Score-based Diffusion Models in Function Space

Jae Hyun Lim[*][†]   LIMJAEHY@MILA.QUEBEC
Université de Montréal

Nikola B. Kovachki[*]   NKOVACHKI@NVIDIA.COM
NVIDIA Corporation

Ricardo Baptista[*]   RSB@CALTECH.EDU
California Institute of Technology

Christopher Beckham   CHRISTOPHER.BECKHAM@MILA.QUEBEC
Polytechnique Montréal

Kamyar Azizzadenesheli   KAMYARA@NVIDIA.COM
NVIDIA Corporation

Jean Kossaifi   JKOSSAIFI@NVIDIA.COM
NVIDIA Corporation

Vikram Voleti   VIKRAM.VOLETI@UMONTREAL.CA
Université de Montréal

Jiaming Song   JIAMINGS@NVIDIA.COM
NVIDIA Corporation

Karsten Kreis   KKREIS@NVIDIA.COM
NVIDIA Corporation

Jan Kautz   JKAUTZ@NVIDIA.COM
NVIDIA Corporation

Christopher Pal   CHRISTOPHER.PAL@MILA.QUEBEC
Polytechnique Montréal & Canada CIFAR AI Chair

Arash Vahdat   AVAHDAT@NVIDIA.COM
NVIDIA Corporation

Anima Anandkumar   ANIMA@CALTECH.EDU
NVIDIA Corporation & California Institute of Technology

Editor: John Shawe-Taylor

Abstract

Diffusion models have recently emerged as a powerful framework for generative modeling. They consist of a forward process that perturbs input data with Gaussian white noise and a reverse process that learns a score function to generate samples by denoising. Despite their tremendous success, they are mostly formulated on finite-dimensional spaces, e.g., Euclidean, limiting their applications to many domains where the data has a functional form, such as in scientific computing and 3D geometric data analysis. This work introduces

---

[*]. Equal contribution.
[†]. Majority of the work was completed while the author was at NVIDIA.

a mathematically rigorous framework called *Denoising Diffusion Operators (DDOs)* for training diffusion models in function space. In DDOs, the forward process perturbs input functions gradually using a Gaussian process. The generative process is formulated by a function-valued annealed Langevin dynamic. Our approach requires an appropriate notion of the score for the perturbed data distribution, which we obtain by generalizing denoising score matching to function spaces that can be infinite-dimensional. We show that the corresponding discretized algorithm generates accurate samples at a fixed cost independent of the data resolution. We theoretically and numerically verify the applicability of our approach on a set of function-valued problems, including generating solutions to the Navier-Stokes equation viewed as the push-forward distribution of forcings from a Gaussian Random Field (GRF), as well as volcano InSAR and MNIST-SDF.[1]

Keywords: Diffusion models, Score matching, Generative models, Operator learning, Function spaces

1. Introduction

Diffusion models (DMs) (Song et al., 2020; Ho et al., 2020; Sohl-Dickstein et al., 2015) have appeared as a highly successful generative approach for various domains, including images (Saharia et al., 2022), 3D data (Poole et al., 2023), audio (Kong et al., 2021), video (Voleti et al., 2022a), machine learning security (Nie et al., 2022), natural language (Li et al., 2022a), proteins (Wu et al., 2024), and molecules (Xu et al., 2022). These models consist of two processes: A forward diffusion process that corrupts input data by gradually adding white noise and a reverse generative process that proceeds by iterative denoising.

Typically, DMs operate on a finite-dimensional space, e.g. $\mathbb{R}^n$, limiting their application in domains where the data is represented by infinite-dimensional objects, e.g. continuous functions. For example, in weather forecasting, data samples are functions of temperature, pressure, and wind, defined on the surface of the globe (Pathak et al., 2022). This also extends to seismology, geophysics, oceanography, aerodynamic vehicle design, and fluid dynamics, where we interact with functional data governed by partial differential equations (PDE) (Yang et al., 2021; Wen et al., 2023). Additionally, in 3D vision and graphics applications, data is represented as functions in the form of radiance fields (Mildenhall et al., 2021) or signed distance functions (SDF) (Park et al., 2019).

Recent attempts at applying DMs to functional data can be grouped into two categories: **(i)** the application of established algorithms on a discretization of functional data on $\mathbb{R}^n$ i.e. conditioning on point-wise values. While this approach can be made to work well at a fixed discretization, the models do not immediately transfer to variable discretizations of the data, and will not scale to higher resolutions (Dutordoir et al., 2023; Zhou et al., 2021), **(ii)** the mapping of input functions to a finite-dimensional latent space and modeling the latent embedding of the data with DMs (Dupont et al., 2022a; Phillips et al., 2022; Hui et al., 2022; Bautista et al., 2022; Chou et al., 2023). Such approaches rely on efficient transformations of the data into compactly representable spaces, which limits their general applicability and are not guaranteed to be discretization-independent/convergent (Kovachki et al., 2023).

The recently proposed infinite-dimensional diffusion model in Kerrigan et al. (2023) is closely related to our work. They consider a Gaussian noise corruption process in Hilbert space and derive a loss function to approximate the conditional mean of the reverse process.

---

1. The code for this project is publicly available at https://github.com/lim0606/ddo
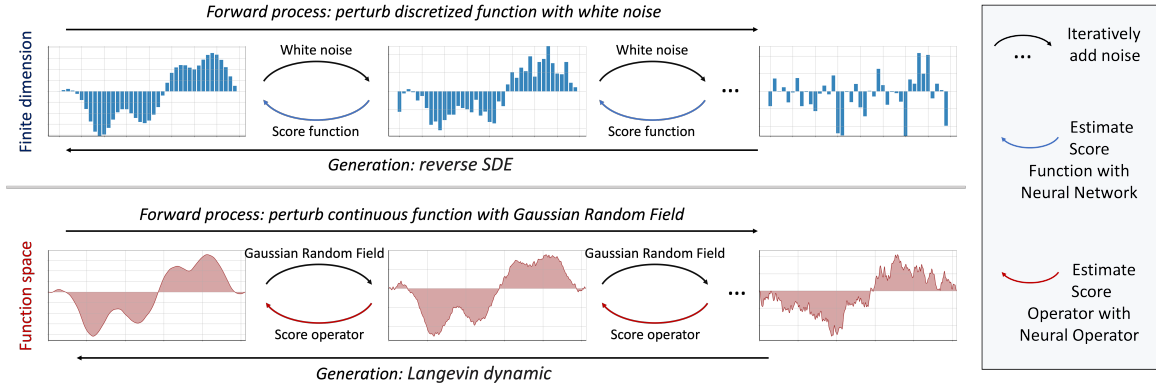
Figure 1: **Overview of our approach**. While in the finite-dimensional case, inputs are discretized, we work directly in function space, on continuous inputs, here 1D functions on $\mathbb{R}$. Noise is first added to the training samples during the forward process. A Neural Operator is used to estimate a score operator (Sec. 4) by minimizing the simplified loss in Eq. Equation 12. Samples are generated using Langevin dynamics (Sec. 4.4). Using structured noise enables efficient learning in function space while white noise does not as the model capacity required grows with the resolution.

