarrow{\mathbf{z}}_t)\mathrm{d}t \middle| \overrightarrow{\mathbf{x}}_s = \boldsymbol{x}_s \right],$$

$$\Gamma_\zeta(\overleftarrow{\mathbf{z}}_t; \overrightarrow{\mathbf{z}}_t) \equiv \frac{1}{2}\|\overleftarrow{\mathbf{z}}_t\|_2^2 + \nabla_\mathbf{v} \cdot (\sqrt{\beta}\overleftarrow{\mathbf{z}}_t - \mathbf{f}_t) + \zeta\overleftarrow{\mathbf{z}}_t^\mathsf{T}\overrightarrow{\mathbf{z}}_t,$$
$$(4)$$

where $\overrightarrow{y}_t = \log \overrightarrow{\psi}(\mathbf{a}_t, t)$ and $\overleftarrow{y}_t = \log \overleftarrow{\varphi}(\mathbf{a}_t, t)$, $\overrightarrow{\mathbf{z}}_t = \sqrt{\beta}\nabla_\mathbf{v} \overrightarrow{y}_t$, $\overleftarrow{\mathbf{z}}_t = \sqrt{\beta}\nabla_\mathbf{v}\overleftarrow{y}_t$, and $\zeta = 1$.

4.2 Linear Approximation via Multivariate Momentum Diffusion

Consider a linear approximation of the forward process (3a) with a fixed matrix $\mathbf{A}_{\mathbf{a},t} = \begin{pmatrix} \mathbf{0}_d & \mathbf{0}_d \\ \mathbf{A}_{\mathbf{x},t} & \mathbf{A}_{\mathbf{v},t} \end{pmatrix} \in$

$\mathbb{R}^{2d \times 2d}$ (referred to as the variational score):

$$\mathrm{d}\overrightarrow{\mathbf{a}}_t = \left[\mathbf{f}(\overrightarrow{\mathbf{a}}_t, t) + \mathbf{g}\mathbf{g}^\mathsf{T}\mathbf{A}_{\mathbf{a},t}\overrightarrow{\mathbf{a}}_t\right]\mathrm{d}t + \mathbf{g}\mathrm{d}\overrightarrow{\mathbf{w}}_t$$

$$= -\frac{1}{2}\mathbf{D}_t\beta\overrightarrow{\mathbf{a}}_t\mathrm{d}t + \mathbf{g}\mathrm{d}\overrightarrow{\mathbf{w}}_t \qquad (5)$$

$$\mathbf{D}_t = \begin{pmatrix} 0 & -1 \\ 1 & \gamma \end{pmatrix} \otimes \mathbf{I}_d - 2\gamma\mathbf{A}_{\mathbf{a},t},$$

where $\mathbf{I} - 2\gamma\mathbf{A}_{\mathbf{x},t}$ and $\mathbf{I} - 2\mathbf{A}_{\mathbf{v},t}$ are restricted to a positive-definite matrix.

The mean and covariance of the augmented linear SDE (5) follow that (Särkkä and Solin, 2019)

$$\frac{\mathrm{d}\boldsymbol{\mu}_{t|0}}{\mathrm{d}t} = -\frac{1}{2}\beta\mathbf{D}_t\boldsymbol{\mu}_{t|0} \qquad (6a)$$

$$\frac{\mathrm{d}\boldsymbol{\Sigma}_{t|0}}{\mathrm{d}t} = -\frac{1}{2}\beta\big(\mathbf{D}_t\boldsymbol{\Sigma}_{t|0} + \boldsymbol{\Sigma}_{t|0}\mathbf{D}_t^\mathsf{T}\big) + \beta\gamma\mathbf{J}_{2d}, \qquad (6b)$$

where $\mathbf{D}_t$ and $\mathbf{J}_{2d}$ are defined in Eq.(5) and (2), respectively. Solving the mean process leads to the solution:

$$\boldsymbol{\mu}_{t|0} = e^{-\frac{1}{2}\beta[\mathbf{D}]_t}\mathbf{x}_0, \qquad (7)$$

where $[\mathbf{D}]_t = \int_0^t \mathbf{D}_s\mathrm{d}s$. The covariance process is a differential Lyapunov matrix equation (Särkkä and Solin, 2019) and can be solved by decomposing $\boldsymbol{\Sigma}_{t|0}$ as $\mathbf{C}_t\mathbf{H}_t^{-1}$, where $\mathbf{C}_t$ and $\mathbf{H}_t$ follow that:

$$\begin{pmatrix} \mathbf{C}_t \\ \mathbf{H}_t \end{pmatrix} = \exp\left[\begin{pmatrix} -\frac{1}{2}\beta[\mathbf{D}]_t & \gamma\beta[\mathbf{J}_{2d}]_t \\ \mathbf{0} & \frac{1}{2}\beta[\mathbf{D}^\mathsf{T}]_t \end{pmatrix} \right] \begin{pmatrix} \boldsymbol{\Sigma}_0 \\ \mathbf{I}_{2d} \end{pmatrix}. \quad (8)$$

Additional speed-ups can be achieved on real-world datasets by avoiding the matrix exponential through the use of a time-invariant and diagonal $\mathbf{D}$, as detailed in Appendix A of Deng et al. (2024b).

Next, we can achieve the simulation-free update of the multivariate momentum diffusion as follows

$$\overrightarrow{\mathbf{a}}_t = \boldsymbol{\mu}_{t|0} + \mathbf{L}_t\boldsymbol{\epsilon}, \qquad (9)$$

where $\boldsymbol{\mu}_{t|0} \sim (7)$, $\mathbf{L}_t$ is a lower-triangular matrix that satisfies $\mathbf{L}_t\mathbf{L}_t^\mathsf{T} = \boldsymbol{\Sigma}_{t|0}$, and $\boldsymbol{\epsilon} \in \mathbb{R}^{2d}$ is a Gaussian vector. The forward PDF follows that

$$\overrightarrow{\rho}_{t|0}(\overrightarrow{\mathbf{a}}_t) \propto \exp\left\{ -\frac{1}{2}(\overrightarrow{\mathbf{a}}_t - \boldsymbol{\mu}_{t|0})^\mathsf{T}\boldsymbol{\Sigma}_{t|0}^{-1}(\overrightarrow{\mathbf{a}}_t - \boldsymbol{\mu}_{t|0}) \right\},$$

which leads to a score function as follows

$$\nabla\log\overrightarrow{\rho}_{t|0}(\overrightarrow{\mathbf{a}}_t) = -\boldsymbol{\Sigma}_{t|0}^{-1}(\overrightarrow{\mathbf{a}}_t - \boldsymbol{\mu}_t) = -\mathbf{L}_t^{-\mathsf{T}}\boldsymbol{\epsilon}. \quad (10)$$

We next resort to a neural network parametrization $s_t(\cdot)$ via the following loss function to learn the score:

$$\nabla_\theta\| - \mathbf{L}_t^{-\mathsf{T}}\boldsymbol{\epsilon} - s_t(\overrightarrow{\mathbf{a}}_t)\|_2^2. \qquad (11)$$

4.2.1 Backward SDE

The backward process follows by taking the time reverse Anderson (1982) of the forward process (5):

$$\mathrm{d}\overleftarrow{\mathbf{a}} = -\frac{1}{2}\mathbf{D}_t\beta\overleftarrow{\mathbf{a}}_t\mathrm{d}t - \mathbf{g}\mathbf{g}^\mathsf{T}s_t(\overleftarrow{\mathbf{a}})\mathrm{d}t + \mathbf{g}\mathrm{d}\overleftarrow{\mathbf{w}}_t, \quad (12)$$

where the prior distribution is restricted to a Gaussian distribution following $\mathbf{a}_T \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{T|0})$ as in Eq.(6b).

4.3 Adaptively Transport-Optimized Diffusion

Among the infinite transportation plans, we aim to obtain the optimal variational scores $\mathbf{A}_{\mathbf{a},t}^\star$ to ensure efficient transport. For that end, we resort to the SOC objective under a linearized SDE constraint:

$$\inf_{\mathbf{A}_{\mathbf{x}},\mathbf{A}_{\mathbf{v}}\in\mathbb{R}^{d\times d}} \mathbb{E}\left\{ \int_0^T \frac{1}{2}\left\|\mathbf{A}_{\mathbf{a},t}\overrightarrow{\mathbf{a}}_t\right\|_2^2 \mathrm{d}t \right\}$$

$$\text{s.t. } \mathrm{d}\overrightarrow{\mathbf{a}}_t = \left[\mathbf{f}(\overrightarrow{\mathbf{a}}_t, t) + \mathbf{g}\mathbf{g}^\mathsf{T}\mathbf{A}_{\mathbf{a},t}\overrightarrow{\mathbf{a}}_t\right]\mathrm{d}t + \mathbf{g}\mathrm{d}\overrightarrow{\mathbf{w}}_t.$$

$$\overrightarrow{\mathbf{a}}_0 \sim \rho_0 := p_{\text{data}} \otimes p_{\mathbf{v}_0}, \quad \overrightarrow{\mathbf{a}}_T \sim \rho_T := p_{\text{prior}} \otimes p_{\mathbf{v}_T}.$$

Since the diffusion from ρ_0 to ρ_T is nonlinear in general, a closed-form solution is often intractable. Bunne et al. (2023) studied the analytic solution of Gaussian SB based on a Langevin prior process, however, the ULD prior process is still not well studied.

To tackle this issue, we first build a loss function through the Feynman-Kac formula in Proposition 1:

$$\overrightarrow{\mathcal{L}}(\mathbf{A}) = -\int_0^T \mathbb{E}_{\overleftarrow{\mathbf{x}}_t \curvearrowright (12)}\left[\Gamma_\zeta(\mathbf{A}_{\mathbf{a},t}\mathbf{a}_t; \overleftarrow{\mathbf{z}}_t^\theta)\mathrm{d}t \Big| \overleftarrow{\mathbf{a}}_T\right], \quad (13)$$

where $\overleftarrow{\mathbf{z}}_t^\theta$ is the approximation of $\overleftarrow{\mathbf{z}}_t$ in Eq.(4).

We next employ stochastic approximation (SA) (Robbins and Monro, 1951; Benveniste et al., 1990) to optimize the variational score $\mathbf{A}_{\mathbf{a},t}$ adaptively for achieving more efficient transportation plans.

(1) Sample $\{\overleftarrow{\mathbf{x}}_{t_i}^{(k+1)}\}_{i=0}^{N-1}$ via the backward SDE (12);

(2) Stochastic approximation of $\{\mathbf{A}_{\mathbf{a},t_i}^{(k)}\}_{i=0}^{N-1}$:

$$\mathbf{A}_{\mathbf{a},t_i}^{(k+1)} = \mathbf{A}_{\mathbf{a},t_i}^{(k)} - \eta_{k+1}\nabla\overrightarrow{\mathcal{L}}_{t_i}(\mathbf{A}_{\mathbf{a},t_i}^{(k)}; \overleftarrow{\mathbf{x}}_{t_i}^{(k+1)}),$$

where η_{k+1} is the step size, $\{t_0, t_1, \cdots, t_{N-1}\}$ is a collection of time discretization through the Euler–Maruyama (EM) or symmetric splitting scheme (Dockhorn et al., 2022), $\nabla\overrightarrow{\mathcal{L}}_{t_i}(\mathbf{A}_{\mathbf{a},t_i}^{(k)}; \overleftarrow{\mathbf{x}}_{t_i}^{(k+1)})$ is the stochastic gradient of Eq.(13) at time t_i given $\overleftarrow{\mathbf{x}}_{t_i}^{(k+1)}$.

