Parametric model reduction of mean-field and stochastic systems via higher-order action matching

Anonymous Author(s)
Affiliation
Address
email

Abstract

The aim of this work is to learn models of population dynamics of physical systems that feature stochastic and mean-field effects and that depend on physics parameters. The learned models can act as surrogates of classical numerical models to efficiently predict the system behavior over the physics parameters. Building on the Benamou-Brenier formula from optimal transport and action matching, we use a variational problem to infer parameter- and time-dependent gradient fields that represent approximations of the population dynamics. The inferred gradient fields can then be used to rapidly generate sample trajectories that mimic the dynamics of the physical system on a population level over varying physics parameters. We show that combining Monte Carlo sampling with higher-order quadrature rules is critical for accurately estimating the training objective from sample data and for stabilizing the training process. We demonstrate on Vlasov-Poisson instabilities as well as on high-dimensional particle and chaotic systems that our approach accurately predicts population dynamics over a wide range of parameters and outperforms state-of-the-art diffusion-based and flow-based modeling that simply condition on time and physics parameters.

1 Introduction

Predicting the behavior of time-dependent processes $X_{t,\mu}$ over time t and across varying physics parameters μ is a key challenge in computational science and engineering [45, 64]. The dynamics of $X_{t,\mu}$ typically are described by systems of (stochastic) differential equations, which are derived from physics models and can be computationally expensive to simulate [39, 31]. Thus, it is desirable to learn reduced or surrogate models that can be rapidly evaluated to predict the system behavior across varying physics parameters [72, 10, 11, 44].

Reduced modeling via learning population dynamics Given a data set of samples, i.e., realizations of the random variable $X_{t,\mu}$ over a suitable domain $\mathcal{X} \subseteq \mathbb{R}^d$,

$$\{X^i_{t_j,\mu_k} \,|\, i = 1, \dots, N_x, \quad j = 1, \dots, N_t, \quad k = 1, \dots, N_\mu\} \subset \mathcal{X}, \tag{1}$$

we aim to learn a dynamical-system reduced model to rapidly predict samples that approximately follow the same law $\rho_{t,\mu}$ as $X_{t,\mu}$ over time t and varying physics parameter μ. We refer to the evolution of $\rho_{t,\mu}$ in time as population dynamics. Learning the population dynamics instead of learning the dynamics of the individual trajectories $t \mapsto X^i_{t,\mu}$ for all $i = 1, \dots, N_x$ and μ can be beneficial: There are cases where $\rho_{t,\mu}$ does not change in time,

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

yet every sample trajectory $t \mapsto X^i_{t,\mu}$ follows complicated dynamics. For example, consider incompressible fluid dynamics with constant density. Samples corresponding to particles that comprise the fluid can have complicated trajectories, whereas on a distribution level, the density of the fluid is constant and so are the population dynamics. Furthermore, learning population dynamics seamlessly treats deterministic and stochastic systems because on the density level $\rho_{t,\mu}$ it is irrelevant if the particles are stochastic or deterministic.

Our approach: Learning parametric minimal energy vector fields that represent population dynamics Building on standard literature on optimal transport theory [8] as well as the so-called action-matching loss introduced in [60], we pose a variational problem to learn gradient fields $\nabla s_{t,\mu}$ so that the continuity equation corresponding to the vector field given by $\nabla s_{t,\mu}$ approximates the population dynamics $\rho_{t,\mu}$ of the samples (1). In the spirit of reduced modeling [72, 10, 11, 44], we seek a vector field $s_{t,\mu}$ that generalizes to different values of the physics parameters μ. We therefore optimize for $s_{t,\mu}$ that minimizes the average objective of the variational problem over all parameters $\mu \sim \nu$, where ν describes the distribution of parameters on the domain $\mathcal{D} \subset \mathbb{R}^p$. We parametrize $s_{t,\mu}$ with a neural network with weight modulation [38, 12] so that it can be evaluated quickly over t and μ.

Rapid sample generation in inference phase Predictions at inference time at new physics parameters μ are made by sampling based on the vector field $\nabla s_{t,\mu}$, which means that our approach represents $\rho_{t,\mu}$ through the application of $\nabla s_{t,\mu}$ on an initial condition. Importantly, time t in the inference step corresponds to the time of the physics problem so that in one inference step a whole sample trajectory is obtained, rather than a sample at one specific time point as in regular conditioning-based methods (see literature review). Thus, we can rapidly generate samples that approximately follow the law $\rho_{t,\mu}$ in the inference phase.

Stabilizing training with higher-order quadrature An important part of our contribution is stabilizing the training procedure by accurately estimating the objective of the variational problems from few data samples. In particular, instead of uniformly mini-batching over the data (1), we introduce an empirical loss that builds on higher-order Gauss-Legendre quadrature [26] in the time direction so that the learned $\nabla s_{t,\mu}$ accurately captures the dynamics over time t. Consequently, we refer to our approach as higher-order action matching (HOAM). Our numerical experiments show that the higher-order quadrature in the empirical loss is key for learning gradient fields $\nabla s_{t,\mu}$ that accurately capture the evolution in time t and that generalize across physics parameters μ.

Literature review We review relevant literature; see Figure 1 for an overview.

Non-intrusive and data-driven surrogate modeling There is a range of surrogate and latent modeling methods that aim to learn the sample dynamics of the realizations rather than the population dynamics, such as dynamic mode decomposition and Koopman-based methods [71, 76, 86, 45, 57, 92, 16] as well as neural network-based methods such as neural ordinary differential equations [19, 27, 47]. There also are extensions to stochastic systems [50, 41, 88, 19, 27, 73]. However, all of these methods ignore physics parameter dependencies and/or aim to learn the sample dynamics, whereas we focus on parametric population dynamics.

Population dynamics and trajectory inference Learning population dynamics has been considered extensively in computational biology in the context of gene expression, where the focus is on learning from independent samples at selected time points rather than from sample trajectories [33, 29, 93, 75, 85, 46]; however, many of these approaches [17, 84] are simulation-based and thus require integrating dynamics during the training or parameterizing the density additionally to the vector field. These works also are not concerned with generalizing over a range of physics parameters in many cases.

Diffusion- and flow-based modeling There is a large body of work on diffusion-based [91, 79, 35, 40, 81, 82] and flow-based modeling [2, 53]; see [1] for a detailed review. These approaches are not taking into account time t because they learn paths between a reference and a target distribution only. There are works that condition on time t and a parameter μ such as [68, 13, 25, 36, 32, 37, 51], but this requires then generating a path for each time step at inference time, which is computationally expensive. Furthermore, the conditioning on time t means that the target distribution $\rho_{t,\mu}$ at each time t and μ is different, and thus a separate

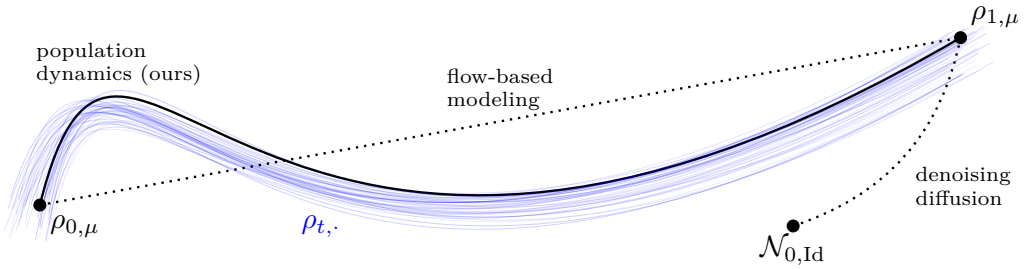

Figure 1: Parametric model reduction with our HOAM seeks to learn vector fields that represent population dynamics $\rho_{t,\mu}$ over time t and that generalize over varying physics parameter μ. In contrast, parametric model reduction with score-based diffusion denoising and flow-based modeling requires conditioning on time t and physics parameter μ, which leads to separate, costly inference steps for each time step and μ of a sample trajectory.

hyper-parameter tuning can be required, which is impractical over many time steps and physics parameters as in our physics problems; see our numerical experiments. The works [14, 78, 49] compute transport-based solutions but parametrize different quantities than our approach, require actively sampling data, and ignore physics parameters μ. We note that there also is work on forecasting with diffusion- and flow-based modeling [68, 61, 18, 20], which is a different task than our task of predicting across varying physics parameters.

Optimal transport Besides the machine learning literature, variational approaches for inferring vector fields are extensively used in optimal transport theory [5, 4]. Of particular importance to us is the formulation by Benamou and Brenier [8]. The Bennamou-Brenier formula describes a joint optimization problem over vector fields and paths in probability space and the action matching loss [60] is the restriction of this optimization problem to the case of a fixed path and the vector field parametrized by a neural network, which are core building blocks for us that we show can be used together with a parameter dependency.

Contributions We summarize our contributions:

(a) Developing a loss to learn population dynamics that remain valid across varying physics parameters by building on optimal transport literature [8] and action matching [60].
(b) Introducing higher-order quadrature schemes based on Gauss-Legendre quadrature for the loss to efficiently couple the gradient fields over time. This leads to lower variance estimators of the loss that critically stabilize training.
(c) Parameterizing the vector fields with networks based on weight modulation to efficiently capture the parameter dependency and to ensure rapid prediction over t and μ.
(d) Demonstrating on a range of physics problems from Vlasov-Poisson instabilities to high-dimensional chaotic systems that our approach leads to (i) accurate predictions and (ii) fast inference of samples due to modeling physics time and so outperforms standard diffusion- and flow-based modeling that condition on time and physics parameters on our examples.

We provide an implementation of our method at `http://github.com/REDACTED`.

2 Method

2.1 Parameter-dependent population dynamics

Continuity equation Let us consider data (1) corresponding to the probability measure $\rho_{t,\mu}$, which is absolutely continuous for $t \in [0,1]$ and $\mu \in \mathcal{D}$. We use the same notation for the measure and its density. The density ν of μ is a continuous function on $\mathcal{D}$. We consider population dynamics of $X_{t,\mu} \sim \rho_{t,\mu}$ that can be described by the continuity equation

$$\partial_t \rho_{t,\mu} = -\nabla \cdot (\rho_{t,\mu} v_{t,\mu}), \qquad \text{for all } t \in [0,1], \mu \in \mathcal{D}, \tag{2}$$

with the initial condition $\rho_{t=0,\mu} =: \rho_{0,\mu}$ and vector field $v_{t,\mu}$. Notice that in our case the continuity equation (2) depends on the physics parameter $\mu \sim \nu$. There can be many vector fields $v_{t,\mu}$ that lead to the same population dynamics (2). For example, if $v_{t,\mu}$ is a

vector field that describes the dynamics of $\rho_{t,\mu}$ via (2), then another vector field is given by $v'_{t,\mu} = v_{t,\mu} + w/\rho_{t,\mu}$ with any other w that satisfies $\nabla \times w = 0$ as long as $\rho_{t,\mu}$ is positive.

