Adjoint Schrödinger Bridge Sampler

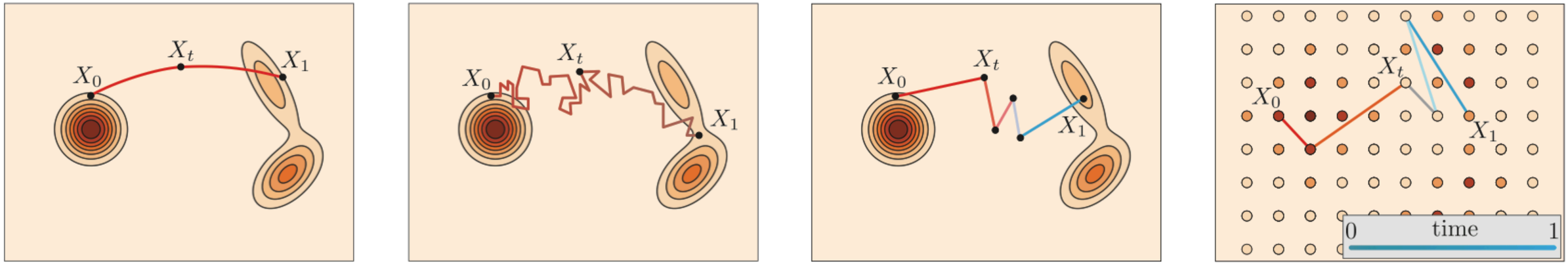
NeurIPS 2025 (Oral)

Guan-Horng Liu*, Jaemoo Choi*, Yongxin Chen, Benjamin Kurt Miller, and Ricky T. Q. Chen

Seungwoo Yoo, KAIST Visual AI Group

Problem Definition

Modern generative models, such as flow and diffusion models, learn to map from one distribution to another via iterative process.



(a) Flow **(b)** Diffusion **(c)** Jump **(d)** CTMC

Lipman *et al.*, Flow Matching Guide and Code, NeurIPS 2024 Tutorial

Problem Definition

Generative Modeling by Estimating Gradients of the Data Distribution

Yang Song
Stanford University
yangsong@cs.stanford.edu

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Abstract

We introduce a new generative model where samples are produced via Langevin dynamics using gradients of the data distribution estimated with score matching. Because gradients can be ill-defined and hard to estimate when the data resides on low-dimensional manifolds, we perturb the data with different levels of Gaussian noise, and jointly estimate the corresponding scores, *i.e.*, the vector fields of gradients of the perturbed data distribution for all noise levels. For sampling, we propose an annealed Langevin dynamics where we use gradients corresponding to gradually decreasing noise levels as the sampling process gets closer to the data manifold. Our framework allows flexible model architectures, requires no sampling during training or the use of adversarial methods, and provides a learning objective that can be used for principled model comparisons. Our models produce samples comparable to GANs on MNIST, CelebA and CIFAR-10 datasets, achieving a new state-of-the-art inception score of 8.87 on CIFAR-10. Additionally, we demonstrate that our models learn effective representations via image inpainting experiments.

1 Introduction

Generative models have many applications in machine learning. To list a few, they have been used to generate high-fidelity images [26, 6], synthesize realistic speech and music fragments [58], improve the performance of semi-supervised learning [28, 10], detect adversarial examples and other anomalous data [54], imitation learning [22], and explore promising states in reinforcement learning [41]. Recent progress is mainly driven by two approaches: likelihood-based methods [17, 29, 11, 60] and generative adversarial networks (GAN [15]). The former uses log-likelihood (or a suitable surrogate) as the training objective, while the latter uses adversarial training to minimize f-divergences [40] or integral probability metrics [2, 55] between model and data distributions.

Although likelihood-based models and GANs have achieved great success, they have some intrinsic limitations. For example, likelihood-based models either have to use specialized architectures to build a normalized probability model (*e.g.*, autoregressive models, flow models), or use surrogate losses (*e.g.*, the evidence lower bound used in variational auto-encoders [29], contrastive divergence in energy-based models [21]) for training. GANs avoid some of the limitations of likelihood-based models, but their training can be unstable due to the adversarial training procedure. In addition, the GAN objective is not suitable for evaluating and comparing different GAN models. While other objectives exist for generative modeling, such as noise contrastive estimation [19] and minimum probability flow [50], these methods typically only work well for low-dimensional data.

In this paper, we explore a new principle for generative modeling based on estimating and sampling from the *(Stein) score* [33] of the logarithmic data density, which is the gradient of the log-density function at the input data point. This is a vector field pointing in the direction where the log data density grows the most. We use a neural network trained with score matching [24] to learn this vector field from data. We then produce samples using Langevin dynamics, which approximately

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Score Matching, NeurIPS 2019 (Oral)

Denoising Diffusion Probabilistic Models

Jonathan Ho
UC Berkeley
jonathanho@berkeley.edu

Ajay Jain
UC Berkeley
ajayj@berkeley.edu

Pieter Abbeel
UC Berkeley
pabbeel@cs.berkeley.edu

Abstract

We present high quality image synthesis results using diffusion probabilistic models, a class of latent variable models inspired by considerations from nonequilibrium thermodynamics. Our best results are obtained by training on a weighted variational bound designed according to a novel connection between diffusion probabilistic models and denoising score matching with Langevin dynamics, and our models naturally admit a progressive lossy decompression scheme that can be interpreted as a generalization of autoregressive decoding. On the unconditional CIFAR10 dataset, we obtain an Inception score of 9.46 and a state-of-the-art FID score of 3.17. On 256x256 LSUN, we obtain sample quality similar to ProgressiveGAN. Our implementation is available at `https://github.com/hojonathanho/diffusion`.

1 Introduction

Deep generative models of all kinds have recently exhibited high quality samples in a wide variety of data modalities. Generative adversarial networks (GANs), autoregressive models, flows, and variational autoencoders (VAEs) have synthesized striking image and audio samples [14, 27, 3, 58, 38, 25, 10, 32, 44, 57, 26, 33, 45], and there have been remarkable advances in energy-based modeling and score matching that have produced images comparable to those of GANs [11, 55].

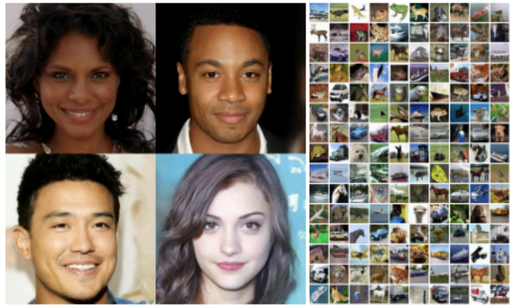


Figure 1: Generated samples on CelebA-HQ 256 × 256 (left) and unconditional CIFAR10 (right)

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Noise Matching, NeurIPS 2020

Flow Matching for Generative Modeling

Yaron Lipman[1,2] **Ricky T. Q. Chen**[1] **Heli Ben-Hamu**[2] **Maximilian Nickel**[1] **Matt Le**[1]
[1]Meta AI (FAIR) [2]Weizmann Institute of Science

Abstract

We introduce a new paradigm for generative modeling built on Continuous Normalizing Flows (CNFs), allowing us to train CNFs at unprecedented scale. Specifically, we present the notion of Flow Matching (FM), a simulation-free approach for training CNFs based on regressing vector fields of fixed conditional probability paths. Flow Matching is compatible with a general family of Gaussian probability paths for transforming between noise and data samples—which subsumes existing diffusion paths as specific instances. Interestingly, we find that employing FM with diffusion paths results in a more robust and stable alternative for training diffusion models. Furthermore, Flow Matching opens the door to training CNFs with other, non-diffusion probability paths. An instance of particular interest is using Optimal Transport (OT) displacement interpolation to define the conditional probability paths. These paths are more efficient than diffusion paths, provide faster training and sampling, and result in better generalization. Training CNFs using Flow Matching on ImageNet leads to consistently better performance than alternative diffusion-based methods in terms of both likelihood and sample quality, and allows fast and reliable sample generation using off-the-shelf numerical ODE solvers.

1 Introduction

Deep generative models are a class of deep learning algorithms aimed at estimating and sampling from an unknown data distribution. The recent influx of amazing advances in generative modeling, *e.g.*, for image generation Ramesh et al. (2022); Rombach et al. (2022), is mostly facilitated by the scalable and relatively stable training of diffusion-based models Ho et al. (2020); Song et al. (2020b). However, the restriction to simple diffusion processes leads to a rather confined space of sampling probability paths, resulting in very long training times and the need to adopt specialized methods (*e.g.*, Song et al. (2020a); Zhang & Chen (2022)) for efficient sampling.

In this work we consider the general and deterministic framework of Continuous Normalizing Flows (CNFs; Chen et al. (2018)). CNFs are capable of modeling arbitrary probability path and are in particular known to encompass the probability paths modeled by diffusion processes (Song et al., 2021). However, aside from diffusion that can be trained efficiently via, *e.g.*, denoising score matching (Vincent, 2011), no scalable CNF training algorithms are known. Indeed, maximum likelihood training (*e.g.*, Grathwohl et al. (2018)) require expensive numerical ODE simulations, while existing simulation-free methods either involve intractable integrals (Rozen et al., 2021) or biased gradients (Ben-Hamu et al., 2022).

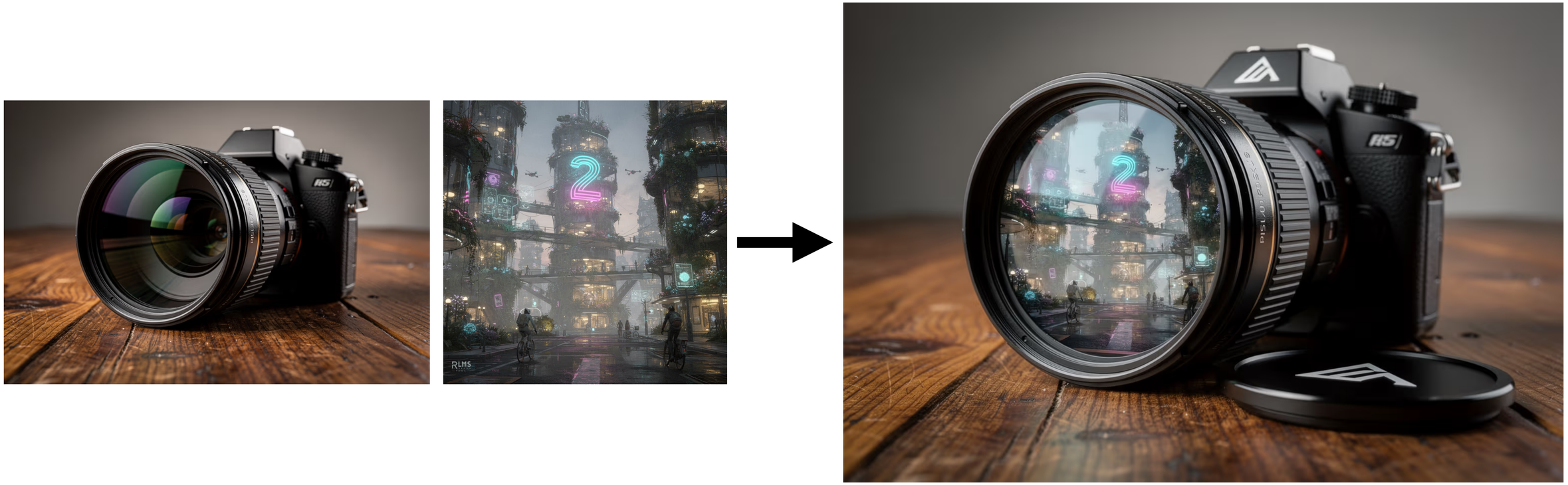
Figure 1: Unconditional ImageNet-128 samples of a CNF trained using Flow Matching with Optimal Transport probability paths.