We expect that under mild assumptions, $\{\mathbf{A}_{\mathbf{a},t}^{(k)}\}_t$ will converge to a local optimum $\{\mathbf{A}_{\mathbf{a},t}^\star\}_t$ that yields suboptimal transport properties and the score function $\{s_t^{\theta_\star}\}$ given $\{\mathbf{A}_{\mathbf{a},t}^\star\}_t$ will be more effective to handle complex transport problems compared to the vanilla $\mathbf{A}_{\mathbf{a},t}^{(k)} \equiv \mathbf{0}$ in CLD.

Connections to Half-bridge Solvers mSB proposes to solve a general half-bridge (left) associated with the forward SDE (2) for optimal transport. For scalability, the linear approximation in Eq.(5) has limited the solution space into a class of generalized Ornstein-Uhlenbeck (gOU) processes (right):

$$\underset{\mathbb{P}\in\mathcal{D}(\rho_{\text{data}},\,\cdot)}{\arg\min}\ \text{KL}(\mathbb{P}\|\mathbb{P}_{2k-1}) \rightarrow \underset{\widehat{\mathbb{P}}\in\text{gOU}(\rho_{\text{data}},\,\cdot)}{\arg\min}\ \text{KL}(\widehat{\mathbb{P}}\|\mathbb{P}_{2k-1})$$

where $\mathcal{D}(\rho_{\text{data}},\cdot)$ and $\text{gOU}(\rho_{\text{data}},\,\cdot)$ denote the classes of path couplings from $t = 0$ to T and the initial marginal follows ρ_{data}. The solution $\widehat{\mathbb{P}}$ acts as a local optimum of the optimal transport solution.

4.4 Stabilization via Damping Transform

We rewrite the forward process (5) as a coupled probability flow ODE (Song et al., 2021)

$$d\mathbf{x}_t = \frac{1}{2}\beta\mathbf{v}_t dt$$
$$d\mathbf{v}_t = -\left[\bar{\gamma}\mathbf{v}_t + \frac{2}{\beta}\bar{\omega}_0^2\mathbf{x}_t + \frac{1}{2}\beta\gamma\nabla_\mathbf{v}\log\overrightarrow{\rho}_{t|0}(\overrightarrow{\mathbf{a}}_t)\right]dt.$$

where $\bar{\gamma} = \frac{1}{2}\beta(\gamma - 2\gamma\mathbf{A}_{\mathbf{v},t})$, $\bar{\omega}_0^2 = \frac{1}{4}\beta^2(1 - 2\gamma\mathbf{A}_{\mathbf{x},t})$.

Regarding the balance between the mass oscillation and damping (McCall, 2011), we rewrite the coupled equations into a second-order differential equation:

$$\frac{d^2\mathbf{x}_t}{dt^2} + \bar{\gamma}\frac{d\mathbf{x}_t}{dt} + \bar{\omega}_0^2\mathbf{x}_t + \frac{1}{2}\beta\gamma\nabla_\mathbf{v}\log\overrightarrow{\rho}_{t|0}(\overrightarrow{\mathbf{a}}_t) = 0.$$

Applying the case of critical damping (McCall, 2011), we have that

$$\bar{\gamma}^2 = 4R\bar{\omega}_0^2,$$

where $R \in (0, 1]$ is a scalar. The trade-off between oscillation and damping w.r.t. different R leads to two algorithms (McCall, 2011; Dockhorn et al., 2022):

- R = 1 corresponds to critical damping (VSCLD);
- R < 1 leads to under-damping (VSULD).

After some transformations, we have that

$$\mathbf{A}_{\mathbf{v},t} = \frac{1}{2} - \frac{1}{\gamma}\sqrt{R(1 - 2\gamma\mathbf{A}_{\mathbf{x},t})}. \tag{14}$$

The above equation indicates that instead of training two modules $\mathbf{A}_{\mathbf{v},t}$ and $\mathbf{A}_{\mathbf{x},t}$, we can solely train one module such as $\mathbf{A}_{\mathbf{v},t}$ and apply the transformation (14) to infer the other. Such a transformation has greatly stabilized the training and alleviated the training cost.

As observed in Dockhorn et al. (2022), under-damping often yields fast mixing while compromising the smoothness of the trajectory. Empirically, we observe that under-damping can be much faster than critical damping and may only slightly decrease the straightness of the trajectories, which motivates us to tune R to obtain the best trade-off. Now we present our algorithm in Algorithm 1.

5 Empirical Studies

5.1 Simulations

We investigate anisotropic generation using two datasets: spiral and checkerboard. Specifically, we stretch the Y-axis of the spiral dataset by a factor of 8 and the X-axis of the checkerboard dataset by a factor of 6, referring to these modified datasets as spiral-8Y and checkerboard-6X, respectively.

Anisotropic Generation We analyze CLD, ULD, VSCLD, and VSULD with various β values, denoting them as CLD-β, ULD-β, VSCLD-β, and VSULD-β. The root mean square error (RMSE) of the probability mass functions (PMFs) between the generated samples and ground-truth samples is measured to assess performance.

Initially, we experiment with CLD-5 and observe that it fails to generate content effectively in the stretched dimension, as shown in Figure 2. In contrast, our VSULD model, with a damping ratio of 0.7, utilizes a faster speed for the stretched dimension and a slower speed for the non-stretched dimension, accurately addressing anisotropic generation.
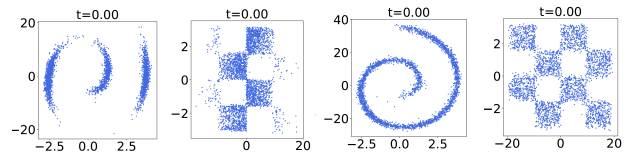

Figure 2: CLD-5 (left two) v.s. VSULD-5 (right two) on spiral-8Y and checkerboard-6X.

Trade-off between Sample Quality and Transport Efficiency To improve the anisotropic generation of CLD, we increase β and observe in Figure 3 that CLD-10 and ULD-10 exhibit comparable generation quality to VSULD-5. Additionally, we find that underdamped models such as ULD-5 and VSULD-5 converge faster than the critically-damped counterparts like CLD-5 and VSCLD-5, yielding slightly better sample quality.

Increasing β significantly enhances anisotropic generation for CLD and ULD. However, a large β results in inefficient transport for the non-stretched dimension

Algorithm 1 Variational Schrödinger Momentum Diffusion (VSMD). The variational scores $\mathbf{A}_{\mathbf{a}}^{(0)}$ are initialized to **0** by default. Specify the diffusion hyperparameters β, γ. The damping ratios R $= 1$ and R < 1 correspond to the VSCLD and VSULD algorithms, respectively, balancing oscillation and damping. Given adaptively optimized $s_t^{(k+1)}$, $\overleftarrow{\mathbf{a}}_0$ can be generated through the backward SDE (12). The continuous dynamics can be empirically discretized through the EM or symmetric splitting scheme.

---

repeat

    Optimization of the Score Function s_t via Cached Dynamics

    Draw $\mathbf{a}_0 \sim p_{\text{data}} \otimes \text{N}(\mathbf{0}, \mathbf{I})$, compute the mean process $\boldsymbol{\mu}_{t|0}$ and $\begin{pmatrix} \mathbf{C}_t \\ \mathbf{H}_t \end{pmatrix}$ by Eq.(7) and (8), respectively.

    Compute the covariance $\boldsymbol{\Sigma}_t = \mathbf{C}_t \mathbf{H}_t^{-1}$ and the Cholesky factor $\mathbf{L}_t^{-\intercal}$, where $\mathbf{L}_t \mathbf{L}_t^{\intercal} = \boldsymbol{\Sigma}_{t|0}$. Store $\boldsymbol{\mu}_{t|0}$, $\boldsymbol{\Sigma}_{t|0}$, and $\mathbf{L}_t^{-\intercal}$ in cache to speedup calculations.

    Draw $\mathbf{a}_t|\mathbf{a}_0 \sim \text{N}(\boldsymbol{\mu}_{t|0}, \boldsymbol{\Sigma}_{t|0})$ and $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \mathbf{I})$. Optimize loss function to learn the score $s_t^{(k+1)}$:

$$\nabla_\theta \| - \mathbf{L}_t^{-\intercal}\boldsymbol{\epsilon} - s_t^{(k+1)}(\overrightarrow{\mathbf{a}}_t)\|_2^2.$$
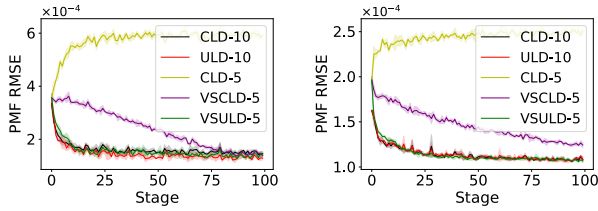
    Stochastic Approximation of Variational Scores $\mathbf{A}_{\mathbf{a},t}$

    Simulate $\overleftarrow{\mathbf{x}}_t^{(k+1)}$ via Eq.(12) and optimize $\mathbf{A}_{\mathbf{x},t}^{(k+1)}$ through the updates:

$$\mathbf{A}_{\mathbf{x},t}^{(k+1)} = \mathbf{A}_{\mathbf{x},t}^{(k)} - \eta_{k+1} \nabla_{\mathbf{A}_\mathbf{x}} \overrightarrow{\mathcal{L}}_t(\mathbf{A}_{\mathbf{a},t}^{(k)}; \overleftarrow{\mathbf{x}}_t^{(k+1)}).$$

    Compute the damping transform $\mathbf{A}_{\mathbf{v},t}^{(k+1)} = \frac{1}{2} - \frac{1}{\gamma}\sqrt{\text{R}(1 - 2\gamma\mathbf{A}_{\mathbf{x},t}^{(k+1)})}$.

until The accuracy meets the criteria.

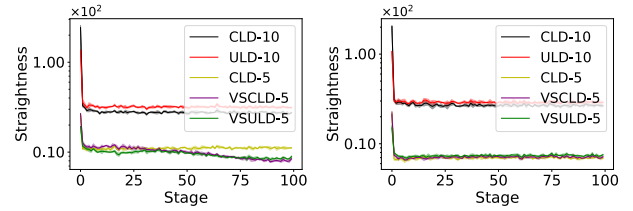(a) Spiral-8Y   (b) Checkerboard-6X

Figure 3: Sample quality evaluation. The damping ratios for ULD and VSULD are both fixed to 0.7.

(a) Spiral-8Y   (b) Checkerboard-6X

Figure 4: Straightness metric of probability flow ODEs on the non-stretched dimension via CLD, ULD, VS-CLD, and VSULD.

(e.g., the X-axis of the spiral dataset). Specifically, evaluating the straightness metric as suggested in Deng et al. (2024b), we observe in Figure 4 that both ULD-10 and CLD-10 show significantly worse straightness compared to models with $\beta = 5$, such as CLD-5, VSCLD-5, and VSULD-5. Furthermore, critically-damped models demonstrate marginally better straightness metrics than underdamped models, indicating a trade-off between convergence speed and transport efficiency.
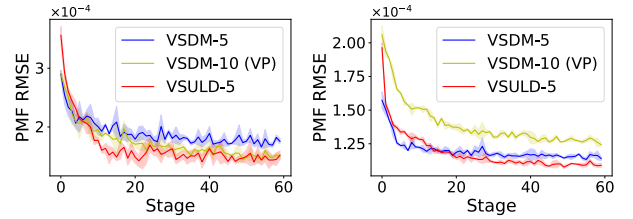
Overdamped v.s. Underdamped We also compare VSULD-5 models with VSDM using a fixed β value ($\beta = 5$, VSDM-5) and the same VPSDE schedule as in Deng et al. (2024b) with $\beta_{\max} = 10$ (VSDM-10 (VP)). Figure 5 shows that VSDM-5 and VSDM-10 (VP) are overall comparable, and VSULD consistently outperforms the overdamped alternatives in terms of accuracy and speed.

