Poisson Flow Generative Models

Yilun Xu*, **Ziming Liu***, **Max Tegmark**, **Tommi Jaakkola**
Massachusetts Institute of Technology
{ylxu, zmliu, tegmark}@mit.edu; tommi@csail.mit.edu

Abstract

We propose a new "Poisson flow" generative model (PFGM) that maps a uniform distribution on a high-dimensional hemisphere into any data distribution. We interpret the data points as electrical charges on the $z = 0$ hyperplane in a space augmented with an additional dimension z, generating a high-dimensional electric field (the gradient of the solution to Poisson equation). We prove that if these charges flow upward along electric field lines, their initial distribution in the $z = 0$ plane transforms into a distribution on the hemisphere of radius r that becomes *uniform* in the $r \to \infty$ limit. To learn the bijective transformation, we estimate the normalized field in the augmented space. For sampling, we devise a backward ODE that is anchored by the physically meaningful additional dimension: the samples hit the (unaugmented) data manifold when the z reaches zero. Experimentally, PFGM achieves current state-of-the-art performance among the normalizing flow models on CIFAR-10, with an Inception score of 9.68 and a FID score of 2.48. It also performs on par with the state-of-the-art SDE approaches while offering $10\times$ to $20\times$ acceleration on image generation tasks. Additionally, PFGM appears more tolerant of estimation errors on a weaker network architecture and robust to the step size in the Euler method. The code is available at https://github.com/Newbeeer/poisson_flow.

1 Introduction

Deep generative models are a prominent approach for data generation, and have been used to produce high quality samples in image [1], text [2] and audio [35], as well as improve semi-supervised learning [20], domain generalization [25] and imitation learning [15]. However, current deep generative models also have limitations, such as unstable training objectives (GANs [1, 12, 17]) and low sample quality (VAEs [21], normalizing flows [6]). New techniques [12, 24] are introduced to stablize the training of CNN-based or ViT-based GAN models. Although recent advances on diffusion [16] and scored-based models [33] achieve comparable sample quality to GAN's without adversarial training, these models have a slow stochastic sampling process. [33] proposes backward ODE samplers (normalizing flow) that speed up the sampling process but these methods have not yet performed on par with the SDE counterparts.

We present a new "Poisson flow" generative model (**PFGM**), exploiting a remarkable physics fact that generalizes to N dimensions. As illustrated in Fig. 1(a), motion in a viscous fluid transforms any planar charge distribution into a uniform angular distribution. Specifically, we interpret N-dimensional data points $\mathbf{x}$ (images, say) as positive electric charges in the $z = 0$ plane of an $N + 1$-dimensional space (see Fig. 1(a)) filled with a viscous liquid (say honey). A positive charge with $z > 0$ will be repelled by the other charges and move in the direction of their repulsive force, eventually crossing an imaginary hemisphere of radius r. We show that, remarkably, if the the original charge distribution is let loose just above $z = 0$, this law of motion will cause a *uniform* distribution for their hemisphere crossings in the $r \to \infty$ limit.

*Equal Contribution.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

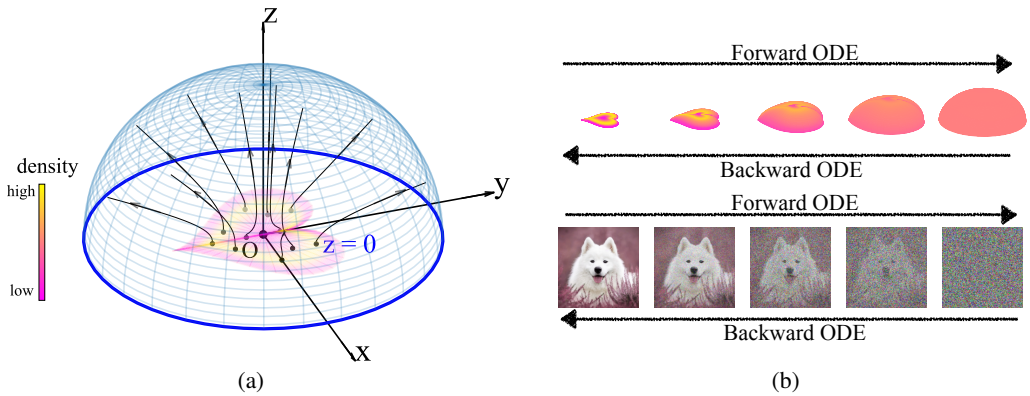


Figure 1: **(a)** 3D Poisson field trajectories for a heart-shaped distribution **(b)** The evolvements of a distribution (**top**) or an (augmented) sample (**bottom**) by the forward/backward ODEs pertained to the Poisson field.

Our Poisson flow generative process reverses the forward process: we generate a uniform distribution of negative charges on the hemisphere, then track their motion back to the $z = 0$ plane, where they will be distributed as the data distribution. A Poisson flow can be viewed as a type of continuous normalizing flows [4, 10, 33] in the sense that it continuously maps between an arbitrary distribution and an easily sampled one: in the previous works an N-dimensional Gaussian and in PFGM a uniform distribution on an N-dimensional hemisphere. In practice, we implement the Poisson flow by solving a pair of forward/backward ordinary differential equations (ODEs) induced by the electric field (Fig. 1(b)) given by the N-dimensional version of Coulomb's law (the gradient of the solution to the Poisson's equation with the data as sources). We will interchangeably refer to this gradient as the *Poisson field*, since electric fields normally refer to the special case $N = 3$.

The proposed generative model PFGM has a stable training objective and empirically outperforms previously state-of-the-art continuous flow methods [30, 33]. As a different iterative method, PFGM offers two advantages compared to score-based methods [32, 33]. First, the ODE process of PFGM achieves faster sampling speeds than the SDE samplers in [33]. while retaining comparable performance. Second, our backward ODE exhibits better generation performance than the reverse-time ODEs of VE/VP/sub-VP SDEs [33], as well as greater stability on a weaker architecture NSCNv2 [32]. The rationale for robustness is that the time variables in these ODE baselines are strongly correlated with the sample norms during training time, resulting in a less error-tolerant inference. In contrast, the tie between the anchored variable and the sample norm in PFGM is much weaker.

Experimentally, we show that PFGM achieves current state-of-the-art performance on CIFAR-10 dataset in the normalizing flow family, with FID/Inception scores of $2.54/9.62$ (w/ DDPM++ [33]) and $2.48/9.68$ (w/ DDPM++ deep [33]). It performs competitively with current state-of-the-art SDE samplers [33] and provides $10\times$ to $20\times$ speed up across datasets. Notably, the backward ODE in PFGM is the *only* ODE-based sampler that can produce decent samples on its own on NCSNv2 [32], while other ODE baselines fail without corrections. In addition, PFGM demonstrates the robustness to the step size in the Euler method, with a varying number of function evaluations (NFE) ranging from 10 to 100. We further showcase the utility of the invertible forward/backward ODEs of the Poisson field on likelihood evaluation and image manipulations, and its scalability to higher resolution images on LSUN bedroom 256×256 dataset.

2 Background and Related works

Poisson equation Let $\mathbf{x} \in \mathbb{R}^N$ and $\rho(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}$ be a *source* function. We assume that the source function has a compact support, $\rho \in \mathcal{C}^0$ and $N \geq 3$. The Poisson equation is

$$\nabla^2 \varphi(\mathbf{x}) = -\rho(\mathbf{x}), \tag{1}$$

where $\varphi(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}$ is called the *potential function*, and $\nabla^2 \equiv \sum_{i=1}^N \frac{\partial^2}{\partial x_i^2}$ is the Laplacian operator. It is usually helpful to define the gradient field $\mathbf{E}(\mathbf{x}) = -\nabla\varphi(\mathbf{x})$ and rewrite the Poisson equation as

$\nabla \cdot \mathbf{E} = \rho$, known in physics as Gauss's law [11]. The Poisson equation is widely used in physics, giving rise to Newton's gravitational theory [9] and the electrostatic theory [11], when $\rho(\mathbf{x})$ is interpreted as mass density or electric charge density, respectively. $\mathbf{E}$ is the N-dimensional analog of the electric field. The Poisson equation Eq. (1) (with zero boundary condition at infinity) admits a unique simple integral solution [2]:

$$\varphi(\mathbf{x}) = \int G(\mathbf{x}, \mathbf{y})\rho(\mathbf{y})d\mathbf{y}, \quad G(\mathbf{x}, \mathbf{y}) = \frac{1}{(N-2)S_{N-1}(1)} \frac{1}{\|\mathbf{x}-\mathbf{y}\|^{N-2}}, \tag{2}$$

where $S_{N-1}(1)$ is a geometric constant representing the surface area of the unit $(N-1)$-sphere [3], and $G(\mathbf{x}, \mathbf{y})$ is the extension of Green's function in N-dimensional space (details in Appendix A.3). The negative gradient field of $\varphi(\mathbf{x})$, referred as *Poisson field* of the source ρ, is

$$\mathbf{E}(\mathbf{x}) = -\nabla\varphi(\mathbf{x}) = -\int \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y})\rho(\mathbf{y})d\mathbf{y}, \quad \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) = -\frac{1}{S_{N-1}(1)} \frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|^{N}}. \tag{3}$$

Qualitatively, the Poisson field $\mathbf{E}(\mathbf{x})$ points away from sources, or equivalently $-\mathbf{E}(\mathbf{x})$ points towards sources, as illustrated in Fig. 1. It is straightforward to check that when $\rho(\mathbf{x}) \to \delta(\mathbf{x}-\mathbf{y})$, we get $\varphi(\mathbf{x}) \to G(\mathbf{x}, \mathbf{y})$ and $\mathbf{E}(\mathbf{x}) \to -\nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y})$. This implies that $G(\mathbf{x}, \mathbf{y})$ and $-\nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y})$ can be interpreted as the potential function and the gradient field generated by a unit point source, *e.g.*, a point charge, located at $\mathbf{y}$. When $\rho(\mathbf{x})$ takes general forms but has bounded support, simple asymptotics exist for $\|\mathbf{x}\| \gg \|\mathbf{y}\|$. To the lowest order, $\mathbf{E}(\mathbf{x}) = \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y})|_{\mathbf{y}=\mathbf{0}} \sim \mathbf{x}/\|\mathbf{x}\|^N$ behaves as if it were generated by a unit point source at $\mathbf{y} = 0$. In physics, the power law decay is considered to be long-range (compared to exponential decay) [11].

Particle dynamics in a Poisson field The Poisson field immediately defines a flow model, where the probability distribution evolves according to the gradient flow $\partial p_t(\mathbf{x})/\partial t = -\nabla \cdot (p_t(\mathbf{x})\mathbf{E}(\mathbf{x}))$. The gradient flow is a special case of the Fokker-Planck equation [28], where the diffusion coefficient is zero. Intuitively we can think of $p_t(\mathbf{x})$ as represented by a population of particles. The corresponding (non-diffusion) case of the Itô process is the forward ODE $\frac{d\mathbf{x}}{dt} = \mathbf{E}(\mathbf{x})$. We can interpret the trajectories of the ODE as particles moving according to the Poisson field $\mathbf{E}(x)$, with initial states drawn from p_0. The physical picture of the forward ODE is a charged particle under the influence of electric fields in the overdamped limit (details in Appendix F).

The dynamics is also *rescalable* in the sense that the particle trajectory remains the same for $\frac{d\mathbf{x}}{dt} = \pm f(\mathbf{x})\mathbf{E}(\mathbf{x})$ for $f(\mathbf{x}) > 0, f(\mathbf{x}) \in \mathcal{C}^1$, because the time rescaling $dt \to f(\mathbf{x}(t))dt$ recovers $\frac{d\mathbf{x}}{dt} = \pm\mathbf{E}(\mathbf{x})$. Note that the dynamics is stiff due to the power law factor in the denominator in Eq. (3), posing computational challenges. Luckily the rescalablility allows us to rescale $\mathbf{E}(\mathbf{x})$ properly to get new ODEs (formally defined later in Section 3.3) that are more amenable for sampling.

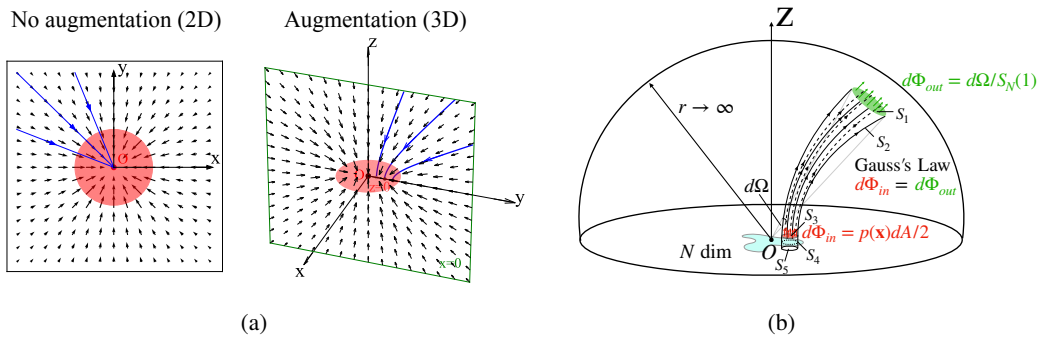
Generative Modeling via ODE Generative modeling can be done by transforming a base distribution to a data distribution via mappings defined by ODEs. The ODE-based samplers allow for adaptive sampling, exact likelihood evaluation and modeling of continuous-time dynamics [4, 33]. Previous works broadly fall into two lines. [4, 3] introduce a continuous-time normalizing flow model that can be trained with maximum likelihood by the instantaneous change-of-variables formula [4]. For sampling, they directly integrate the learned invertiable mapping over time. Another work [33] unifies the scored-based model [31, 32] and diffusion model [16] into a general diffusion process, and uses the reverse-time ODE of the diffusion process for sampling. They show that the reverse-time ODE produces high quality samples with improved architecture.

3 Poisson Flow Generative Models

In this section, we start with the properties of the Poisson flow in the augmented space and show how to draw samples from the data distribution by following the backward ODE of the Poisson flow (Section 3.1). We then discuss how to actually learn a normalized Poisson field from data samples through simulations of the forward ODE (Section 3.2) and present an equivalent backward ODE that allows for exponentially decay on z (Section 3.3).

[2] Eq. (2) is valid for $N \geq 3$. When $N = 2$, the Green's function is $G(\mathbf{x}, \mathbf{y}) = -\log(\|\mathbf{x}-\mathbf{y}\|)/2\pi$. We assume $N \geq 3$ since N is typically large in the relevant applications.