The goal of this work is to propose Flow Matching (FM), an efficient simulation-free approach to training CNF models, allowing the adoption of general probability paths to supervise CNF training. Importantly, FM breaks the barriers for scalable CNF training beyond diffusion, and sidesteps the need to reason about diffusion processes to directly work with probability paths.

Flow Matching, ICLR 2023 (Spotlight)

Problem Definition

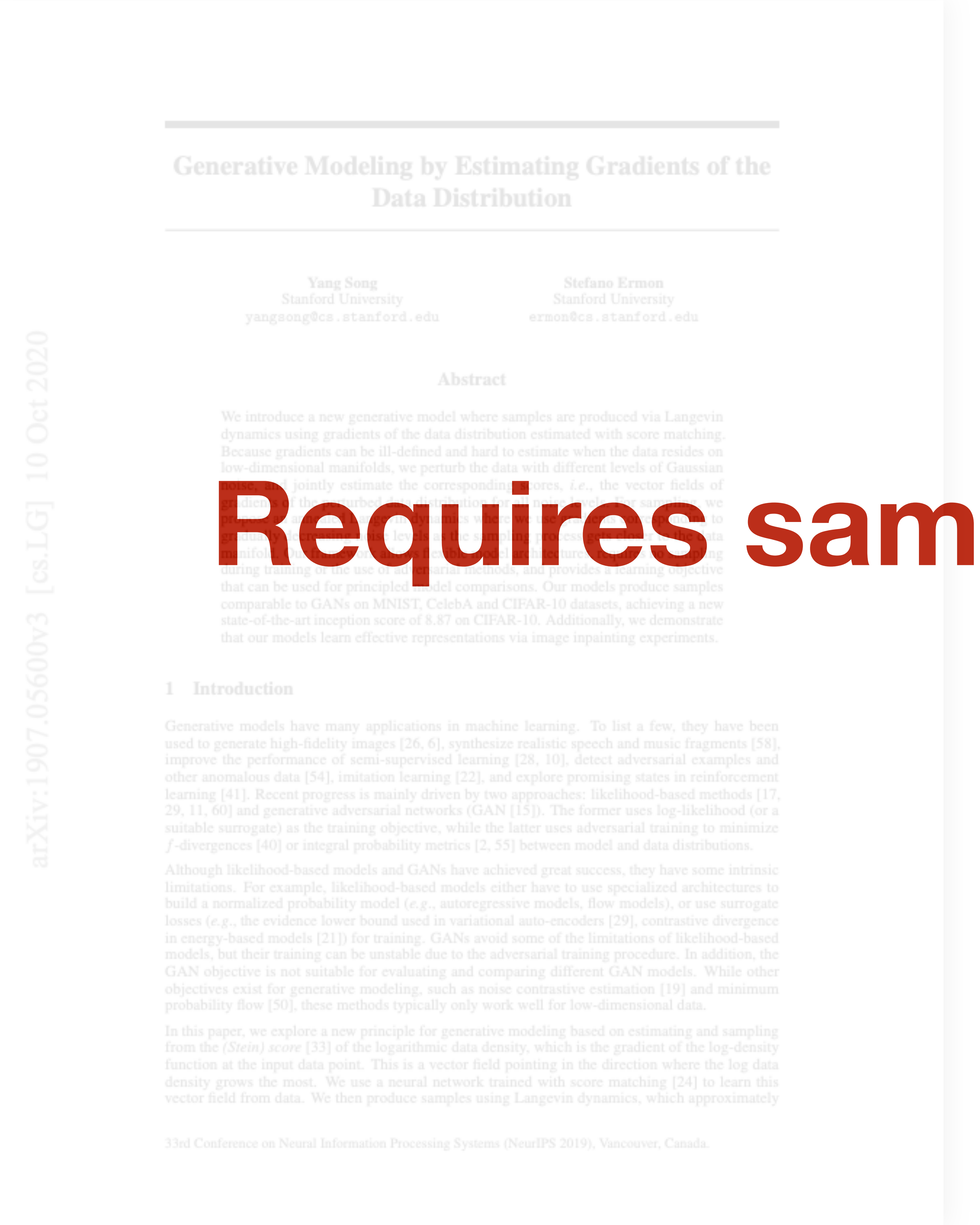
These approaches scale effectively to Internet-scale datasets and have demonstrated practical success in data domains such as images and videos.



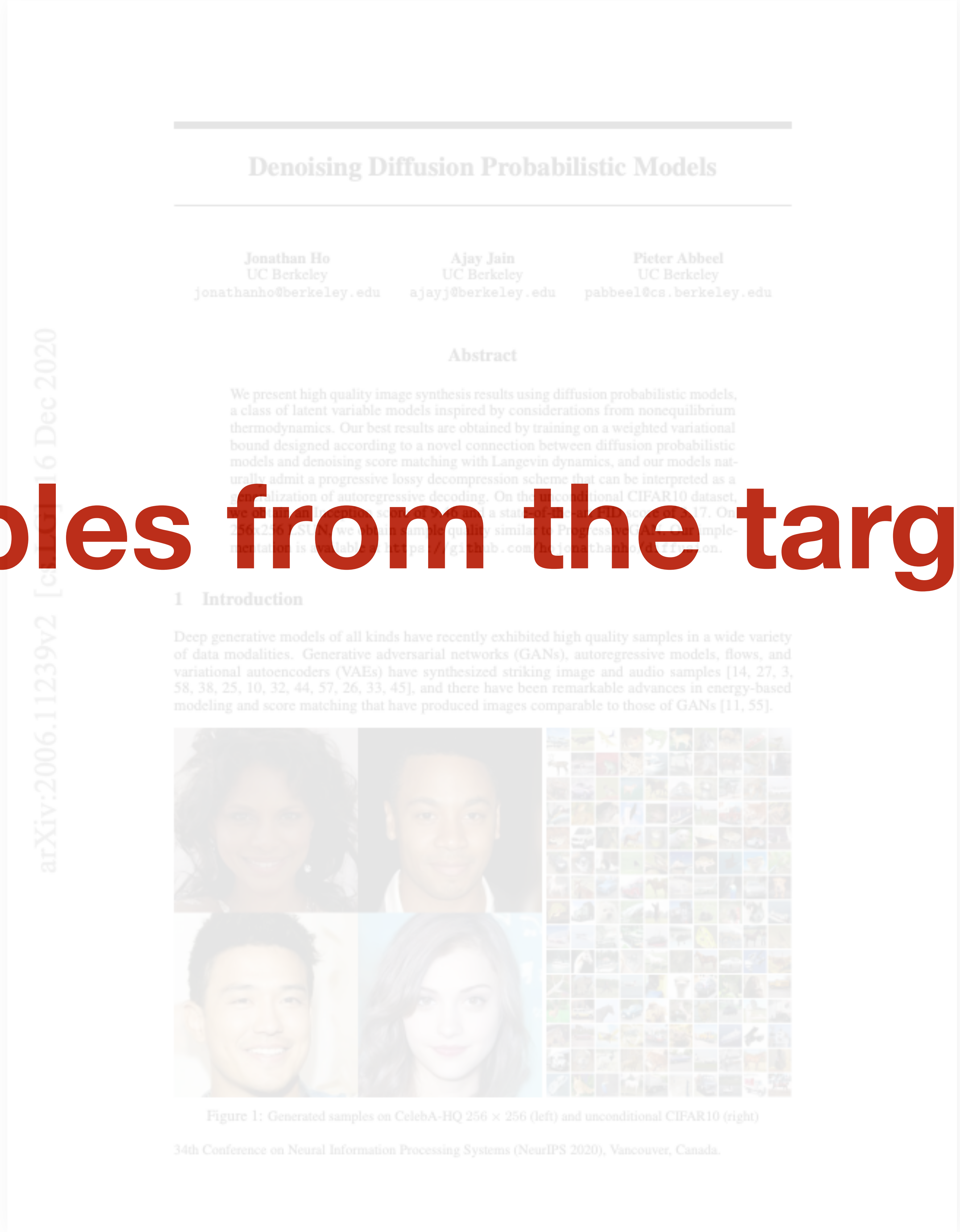
Black Forest Labs, FLUX.2

Problem Definition

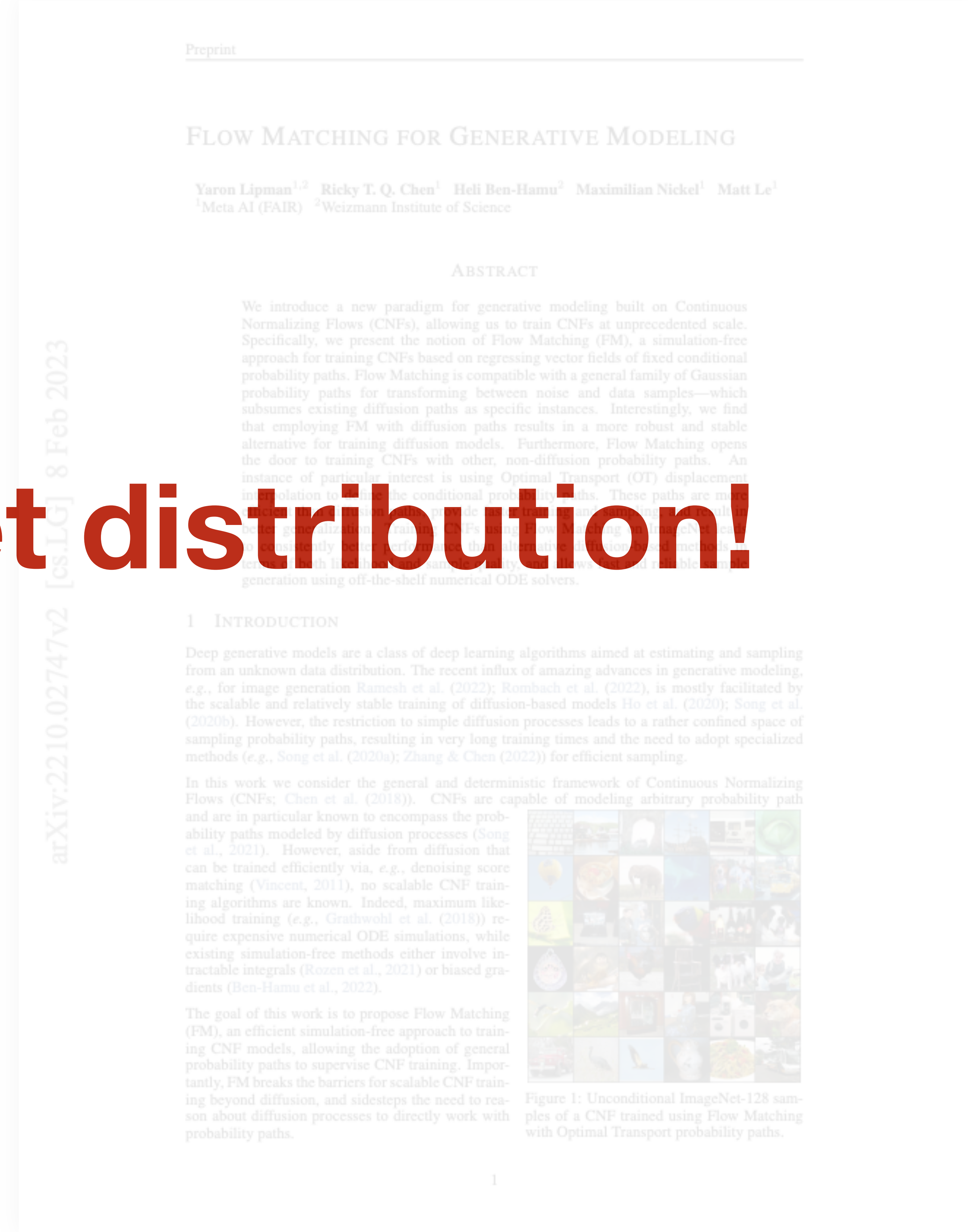
Requires samples from the target distribution!