Uniqueness via gradient fields and the corresponding elliptic problems Because we aim to learn a vector field from sample data (1) that describes the population dynamics (2) of the corresponding law $\rho_{t,\mu}$, it is helpful to remove this non-uniqueness. One way to do so is to restrict the vector field to $v_{t,\mu} = \nabla s_{t,\mu}$ so that it is a gradient field [4, p. 45]. Plugging $v_{t,\mu} = \nabla s_{t,\mu}$ into (2), together with the assumptions $\rho_{t,\mu} > 0$ and $\int_{\mathcal{X}} \partial_t \rho_{t,\mu} \mathrm{d}x = 0$, leads to parametric elliptic problems in $s_{t,\mu}$

$$-\nabla \cdot (\rho_{t,\mu} \nabla s_{t,\mu}) = \partial_t \rho_{t,\mu}\,, \tag{3}$$

with coefficient function $\rho_{t,\mu}$, right-hand side (source term) $\partial_t \rho_{t,\mu}$, and homogeneous Neumann boundary conditions $\rho_{t,\mu} \nabla s_{t,\mu} \cdot \hat{n} = 0$ on $\partial\mathcal{X}$ with normal vector $\hat{n}$ for all $t \in [0,1]$ and $\mu \in \mathcal{D}$. The weak forms of the elliptic problems (3) lead to energy minimization problems that are useful when aiming to learn a gradient field $s_{t,\mu}$ via optimization,

$$\min_{s \in H^1(\rho_{t,\mu}, \mathcal{X})} E_{t,\mu}(s) := \min_{s \in H^1(\rho_{t,\mu}, \mathcal{X})} \frac{1}{2} \int_{\mathcal{X}} |\nabla s|^2 \rho_{t,\mu} \mathrm{d}x - \int_{\mathcal{X}} \partial_t \rho_{t,\mu} s \mathrm{d}x \tag{4}$$

for each $t \in [0,1]$ and $\mu \in \mathcal{D}$. The space $H^1(\rho_{t,\mu}, \mathcal{X})$ contains functions s with $\int_{\mathcal{X}} |s|^2 \rho_{t,\mu} \mathrm{d}x < \infty$, which is the energy norm corresponding to the $\rho_{t,\mu}$-weighted inner product [28, Sec. 2.3.2].

Optimal transport Standard elliptic theory guarantees unique solutions up to constants of (4) in the Sobolev space $H^1(\mathcal{X})$ under strong assumptions on $\rho_{t,\mu}$ such as uniform boundedness by a positive constant for all t and μ; see [28, Proposition 2.2] and [11, Section 3.2]. The theory of optimal transport allows treating the much more general case when $\rho_{t,\mu}$ is not uniformly bounded away from zero; we refer to [8] and [74, Section 5.3.1] for details. To see the connection, notice that among all vector fields $v_{t,\mu}$ that are compatible to $\rho_{t,\mu}$ in the sense of (2), gradient fields $\nabla s_{t,\mu}$ have the smallest associated kinetic energy $\frac{1}{2} \int_{\mathcal{X}} |v|^2 \rho_{t,\mu} \mathrm{d}x$, which is the objective considered in [8]. In the language of optimal transport and in particular the formalism of [62], vector fields with minimal kinetic energy describe tangent vectors to the curve $t \mapsto \rho_{t,\mu}$. The metric is the inner product of $L^2(\rho_{t,\mu}, \mathcal{X}, \mathbb{R}^d)$. This is the weak Riemannian structure of $\mathcal{P}(\mathcal{X})$ equipped with the Kantorovich-Rubinstein metric and described in detail in [5, Chapter 8]. We give a short description in Appendix E.

Energy functional with entropy term Instead of the energy (4), we can also use other choices of the energy to select gradient fields, as long as energy functions are convex to maintain uniqueness. We consider an energy that is based on a different notion of distance on $\mathcal{P}(\mathcal{X})$, which is commonly referred to as the entropic optimal transport or Schrödinger bridge problem [77, 55],

$$E^\epsilon_{t,\mu}(s) = \frac{1}{2} \int_{\mathcal{X}} |\nabla (s - \frac{\epsilon^2}{2} \log \rho_{t,\mu})|^2 \rho_{t,\mu} \mathrm{d}x - \int_{\mathcal{X}} \partial_t \rho_{t,\mu} s \mathrm{d}x\,, \tag{5}$$

which depends on $\epsilon \geq 0$. The energy $E^\epsilon_{t,\mu}$ is of particular interest for two reasons: One, the Euler-Lagrange equation of (5) in strong form is the Fokker-Planck equation for $s^\epsilon_{t,\mu}$: $\partial_t \rho_{t,\mu} = -\nabla \cdot (\rho_{t,\mu} \nabla s^\epsilon_{t,\mu}) + \frac{\epsilon^2}{2} \Delta \rho_{t,\mu}$, again with homogeneous Neumann boundary conditions for all $t \in [0,1]$ and $\mu \in \mathcal{D}$; see Appendix C. This means we can efficiently generate samples after learning $s^\epsilon_{t,\mu}$ via corresponding stochastic differential equations (SDEs). Two, it can be interpreted as regularizing the field s, which we discuss in Appendix C.

2.2 Loss for learning vector fields over time t and physics parameter μ

Variational formulation over t and μ So far we just carried along time t and physics parameter μ but did not address them in the variational problems, i.e., we had separate variational problems (4) for all $t \in [0,1]$ and $\mu \sim \nu$. We now propose to consider the average energy over t and μ to infer a map $s : [0,1] \times \mathcal{D} \to H^1(\rho_{t,\mu}, \mathcal{X}), (t,\mu) \mapsto s_{t,\mu}$, which is called a solution map in reduced modeling [72, 10, 11, 44],

$$\min_{s : [0,1] \times \mathcal{D} \to H^1(\rho_{t,\mu}, \mathcal{X})} E(s) := \min_s \int_{\mathcal{D}} \int_0^1 E^\epsilon_{t,\mu}(s) \,\mathrm{d}t \,\mathrm{d}\nu(\mu). \tag{6}$$

Notice that time t and physics parameter μ have two different effects on the gradient field $\nabla s_{t,\mu}$: Time t couples the elliptic problems (i.e., (3) for $\epsilon = 0$) via the time derivative $\partial_t \rho_{t,\mu}$; see Appendix D. In contrast, the elliptic problems are uncoupled over μ and can be considered separately. This means that to compute the solution to an elliptic problem (i.e., (3) in case of $\epsilon = 0$) for one value of $\mu \in \mathcal{D}$, one does not need to consider any other $\mu' \in \mathcal{D}$. This will allow us to sample the physics parameters over $\mathcal{D}$ independently from each other when estimating the corresponding loss, whereas we will use higher-order quadrature to obtain an accurate approximation of the time integral to ensure the coupling between the time points is reflected in $s_{t,\mu}$; see Section 2.3.

Loss for learning gradient fields from samples over t and μ The energy $E_{t,\mu}$ defined in (4) as well as the energy $E^\epsilon_{t,\mu}$ defined in (5) leads to a loss that can be estimated from samples (1). The quantity $\partial_t \rho_{t,\mu}$ appears in (4) and (5), which is typically unavailable when we have access to data samples (1) only. Integration by parts of the term involving $\partial_t \rho_{t,\mu}$ eliminates it, see also Appendix D. We arrive at

$$E^\epsilon(s) = \int_{\mathcal{D}} \int_0^1 \int_{\mathcal{X}} \frac{1}{2}|\nabla s_{t,\mu}|^2 + \partial_t s_{t,\mu} + \frac{\epsilon^2}{2}\Delta s_{t,\mu} \rho_{t,\mu} \mathrm{d}x\mathrm{d}t + \int_{\mathcal{X}} s_{t,\mu}\rho_{t,\mu}\mathrm{d}x \Big|_{t=0}^{t=1} \mathrm{d}\nu(\mu)\,, \quad (7)$$

which corresponds to $E_{t,\mu}$ for $\epsilon = 0$ and to $E^\epsilon_{t,\mu}$ for $\epsilon > 0$.

Remark 1. *Loss functions of the form as* (7) *but without the parameter dependence have been used in [60] and [46, Theorem 2.1]. In fact, the case with $\epsilon = 0$ appears already in [8, Equation 35] and [63, Section 3]. We build on these results but work with population dynamics that depend on physics parameters, which leads the loss shown in* (7).

2.3 Parameterizing the vector field, estimating the loss from data, sampling

Parametrizing $s_{t,\mu}$ with weight modulations We parametrize the vector field $s_{t,\mu}$ via a neural network with continuous versions of low-rank adaptation layers [38], which have been successfully used for parametric model reduction of deterministic time-dependent dynamical systems [12]. The layers have the form $\mathcal{C}(x) = Wx + \phi(t,\mu)ABx + b$, where W is a weight matrix, A, B are low-rank matrices, b is a bias vector, and $\phi(t,\mu) \in \mathbb{R}$ is a scalar weight modulation; see Appendix B. Only the weight modulations $\phi(t,\mu)$ depend on time t and physics parameter μ. We use a hyper-network $h : [0,1] \times \mathcal{D} \times \Psi \to \mathbb{R}$ that depends on the weight vector $\psi \in \Psi \subseteq \mathbb{R}^q$ to map t and μ to the modulation weights $\phi(t,\mu) = h(t,\mu;\psi)$. The weights W, A, B, b, which are are independent of t and μ, over all layers are collected into the weight vector $\theta \in \Theta \subseteq \mathbb{R}^{q'}$. Typically $q \ll q'$. Using the hyper-network encourages continuity of $s_{t,\mu}$ in time t, which is key for many physics problems [12].