While the loss function is formulated using infinite-dimensional measures, the difference between the true and approximate means does not satisfy the strict range conditions that are required to have non-singular measures, and thus yields a loss that is almost surely infinite. Numerically, this effect can only be seen through progressive grid refinement which the work does not consider. For further discussion, see Appendix H.

Developing a diffusion-based generative framework for functions requires solving several technical challenges. First, instead of the commonly used Gaussian white noise, a new function-valued corruption process must be introduced to gradually map the data functions into random functions. Second, sample generation requires an appropriate notion of the score since infinite-dimensional distributions do not have standard probability density functions (pdf). Finally, approximating the score requires both careful analysis and generalization of finite-dimensional techniques in order to obtain a well-defined optimization problem as well as approximation architectures that are consistent as mappings between function spaces.

In our approach, we introduce a rigorous framework termed denoising diffusion operators (DDOs) that addresses these challenges. DDOs use a Hilbert space-valued Gaussian process to perturb the input data. To define an appropriate notion of the score, we first consider densities with respect to a Gaussian measure (as opposed to the Lebesgue measure). The derivative of this density for certain perturbations of the Gaussian measure defines the score *operator*. To approximate this score in practice, we generalize the denoising score matching objective of Vincent (2011) to our setting, and show how samples can be generated using Langevin dynamics with a learned score operator.

For learning the score, we utilize the neural operators (Li et al., 2020b,a; Kovachki et al., 2023), which provide a consistent architecture in function space. We theoretically prove that approximating the score operator using neural operators is feasible. Figure 1 provides an overview of our approach. By working directly in the function space and discretizing only

later for the purposes of computation, we obtain scalable and discretization-independent algorithms for generative models in function spaces.

Our primary contributions are summarized below:

1. We develop a mathematically rigorous framework for denoising score matching with function-valued data called DDO by formulating and extending all necessary theory to the abstract Hilbert space setting.

2. We propose a diffusion model for incrementally sampling from the data distribution by discretizing an infinite-dimensional Langevin equation with a hierarchy of noise corruption Gaussian processes, generalizing several popular finite-dimensional frameworks.

3. We empirically show DDO learns distributions of function-valued data on various datasets, including generating solutions to the Navier-Stokes equation viewed as the push-forward distribution of forcings from a Gaussian Random Field (GRF), as well as volcano Interferometric Synthetic Aperture Radar (InSAR) (Rosen et al., 2012) and MNIST-SDF (Sitzmann et al., 2020).

4. We empirically verify DDO's invariance to spatial discretization with fixed model capacity, and demonstrate accurate sample generation of a non-Gaussian distribution from the pushforward of random forcings from a GRF under the Navier-Stokes solution operator.

2. Related Works

Our approach is broadly related to generative models formulated directly in function space instead of finite-dimensional Euclidean space (Rahman et al., 2022). Approaches for dealing with functional data include Gaussian processes (Rasmussen, 2004), and neural operators (Li et al., 2020b,a; Nelsen and Stuart, 2021). These methods aim to define deep learning models in function spaces, generalizing traditional neural networks.

In the context of generative diffusion models, this complication enters the model complexity and the number of time steps that typically need to grow with the data dimension. To improve sample quality and reduce the cost of sample generation in high dimensions, yet finite, several methods propose to use diffusion models in transformed spaces. These include latent spaces (Vahdat et al., 2021), hierarchically defined subspaces (Jing et al., 2022), spectral decompositions (Phillips et al., 2022), and extend to multi-scale wavelet transformations (Guth et al., 2022). Compared to score-based models operating in the original domain, the latter approach shows that the time complexity (i.e., the number of time steps required to achieve a fixed error) grows linearly with the image dimension. However, these models are not formulated in an infinite-dimensional space.

Neural Processes (NP) (Garnelo et al., 2018; Kim et al., 2019; Bruinsma et al., 2021) aim to model distributions consistent with arbitrary discretizations, and Dutordoir et al. (2023) have examined their extension to Neural Diffusion Processes (NDP). While the NP framework can process arbitrary sets of inputs, they inherit the limitations of using finite-dimensional latent variables; thus, consistency breaks in practice as the resolution grows. Moreover, the induced model distributions in NDP do not exist in function space due to independent noise in the noise process.

An earlier attempt to learn measures on function spaces deploys sequences of delta functions to fully memorize the data points (Craswell, 1965). Such a method is based on pure memorization and ignores possible underlying structures of the data measure. Kernel density estimation was proposed as a heuristic approach in infinite dimensional spaces (Dabo-Niang, 2004), though requires smoothness, extra regularity, and continuity with respect to an unspecified measure (Dabo-Niang, 2004). Alternative methods treat a discretized function as a point cloud and aim to maximize the likelihood of the point values (Garnelo et al., 2018), similarly to NPs.

Leveraging neural operators, Rahman et al. (2022) propose the generative adversarial neural operator (GANO) for learning function data distribution. As such, it enables learning of the distribution in function spaces through learning a mapping from infinite dimension spaces of Gaussian Random Fields (GRFs) to distribution in function spaces data. However, GANO inherently suffers from the major drawbacks of adversarial training, such as limited stability, optimization, and flexibility, as pointed out in prior works (Arjovsky and Bottou, 2017; Lin et al., 2018; Song and Ermon, 2019; Berard et al., 2020).