Score Matching, NeurIPS 2019 (Oral)



Noise Matching, NeurIPS 2020



Flow Matching, ICLR 2023 (Spotlight)

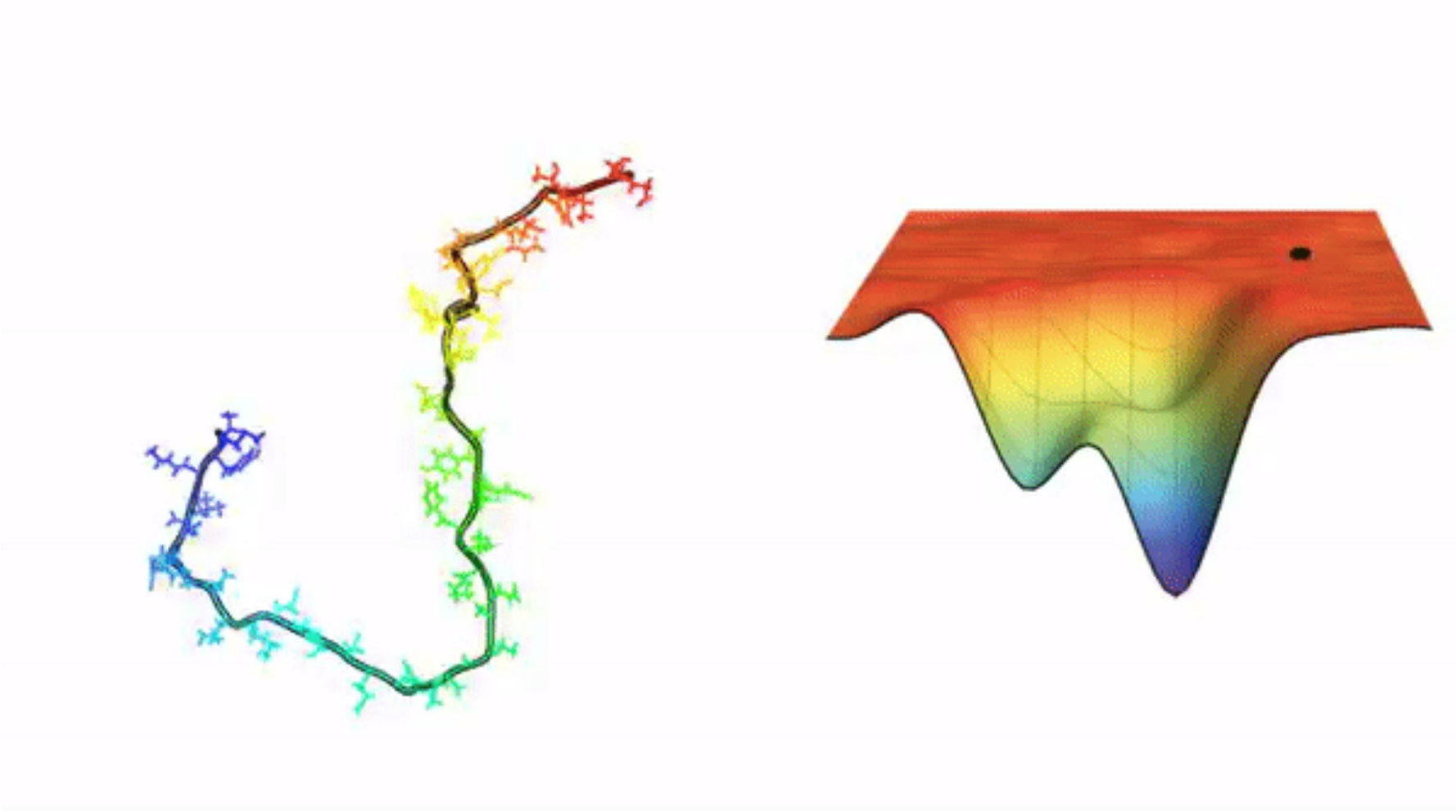
Problem Definition

In Bayesian inference and scientific applications, one is often interested in **sampling from a Boltzmann distribution**:

$$\nu(x) = \frac{e^{-E(x)}}{Z}, \quad Z = \int_{\mathcal{X}} e^{-E(x)} dx$$

which is characterized by an *energy function* $E(x)$, with the normalizer Z being intractable. Here, a lower energy indicates higher likelihood of a sample x.

Problem Definition

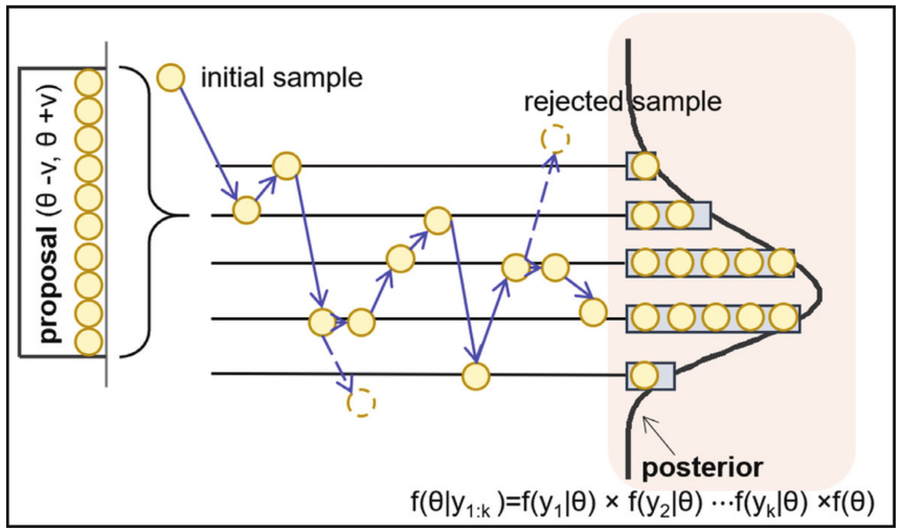


The Protein Folding Problem, Aryan Misra

Problem Definition

Classical methods rely on Markov Chain Monte Carlo (MCMC) algorithms, which run a Markov chain whose stationary distribution is $\nu(x)$, but suffer from

- **Slow mixing time;**
- **Requiring many evaluations of $E(x)$, which is often expensive.**



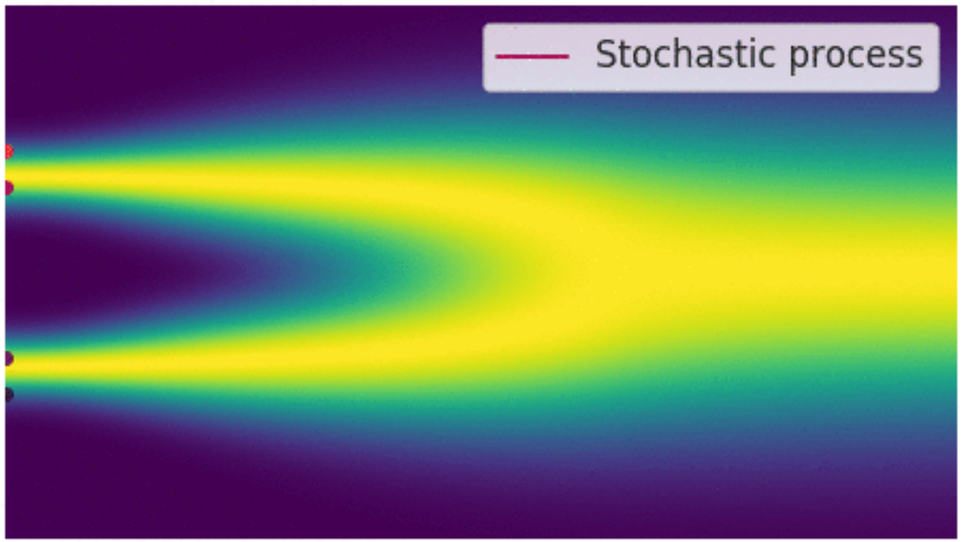
A visual schematic of MCMC algorithm

Problem Definition

Inspired by the recent success of diffusion models, *Diffusion Samplers* have gained attention, which consider stochastic differential equations (SDEs)

$$dX_t = \left(f(X_t) + \sigma_t u_t^\theta(X_t)\right) dt + \sigma_t dW_t \text{ s.t. } X_0 \sim \mu(X_0), X_1 \sim \nu(X_1)$$

as **a bridge that transports samples to the target distribution** $\nu(x)$ at $t = 1$.



Song *et al.*, Score-Based Generative Modeling through Stochastic Differential Equations, ICLR 2021 (Oral)

Problem Definition

Inspired by the recent success of diffusion models, *Diffusion Samplers* have gained attention, which consider stochastic differential equations (SDEs)

$$dX_t = \left(f(X_t) + \sigma_t u_t^{\theta}(X_t)\right) dt + \sigma_t dW_t \text{ s.t. } X_0 \sim \mu(X_0), X_1 \sim \nu(X_1)$$

as **a bridge that transports samples to the target distribution** $\nu(x)$ at $t = 1$.

Prescribed when defining a problem instance and remain fixed!