Combining higher-order quadrature and Monte Carlo sampling for estimating the loss from sample data Estimating the loss (7) from data can be challenging because of the three nested integrals (expectations) over the samples $X^i_{t,\mu}$, time t, and physics parameter μ: Recall that $N = N_x \times N_t \times N_\mu$ is the number of samples in the training data set (1). Using mini-batching during training, we can use $N^b \ll N$ samples with $N^b = N^b_x \times N^b_t \times N^b_\mu$. Plain uniform sampling over the data set (1) for mini-batching can lead to poor estimates of the loss. One reason is that typically $N_\mu \ll N_t \ll N_x$ and when uniformly sub-sampling for mini-batching then the samples are unbalanced between x, t, μ, which means that one of the three nested integrals can be poorly approximated. Poor estimation of the loss then quickly leads to instabilities in the training; see Section 3 and Figure 2.

We propose a combination of higher-order numerical quadrature and Monte Carlo sampling to estimate the loss (7). In the direction of time t, we propose to use Gauss-Legendre quadrature [26], which requires data at few time points so that many of the mini-batch data points can be taken over the samples $i = 1, \ldots, N_x$. It is important to accurately estimate the time integral because it ensures the coupling between the time points as well as the coupling to the boundary terms to match the path from $\rho_{0,\mu}$ at time $t = 0$ to $\rho_{1,\mu}$ at time $t = 1$. If the data set (1) does not contain samples at the Gauss-Legendre time nodes then we linearly interpolate the data at the Gauss-Legendre nodes. While we could use higher-order interpolation scheme at higher costs, note that the time integral is taken over estimates of

expectations over the samples, which already are crude estimates and thus interpolating them with high accuracy is unnecessary. Because the loss is decoupled over the physics parameter μ, we use one μ_k, $N_\mu^b = 1$, per mini-batch over all $\{\mu_1, \ldots, \mu_{N_\mu}\}$ in the training set. This gives the empirical loss, which is written out in Appendix A. The numerical experiments will show that the stabilization of the training via Gauss-Legendre quadrature is key.

Rapid predictions (inference) with learned reduced models Making predictions in the inference step means drawing samples that follow the law represented by the learned gradient field $\nabla s_{t,\mu}$, which approximates the law $\rho_{t,\mu}$ of $X_{t,\mu}$. Because we train with the loss (7), we integrate the SDE $\mathrm{d}\hat{X}_{t,\mu} = \nabla s_{t,\mu}(\hat{X}_{t,\mu})\mathrm{d}t + \epsilon \mathrm{d}W_t$, where W_t are Wiener processes and ϵ is the same ϵ that is used in the training loss (7); see Appendix C. As initial condition, we use samples from $\rho_{0,\mu}$ at time $t = 0$. Of course other sampling schemes can be used [70].

Notice that the time t in the SDE used for generating samples is the same time as of the physics problem and thus of the sample trajectory. This means that the costs of the inference step of our HOAM for generating a trajectory of length K scales as $\mathcal{O}(K)$. In contrast, introducing a conditioning on time and physics parameter in, e.g., noise-conditioned score matching (NCSM) [80] and conditional flow matching (CNF) or stochastic interpolants [2, 53] requires inferring a separate sampling path for each t and μ pair of interest. In particular, the inference costs of CFM scale as $\mathcal{O}(K\tau)$, where τ is the number of steps taken in the differential equation for generating one sample at one time point. For NCSM with annealed Langevin sampling, the inference costs scale as $\mathcal{O}(K\tau\sigma)$, where σ is the number of annealing steps. Contrasting this to the scaling of $\mathcal{O}(K)$ of our HOAM approach shows that HOAM is well suited for fast predictions over t and μ as required in parametric model reduction.

3 Numerical experiments

Examples We consider the following parametric dynamical systems; details in Appendix B.

1. Bimodal Duffing oscillator A collection of particles evolves in two-dimensional phase-space governed by the bimodal Duffing oscillator dynamics [66, Sec. 4.2.2]. After a transient phase, the particles are distributed according to a bimodal distribution. The parameter $\mu \in [0.1, 0.3]$ determines the position of the modes of the stationary distribution.

2. Two-stream instability We numerically solve the Vlasov-Poisson partial differential equations using a particle-in-cell method to generate samples (1). We consider the two-stream instability [21, 42] in a 1D1V configuration with collisions [87, Sec 2(b)(i)], with $\beta = 10^{-3}$ and $v_0 = 1$ as in [48]. This makes the problem stochastic. The parameter $\mu \in [1.2, 1.9]$ is a normalization constant related to the Debye length [83] that controls the wave number.

3. Bump-on-tail instability Using the same numerical setup of the Vlasov-Poisson equation as for the two-stream instability, we also consider the the bump-on-tail instability [7, 34, 42]. The parameter varies as $\mu \in [1.3, 2.0]$.

4. Strong Landau damping We consider the strong Landau damping phenomenon that is governed by Vlasov-Poisson partial differential equations again but now in a 3D3V (six-dimensional) setup. A perturbation in the x_1-direction leads to the formation of phase-space structures [58]. The parameter $\mu \in [0.5, 1.5]$ is the mass of the charged particles.

5. High-dimensional chaos A Rayleigh–Bénard convection leads to a density gradient that sets a fluid in motion. We consider a nine-dimensional dynamical system that is derived from such a flow, which exhibits cascades that lead to chaos [69]. The parameter $\mu \in [13.7, 14.4]$ is the reduced Rayleigh number.

6. Particles in aharmonic trap We consider 50 particles in an aharmonic trap [15], which lead to 100-dimensional samples $X_{t,\mu}^i$ that encode the positions of the particles. The particle positions are governed by a stochastic differential equation. The parameter $\mu \in [0.3, 0.9]$ controls the velocity of the trap.

Setup We compare our higher-order action matching (HOAM) to the original version of action matching (AM) [60], where we handle the parameter dependence on μ in the same way as in our approach. Additionally, we compare to noise-conditioned score matching (NCSM) where samples are generated via annealed Langevin dynamics [80] and conditional flow matching (CFM) [2, 53], for which we condition on time t and μ; see Appendix B.

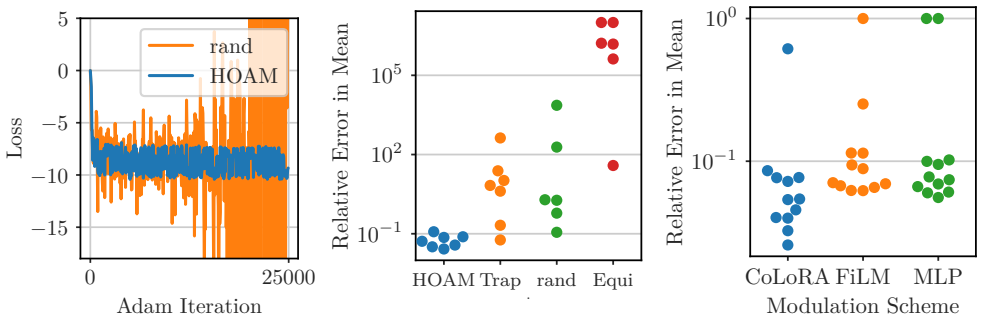

Figure 2: **L**: Gauss quadrature in HOAM gives more accurate loss estimates compared to uniform sampling as in AM. **M**: HOAM with Gauss quadrature stabilizes training. **R**: HOAM based on CoLoRA outperforms other modulation schemes for model reduction task.

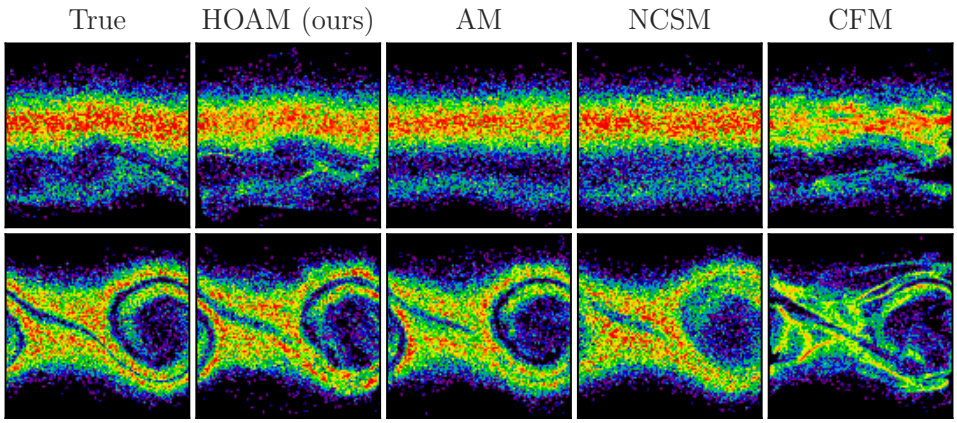

Figure 3: Histograms of samples of bump-on-tail (top, $t = 20$) and two-stream (bottom, $t = 20$) instability. HOAM accurately predicts the fine scale features and multi-modality of the population density.

HOAM stabilizes training with higher-order quadrature Figure 2(left) demonstrates on the bimodal Duffing oscillator example that using higher-order Gauss-Legendre quadrature in HOAM for estimating the time integral of the loss (7) in each mini-batch stabilizes training compared to uniform sampling over the data (1); see also Section 2.3. Whereas uniform sub-sampling leads to large oscillations in the loss over the optimization iterations, the loss estimated with Gauss-Legendre quadrature in HOAM remains stable over time. Figure 2(middle) provides further evidence of the importance of estimating the time integral well by showing the relative error of the mean (17) of the generated samples with HOAM based on Gauss-Legendre quadrature versus other quadrature rules (trapezoidal, equidistant [26]) and uniform Monte Carlo estimators. Over seven random initializations, the gradient fields learned with the loss based on Gauss-Legendre quadrature achieve up to 1–2 orders of magnitude lower relative mean errors. Additionally, Figure 2(right) shows that parameterizing the vector field $s_{t,\mu}$ with CoLoRA layers (see Section 2.3) achieves the lowest relative error of the mean, which motivates the use of the CoLoRA modulation scheme [38, 12] in our task of model reduction; see Appendix B.5 for FiLM [65] and MLPs.