The use of GRFs in denoising diffusion models has been discussed but is yet to be explored in the domain of function spaces (Voleti et al., 2022b). The generative adversarial neural network framework (Goodfellow et al., 2020) was recently used in conjunction with implicit neural network representations of data (Dupont et al., 2022b; Anokhin et al., 2021; Skorokhodov et al., 2021; Chen et al., 2021). These methods are not discretization invariant and fail as the discretization of the data changes (Rahman et al., 2022). Dupont et al. (2022a) embeds discretized data in function space using implicit neural network representations, but it still inherits the drawbacks of using finite-dimensional latent spaces to encode infinite-dimensional data.

Recently, several continuous-time diffusion models in function space have also been proposed by Pidstrigach et al. (2024); Baldassari et al. (2023); Hagemann et al. (2023). These works define a forward and backward process by a pair of stochastic differential equations (SDEs) where the score operator is given as a conditional expectation, depending on the forward process. Our work offers an alternative viewpoint with the score defined as a logarithmic derivative of a perturbed measure and sampling done by a Langevin process and its annealed version. This allows us to make clearly interpretable assumptions on the data measure that are needed to guarantee convergence in the infinite-dimensional setting and furthermore allows us to study the interplay between the regularity of noise and the data.

3. Background: Denoising Score Matching in Finite Dimensions

Historically, *score matching* refers to the notion of approximating the score (i.e., the logarithmic derivative) of some unknown or computationally intractable distribution for the purposes of sampling, testing, or density estimation. Let $p \colon \mathbb{R}^d \to \mathbb{R}$ denote the pdf of a d-dimensional distribution and let $s_\theta \colon \mathbb{R}^d \to \mathbb{R}^d$ be a parametric mapping with parameters $\theta \in \mathbb{R}^m$. Ideally, score matching aims to solve

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{x \sim p(x)} \|s_\theta(x) - \nabla \log p(x)\|_2^2. \tag{1}$$

In many applications, we are only given samples from p, but do not know its analytic form. Therefore, solving Equation 1 is intractable. Using integration by parts on the objective, Hyvärinen (2005) showed that the minimizer of Equation 1 can be found by optimizing

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{x \sim p(x)} \left[\text{Tr}\big(\nabla s_\theta(x)\big) + \|s_\theta(x)\|_2^2 \right]. \tag{2}$$

Remarkably, the objective in Equation 2 can be minimized using a Monte-Carlo approximation to the expectation. It was later noted in Vincent (2011) that, up to a perturbation of the data distribution, the optimization problem is equivalent to *denoising score matching* where the objective depends on the analytically tractable score of the conditional perturbed distribution and no derivatives of the approximating function. In particular, for a Gaussian perturbation of variance σ^2, Equation 1 is equivalent to optimizing

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{\eta \sim N(0,\sigma^2 I)} \mathbb{E}_{x \sim p(x)} \left\| \frac{\eta}{\sigma^2} + s_\theta(x + \eta) \right\|_2^2, \tag{3}$$

where s_θ is now an approximation to the score of the perturbed distribution. Since Equation 3 does not require knowledge of p or computation of any derivatives, denoising score matching is attractive for problems in high dimensions where computing derivatives is costly. Furthermore, it is argued in Song and Ermon (2019), that for many practical applications, for example, photorealistic image generation, p is supported on a lower dimensional manifold and thus approximating the score on the ambient space can be unstable. Thus perturbing the data distribution gives *both* a more computationally tractable optimization problem and acts as a regularizer by spreading the support of p to the entire space.

We build on this framework by generalizing the notion of score and denoising score matching to infinite dimensions. By working directly in the infinite-dimensional setting, we derive a methodology that is consistent and generalizable across different discretizations of the data.

4. Denoising Diffusion Operators (DDO)

We introduce DDO to perform denoising score matching in function space. We work on an infinite-dimensional, real, separable Hilbert space $(H, \langle \cdot, \cdot \rangle, \| \cdot \|)$ with the Borel σ-algebra of measurable sets denoted $\mathcal{B}(H)^2$. Since there is no Lebesgue measure in infinite dimensions, there is no standard notion of a probability density; we therefore adopt the more general, measure-theoretic notation to introduce our setting. We denote by μ a probability measure on $\mathcal{B}(H)$ which we will call our *data measure*. In particular, we assume to have a dataset of samples $\{u_j\}_{j=1}^N$ where $u_j \sim \mu$ are i.i.d. random variables. These samples are considered to be infinite-dimensional objects, i.e. functions or infinite sequences, before any finite-dimensional discretization is done for the purposes of computation.

For a corruption process, we consider additive Gaussian perturbations to the data in the form of function-valued GRF perturbations. This choice is motivated by the availability of analytical results related to Gaussian measures, the ease and efficiency of sampling Gaussians in infinite dimensions by means of the Karhunen-Loéve expansion (Lord et al., 2014) (see

---

2. While a more general formulation on Banach or even locally convex spaces is possible, explicit computations for Gaussian measures on Hilbert spaces are more readily available and thus we consider this setting.

Appendix G), and the plethora of empirically successful results for denoising score matching with Gaussians in finite dimensions (Song et al., 2021; Ho et al., 2020). We employ the centered Gaussian measure on H denoted by $\mu_0 = N(0, C)$ with a covariance operator $C : H \to H$ to be self-adjoint, non-negative, and trace-class (nuclear). Indeed, these conditions on C are necessary and sufficient for μ_0 to be Gaussian on H (Da Prato et al., 1992). We note that since trace-class implies compact, the identity covariance operator is ruled-out as H is infinite-dimensional. In particular, white noise does not live in H but must rather be defined on a larger space (Da Prato et al., 1992). We show empirically that by working with noise defined on H our method remains discretizationally invariant with respect to the data. On the other hand, working with white noise breaks this property precisely because white noise samples are not regular compared to the elements of H.