(a) Spiral-8Y   (b) Checkerboard-6X

Figure 5: Overdamped versus underdamped models.

5.2 Time Series Forecasting

We demonstrate our models ability in a real world multivariate probabilistic forecasting. Given a sequence $x_{1:N} = \{(t_i, x_i)\}_{i=1}^N$ where $t_i \in \mathbb{R}$ is a time variable and $x_i \in \mathbb{R}^d$. Our goal is to predict the next elements of

this sequence, that is predict $x_{N+1}, \ldots, x_{N+P}$ for some time points $t_{N+1}, \ldots t_{N+P}$.

We follow the same approach as in Deng et al. (2024b) and encode the sequence $x_{1:N}$ into a vector $h_i \in \mathbb{R}^h$. We then train a conditional diffusion model to predict $x_{n+1}|h_n$. Such a model allows generating the entire prediction sequence in an auto-regressive fashion as in Rasul et al. (2021).

We utilize a similar U-Net architecture as as in Deng et al. (2024b). We use a second order Heunn method as introduced in Karras et al. (2022). To the best of our knowledge the use of second order samplers had not been explored in the time series forecasting problem. As expected, this change significantly improves the forecasts. Due to the autoregressive nature of the method it is important to reduce the error in early stages to prevent the model from drifting away. The training is performed on a laptop Geforce RTX 4070 with $8GB$ of VRAM.

	Electricity	Exchange	Solar
CLD	0.2115	0.0069	0.489
VSDM	0.038	0.008	0.395
VSCLD	**0.029**	0.0085	0.471
VSULD	0.039	**0.0060**	**0.3585**

Table 1: Performance comparison on Electricity, Exchange, and Solar in the CRPS-Sum metric

We test on the exchange rate dataset which contains 6071 8-dimensional measurements every day. The solar dataset is a 137 dimensional dataset with 7009 values measured every hour. Finally the electricity dataset is an hourly dataset with 370 dimensions with 5833 measurements. In Table 1 we demonstrate the value of CRPS-Sum of our method. We compare against CLD, and VDSM using the same architecture and the improved sampler. We present forecasts in the first three dimensions for the solar dataset in figure 6, forecasts for other datasets and methods in the appendix.
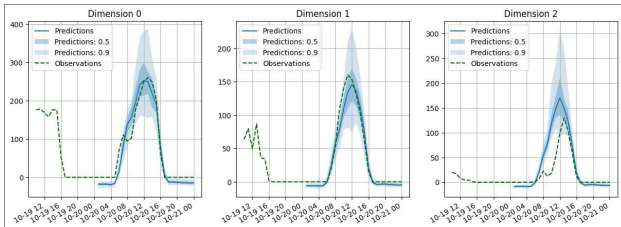

Figure 6: Sample forecasts of our method in the solar dataset

5.3 Image experiments

Experiment Setup We test the scalability of our method by training an unconditional generative model on the CIFAR-10 dataset. We make use of the critical damping transformation to perform this experiment. We train our model in 8 NVIDIA V100-16GB GPUs with a batch size of 256. We follow standard practices and use the EMA during inference where we made use of the second order Heun's method to discretize the probability flow ODE. We present some sample images in Appendix C.

One natural concern is that when the variational score gets updated, this changes the dynamics and the target distribution of the backward process. Then the score needs to be correctly updated to revert for these changes. To circumvent this we make use of a step learning rate schedule with parameter .99. Doing so allows to keep the variational scores from changing drastically towards the later parts of training. This in combination with the stochastic approximation technique described in 4.2.1 allows for a stable training and annealing of the variational scores. We present a comparison of fid on Table 2, we achieve competitive results without requiring warm-up starts. We refer the reader to Appendix C for more details on the choice of hyperparameters.

Table 2: CIFAR10 evaluation using sample quality (FID)

Class	Method	FID $\downarrow$
	VSCLD (Ours)	2.89
	VSDM (Deng et al. (2024b))	2.28
OT	SB-FBSDE (Chen et al. (2022))	3.01
	DOT (Tanaka (2019))	15.78
	DGflow (Ansari et al. (2020))	9.63
	SDE (Song et al. (2021))	2.92
	CLD (Dockhorn et al. (2022))	2.23
SGMs	VDM (Kingma et al. (2021))	4.00
	LSGM (Vahdat et al. (2021))	2.10
	EDM (Karras et al. (2022))	1.97

6 Conclusions and Future Works

Momentum Schrödinger bridge diffusion models provide a principled framework for studying generative models with optimal transport properties. However, achieving optimal transportation plans is often prohibitively expensive in real-world scenarios. To address the scalability issue, we propose the Variational Schrödinger Momentum Diffusion (VSMD) model, a scalable multivariate diffusion model that enables simulation-free training of backward scores, and the forward scores are optimized adaptively for more efficient transportation

plans. Motivated by kinetic (second-order) Langevin dynamics, the inclusion of velocity components enhances training and sampling efficiency and eliminates the need for a complex denoising process. For future work, we aim to further simplify the forward diffusion process while maintaining efficient transportation plans to support more scalable applications.

Acknowledgements

References

Albergo, M. S., Bof, N. M., and Vanden-Eijnden, E. (2023). Stochastic Interpolants: A Unifying Framework for Flows and Diffusions. *arXiv:2303.08797v1*, pages 1–48.

Albergo, M. S. and Vanden-Eijnden, E. (2023). Building Normalizing Flows with Stochastic Interpolants. In *International Conference on Learning Representation (ICLR)*.

Anderson, B. D. (1982). Reverse-time Diffusion Equation Models. *Stochastic Processes and Their Applications*, 12(3):313–326.

Ansari, A. F., Ang, M. L., and Soh, H. (2020). Refining Deep Generative Models via Discriminator Gradient Flow. In *International Conference on Learning Representations*.

Bartosh, G., Vetrov, D., and Naesseth, C. A. (2024). Neural Flow Diffusion Models: Learnable Forward Process for Improved Diffusion Modelling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Benveniste, A., Métivier, M., and Priouret, P. (1990). *Adaptive Algorithms and Stochastic Approximations*. Berlin: Springer.

Bunne, C., Hsieh, Y.-P., Cuturi, m., and Krause, A. (2023). The Schrödinger Bridge between Gaussian Measures has a Closed Form. In *AISTATS*.

Caluya, K. F. and Halder, A. (2022). Wasserstein Proximal Algorithms for the Schrödinger Bridge Problem: Density Control with Nonlinear Drift. *IEEE Transactions on Automatic Control*, 67(3):1163–1178.

Chen, C., Ding, N., and Carin, L. (2015). On the Convergence of Stochastic Gradient MCMC Algorithms with High-order Integrators. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2278–2286.

Chen, H., Lee, H., and Lu, J. (2023a). Improved Analysis of Score-based Generative Modeling: User-friendly Bounds under Minimal Smoothness Assumptions. In *International Conference on Machine Learning*, pages 4735–4763.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. (2018). Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2023b). Sampling is as Easy as Learning the Score: Theory for Diffusion Models with Minimal Data Assumptions. *International Conference on Learning Representation (ICLR)*.

Chen, T., Gu, J., Dinh, L., Theodorou, E. A., Susskind, J., and Zhai, S. (2024). Generative Modeling with Phase Stochastic Bridges. In *International Conference on Learning Representation (ICLR)*.

Chen, T., Liu, G.-H., Tao, M., and Theodorou, E. A. (2023c). Deep Momentum Multi-Marginal Schrödinger Bridge. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, T., Liu, G.-H., and Theodorou, E. A. (2022). Likelihood Training of Schrödinger Bridge using Forward-Backward SDEs Theory. In *International Conference on Learning Representation (ICLR)*.

Chen, Y., Deng, W., Fang, S., Li, F., Yang, N., Zhang, Y., Rasul, K., Zhe, S., Schneider, A., and Nevmyvaka, Y. (2023d). Provably Convergent Schrödinger Bridge with Applications to Probabilistic Time Series Imputation. In *International Conference on Machine Learning (ICML)*.

Chen, Y. and Georgiou, T. (2016). Stochastic Bridges of Linear Systems. *IEEE Transactions on Automatic Control*, 61(2).

Chen, Y., Georgiou, T. T., and Pavon, M. (2021). Stochastic Control Liaisons: Richard Sinkhorn Meets Gaspard Monge on a Schrödinger Bridge. *SIAM Review*, 63(2):249–313.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2017). Underdamped Langevin MCMC: A Non-Asymptotic Analysis. In *Proc. of Conference on Learning Theory (COLT)*.

Dalalyan, A. S. and Karagulyan, A. G. (2019). User-friendly Guarantees for the Langevin Monte Carlo with Inaccurate Gradient. *Stochastic Processes and their Applications*, 129:12:5278–5311.

Dalalyan, A. S. and Riou-Durand, L. (2020). On Sampling from a Log-concave Density using Kinetic Langevin Diffusions. *Bernoulli*, 26(3):1956–1988.

De Bortoli, V., Korshunova, I., Mnih, A., and Doucet, A. (2024). Schrödinger Bridge Flow for Unpaired Data Translation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion Schrödinger Bridge with Applications to Score-Based Generative Modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Deng, W., Chen, Y., Yang, N. T., Du, H., Feng, Q., and Chen, R. T. Q. (2024a). Reflected Schrödinger Bridge for Constrained Generative Modeling . In *Proc. of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Deng, W., Luo, W., Tan, Y., Biloš, M., Chen, Y., Nevmyvaka, Y., and Chen, R. T. Q. (2024b). Variational Schrödinger Diffusion Models. In *Proc. of the International Conference on Machine Learning (ICML)*.

Dhariwal, P. and Nichol, A. (2022). Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dockhorn, T., Vahdat, A., and Kreis, K. (2022). Score-Based Generative Modeling with Critically-Damped Langevin Diffusion . In *Proc. of the International Conference on Learning Representation (ICLR)*.

Eyring, L., Klein, D., Uscidda, T., Palla, G., Kilbertus, N., Akata, Z., and Theis, F. (2024). Unbalancedness in Neural Monge Maps Improves Unpaired Domain Translation. In *ICLR*.

Gushchin, N., Kolesov, A., Korotin, A., Vetrov, D., and Burnaev, E. (2023). Entropic Neural Optimal Transport via Diffusion Processes. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hairer, E., Lubich, C., and Wanner, G. (2006). *Geometric Numerical Integration*. Springer, Berlin, 2nd edition.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., and Salimans, T. (2022). Imagen Video: High Definition Video Generation with Diffusion Models. In *arXiv:2210.02303*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Karras, T., Aittala, M., Aila, T., and Laine, S. (2022). Elucidating the Design Space of Diffusion-Based Generative Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kim, B., Kwon, G., Kim, K., and Ye, J. C. (2024). Unpaired Image-to-Image Translation via Neural Schrödinger Bridge. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Kingma, D. P., Salimans, T., Poole, B., and Ho, J. (2021). Variational Diffusion Models. *ArXiv*, abs/2107.00630.

Kong, Z., Ping, W., Huang, J., Zhao, K., and Catanzaro, B. (2021). DiffWave: A Versatile Diffusion Model for Audio Synthesis . In *Proc. of the International Conference on Learning Representation (ICLR)*.

Korotin, A., Gushchin, N., and Burnaev, E. (2024). Light Schrödinger Bridge. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Lavenant, H. and Santambrogio, F. (2022). The Flow Map of the Fokker–Planck Equation Does Not Provide Optimal Transport. *Applied Mathematics Letters*, 133.

Lee, H., Lu, J., and Tan, Y. (2022). Convergence for Score-based Generative Modeling with Polynomial Complexity. *Advances in Neural Information Processing Systems (NeurIPS)*.