Accurate predictions with speedups for Vlasov-Poisson equations Our Vlasov-Poisson problems describe the interaction of charged particles with dynamics that depend on all other particles, which leads to mean-field dynamics for large numbers of particles N_x. Thus, reduced modeling with HOAM is well suited for this problem because the natural dynamics to learn from such a system are the population dynamics $\rho_{t,\mu}$ rather than the sample dynamics; see Appendix B.2. We observe the particles computed with a particle-in-cell method and learn the gradient field $\nabla s_{t,\mu}$ with the proposed HOAM approach. For a test

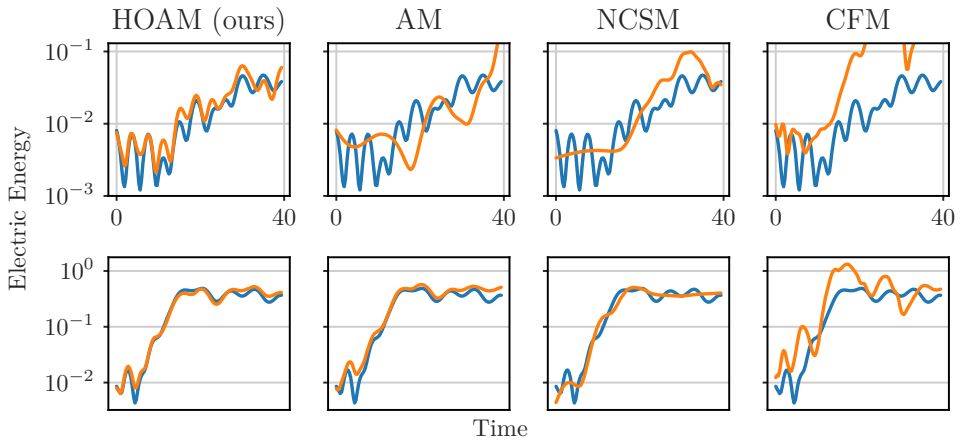

Figure 4: Electric energy of bump-on-tail (top) and two-stream (bottom) instability. HOAM accurately predicts the energy growth in the transient regime and oscillations at later times.

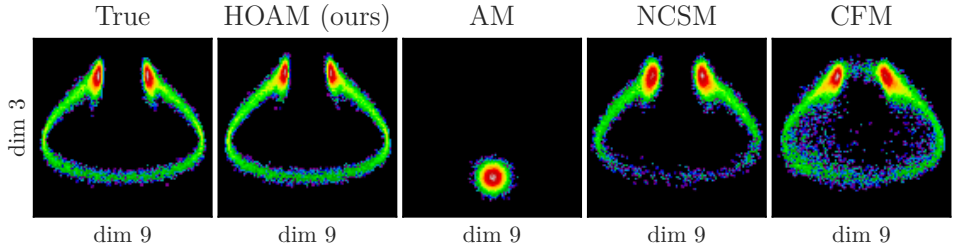

Figure 5: HOAM accurately predicts the low probability region that connects the two high probability regions. The plot shows the projection of dimension three and nine over the nine-dimensional chaotic system [69]. Time is $t = 3.7$.

physics parameter μ that controls the wave number, we then generate samples with $\nabla s_{t,\mu}$ and plot a histogram in Figure 3 for the bump-on-tail (top) and two-stream (bottom) instability. Our approach approximates well the histogram obtained with the classical particle-in-cell method and our HOAM reduced model is additionally about 20× faster (6sec vs. 2min).

The quantity of interest in both examples is the electrical (potential) energy. We compute the electric energy from the generated samples over time t for the test physics parameters, which we plot in Figure 4 and its relative error averaged over time (e.e.) in Table 1 (see (16)). Our approach approximates the electric energy well at later times, whereas NCSM and CFM lead to poorer approximations at later time t. This is relevant because this non-linear regime is where numerical solvers become important; the initial (linear) growth regime can be approximated well by analytical perturbation theory. Also for the six-dimensional strong Landau damping problem, our HOAM approach provides accurate predictions of the electric energy with orders of magnitude speedups; see Table 1 as well as Figure 9 in Appendix B.7.

Speedups in inference step (predictions) Recall two limitations of introducing a time and physics parameter dependence in NCSM/CFM via conditioning (see page 2 and Section 2.3): (i) For each t and μ, a separate sampling path has to be computed, which leads to orders of magnitude higher inference runtimes than in HOAM; see Table 1, Section 2.3. (ii) For each t and μ pair, the target distribution $\rho_{t,\mu}$ is different, which can require t- and μ-specific tuning of hyper-parameters of the inference step, which is impractical and thus can lead to a deterioration of accuracy compared to our HOAM approach; see Figure 3–4.

Predicting statistics of chaotic and particle dynamics in high dimensions We now consider the nine-dimensional dynamical system introduced in [69], which leads to chaotic behavior. We show in Figure 5 the sample histogram corresponding to a test physics parameter that represents the Rayleigh number. At time $t = 3.7$ and projecting onto

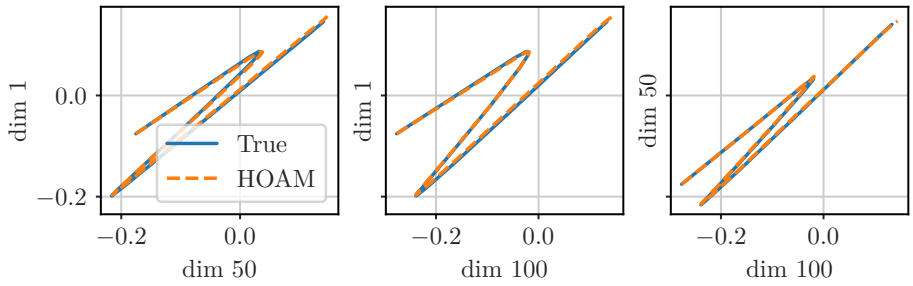

Figure 6: HOAM accurately predicts the time evolution of the mean position of a 100-dimensional particle system in an aharmonic moving trap.

example:	**two-stream**		**bump-on-tail**		**strong Landau**		**9D chaos**	
metric:	e.e.	r.t. [s]	e.e.	r.t. [s]	e.e.	r.t. [s]	sinkhorn	r.t. [s]
CFM [2, 53]	1.44	139	5.52	141	0.629	161	0.259	36
NCSM [80]	0.245	1142	0.626	1133	4.06	4531	0.869	1109
AM [60]	0.275	6	0.892	6	NaN	-	80.1	7
HOAM (ours)	**0.208**	6	**0.429**	6	**0.447**	7	**0.217**	7

Table 1: HOAM outperforms state-of-the-art methods w.r.t. inference runtime (r.t.) with comparable errors when applied to various physics problems for parametric model reduction. Metrics: e.e. is the relative error in electric energy, see (16); for sinkhorn see Appendix B.6.

dimension three and nine, the plots in Figure 5 show that the proposed HOAM accurately matches the low probability region that connects the two high probability regions. Action matching without higher-order quadrature fails to train on this example, despite extensive parameter sweeps. Consider now the example of the particles in an aharmonic trap, which leads to 100-dimensional samples $X^i_{t,\mu}$. For a test physics parameter that corresponds to the velocity of the moving trap, we plot the predicted mean particle positions over time for various dimensions in Figure 6. In all cases, the proposed HOAM predicts the mean well.

4 Conclusions, limitations, and future work

For parametric model reduction, learning population dynamics via minimal-energy vector fields over time t and physics parameter μ with our variational approach helps reduce inference runtime compared to standard diffusion- and flow-based modeling that condition on t and μ and therefore have to solve a separate inference problem for each time step and physics parameter at test time. Because we learn the dynamics over time t, it is critical to accurately capture the coupling over the time steps, for which we propose to use higher-order quadrature schemes when estimating time integrals in the training loss. The higher-order quadrature of the time integrals considerably improves training stability. Our numerical experiments indicate that our approach achieves errors that are comparable to state-of-the-art methods while at the same time reducing inference runtime by 1–2 orders of magnitude.

Limitations: First, we assume access to a rather dense set of time points for the Gauss-Legendre quadrature, which is in line with our application to parametrized partial differential equations, where the data usually comes from querying high-accuracy solvers that operate on a fine time grid. We do not claim that this is applicable when only very few samples in time are available such as in computational biology [22, 9]. Second, we currently seek a vector field that minimizes the kinetic energy. However, there are examples where the vector field with minimal kinetic energy is more "complicated" than other vector fields that lead to the same population dynamics. For example, the population dynamics of the oscillator example can be described with a constant vector field, whereas the minimal-energy field varies with time. Understanding what energies to use for which problems remains an open challenge.

We do not expect that this work has negative societal impacts.

References

[1] M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv*, 2303.08797, 2023.

[2] M. S. Albergo and E. Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023.

[3] L. Ambrosio and W. Gangbo. Hamiltonian ODEs in the Wasserstein space of probability measures. *Communications on Pure and Applied Mathematics*, 61(1):18–53, Jan. 2008.

[4] L. Ambrosio and N. Gigli. A User's Guide to Optimal Transport. In *Modelling and Optimisation of Flows on Networks*, volume 2062, pages 1–155. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. Series Title: Lecture Notes in Mathematics.

[5] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows*. Lectures in Mathematics ETH Zürich. Birkhäuser-Verlag, Basel, 2005.

[6] V. Arnold. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Annales de l'institut Fourier*, 16(1):319–361, 1966.

[7] J. W. Banks and J. A. F. Hittinger. A new class of nonlinear finite-volume methods for Vlasov simulation. *IEEE Transactions on Plasma Science*, 38(9 PART 1):2198 – 2207, 2010.

[8] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, Jan. 2000.

[9] J.-D. Benamou, T. O. Gallouët, and F.-X. Vialard. Second-Order Models for Optimal Transport and Cubic Splines on the Wasserstein Space. *Foundations of Computational Mathematics*, 19(5):1113–1143, Oct. 2019.

[10] P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review*, 57(4):483–531, 2015.

[11] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox, editors. *Model Reduction and Approximation: Theory and Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, July 2017.

[12] J. Berman and B. Peherstorfer. CoLoRA: Continuous low-rank adaptation for reduced implicit neural modeling of parameterized partial differential equations. *arXiv*, 2402.14646, 2024.

[13] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

[14] N. M. Boffi and E. Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, jul 2023.

[15] J. Bruna, B. Peherstorfer, and E. Vanden-Eijnden. Neural Galerkin schemes with active learning for high-dimensional evolution equations. *Journal of Computational Physics*, 496:112588, 2024.

[16] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz. Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control. *PLOS One*, 11(2):e0150171, 2016.