4.1 Denoising score matching on function spaces

We consider the perturbation to the data samples

$$v = u + \eta, \qquad u \sim \mu, \ \eta \sim \mu_0, \tag{4}$$

with $u \perp \eta$ and denote by ν the probability measure induced by the random variable v i.e. the convolution $\nu = \mu * \mu_0$; see Appendix B.1 for more details. We show in Lemma 10, that when the noise η is small in an approximate sense, μ and ν are close as measures in the Wasserstein metric. It is therefore reasonable to approximate ν instead of μ as is done in denoising score matching since in the limit of vanishing noise, the two become identical.

We define the score of ν via an appropriate notion of density, which is defined with respect to a reference measure. In infinite dimensions, much work has been focused on studying densities defined with respect to Gaussian measures (as opposed to the Lebesgue measure in finite dimensions) as doing so has natural applications in statistics, inverse problems, and quantum field theory (Ghosal and van der Vaart, 2017; Stuart, 2010; Kupiainen, 2016). We also take this approach as it leads to a well-defined notion of the score that is analytically tractable and comes with an associated Langevin equation which can be solved to produce samples from ν. We choose the reference to be perturbing measure μ_0, which is natural in this setting since the conditional $v|u$ is Gaussian with the same covariance as μ_0. A density is then be obtained by the Radon–Nikodym Theorem under the assumption that ν is absolutely continuous with respect to μ_0, i.e., $\nu \ll \mu_0$ (Halmos, 1976).

To satisfy the absolute continuity condition with respect to Gaussian μ_0, it is reasonable to expect that the data measure μ must satisfy certain assumptions. The assumption we make is that $\mu(H_{\mu_0}) = 1$, i.e., μ is fully supported on the Cameron-Martin space of μ_0 that is denoted by $H_{\mu_0} := C^{1/2}(H)$. Cameron-Martin spaces play a crucial role in the theory of Gaussian measures as they are an invariant of the measure that gives it meaning outside the ambient space H (Bogachev, 2015). We remark that this assumption can make precise the "manifold hypothesis" in Song and Ermon (2019) that is used to justify the perturbation since H_{μ_0} is a proper subspace of H and, in fact, $\mu_0(H_{\mu_0}) = 0$; see Section 6 of Stuart (2010) for more details. In particular, data samples lie on a measure-zero set of the perturbing measure. The addition of noise thereby spreads out the samples to the whole space. We note that when this assumption is not satisfied, we can still apply our framework using a different form of the perturbation in Equation 4; see Section 4.2. Some example measure that satisfies $\mu(H_{\mu_0}) = 1$ are listed in Appendix F.

Under the condition above on the data perturbations, we can now state the following theorem.

Theorem 1 (Measure Equivalence) *The perturbed measure ν and the centered Gaussian μ_0 are equivalent in the sense of measures, which we denote by $\nu \sim \mu_0$.*

A more general statement and proof of this result are given in Appendix B.1. The importance of Theorem 1 is that it allows us to obtain a density. Indeed, it verifies the assumption of the Radon–Nikodym Theorem, which we apply to obtain a strictly positive density of ν with respect to μ_0. In particular, there exists a Borel measurable mapping $\Phi \colon H \to \mathbb{R}$ such that

$$\frac{d\nu}{d\mu_0}(w) = \exp\big(\Phi(w)\big), \qquad \mu_0\text{-a.s. } w \in H. \tag{5}$$

We will assume that Φ is Fréchet differentiable along the Cameron-Martin space H_{μ_0} which is itself a Hilbert space continuously embedded in H. This is a reasonable assumption since the vectors of differentiability of any Gaussian are precisely those in its Cameron-Martin space and ν is equivalent, in the sense of measures, to the Gaussian μ_0 (Bogachev, 2015). While in finite dimensions differentiability is always ensured since Gaussians have infinitely smooth density and convolutions preserve this regularity, in infinite dimensions, this need not always be the case. We therefore make it an assumption, however, the following example shows that it is true of any Gaussian data measure.

Example 1 *Suppose $\mu = N(0, Q)$ for some self-adjoint, non-negative, and trace-class operator $Q : H \to H$. It follows by non-negativity that*

$$\langle h, Ch \rangle \leq \langle h, (C + Q)h \rangle \qquad \forall\, h \in H.$$

Therefore by Lemma 6.15 in (Stuart, 2010), $C^{1/2}(H) \subseteq (C + Q)^{1/2}(H)$. From definition $\nu = N(0, C + Q)$ and, by Proposition 5.1.6. in (Bogachev, 2015), ν is differentiable along its Cameron-Martin space $(C + Q)^{1/2}(H)$. Therefore ν is differentiable along $C^{1/2}(H) = H_{\mu_0}$.

We define the score precisely as the Fréchet derivative of Φ in the direction of H_{μ_0} and denote it $D_{H_{\mu_0}}\Phi \colon H \to H_{\mu_0}^*$ where $H_{\mu_0}^*$ is the topological (continuous) dual of H_{μ_0}. In other words, the score of ν with respect to μ_0 is the Fréchet derivative of the logarithm of the density of ν with respect to μ_0,

$$D_{H_{\mu_0}}\Phi = D_{H_{\mu_0}} \log \frac{d\nu}{d\mu_0}. \tag{6}$$

We refer the reader to Chapter 5 in Bogachev (2015) for a general discussion of differentiability in infinite dimensions.

Having appropriately defined the score of ν, we can introduce a score matching objective. Let $G_\theta \colon H \to H_{\mu_0}^*$ be a parametric mapping with parameters $\theta \in \mathbb{R}^p$. We consider the learning problem

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{v \sim \nu} \| D_{H_{\mu_0}}\Phi(v) - G_\theta(v) \|_{H_{\mu_0}^*}^2. \tag{7}$$

Since $D_{H_{\mu_0}}\Phi$ is unknown to us, solving Equation 7 is computationally intractable.