[3] The N-sphere with radius r is defined as $\{\mathbf{x} \in \mathbb{R}^{N+1}, \|\mathbf{x}\| = r\}$



(a) (b)

Figure 2: **(a)** Poisson field (black arrows) and particle trajectories (blue lines) of a 2D uniform disk (red). **Left** (no augmentation, 2D): all particles collapse to the disk center. **Right** (augmentation, 3D): particles hit different points on the disk. **(b)** Proof idea of Theorem 1. By Gauss's Law, the outflow flux $d\Phi_{out}$ equals the inflow flux $d\Phi_{in}$. The factor of two in $p(\mathbf{x})dA/2$ is due to the symmetry of Poisson fields in $z < 0$ and $z > 0$.

3.1 Augment the data with additional dimension

We wish to generate samples $\mathbf{x} \in \mathbb{R}^N$ from a distribution $p(\mathbf{x})$ supported on a bounded region. We may set the source $\rho(\mathbf{x}) = p(\mathbf{x}) \in \mathcal{C}^0$ [4] and compute the resulting gradient field $\mathbf{E}(\mathbf{x})$ from Eq. (3). Since $-\mathbf{E}(\mathbf{x})$ points towards sources, the backward ODE $d\mathbf{x}/dt = -\mathbf{E}(\mathbf{x})$ will take samples close to the sources. One may naively hope that the backward ODE is a generative model that recovers $p(\mathbf{x})$. Unfortunately, the backward ODE has the problem of mode collapse. We illustrate this phenomenon with a 2D uniform disk. The reverse Poisson field $-\mathbf{E}(\mathbf{x})$ on the 2D (x, y)-plane points towards the center of the disk O (Fig. 2(a) left), so all particle trajectories (blue lines) will eventually hit O. If we instead add an additional dimension z (Fig. 2(a) right), particles can hit different points on the disk and faithfully recover the data distribution.

Consequently, instead of solving the Poisson equation $\nabla^2\varphi(\mathbf{x}) = -p(\mathbf{x})$ in the original data space, we solve the Poisson equation in an augmented space $\tilde{\mathbf{x}} = (\mathbf{x}, z) \in \mathbb{R}^{N+1}$ with an additional variable $z \in \mathbb{R}$. We augment the training data $\tilde{\mathbf{x}}$ in the new space by setting $z = 0$ such that $\tilde{\mathbf{x}} = (\mathbf{x}, 0)$. As a consequence, the data distribution in the augmented space is $\tilde{p}(\tilde{\mathbf{x}}) = p(\mathbf{x})\delta(z)$, where δ is the Dirac delta function. By Eq. (3), the Poisson field by solving the new Poisson equation $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$ has an analytical form:

$$\forall \tilde{\mathbf{x}} \in \mathbb{R}^{N+1}, \mathbf{E}(\tilde{\mathbf{x}}) = -\nabla\varphi(\tilde{\mathbf{x}}) = \frac{1}{S_N(1)} \int \frac{\tilde{\mathbf{x}} - \tilde{\mathbf{y}}}{\|\tilde{\mathbf{x}} - \tilde{\mathbf{y}}\|^{N+1}} \tilde{p}(\tilde{\mathbf{y}}) d\tilde{\mathbf{y}} \tag{4}$$

The associated forward/backward ODEs of the Poisson field are $d\tilde{\mathbf{x}}/dt = \mathbf{E}(\tilde{\mathbf{x}}), d\tilde{\mathbf{x}}/dt = -\mathbf{E}(\tilde{\mathbf{x}})$. Intuitively, theses ODEs uniquely define trajectories of particles between the $z = 0$ hyperplane and an enclosing hemisphere (cf. Fig. 1(a)). In the following theorem, we show that the backward ODE defines a transformation between the uniform distribution on an infinite hemisphere and the data distribution $\tilde{p}(\tilde{\mathbf{x}})$ in the $z = 0$ plane. We present the formal proof to Appendix A, illustrated by Fig. 2(b). The proof is based on the idea that when the radius of hemisphere $r \to \infty$, the data distribution $\tilde{p}(\tilde{\mathbf{x}})$ can be effectively viewed as a delta distribution at origin. Consequently, the Poisson field points in the radial direction at $r \to \infty$, perpendicular to $S_N^+(r)$ (Green arrows in Fig. 2(b)).

Theorem 1. *Suppose particles are sampled from a uniform distribution on the upper ($z > 0$) half of the sphere of radius r and evolved by the backward ODE $\frac{d\tilde{\mathbf{x}}}{dt} = -\mathbf{E}(\tilde{\mathbf{x}})$ until they reach the $z = 0$ hyperplane, where the Poisson field $\mathbf{E}(\tilde{\mathbf{x}})$ is generated by the source $\tilde{p}(\tilde{\mathbf{x}})$. In the $r \to \infty$ limit, under some mild conditions detailed in Appendix A, this process generates a particle distribution $\tilde{p}(\tilde{\mathbf{x}})$, i.e., a distribution $p(\mathbf{x})$ in the $z = 0$ hyperplane.*

Proof sketch. Suppose the flux of the backward ODE connects a solid angle $d\Omega$ (on $S_N^+(r)$) with an area dA (on $\text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$). According to Gauss's law, the outflow flux $d\Phi_{out} = d\Omega/S_N(1)$ on the

[4] A probability distribution $p(\mathbf{x})$ is a special case of "charge density" $\rho(x)$ because $p(\mathbf{x})$ need to be non-negative and integrates to unity. Here we focus on applications to probability distribution of data, which is the objective to be modeled in generative modeling.

hemisphere (Green arrows in Fig. 2(b)) equals the inflow flux $d\Phi_{in} = p(\mathbf{x})dA/2$ on $\text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$ (Red arrows in Fig. 2(b)). $d\Phi_{in} = d\Phi_{out}$ gives $d\Omega/dA = p(\mathbf{x})S_N(1)/2 \propto p(\mathbf{x})$. Together, by change-of-variable, we conclude that the final distribution in the $z = 0$ hyperplane is $p(\mathbf{x})$. □

The theorem states that starting from an infinite hemisphere, one can recover the data distribution $\tilde{p}$ by following the inverse Poisson field $-\mathbf{E}(\tilde{\mathbf{x}})$. We defer the formal proof and technical assumptions of the theorem to Appendix A. The property allows generative modeling by following the Poisson flow of $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$.

3.2 Learning the normalized Poisson Field

Given a set of training data $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ i.i.d sampled from the data distribution $p(\mathbf{x})$, we define the empirical version of the Poisson field (Eq. (4)) as follows:

$$\hat{\mathbf{E}}(\tilde{\mathbf{x}}) = c(\tilde{\mathbf{x}}) \sum_{i=1}^{n} \frac{\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i}{||\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i||^{N+1}}$$

where the gradient field is calculated on n *augmented* datapoints $\{\tilde{\mathbf{x}}_i = (\mathbf{x}_i, 0)\}_{i=1}^n$, and $c(\tilde{\mathbf{x}}) = 1/\sum_{i=1}^n \frac{1}{||\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i||^{N+1}}$ is the multiplier for numerical stability. We further normalize the field to resolve the variations in the magnitude of the norm $\parallel \hat{\mathbf{E}}(\tilde{\mathbf{x}}) \parallel_2$, and fit the neural network to the more amenable normalized field $\mathbf{v}(\tilde{\mathbf{x}}) = \sqrt{N}\hat{\mathbf{E}}(\tilde{\mathbf{x}})/\parallel \hat{\mathbf{E}}(\tilde{\mathbf{x}}) \parallel_2$. The Poisson field is rescalable (cf. Section 2) and thus trajectories of its forward/backward ODEs are invariant under normalization. We denote the empirical field calculated on batch data $\mathcal{B}$ by $\hat{\mathbf{E}}_{\mathcal{B}}$ and the normalized field as $\mathbf{v}_{\mathcal{B}}(\tilde{\mathbf{x}}) = \sqrt{N}\hat{\mathbf{E}}_{\mathcal{B}}(\tilde{\mathbf{x}})/\parallel \hat{\mathbf{E}}_{\mathcal{B}}(\tilde{\mathbf{x}}) \parallel_2$.

Similar to the scored-based models, we sample points inside the hemisphere by perturbing the augmented training data. Given a training point $\mathbf{x} \in \mathcal{D}$, we add noise to its augmented version $\{\tilde{\mathbf{x}}_i = (\mathbf{x}_i, 0)\}_{i=1}^n$ to construct the perturbed point $(\mathbf{y}, z)$:

$$\mathbf{y} = \mathbf{x} + \parallel \epsilon_{\mathbf{x}} \parallel (1+\tau)^m \mathbf{u}, \quad z = |\epsilon_z|(1+\tau)^m \tag{5}$$

where $\epsilon = (\epsilon_{\mathbf{x}}, \epsilon_z) \sim \mathcal{N}(0, \sigma^2 I_{N+1\times N+1})$, $\mathbf{u} \sim \mathcal{U}(S_N(1))$ and $m \sim \mathcal{U}[0, M]$. The upper limit M, standard deviation σ and τ are hyper-parameters. With fixed ϵ and $\mathbf{u}$, the added noise increases exponentially with m. The rationale behind the design is that points farther away from the data support play a less important role in generative modeling, sharing a similar spirit with the choice of noisy scales in score-based models [32, 33].

In practice, we sample the points by perturbing a mini-batch data $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{B}|}$ in each iteration. We uniformly sample the power m in $[0, M]$ for each datapoint. We select a large M (typically around 300) to ensure the perturbed points can reach a large enough hemisphere. We use a larger batch $\mathcal{B}_L$ for the estimation of normalized field since the empirical normalized field is biased, , which empirically gives better results. Denoting the set of perturbed points as $\{\tilde{\mathbf{y}}_i\}_{i=1}^{|\mathcal{B}|}$, we train the neural network f_θ on these points to estimate the normalized field by minimizing the following loss:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \parallel f_\theta(\tilde{\mathbf{y}}_i) - \mathbf{v}_{\mathcal{B}_L}(\tilde{\mathbf{y}}_i) \parallel_2^2$$

We summarize the training process in Algorithm 1. In practice, we add a small constant γ to the denominator of the normalized field to overcome the numerical issue when $\exists i, ||\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i|| \approx 0$.

3.3 Backward ODE anchored by the additional dimension

After estimating the normalized field $\mathbf{v}$, we can sample from the data distribution by the backward ODE $d\tilde{\mathbf{x}} = -\mathbf{v}(\tilde{\mathbf{x}})dt$. Nevertheless, the boundary condition of the above ODE is unclear: the starting and terminal time t of the ODE are both unknown. To remedy the issue, we propose an equivalent backward ODE in which $\mathbf{x}$ evolves with the augmented variable z:

$$d(\mathbf{x}, z) = -(\frac{d\mathbf{x}}{dt}\frac{dt}{dz}dz, dz) = -(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}, 1)dz$$

Algorithm 1 Learning the normalized Poisson Field

Input: Training iteration T, Initial model f_θ, dataset $\mathcal{D}$, constant γ, learning rate η.
for $t = 1 \dots T$ **do**
  Sample a large batch $\mathcal{B}_L$ from $\mathcal{D}$ and subsample a batch of datapoints $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^{|\mathcal{B}|}$ from $\mathcal{B}_L$
  Simulate the ODE: $\{\tilde{\mathbf{y}}_i = \text{perturb}(\mathbf{x}_i)\}_{i=1}^{|\mathcal{B}|}$
  Calculate the normalized field by $\mathcal{B}_L$: $\mathbf{v}_{\mathcal{B}_L}(\tilde{\mathbf{y}}_i) = \sqrt{N}\hat{\mathbf{E}}_{\mathcal{B}_L}(\tilde{\mathbf{y}}_i)/(\| \hat{\mathbf{E}}_{\mathcal{B}_L}(\tilde{\mathbf{y}}_i) \|_2 + \gamma), \forall i$
  Calculate the loss: $\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|} \| f_\theta(\tilde{\mathbf{y}}_i) - \mathbf{v}_{\mathcal{B}_L}(\tilde{\mathbf{y}}_i) \|_2^2$
  Update the model parameter: $\theta = \theta - \eta\nabla\mathcal{L}(\theta)$
end for
return f_θ

Algorithm 2 perturb($\mathbf{x}$)

Sample the power $m \sim \mathcal{U}[0, M]$
Sample the initial noise $(\epsilon_{\mathbf{x}}, \epsilon_z) \sim \mathcal{N}(0, \sigma^2 I_{(N+1)\times(N+1)})$
Uniformly sample the vector from the unit ball $\mathbf{u} \sim \mathcal{U}(S_N(1))$
Construct training point $\mathbf{y} = \mathbf{x} + \| \epsilon_{\mathbf{x}} \| (1+\tau)^m \mathbf{u}$, $z = |\epsilon_z|(1+\tau)^m$
return $\tilde{\mathbf{y}} = (\mathbf{y}, z)$

where $\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}, \mathbf{v}(\tilde{\mathbf{x}})_z$ are the corresponding components of $\mathbf{x}, z$ in vector $\mathbf{v}(\tilde{\mathbf{x}})$. In the new ODE, we replace the time variable t with the physically meaningful variable z, permitting explicit starting and terminal conditions: when $z = 0$, we arrive at the data distribution and we can freely choose a large z_{max} as the starting point in the backward ODE. The backward ODE is compatible with general-purpose ODE solvers, *e.g.*, RK45 method [23] and forward Euler method. The popular black-box ODE solvers, such as the one in Scipy library [37], typically use a common starting time for the same batch of samples. Since the distribution on the $z = z_{\text{max}}$ hyperplane is no longer uniform, we derive the prior distribution by radially projecting uniform distribution on the hemisphere with radius $r = z_{\text{max}}$ to the $z = z_{\text{max}}$ hyperplane:

$$p_{\text{prior}}(\mathbf{x}) = \frac{2z_{\text{max}}^{N+1}}{S_N(z_{\text{max}})(\| \mathbf{x} \|_2^2 + z_{\text{max}}^2)^{\frac{N+1}{2}}} = \frac{2z_{\text{max}}}{S_N(1)(\| \mathbf{x} \|_2^2 + z_{\text{max}}^2)^{\frac{N+1}{2}}}$$

where $S_N(r)$ is the surface area of N-sphere with radius r. The reason behind the radial projection is that the Poisson field points in the radial direction at $r \to \infty$. The new backward ODE also defines a bijective transformation between $p_{\text{prior}}(\mathbf{x})$ on the infinite hyperplane ($z_{\text{max}} \to \infty$) and the data distribution $\tilde{p}(\tilde{\mathbf{x}})$, analogous to Theorem 1. In order to sample from $p_{\text{prior}}(\mathbf{x})$, it is suffice to sample the norm from the distribution: $p_{\text{prior}}(\| \mathbf{x} \|_2) \propto \| \mathbf{x} \|_2^{N-1}/(\| \mathbf{x} \|_2^2 + z_{\text{max}}^2)^{\frac{N+1}{2}}$ and then uniformly sample its angle. Detailed derivations are deferred to Appendix A.4. We further achieve exponentially decay on the z dimension by introducing a new variable t':

$$[\text{Backward ODE}] \quad d(\mathbf{x}, z) = -(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}z, z)dt' \tag{6}$$

The z component in the backward ODE, *i.e.*, $dz = -zdt'$, can be solved by $z = e^{-t'}$. Since z reaches zero as $t' \to \infty$, we instead choose a tiny positive number z_{min} as the terminal condition. The corresponding starting/terminal time of the variable t' are $-\log z_{\text{max}}/-\log z_{\text{min}}$ respectively. Empirically, this simple change of variable leads to 2× faster sampling with almost no harm to the sample quality. We defer more details to Appendix B.2.