[17] C. Bunne, L. Papaxanthos, A. Krause, and M. Cuturi. Proximal optimal transport modeling of population dynamics. In G. Camps-Valls, F. J. R. Ruiz, and I. Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 6511–6528. PMLR, 28–30 Mar 2022.

[18] S. R. Cachay, B. Zhao, H. James, and R. Yu. DYffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[19] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[20] Y. Chen, M. Goldstein, M. Hua, M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and Föllmer processes. *arXiv*, 2403.13724, 2024.

[21] Y. Cheng, A. J. Christlieb, and X. Zhong. Energy-conserving discontinuous Galerkin methods for the Vlasov–Ampère system. *Journal of Computational Physics*, 256:630–655, 2014.

[22] S. Chewi, J. Clancy, T. Le Gouic, P. Rigollet, G. Stepaniants, and A. Stromme. Fast and Smooth Interpolation on Wasserstein Space. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3061–3069. PMLR, Apr. 2021.

[23] S.-N. Chow, W. Li, and H. Zhou. Wasserstein Hamiltonian flows. *Journal of Differential Equations*, 268(3):1205–1219, Jan. 2020.

[24] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[25] A. Davtyan, S. Sameni, and P. Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23263–23274, October 2023.

[26] P. Deuflhard and A. Hohmann. *Numerical Analysis in Modern Scientific Computing*. Springer, 2003.

[27] E. Dupont, A. Doucet, and Y. W. Teh. Augmented neural ODEs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[28] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer New York, New York, NY, 2004.

[29] J. A. Farrell, Y. Wang, S. J. Riesenfeld, K. Shekhar, A. Regev, and A. F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, June 2018.

[30] I. Gentil, C. Léonard, and L. Ripani. Dynamical aspects of the generalized Schrödinger problem via Otto calculus – A heuristic point of view. *Revista Matemática Iberoamericana*, 36(4):1071–1112, Jan. 2020.

[31] R. G. Ghanem and P. D. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer, 1991.

[32] W. Harvey, S. Naderiparizi, V. Masrani, C. D. Weilbach, and F. Wood. Flexible diffusion modeling of long videos. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[33] T. Hashimoto, D. Gifford, and T. Jaakkola. Learning population-level diffusions with generative RNNs. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2417–2426, New York, New York, USA, 20–22 Jun 2016. PMLR.

[34] J. Hittinger and J. Banks. Block-structured adaptive mesh refinement algorithms for Vlasov simulation. *Journal of Computational Physics*, 241:118–140, 2013.

[35] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[36] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646. Curran Associates, Inc., 2022.

[37] B. Holzschuh, S. Vegetti, and N. Thuerey. Solving inverse physics problems with score matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[38] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[39] T. J. R. Hughes. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Dover Publications, 2012.

[40] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[41] P. Kidger, J. Foster, X. Li, and T. J. Lyons. Neural SDEs as infinite-dimensional GANs. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5453–5463. PMLR, 18–24 Jul 2021.

[42] K. Kormann and A. Yurova. A generalized Fourier–Hermite method for the Vlasov–Poisson system. *BIT Numerical Mathematics*, 61(3):881–909, Sept. 2021.

[43] T. Koshizuka and I. Sato. Neural Lagrangian Schrödinger Bridge: Diffusion Modeling for Population Dynamics. In *The Eleventh International Conference on Learning Representations*, 2023.

[44] B. Kramer, B. Peherstorfer, and K. E. Willcox. Learning nonlinear reduced models from data with operator inference. *Annual Review of Fluid Mechanics*, 56(Volume 56, 2024):521–548, 2024.

[45] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor. *Dynamic mode decomposition: data-driven modeling of complex systems*. SIAM, 2016.

[46] H. Lavenant, S. Zhang, Y.-H. Kim, and G. Schiebinger. Toward a mathematical theory of trajectory inference. *The Annals of Applied Probability*, 34(1A):428 – 500, 2024.

[47] K. Lee and E. J. Parish. Parameterized neural ordinary differential equations: applications to computational physics problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2253):20210162, 2021.

[48] A. Lenard and I. B. Bernstein. Plasma Oscillations with Diffusion in Velocity Space. *Physical Review*, 112(5):1456–1459, Dec. 1958.

[49] L. Li, S. Hurault, and J. Solomon. Self-consistent velocity matching of probability flows. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[50] X. Li, T.-K. L. Wong, R. T. Q. Chen, and D. K. Duvenaud. Scalable gradients and variational inference for stochastic differential equations. In C. Zhang, F. Ruiz, T. Bui, A. B. Dieng, and D. Liang, editors, *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pages 1–28. PMLR, 08 Dec 2020.

[51] M. Lienen, D. Lüdke, J. Hansen-Palmus, and S. Günnemann. From zero to turbulence: Generative modeling for 3d flow simulation. In *The Twelfth International Conference on Learning Representations*, 2024.

[52] E. M. Lifshitz and L. P. Pitaevski. Chapter III - Collisionless Plasmas. In E. M. Lifshitz and L. P. Pitaevski, editors, *Physical Kinetics*, volume 10 of *Course of Theoretical Physics*, pages 115–167. Pergamon, Amsterdam, Jan. 1981.

[53] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[54] J. Lott. Some Geometric Calculations on Wasserstein Space. *Communications in Mathematical Physics*, 277(2):423–437, Nov. 2007.

[55] C. Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems - A*, 34(4):1533–1574, 2014.

[56] D. B. Melrose. *Instabilities in Space and Laboratory Plasmas*. Aug. 1986. Publication Title: Instabilities in Space and Laboratory Plasmas ADS Bibcode: 1986islp.book.....M.

[57] I. Mezić. Spectral properties of dynamical systems, model reduction and decompositions. *Nonlinear Dynamics*, 41(1-3):309–325, 2005.

[58] C. Mouhot and C. Villani. On Landau damping. *Acta Mathematica*, 207(1):29–201, 2011.

[59] A. Muntean, J. Rademacher, and A. Zagaris, editors. *Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity*, volume 3 of *Lecture Notes in Applied Mathematics and Mechanics*. Springer International Publishing, Cham, 2016.

[60] K. Neklyudov, R. Brekelmans, D. Severo, and A. Makhzani. Action Matching: Learning Stochastic Dynamics from Samples. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 25858–25889. PMLR, July 2023.

[61] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi. Machine learning for molecular simulation. *Annual Review of Physical Chemistry*, 71(Volume 71, 2020):361–390, 2020.

[62] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, Jan. 2001.

[63] F. Otto and C. Villani. Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400, June 2000.

[64] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

[65] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[66] L. Pichler, A. Masud, and L. A. Bergman. Numerical Solution of the Fokker–Planck Equation by Finite Difference and Finite Element Methods—A Comparative Study. In M. Papadrakakis, G. Stefanou, and V. Papadopoulos, editors, *Computational Methods in Stochastic Dynamics*, volume 26, pages 69–85. Springer Netherlands, Dordrecht, 2013. Series Title: Computational Methods in Applied Sciences.

[67] S. Possanner, F. Holderied, Y. Li, B. K. Na, D. Bell, S. Hadjout, and Y. Güçlü. High-Order Structure-Preserving Algorithms for Plasma Hybrid Models. In F. Nielsen and F. Barbaresco, editors, *Geometric Science of Information*, volume 14072, pages 263–271. Springer Nature Switzerland, Cham, 2023. Series Title: Lecture Notes in Computer Science.

[68] K. Rasul, C. Seward, I. Schuster, and R. Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8857–8868. PMLR, 18–24 Jul 2021.

[69] P. Reiterer, C. Lainscsek, F. Schürrer, C. Letellier, and J. Maquet. A nine-dimensional lorenz system to study high-dimensional chaos. *Journal of Physics A: Mathematical and General*, 31(34):7121, aug 1998.

[70] C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2004.

[71] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. Spectral analysis of nonlinear flows. *Journal of Fluid Mechanics*, 641:115–127, 2009.

[72] G. Rozza, D. Huynh, and A. Patera. Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. *Archives of Computational Methods in Engineering*, 15(3):1–47, 2007.

[73] C. Salvi, M. Lemercier, and A. Gerasimovics. Neural stochastic pdes: Resolution-invariant learning of continuous spatiotemporal dynamics. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1333–1344. Curran Associates, Inc., 2022.

[74] F. Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Springer International Publishing, Cham, 2015.

[75] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, L. Lee, J. Chen, J. Brumbaugh, P. Rigollet, K. Hochedlinger, R. Jaenisch, A. Regev, and E. S. Lander. Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176(4):928–943.e22, Feb. 2019.

[76] P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

[77] E. Schrödinger. Über die Umkehrung der Naturgesetze. Technical Report 1931 IX, Akademie der Wissenschaften, Berlin, 1931.

[78] Z. Shen, Z. Wang, S. Kale, A. Ribeiro, A. Karbasi, and H. Hassani. Self-consistency of the fokker planck equation. In P.-L. Loh and M. Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 817–841. PMLR, 02–05 Jul 2022.

[79] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[80] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[81] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, page 204, 2019.

[82] 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*, 2021.

[83] E. Sonnendrücker, A. Wacher, R. Hatzky, and R. Kleiber. A split control variate scheme for PIC simulations with collisions. *Journal of Computational Physics*, 295:402–419, Aug. 2015.

[84] A. Tong, J. Huang, G. Wolf, D. Van Dijk, and S. Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9526–9536. PMLR, 13–18 Jul 2020.

[85] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology*, 32(4):381–386, Apr. 2014.

[86] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014.

[87] T. M. Tyranowski. Stochastic variational principles for the collisional Vlasov–Maxwell and Vlasov–Poisson equations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2252):20210167, Aug. 2021.

[88] B. Tzen and M. Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv*, 1905.09883, 2019.

[89] C. Villani. *Optimal transport: old and new*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[90] C. Villani. *Topics in optimal transportation*. Number 58 in Graduate studies in mathematics. American Mathematical Society, Providence, Rhode Island, reprinted with corrections edition, 2016.

[91] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[92] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. A data–driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, 25(6):1307–1346, 2015.

[93] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, Dec. 2019.