To obtain a tractable problem, we generalize the conditioning theorem in Vincent (2011). Let us first notice that the measure induced by the conditional $v|u$ is the Gaussian $N(u, C) :=$

γ^u for μ-almost any $u \in H$. Since $\mu(H_{\mu_0}) = 1$, the Feldman–Hájek Theorem implies that $\gamma^u \sim \mu_0$ (Da Prato et al., 1992). In particular, we may compute explicitly that, for μ_0-almost any $w \in H$ and μ-almost any $u \in H_{\mu_0}$,

$$
\begin{aligned}
\frac{d\gamma^u}{d\mu_0}(w) &= \exp\left(\sum_{j=1}^{\infty} \lambda_j^{-1}\langle w, \varphi_j\rangle\langle u, \varphi_j\rangle - \frac{1}{2}\|C^{-1/2}u\|^2\right) \\
&:= \exp\big(\Psi(w; u)\big),
\end{aligned}
\tag{8}
$$

where $C\varphi_j = \lambda_j\varphi_j$ for $j \in \mathbb{N}$ is an eigendecomposition of C and $C^{-1/2}$ denotes the inverse of $C^{1/2}$ on H_{μ_0}, see Theorem 2.23 in Da Prato et al. (1992). The score of each conditional γ^u is given as the Fréchet derivative (in the first argument) of the potential $\Psi : H \times H_{\mu_0} \to \mathbb{R}$ in the direction of H_{μ_0}. We can now state the following (informal) theorem relating Equation 7 to the solution of a tractable problem.

Theorem 2 (Denoising Score Matching) *Under some integrability assumptions on $D_{H_{\mu_0}}\Phi$ and G_θ, the minimizers of Equation 7 are the same as the minimizers of*

$$
\min_{\theta} \mathbb{E}_{u\sim\mu}\mathbb{E}_{w\sim\gamma^u}\|D_{H_{\mu_0}}\Psi(w; u) - G_\theta(w)\|^2_{H^*_{\mu_0}}.
\tag{9}
$$

The more general statement (for a broader class of perturbations than Equation 4) and proof are given in Appendix B.2. Equation 9 gives us an infinite-dimensional analog of Equation 3, where we can compute $D_{H_{\mu_0}}\Psi(w; u)$ from Equation 8. That is,

$$
D_{H_{\mu_0}}\Psi(w; u) = \sum_{j=1}^{\infty} \lambda_j^{-1}\langle u, \varphi_j\rangle\varphi_j,
\tag{10}
$$

where we interpret Equation 10 as

$$
D_{H_{\mu_0}}\Psi(w; u)z = \sum_{j=1}^{\infty} \lambda_j^{-1}\langle z, \varphi_j\rangle\langle u, \varphi_j\rangle,
\tag{11}
$$

for any $z \in H_{\mu_0}$. Indeed, Lemma 12 shows that $D_{H_{\mu_0}}\Psi(w; u) \in H^*_{\mu_0}$ as defined by Equation 10.

Recall that our objective is to approximate $D_{H_{\mu_0}}\Phi$ by solving Equation 7, which we have shown is equivalent to Equation 9. Given such an approximation, we can then solve a Langevin equation with the learned score in order to obtain samples from ν. As we will show in the next section, this Langevin equation requires only knowledge of the $D_{H_{\mu_0}}\Phi$ projected onto H_{μ_0}. We can thus simplify the optimization problem in Equation 9 by considering the Reisz map $R : H^*_{\mu_0} \to H_{\mu_0}$, which is the canonical isometric isomorphism between the Hilbert spaces $H^*_{\mu_0}$ and H_{μ_0}. Using the isometric property, we find

$$
\|D_{H_{\mu_0}}\Psi(v; u) - G_\theta(v)\|^2_{H^*_{\mu_0}} = \|C^{-1/2}\big(u - RG_\theta(v)\big)\|^2
$$

by noting that R acts as C to elements of $H^*_{\mu_0}$ that are not in H_{μ_0} and using Equation 10. In particular, we have shown that minimizing Equation 9 is equivalent to minimizing

$$
\min_{\theta\in\mathbb{R}^p} \mathbb{E}_{u\sim\mu}\mathbb{E}_{\eta\sim\mu_0}\|C^{-1/2}\big(u - RG_\theta(u + \eta)\big)\|^2,
\tag{12}
$$

which is a de-noising problem pre-conditioned by $C^{-1/2}$. Note that Equation 12 is almost surely finite since $u - RG_\theta(w) \in H_{\mu_0}$ for any $w \in H$ by our assumption that $\mu(H_{\mu_0}) = 1$.

To that end, supposing that C is positive, then $\ker(C^{-1/2}) = \{0\}$. Therefore, optimizing Equation 12 is equivalent to optimizing

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta \sim \mu_0} \|u - RG_\theta(u + \eta)\|^2. \tag{13}$$

4.2 Smoothing Operators

When the assumption $\mu(H_{\mu_0}) = 1$ is not satisfy, we may consider a different form of the perturbation in Equation 4 to remove this regularity assumption. To that end, let $A \colon H \to H$ be a linear operator with the property that $A(H) \subseteq H_{\mu_0}$. Consider the data perturbation

$$v = Au + \eta, \qquad u \sim \mu, \ \eta \sim \mu_0. \tag{14}$$

We re-define the measures ν and γ^u appropriately according to Equation 14. Corollary 7 and the Feldman–Hájek Theorem imply that $\nu \sim \mu_0$ and $\gamma^u \sim \mu_0$ for μ-almost every $u \in H$. Therefore the results of the previous section hold with the mapping $u \mapsto Au$ implemented in all formulae. Crucially, the learning problem in Equation 13 becomes

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta \sim \mu_0} \|Au - RG_\theta(Au + \eta)\|^2. \tag{15}$$

Here A acts as a smoothing operator, bringing the data into a regular enough space for the required absolute continuity to hold. This makes mathematically precise diffusion models which use heat-dissipation or blurring as a forward operator (Rissanen et al., 2023; Hoogeboom and Salimans, 2022). We expand on this idea in Section 4.5 (See also Appendix I).

4.3 Approximation Theory

We have shown that the pre-conditioned score operator necessary for sampling is a non-linear mapping of the Hilbert space H into itself. We therefore need architectures which can approximate such mapping. We employ the neural operator framework of Kovachki et al. (2023). The following approximation result then follows by Theorems 11 and 13 in Kovachki et al. (2023) and the proof methods therein.