Problem Definition

Inspired by the recent success of diffusion models, *Diffusion Samplers* have gained attention, which consider stochastic differential equations (SDEs)

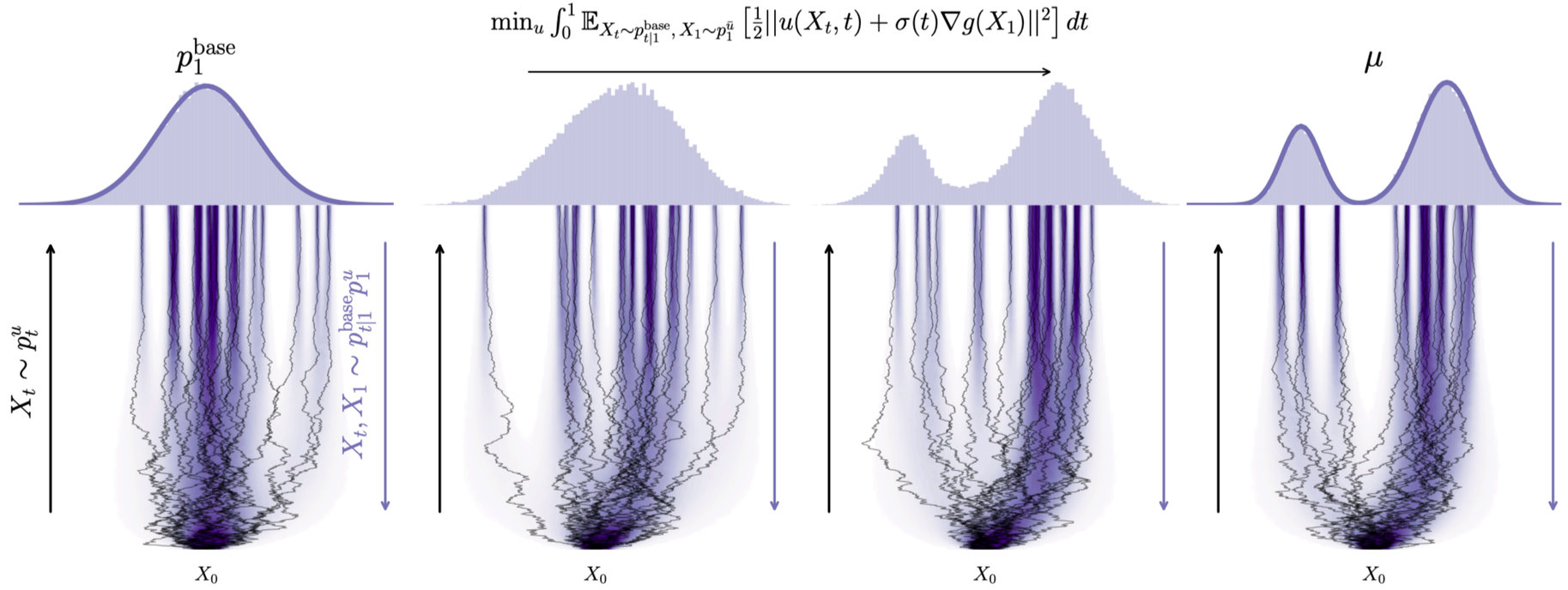
$$dX_t = \left(f(X_t) + \sigma_t u_t^\theta(X_t)\right) dt + \sigma_t dW_t \text{ s.t. } X_0 \sim \mu(X_0), X_1 \sim \nu(X_1)$$

as **a bridge that transports samples to the target distribution** $\nu(x)$ at $t = 1$.

What would be a "scalable algorithm" for learning $u_t^\theta(X_t)$?

Problem Definition

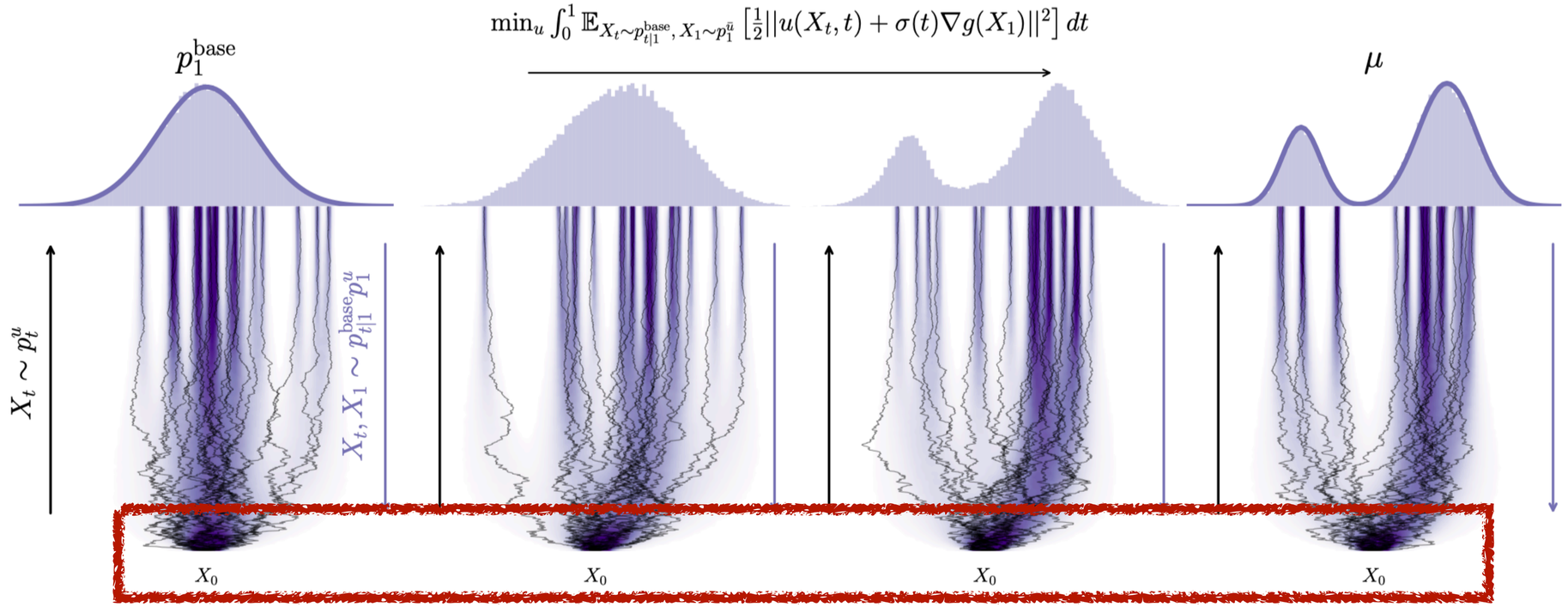
Recently, [Havens *et al.*, ICML 2025] introduced **Adjoint Sampling (AS)**, a class of diffusion samplers based on **stochastic optimal control (SOC) theory** relying only on on-policy samples.



Havens *et al.*, Adjoint Sampling, ICML 2025

Problem Definition

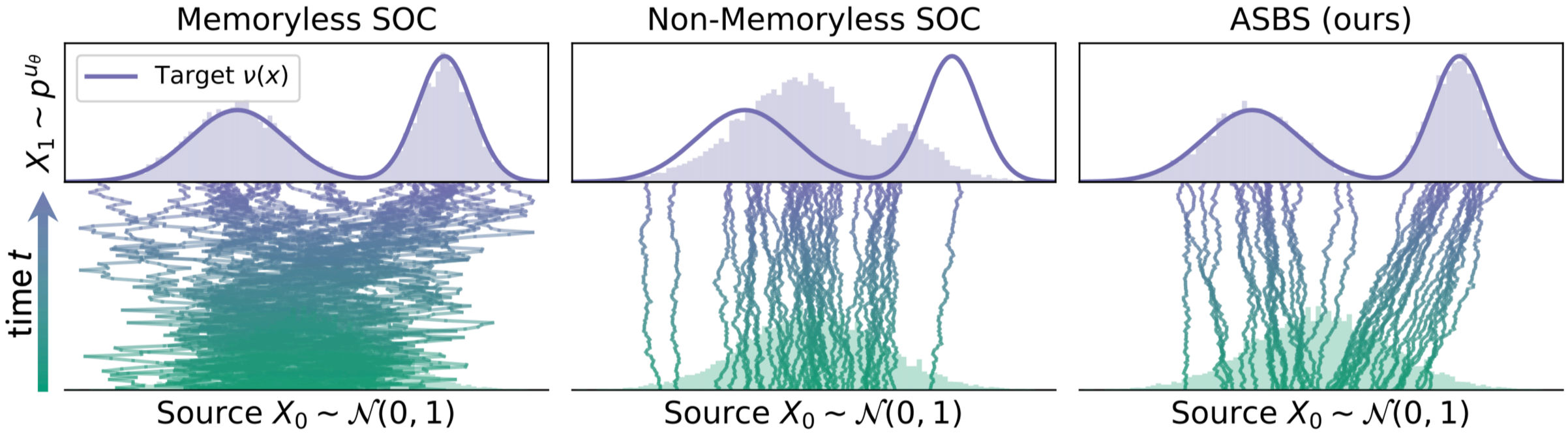
However, Adjoint Matching heavily relies on the **memoryless condition** that restricts the source distribution $\mu(X_0)$ to be Dirac delta, precluding the use of common priors such as Gaussian.



Havens *et al.*, Adjoint Sampling, ICML 2025

Problem Definition

This work proposes **Adjoint Schrödinger Bridge Sampler (ASBS)**, an extension of AS that eliminates the dependency on the memoryless condition.



Liu *et al.*, Adjoint Schrödinger Bridge Sampler (ASBS), NeurIPS 2025 (Oral)

Problem Definition

Formally, ASBS casts learning u_t^θ as a distributionally constrained optimization, known as the Schrödinger Bridge (SB) problem:

$$\min_u D_{KL}(p^u \| p^{base}) = \mathbb{E}_{X\sim p^u}\left[\int_0^1 \frac{1}{2}\|u_t^\theta(X_t)\|^2 dt\right],$$

such that

$$dX_t = \left[f_t(X_t) + \sigma_t u_t^\theta(X_t)\right] dt + \sigma_t dW_t,\ X_0 \sim \mu(X_0), X_1 \sim \nu(X_1).$$

Problem Definition

Formally, ASBS casts learning u_t^θ as a distributionally constrained optimization, known as the Schrödinger Bridge (SB) problem:

$$\min_u D_{KL}(p^u \| p^{base}) = \mathbb{E}_{X\sim p^u}\left[\int_0^1 \frac{1}{2}\|u_t^\theta(X_t)\|^2 dt\right],$$

such that

$$dX_t = \underbrace{[f_t(X_t)}_{\mathbf{0}} + \sigma_t u_t^\theta(X_t)]\, dt + \sigma_t dW_t,\ X_0 \sim \underbrace{\mu}_{\delta(\,\cdot\,)}(X_0), X_1 \sim \nu(X_1).$$

Adjoint Sampling

Key Contributions

Specifically, this paper presents:

1. ASBS, an **SB-based diffusion sampler** capable of sampling target distributions using only unnormalized energy functions;

Key Contributions

Specifically, this paper presents:

1. ASBS, an **SB-based diffusion sampler** capable of sampling target distributions using only unnormalized energy functions;

2. A theoretical framework that **relaxes the memoryless constraint from AS**, while **retaining scalability** of the matching-based algorithm;

Key Contributions

Specifically, this paper presents:

1. ASBS, an **SB-based diffusion sampler** capable of sampling target distributions using only unnormalized energy functions;

2. A theoretical framework that **relaxes the memoryless constraint from AS**, while **retaining scalability** of the matching-based algorithm;

3. **Extensive comparisons against prior methods** spanning Boltzmann distributions of classical energy fnctions, molecular energy potentials.