Leimkuhler, B. and Matthews, C. (2013). Rational construction of stochastic numerical methods for molecular sampling. *Applied Mathematics Research eXpress*, 2013(1):34–56.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). Flow Matching for Generative Modeling. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Liu, G.-H., Chen, T., So, O., and Theodorou, E. A. (2022). Deep Generalized Schrödinger Bridge. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Liu, Q. (2022). Rectified Flow: A Marginal Preserving Approach to Optimal Transport. *arXiv:2209.14577*.

Liu, X., Gong, C., and Liu, Q. (2023). Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In *ICLR*.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. (2022). DPM-Solver: A Fast ODE Solver for Diffusion Probabilistic Model Sampling in Around 10 Steps. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ma, J. and Yong, J. (2007). *Forward-Backward Stochastic Differential Equations and their Applications*. Springer.

Ma, Y.-A., Chatterji, N. S., Cheng, X., Flammarion, N., Bartlett, P. L., and Jordan, M. I. (2021). Is there an analog of nesterov acceleration for gradient-based mcmc? *Bernoulli*, 27(3).

Mangoubi, O. and Smith, A. (2021). Mixing of Hamiltonian Monte Carlo on Strongly Log-concave Distributions: Continuous Dynamics. *The Annals of Applied Probability*, 31(5):2019–2045.

Mangoubi, O. and Vishnoi, N. K. (2018). Dimensionally Tight Running Time Bounds for Second-order

Hamiltonian Monte Carlo. In *Advances in Neural Information Processing Systems (NeurIPS)*.

McCall, M. W. (2011). *Classical Mechanics: From Newton to Einstein: A Modern Introduction (Second Edition)*. John Wiley & Sons.

McCann, R. J. (1997). A Convexity Principle for Interacting Gases. *Advances in mathematics*, 128(1):153–179.

Neal, R. M. (2012). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo*, volume 54, pages 113–162.

Neklyudov, K., Brekelmans, R., Tong, A., Atanackovic, L., Liu, Q., and Makhzani, A. (2024). A Computational Framework for Solving Wasserstein Lagrangian Flows. In *Proc. of the International Conference on Machine Learning (ICML)*.

Pavon, M., Tabak, E. G., and Trigila, G. (2021). The Data-driven Schrödinger Bridge. *Communications on Pure and Applied Mathematics*, 74:1545–1573.

Peluchetti, S. (2023). Diffusion Bridge Mixture Transports, Schrödinger Bridge Problems and Generative Modeling. *Journal of Machine Learning Research*.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. T. Q. (2023). Multisample Flow Matching: Straightening Flows with Minibatch Couplings. In *Proc. of the International Conference on Machine Learning (ICML)*.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. In *arXiv:2204.06125v1*.

Rasul, K., Seward, C., Schuster, I., and Vollgraf, R. (2021). Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *International Conference on Machine Learning*.

Robbins, H. and Monro, S. (1951). A Stochastic Approximation Method. *Annals of Mathematical Statistics*, 22:400–407.

Salimans, T. and Ho, J. (2022). Progressive Distillation for Fast Sampling of Diffusion Models. In *ICLR*.

Särkkä, S. and Solin, A. (2019). *Applied Stochastic Differential Equations*. Cambridge University Press.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. (2023). Diffusion Schrödinger Bridge Matching. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Singhal, R., Goldstein, M., and Ranganath, R. (2023). Where to Diffuse, How to Diffuse, and How to Get Back: Automated Learning for Multivariate Diffusions. In *Proc. of the International Conference on Learning Representation (ICLR)*.

Somnath, V. R., Pariset, M., Hsieh, Y.-P., Martinez, M. R., Krause, A., and Bunne, C. (2023). Aligned Diffusion Schrödinger Bridges. In *Conference on Uncertainty in Artificial Intelligence*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-Based Generative Modeling through Stochastic Differential Equations . In *International Conference on Learning Representation (ICLR)*.

Tanaka, A. (2019). Discriminator Optimal Transport. In *Neural Information Processing Systems*.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. (2024). Improving and Generalizing Flow-based Generative Models with Minibatch Optimal Transport. *Transactions on Machine Learning Research*.

Tuckerman, M. E. (2010). *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press, New York.

Vahdat, A., Kreis, K., and Kautz, J. (2021). Score-based Generative Modeling in Latent Space. *Advances in Neural Information Processing Systems*, 34:11287–11302.

Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. (2021). Solving Schrödinger Bridges via Maximum Likelihood. *Entropy*, 23(9):1134.

Wang, G., Jiao, Y., Xu, Q., Wang, Y., and Yang, C. (2021). Deep Generative Learning via Schrödinger Bridge. In *International Conference on Machine Learning (ICML)*.

Supplementary Material for "Variational Schrödinger Momentum Diffusion"

A Kinetic (second-order) Langevin Dynamics

A.1 Different Damping Regimes

We follow Dockhorn et al. (2022) and study the different damping regimes for the multivariate kinetic Langevin diffusion process:

$$\begin{pmatrix} \mathrm{d}\mathbf{x}_t \\ \mathrm{d}\mathbf{v}_t \end{pmatrix} = \underbrace{\frac{\beta}{2} \begin{pmatrix} \mathbf{v}_t \\ -(1 - 2\gamma\mathbf{A}_{\mathbf{x},t})\mathbf{x}_t \end{pmatrix} \mathrm{d}t}_{\text{Hamiltonian component}} + \underbrace{\frac{\beta\gamma}{2} \begin{pmatrix} \mathbf{0} \\ -\left(1 - 2\mathbf{A}_{\mathbf{v},t}\right)\mathbf{v}_t \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \mathbf{0} \\ \sqrt{\beta\gamma}\mathbf{I}_d \end{pmatrix} \mathrm{d}\mathbf{w}_t}_{\text{Ornstein-Uhlenbeck process: O}}. \tag{15}$$

- overdamped Langevin dynamics (LD): a high friction limit of (15) without momentum (Hamiltonian component) acceleration. LD requires $\Omega(d/\epsilon^2)$ iterations to achieve an ϵ error in 2-Wasserstein (W2) distance for strongly log-concave distributions (Dalalyan and Karagulyan, 2019).

- critically-damped Langevin dynamics (CLD) via R = 1 in Eq.(14): theoretically optimal trade-off between mass oscillation and damping (McCall, 2011). However, in practice, we may need to tune the damping ratio R to achieve the best balance between acceleration and transport efficiency.

- underdamped Langevin dynamics (ULD) via R < 1: the Hamiltonian component plays a crucial role and also induces more oscillatory behavior. ULD requires only $\Omega(\sqrt{d}/\epsilon)$ (instead of $\Omega(d/\epsilon^2)$ via LD) iterations to achieve an ϵ error in W2 for strongly log-concave distributions (Cheng et al., 2017).

A.2 Numerical Schemes

The Euler–Maruyama scheme for the backward kinetic Langevin diffusion in Eq.(12) follows that

$$\begin{aligned} \overleftarrow{\mathbf{x}}_{(n-1)h} &= \overleftarrow{\mathbf{x}}_{nh} - \frac{h}{2}\beta\overleftarrow{\mathbf{v}}_{nh} \\ \overleftarrow{\mathbf{v}}_{(n-1)h} &= \overleftarrow{\mathbf{v}}_{nh} + \frac{h}{2}\beta(1 - 2\gamma\mathbf{A}_{\mathbf{x},nh}) + \frac{h}{2}\beta\gamma\left(1 - 2\mathbf{A}_{\mathbf{v},nh}\right)\overleftarrow{\mathbf{v}}_{nh} + h\beta\gamma s_{nh}(\overleftarrow{\mathbf{a}}_{nh}) + \sqrt{\beta\gamma h}\overleftarrow{\boldsymbol{\xi}}_{nh}, \end{aligned} \tag{16}$$

where h is the learning rate.

Theoretically, the Euler–Maruyama scheme (16) suffers from instability with large discretization step sizes. Motivated by the symplectic Euler for the Hamiltonian system, we consider the symmetric splitting (S2) scheme (Tuckerman, 2010; Leimkuhler and Matthews, 2013; Dockhorn et al., 2022) for the kinetic Langevin dynamics to ensure better stability. To that end, we first compose the Hamiltonian component in Eq.(15) into two parts:

$$\begin{pmatrix} \mathrm{d}\overleftarrow{\mathbf{x}}_t \\ \mathrm{d}\overleftarrow{\mathbf{v}}_t \end{pmatrix} = \underbrace{\frac{\beta}{2} \begin{pmatrix} \overleftarrow{\mathbf{v}}_t \\ \mathbf{0} \end{pmatrix} \mathrm{d}t}_{\text{A}} + \underbrace{\frac{\beta}{2} \begin{pmatrix} \mathbf{0} \\ -(1 - 2\gamma\mathbf{A}_{\mathbf{x},t})\overleftarrow{\mathbf{x}}_t \end{pmatrix} \mathrm{d}t}_{\text{B}} + \underbrace{\frac{\beta\gamma}{2} \begin{pmatrix} \mathbf{0} \\ -\left(1 - 2\mathbf{A}_{\mathbf{v},t}\right)\overleftarrow{\mathbf{v}}_t - 2s_t(\overleftarrow{\mathbf{a}}_t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} \mathbf{0} \\ \sqrt{\beta\gamma}\mathbf{I}_d \end{pmatrix} \mathrm{d}\overleftarrow{\mathbf{w}}_t}_{\text{O}}, \tag{17}$$

where each part yields an "analytic" form and the underlying Kolmogorov (Fokker-Planck) operators are denoted by $\mathcal{L}_A$, $\mathcal{L}_B$, and $\mathcal{L}_O$, respectively.

The stochastic discretization schemes for kinetic Langevin are well-studied and lead to different formulations. The main difference lies in the approximations of the Hamiltonian component (Leimkuhler and Matthews, 2013)

- the BAOAB method: $\Phi_{\text{BAOAB}}^h = \exp(\frac{h}{2}\mathcal{L}_B)\exp(\frac{h}{2}\mathcal{L}_A)\exp(h\mathcal{L}_O)\exp(\frac{h}{2}\mathcal{L}_A)\exp(\frac{h}{2}\mathcal{L}_B)$;

- the ABOBA method: $\Phi_{\text{ABOBA}}^h = \exp(\frac{h}{2}\mathcal{L}_A)\exp(\frac{h}{2}\mathcal{L}_B)\exp(h\mathcal{L}_O)\exp(\frac{h}{2}\mathcal{L}_B)\exp(\frac{h}{2}\mathcal{L}_A)$.

In particular, the ABOBA method based on the symmetric splitting scheme follows that:

$$\overleftarrow{\mathbf{x}}_{(k-\frac{1}{2})h} = \overleftarrow{\mathbf{x}}_{nh} - \frac{h}{4}\beta\overleftarrow{\mathbf{v}}_{nh}$$

$$\overleftarrow{\mathbf{v}}_{(k-\frac{1}{2})h} = \overleftarrow{\mathbf{v}}_{nh} + \frac{h}{4}\beta(1 - 2\gamma\mathbf{A}_{\mathbf{x},nh})$$

$$\overleftarrow{\mathbf{v}}_{(k-\frac{1}{2})h} = \overleftarrow{\mathbf{v}}_{(k-\frac{1}{2})h} + \frac{h}{2}\beta\gamma\big(1 - 2\mathbf{A}_{\mathbf{v},nh}\big)\overleftarrow{\mathbf{v}}_{nh} + h\beta\gamma s_{nh}(\overleftarrow{\mathbf{a}}_{nh}) + \sqrt{\beta\gamma h}\,\overleftarrow{\boldsymbol{\xi}}_{nh}$$

$$\overleftarrow{\mathbf{v}}_{(n-1)h} = \overleftarrow{\mathbf{v}}_{(k-\frac{1}{2})h} + \frac{h}{4}\beta(1 - 2\gamma\mathbf{A}_{\mathbf{x},(k-\frac{1}{2})h})$$

$$\overleftarrow{\mathbf{x}}_{(n-1)h} = \overleftarrow{\mathbf{x}}_{(k-\frac{1}{2})h} - \frac{h}{4}\beta\overleftarrow{\mathbf{v}}_{(n-1)h}.$$

The BAOAB method can derived similarly. The numerical study of the invariant measure is based on the Baker–Campbell–Hausdorff (BCH) expansion (Hairer et al., 2006). The symmetric splitting scheme is a second-order integrator and is known to yield an approximation error of $O(h^3)$ (Leimkuhler and Matthews, 2013; Chen et al., 2015; Dockhorn et al., 2022). In contrast, the Euler–Maruyama scheme has a weaker error of $O(h^2)$ (Chen et al., 2015, 2023b). Nonetheless, the empirical advantage of the symmetric splitting scheme mainly holds with a small learning rate (Leimkuhler and Matthews, 2013) and may not necessarily reduce the number of function evaluations in practice. For the sake of convenience in theoretical analysis, we will focus on the Euler–Maruyama scheme similar to the analysis in Chen et al. (2023b).