4 Generative Modeling via the Backward ODE

In this section, we demonstrate the effectiveness of the backward ODE associated with PFGM on image generation tasks. In Section 4.1, we show that PFGM achieves currently best in class performance in the normalizing flow family. In comparison to the existing state-of-the-art SDE or MCMC approaches, PFGM exhibits 10× or 20× acceleration while maintaining competitive or higher generation quality. Meanwhile, unlike existing ODE baselines that heavily rely on corrector to generate decent samples on weaker architectures, PFGM exhibits greater stability against error (Section 4.2). Finally, we show that PFGM is robust to the step size in the Euler method (Section 4.3),

Table 1: CIFAR-10 sample quality (FID, Inception) and number of function evaluation (NFE).

	Invertible?	Inception ↑	FID ↓	NFE ↓
PixelCNN [36]	✗	4.60	65.9	1024
IGEBM [8]	✗	6.02	40.6	60
ViTGAN [24]	✗	9.30	6.66	1
StyleGAN2-ADA [17]	✗	9.83	2.92	1
StyleGAN2-ADA (cond.) [17]	✗	10.14	2.42	1
NCSN [31]	✗	8.87	25.32	1001
NCSNv2 [32]	✗	8.40	10.87	1161
DDPM [16]	✗	9.46	3.17	1000
NCSN++ VE-SDE [33]	✗	9.83	2.38	2000
NCSN++ deep VE-SDE [33]	✗	9.89	2.20	2000
Glow [19]	✓	3.92	48.9	1
DDIM, T=50 [30]	✓	-	4.67	50
DDIM, T=100 [30]	✓	-	4.16	100
NCSN++ VE-ODE [33]	✓	9.34	5.29	194
NCSN++ deep VE-ODE [33]	✓	9.17	7.66	194
DDPM++ backbone				
VP-SDE [33]	✗	9.58	2.55	1000
sub-VP-SDE [33]	✗	9.56	2.61	1000
VP-ODE [33]	✓	9.46	2.97	134
sub-VP-ODE [33]	✓	9.30	3.16	146
PFGM (ours)	✓	**9.62**	**2.54**	**110**
DDPM++ deep backbone				
VP-SDE [33]	✗	9.68	2.41	1000
sub-VP-SDE [33]	✗	9.57	2.41	1000
VP-ODE [33]	✓	9.47	2.86	134
sub-VP-ODE [33]	✓	9.40	3.05	146
PFGM (ours)	✓	**9.68**	**2.48**	**110**

and its associated ODE allows for likelihood evaluation and image manipulation by editing the latent space (Section 4.4).

4.1 Efficient image generation by PFGM

Setup For image generation tasks, we consider the CIFAR-10 [22], CelebA 64×64 [38] and LSUN bedroom 256×256 [39]. Following [32], we first center-crop the CelebA images and then resize them to 64×64. We choose $M = 291$ (CIFAR-10 and CelebA)/356 (LSUN bedroom), $\sigma = 0.01$ and $\tau = 0.03$ for the perturbation Algorithm 2, and $z_{\text{min}} = 1e-3$, $z_{\text{max}} = 40$ (CIFAR-10 and CelebA)/100 (LSUN bedroom) for the backward ODE. We adopt the DDPM++ and DDPM++ deep architectures [33] as our backbones. We add the scalar z (resp. predicted direction on z) as input (resp. output) to accommodate the additional dimension. We take the same set of hyper-parameters, such as batch size, learning rate and training iterations from [33]. We provide more training details in Appendix B.1.

Baselines We compare PFGM to modern autoregressive model [36], GAN [17, 24], normalizing flow [19] and EBM [8]. We also compare with variants of score-based models such as DDIM [30] and current state-of-the-art SDE/ODE methods [33]. We denote the methods that use forward-time SDEs in [33] such as Variance Exploding (**VE**) SDE/Variance Preserving (**VP**) SDE/ sub-Variance Preserving (**sub-VP**), and the corresponding backward SDE/ODE, as **A-B**, where A $\in$ {VE, VP, sub-VP} and B $\in$ {SDE, ODE}. We follow the model selection protocol in [33], which selects the checkpoint with the smallest FID score over the course of training every 50k iterations.

Numerical Solvers The backward ODE (Eq. (6)) is compatible with any general purpose ODE solver. In our experiments, the default solver of ODEs is the black box solver in the Scipy library [37] with the RK45 [7] method (**RK45**), unless otherwise specified. For VE/VP/subVP-SDEs, we

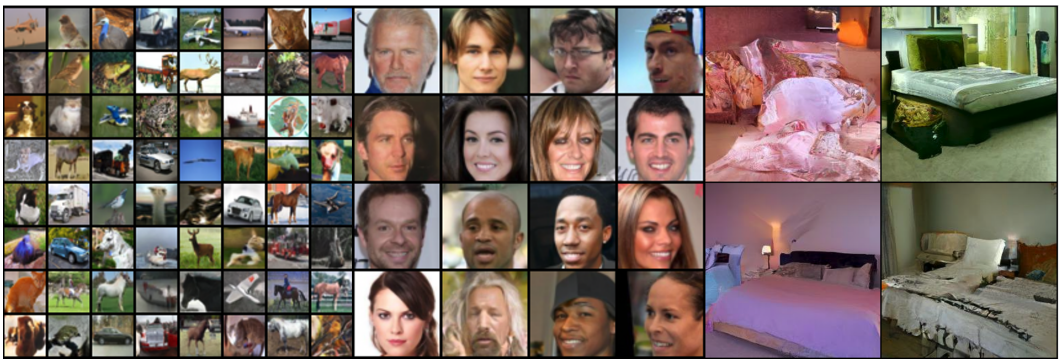


Figure 3: Uncurated samples on datasets of increasing resolution. From left to right: CIFAR-10 32×32, CelebA 64×64 and LSUN bedroom 256×256.

use the predictor-corrector (**PC**) sampler introduced in [33]. For VP/sub-VP-SDEs, we apply the predictor-only sampler, because its performance is on par with the PC sampler while requiring half computation.

Results For quantitative evaluation on CIFAR-10, we report the Inception [29] (higher is better) and FID [13] scores (lower is better) in Table 1. We also include our preliminary experimental results on a weaker architecture NCSNv2 [32] in Appendix D.2. We measure the inference speed by the average NFE (number of function evaluation). We also explicitly indicate which methods belong to the invertible flow family.

Our main findings are: **(1) PFGM achieves the best Inception scores and FID scores among the normalizing flow models.** Specifically, PFGM obtains an Inception score of 9.68 and a FID score of 2.48 using the DDPM++ deep architecture. To our best knowledge, these are the highest FID and Inception scores by flow models on CIFAR-10. **(2) PFGM achieves a** $10\times \sim 20\times$ **faster inference speed than the SDE methods using the similar architectures, while retaining comparable sample quality.** As shown in Table 1, PFGM requires NFEs of 110 whereas the SDE methods typically use $1000 \sim 2000$ inference steps. PFGM outperforms all the baselines on DDPM++ in all metrics. In addition, PFGM generally samples faster than other ODE baselines with the same RK45 solver. **(3) The backward ODE in PFGM is compatible with architectures with varying capacities.** PFGM consistently outperforms other ODE baselines on DDPM++ (Table 1) or NCSNv2 (Appendix D.2) backbones. **(4) PFGM shows scalability to higher resolution datasets.** In Appendix D.1, we show that PFGM are capable of scale-up to LSUN bedroom 256×256. In particular, PFGM has comparable performance with VE-SDE with $15\times$ fewer NFE.

In Fig. 3, we visualize the uncurated samples from PFGM on CIFAR-10, CelebA 64×64 and LSUN bedroom 256×256. We provides more samples in Appendix E.

4.2 Failure of VE/VP-ODEs on NCSNv2 architecture

In our preliminary experiments on NCSNv2 architectures, we empirically observe that the VE/VP-ODEs have FID scores greater than 90 on CIFAR-10. In particular, VE/VP-ODEs can only generate decent samples when applying the Langevin dynamics corrector, and even then, their performances are still inferior to PFGM (Table 8, Table 9). The poor performance on NCSNv2 stands in striking contrast to their high sample quality on NCSN++/DDPM++ in [33]. **It indicates that the VE/VP-ODEs are more susceptible to estimation errors than PFGM.** We hypothesize that the strong norm-σ correlation seen during the training of score-based models causes the problem.

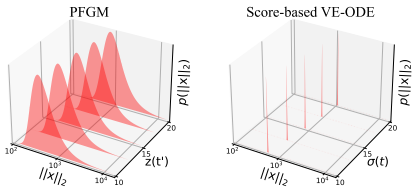
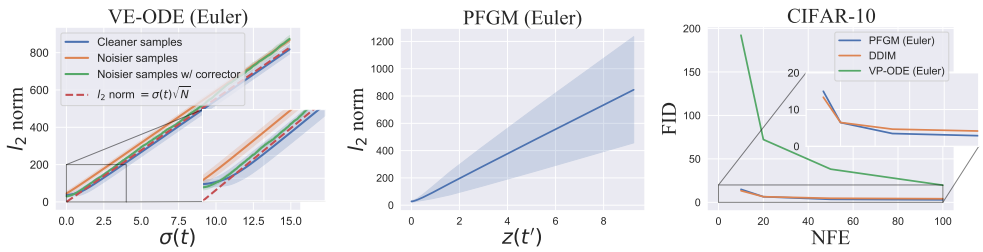


Figure 4: Sample norm distributions with varying time variables ($\sigma(t)$ for VE-ODE and $z(t')$ for PFGM)

For score-based models, the l_2 norms of perturbed training samples and the standard deviations $\sigma(t)$ of Gaussian noises have strong correlation, *e.g.*, l_2 norm



(a) Norm-$\sigma(t)$ in VE-ODE (b) Norm-$z(t')$ in PFGM (c) FID vs. NFE on CIFAR-10

Figure 5: **(a)** Norm-$\sigma(t)$ relation during the backward sampling of VE-ODE (Euler). **(b)** Norm-$z(t')$ relation during the backward sampling of PFGM (Euler). The shaded areas mean the standard deviation of norms. **(c)** Number of steps versus FID score.

$\approx \sigma(t)\sqrt{N}$ for large $\sigma(t)$ in VE [33]. In contrast, as shown in Fig. 4, PFGM allocates high mass across a wide spectrum of the training sample norms. During sampling, VE/VP-ODEs could break down when the trajectories of backward ODEs deviate from the norm-$\sigma(t)$ relation to which most training samples pertain. The weaker NCSNv2 backbone incurs larger errors and thus leads to their failure. The PFGM is more resistant to estimate errors because of the greater range of training sample norms.

To further verify the hypothesis above, we split a batch of VE-ODE samples into cleaner and noisier samples according to visual quality (Fig. 9(a)). In Fig. 5(a), we investigate the relation for cleaner and noisier samples during the forward Euler simulation of VE-ODE when $\sigma(t) < 15$. We can see that the trajectory of cleaner samples stays close to the norm-$\sigma(t)$ relation (the red dash line), whereas that of the noisier samples diverges from the relation. The Langevin dynamics corrector changes the trajectory of noisier samples to align with the relation. Fig. 5(b) further shows that the anchored variable $z(t')$ and the norms in the backward ODE of PFGM are not strongly correlated, giving rise to the robustness against the imprecise estimation on NCSNv2. We defer more details to Appendix C.

4.3 Effects of step size in the forward Euler method

In order to accelerate the inference speed of ODEs, we can increase the step size (decrease the NFEs) in numerical solvers such as the forward Euler method. It also enables the trade-off between sample quality and computational efficiency in real-world deployment. We study the effects of increasing step size on PFGM, VP-ODE and DDIM [30] using the forward Euler method, with a varying NFE ranging from 10 to 100.

In Fig. 5(c), we report the sample quality measured by FID scores on CIFAR-10. As expected, all the methods have higher FID scores when decreasing the NFE. We observe that the sample quality of PFGM degrades gracefully as we decrease the NFE. Our method shows significantly better robustness to step sizes than the VP-ODE, especially when only taking a few Euler steps. In addition, PFGM obtains better FID scores than DDIM on most NFEs except for 10 where PFGM is marginally worse. This suggests that the PFGM is a promising method for accommodating instantaneous resource availability, as high-quality samples can be generated in limited steps.

4.4 Utilities of ODE: likelihood evaluation and latent representation

Similar to the family of discrete normalizing flows [6, 19, 14] and continuous probability flow [33], the forward ODE in PFGM defines an invertible mapping between the data space and latent space with a known prior. Formally, we define the invertible forward $\mathcal{M}$ mapping by integrating the corresponding forward ODE $d(\mathbf{x}, z) = (\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}z, z)dt'$ of Eq. (6):

$$\mathbf{x}(\log z_{\max}) = \mathcal{M}(\mathbf{x}(\log z_{\min})) \equiv \mathbf{x}(\log z_{\min}) + \int_{\log z_{\min}}^{\log z_{\max}} \mathbf{v}(\mathbf{x}(t'))_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}}(t'))_z^{-1}e^{t'}\,dt'$$

where $\log z_{\min}$/$\log z_{\max}$ are the starting/terminal time in the forward ODE. The forward mapping transfers the data distribution to the prior distribution p_{prior} on the $z = z_{\max}$ hyperplane (cf. Section 3.3): $p_{\text{prior}}(\mathbf{x}(\log z_{\max})) = \mathcal{M}(p(\mathbf{x}(\log z_{\min})))$. The invertibility enables likelihood evaluation and creates

a meaningful latent space on the $z = z_{\max}$ hyperplane. In addition, we can adapt to the computational constraints by adjusting the step size or the precision in numerical ODE solvers.