Theorem 3 (Score Approximation) *Let $D \subset \mathbb{R}^d$ be a bounded open set with Lipschitz boundary and consider $H = L^2(D; \mathbb{R})$. Suppose $\mathcal{I} \subset \mathbb{R}^n$ is compact and let $RD_{H_{\mu_t}}\Phi(\cdot, t) \colon H \to H$ be the pre-conditioned score of the perturbation in Equation 20 for each $t \in \mathcal{I}$. Suppose ν_t has a finite second-moment for each $t \in \mathcal{I}$ and the map $t \mapsto RD_{H_{\mu_t}}\Phi(\cdot, t)$ is uniformly continuous. Then, for any $\epsilon > 0$, there exists a number $p = p(\epsilon) \in \mathbb{N}$ and a parameter vector $\theta = \theta(\epsilon) \in \mathbb{R}^p$ such that a neural operator $G_\theta \colon H \times \mathcal{I} \to H$ satisfies*

$$\sup_{t \in \mathcal{I}} \mathbb{E}_{u \sim \nu_t} \|RD_{H_{\mu_t}}\Phi(u, t) - G_\theta(u, t)\|^2 < \epsilon.$$

Remark 4 *In Theorem 3, we crucially work in a setting where the map $t \mapsto RD_{H_{\mu_t}}\Phi(\cdot, t)$ is uniformly continuous and the score is well-defined for every $t \in \mathcal{I}$ i.e., the perturbing noise*

has a non-zero covariance uniformly across $\mathcal{I}$. This is important in avoiding the well-known singularity in the conditional score in the limit of vanishing noise. See Kim et al. (2022) for numerical methods for accurately approximating the score at small times for score-based models in finite dimensions.

Theorem 3 suggests that approximating score operators in infinite dimensions is feasible using neural operators. We demonstrate this numerically in the next section.

4.4 Langevin Dynamics

To sample from ν, we consider the infinite-dimensional, pre-conditioned, Langevin equation,

$$\frac{du}{dt} = -u + RD_{H_{\mu_0}}\Phi(u) + \sqrt{2}\frac{dW}{dt}, \quad u(0) = u_0 \tag{16}$$

for some $u_0 \in H$ where $RD_{H_{\mu_0}}\Phi : H \to H_{\mu_0}$ and W is a C-Wiener process (Da Prato et al., 1992). It is shown in Dashti and Stuart (2017) that, under appropriate boundedness assumptions on $RD_{H_{\mu_0}}\Phi$, equation Equation 16 has a unique strong solution with continuous paths and an invariant measure ν. In particular, samples from ν can be obtained as the long-time solutions of Equation 16. We will approximate Equation 16 by using the learned score RG_θ and discretizing in time using the Euler–Maruyama scheme with step-size $h > 0$. This gives us the update

$$u_{n+1} = u_n + h(RG_\theta(u_n) - u_n) + \sqrt{2h}\xi_n, \tag{17}$$

for any $n \in \mathbb{N}$ where $\xi_n \sim N(0, C)$ are i.i.d. random variables. Equation Equation 17 also suggests that instead of looking for the map RG_θ, we can re-parameterize and instead directly find the mapping $v \mapsto RG_\theta(v) - v$. Defining $F_\theta : H \to H$ by $F_\theta(v) = RG_\theta(v) - v$, optimizing Equation 15 is equivalent to

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu}\mathbb{E}_{\eta \sim \mu_0}\|\eta + F_\theta(Au + \eta)\|^2, \tag{18}$$

which simplifies the sampling update in Equation 17 to

$$u_{n+1} = u_n + hF_\theta(u_n) + \sqrt{2h}\xi_n. \tag{19}$$

Note that this re-parameterization is only valid when C is positive, otherwise Equation 12 and Equation 13 are not equivalent and $\text{Im}(F_\theta) = H$ while $\text{dom}(C^{-1/2}) = H_{\mu_0}$. In particular, for general C, we may optimize Equation 12 and sample with Equation 17, while for C positive, we can alternatively optimize Equation 18 and sample with Equation 19. The advantage of Equation 18 is that we can parameterize F_θ as an arbitrary $H \to H$ mapping without any restrictions on its range space. Furthermore current empirical evidence suggests that learning the noise from the signal instead of the signal from the noise yields better sample quality (Song and Ermon, 2020; Ho et al., 2020). Since C is a choice in our method that can be tuned, we always pick it positive and thus utilize this re-parametrization in our experiments.

We remark that we have only picked the Euler–Maruyama discretization of Equation 16 here for the sake of clarity in exposition. Many other choices such as Cotter et al. (2013) are possible; see Appendix D for details.

4.5 Multiple Noise Scales and Annealed Langevin Dynamics

As argued in Song and Ermon (2019), the mixing times of Langevin dynamics such as Equation 16 may be slow. It therefore of practical interest to consider multiple noise processes over different scales and thereby an annealing process for discretizing Equation 16. To that end, let $\mathcal{I}$ be some (possibly uncountable) index set and consider the data perturbations

$$v_t = A_t u + \eta_t, \qquad u \sim \mu, \ \eta_t \sim \mu_t \tag{20}$$

for a family of linear operators $\{A_t \colon H \to H\}_{t \in \mathcal{I}}$ and Gaussian measures $\{\mu_t\}_{t \in \mathcal{I}}$. Let ν_t be the measure for v_t.

Let us first consider the case $\mu(H_{\mu_0}) = 1$. Moreover, we assume $\mu_t = N(0, C_t)$ where $C_t = g(t)C$ with C as before, $A_t = f(t)I$ where I is the identity operator, and $f, g : \mathcal{I} \to \mathbb{R}$ are mappings bounded from above and below away from zero. Lemma 13 shows that $A_t(H) \subseteq H_{\mu_t}$ for all $t \in \mathcal{I}$ and therefore our previous theory holds. The choices $\mathcal{I} = [T]$ for some $T \in \mathbb{N}$ and $f(t) = 1$, $g(t) = \sigma_t^2$ for some sequence $0 < \sigma_T \le \cdots \le \sigma_1$ recovers the NCSN framework of Song and Ermon (2019). Similarly, let $0 < \beta_1 \le \cdots \le \beta_T < 1$ be some sequence and define $\alpha_t = \prod_{s=1}^{t}(1 - \beta_s)$. Then, setting $f(t) = \sqrt{\alpha_t}$ and $g(t) = 1 - \alpha_t$ recovers the DDPM framework of Ho et al. (2020); see Appendix E for more details on this connection. In particular, we generalize two widely used diffusion models in infinite-dimensions, up to the method selected for generating samples.