B Convergence Theory

We follow the methodology outlined in Deng et al. (2024b) and utilize stochastic approximation (SA) techniques (Robbins and Monro, 1951) to evaluate the generation quality based on the adaptive momentum diffusion models. By employing simulated backward trajectories, we optimize the variational scores. Consequently, the optimized forward process becomes not only simulation-free but also more transport-efficient. The iterates are conducted alternatingly and eventually yield more accurate backward score functions.

Algorithm 2 The SA formulation of the variational Schrödinger diffusion models. We approximate $\nabla_{\mathbf{v}}\log\overrightarrow{\rho}_t^{(k)}$ through the parametrized score estimation $s_t^{(k+1)}$ at each stage k and time t.

---

repeat

    Simulation: Draw approximate samples $(\overleftarrow{\mathbf{x}}_{(n-1)h}^{(k+1)}, \overleftarrow{\mathbf{v}}_{(n-1)h}^{(k+1)})$ from the backward process (16) with fixed $\mathbf{A}_{\mathbf{x},nh}^{(k)}$ and $\mathbf{A}_{\mathbf{v},nh}^{(k)}$), where $(\overleftarrow{\mathbf{x}}_{(N-1)h}^{(k+1)}, \overleftarrow{\mathbf{v}}_{(N-1)h}^{(k+1)}) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_{(N-1)h|0}^{(k)})$ and $\boldsymbol{\Sigma}_{(N-1)h|0}^{(k)}$ is defined in Eq.(6), $n \in \{1, 2, \cdots, N-1\}$.

    Optimization: Minimize the transport cost via the forward loss function (11):

$$\mathbf{A}_{\mathbf{a},nh}^{(k+1)} = \mathbf{A}_{\mathbf{a},nh}^{(k)} - \eta_{k+1}\nabla\overrightarrow{\mathcal{L}}_{nh}(\mathbf{A}_{\mathbf{a},nh}^{(k)}; \overleftarrow{\mathbf{x}}_{nh}^{(k+1)}), \tag{18}$$

    where η_k denotes the step size and $n \in \{0, 1, \cdots, N-1\}$.

until $k = k_{\max}$

In our convergence study, we assume a single step of sampling and a single step of optimization and conduct the iterates in Eq.(18) for the coupled score function $\mathbf{A}_{\mathbf{a},t}$ instead of $\mathbf{A}_{\mathbf{x},t}$ for theoretical convenience. However, this simplification is not required in practical applications to boost the performance.

The SA iterates (18) can be viewed as a stochastic numerical scheme of an ODE system as follows

$$\mathrm{d}\mathbf{A}_{\mathbf{a},t} = \nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_{\mathbf{a},t})\mathrm{d}s, \tag{19}$$

where $\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_{\mathbf{a},t})$ is the mean-field aggregated from random-field functions $\nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_{\mathbf{a},t}; \overleftarrow{\mathbf{a}}_t^{(\cdot)})$:

$$\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_{\mathbf{a},t}) = \int_{\mathcal{X}} \nabla\overrightarrow{\mathcal{L}}_t(\mathbf{A}_{\mathbf{a},t}; \overleftarrow{\mathbf{a}}_t^{(\cdot)})\overleftarrow{\rho}_t(\mathrm{d}\overleftarrow{\mathbf{a}}_t^{(\cdot)}). \tag{20}$$

We aim to find the solution of $\nabla \overrightarrow{\mathbf{L}}_t(\mathbf{A}_{\mathbf{a},t}^\star) = \mathbf{0}$ through approximate samples $(\overleftarrow{\mathbf{x}}_{(n-1)h}^{(k+1)}, \overleftarrow{\mathbf{v}}_{(n-1)h}^{(k+1)})$ from the backward process (16).

Next, we present the underlying assumptions for regularity conditions for the solution $\mathbf{A}_{\mathbf{a},t}^\star$ and the neighborhood.

Assumption A1 (Positive Definiteness). *Both $\mathbf{I}_d - 2\gamma\mathbf{A}_{\mathbf{x},t}$ and $\mathbf{I}_d - 2\mathbf{A}_{\mathbf{v},t}$ are symmetric and positive-definite. Morever, $\|\mathbf{D}\|_{op} \lesssim O(1)$, where $\|\cdot\|_{op}$ denotes the operator norm of a matrix.*

Assumption A2 (Locally strong convexity). *For any stable equilibrium $\mathbf{A}_{\mathbf{a}}^\star$ with $\nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A}_{\mathbf{a}}^\star) = \mathbf{0}$, there exists a convex set $\boldsymbol{\Theta}$ s.t. $\mathbf{A}_{\mathbf{a}}^\star \in \boldsymbol{\Theta} \subset \mathcal{A}$ and $m\mathbf{I} \preccurlyeq \frac{\partial^2 \overrightarrow{\mathbf{L}}_t}{\partial \mathbf{A}^2}(\mathbf{A}) \preccurlyeq M\mathbf{I}$ for $\forall \mathbf{A} \in \boldsymbol{\Theta}$ and some fixed constants $M > m > 0$.*

The following assumes the smoothness of the score functions with respect to the input $\mathbf{x}, \mathbf{y}$ and variational scores $\mathbf{A}_{\mathbf{a}_1}, \mathbf{A}_{\mathbf{a}_2}$ and similar ones have been widely used in Lee et al. (2022); Chen et al. (2023b,a); Deng et al. (2024b).

Assumption A3 (Smoothness). *There exists a fixed constant L such that for any $t \in [0, T]$, $\mathbf{A}_{\mathbf{a}_1}, \mathbf{A}_{\mathbf{a}_2} \in \mathcal{A}$ and $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, the score functions $\nabla_{\mathbf{v}} \log \overrightarrow{\rho}_{\mathbf{a}_1,t}$ and $\nabla_{\mathbf{v}} \log \overrightarrow{\rho}_{\mathbf{a}_2,t}$ w.r.t. $\mathbf{A}_{\mathbf{a}_1}$ and $\mathbf{A}_{\mathbf{a}_2}$ satisfy*

$$\|\nabla_{\mathbf{v}} \log \overrightarrow{\rho}_{\mathbf{a}_1,t}(\mathbf{x}) - \nabla_{\mathbf{v}} \log \overrightarrow{\rho}_{\mathbf{a}_2,t}(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2 + L\|\mathbf{A}_{\mathbf{a}_1} - \mathbf{A}_{\mathbf{a}_2}\|,$$

where $\|\cdot\|_2$ is the Euclidean norm and $\|\cdot\|$ denotes the standard matrix norm.

Assumption A4 (Bounded Second Moment). *The second moment of ρ_{data} is upper bounded by $\mathfrak{m}_2^2$.*

Assumption A5 (Estimation of Score Functions). *For all $t \in [0, T]$, and any $\mathbf{A}_{\mathbf{a}}$, the estimation error of the score functions is upper bounded by ϵ_{score}^2:*

$$\mathbb{E}_{\overrightarrow{\rho}_t}[\|s_t - \nabla_{\mathbf{v}} \log \overrightarrow{\rho}_{\mathbf{a},t}\|_2^2] \leq \epsilon_{score}^2.$$

Proof Sketch   Similar to Deng et al. (2024b), the understanding of the quality of the adaptively generated data hinges on the stochastic approximation framework and can be decomposed into three steps:

- **Fixed Generation Quality**: We first show that given a fixed $\mathbf{A}_{\mathbf{a},t}^{(k)}$, the generated data is approximately close to the real data in Theorem 1;

- **Convergence of Variational Scores**: We next prove the convergence of $\mathbf{A}_{\mathbf{a},t}^{(k)}$ to the optimal $\mathbf{A}_{\mathbf{a},t}^\star$ via stochastic approximation in Theorem 2;

- **Adaptive Generation Quality:** In the limit of infinite iterations, we show the generated data is close to the real data in distribution given the optimal $\mathbf{A}_{\mathbf{a},t}^\star$ in Theorem 3.

In particular, the stochastic approximation part is standard and inherited from Deng et al. (2024b). The major novelty lies in the extension of single-variate kinetic Langevin diffusion to multi-variate kinetic Langevin diffusion through a customized Lyapunov function in Eq.(22). As such, the adaptive momentum diffusion differs from the vanilla CLD in that the transportation plans are optimized locally in particular tailored to the data, moreover, the training maintains the same efficiency as CLD due to the simulation-free property of forward processes.

B.1   Fixed Generation Quality

In this section, we first study the generation quality based on time-invariant $\mathbf{A}_{\mathbf{x},t} := \mathbf{A}_{\mathbf{x}}$ and $\mathbf{A}_{\mathbf{v},t} := \mathbf{A}_{\mathbf{v}}$ and discuss the extensions to time-variant cases.

$$d\overrightarrow{\mathbf{a}}_t = -\frac{1}{2}\mathbf{D}\beta\overrightarrow{\mathbf{a}}_t dt + \mathbf{g}d\overrightarrow{\mathbf{w}}_t \tag{21}$$

$$\mathbf{D} = \begin{pmatrix} 0 & -1 \\ 1 & \gamma \end{pmatrix} \otimes \mathbf{I}_d - 2\gamma \begin{pmatrix} 0 & 0 \\ \mathbf{A}_{\mathbf{x}} & \mathbf{A}_{\mathbf{v}} \end{pmatrix}.$$

We denote the distribution of $\overrightarrow{\mathbf{a}}_t$ by $\overrightarrow{\rho}_t$ and its $\mathbf{x}$ and $\mathbf{v}$-marginal by $\overrightarrow{\rho}_{\mathbf{x},t}$ and $\overrightarrow{\rho}_{\mathbf{v},t}$ respectively. To ensure an exponential convergence of CLD in (21) to its invariant measure, we make the following assumption:

Lemma 1 (Invariant Measure)**.** *If Assumption A1 holds, the invariant measure of (21) is given by*

$$\mu = \mathrm{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{B}_1^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}_2^{-1} \end{pmatrix}\right).$$

where $\mathbf{B}_1 = (\mathbf{I}_d - 2\gamma\mathbf{A_x})^T(\mathbf{I}_d - 2\mathbf{A_v})$ *and* $\mathbf{B}_2 = \mathbf{I}_d - 2\mathbf{A_v}$