Likelihood evaluation We evaluate the data likelihood by the instantaneous change-of-variable formula [4, 33]. In Table 6, we report the bits/dim on the uniformly dequantized CIFAR-10 test set and compare with existing baselines that use the same setup. We observe that PFGM achieves better likelihoods than discrete normalizing flow models, even without maximum likelihood training. Among the continuous flow models, sub-VP-ODE shows the lowest bits/dim, although its sample quality is worse than VP-ODE and PFGM (Table 1). The exploration of the seeming trade-off between likelihood and sample quality is left for future works.

Figure 6: Bits/dim on CIFAR-10

	bits/dim ↓
RealNVP [6]	3.49
Glow [19]	3.35
Residual Flow [3]	3.28
Flow++ [14]	3.29
DDPM (L) [16]	$\leq 3.70^*$
DDPM++ backbone	
VP-ODE [33]	3.20
sub-VP-ODE [33]	**3.02**
PFGM (ours)	3.19

Latent representation Since the samples are uniquely identifiable by their latents via the invertible mapping $\mathcal{M}$, PFGM further supports image manipulation using its latent representation on the $z = z_{\max}$ hyperplane. We include the results of image interpolation and the temperature scaling [6, 19, 33] to Appendix D.4 and Appendix D.5. For interpolation, it shows that we can travel along the latent space to obtain perceptually consistent interpolations between CelebA images.

5 Conclusion

We present a new deep generative model by solving the Poisson equation whose source term is the data distribution. We estimate the normalized gradient field of the solution in an augmented space with an additional dimension. For sampling, we devise a backward ODE that exponential decays on the physically meaningful additional dimension. Empirically, our approach has currently best performance over other normalizing flow baselines, and achieving $10\times$ to $20\times$ acceleration over the stochastic methods. Our backward ODE shows greater stability against errors than popular ODE-based methods, and enables efficient adaptive sampling. We further demonstrate the utilities of the forward ODE on likelihood evaluation and image interpolation. Future directions include improving the normalization of Poisson fields. More principled approaches can be used to get around the divergent near-field behavior. For example, we may exploit renormalization, a useful tool in physics, to make the Poisson field well-behaved in near fields.

Acknowledgements

We are grateful to Shangyuan Tong, Timur Garipov and Yang Song for helpful discussion. We would like to thank Octavian Ganea and Wengong Jin for reviewing an early draft of this paper. YX and TJ acknowledge support from MIT-DSTA Singapore collaboration, from NSF Expeditions grant (award 1918839) "Understanding the World Through Code", and from MIT-IBM Grand Challenge project. ZL and MT would like to thank the Center for Brains, Minds, and Machines (CBMM) for hospitality. ZL and MT are supported by The Casey and Family Foundation, the Foundational Questions Institute, the Rothberg Family Fund for Cognitive Science and IAIFI through NSF grant PHY-2019786.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *ArXiv*, abs/1809.11096, 2019.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[3] Ricky T. Q. Chen, Jens Behrmann, David Kristjanson Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *ArXiv*, abs/1906.02735, 2019.

[4] Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David Kristjanson Duvenaud. Neural ordinary differential equations. *ArXiv*, abs/1806.07366, 2018.

[5] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *ArXiv*, abs/1605.08803, 2017.

[7] J. R. Dormand and P. J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6:19–26, 1980.

[8] Yilun Du and Igor Mordatch. Implicit generation and generalization in energy-based models. *ArXiv*, abs/1903.08689, 2019.

[9] Herbert Goldstein, Charles Poole, and John Safko. Classical mechanics, 2002.

[10] Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Kristjanson Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *ArXiv*, abs/1810.01367, 2019.

[11] David J Griffiths. Introduction to electrodynamics, 2005.

[12] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. In *NIPS*, 2017.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[14] Jonathan Ho, Xi Chen, A. Srinivas, Yan Duan, and P. Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. *ArXiv*, abs/1902.00275, 2019.

[15] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *NIPS*, 2016.

[16] Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020.

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *ArXiv*, abs/2006.06676, 2020.

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[19] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018.

[20] Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. *ArXiv*, abs/1406.5298, 2014.

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[22] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009.

[23] F Shampine Lawrence. Some practical runge-kutta formulas. *Mathematics of Computation*, 46:135–150, 1986.

[24] Kwonjoon Lee, Huiwen Chang, Lu Jiang, Han Zhang, Zhuowen Tu, and Ce Liu. Vitgan: Training gans with vision transformers. *ArXiv*, abs/2107.04589, 2021.

[25] Daiqing Li, Junlin Yang, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Semantic segmentation with generative models: Semi-supervised learning and strong out-of-domain generalization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8296–8307, 2021.

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *ArXiv*, abs/1802.05957, 2018.

[27] Henry Ricardo. A modern introduction to differential equations. 2002.

[28] Hannes Risken. Fokker-planck equation. 1984.

[29] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *ArXiv*, abs/1606.03498, 2016.

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2021.

[31] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *ArXiv*, abs/1907.05600, 2019.

[32] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *ArXiv*, abs/2006.09011, 2020.

[33] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *ArXiv*, abs/2011.13456, 2021.

[34] Christian Szegedy, V. Vanhoucke, S. Ioffe, Jon Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016.

[35] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, 2016.

[36] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. Conditional image generation with pixelcnn decoders. In *NIPS*, 2016.

[37] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, Ilhan Polat, Yu Feng, Eric W. Moore, J. Vanderplas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Daniel Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Alessandro Pietro Alex Andreas Andreas Anthony Ant Vijaykumar Bardelli Rothberg Hilboll Kloeckner Sco, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andre Kloeckner, Anthony M. Scopatz, Antony Lee, Ariel S. Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald,

David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, Gert Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jorg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavi, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph E. Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Lee Tritz, Martin Dr Thoma, Matt Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per Andreas Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Taku Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis Raymond Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, U. Upadhyay, Yaroslav O. Halchenko, and Y. Vázquez-Baeza. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261 – 272, 2020.

[38] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial parts responses to face detection: A deep learning approach. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3676–3684, 2015.

[39] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *ArXiv*, abs/1506.03365, 2015.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Appendix G.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appendix H.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the abstract.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix B.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] All the experiments are run on a single NVIDIA A100 GPU.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] The assets are public/open-source datasets and codes.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

A Proofs

A.1 Formal Proof of Theorem 1

Before proceeding to Theorem 1, we show a technical lemma that guarantees the existence-uniqueness of the solution to the Poisson equation, under some mild conditions.

Lemma 1. *Given $\Omega = \mathbb{R}^N, N \geq 3$, assume that the source function $\rho \in \mathcal{C}^0(\Omega)$, and ρ has a compact support. Then the the Poisson equation $\nabla^2\varphi(\mathbf{x}) = -\rho(\mathbf{x})$ on Ω with zero boundary condition at infinity ($\lim_{\|\mathbf{x}\|_2\to\infty}\varphi(\mathbf{x}) = 0$) has a unique solution $\varphi(\mathbf{x}) \in \mathcal{C}^2(\Omega)$ up to a constant.*

Proof. For the existence of the solution, one can verify that the analytical construction using the extension of Green's function in $N \geq 3$ dimensional space (Lemma 4), *i.e.*, $\varphi(\mathbf{x}) = \int G(\mathbf{x},\mathbf{y})\rho(\mathbf{y})d\mathbf{y}, G(\mathbf{x},\mathbf{y}) = \frac{1}{(N-2)S_{N-1}(1)}\frac{1}{\|\mathbf{x}-\mathbf{y}\|^{N-2}}$, is one possible solution to the Poisson equation $\nabla^2\varphi(\mathbf{x}) = -\rho(\mathbf{x})$. Since $\rho \in \mathcal{C}^0(\Omega)$ and $\nabla^2\varphi(\mathbf{x}) = -\rho(\mathbf{x})$, we conclude that $\varphi(\mathbf{x}) \in \mathcal{C}^2(\Omega)$.

The proof idea of the uniqueness is similar to the uniqueness theorems in electrostatics. Suppose we have two different solutions $\varphi_1, \varphi_2 \in \mathcal{C}^2$ which satisfy

$$\nabla^2\varphi_1(\mathbf{x}) = -\rho(\mathbf{x}), \nabla^2\varphi_2(\mathbf{x}) = -\rho(\mathbf{x}). \tag{7}$$

We define $\tilde{\varphi}(\mathbf{x}) \equiv \varphi_2(\mathbf{x}) - \varphi_1(\mathbf{x})$. Subtracting the above two equations gives

$$\nabla^2\tilde{\varphi}(\mathbf{x}) = 0, \forall \mathbf{x} \in \Omega. \tag{8}$$

By the vector differential identity we have

$$\tilde{\varphi}(\mathbf{x})\nabla^2\tilde{\varphi}(\mathbf{x}) = \nabla\cdot(\tilde{\varphi}(\mathbf{x})\nabla\tilde{\varphi}(\mathbf{x})) - \nabla\tilde{\varphi}(\mathbf{x})\cdot\nabla\tilde{\varphi}(\mathbf{x}), \tag{9}$$

By the divergence theorem we have

$$\int_\Omega \nabla\cdot(\tilde{\varphi}(\mathbf{x})\nabla\tilde{\varphi}(\mathbf{x}))d^N\mathbf{x} = \oiint_{\partial\Omega} \tilde{\varphi}(\mathbf{x})\nabla\tilde{\varphi}(\mathbf{x})\cdot d^{N-1}\mathbf{S} = 0, \tag{10}$$

where $d^{N-1}\mathbf{S}$ denotes an $N-1$ dimensional surface element at infinity, and the second equation holds due to zero boundary condition at infinity. Combining Eq. (8)(9)(10), we have

$$\int_\Omega \nabla\cdot(\tilde{\varphi}(\mathbf{x})\nabla\tilde{\varphi}(\mathbf{x}))d^N\mathbf{x} = \int_\Omega \|\nabla\tilde{\varphi}(\mathbf{x})\|^2 d^N\mathbf{x} = 0, \tag{11}$$

since this is an integral of a positive quantity, we must have $\nabla\tilde{\varphi}(\mathbf{x}) = \mathbf{0}$, or $\tilde{\varphi}(\mathbf{x}) = c$, $\forall \mathbf{x} \in \Omega$. This means φ_1 and φ_2 differ at most by a constant, but a constant does not affect gradients, so $\nabla\varphi_1(\mathbf{x}) = \nabla\varphi_2(\mathbf{x})$. □

In our method section (Section 3.1), we augmented the original N-dimensional data with an extra dimension. The new data distribution in the augmented space is $\tilde{p}(\tilde{\mathbf{x}}) = p(\mathbf{x})\delta(z)$, where δ is the Dirac delta function. The support of the data distribution is in the $z = 0$ hyperplane. In the following lemma, we show the existence and uniqueness of the solution to $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$ outside the data support.

Lemma 2. *Assume the support of the data distribution in the augmented space (supp$(\tilde{p}(\tilde{\mathbf{x}}))$) is a compact set on the $z = 0$ hyperplane, $p(\mathbf{x}) \in \mathcal{C}^0$ and $N \geq 3$. The Poisson equation $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$ with zero boundary condition at infinity ($\lim_{\|\mathbf{x}\|_2\to\infty}\varphi(\tilde{\mathbf{x}}) = 0$) has a unique solution $\varphi(\tilde{\mathbf{x}}) \in \mathcal{C}^2$ for $\tilde{x} \in \mathbb{R}^{N+1} \setminus$ supp$(\tilde{p}(\tilde{\mathbf{x}}))$, up to a constant.*

Proof. Similar to the proof in Lemma 1, one can easily verify that the analytical construction using Green's method, *i.e.*, $\varphi(\tilde{\mathbf{x}}) = \int G(\tilde{\mathbf{x}},\tilde{\mathbf{y}})\tilde{p}(\tilde{\mathbf{x}})d\tilde{\mathbf{y}}, G(\tilde{\mathbf{x}},\tilde{\mathbf{y}}) = \frac{1}{(N-1)S_N(1)}\frac{1}{\|\tilde{\mathbf{x}}-\tilde{\mathbf{y}}\|^{N-1}}$, is one possible solution to the Poisson equation $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$. Since $\tilde{p}(\tilde{\mathbf{x}}) = 0$ for $\tilde{\mathbf{x}} \in \mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$ and $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$, we conclude that $\varphi(\tilde{\mathbf{x}}) \in \mathcal{C}^2(\mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}})))$.

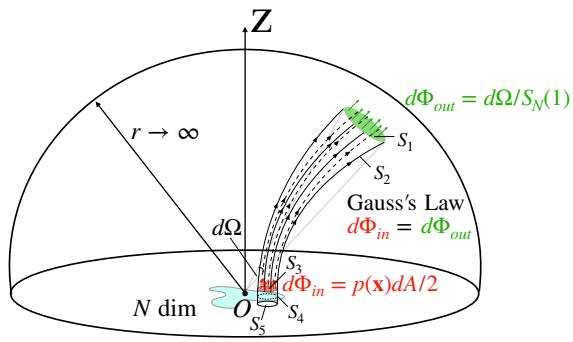


Figure 7: Proof idea of Theorem 2. By Gauss's Law, the outflow flux $d\Phi_{out}$ equals the inflow flux $d\Phi_{in}$. The factor of two in $p(\mathbf{x})dA/2$ is due to the symmetry of Poisson fields in $z < 0$ and $z > 0$.