Preliminaries: Stochastic Optimal Control

Stochastic optimal control (SOC) spans general optimization problems over SDEs. Among them, our particular interest is to solve:

$$\min_{u} = \mathbb{E}_{X \sim p^u} \left[\int_0^1 \frac{1}{2} \| u_t^\theta(X_t) \|^2 dt + g(X_1) \right],$$

where X_t is an intermediate sample along a trajectory governed by an SDE:

$$dX_t = \left[f_t(X_t) + \sigma_t u_t^\theta(X_t) \right] dt + \sigma_t dW_t,\ X_0 \sim \mu(X_0)$$

Preliminaries: Stochastic Optimal Control

For this problem, the optimal distribution p^* induced by the optimal control velocity u^* is analytically known as:

The distribution induced by the base SDE ($u = 0$)

$$p^*(X_0, X_1) = p^{\text{base}}(X_0, X_1) e^{-g(X_1)+V_0(X_0)}$$

where $V_0(X_0)$ is called the (intractable) *initial value function*, defined as:

$$V_0(X_0) = -\log \int p_{1|0}(X_1 \mid X_0) e^{-g(X_1)} dX_1.$$

Preliminaries: Stochastic Optimal Control

For this problem, the optimal distribution p^* induced by the optimal control velocity u^* is analytically known as:

$$p^*(X_0, X_1) = p^{\text{base}}(X_0, X_1) e^{-g(X_1)+V_0(X_0)}$$

where $V_0(X_0)$ is called the (intractable) *initial value function*, defined as:

$$V_0(X_0) = -\log \int p_{1|0}(X_1 \,|\, X_0) e^{-g(X_1)} dX_1.$$

p^* is tilted by the terminal cost $g(\,\cdot\,)$

Preliminaries: Memoryless Condition

However, marginalizing $p^*(X_0, X_1)$ does NOT yield the desired distribution

$$p^*(X_1) \propto p^{\text{base}}(X_1) e^{-g(X_1)} = \nu(X_1),$$

since the initial value function $V_0(X_0)$, serving as a bias, does NOT vanish during marginalization due to the correlation between X_0 and X_1.

Preliminaries: Memoryless Condition

A common approach for mitigating this issue is to assume that the base process is *memoryless* by carefully choosing (f_t, σ_t, μ). This assumption gives us:

$$p^{\text{base}}(X_0, X_1) = p^{\text{base}}(X_0)p^{\text{base}}(X_1).$$

Statistical Independence between (X_0, X_1)

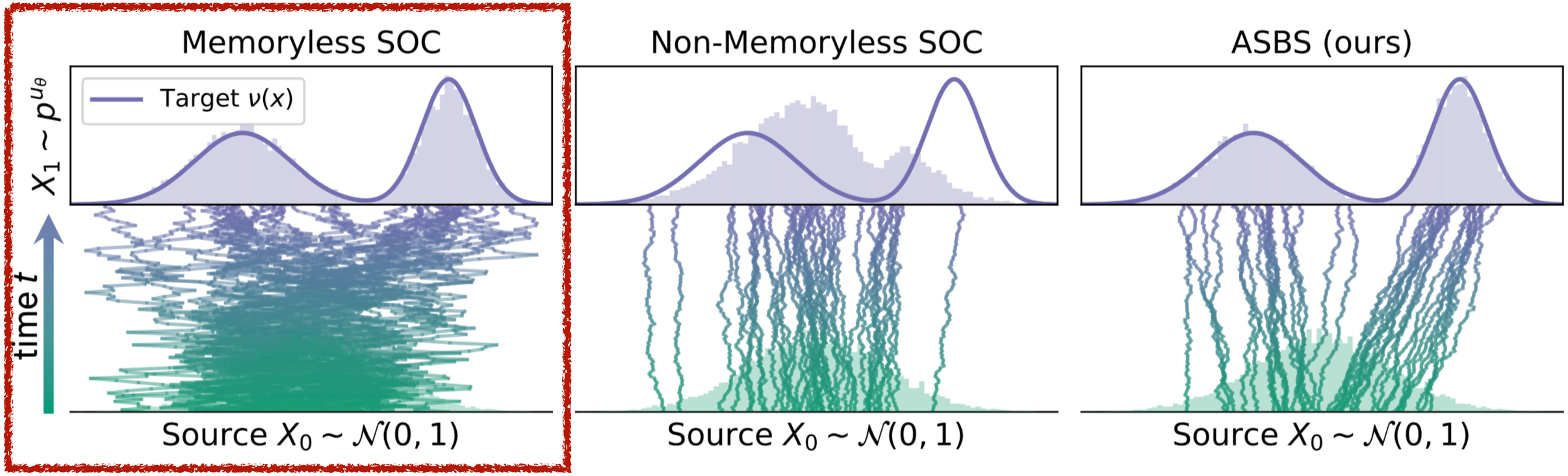
With this condition, marginalization recovers the target distribution μ:

$$p^*(X_1) = \int p^{\text{base}}(X_0)p^{\text{base}}(X_1)e^{-g(X_1)+V_0(X_0)}dX_0 \propto p^{\text{base}}(X_1)e^{-g(X_1)} = \nu(X_1),$$

where the last equality is obtained by setting $g(X_1) = \log \frac{p^{\text{base}}(X_1)}{\nu(X_1)}$.

Preliminaries: Memoryless Condition

For instance, the variance-preserving (VP) process requires a linear base drift f_t, a noise schedule σ_t that grows significantly with time, and a Gaussian prior μ.



Liu *et al.*, Adjoint Schrödinger Bridge Sampler (ASBS), NeurIPS 2025 (Oral)

Preliminaries: Memoryless Condition

Likewise, AS [Havens *et al.*, ICML 2025] limits itself to an SOC problem of form:

$$\min_{u} \mathbb{E}_{X\sim p^{u}}\left[\int_{0}^{1}\frac{1}{2}\|u_t(X_t)\|^2 dt + \log\frac{p^{\text{base}}(X_1)}{\nu(X_1)}\right],$$

with X_t's following a specific instance of the aforementioned general SDE:

$$dX_t = \sigma_t u_t^{\theta}(X_t)dt + \sigma_t dW_t,\ X_0 = 0$$

$$dX_t = \left[f_t(X_t) + \sigma_t u_t^{\theta}(X_t)\right]dt + \sigma_t dW_t,\ X_0 \sim \mu(X_0)$$

Adjoint Schrödinger Bridge Sampler

Instead of SOC-based construction, this paper proposes to learn a diffusion sampler by **solving the Schrödinger Bridge (SB) problem**:

$$\min_{u} D_{KL}(p^{u} \| p^{base}) = \mathbb{E}_{X \sim p^{u}} \left[\int_{0}^{1} \frac{1}{2} \| u_t^{\theta}(X_t) \|^2 dt \right],$$

such that

$$dX_t = \left[f_t(X_t) + \sigma_t u_t^{\theta}(X_t) \right] dt + \sigma_t dW_t,\ X_0 \sim \mu(X_0), X_1 \sim \nu(X_1).$$

Adjoint Schrödinger Bridge Sampler

Specifically, the optimal drift u_t^* satisfies the following optimality equations:

$$u_t^*(x) = \sigma_t \nabla \log \varphi_t(x),$$

where

SB Potentials

$$\begin{cases} \varphi_t(x) = \int p_{1|t}^{\text{base}}(y \,|\, x)\varphi_1(y)dy, & \varphi_0(x)\hat{\varphi}_0(x) = \mu(x) \\ \hat{\varphi}_t(x) = \int p_{t|0}^{\text{base}}(x \,|\, y)\hat{\varphi}_0(y)dy, & \varphi_1(x)\hat{\varphi}_1(x) = \nu(x) \end{cases},$$

and $p_{t|s}^{\text{base}}(y \,|\, x) = p^{\text{base}}(X_t = y \,|\, X_s = x)$ is the transition kernel of the base process.

Adjoint Schrödinger Bridge Sampler

Specifically, the optimal drift u_t^* satisfies the following optimality equations:

$$u_t^*(x) = \sigma_t \nabla \log \varphi_t(x),$$

where

Intractable Integrals!

$$\begin{cases} \varphi_t(x) = \int p_{1|t}^{\text{base}}(y \,|\, x)\varphi_1(y)dy, & \varphi_0(x)\hat{\varphi}_0(x) = \mu(x) \\ \hat{\varphi}_t(x) = \int p_{t|0}^{\text{base}}(x \,|\, y)\hat{\varphi}_0(y)dy, & \varphi_1(x)\hat{\varphi}_1(x) = \nu(x) \end{cases},$$

and $p_{t|s}^{\text{base}}(y \,|\, x) = p^{\text{base}}(X_t = y \,|\, X_s = x)$ is the transition kernel of the base process.

Adjoint Schrödinger Bridge Sampler

Rather than directly solving the numerically intractable SB problem, the authors show that **its optimal solution u^* is the solution of another SOC problem**:

$$\min_u \mathbb{E}_{X\sim p^u}\left[\int_0^1 \frac{1}{2}\|u_t(X_t)\|^2 dt + \log\frac{\hat{\varphi}(X_1)}{\nu(X_1)}\right],$$

such that

$$dX_t = \left[f_t(X_t) + \sigma_t u_t^\theta(X_t)\right]dt + \sigma_t dW_t,\ X_0 \sim \mu(X_0).$$

Adjoint Schrödinger Bridge Sampler

AS

$$\min_u \mathbb{E}_{X\sim p^u}\left[\int_0^1 \frac{1}{2}\|u_t(X_t)\|^2 dt + \log \frac{p^{\text{base}}(X_1)}{\nu(X_1)}\right] \text{ s.t.}$$

$$dX_t = \sigma_t u_t^\theta(X_t)dt + \sigma_t dW_t,\ X_0 = 0$$

ASBS

$$\min_u \mathbb{E}_{X\sim p^u}\left[\int_0^1 \frac{1}{2}\|u_t(X_t)\|^2 dt + \log \frac{\hat{\varphi}(X_1)}{\nu(X_1)}\right] \text{ s.t.}$$

$$dX_t = \left[f_t(X_t) + \sigma_t u_t^\theta(X_t)\right] dt + \sigma_t dW_t,\ X_0 \sim \mu(X_0).$$

Adjoint Schrödinger Bridge Sampler

Similar to the SOC considered in AS [Havens *et al.*, ICML 2025], the optimal distribution $p^*(X_0, X_1)$ achievable by solving the problem is:

$$p^*(X_0, X_1) = p^{\text{base}}(X_0, X_1)\exp\left(-\log\frac{\hat{\varphi}_1(X_1)}{\nu(X_1)} - \log\varphi_0(X_0)\right),$$

where "$-\log\varphi_0(X_0)$" is the *initial value function* in this case.

Adjoint Schrödinger Bridge Sampler

Interestingly, by marginalizing this density over X_1, one can show that

$$p^*(X_1) = \frac{\nu(X_1)}{\hat{\varphi}_1(X_1)} \int p^{\text{base}}(X_0, X_1) \frac{1}{\varphi_0(X_0)} dX_0.$$

Leveraging the definitions of the SB potentials φ_t, $\hat{\varphi}_t$, the expression further simplifies to:

$$p^*(X_1) = \frac{\nu(X_1)}{\hat{\varphi}_1(X_1)} \int p^{\text{base}}(X_1 \,|\, X_0) \hat{\varphi}_0(X_0) dX_0 = \nu(X_1).$$

That is, **we now have a theoretical guarantee that solving this SOC will allow us to learn the unbiased target distribution $\nu(X_1)$.**

How can we learn such u?

Recap: Adjoint Matching

The most naive approach is gradient-descent, which iteratively optimizes

$$\min_{\theta} \mathscr{L}(u_t^{\theta}; \mathbf{X}) = \int_0^1 \left(\frac{1}{2} \|u_t^{\theta}(X_t)\|^2 + f(X_t) \right) dt + g(X_1),$$

with $g(X_1) = \log \dfrac{\varphi(X_1)}{\nu(X_1)}$, using autodifferentiation to compute $\dfrac{\partial \mathscr{L}}{\partial \theta}$.