Proof   Recall the Fokker Planck Equation gives us

$$\partial_t \overrightarrow{\rho}_t(\mathbf{a}) = \nabla \cdot \left(\overrightarrow{\rho}_t(\mathbf{a})(\frac{1}{2}\mathbf{D}\beta\mathbf{a} + \frac{1}{2}\mathbf{g}\mathbf{g}^T\nabla \ln \overrightarrow{\rho}_t(\mathbf{a})) \right),$$

where $\overrightarrow{\rho}_t$ is the density of $\overrightarrow{\mathbf{a}}_t$. Note that $\mu(\mathbf{a}) = \mu(\mathbf{x}, \mathbf{v}) \propto \exp(-\frac{\mathbf{x}^T\mathbf{B}_1\mathbf{x}}{2} - \frac{\mathbf{v}^T\mathbf{B}_2\mathbf{v}}{2})$. Hence if $\overrightarrow{\rho}_t = \mu$, then

$$
\begin{aligned}
\partial \overrightarrow{\rho}_t(\mathbf{a}) &= \nabla \cdot \left(\mu(\mathbf{a})\left(\frac{1}{2}\mathbf{D}\beta\mathbf{a} + \frac{1}{2}\mathbf{g}\mathbf{g}^T\nabla \ln \mu(\mathbf{a})\right) \right) \\
&= \left\langle \nabla\mu(\mathbf{a}), \frac{1}{2}\mathbf{D}\beta\mathbf{a} + \frac{1}{2}\mathbf{g}\mathbf{g}^T\nabla \ln \mu(\mathbf{a}) \right\rangle + \mu(\mathbf{a})\nabla \cdot \left(\frac{1}{2}\mathbf{D}\beta\mathbf{a} + \frac{1}{2}\mathbf{g}\mathbf{g}^T\nabla \ln \mu(\mathbf{a}) \right) \\
&= \left\langle \mu(\mathbf{a})\begin{pmatrix} -\mathbf{B}_1\mathbf{x} \\ -\mathbf{B}_2\mathbf{v} \end{pmatrix}, \frac{\beta}{2}\begin{pmatrix} 0 & -\mathbf{I}_d \\ \mathbf{I}_d - 2\gamma\mathbf{A_x} & \gamma\mathbf{B}_2 \end{pmatrix}\begin{pmatrix} \mathbf{x} \\ \mathbf{v} \end{pmatrix} + \frac{\beta}{2}\begin{pmatrix} 0 \\ \gamma\mathbf{B}_2\mathbf{v} \end{pmatrix} \right\rangle + \frac{\beta}{2}\mu(\mathbf{a})(\mathrm{Tr}(\mathbf{D}) - \mathrm{Tr}(\mathbf{B}_2)) \\
&= \frac{\beta}{2}\mu(\mathbf{a})\left\langle \begin{pmatrix} -\mathbf{B}_1\mathbf{x} \\ -\mathbf{B}_2\mathbf{v} \end{pmatrix}, \begin{pmatrix} -\mathbf{v} \\ (\mathbf{I}_d - 2\gamma\mathbf{A_x})\mathbf{x} \end{pmatrix} \right\rangle \\
&= 0.
\end{aligned}
$$

Therefore μ is an invariant measure of (21). This invariant measure is unique since (21) is a linear SDE.

We denote the distribution of the numerical reverse process (16) by $\overleftarrow{\rho}_t$ and its $\mathbf{x}$ and $\mathbf{v}$-marginal distribution by $\overleftarrow{\rho}_{\mathbf{x},t}$ and $\overleftarrow{\rho}_{\mathbf{v},t}$ respectively. We aim to bound $\mathrm{TV}(\overleftarrow{\rho}_{\mathbf{x},0}, p_{\mathrm{data}})$. To achieve this, we first bound $\mathrm{TV}(\overleftarrow{\rho}_0, \overrightarrow{\rho}_0)$ and then apply the Data-Processing Inequality. We lay out three standard assumptions following Chen et al. (2023b) to conduct our analysis.

Theorem 1 (Fixed Generation Quality)**.** *Assume assumptions A1, A3, A4 and A5 hold. The generated data distribution is close to the data distributions p_{data} such that*

$$\mathrm{TV}(\overleftarrow{\rho}_{\mathbf{x},0}, p_{data}) \lesssim \underbrace{\sqrt{\mathrm{KL}(p_{data}\|\mu_{\mathbf{x}}) + \mathrm{FI}(p_{data}\|\mu_{\mathbf{x}})}\exp(-T)}_{\textit{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\textit{discretization error}} + \underbrace{\epsilon_{score}\sqrt{T}}_{\textit{score estimation}},$$

where $\mu_{\mathbf{x}}$ is the $\mathbf{x}$-marginal distribution of μ defined in Lemma 1.

Proof   Following Chen et al. (2023a), we employ the chain rule for KL divergence and obtain:

$$\mathrm{KL}(\overrightarrow{\rho}_0\|\overleftarrow{\rho}_0) \leq \mathrm{KL}(\overrightarrow{\rho}_T\|\overleftarrow{\rho}_T) + \mathbb{E}_{\overrightarrow{\rho}_T(\mathbf{a})}[\mathrm{KL}(\overrightarrow{\rho}_{0|T}(\cdot\|\mathbf{a})|\overleftarrow{\rho}_{0|T}(\cdot\|\mathbf{a})],$$

where $\overrightarrow{\rho}_{0|T}$ is the conditional distribution of $\mathbf{a}_0$ given $\mathbf{a}_T$ and likewise for $\overleftarrow{\rho}_{0|T}$. Note that the two terms correspond to the convergence of the forward and reverse process respectively. We proceed to prove that

Part I: Forward process $\mathrm{KL}(\overrightarrow{\rho}_T\|\overleftarrow{\rho}_T) \lesssim (\mathrm{KL}(p_{\mathrm{data}}\|\mu_{\mathbf{x}}) + \mathrm{FI}(p_{\mathrm{data}}\|\mu_{\mathbf{x}}))e^{-T},$

Part II: Backward process $\mathbb{E}_{\overrightarrow{\rho}_T(\mathbf{x})}[\mathrm{KL}(\overrightarrow{\rho}_{0|T}(\cdot|\mathbf{x})\|\overleftarrow{\rho}_{0|T}(\cdot|\mathbf{x})] \lesssim (L^2dh + L^2m_2^2h^2)T + \epsilon_{\mathrm{score}}^2 T.$

Part I: Note that $\overleftarrow{\rho}_T = \mu$, where μ is the invariant measure of (21). Following Ma et al. (2021), we construct the Lyapunov Function

$$\mathrm{L}(\overrightarrow{\rho}_t) := \mathrm{KL}(\overrightarrow{\rho}_t|\mu) + \mathbb{E}_{\overrightarrow{\rho}_t}\left[\left\langle \nabla \ln \frac{\overrightarrow{\rho}_t}{\mu}, S\nabla \ln \frac{\overrightarrow{\rho}_t}{\mu} \right\rangle \right] \tag{22}$$

for some positive definite matrix S. Since the Gaussian distribution μ satisfies the log-Sobolev inequality, one can show that there exists a constant $c > 0$ such that $\frac{d}{dt}\mathrm{L}(\overrightarrow{\rho}_t) \leq -c\mathrm{L}(\overrightarrow{\rho}_t)$. Here c depends on β, γ and the log-sobolev constant of μ. Thus,

$$\mathrm{KL}(\overrightarrow{\rho}_T|\mu) \leq \mathrm{L}(\overrightarrow{\rho}_T) \leq \mathrm{L}(\overrightarrow{\rho}_0)e^{-cT} \lesssim (\mathrm{KL}(p_{\mathrm{data}}\|\mu_{\mathbf{x}}) + \mathrm{FI}(p_{\mathrm{data}}\|\mu_{\mathbf{x}}))e^{-T}.$$

For the detailed proof, we refer readers to Appendix C of Ma et al. (2021), as the argument closely follows similar reasoning. For brevity, we omit it here.

Part II: The proof for the convergence of the reverse process is essentially identical to Theorem 15 of Chen et al. (2023b), with the only potential replacements being instances of $\left\| \begin{pmatrix} 0 & \mathbf{I}_d \\ \mathbf{I}_d & \gamma\mathbf{I}_d \end{pmatrix} \right\|_{op}$ with $\|\mathbf{D}\|_{op}$. However, they are equivalent due to Assumption A1. Therefore, we omit the proof here.

Combining the results from Parts I and II, and applying Pinsker's Inequality, we obtain

$$\mathrm{TV}(\overleftarrow{\rho}_0, \overrightarrow{\rho}_0) \lesssim \sqrt{\mathrm{KL}(p_{\mathrm{data}}\|\mu_{\mathbf{x}}) + \mathrm{FI}(p_{\mathrm{data}}\|\mu_{\mathbf{x}})}\exp(-T) + (L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T} + \epsilon_{\mathrm{score}}\sqrt{T}.$$

The final result follows by applying the Data-Processing Inequality to transition from $\overrightarrow{\rho}_0$ and $\overleftarrow{\rho}_0$ to their respective $\mathbf{x}$-marginals.

Remark 1. *The proof could be potentially generalized to the case where $\mathbf{A}_{\mathbf{x},t}$ and $\mathbf{A}_{\mathbf{v},t}$ are time-varying. First, for the backward process, note that the proof of Theorem 15 in Chen et al. (2023b) relies only on the score estimation and the Lipschitz property of the score function, which does not require the drift of the forward process to be time-invariant. Second, for the forward process, for a fixed $T > 0$, one can always consider a modified version of (5) with time-averaged drift given by (21), where $\mathbf{D} = \frac{1}{T}\int_0^T \mathbf{D}_t, \mathrm{d}t$. Thus, for the same initial condition $\mathbf{a}_0$, (5) and (21) will generate the same distribution at time T.*

B.2 Convergence of Variational Scores

$\mathbf{A}_{\mathbf{a},t}^{(k)}$ tracks a mean-field ODE, which converges to the equilibrium $\mathbf{A}_{\mathbf{a},t}^{\star}$ if we can establish the stability condition of the mean-field ODE. As such, we can ensure that the perturbations caused by errors in Theorem 1 result in, at most, similar variations in subsequent iterates.

The following is a restatement of Lemma 2 in Deng et al. (2024b)

Lemma 2 (Local stabilty). *Given assumptions A2 and A3, we can identify a local stability condition for any $\mathbf{A} \in \boldsymbol{\Theta}$ such that*

$$\langle \mathbf{A} - \mathbf{A}_{\mathbf{a},t}^{\star}, \nabla\overrightarrow{\mathbf{L}}_t(\mathbf{A})\rangle \geq m\|\mathbf{A} - \mathbf{A}_{\mathbf{a},t}^{\star}\|_2^2.$$

Next, we assume the step size η_k follows the tradition in stochastic approximation (Benveniste et al., 1990).

Assumption A6 (Step size).

$$0 < \eta_{k+1} < \eta_k, \quad \sum_{k=1}^{\infty}\eta_k = +\infty, \quad \sum_{k=1}^{\infty}\eta_k^{2\alpha} < \infty, \quad \alpha \in \left(\frac{1}{2}, 1\right].$$

The next result is a restatement of Theorem 2 in Deng et al. (2024b) to prove the convergence of the variational scores.