For the uniqueness, suppose we have two different solutions $\varphi_1, \varphi_2 \in \mathcal{C}^2(\mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}})))$ which satisfy

$$\nabla^2\varphi_1(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}}), \nabla^2\varphi_2(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}}). \tag{12}$$

We define $\tilde{\varphi}(\tilde{\mathbf{x}}) \equiv \varphi_2(\tilde{\mathbf{x}}) - \varphi_1(\tilde{\mathbf{x}})$. Subtracting the above two equations gives

$$\nabla^2\tilde{\varphi}(\tilde{\mathbf{x}}) = 0, \forall \tilde{\mathbf{x}} \in \mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}})). \tag{13}$$

By the vector differential identity we have

$$\tilde{\varphi}(\tilde{\mathbf{x}})\nabla^2\tilde{\varphi}(\tilde{\mathbf{x}}) = \nabla \cdot (\tilde{\varphi}(\tilde{\mathbf{x}})\nabla\tilde{\varphi}(\tilde{\mathbf{x}})) - \nabla\tilde{\varphi}(\tilde{\mathbf{x}}) \cdot \nabla\tilde{\varphi}(\tilde{\mathbf{x}}), \tag{14}$$

By the divergence theorem we have

$$\int_{\mathbb{R}^{N+1}} \nabla \cdot (\tilde{\varphi}(\tilde{\mathbf{x}})\nabla\tilde{\varphi}(\tilde{\mathbf{x}}))d^{N+1}\tilde{\mathbf{x}} = \oiint_{\partial\mathbb{R}^{N+1}} \tilde{\varphi}(\tilde{\mathbf{x}})\nabla\tilde{\varphi}(\tilde{\mathbf{x}}) \cdot d^N\mathbf{S} = 0, \tag{15}$$

where $d^N\mathbf{S}$ denotes an N dimensional surface element at infinity, and the second equation holds due to zero boundary condition at infinity. Combining Eq. (13)(14)(15), we have

$$\begin{aligned}\int_{\mathbb{R}^{N+1}} \nabla \cdot (\tilde{\varphi}(\tilde{\mathbf{x}})\nabla\tilde{\varphi}(\tilde{\mathbf{x}}))d^{N+1}\tilde{\mathbf{x}} &= \int_{\mathbb{R}^{N+1}\setminus\text{supp}(\tilde{p}(\tilde{\mathbf{x}}))} \nabla \cdot (\tilde{\varphi}(\tilde{\mathbf{x}})\nabla\tilde{\varphi}(\tilde{\mathbf{x}}))d^{N+1}\tilde{\mathbf{x}} \\ &= \int_{\mathbb{R}^{N+1}\setminus\text{supp}(\tilde{p}(\tilde{\mathbf{x}}))} \|\nabla\tilde{\varphi}(\tilde{\mathbf{x}})\|^2 d^{N+1}\tilde{\mathbf{x}} = 0,\end{aligned}$$

The first equation holds because Lebesgue measure of $\text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$ is zero. Since $\|\nabla\tilde{\varphi}(\tilde{\mathbf{x}})\|^2$ is an integral of a positive quantity, we must have $\nabla\tilde{\varphi}(\tilde{\mathbf{x}}) = \mathbf{0}$, or $\tilde{\varphi}(\tilde{\mathbf{x}}) = c$, $\forall \tilde{\mathbf{x}} \in \mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$. This means φ_1 and φ_2 differ at most by a constant function, but a constant does not affect gradients, so $\nabla\varphi_1(\tilde{\mathbf{x}}) = \nabla\varphi_2(\tilde{\mathbf{x}})$. □

As illustrated in Fig. 7, there is a bijective mapping between the upper hemisphere of radius r and the $z = 0$ plane, where each pair of corresponding points is connected by an electric field line. We will now formally prove that, in the $r \to \infty$ limit, this mapping transforms the arbitrary charge distribution in the source plane (that generated the electric field) into a uniform distribution on the hemisphere.

Theorem 2. *Suppose particles are sampled from a uniform distribution on the upper ($z > 0$) half of the sphere of radius r and evolved by the backward ODE $\frac{d\tilde{\mathbf{x}}}{dt} = -\mathbf{E}(\tilde{\mathbf{x}})$ until they reach the $z = 0$ hyperplane, where the Poisson field $\mathbf{E}(\tilde{\mathbf{x}})$ is generated by the source $\tilde{p}(\tilde{\mathbf{x}})$. In the $r \to \infty$ limit, under the conditions in Lemma 2, this process generates a particle distribution $\tilde{p}(\tilde{\mathbf{x}})$, i.e., a distribution $p(\mathbf{x})$ in the $z = 0$ hyperplane.*

Proof. By Lemma 2, we know that with zero boundary at infinity, the Poisson equation $\nabla^2\varphi(\tilde{\mathbf{x}}) = -\tilde{p}(\tilde{\mathbf{x}})$ has a unique solution $\varphi(\tilde{\mathbf{x}}) \in \mathcal{C}^2$ for $\tilde{\mathbf{x}} \in \mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$. Hence $\mathbf{E}(\tilde{\mathbf{x}}) = -\nabla\varphi(\tilde{\mathbf{x}}) \in \mathcal{C}^1$, guaranteeing the existence-uniqueness of the solution to the ODE $\frac{d\tilde{\mathbf{x}}}{dt} = -\mathbf{E}(\tilde{\mathbf{x}})$ according to Theorem 2.8.1 in [27].

Consider the tube in Fig. 7 connecting an area on dA in the $z = \epsilon \to 0^+$ hyperplane (S_3) to a solid angle $d\Omega$ on the hemisphere (S_1), with S_2 as its side. The tube is the space swept by dA following electric field $\mathbf{E}$, so by definition the electric field is parallel to the tangent space of the tube sides S_2. The bottom of the tube S_3 is located at $z = \epsilon \to 0^+$, a bit above the $z = 0$ plane, so the tube does not enclose any charges. We note that the divergence of Poisson field is zero in $\mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$:

$$\nabla \cdot \mathbf{E}(\tilde{\mathbf{x}}) = -\nabla^2 \varphi(\tilde{\mathbf{x}}) = \tilde{p}(\tilde{\mathbf{x}}) = 0, \forall \tilde{\mathbf{x}} \in \mathbb{R}^{N+1} \setminus \text{supp}(\tilde{p}(\tilde{\mathbf{x}}))$$

Denote the volume and surface of the tube as V and $\mathbf{B}$. According to divergence theorem, $\oiint \mathbf{E}(\tilde{\mathbf{x}}) \cdot d\mathbf{B} = \int_V \nabla \cdot \mathbf{E}(\tilde{\mathbf{x}}) dV = 0$. Hence the net flux leaving the tube is zero:

$$\Phi_{S_1} + \Phi_{S_2} + \Phi_{S_3} = 0, \quad \Phi_{S_i} \equiv \oiint_{S_i} \mathbf{E}(\tilde{\mathbf{x}}) \cdot d\mathbf{B} \quad (i = 1, 2, 3) \tag{16}$$

There is no flux through the sides, i.e., $\Phi_{S_2} = 0$, since $\mathbf{E}(\tilde{\mathbf{x}})$ is orthogonal to the surface element $d\mathbf{B}$ on the tube sides by definition. As a result, the flux Φ_{S_3} entering from below must equal the flux Φ_{S_1} leaving the other end. Denote the l_2 norm of the vector $\mathbf{r}$ as r. We first calculate the influx Φ_{S_3}. To do so, we study a Gaussian pillbox whose top, side and bottom are S_3, S_4 and S_5. S_3 and S_5 are located at $z = \epsilon$ and $z = -\epsilon$ ($\epsilon \to 0^+$). Denote the volume and surface of the pillbox as V' and $\mathbf{B}'$. The pillbox contains charge $p(\mathbf{x})dA$, so according to Gauss's law $\oiint \mathbf{E}(\tilde{\mathbf{x}}) \cdot d\mathbf{B}' = \int_{V'} \nabla \cdot \mathbf{E}(\tilde{\mathbf{x}}) dV' = \int_{V'} \tilde{p}(\tilde{\mathbf{x}}) dV' = p(\mathbf{x}) dA$, i.e.,

$$\Phi'_{S_3} + \Phi'_{S_4} + \Phi'_{S_5} = p(\mathbf{x})dA, \quad \Phi'_{S_i} \equiv \oiint_{S_i} \mathbf{E}(\tilde{\mathbf{x}}) \cdot d\mathbf{B}' \quad (i = 3, 4, 5) \tag{17}$$

The flux on the sides $\Phi'_{S_4} \propto \epsilon \to 0$, and $\Phi'_{S_3} = \Phi'_{S_5}$ due to mirror symmetry of $z = 0$. So $\Phi'_{S_3} = \Phi'_{S_5} = p(\mathbf{x})dA/2$. Note on the S_3 surface, the outflux of the pillbox is exactly the influx of the tube, so we have:

$$\Phi_{S_3} = -\Phi'_{S_3} = -p(\mathbf{x})dA/2, \tag{18}$$

inserting which and $\Phi_{S_2} = 0$ to Eq. (16) gives

$$\Phi_{S_1} = -\Phi_{S_3} = p(\mathbf{x})dA/2. \tag{19}$$

On the other hand, in the far-field limit $r \to \infty$, since $\text{supp}(p(\mathbf{x}))$ is bounded, the data distribution can be effectively seen as a point charge (see Appendix A.2). By Lemma 3, we have $\lim_{r\to\infty} \mathbf{E}(\mathbf{r}) = -\lim_{r\to\infty} \nabla\varphi(\mathbf{r}) = \frac{\mathbf{r}}{S_N(1) r^{N+1}}$. The resulting outflux on the hemisphere is

$$\Phi_{S_1} = E_r r^N d\Omega = d\Omega / S_N(1) \tag{20}$$

where $E_r \equiv \mathbf{E}(\mathbf{r}) \cdot \mathbf{r}/r$ is the radial component of $\mathbf{E}$. Comparing Eq. (19) and Eq. (20) yields $d\Omega/dA = p(\mathbf{x}) S_N(1)/2 \propto p(\mathbf{x})$. In other words, the mapping from the $z = 0$ hyperplane to the hemisphere dilutes the charge density $p(\mathbf{x})$ up to a constant factor. Thus by change-of-varible, we conclude that the mapping transforms the data distribution into a uniform distribution on the infinite hemisphere. Since the ODE is reversible, the backward ODE transforms the uniform distributoin on the infinite hemisphere to the distribution $\tilde{p}(\tilde{\mathbf{x}})$. □

A.2 Multipole Expansion

We discuss the behaviors of the potential function in Poisson equation (Eq. (1)) under different scenarios, utilizing the multipole expansion. Suppose we have a unit point charge $q = 1$ located at $\mathbf{x} \in \mathbb{R}^N$. We know that the potential function at another point $\mathbf{y} \in \mathbb{R}^N$ is $\varphi(\mathbf{y} - \mathbf{x}) = 1/\|\mathbf{y} - \mathbf{x}\|^{N-2}$ (ignoring a constant factor). Now we assume that $\mathbf{x}$ is close to the origin such that we can Taylor expand around $\mathbf{x} = 0$:

$$\varphi(\mathbf{y} - \mathbf{x}) = \varphi(\mathbf{y}) - \sum_{\alpha=1}^{N} \mathbf{x}_\alpha \varphi_\alpha(\mathbf{y}) + \frac{1}{2} \sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} \mathbf{x}_\alpha \mathbf{x}_\beta \varphi_{\alpha\beta}(\mathbf{y}) - \ldots \tag{21}$$

where

$$\begin{aligned} \varphi_\alpha(\mathbf{y}) &= \left(\frac{\partial \varphi(\mathbf{y} - \mathbf{x})}{\partial \mathbf{x}_\alpha} \right)_{\mathbf{x}=0} = (N-2) \frac{\mathbf{y}_\alpha}{\|\mathbf{y}\|^N} \\ \varphi_{\alpha\beta}(\mathbf{y}) &= \left(\frac{\partial^2 \varphi(\mathbf{y} - \mathbf{x})}{\partial \mathbf{x}_\alpha \partial \mathbf{x}_\beta} \right)_{\mathbf{x}=0} = (N-2) \frac{N \mathbf{y}_\alpha \mathbf{y}_\beta - \|\mathbf{y}\|^2 \delta_{\alpha\beta}}{\|\mathbf{y}\|^{N+2}} \end{aligned} \tag{22}$$

In the case where the source is a distribution $p(\mathbf{x})$, the potential $\varphi(\mathbf{y})$ can again be Taylor expanded:

$$\varphi(\mathbf{y}) = q\varphi(\mathbf{y}) + \sum_{\alpha=1}^{N} q_\alpha \varphi_\alpha(\mathbf{y}) + \sum_{\alpha=1}^{N}\sum_{\beta=1}^{N} q_{\alpha\beta}\varphi_{\alpha\beta}(\mathbf{y}) - \dots \tag{23}$$

where

$$q = \int p(\mathbf{x})d\mathbf{x}, q_\alpha = \int p(\mathbf{x})\mathbf{x}_\alpha d\mathbf{x}, q_{\alpha\beta} = \int p(\mathbf{x})\mathbf{x}_\alpha\mathbf{x}_\beta d\mathbf{x}, \tag{24}$$

which are called monopole, dipole and quadrupole in physics, respectively. The gradient field $\mathbf{E}(y) = \nabla\Phi(\mathbf{y})$ can be expanded in the same such that

$$\mathbf{E}(\mathbf{y}) = \mathbf{E}^{(0)}(\mathbf{y}) + \mathbf{E}^{(1)}(\mathbf{y}) + \mathbf{E}^{(2)}(\mathbf{y}) + \dots \tag{25}$$

It is easy to check that $||\mathbf{E}^{(i)}(\mathbf{y})||$ decays as $1/||\mathbf{y}||^{N-2+i}$, which means higher-order corrections decay faster than leading terms. So when $||\mathbf{y}|| \to \infty$, only the monopole term $||\mathbf{E}^{(0)}(\mathbf{y})||$ matters, which behaves like a point source.

In a more realistic setup, we only have a large but finite $||\mathbf{y}||$, so the question is: under what condition is the point source approximation valid? We examine $\varphi^{(0)}$, $\varphi^{(1)}$ and $\varphi^{(2)}$ more carefully:

$$\begin{aligned}
\varphi^{(0)} &= \frac{1}{||\mathbf{y}||^{N-2}} \\
\varphi^{(1)} &= \sum_{\alpha=1}^{N}(N-2)\frac{\mathbf{y}_\alpha\mathbf{x}_\alpha}{||\mathbf{y}||^N} = (N-2)\frac{\mathbf{x}^T\mathbf{y}}{||\mathbf{y}||^N} \\
\varphi^{(2)} &= \frac{1}{2}\sum_{\alpha=1}^{N}\sum_{\beta=1}^{N}(N-2)\frac{N\mathbf{y}_\alpha\mathbf{y}_\beta - ||\mathbf{y}||^2\delta_{\alpha\beta}}{||\mathbf{y}||^{N+1}}\mathbf{x}_\alpha\mathbf{x}_\beta = \frac{N-2}{2}\frac{N(\mathbf{x}^T\mathbf{y})^2 - ||\mathbf{x}||^2||\mathbf{y}||^2}{||\mathbf{y}||^{N+2}}
\end{aligned} \tag{26}$$

Since $\varphi^{(1)}$ is an odd function of $\mathbf{x}$, integrating $\varphi^{(1)}$ over $\mathbf{x}$ leads to zero (samples are normalized to zero mean). In machine learning applications, N is usually a large number (although in physics N is merely 3). If $\mathbf{y}$ is a random vector of length $||\mathbf{y}||$, then $\mathbf{x}^T\mathbf{y} \sim (\frac{1}{\sqrt{N}} \pm \frac{1}{N})||\mathbf{x}||||\mathbf{y}||$. So Eq. (26) can be approximated as

$$\varphi^{(0)} \sim \frac{1}{||\mathbf{y}||^{N-2}}, \varphi^{(2)} \sim \frac{\sqrt{N}}{2}\frac{||\mathbf{x}||^2}{||\mathbf{y}||^N} \tag{27}$$

Requiring $\int \varphi^{(0)}p(\mathbf{x})d\mathbf{x} \gg \int \varphi^{(2)}p(\mathbf{x})d\mathbf{x}$ gives $||\mathbf{y}||^2 \gg \sqrt{N}\,\mathbb{E}_{p(x)}\,||\mathbf{x}||^2$. So the condition for the point source approximation to be valid is:

$$\kappa = \frac{||\mathbf{y}||^2}{\sqrt{N}\,\mathbb{E}_{p(x)}\,||\mathbf{x}||^2} \gg 1 \tag{28}$$

Based on this condition, we can partition space into three zones: (1) the far zone $\kappa \gg 1$, where the point source approximation is valid; (2) the intermediate zone $\kappa \sim O(1)$, where the gradient field has moderate curvature; (3) the near zone $\kappa \ll 1$, where the gradient field has high curvature. In practice, the initial value $||\mathbf{y}||$ is greater than 1000 (hence $\kappa \gg 1$) with high probability on CIFAR-10 and CelebA datasets, incidating that the initial samples lie in the far zone and gradually move toward the near zone where $||\mathbf{y}|| \approx ||\mathbf{x}||$ ($\kappa \ll 1$).