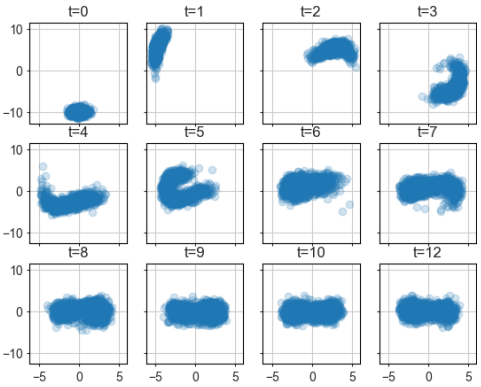

Figure 7: We plot the particles of the bimodal Duffing oscillator at different times.

A Additional details about loss

Following the notation of Section 2.3, the empirical loss is

$$E^{\epsilon}_{M_x^b,M_t^b,M_\mu^b}(\theta,\psi) = \sum_{j=1}^{M_t^b} w_j \sum_{k=1}^{M_x^b} \frac{1}{2}|\nabla s(X^{i_k}_{t_j,\mu};h(t_j,\mu;\phi),\theta)|^2 + \partial_t s(X^{i_k}_{t_j,\mu};h(t_j,\mu;\phi),\theta) + \frac{\epsilon^2}{2}\Delta s(X^{i_k}_{t_j,\mu};h(t_j,\mu;\phi),\theta) + s(X^{i_k}_{0,\mu},h(0,\mu;\psi),\theta) - s(X^{i_k}_{1,\mu},h(1,\mu;\psi),\theta)\,, \quad (8)$$

with μ ($M_\mu^b = 1$) uniformly sampled from $\{\mu_1,\dots,\mu_{M_\mu}\} \sim \nu$, indices $i_1,\dots,i_k$ uniformly sampled from $\{1,\dots,N_x\}$, and $w_1,\dots,w_{M_t^b}$ Gauss-Legendre weights and $t_1,\dots,t_{M_t^b}$ Gauss-Legendre nodes [26].

B Details about numerical examples

B.1 Bimodal Duffing oscillator

The equation of motion is given by Equation (5.21) in [66] for $X = [X_1, X_2]$:

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}X_1\\X_2\end{bmatrix} = \begin{bmatrix}X_2\\-2\xi\omega X_2 + \omega^2 X_1 - \omega^2\mu X_1^3\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} w(t)\,. \quad (9)$$

We let $\xi = 0.2$, $w \equiv 1$, $\mu_{\text{train}} \in \{0.10, 0.15, 0.25, 0.30\}$, and $\mu_{\text{test}} = 0.20$. The initial configuration in phase-space is given by a Gaussian centered at $(0,-10)$ with width 0.5. The stationary solution of this system is given by

$$\rho_{\infty,\mu}(x_1,x_2) \propto \exp\left(-\frac{1}{2\sigma_1^2}\left(-x_1^2 + \frac{\mu}{2}x_1^4\right) - \frac{\sigma_2^2}{2}x_2^2\right) \quad (10)$$

with constants $\sigma_{1,2}$. The maximum values of the stationary distribution are located at $(\pm\mu^{-1/2}, 0)$. Importantly, while individual sample (particle) trajectories continue to evolve at the stationary state, the population dynamics are constant in the sense that the gradient field s remains constant. We integrate 25000 particles of the system up to $T = 12$ using the Euler-Maruyama scheme with time step size equal to $1e-2$.

B.2 Vlasov-Poisson problems

Mean field approximations The Vlasov-Poisson system describes the interaction of charged particles. Due to the presence of the Coulomb force, the dynamics of a single particle depend on the position of all other particles. Assuming N particles in the system, this means $\frac{\mathrm{d}}{\mathrm{d}t}X^i_{t,\mu} = v(t, X^i_{t,\mu}; \mu, X^1_{t,\mu}, \dots, X^N_{t,\mu})$. Given the fact that N is in practice extremely large, it is natural to pass to the *mean-field limit*. Assuming the particles are indistinguishable, the result is a PDE of the form $\partial_t\rho_{t,\mu} + \nabla\cdot(\rho_{t,\mu}v_{\text{mf}}(t,\cdot;\mu,\rho_{t,\mu})) = 0$ that describes the evolution

of the collection (or population, ensemble) of particles denoted by $\rho_{t,\mu}$. In the specific case of the Vlasov-Poisson problem, Coulomb interactions in the mean-field limit give rise to a Poisson equation determining an electric field that is generated by the collection of particles and influences its dynamics. For completeness sake, we mention that the singularity of the Coulomb interaction poses a considerable technical challenge when passing to this limit. We refer to [52, 58] for the derivation of the Vlasov-Poisson equation and [59] for more examples of mean-field systems. The theory behind the test-cases we run in this work can be found in [56], Chapter 3.

Governing equation We slightly change the notation here to be consistent with the references. $f : \mathcal{X}_x \times \mathbb{R}^d \times \mathbb{R} \times \mathcal{D} \to \mathbb{R}$, $d \in \{1, 2, 3\}$, denotes the distribution function governed by the Vlasov-Poisson system

$$\partial_t f(x, v, t; \mu) = -v \cdot \nabla_x f(x, v, t; \mu) - \nabla\phi(x, t) \cdot \nabla_v f(x, v, t; \mu) = 0\,, \tag{11}$$

$$-\Delta\phi(x, t; \mu) = 1 - \int_{\mathbb{R}^d} f(x, v, t; \mu)\mathrm{d}v\,. \tag{12}$$

In the notation of the rest of this work, $f(\cdot, \cdot, t; \mu) = \rho_{t,\mu}$, $\mathcal{X}_x \times \mathbb{R}^d = \mathcal{X}$. The spatial domain $\mathcal{X}_x$ is a subset of $\mathbb{R}^d$, in all our examples it is of the form $[0, l_1] \times [0, l_2] \times [0, l_3]$ with periodic boundary conditions in the spatial coordinate.

Two-stream instability In this case, $d = 2$, so the the particle positions $x = x_1$ vary in $\mathcal{X}_x = [0, l_1]$ with periodic boundary conditions and the velocity v evolves in $\mathbb{R}$. For the two-stream instability, we set the initial distribution to

$$f_0(x, v) := \frac{1}{2\sqrt{2\pi}} \left(1 + \alpha \cos\left(2\pi \frac{x}{l_1}\right)\right) \left(\exp\left(-\frac{(v - v_0)^2}{2}\right) + \exp\left(-\frac{(v + v_0)^2}{2}\right)\right), \tag{13}$$

with $\alpha = 0.05, l_1 = 50, v_0 = 3$. The parameter μ varies as $\mu_{\text{train}} \in \{1.2, 1.3, \ldots, 1.9\}$ and $\mu_{\text{test}} \in \{1.25, 1.85\}$. We use a particle-in-cell method for generating the data based on the repository `https://github.com/pmocz/pic-python`. The number of marker particles is $N = 25000$ and for the sake of computing the electric field, a uniform grid of $N/8$ cells is used. Integration in time is done via a Störmer-Verlet splitting over $t \in [0, 40]$ with time-step size $1e - 2$.

Bump-on tail We consider the initial distribution

$$f_0(x, v) = \frac{1}{\sqrt{2\pi}} \left(1 + \alpha \cos\left(2\pi \frac{x}{l_1}\right)\right) \left(\frac{\delta}{\sigma_1} \exp\left(-\frac{v^2}{2\sigma_1^2}\right) + \frac{1 - \delta}{\sigma_2} \exp\left(-\frac{(v - v_b)^2}{2\sigma_2^2}\right)\right), \tag{14}$$

with $\alpha = 0.05, l_1 = 50, v_b = 4, \delta = 9/10, \sigma_1 = 1, \sigma_2 = 1/\sqrt{2}$. The parameter μ varies as $\mu_{\text{train}} \in \{1.3, 1.4, \ldots, 2.0\}$ and $\mu_{\text{test}} \in \{1.35, 1.95\}$. The other parameters are the same as in the two-stream case.

Strong Landau damping In this case, $d = 6$ and

$$f_0(x, v) = \frac{1}{\sqrt{2\pi}^3} \left(1 + \alpha \cos\left(2\pi \frac{x_1}{l_1}\right)\right) \exp\left(-\frac{|v|^2}{2}\right), \tag{15}$$

with $l_1 = 4\pi$ and $l_2 = l_3 = 1$. The data is generated using the Struphy package [67], the exact specifications of the simulation are available at `https://gitlab.mpcdf.mpg.de/struphy` as an example problem. The physics parameter we vary is the mass of the charged particles, which has the effect of changing the strength of the inertial term accelerating the particles relative to the advection term $v \cdot \nabla_x f$. This implies $\mu \in \{0.5, 0.6, \ldots, 1.5\}$, where $\mu = 1.0$ corresponds to the default settings. This $\mu = 1.0$ is also the test parameter and is excluded from the training set.

The high-fidelity data we generate is using a control variate approach in order to reduce numerical noise introduced by the finite number of marker particles. Since we require the particles to be identical for our method, we assume they are all weighted equally when

re-constructing the electric potential. This biases our reconstructed potential in comparison to the physical one, but we observe in practice that this is only by a multiplicative constant. We save 10^5 marker particles from the high-order simulations and use $N = 25000$ of them as input data for our method. We integration in time over $t \in [0, 8.75]$

B.3 High-dimensional chaos

We consider the dynamical system introduced in [69]. We generate samples by initializing a 9 dimensional Gaussian centered at the origin with width equal to $2e-2$. We then integrate these samples forward as an SDE whose drift is given by the 9-dimensional system of ODEs described in [69] and the diffusion term is given as diagonal noise equal to $5e-2$. We integrate 25000 particles of the system up to $T = 20$ using the Euler-Maruyama scheme with time step size equal to $1e-2$.

B.4 Particles in aharmonic trap

We consider the evolution of interacting particles in an aharmonic trap [15]. The two-dimensional particle positions $Z_1(t,\mu), \ldots, Z_M(t,\mu)$ are governed by an SDE

$$\mathrm{d}Z_i = g(t, Z_i)\mathrm{d}t + \sum_{j=1}^{M} K(Z_i, Z_j)\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}W_i\,, \qquad i = 1, \ldots, M\,,$$

where $\gamma > 0$ is the diffusion coefficient and W_i are independent Wiener processes. The function $g(t, Z) = (a(t) - Z)^3$ describes a time-dependent one-body force, where $a(t) = 5/4(\sin(\pi t)+3/2))+\mu\cos(t\pi 2)$ is the position of the trap. The function $K(Z, Z') = \frac{\alpha}{M}(Z'-Z)$ describes a pairwise interaction term. We set $\alpha = -1/4$ and $\gamma = 10^{-2}$. The parameter μ is in the range $\mathcal{D} = [0.3, 0.9]$ and modifies the position of the trap. A sample $X^i_{t,\mu}$ corresponds to a vector $[Z_1(t,\mu), \ldots, Z_M(t,\mu)]^T$ of dimension 100, because we have $M = 50$ particles and each position $Z_j(t,\mu)$ as two dimensions. We generate samples via Monte Carlo by using the Euler-Maruyama scheme. The time step size is $\delta t = 1e-3$ and we integrate up to final time 2.