Recap: Adjoint Matching

The most naive approach is gradient-descent, which iteratively optimizes

$$\min_\theta \mathscr{L}(u_t^\theta; \mathbf{X}) = \int_0^1 \left(\frac{1}{2} \|u_t^\theta(X_t)\|^2 + f(X_t) \right) dt + g(X_1),$$

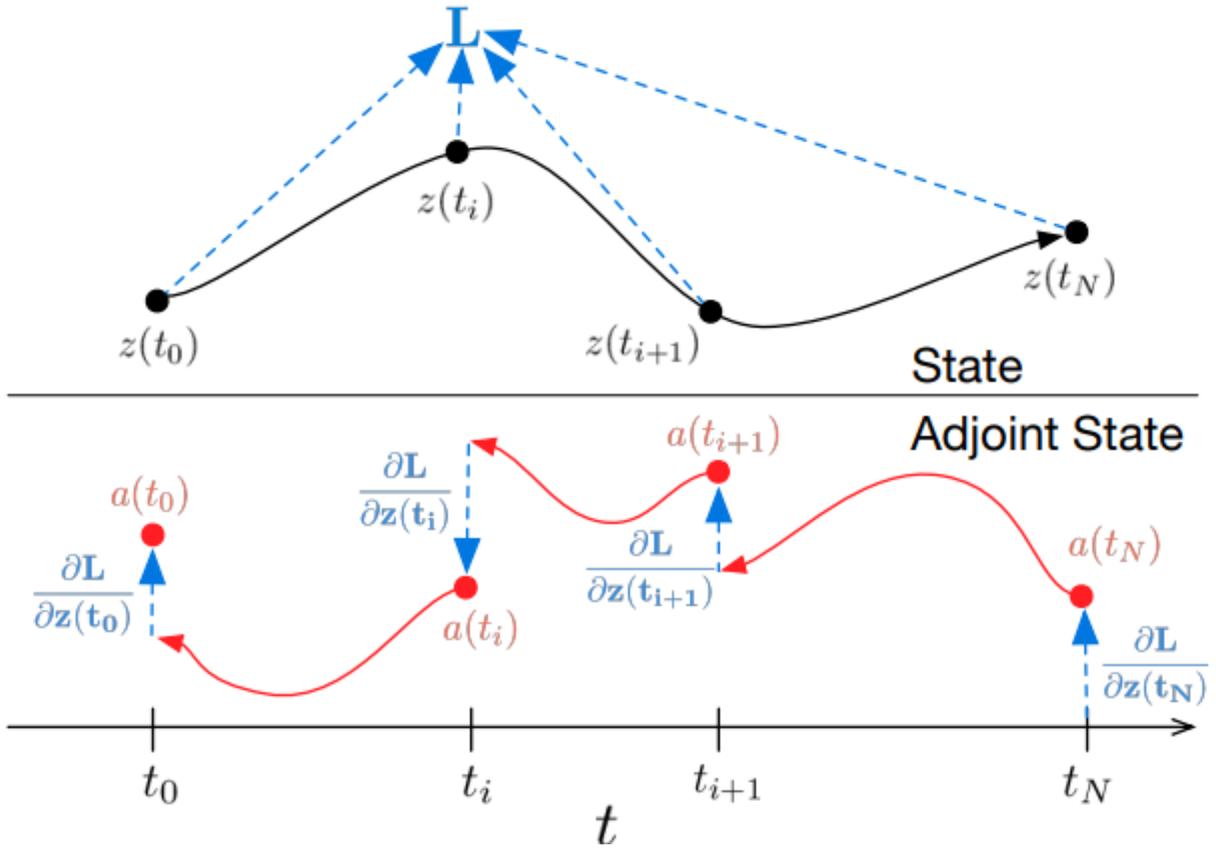
with $g(X_1) = \log \frac{\varphi(X_1)}{\nu(X_1)}$, using autodifferentiation to compute $\frac{\partial \mathscr{L}}{\partial \theta}$.

Backpropagating through “hundreds” of forward passes

Quickly runs out of memory

Recap: Adjoint Matching

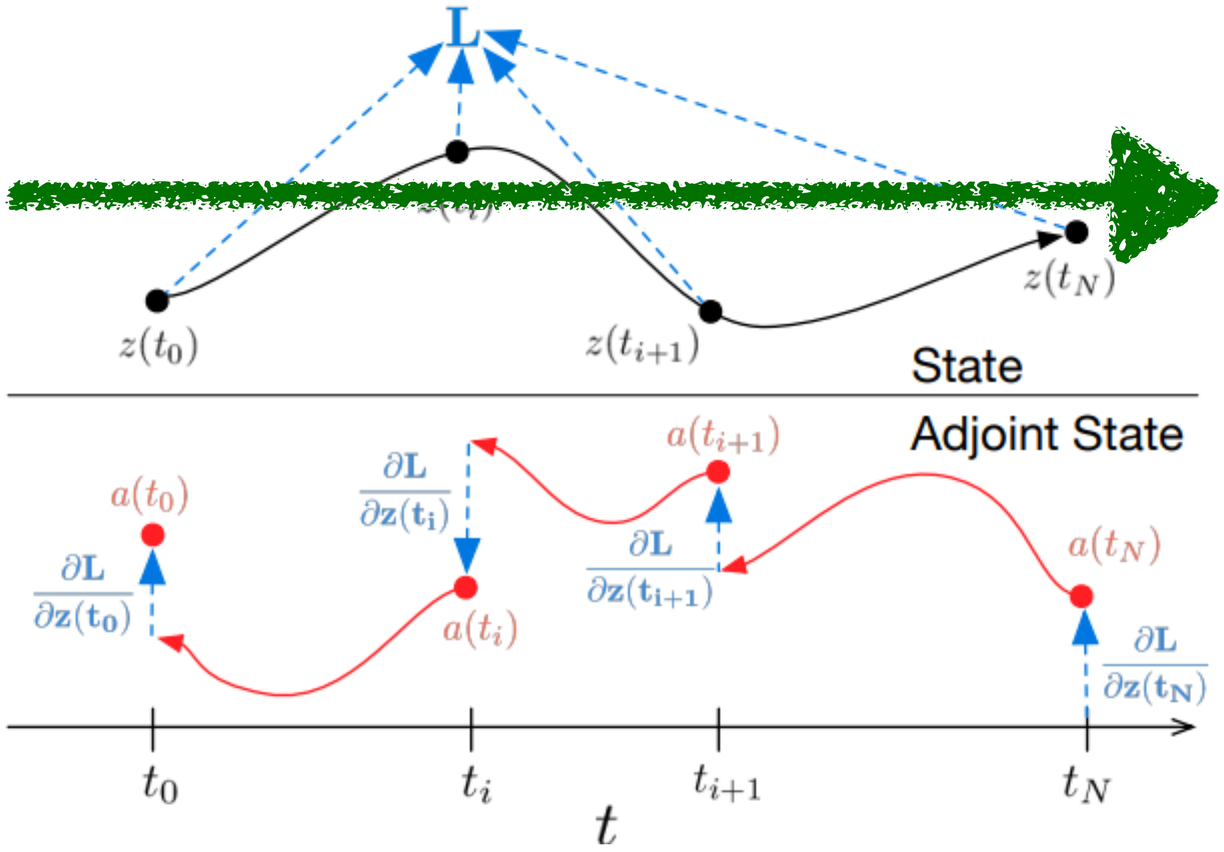
The **adjoint method** [Pontryagin *et al.*, 1962] enables gradient computation with constant memory complexity, at the expense of increased computational cost.



Chen *et al.*, Neural Ordinary Differential Equations, NeurIPS 2018 (Best Paper Award)

Recap: Adjoint Matching

The **adjoint method** [Pontryagin *et al.*, 1962] enables gradient computation with constant memory complexity, at the expense of increased computational cost.

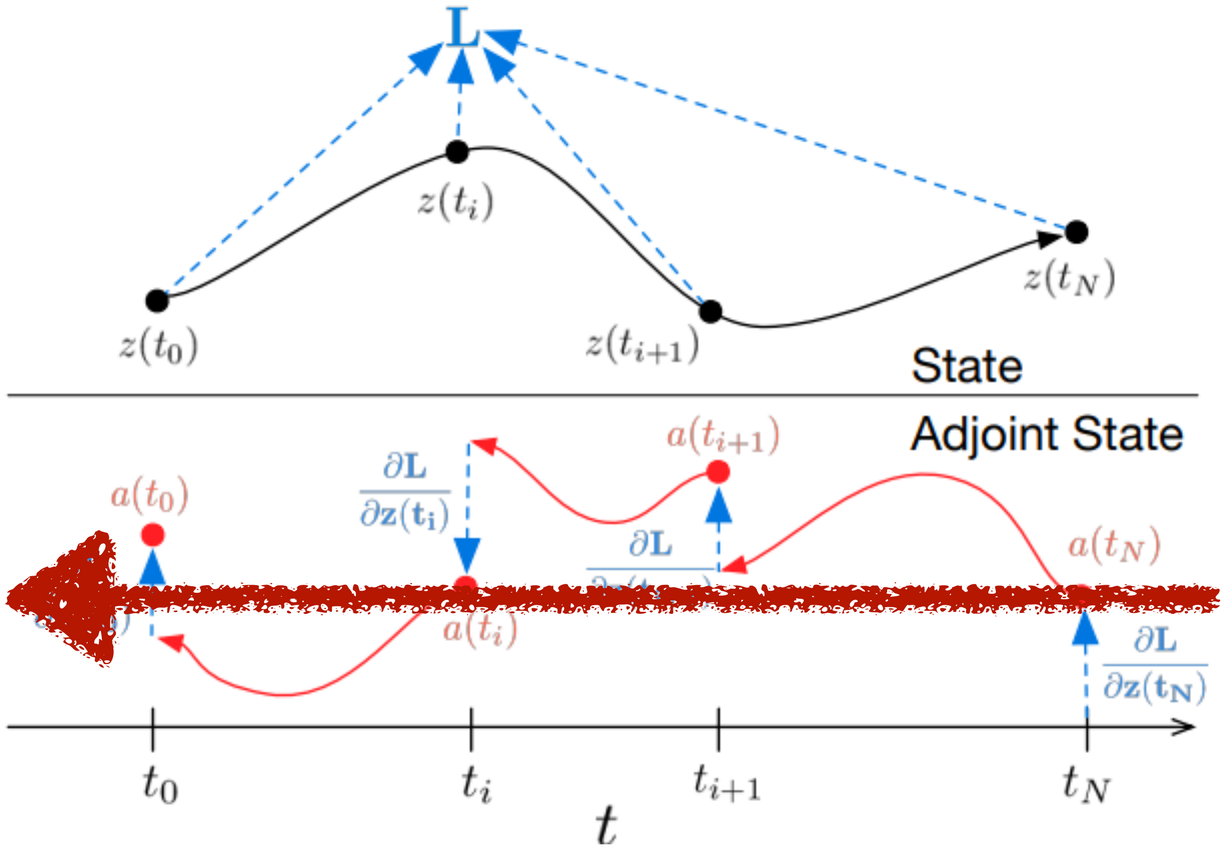


Forward simulation to compute $g(X_1)$

Chen *et al.*, Neural Ordinary Differential Equations, NeurIPS 2018 (Best Paper Award)

Recap: Adjoint Matching

The **adjoint method** [Pontryagin *et al.*, 1962] enables gradient computation with constant memory complexity, at the expense of increased computational cost.