Theorem 2 (Convergence in L^2). *Given assumptions A2 - A6 and a large enough k, the variational score $\mathbf{A}_{\mathbf{a},t}^{(k)}$ in algorithm 2 converges to a local equilibrium $\mathbf{A}_{\mathbf{a},t}^{\star}$ that motivates efficient transport such that*

$$\mathbb{E}_{\overleftarrow{\rho}_{\mathbf{a},t}^{(k)}}[\|\mathbf{A}_{\mathbf{a},t}^{(k)} - \mathbf{A}_{\mathbf{a},t}^{\star}\|_2^2] \leq 2\eta_k.$$

B.3 Adaptive Generation Quality

Theorem 2 shows that the non-optimized $\mathbf{A}_{\mathbf{a},t}^{(k)}$ converges to the equilibrium $\mathbf{A}_{\mathbf{a},t}^{\star}$, where the latter yields efficient transportation plans. Combining the study of the sample quality in Theorem 1 based on a fixed $\mathbf{A}_{\mathbf{a},t}^{(k)}$, we can evaluate the adaptive sample quality based on $\mathbf{A}_{\mathbf{a},t}^{\star}$, which yields more and more efficient transportation plans in the long-time limit. The following is a natural extension of Theorem 3 in Deng et al. (2024b) since both algorithms follow from the framework of multivariate diffusion:

Theorem 3. *Assume assumptions A1-A6 hold. The adaptively generated sample at stage k based on the equilibrium $\mathbf{A}_t^{\star}$ with efficient transportation plans is close in total variation (TV) distance to the real sample such that*

$$\mathrm{TV}(\overleftarrow{\rho}_{0,\mathbf{x}}^{\star}, \rho_{data}) \lesssim \underbrace{\sqrt{\mathrm{KL}(\rho_{data}\|\mu^{\mathbf{x}}) + \mathrm{FI}(\rho_{data}\|\mu^{\mathbf{x}})}\exp(-T)}_{convergence\ of\ forward\ process} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{discretization\ error} + \underbrace{(\epsilon_{score} + \sqrt{\eta_k})\sqrt{T}}_{adaptive\ score\ estimation}.$$

C   Experimental Details

We present more details on the images experiments. We consider the same U-net architecture as that used in Dhariwal and Nichol (2022) and implemented by Karras et al. (2022). We normalize the images to the interval [-1,1] and use a horizontal flip as only data augmentation. The total training time is 1 day and 19 hours. We present a table of hyperparameters used during training (3):

Table 3: Table of hyperparameters used during training

Parameter	Value
Forward Score learning rate	3e-4
Backward Score learning rate	3e-6
Num Iters	280K
EMA Beta	.9999
Sampling Time Steps	100
Batch Size	256
Damping parameter	1.

Our method results in the following table of FID values (2). Our method achieves competitive results among the compared methods. Additionally, results could be further improved by using more advanced preconditioning and data augmentation techniques like those presented in Karras et al. (2022). We delay this detailed investigation for future work. Our current result demonstrates the scalability of the method in high dimensions. We now present samples from different pretrained models and ours. We show that our method can product crisp and clean images. We use the EDM repository for samples from EDM Karras et al. (2022) and VP Song et al. (2021) to generate samples from baselines.

C.1   CIFAR-10 Samples from different methods

Figure 7: Unconditional Samples generated by our method on CIFAR-10

D   Time Series Forecasts

In this section, we present more forecasts generated using different methods
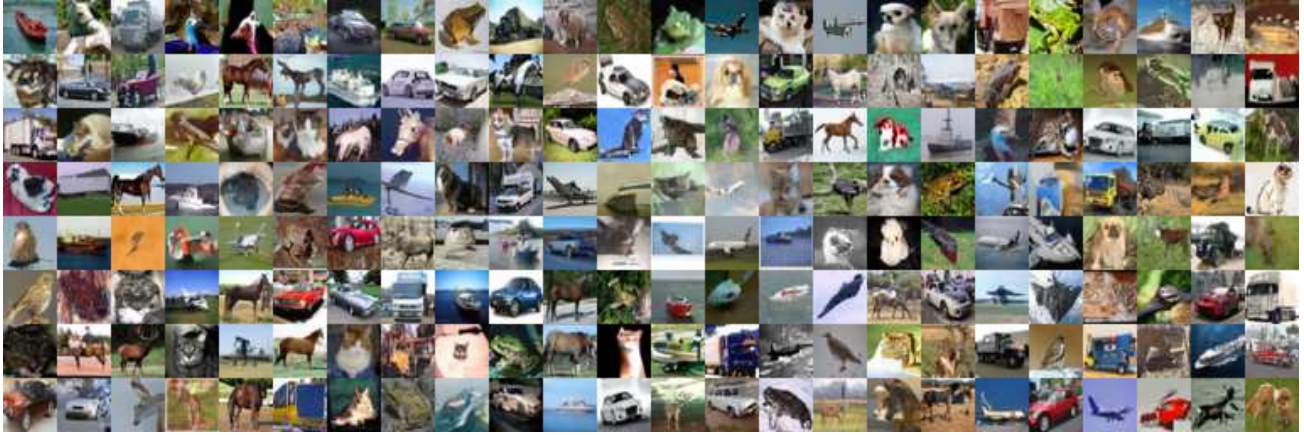
Figure 8: Unconditional Samples generated by VP on CIFAR-10

Figure 9: Unconditional Samples generated by EDM on CIFAR-10
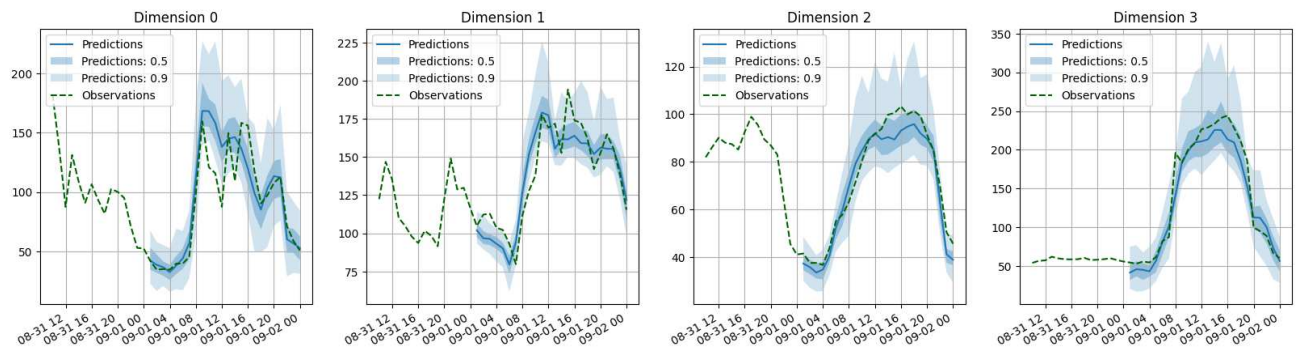
Figure 10: Sample forecasts of VSCLD in the electricity dataset
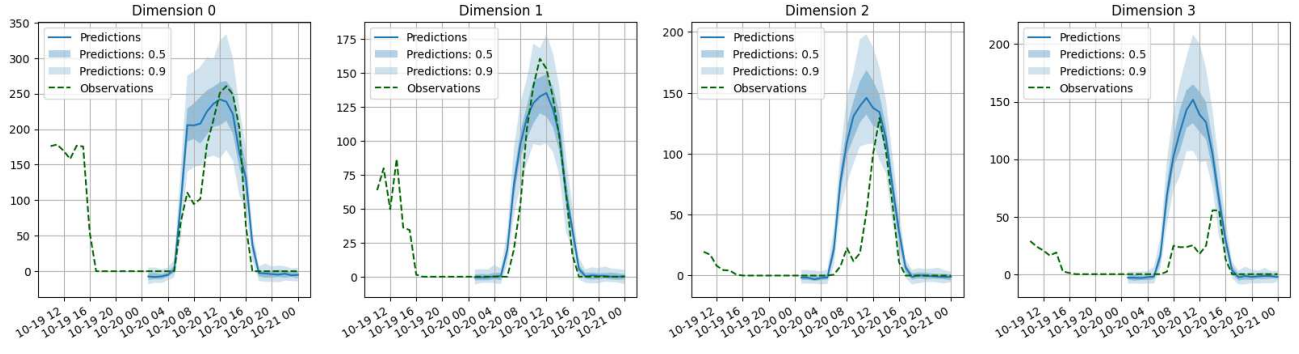
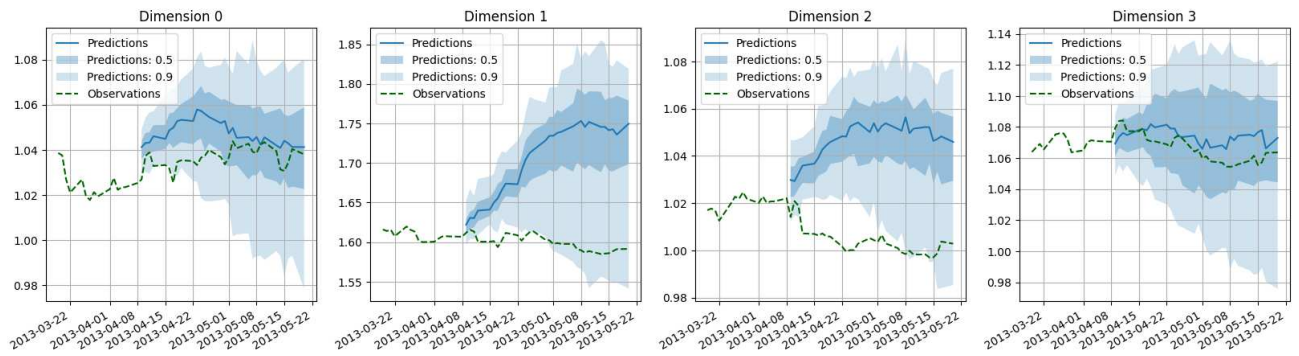Figure 11: Sample forecasts of VSCLD in the solar dataset