B.5 Modulation schemes

The other two modulation schemes that are compared in Figure 2 are FiLM [65] and MLP. For the MLP the inputs x, t, μ are concatenated together and input directly to the model. There is no hyper-network or modulation scheme. For FiLM, we closely follow the original paper. The main network takes in x as input and the hyper-network t, μ as input. The hyper-network and main network have the same parameter counts as in the CoLoRA experiments. The output of the hyper-network then directly modulates the activation of each layer of the main network as detailed in the original FiLM paper [65].

B.6 Other details about numerical experiments

In terms of network architecture, we follow [12] closely because we use their network architecture. We use MLPs to parameterize both the main network and the hyper-network with swish activation functions. The main network is depth 7 and width 64 linear layers while the hyper-network is depth 3 with width 15 linear layers. Identical CoLoRA architectures are used for all HOAM experiments as well as the comparisons with AM, NCSM, and CFM. The only difference is the size of the output layer for NCSM and CFM whose outputs must be the same dimensionality as their inputs.

For all experiments we use an Adam optimizer at a $2e-3$ learning rate with a cosine learning rate scheduler. For all experiments unless otherwise noted, we take a batch size of 256 particles over 256 time points.

The results were computed on NVIDIA Quadro RTX 8000 GPUs. All code was implemented in Python using the JAX library with JIT complication where possible.

Hyper-parameter ϵ in the loss (7) searched over $[0.0, 1e-2, 5e-2]$ for both HOAM and AM.

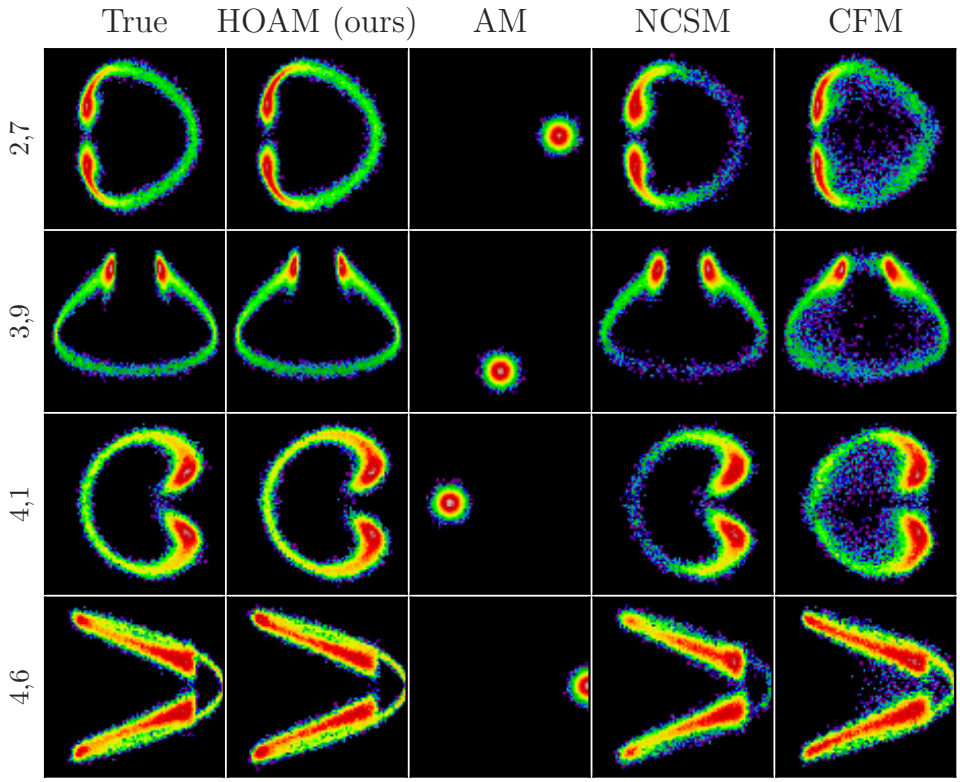

Figure 8: Shows the projections of other dimensions of the nine-dimensional chaotic system [69]; see also Figure 5.

The relative error in the electric energy is computed as

$$\frac{1}{T}\sum_{t=1}^{T}\frac{|e_{\text{true}}(t)-e_{\text{predict}}(t)|}{|e_{\text{true}}(t)|}, \tag{16}$$

where $e_{\text{true}}(t)$ is the electric energy predicted by the high-fidelity numerical simulations at time t and $e_{\text{predict}}(t)$ is the electric energy computed from samples of either HOAM (ours), AM, NCSM, or CFM. The relative error in the mean is

$$\frac{1}{T}\sum_{t=1}^{T}\frac{|\mathbb{E}[\rho_{\text{true}}(t)]-\mathbb{E}[\rho_{\text{predict}}(t)]|}{|\mathbb{E}[\rho_{\text{true}}(t)]|}, \tag{17}$$

where the expected values are estimated via Monte Carlo from the generated samples.

The Sinkhorn distance is computed with `https://ott-jax.readthedocs.io/en/latest/` with threshold 10^{-3}; see also [24].

B.7 Additional numerical results

In Figure 8 we show the various projections at time $t = 3.7$ of the sample distribution corresponding to the nine-dimensional chaotic system [69]; see also Figure 5 which shows the projection onto dimension three and nine.

In Figure 9 we show the particle histograms and the electric energy curves for the six-dimensional Vlasov-Poisson problem corresponding to strong Landau damping.

C Calculations regarding the entropic loss

In the following, assume that $\rho \in \mathcal{P}(\mathcal{X})$ is a smooth density bounded away from zero. We begin by showing some calculation rules of the operator $-\Delta_\rho : s \mapsto -\nabla\cdot(\rho\nabla s)$ with

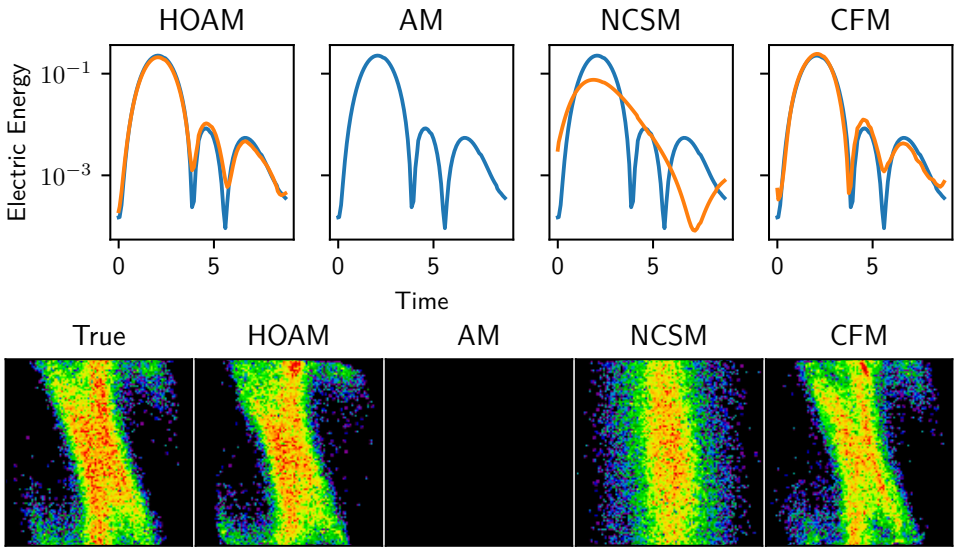

Figure 9: Electric energy and solution field at time $t = 4$ for the 6 dimensional strong Landau example.

homgeneous Neumann boundary conditions. In its weak form, it reads

$$-\int_{\mathcal{X}} f \Delta_\rho s \mathrm{d}x = \int_{\mathcal{X}} \nabla f \cdot \nabla s \, \rho \mathrm{d}x \quad \forall f \in \mathcal{C}^\infty(\mathcal{X}). \tag{18}$$

With the choice $f = \log \rho$, we find the useful identity $\Delta_\rho \log \rho = \Delta \rho$. Next, recall the objective E^ϵ from (5):

$$E^\varepsilon(s) = \frac{1}{2} \int_{\mathcal{X}} \left| \nabla \left(s - \frac{\varepsilon^2}{2} \log \rho \right) \right|^2 \rho \mathrm{d}x - \int_{\mathcal{X}} \partial_t \rho s \mathrm{d}x.$$

Now denote by δs an arbitrary element of $\mathcal{C}^\infty(\mathcal{X})$. Then, if s^ϵ is a minimizer of the (strictly convex) objective, it holds that

$$0 \stackrel{!}{=} \frac{\mathrm{d}}{\mathrm{d}\tau} E^\varepsilon(s^\epsilon + \tau \delta s)\Big|_{\tau=0} = -\int_{\mathcal{X}} \delta s \Delta_\rho \left(s^\epsilon - \frac{\varepsilon^2}{2} \log \rho \right) \mathrm{d}x - \int_{\mathcal{X}} \partial_t \rho \delta s \mathrm{d}x \quad \forall \delta s. \tag{19}$$

Hence,

$$0 = \Delta_\rho \left(s^\epsilon - \frac{\varepsilon^2}{2} \log \rho \right)^2 + \partial_t \rho = \nabla \cdot (\rho \nabla s^\epsilon) - \frac{\varepsilon^2}{2} \Delta \rho + \partial_t \rho. \tag{20}$$

Furthermore, note that (5) is identical to

$$E^\epsilon_{t,\mu}(s) = \frac{1}{2} \int_{\mathcal{X}} \left(|\nabla s|^2 + \epsilon^2 \Delta s \right) \rho_{t,\mu} \mathrm{d}x - \int_{\mathcal{X}} \partial_t \rho_{t,\mu} s \mathrm{d}x + \frac{\epsilon^2}{8} \int_{\mathcal{X}} |\nabla \log \rho_{t,\mu}|^2 \rho_{t,\mu} \mathrm{d}x \tag{21}$$

after integration by parts. The second term can be interpreted as a regularizing factor that leads to more regular solutions $s_{t,\mu}$. We have found this to be beneficial in practice. The last term is the Fisher information of the data at t, μ and plays no role in the optimization.