Backward simulation to compute adjoints $a(t)$

Chen *et al.*, Neural Ordinary Differential Equations, NeurIPS 2018 (Best Paper Award)

Recap: Adjoint Matching

Crucially, [Domingo-Enrich *et al.*, ICLR 2025] have proven that one can convert the naive SOC objective to the **Adjoint Matching loss** of form:

$$\mathscr{L}_{\text{Adj-Match}}(u_t^\theta; \mathbf{X}) = \frac{1}{2}\int_0^1 \|u_t^\theta(X_t) + \sigma_t^\intercal \tilde{a}(t; \mathbf{X})\|^2 dt, \quad \mathbf{X} \sim p^{\bar{u}_t^\theta}$$

where $\frac{d}{dt}\tilde{a}(t; \mathbf{X}) = -\left(\tilde{a}(t; \mathbf{X})^\intercal \nabla_{X_t} f(X_t)\right)$, $\tilde{a}(1; \mathbf{X}) = \nabla_{X_1} g(X_1)$,

and $p^{\bar{u}_t^\theta}$ is the path distribution induced by $\bar{u}_t^\theta = \text{stopgrad}(u_t^\theta)$.

Adjoint Schrödinger Bridge Sampler

Applying Adjoint Matching, our loss for solving the SB problem becomes:

$$\mathbb{E}_{p^{base}_{t|0,1}, p^{\bar{u}}_{0,1}} \left[\left\| u_t(X_t) + \sigma_t \left(\nabla E + \nabla \log \hat{\varphi}_1 \right)(X_1) \right\|^2 \right], \quad \bar{u} = \text{stopgrad}(u)$$

where the *corrector* gradient $\nabla \log \hat{\varphi}_1(x)$ is the solution of another problem:

$$\nabla \log \hat{\varphi}_1 = \underset{h}{\text{argmin}} \mathbb{E}_{p^{u^*}_{0,1}} \left[\left\| h(X_1) - \nabla_{X_1} \log p^{base}(X_1 \,|\, X_0) \right\|^2 \right].$$

Adjoint Schrödinger Bridge Sampler

Applying Adjoint Matching, our loss for solving the SB problem becomes:

$$\mathbb{E}_{p^{base}_{t|0,1}, p^{\bar{u}}_{0,1}} \left[\left\| u_t(X_t) + \sigma_t \left(\nabla E + \nabla \log \hat{\varphi}_1 \right)(X_1) \right\|^2 \right], \quad \bar{u} = \text{stopgrad}(u)$$

where the *corrector* gradient $\nabla \log \hat{\varphi}_1(x)$ is the solution of another problem:

$$\nabla \log \hat{\varphi}_1 = \underset{h}{\text{argmin}} \mathbb{E}_{p^{u^*}_{0,1}} \left[\left\| h(X_1) - \nabla_{X_1} \log p^{base}(X_1 \mid X_0) \right\|^2 \right].$$

Adjoint Schrödinger Bridge Sampler

Applying Adjoint Matching, our loss for solving the SB problem becomes:

$$\mathbb{E}_{p^{base}_{t|0,1}, p^{\bar{u}}_{0,1}} \left[\left\| u_t(X_t) + \sigma_t \left(\nabla E + \nabla \log \hat{\varphi}_1 \right)(X_1) \right\|^2 \right], \quad \bar{u} = \text{stopgrad}(u)$$

where the *corrector* gradient $\nabla \log \hat{\varphi}_1(x)$ is the solution of another problem:

$$\nabla \log \hat{\varphi}_1 = \underset{h}{\text{argmin}} \mathbb{E}_{p^{u^*}_{0,1}} \left[\left\| h(X_1) - \nabla_{X_1} \log p^{base}(X_1 \mid X_0) \right\|^2 \right].$$

Adjoint Schrödinger Bridge Sampler

Applying Adjoint Matching, our loss for solving the SB problem becomes:

$$\mathbb{E}_{p^{base}_{t|0,1}, p^{\bar{u}}_{0,1}} \left[\left\| u_t(X_t) + \sigma_t \left(\nabla E + \nabla \log \hat{\varphi}_1 \right)(X_1) \right\|^2 \right], \quad \bar{u} = \text{stopgrad}(u)$$

where the *corrector* gradient $\nabla \log \hat{\varphi}_1(x)$ is the solution of another problem:

$$\nabla \log \hat{\varphi}_1 = \underset{h}{\text{argmin}} \mathbb{E}_{p^{u^*}_{0,1}} \left[\left\| h(X_1) - \nabla_{X_1} \log p^{base}(X_1 \,|\, X_0) \right\|^2 \right].$$

Algorithm 1 Adjoint Schrödinger Bridge Sampler (ASBS)

Require: Sample-able source $X_0 \sim \mu$, differentiable energy $E(x)$, parametrized $u_\theta(t, x)$ and $h_\phi(x)$

1: Initialize $h_\phi^{(0)} := 0$
2: **for** stage k **in** $1, 2, \ldots$ **do**
3: Update drift $u_\theta^{(k)}$ by solving (14) ▷ adjoint matching
4: Update corrector $h_\phi^{(k)}$ by solving (15) ▷ corrector matching
5: **end for**

Adjoint Schrödinger Bridge Sampler

The alternating optimization produces a sequence of updates

$$(u^{(0)}, h^{(0)}) \to \cdots \to (u^{(k)}, h^{(k)})$$

which can be interpreted as the coordinate descent between u and h. Furthermore, the authors show that **this optimization scheme converges to the true solution of the SB problem**.

Experiments

ASBS is evaluated on three classes of multi-particle energy functions.

Synthetic Energy Functions (Analytically Known Potentials)

1. 2D 4-particle Double-Well potential (DW-4)
2. 1D 5-particle Many-Well potential (MW-5)
3. 3D 13-particle Lennard-Jones potential (LJ-13)
4. 3D 55-particle Lennard-Jones potential (LJ-55)

Experiments

ASBS is evaluated on three classes of multi-particle energy functions.

Alanine Diepeptide (Molecule with 22 atoms in 3D)

1. Samples from the Boltzmann distribution of the molecules in a solvent;

2. Uses an energy function $E(x)$ from the OpenMM library to poulate 10^7 GT configurations.

Experiments

ASBS is evaluated on three classes of multi-particle energy functions.

Amortized Conformer Generation

1. Conformer: Molecule configurations at the local minima of the molecule's potential energy surface;

2. Samples $\nu(x \mid g) \propto e^{-\frac{1}{\tau}E(x|g)}$ where g is the molecular topology;

3. Training and test sets include 25K and 80 molecular toplogies, respectively;

4. Employs *eSEN* [Fu *et al.*, 2025], a neural network approximating $E(x \mid g)$.

Experiments

ASBS is compared against previous diffusion sampler, including

1. PIS [Zhang and Chen, ICLR 2022]
2. DDS [Vargas *et al.*, ICLR 2023]
3. PDDS [Phillips *et al.*, ICML 2024]
4. SCLD [Chen *et al.*, ICLR 2025]
5. LV [Richter and Berner, ICLR 2024]
6. iDEM [Akhound-Sadegh *et al.*, ICML 2024]
7. AS [Havens *et al.*, ICML 2025]

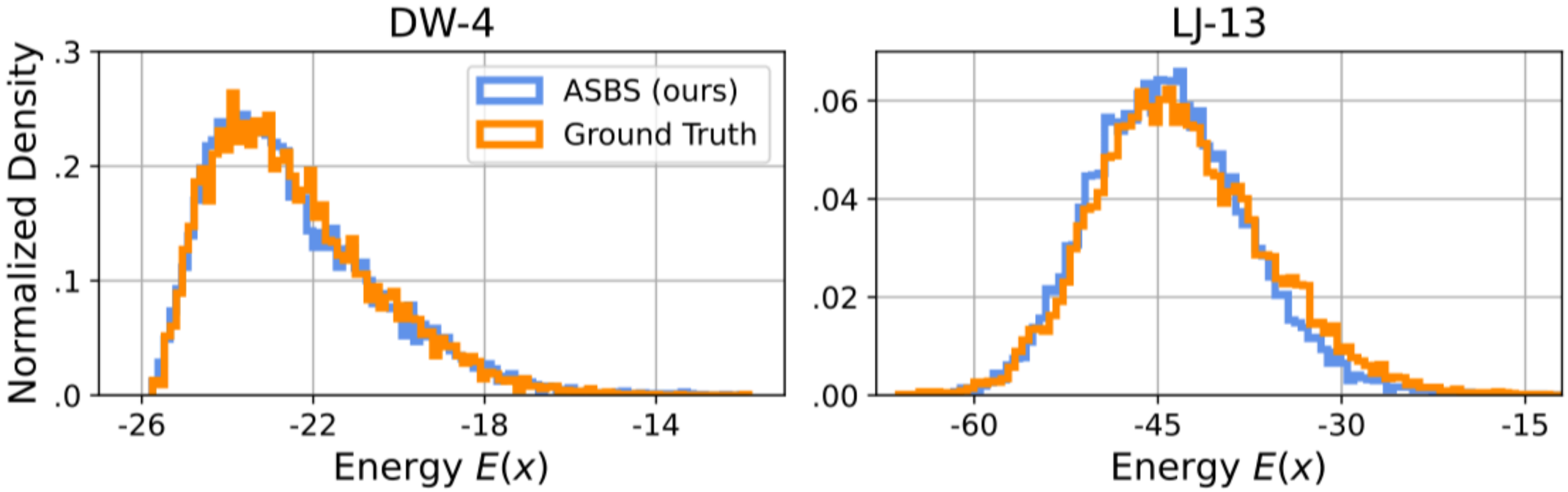
Experiments

On synthetic energy functions, ASBS outperforms all previous diffusion samplers. ($\mathscr{W}_2$/$E(\cdot)\mathscr{W}_2$: Wasserstein-2 distances w.r.t samples / energies)

	MW-5 (d=5)	DW-4 ($d = 8$)		LJ-13 ($d = 39$)		LJ-55 ($d = 165$)	
Method	Sinkhorn ↓	$\mathcal{W}_2$ ↓	$E(\cdot)$ $\mathcal{W}_2$ ↓	$\mathcal{W}_2$ ↓	$E(\cdot)$ $\mathcal{W}_2$ ↓	$\mathcal{W}_2$ ↓	$E(\cdot)$ $\mathcal{W}_2$ ↓
PDDS (Phillips et al., 2024)	—	$0.92_{\pm 0.08}$	$0.58_{\pm 0.25}$	$4.66_{\pm 0.87}$	$56.01_{\pm 10.80}$	—	—
SCLD (Chen et al., 2025)	$0.44_{\pm 0.06}$	$1.30_{\pm 0.64}$	$0.40_{\pm 0.19}$	$2.93_{\pm 0.19}$	$27.98_{\pm 1.26}$	—	—
PIS (Zhang and Chen, 2022)	$0.65_{\pm 0.25}$	$0.68_{\pm 0.28}$	$0.65_{\pm 0.25}$	$1.93_{\pm 0.07}$	$18.02_{\pm 1.12}$	$4.79_{\pm 0.45}$	$228.70_{\pm 131.27}$
DDS (Vargas et al., 2023)	$0.63_{\pm 0.24}$	$0.92_{\pm 0.11}$	$0.90_{\pm 0.37}$	$1.99_{\pm 0.13}$	$24.61_{\pm 8.99}$	$4.60_{\pm 0.09}$	$173.09_{\pm 18.01}$
LV-PIS (Richter and Berner, 2024)	—	$1.04_{\pm 0.29}$	$1.89_{\pm 0.89}$	—	—	—	—
iDEM (Akhound-Sadegh et al., 2024)	—	$0.70_{\pm 0.06}$	$0.55_{\pm 0.14}$	$1.61_{\pm 0.01}$	$30.78_{\pm 24.46}$	$4.69_{\pm 1.52}$	$93.53_{\pm 16.31}$
AS (Havens et al., 2025)	$0.32_{\pm 0.06}$	$0.62_{\pm 0.06}$	$0.55_{\pm 0.12}$	$1.67_{\pm 0.01}$	$2.40_{\pm 1.25}$	$4.04_{\pm 0.05}$	$30.83_{\pm 8.19}$
ASBS (**Ours**)	$0.15_{\pm 0.02}$	$0.43_{\pm 0.05}$	$0.20_{\pm 0.11}$	$1.59_{\pm 0.03}$	$1.99_{\pm 1.01}$	$4.00_{\pm 0.03}$	$28.10_{\pm 8.15}$