Figure 12: Sample forecasts of VSCLD in the exchange rate dataset
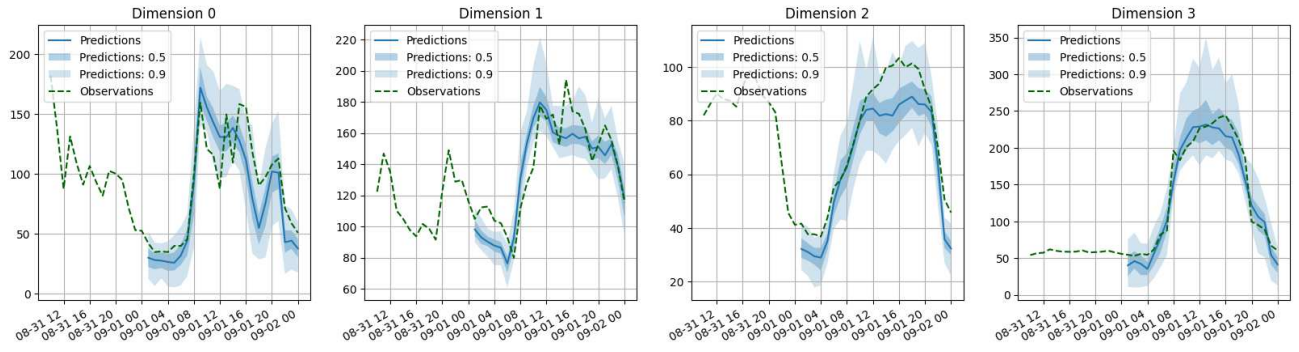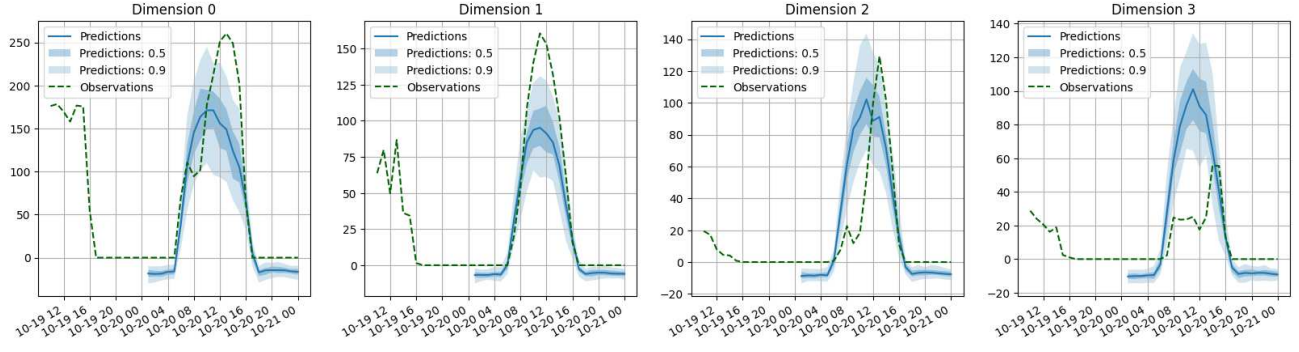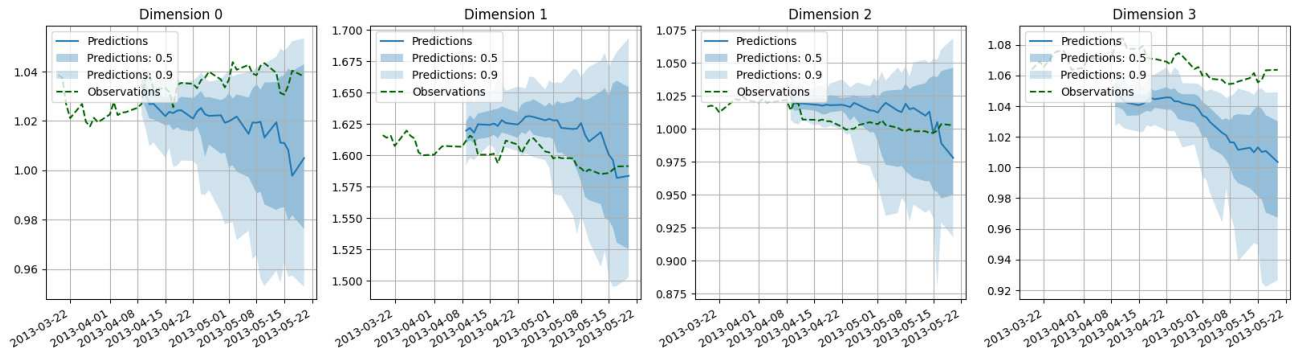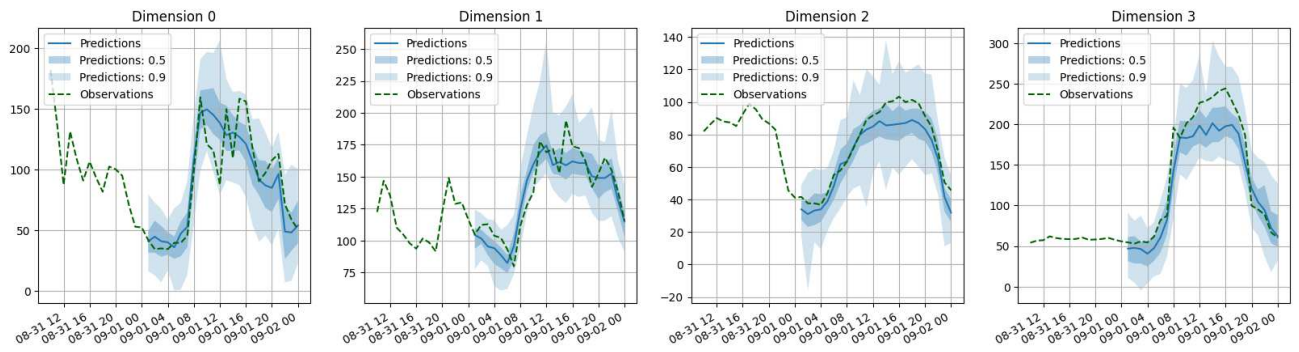

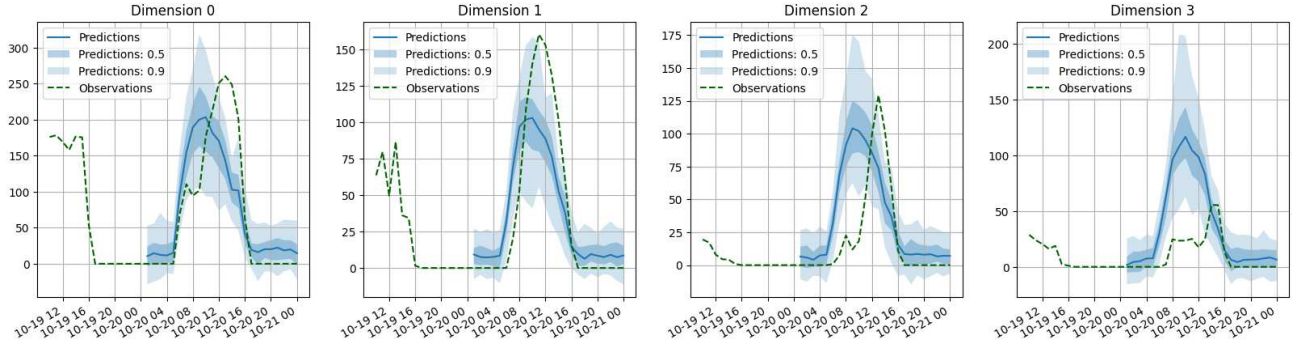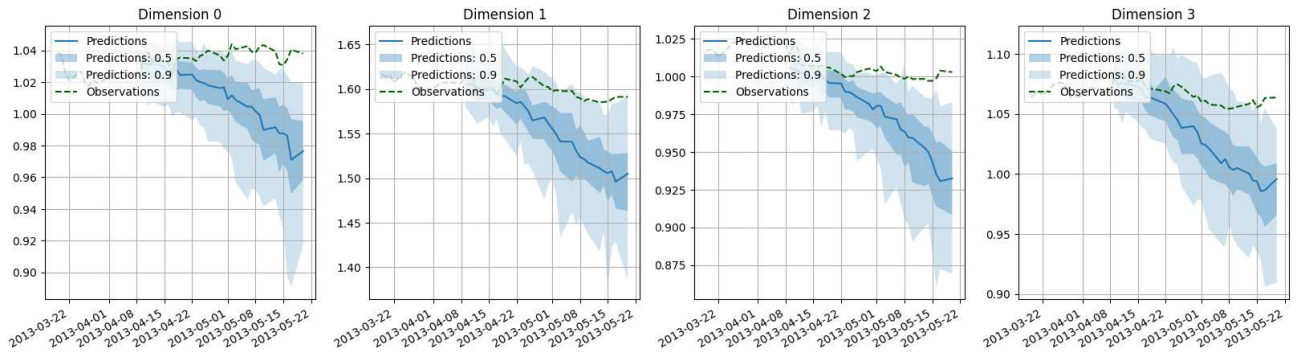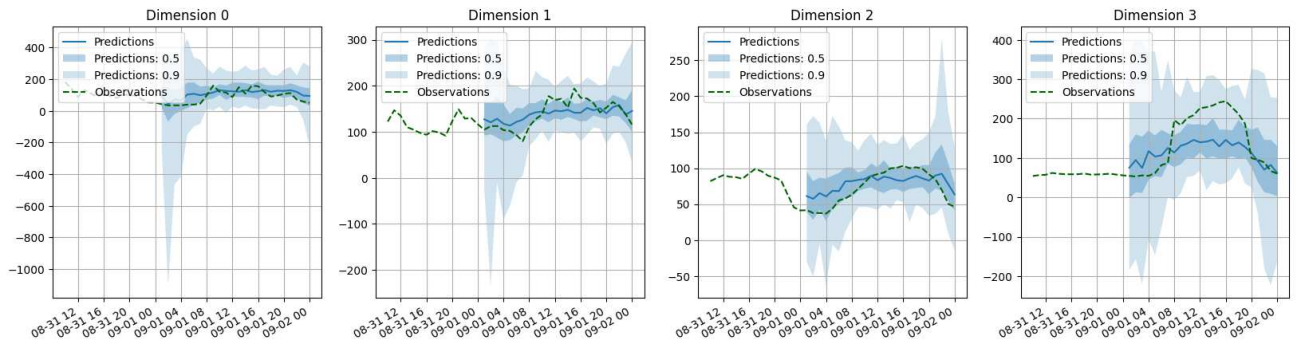Figure 13: Sample forecasts of VSULD in the electricity dataset

D.1   Samples for VSCLD

D.2   Samples for VSULD

D.3   Forecasts for VSDM

D.4   Forecasts for CLD

Figure 14: Sample forecasts of VSULD in the solar dataset

Figure 15: Sample forecasts of VSULD in the exchange rate dataset

Figure 16: Sample forecasts of VSDM in the electricity dataset

Figure 17: Sample forecasts of VSDM in the solar dataset

Figure 18: Sample forecasts of VSDM in the exchange rate dataset

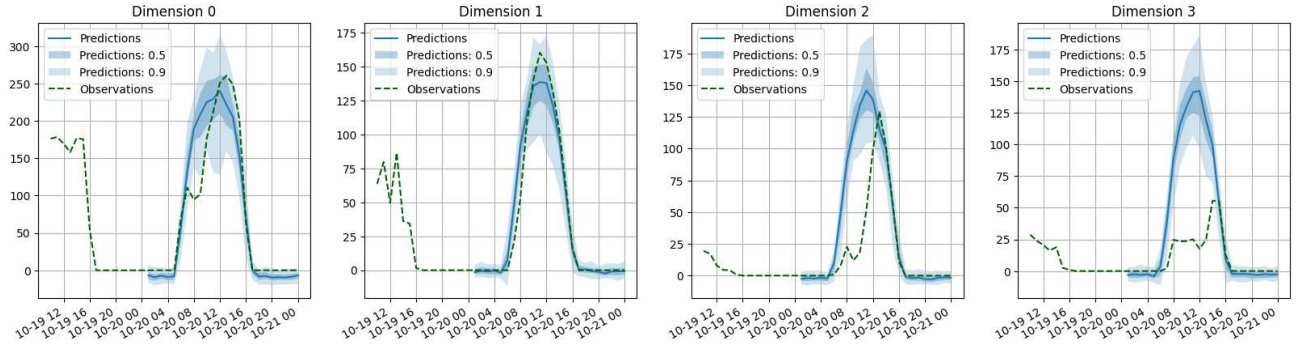Figure 19: Sample forecasts of CLD in the electricity dataset

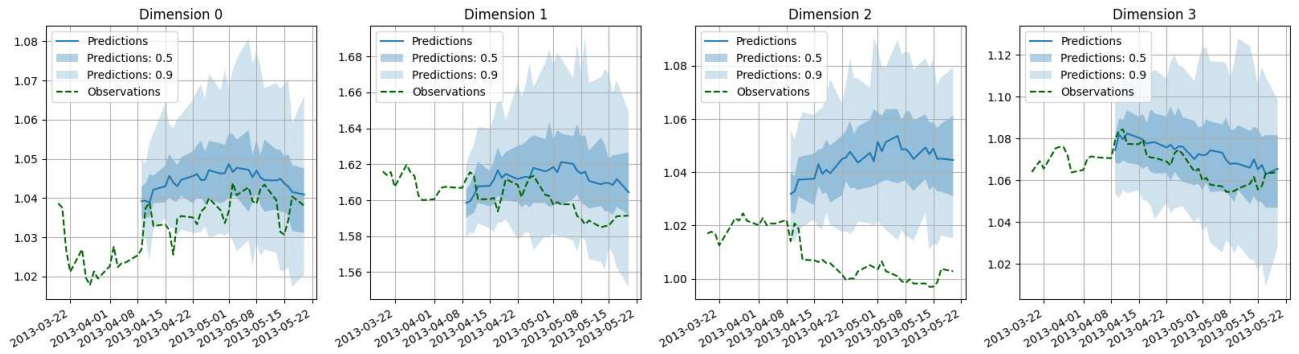Figure 20: Sample forecasts of CLD in the solar dataset

Figure 21: Sample forecasts of CLD in the exchange rate dataset