D Motivating the partial integration in time in the loss

Note that while t plays the role of a parameter in Equation (3), the problems corresponding to different values of t are coupled through the term $\partial_t \rho_{t,\mu}$. This is most apparent when one discretizes the equation in time. Denote by $\{t_i\}_{i=0}^M$ a strictly increasing sequence with

$t_0 = 0, t_M = 1$, and $t_{i+1} - t_i = \delta t_i$. Then, for fixed but arbitrary μ, we obtain M coupled problems of the form

$$\min_{s_{t_i} \in H^1(\rho_{t_i,\mu}, \mathcal{X})} \frac{1}{2} \int_{\mathcal{X}} |\nabla s_{t_i,\mu}|^2 \rho_{t_i,\mu} \mathrm{d}x - \frac{1}{\delta t} \int_{\mathcal{X}} (\rho_{t_{i+1},\mu} - \rho_{t_i,\mu}) s_{t_i,\mu} \mathrm{d}x \quad \forall i, \mu. \tag{22}$$

Adding these problems and shifting the indices, one can eliminate $\rho_{t_{i+1},\mu}$, explicitly coupling $s_{t_i,\mu}$ and $s_{t_{i+1},\mu}$. The continuous equivalent of this of course is just an integration over t, followed by an integration by parts.

E Geometric picture of the optimization problem

We omit the dependence on the parameter μ here for the sake of simpler notation and write $\mathrm{d}\rho$ for $\rho \, \mathrm{d}x$ for brevity. Note that the following considerations are purely formal. They are meant to illustrate a geometric picture of the optimization problems we consider. We claim no originality of these ideas; the exposition is based on Chapter 7 of [89] as well as [23, 54].

Otto calculus Based on the identification of the tangent space of $P(\mathcal{X})$ with the space of gradients (more rigorously, at point $\rho_t \in P(\mathcal{X})$, the closure of $\{\nabla f : f \in \mathcal{C}^\infty(\mathcal{X})\}$ in $L^2(\mathcal{X}, \rho_t, \mathbb{R}^d)$, see Definition 8.4.1 in [5]), one can view $\mathcal{P}(\mathcal{X})$ formally as a Riemannian manifold:

Definition 1 ([62])**.** *Let $\tau \mapsto \rho_\tau^1$ and $\tau \mapsto \rho_\tau^2$ be two curves valued in $\mathcal{P}(\mathcal{X})$ for $\tau \in (t-\epsilon, t+\epsilon)$ such that $\rho_\tau^1\big|_{\tau=t} = \rho_\tau^2\big|_{\tau=t} = \rho_t$. The optimal transport metric on $T\mathcal{P}(\mathcal{X})$ at $\rho_t \in P(\mathcal{X})$ is given by*

$$g(\rho_t)(\partial_\tau \rho_\tau^1\big|_{\tau=t}, \partial_\tau \rho_\tau^2\big|_{\tau=t}) = \int_{\mathcal{X}} (\nabla s_t^1 \cdot \nabla s_t^2) d\rho_t : \\ \partial_\tau \rho_\tau^1 + \nabla \cdot (\rho_t \nabla s_t^1) = 0, \partial_\tau \rho_\tau^2 + \nabla \cdot (\rho_t \nabla s_t^2) = 0. \tag{23}$$

This formalism is commonly named after the author of [62] and is closely linked to Arnold's considerations on geometric hydrodynamics [6][1] As both the identification of s_t from $\partial_t \rho_t$ and the metric depend on ρ_t, the geometry defined on $\mathcal{P}(\mathcal{X})$ in this way is non-trivial.

Action of a curve The optimization Equation (3) has an appealing physical interpretation: The vector field we define as tangent to the curve is, among all compatible ones, the one with the smallest integrated kinetic energy. In analogy with the physical literature, we call $\frac{1}{2} \int_0^1 \int_{\mathcal{X}} |\nabla s_t|^2 \mathrm{d}\rho_t$ the *action* of the curve $t \mapsto \rho_t$ with tangent velocity ∇s_t. We want to stress that while this procedure is reminiscent of physical action principles, in the latter a solution corresponds to a *stationary point* given boundary conditions at the beginning and end of the curve. The problem we consider in Equation (6) is more narrow and concerned with finding ∇s_t that matches a *given* curve $t \mapsto \rho_t$. Determining curves of minimal action in $\mathcal{P}(\mathcal{X})$, leads to the Benamou-Brenier formula ([4], Proposition 2.30)):

$$\frac{1}{2} W_2^2(\rho_0, \rho_1) = \inf_{\rho, s} \left(\frac{1}{2} \int_0^1 \int_{\mathcal{X}} |\nabla s_t|^2 d\rho_t \, dt : \partial_t \rho_t + \nabla \cdot (\rho_t \nabla s_t) = 0, \rho_{t=0} = \rho_0, \rho_{t=1} = \rho_1 \right), \tag{24}$$

with W_2 the Wasserstein (or Kantorochiv-Rubinstein) distance.

Lagrangian functions The selection criterion based on kinetic energy alone is not without alternatives. In [23], the relation $\partial_t \rho = -\Delta_\rho s$ is interpreted as a form of Legendre transform, hence s plays the role of a momentum and $L(\rho_t, \partial_t \rho_t, t) = \int_{\mathcal{X}} |\nabla \Delta_\rho^\dagger \partial_t \rho|^2 \mathrm{d}\rho$ that of a Lagrangian. Here, we introduced the notation $\Delta_\rho^\dagger$ to denote the pseudo inverse operator. Note that, formally, it is sensible to consider $\partial_t \rho$ as an element of the tangent space of $\mathcal{P}(\mathcal{X})$. After all, $\rho + \tau \partial_t \rho \in \mathcal{P}(\mathcal{X})$ for ρ strictly positive and τ small enough. In this picture, s is an

[1] The derivation of fluid dynamics from variational principles is, of course, much older and goes back as far as Langrange's Mécanique analytique published in 1789.

element of the cotangent space. The introduction of [62] addresses the two concepts and how they relate.

Any function $L : (\rho, \partial_t\rho, t) \mapsto L(\rho, \partial_t\rho, t)$, strictly convex and superlinear in its second argument, can be chosen to define the minimization objective.[2] Details can be found in Chapter 7 of [89], which also features a comprehensive discussion of the history and applications of this problem. In recent years, this formulation has been applied for modeling purposes, e.g. in [43]. To give an example, the choice $L(\rho, \partial_t\rho, t) = \frac{1}{2}\int_{\mathcal{X}} |\nabla\Delta_\rho^\dagger\partial_t\rho|^2 \mathrm{d}\rho - \int_{\mathcal{X}} V d\rho$ for a potential $V : \mathcal{X} \to \mathbb{R}$ can be used to model obstacles in the path of the samples. There exist a number of partial differential equations whose solutions ρ_t can be described as curves of stationary action with respect to such Lagrangians, described in [3, 23], as well as [89], Chapter 23, and [90], Chapter 8.

Schrödinger Bridge The objective defined in Equation (5) corresponds to the choice

$$L^\epsilon(\rho, \partial_t\rho, t) := \frac{1}{2}\int_{\mathcal{X}} \left|\nabla\left(-\Delta_\rho^\dagger\partial_t\rho + \frac{\epsilon^2}{2}\log\rho\right)\right|^2 \mathrm{d}\rho. \tag{25}$$

The associated momentum s^ϵ therefore satisfies $s^\epsilon = \frac{\delta L^\epsilon}{\delta(\partial_t\rho)}$, hence $\Delta_\rho s^\epsilon + \frac{\epsilon^2}{2}\Delta\rho = \partial_t\rho$, a Fokker-Planck equation. Furthermore, the action of the curve $t \mapsto \rho_t$ is given by

$$\begin{aligned}\int_0^1 L^\epsilon(\rho_t, \partial_t\rho, t)\mathrm{d}t = \int_0^1 \left(\frac{1}{2}\int_{\mathcal{X}} |\nabla\Delta_\rho^\dagger\partial_t\rho|^2 \mathrm{d}\rho + \frac{\epsilon^4}{8}\int_{\mathcal{X}} |\nabla\log\rho_t|^2 \, d\rho_t\right) dt \\ + \frac{\epsilon^2}{2}\left(\int_{\mathcal{X}} \log\rho_t \mathrm{d}\rho_t\Big|_{t=1} - \int_{\mathcal{X}} \log\rho_t \mathrm{d}\rho_t\Big|_{t=0}\right).\end{aligned} \tag{26}$$

This expression is known as the dual formulation of the Kantorovich-Schrödinger problem ([30], Theorem 36, except for the fact that the ϵ therein corresponds to $\epsilon^2/2$ here). While the classical optimal transport problem is concerned with the path connecting ρ_0 and ρ_1 minimizing the time integral of the kinetic energy (which coincides with the transport cost), the Schrödinger-Bridge problem is concerned with finding the most likely configuration at intermediate times, subject to the information that the configuration is given at times 0 and 1 and assuming that the particles X_t undergo Brownian motion with diffusivity $\varepsilon^2/2$. Unless ρ_1 is the result of a convolution of ρ_0 with a Gaussian kernel of width ε, the evolution of the system towards ρ_1 is a rare event and the most likely solution is to be understood conditional on the observation of this event.

Rigorous results can be found in Section 5 of [30]. Another derivation of the loss function from Equation (5), starting from the static formulation and linking to the dynamical picture presented here, can also be found in [46], Theorem 2.1. In their notation, $\Psi = -s$.

[2] The variables ρ and ∂_t here denote any probability density and a scalar field on $\mathcal{X}$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 4.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Section 2.1, Appendix C–E.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 3, Appendix A–B, code implementation link in Section 1 (retracted for review).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be

possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Section 1 provides link to code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3, Appendix A–B, code publication discussed in Section 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Figure 2 shows replicates, results reported in Figure 3–6, Table 1 are based on thousands of samples.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table 1, Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Only data from numerical simulations are used. We do not expect that this work has negative societal impacts.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

Justification: Section 4.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Appendix B.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Section 1 and Appendix B.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.