Experiments

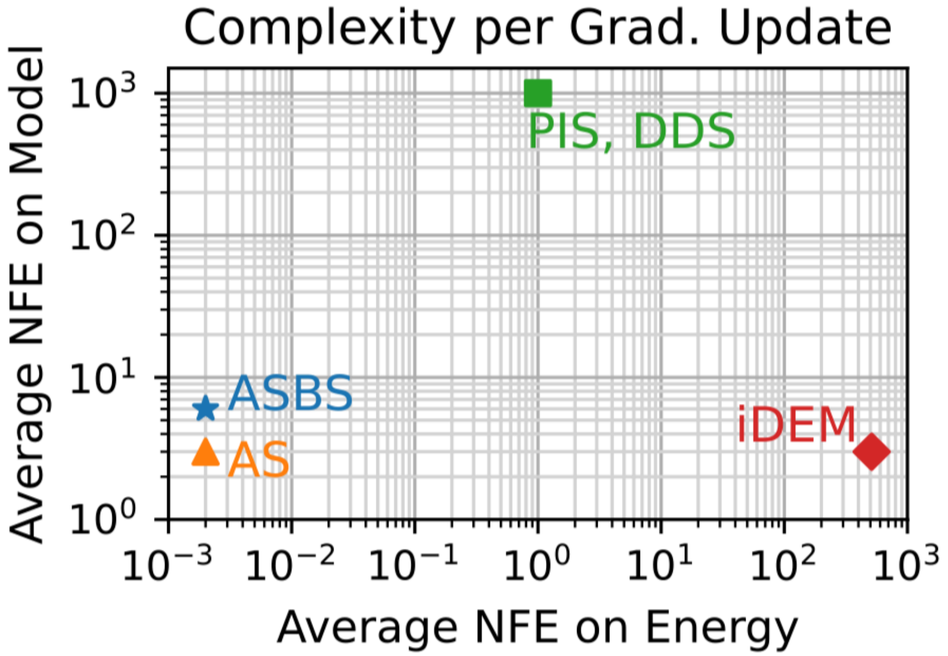
Notably, on DW-4 and LJ-13 energies, energy histograms of ASBS samples closely resemble those of MCMC samples, treated as the ground truth.



Energy histograms of DW-4 and LJ-13 comparing ASBS against MCMC (GT).

Experiments

Furthermore, ASBS retains the scalability of AS [Havens *et al.*, ICML 2025], requiring far less number of energy function evaluations.



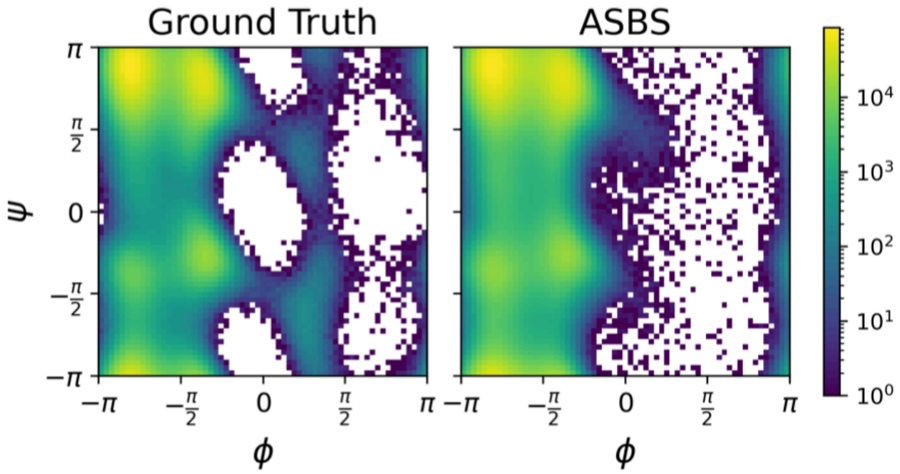
Complexity with respect to the model and energy NFE on LJ-13 potential.

Experiments

On the task of sampling the Boltzmann distribution of the alanine dipeptide, ASBS samples achieves the lowest KL divergence and Wasserstein-2 distance.

	without relaxation				with relaxation			
	SPICE		GEOM-DRUGS		SPICE		GEOM-DRUGS	
Method	Coverage ↑	AMR ↓	Coverage ↑	AMR ↓	Coverage ↑	AMR ↓	Coverage ↑	AMR ↓
RDKit ETKDG (Riniker and Landrum, 2015)	$56.94_{\pm 35.82}$	$1.04_{\pm 0.52}$	$50.81_{\pm 34.69}$	$1.15_{\pm 0.61}$	$70.21_{\pm 31.70}$	$0.79_{\pm 0.44}$	$62.55_{\pm 31.67}$	$0.93_{\pm 0.53}$
AS (Havens et al., 2025)	$56.75_{\pm 38.15}$	$0.96_{\pm 0.26}$	$36.23_{\pm 33.42}$	$1.20_{\pm 0.43}$	$82.41_{\pm 25.85}$	$0.68_{\pm 0.28}$	$64.26_{\pm 34.57}$	$0.89_{\pm 0.45}$
ASBS w/ Gaussian prior (**Ours**)	$73.04_{\pm 31.95}$	$0.83_{\pm 0.24}$	$50.23_{\pm 35.98}$	$1.05_{\pm 0.43}$	$88.26_{\pm 20.57}$	$0.60_{\pm 0.24}$	$72.32_{\pm 29.68}$	$0.77_{\pm 0.35}$
ASBS w/ harmonic prior (**Ours**)	$74.05_{\pm 31.61}$	$0.82_{\pm 0.23}$	$53.14_{\pm 35.69}$	$1.03_{\pm 0.42}$	$88.71_{\pm 18.63}$	$0.59_{\pm 0.24}$	$72.77_{\pm 29.94}$	$0.78_{\pm 0.35}$
AS +RDKit warmup (Havens et al., 2025)	$72.21_{\pm 30.22}$	$0.84_{\pm 0.24}$	$52.19_{\pm 35.20}$	$1.02_{\pm 0.34}$	$87.84_{\pm 19.20}$	$0.60_{\pm 0.23}$	$73.88_{\pm 28.63}$	$0.76_{\pm 0.34}$
ASBS +RDKit warmup (**Ours**)	$77.84_{\pm 28.37}$	$0.79_{\pm 0.23}$	$57.19_{\pm 35.14}$	$0.98_{\pm 0.40}$	$88.08_{\pm 18.84}$	$0.58_{\pm 0.24}$	$73.18_{\pm 30.09}$	$0.76_{\pm 0.37}$

Quantitative comparisons on Alanine Dipeptide's Boltzmann distribution.



Ramachandran plots.

Experiments

ASBS outperforms AS, which is restricted to Dirac-Delta priors, benefiting from the use of Gaussian and harmonic prior distributions.

	without relaxation				with relaxation			
	SPICE		GEOM-DRUGS		SPICE		GEOM-DRUGS	
Method	Coverage ↑	AMR ↓	Coverage ↑	AMR ↓	Coverage ↑	AMR ↓	Coverage ↑	AMR ↓
RDKit ETKDG (Riniker and Landrum, 2015)	$56.94_{\pm 35.82}$	$1.04_{\pm 0.52}$	$50.81_{\pm 34.69}$	$1.15_{\pm 0.61}$	$70.21_{\pm 31.70}$	$0.79_{\pm 0.44}$	$62.55_{\pm 31.67}$	$0.93_{\pm 0.53}$
AS (Havens et al., 2025)	$56.75_{\pm 38.15}$	$0.96_{\pm 0.26}$	$36.23_{\pm 33.42}$	$1.20_{\pm 0.43}$	$82.41_{\pm 25.85}$	$0.68_{\pm 0.28}$	$64.26_{\pm 34.57}$	$0.89_{\pm 0.45}$
ASBS w/ Gaussian prior (**Ours**)	$73.04_{\pm 31.95}$	$0.83_{\pm 0.24}$	$50.23_{\pm 35.98}$	$1.05_{\pm 0.43}$	$88.26_{\pm 20.57}$	$0.60_{\pm 0.24}$	$72.32_{\pm 29.68}$	$0.77_{\pm 0.35}$
ASBS w/ harmonic prior (**Ours**)	$74.05_{\pm 31.61}$	$0.82_{\pm 0.23}$	$53.14_{\pm 35.69}$	$1.03_{\pm 0.42}$	$88.71_{\pm 18.63}$	$0.59_{\pm 0.24}$	$72.77_{\pm 29.94}$	$0.78_{\pm 0.35}$
AS +RDKit warmup (Havens et al., 2025)	$72.21_{\pm 30.22}$	$0.84_{\pm 0.24}$	$52.19_{\pm 35.20}$	$1.02_{\pm 0.34}$	$87.84_{\pm 19.20}$	$0.60_{\pm 0.23}$	$73.88_{\pm 28.63}$	$0.76_{\pm 0.34}$
ASBS +RDKit warmup (**Ours**)	$77.84_{\pm 28.37}$	$0.79_{\pm 0.23}$	$57.19_{\pm 35.14}$	$0.98_{\pm 0.40}$	$88.08_{\pm 18.84}$	$0.58_{\pm 0.24}$	$73.18_{\pm 30.09}$	$0.76_{\pm 0.37}$

Conclusion

To conclude, this paper:

1. introduces **Adjoint Schrödinger Bridge Sampler (ASBS)**, a novel diffusion sampler that solves general SB problems given only energy functions;

Conclusion

To conclude, this paper:

1. introduces **Adjoint Schrödinger Bridge Sampler (ASBS)**, a novel diffusion sampler that solves general SB problems given only energy functions;

2. provides **a theoretical analysis of previous SOC-based approaches**, a **matching objective** that solves a SB problem, proves its global convergence;

Conclusion

To conclude, this paper:

1. introduces **Adjoint Schrödinger Bridge Sampler (ASBS)**, a novel diffusion sampler that solves general SB problems given only energy functions;

2. provides **a theoretical analysis of previous SOC-based approaches**, a **matching objective** solving a SB problem, and proves its global convergence;

3. **demonstrates the superior performance** over baselines on various energy functions, including molecular conformer generation.

Adjoint Schrödinger Bridge Sampler

NeurIPS 2025 (Oral)

Guan-Horng Liu*, Jaemoo Choi*, Yongxin Chen, Benjamin Kurt Miller, and Ricky T. Q. Chen

Seungwoo Yoo, KAIST Visual AI Group