Let us now consider a case where we do not make assumptions on the data measure μ. For our previous theory to hold, we need that $A_t(H) \subseteq C_t^{1/2}(H)$ for all $t \in \mathcal{I}$. This can be accomplished with various choices of A_t, see Appendix C. For the current discussion, we will take $H = \dot{L}^2(\mathbb{T}^d; \mathbb{R})$ with μ_t as before where C has the form Equation 35. Let $\mathcal{I} = [T_0, T]$ for some $0 < T_0 < T < \infty$ and choose $A_t = f(t)e^{t\Delta}$ with the same boundedness assumptions on f, g. In particular, the family $\{A\}_{t \in \mathcal{I}}$ is a re-scaled subset of the semi-group associated to the solution operator of the heat equation (assuming f is continuous so that $t \mapsto A_t$ is continuous). Classical results on the heat equation show that for any $u \in H$, $A_t u \in \dot{H}^s(\mathbb{T}^d; \mathbb{R})$ for any $s > 0$ Evans (2010). In particular, by choosing $\alpha_1 > d/2$ in Equation 35 so that C is trace-class, we find that $A_t(H) \subset C_t^{1/2}(H) = \dot{H}^{\alpha_1}(\mathbb{T}^d; \mathbb{R})$. We have thus exhibited an infinite-dimensional generalization to the "inverse heat-dissipation" framework of Rissanen et al. (2023).

Algorithm 1 Annealed Langevin Dynamics

Input: F_θ, $u_0 \in H$, $\{\sigma_t\}_{t=1}^{T}$, $M \in \mathbb{N}$, $\epsilon > 0$
for $t = 1$ **to** T **do**
  $h_t = \epsilon \sigma_t^2 / \sigma_T^2$.
  for $n = 0$ **to** $M - 1$ **do**
    $\eta_n^{(t)} \sim N(0, C)$
    $u_{n+1} = u_n + h_t F_\theta(u_n, t) + \sqrt{2h_t}\eta_n^{(t)}$
  end for
  $u_0 = u_M$
end for

To sample $\{\nu_t\}_{t \in \mathcal{I}}$ and thus eventually sample ν_T, we can apply to the Euler-Maruyama scheme for each $t \in \mathcal{I}$ to obtain the iteration

$$u_{n+1} = u_n + h_t F(u_n, t) + \sqrt{2h_t} \eta_n^{(t)} \tag{21}$$

for any $n \in \mathbb{N}$, where $\eta_n^{(t)} \sim \mu_t$ form an i.i.d. sequence, with $h_t > 0$ and $t_0 \in \mathcal{I}$ are fixed. Here, $F : H \times \mathcal{I} \to H$ is a model defined by $F(u, t) = -u + R D_{H_{\mu_t}} \Phi(u, t)$ as discussed in Equation 19. For any $v \in H$, the iteration in Equation 21 starting with $u_0 = v$ at $t = t_0$ transforms v to an approximate sample of ν_{t_0}. We denote this sample by v_{t_0}. Now fix $t_1 \in \mathcal{I}$. We again run the iteration Equation 21 with $t = t_1$ and $u_0 = v_{t_0}$. This will transform v_{t_0} into an approximate sample from ν_{t_1} which we denote v_{t_1}. If $\mathcal{I} = \{1, \dots, T\}$ for some $T \in \mathbb{N}$ then repeating this process yields v_T, which is approximately distributed according to ν_T. Moreover, according to Lemma 10, v_T is approximately distributed according to our original data measure μ. We outline this annealing process in Algorithm 1.

4.6 Conditional Sampling

Finally, we demonstrate the straightforward extension of our approach to conditional simulation. We apply DDO to sample the conditional distribution $\mu(\cdot|y)$ for a parameter u supported on H given a relevant observation $y \in \mathbb{R}^m$ for inferring u. Specifically, we aim to solve a Bayesian inverse problem—modeling a posterior distribution—where the observation y is typically assumed to arise from the additive noise model $y = \mathcal{F}(u) + \eta$ where $\mathcal{F} : H \to \mathbb{R}^m$ is a forward operator and $\eta \in \mathbb{R}^m$ is a noise random variable that is independent of u. The noise model induces a likelihood function $\mu(y|u)$ that together with a prior measure $\mu(u)$ for the parameter yields the posterior measure from Bayes' rule as: $\mu(u|y) \propto \mu(y|u)\mu(u)$.

To extend the DDO framework to sample conditionally, we first consider the perturbations to the data samples in Equation 4 with the data sample drawn from the posterior distribution. That is,

$$v = u + \eta, \quad u \sim \mu(\cdot|y), \quad \eta \sim \mu_0.$$

The resulting random variable v has probability measure $\nu(\cdot|y) = \mu(\cdot|y) * \mu_0(\cdot)$ that depends on the observation y. Under the same assumptions as in Section 4.1, one can define the logarithmic derivative of $\nu(\cdot|y)$ denoted by $D_{H_{\mu_0}} \Phi(\cdot; y) = D_{H_{\mu_0}} \log \frac{d\nu(\cdot;y)}{d\mu_0}$. The following result shows that we can approximate the logarithmic derivatives of these measures (depending by y) using a parametric mapping $G_\theta : H \times \mathbb{R}^m \to H_{\mu_0}^*$ by solving a denoising score matching problem. The proof follows identically from the one for Theorem 2 and hence is omitted.

Theorem 5 *Under integrability assumptions on $D_{H_{\mu_0}} \Phi$ and G_θ, the minimizers of the score matching problem for conditional sampling*

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{y \sim \mu(y)} \mathbb{E}_{v \sim \nu(\cdot|y)} \|D_{H_{\mu_0}} \Phi(v) - G_\theta(v, y)\|_{H_{\mu_0}^*}^2. \tag{22}$$

are identical to the minimizers of the problem

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{y \sim \mu(y)} \mathbb{E}_{u \sim \mu(\cdot|y)} \mathbb{E}_{w \sim \gamma^u} \|D_{H_{\mu_0}} \Psi(w; u) - G_\theta(w, y)\|_{H_{\mu_0}^*}^2, \tag{23}$$

where $D_{H_{\mu_0}} \Psi$ has the form in Equation 10.