We summarize above observations in the following lemma in the $||\mathbf{y}|| \to \infty$ limit:

Lemma 3. *Assume the data distribution $p(\mathbf{x}) \in \mathcal{C}^0$ has a compact support in $\mathbb{R}^N$, then the solution φ to the Poisson equation $\nabla^2\varphi(\mathbf{x}) = -p(\mathbf{x})$ with zero boundary condition at infinity satisfies $\lim_{\|\mathbf{x}\|_2\to\infty}\nabla\varphi(\mathbf{x}) = -\frac{1}{S_{N-1}(1)}\frac{\mathbf{x}}{\|\mathbf{x}\|_2^N}$.*

Proof. By Lemma 1, the gradient of the solution has the following form:

$$\nabla\varphi(\mathbf{x}) = \int \nabla_{\mathbf{x}}G(\mathbf{x},\mathbf{y})p(\mathbf{y})d\mathbf{y}, \quad \nabla_{\mathbf{x}}G(\mathbf{x},\mathbf{y}) = -\frac{1}{S_{N-1}(1)}\frac{\mathbf{x}-\mathbf{y}}{||\mathbf{x}-\mathbf{y}||^N}.$$

Since $p(\mathbf{x})$ has a bounded support, we assume $\max\{\| \mathbf{x} \|_2 \colon p(\mathbf{x}) \neq 0\} < B$. On the other hand, we have

$$\lim_{\|\mathbf{x}\|_2\to\infty}\nabla_{\mathbf{x}}G(\mathbf{x},\mathbf{y}) = \lim_{\|\mathbf{x}\|_2\to\infty} -\frac{1}{S_{N-1}(1)}\frac{\mathbf{x}-\mathbf{y}}{||\mathbf{x}-\mathbf{y}||^N} = \lim_{\|\mathbf{x}\|_2\to\infty} -\frac{1}{S_{N-1}(1)}\frac{\mathbf{x}}{||\mathbf{x}||^N}$$

for $\forall y$ such that $\| y \|_2 < B$. Hence,

$$\lim_{\|\mathbf{x}\|_2 \to \infty} \nabla \varphi(\mathbf{x}) = \lim_{\|\mathbf{x}\|_2 \to \infty} \int \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) p(\mathbf{y}) d\mathbf{y} = \int \lim_{\|\mathbf{x}\|_2 \to \infty} \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) p(\mathbf{y}) d\mathbf{y}$$
$$= -\frac{1}{S_{N-1}(1)} \frac{\mathbf{x}}{\| \mathbf{x} \|_2^N}$$

□

A.3 Extension of Green's Function in N-dimensional Space

In this section, we show that the function $G(\mathbf{x}, \mathbf{y})$ defined in Eq. (2) is the N-dimensional extension of the Green's function, $\varphi(\mathbf{x}) = \int G(\mathbf{x}, \mathbf{y}) \rho(\mathbf{y}) d\mathbf{y}$ solves the Poisson equation $\nabla^2 \varphi(\mathbf{x}) = -\rho(\mathbf{x})$.

Lemma 4. *Assume the dimension $N \geq 3$, and the source term satisfies $\rho \in \mathcal{C}^0(\Omega)$, $\int_{\mathbb{R}^N} \rho^2(\mathbf{x}) d\mathbf{x} < +\infty$, $\lim_{\|\mathbf{x}\|_2 \to \infty} \rho(\mathbf{x}) = 0$. The extension of Green's function $G(\mathbf{x}, \mathbf{y}) = \frac{1}{(N-2)S_{N-1}(1)} \frac{1}{\|\mathbf{x}-\mathbf{y}\|^{N-2}}$ solves the Poisson equation $\nabla_{\mathbf{x}}^2 G(\mathbf{x}, \mathbf{y}) = -\delta(\mathbf{x} - \mathbf{y})$. In addition, with zero boundary condition at infinity ($\lim_{\|\mathbf{x}\|_2 \to \infty} \varphi(\mathbf{x}) = 0$), $\varphi(\mathbf{x}) = \int G(\mathbf{x}, \mathbf{y}) \rho(\mathbf{y}) d\mathbf{y}$ solves the Poisson equation $\nabla^2 \varphi(\mathbf{x}) = -\rho(\mathbf{x})$.*

Proof. It is convenient to denote $\mathbf{r} = \mathbf{x} - \mathbf{y}$, $r = \|\mathbf{r}\|$ and notice $\partial r / \partial \mathbf{x} = \mathbf{r}/r$. Firstly, we calculate $\nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y})$:

$$\begin{aligned} \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) &= \frac{1}{(N-2)S_{N-1}(1)} \nabla_{\mathbf{x}} \left(\frac{1}{r^{N-2}}\right) \\ &= \frac{1}{(N-2)S_{N-1}(1)} \frac{\partial}{\partial r} \left(\frac{1}{r^{N-2}}\right) \nabla_{\mathbf{x}} r \\ &= -\frac{1}{S_{N-1}(1)} \frac{\mathbf{r}}{r^N} \end{aligned} \tag{29}$$

Then we calculate $\nabla_{\mathbf{x}}^2 G(\mathbf{x}, \mathbf{y})$:

$$\begin{aligned} \nabla_{\mathbf{x}}^2 G(\mathbf{x}, \mathbf{y}) &\equiv \nabla_{\mathbf{x}} \cdot \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) \\ &= -\frac{1}{S_{N-1}(1)} \nabla_{\mathbf{x}} \cdot \frac{\mathbf{r}}{r^N} \\ &= -\frac{1}{S_{N-1}(1)} \left(\nabla_{\mathbf{x}} \left(\frac{1}{r^N}\right) \cdot \mathbf{r} + \frac{1}{r^N} \nabla_{\mathbf{r}} \cdot \mathbf{r}\right) \\ &= -\frac{1}{S_{N-1}(1)} \left(-\frac{N}{r^N} + \frac{N}{r^N}\right) \\ &= -\frac{0}{S_{N-1}(1) r^N} \end{aligned} \tag{30}$$

which is 0 for $r > 0$, but undermined for $r = 0$. So we are left with proving

$$\int_{S_\epsilon(\mathbf{y})} \nabla_{\mathbf{x}}^2 G(\mathbf{x}, \mathbf{y}) d^N \mathbf{x} = -1, \tag{31}$$

where $S_\epsilon(\mathbf{y})$ denotes a ball centered at $\mathbf{y}$ with a radius $\epsilon \to 0^+$. With the divergence theorem, we have

$$\int_{S_\epsilon(\mathbf{y})} \nabla_{\mathbf{x}}^2 G(\mathbf{x}, \mathbf{y}) d^N \mathbf{x} = \oiint_{\partial S_\epsilon(\mathbf{y})} \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) \cdot d^{N-1} \mathbf{B} \tag{32}$$

where the surface integral can be computed

$$\oiint_{\partial S_\epsilon(\mathbf{y})} \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) \cdot d^{N-1} \mathbf{B} = \oiint_{\partial S_\epsilon(\mathbf{y})} \left(-\frac{1}{S_{N-1}(1)} \frac{\mathbf{r}}{r^N}\right) \cdot d^{N-1} \mathbf{B} = -\frac{1}{S_{N-1}(1)} \frac{S_{N-1}(\epsilon)}{\epsilon^{N-1}} = -1 \tag{33}$$

in which we used $\oiint_{\partial S_\epsilon(\mathbf{y})} \mathbf{r} \cdot d^{N-1} \mathbf{B} = \epsilon S_{N-1}(\epsilon)$. Together, we conclude that

$$\nabla_{\mathbf{x}}^2 G(\mathbf{x}, \mathbf{y}) = -\delta(\mathbf{x} - \mathbf{y}) \tag{34}$$

Next we show that $\varphi(\mathbf{x}) = \int G(\mathbf{x},\mathbf{y})\rho(\mathbf{y})d\mathbf{y}$ solves $\nabla^2\varphi(\mathbf{x}) = -\rho(\mathbf{x})$. Taking the Laplacian operator of both sides gives:

$$\begin{aligned}\nabla^2_{\mathbf{x}}\varphi(\mathbf{x}) &= \nabla^2_{\mathbf{x}}\int G(\mathbf{x},\mathbf{y})\rho(\mathbf{y})d\mathbf{y}\\ &= \int \nabla^2_{\mathbf{x}}G(\mathbf{x},\mathbf{y})\rho(\mathbf{y})d\mathbf{y}\\ &= \int -\delta(\mathbf{x}-\mathbf{y})\rho(\mathbf{y})d\mathbf{y} \quad \text{(By Eq. (34))}\\ &= -\rho(\mathbf{x})\end{aligned}$$

In addition, we show that $\varphi(\mathbf{x})$ is zero at infinity. Since $\rho(\mathbf{x}) \in \mathcal{C}^0$ and has compact support, we know that $\rho(\mathbf{x})$ is bounded, and let $|\rho(\mathbf{x})| < B$.

$$\begin{aligned}\lim_{\|\mathbf{x}\|_2\to\infty}\varphi(\mathbf{x}) &= \lim_{\|\mathbf{x}\|_2\to\infty}\int G(\mathbf{x},\mathbf{y})\rho(\mathbf{y})d\mathbf{y}\\ &\le B\lim_{\|\mathbf{x}\|_2\to\infty}\int_{\text{supp}(\rho)}\frac{1}{(N-2)S_{N-1}(1)}\frac{1}{\|\mathbf{x}-\mathbf{y}\|^{N-2}}d\mathbf{y}\\ &= 0\end{aligned}$$

The last equality holds since $\text{supp}(\rho)$ is a compact set. □

A.4 Proof for the Prior Distribution on $z = z_{\text{max}}$ Hyperplane

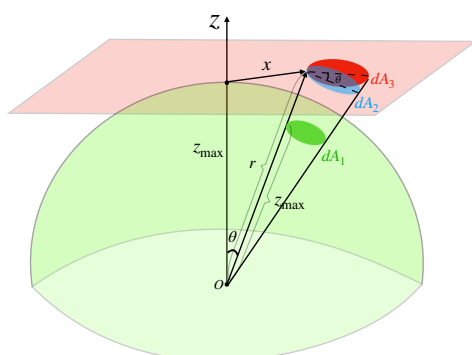


Figure 8: Diagram of the deviation in Proposition 1

We obtain the prior distribution p_{prior} by projecting the uniform distribution $\mathcal{U}(S^+_N(z_{\text{max}}))$ on the hemisphere $S^+_N(z_{\text{max}})$ to the $z = z_{\text{max}}$ hyperplane. In the following proposition, we show that the projected distribution is $p_{\text{prior}}(\mathbf{x}) = \frac{2z_{\text{max}}}{S_N(1)r^{N+1}}$.

Proposition 1. *The radial projection of $\mathcal{U}(S^+_N(z_{max}))$ on the hemisphere $S^+_N(z_{max})$ to the $z = z_{max}$ hyperplane is $p_{prior}(\mathbf{x}) = \frac{2z_{max}}{S_N(1)r^{N+1}}$.*

Proof. We calculate the change-of-variable ratio by comparing two associate areas. As illustrated in Fig. 8, an area dA_1 on $S^+_N(z_{\text{max}})$ is projected to an area dA_3 on the hyperplane in the $(\mathbf{x}, z_{\text{max}})$ direction, and we have

$$\mathcal{U}(S^+_N(z_{\text{max}}))dA_1 = p_{\text{prior}}(\mathbf{x})dA_3$$

We aim to calculate the ratio dA_1/dA_3 below. We define the angle between $(\mathbf{0}, z_{\text{max}})$ and $\tilde{\mathbf{x}} = (\mathbf{x}, z_{\text{max}})$ to be θ. We project dA_3 to the hyperplane orthogonal to $\tilde{\mathbf{x}}$ to get $dA_2 = dA_3\cos\theta = dA_3 z_{\text{max}}/r$ where $r \equiv \|\tilde{\mathbf{x}}\| = \sqrt{\|\mathbf{x}\|^2 + z^2_{\text{max}}}$. Since dA_1 is parallel to dA_2 and they lie in the same cone from the origin O, we have $dA_2/dA_1 = (r/z_{\text{max}})^N$. Combining all the results gives

$$p_{\text{prior}}(\mathbf{x}) = \mathcal{U}(S^+_N(z_{\text{max}}))\frac{dA_1}{dA_3} = \mathcal{U}(S^+_N(z_{\text{max}}))\frac{dA_1}{dA_2}\frac{dA_2}{dA_3} = \frac{2}{S_N(1)z^N_{\text{max}}}(\frac{z_{\text{max}}}{r})^N\frac{z_{\text{max}}}{r} = \frac{2z_{\text{max}}}{S_N(1)r^{N+1}}.$$

□

B Experimental Details

B.1 Training

In this section we include more details about the training of PFGM and other baselines. We show the hyper-parameters settings for all the baselines (Appendix B.1.1). All the experiments are run on a single NVIDIA A100 GPU.

B.1.1 Additional Settings

PFGM We set the hyper-parameters $\gamma = 5$, the larger batch size for calculating normalize field $|\mathcal{B}_L| = 2048$ (CIFAR-10), 256 (CelebA), 64 (LSUN bedroom) in Algorithm 1, and $M = 291$ (CIFAR-10, CelebA)/356 (LSUN bedroom), $\sigma = 0.01$ and $\tau = 0.03$ in Algorithm 2. We use the a batch size of $|\mathcal{B}| = 128$ (CIFAR-10, CelebA)/32 (LSUN bedroom), the same Adam optimizer and exponential moving average in [33]. We center the data around the origin. The initial z components in the normalized field are approximately zero with small initial $|\epsilon_z|$ values in Algorithm 2. In this case, the trajectories of the forward ODE terminate at points that are unlikely traversed by the backward ODE, *i.e.*, points with large $\| x \|_2$ and small z. In light of this, we heuristically confine the maximum sampling step to $M = 200$ (CIFAR-10, CelebA)/250 (LSUN bedroom) for points with the initial $|\epsilon_z|$ smaller than 0.005. More principal solutions are left for future works.

Since we are operating in the augmented space, we add minor modifications to the DDPM++/DDPM++ deep architectures to accommodate the extra dimension. More specifically, we replace the conditioning time variable in VP/sub-VP with the additional dimension z in PFGM as the input to the positional embedding. We also need to add an extra scalar output representing the z direction. To this end, we add an additional output channel to the final convolution layer and take the global average pooling of this channel to obtain the scalar. For LSUN bedroom dataset, we both experiments with the channel configurations suggested in NSCN++ [33] and DDPM [16].

VE/VP/sub-VP We use the same set of hyper-parameters and the NCSN++/DDPM++ (deep) backbone and the continuous-time training objectives for forward SDEs in [33].

B.2 Sampling

We provide more details of PFGM and VE/VP sampling implementations in Appendix B.2.1. We further discuss two techniques used in PFGM ODE sampler: change-of-variable formula (Appendix B.2.2) and the substitution of ground-truth Poisson field direction on z (Appendix B.2.3).

B.2.1 Additional settings

PFGM For RK-45 sampler, we use the function implemented in `scipy.integrate.solve_ivp` with `atol`=$1e-4$, `rtol`=$1e-4$. For forward Euler method, we discretize the ODE with constant step size determined by the number of steps, *i.e.*, step size = $(\log z_{\text{max}} - \log z_{\text{min}})$/number of steps for the backward ODE (Eq. (6)). As in [1], we set the terminal value of $z_{min} = 1e-3$. We choose $z_{\text{max}} = 40$ (CIFAR-10, CelebA), 100 (LSUN bedroom) to satisfy the condition $\kappa \gg 1$ by the multipole expansion analysis in Appendix A.2. The condition ensures that the data distribution can be viewed roughly as a point source at origin. For example, we set $z_{max} = 40$ on CIFAR-10, and the corresponding κ is greater than 50 with high probability. The hyperparameters work well without further fine tuning. Hence, we hypothesize that PFGM is insensitive to the choice of hyperparameters in a reasonable range, as shown in Table 2. We clip the norms of initial samples into $(0, 3000)$ for CIFAR-10, $(0, 6000)$ for CelebA and $(0, 30000)$ for LSUN bedroom.

Table 2: FID scores versus z_{max} on PFGM w/ DDPM++

z_{max}	30	40	50
FID score	2.54	2.54	2.53

VE/VP/sub-VP For the PC sampler in VE, we follow [33] to set the reverse diffusion process as the predictor and the Langevin dynamics (MCMC) as the corrector. For VP/sub-VP, we drop the corrector in PC sampler since it only gives slightly better results [33].

B.2.2 Exponential Decay on z Dimension

Recall that in Section 3.3, we replace the vanilla backward ODE with a new ODE anchored by z:

$$d(\mathbf{x}, z) = -(\frac{d\mathbf{x}}{dt}\frac{dt}{dz}dz, dz) = -(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}, 1)dz$$

We further use the change-of-variable formula, *i.e.*, $t' = -\log z$, to achieve exponential decay on the z dimension:

$$d(\mathbf{x}, z) = -(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}z, z)dt'$$

The trajectories of the two ODEs above are the same when $dt, dt' \to 0$. We compare the NFE and the sample quality of different ODEs in Table 3. We measure the NFE/FID of generating 50000 CIFAR-10 samples with the RK45 method in Scipy package [37]. The batch size is set to 1000. All the numbers are produced on a single NVIDIA A100 GPU. We observe that the ODE with the anchor variable t' not only accelerates the vanilla by 2 times, but has almost no harm to the sample quality measured by FID score.

Table 3: NFE and FID scores of different backward ODEs in PFGM

Algorithm	$d(\mathbf{x}, z)/dz$	$d(\mathbf{x}, z)/dt'$
NFE	242	110
FID score	2.53	2.54

B.2.3 Substitute the Predicted z Direction with the Ground-truth

Since the neural network cannot perfectly learn the ground-truth z direction, we replace the predicted $f_\theta(x)_z$ with the ground-truth direction when z is small. More specifically, given $\tilde{\mathbf{x}} = (\mathbf{x}, z) \in \mathbb{R}^{N+1}$, recall that the empirical field is $\hat{\mathbf{E}}(\tilde{\mathbf{x}}) = c(\tilde{\mathbf{x}})\sum_{i=1}^n \frac{\tilde{\mathbf{x}}-\tilde{\mathbf{x}}_i}{\|\tilde{\mathbf{x}}-\tilde{\mathbf{x}}_i\|^{N+1}}$ where $c(\tilde{\mathbf{x}}) = 1/\sum_{i=1}^n \frac{1}{\|\tilde{\mathbf{x}}-\tilde{\mathbf{x}}_i\|^{N+1}}$. Hence we can rewrite the empirical field as

$$\hat{\mathbf{E}}(\tilde{\mathbf{x}}) = \sum_{i=1}^n w(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_i)(\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i)$$

where $\sum_{i=1}^n w(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_i) = \sum_{i=1}^n \frac{\frac{1}{\|\tilde{\mathbf{x}}-\tilde{\mathbf{x}}_i\|^{N+1}}}{\sum_{j=1}^n \frac{1}{\|\tilde{\mathbf{x}}-\tilde{\mathbf{x}}_j\|^{N+1}}} = 1$. Furthermore we have $\forall i, (\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i)_z = z - 0 = z$. Together, the z component in the empirical field is $\hat{\mathbf{E}}(\tilde{\mathbf{x}})_z = \sum_{i=1}^n w(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_i)(\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i)_z = z$. The predicted normalized field is trained to approximate the normalized field, *i.e.*, $f_\theta(\tilde{\mathbf{x}}) \approx \sqrt{N}\hat{\mathbf{E}}(\tilde{\mathbf{x}})/(\| \hat{\mathbf{E}}(\tilde{\mathbf{x}}) \|_2 + \gamma)$. Solving for $\| \hat{\mathbf{E}}(\tilde{\mathbf{x}}) \|_2$, we get $\| \hat{\mathbf{E}}(\tilde{\mathbf{x}}) \|_2 \approx \frac{\gamma\|f_\theta(\tilde{\mathbf{x}})\|_2/\sqrt{N}}{1-\|f_\theta(\tilde{\mathbf{x}})\|_2/\sqrt{N}}$. Hence the z component in the normalized field after substituting the ground-truth $\hat{\mathbf{E}}(\tilde{\mathbf{x}})_z$ is $\hat{\mathbf{E}}(\tilde{\mathbf{x}})_z/(\| \hat{\mathbf{E}}(\tilde{\mathbf{x}}) \|_2 + \gamma) = z/(\frac{\gamma\|f_\theta(\tilde{\mathbf{x}})\|_2/\sqrt{N}}{1-\|f_\theta(\tilde{\mathbf{x}})\|_2/\sqrt{N}} + \gamma)$. In our experiments, we therefore replace the original prediction $f_\theta(\tilde{\mathbf{x}})_z$ with $z/(\frac{\gamma\|f_\theta(\tilde{\mathbf{x}})\|_2/\sqrt{N}}{1-\|f_\theta(\tilde{\mathbf{x}})\|_2/\sqrt{N}} + \gamma)$ when $z < 5$ during the backward ODE sampling.

Table 4 reports the NFE and FID score w/o and w/ the above substitution. We observe that the usage of ground-truth z direction in the near field accelerates the sampling speed.

B.3 Evaluation

We use FID [13] and Inception scores [29] to quantitatively measure the sample quality, and NFE (number of evaluation steps) for the inference speed. FID (Fréchet Inception Distance) score is the Fréchet distance between two multivariate Gaussians, whose means and covariances are estimated

Table 4: NFE and FID scores of w/ and w/o substitution

Algorithm	w/o substitution	w/ substitution
NFE	134	110
FID score	2.54	2.54

from the 2048-dimensional activations of the Inception-v3 [34] network for real and generated samples respectively. Inception score is the exponential mutual information between the predicted labels of the Inception network and the images. We also report bits/dim for likelihood evaluation. It is computed by dividing the negative log-likelihood by the data dimension, *i.e.*, bits/dim $= -\log p_{\text{prior}}(\mathbf{x})/N$.

For CIFAR-10, we compute the Fréchet distance between 50000 samples and the pre-computed statistics of CIFAR-10 dataset in [13]. For CelebA 64×64, we follow the setting in [32] where the distance is computed between 10000 samples and the test set. For model selection, we follow [32] and pick the checkpoint with smallest FID every 50k iterations on 10k samples for computing all the scores.

B.4 Effects of Step Size: FID versus NFE

For preciseness, Table 5 reports the exact numbers in Fig. 5(c).

Table 5: The FID scores in Fig. 5(c) of different methods and NFE.

Method / NFE	10	20	50	100
VP-ODE	192.36	72.25	38.18	19.73
DDIM	13.36	6.48	4.67	4.16
PFGM	14.98	6.46	3.48	2.89

Since in the ODE $d(\mathbf{x}, z) = -(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}z, z)dt'$ of PFGM, the z variable is a function of t' ($z = e^{t'}$), we integrate the z in the Euler method to reduce the discretization error. The vanilla update from time t'_i to time t'_{i+1} is $(\mathbf{x}_{i+1}, z_{i+1}) = (\mathbf{x}_i, z_i) - (\mathbf{v}(\tilde{\mathbf{x}}_i)_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}}_i)_{z_i}^{-1}z_i, z_i)(t'_{i+1} - t'_i)$, and the new update is $(\mathbf{x}_{i+1}, z_{i+1}) = (\mathbf{x}_i, z_i) - (\mathbf{v}(\tilde{\mathbf{x}}_i)_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}}_i)_{z_i}^{-1}\int_{t'_i}^{t'_{i+1}} z(t')dt', \int_{t'_i}^{t'_{i+1}} z(t')dt')$. We empirically observe that the new update scheme significantly improve the FID score.

C Failure of VE/VP-ODE on NCSNv2 backbone

In Fig. 5(a), we demonstrate the trajectories of cleaner samples/noisier samples/noisier samples w/ corrector. We visualize these three groups in Fig. 9(a) and Fig. 9(b). The noisier samples are marked with red boxes in Fig. 9(a) and the remaining images in Fig. 9(a) are cleaner samples. The samples within green boxes in Fig. 9(b) are noisier samples w/ corrector. Samples on the same spatial locations in the two figures are generated by identical initial latents.

The Gaussian kernels in score-based models are $\mathcal{N}(\mathbf{x}, \sigma(t)^2)$ (VE) and $\mathcal{N}(\sqrt{1-\sigma(t)^2}\mathbf{x}, \sigma(t)^2)$ (VP) [33]. When $\sigma(t)$ is large, the norms of perturbed samples are approximately $\sqrt{N}\sigma(t)$. The backward ODE could break down if the trajectories diverge from the norm-$\sigma(t)$ relation, as shown by the noisier samples' trajectories in Fig. 5(a). In contrast, the norm distributions of PFGM is approximately $p(\|\mathbf{x}\|) \propto \|\mathbf{x}\|_2^{N-2}/(\|\mathbf{x}\|_2^2 + z^2)^{\frac{N}{2}}$ when z is large (see deviation for p_{prior} in Appendix A.4), which have a wider span for high density region (see Fig. 4). The weak correlation between norm and z makes PFGM more robust on the lighter NCSNv2 backbone.

(a) Samples from VE-ODE (Euler)

(b) Samples from VE-ODE (Euler w/ corrector)

Figure 9: **(a)** Samples from VE-ODE (Euler w/o corrector). We highlight the noisier images with red boxes. The rest are cleaner images. **(b)** Samples from VE-ODE (Euler w/ corrector). We mark the noisier samples after correction with green boxes.

D Extra Experiments

D.1 LSUN Bedroom 256×256

We report the FID scores and NFEs for LSUN bedroom dataset in Table 6. We adopt the code base of [33] in our experiments. In [33], they experimented on the LSUN bedroom 256× 256 dataset only on VE-SDE using a deeper NCSN++ backbone. In our DDPM++ architecture, we directly borrow the configuration of channels from the NCSN++ architecture [33] in each residual block (PFGM w/ NCSN++ channel). We further change z_{max} to 100, as it empirically gives better sample quality.

We also evaluate the performance when using the configuration of channels in the DDPM [16] architecture (PFGM w/ DDPM channel). We use the RK45 [7] solver in the Scipy library [37] for PFGM sampling. We report the FID score using the evaluation protocol in [5].

Table 6: FID/NFE on LSUN bedroom 256×256

	FID ↓	NFE ↓
StyleGAN [18]	**2.65**	**1**
DDPM [16]	6.86	1000
VE-SDE [33]	11.75	2000
PFGM w/ NCSN++ channel	17.01	134
PFGM w/ DDPM channel	13.66	122

Table 6 shows that PFGM has comparable performance with VE-SDE when using DDPM channel, while achieving around 15× acceleration. We observe that PFGM achieves a better FID score using the similar configuration in the DDPM model, and converges faster — 150k over the total 2.4M training iterations suggested in [33]. Remarkably, the VE-ODE baseline — the method most comparable to ours — only produces noisy samples on this dataset. It suggests that PFGM is able to scale up to high resolution images when using advanced architectures. We also compare with the number reported in [16] using similar architecture. Note that DDPM requires 1000 NFE during sampling, and doesn't possess invertibility compared to flow models.

D.2 Results on NCSNv2 Architecture

In this section, we demonstrate the image generation on CIFAR-10 and CelebA 64×64, using NCSNv2 architecture [32], which is the predecessor of NCSN++ and DDPM++ [33] and has smaller capacity. Since the VE/VP-ODE has poor performance (FID greater than 90), with the RK45 solver, we also apply the forward Euler method (**Euler**) with fixed number of steps. We explicitly name the sampler, with forward Euler method as predictor and Langevin dynamics as corrector, as **Euler w/ corrector**. For Euler w/ corrector in VE/VP-ODE, we use the probability flow ODE (reverse-time ODE) as the predictor and the Langevin dynamics (MCMC) as the corrector. We borrow all the hyper-parameters from [33] except for the signal-to-noise ratio. We empirically observe the new configurations in Table 7 give better results on the NCSNv2 architecture.

To accommodate the extra dimension z on NCSNv2, we concatenate the image with an additional constant channel with value z and thus the first convolution layer takes in four input channels. We also add an additional output channel to the final convolution layer and take the global average pooling of this channel to obtain the direction on z.

Table 7: Signal-to-noise ratio of different dataset-method pairs

Dataset-Method	CIFAR-10 - VE	CIFAR-10 - VP	CelebA - VE	CelebA - VP
signal-to-noise ratio	0.16	0.27	0.12	0.27

D.2.1 CIFAR-10

Table 8 reports the image quality measured by Inception/FID scores and the inference speed measured by NFE on CIFAR-10. We show that PFGM with the RK45 solver has competitive FID/Inception scores with the Langevin dynamics, which was the best model on the NCSNv2 architecture before, and requires $10\times$ less NFE. In addition, PFGM performs better than all the other ODE samplers. Our method is more tolerant of sampling error. Among the compared ODEs, our backward ODE (Eq. (6)) is the only one that successfully generates high quality samples while the VE/VP-ODE fail w/o the Langevin dynamics corrector. The backward ODE still beats the baselines w/ corrector.

Table 8: CIFAR-10 sample quality (FID, Inception) and number of function evaluation (NFE). All the methods below the *NCSNv2 backbone* separator use the NCSNv2 [32] network architecture as the backbone.

	Inception ↑	FID ↓	NFE ↓
PixelCNN [36]	4.60	65.93	1024
IGEBM [8]	6.02	40.58	60
WGAN-GP [12]	7.86 ± .07	36.4	1
SNGAN [26]	8.22 ± .05	21.7	1
NCSN [31]	**8.87 ± .12**	25.32	1001
NCSNv2 backbone			
Langevin dynamics [32]	8.40 ± .07	**10.87**	1161
VE-SDE [33]	8.23 ± .02	10.94	1000
VP-SDE [33]	6.85 ± .01	44.05	1000
VE-ODE (Euler w/ corrector)	8.05 ± .03	11.33	1000
VP-ODE (Euler w/ corrector)	7.33 ± .07	37.74	1000
PFGM (Euler)	8.00 ± .09	11.78	200
PFGM (RK45)	8.30 ± .05	11.22	**118**

D.2.2 CelebA

In Table 9, we report the quality of images generated by models trained on CelebA 64×64, as measured by the FID scores, and the sampling speed, as measured by NFE. As shown in Table 9, PFGM achieves best FID scores than all the baselines on CelebA dataset, while accelerating the

inference speed around 20×. Remarkably, PFGM outperforms the Langevin dynamics and reverse-time SDE samplers, which are usually considered better than their deterministic counterparts.

Remark: On the FID scores on CelebA 64×64 One interesting observation is that the samples of PFGM (RK45) (Fig. 10(b)) contain more obvious artifacts than Langevin dynamics (Fig. 10(a)), although PFGM has a lower FID score on the same architecture. We hypothesize that the diversity of samples has larger effects on the FID scores than the artifacts. As shown in Fig. 10(a) and Fig. 10(b), samples generated by PFGM have more diverse background colors and hair colors than samples of Langevin dynamics. In addition, we evaluate the performance of PFGM on the DDPM++ architecture. We show that the FID score can be further reduced to 4.04 using the more advanced DDPM++ architecture. By examining the generated samples of PFGM on DDPM++ (Fig. 14), we observe that the samples are diverse and exhibit fewer artifacts than PFGM on NCSNv2. It suggests that by using a more powerful architecture like DDPM++, we can remove the artifacts while retaining the diversity in PFGM.

(a) Langevin dynamics [31]

(b) PFGM (RK45)

Figure 10: Uncurated samples from Langevin dynamics [31] and PFGM (RK45), both using the NCSNv2 architecture.

Table 9: FID/NFE on CelebA 64×64

	FID ↓	NFE ↓
NCSN [31]	26.89	1001
NCSNv2 backbone		
Langevin dynamics [32]	10.23	2501
VE-SDE [33]	8.15	1000
VP-SDE [33]	34.52	1000
VE-ODE (Euler w/ corrector)	8.30	200
VP-ODE (Euler w/ corrector)	41.81	200
PFGM (Euler)	7.85	**100**
PFGM (RK45)	7.93	110

D.3 Wall-clock Sampling Time

The main bottleneck of sampling time in each ODE step is the function evaluation of the neural network. Hence, for different ODE equations using similar neural network architectures, their inference times per ODE step are approximately the same.

We implement PFGM on the NCSNv2 [32], DDPM++ [33], and DDPM++ deep [33] architectures, with sight modifications to account for the extra dimension z. In Table 10, we report the sampling time per ODE step method with the DDPM++ backbone, as well as the total sampling time. We measure the sampling time of generating a batch of 1000 images on CIFAR-10. We compare PFGM, VP/sub-VP ODEs using the RK45 solver. As a reference, we also report the results of VP-SDE using the predictor-corrector sampler [33]. All the numbers are produced on a single NVIDIA A100 GPU.

Table 10: Wall-clock sampling time (second)

Method	PFGM	VP-ODE	sub-VP-ODE	VP-SDE (PC)
NFE	110	134	146	1000
Wall-clock time per step	0.526	0.522	0.520	0.491
Total wall-clock time	57.81	69.97	75.92	490.65

As expected, ODEs using similar architectures and the same solver have nearly the same wall-clock time per ODE step. The table also shows that PFGM achieves the smallest total wall-clock sampling time.

D.4 Image Interpolations

The invertibility of the ODE in PFGM enables the interpolations between pairs of images. As shown in Fig. 11, we adopt the spherical interpolations between the latent representations of the images in the first and last column.

Figure 11: Interpolation on CelebA 64×64 by PFGM

D.5 Temperature Scaling

To demonstrate more utilities of the meaningful latent space of PFGM, we include the experiments of temperature scaling on CelebA 64×64 dataset. We linearly increase the norm of latent codes from 1000 to 6000 to get the samples in Fig. 12.

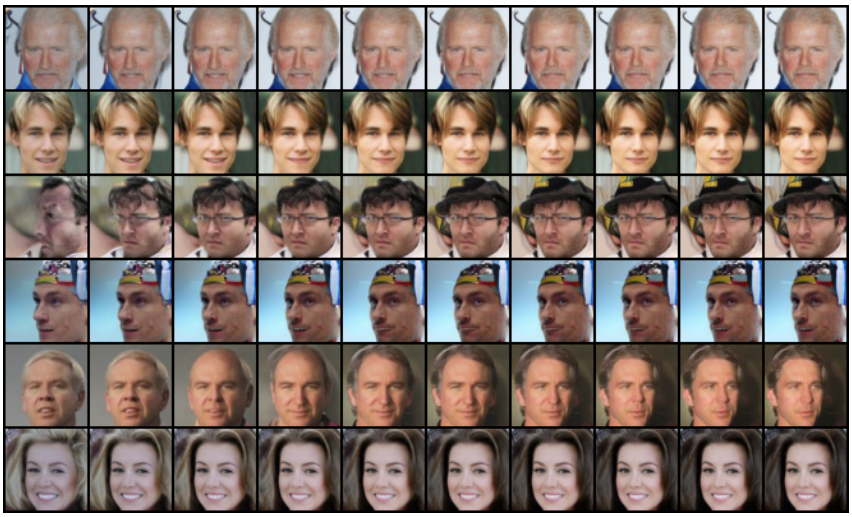


Figure 12: Temperature scaling on CelebA 64×64 by PFGM

E Extended Examples

We provide extended samples from PFGM on CIFAR-10 (Fig. 13), CelebA 64×64 (Fig. 14) and LSUN bedroom 256×256 (Fig. 15) datasets.

F Physical Interpretation of the ODEs in PFGM

In Section 2, in order to move the particles along the electric lines, we set the time derivative of x to the Poisson field $\mathbf{E}(x)$:

$$[q = 1, \text{forward ODE}] \quad \frac{d\mathbf{x}}{dt} = \mathbf{E}(\mathbf{x}), \quad [q = -1, \text{backward ODE}] \quad \frac{d\mathbf{x}}{dt} = -\mathbf{E}(\mathbf{x}) \tag{35}$$

We give the interpretation of the ODEs from a physical perspective. Newton's law implies that the external force is proportional to the acceleration of the particle. In the overdamped limit, e.g., when the particle is moving in honey, the external force is instead proportional to the velocity of the particle, making the equation of motion a first-order ODE. Denoting the viscosity of the fluid as γ, the dynamics of the particle under the influence of the electric field of the source $\rho(\mathbf{x})$ is

$$m\frac{d^2\mathbf{x}}{dt^2} = -\gamma\frac{d\mathbf{x}}{dt} + q\mathbf{E}(\mathbf{x}),$$

which has an overdamped limit $\frac{d\mathbf{x}}{dt} = q\mathbf{E}(\mathbf{x})$ when we set $t \to \gamma t$ and $\gamma \to \infty$. In this case, a particle with mass $m = 1$ and charge $q = 1$ would follow the electric field with velocity equal to $\mathbf{E}$, justifying Eq. (35).

G Limitations and Future Directions

In Section 3.2 we discuss the training paradigm of PFGM, including the normalized Poisson field and the discretized forward ODE. There are several potential improvements. First, the normalized field on mini-batch is biased. In this paper, we directly alleviate the bias by using a larger training batch. However, it does not solve the problem fundamentally. Some potential directions are incorporating more physical tools: we can exploit renormalization to make the Poisson field well-behaved in near fields. Another possibility is to replace a point charge with a quantum particle, whose position uncertainty fills the empty space among nearest neighbor data samples and makes the data manifold smoother.

H Potential Social Impact

Generative models is a rapidly growing field of study with far-reaching implications for science and society. Our work proposes a new generative model that allows for high-quality samples, quick inference and adaptivity. Many downstream applications benefit from our PFGM models' powerful expressive capabilities, particularly those that need fast inference speed and good sample quality at the same time. The usage of these models might have both positive and negative outcomes depending on the downstream use case. For example, PFGM can be incorporated in producing good image/audio samples by the fast backward ODE. This, on the other hand, promotes *deepfake* technology and leads to social scams. Generative models are also brittle and susceptible to backdoor adversarial attacks on publicly available training data, causing unanticipated failure. Addressing the above concerns requires further research in providing robustness guarantees for generative models as well as close collaborations with researchers in socio-technical disciplines.

Figure 13: CIFAR-10 samples from PFGM (RK45)

Figure 14: CelebA 64×64 samples from PFGM (RK45, NCSNv2 architecture)

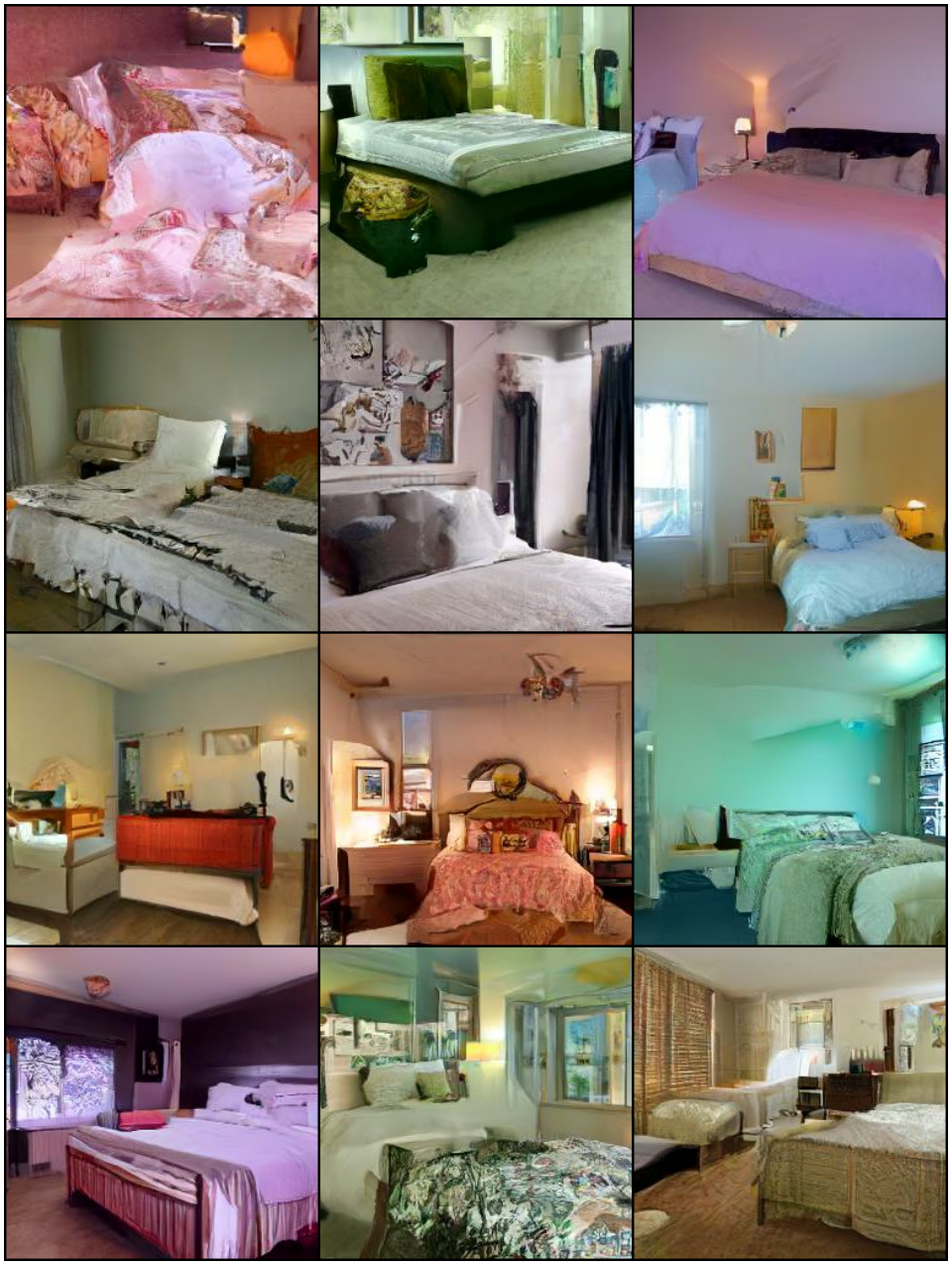


Figure 15: LSUN bedroom 256×256 samples from PFGM (RK45) using DDPM channel configuration.