13

After identifying the parametric approximation to the logarithmic derivative of the conditional measure by solving Equation 23, one can use the resulting parametric map to construct a Langevin sampling algorithm whose stationary distribution is (approximately) $\nu(\cdot|y)$. Moreover, as discussed in Sections 4.2 and 4.5, we can introduce smoothing operators in the learning problem and consider multiple noise scales to sample our target posterior measure via a sequence of less noisy target distributions using annealed Langevin dynamics.

5. Numerical Experiments

In all examples, we use the Fourier neural operator (FNO) (Li et al., 2020a), U-shaped neural operator(UNO) (Rahman et al., 2023) as they are well-defined architecture for maps between Hilbert spaces Li et al. (2020a); Kovachki et al. (2021). The goal of our numerics is to showcase the simple message that by employing trace-class noise and a consistent architecture for function space data, we obtain dimension (i.e., resolution)-independent results, observed by varying the discretization of the data. All experiments are done by solving Equation 16 in a way similar to Song and Ermon (2019), generalized to function spaces; see Appendix D.
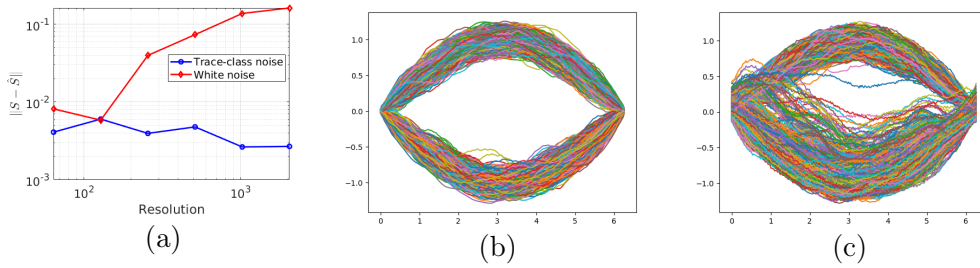

Figure 2: **Gaussian mixture** (Section 5.1): **(a)** Uniform-norm error in the average spectra of samples when using trace-class noise vs. white noise. **(b)** Generated samples at a resolution of 256 with trace-class noise and **(c)** with white noise.
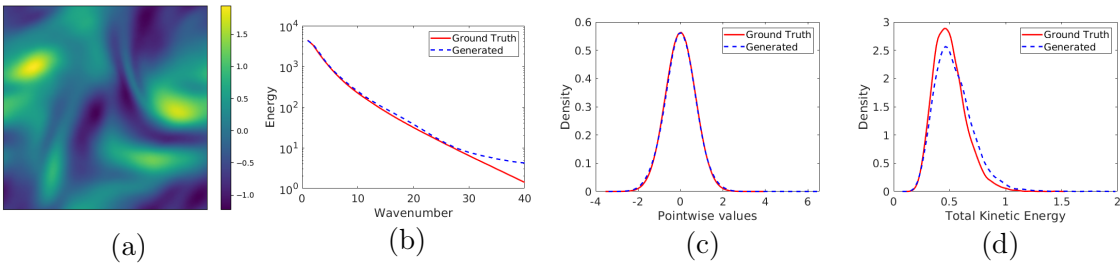

Figure 3: **Navier-Stokes** (Section 5.2): **(a)** Generated sample at the resolution 1024×1024 with a model trained at 128×128, **(b)** Spectrum, **(c)** Pointwise value density, **(d)** Total kinetic energy density of samples from the model vs. the data.

5.1 Gaussian Mixture

We consider a Gaussian mixture model by sampling a Gaussian random field (GRF) on the domain $(0, 2\pi)$ and assigning it one of two mean functions with a fixed probability. Details

on the precise construction can be found in Appendix J.1. We fix a FNO model architecture and train DDO on various discretizations of the data using either trace-class noise or white noise. In Figure 2(a) we compare the uniform (or sup) norm error in the spectrum of the true and generated data for the two types of noise. We see that while white noise achieves small errors at low resolutions, its error grows as we refine the resolution. On the other hand, trace-class noise achieves a consistent error across many resolutions. Indeed even at a resolution of 256, the trace-class noise model captures the right distribution in Figure 2(b), unlike the white noise model in Figure 2(c); see Appendix J.1 for further visualizations and Appendix I for an example that uses the smoothing operators described in Section 4.2. This is because as we refine the resolution, the model trained with white noise has to capture progressively higher frequency functions and thus it fails to do so with a fixed capacity model. Trace-class noise, on the other hand, has a convergent Fourier spectrum that the model can capture independently of the discretization. The white noise issue can be fixed by designing larger architectures and more sampling steps, but this yields a model where both the number of parameters and sampling steps need to increase with dimension. Therefore algorithms designed with white noise cannot be expected to scale to arbitrarily large resolutions.

5.2 Navier-Stokes

Next, we consider the vorticity form of the Navier-Stokes equation on the 2D-torus with a Reynolds number of 500. We develop a solver for this problem and solve it up to a fixed time for a fixed initial condition with different forcing functions generated by a GRF. The data distribution is therefore a pushforward of a Guassian under a non-linear map and is therefore non-Gaussian. Details are given in Appendix J.2. We train a DDO with FNO based model with data at a fixed 128×128 resolution and $L^2(\mathbb{T}^2; \mathbb{R})$ valued noise. We observe that the trained DDO accurately generate the function valued data learned from the underlying distribution. In Figure 3(b-d), we compare statistics relevant for turbulence analysis from the data and the samples from the model, verifying that we are able to capture the true distribution Li et al. (2022b). In Figure 3(a) we show a sample from the model generated at a 1024×1024 resolution without any re-training; more samples are visualized in Appendix J.2. In particular, our model generalizes to high resolutions at no extra cost, performing super-resolution natively. Such a method has powerful applications for learning the invariant measures of dissipative dynamical systems which can used for turbulance analysis and climate science Temam (1988).

5.3 Volcano Dataset

For the following experiments we use the volcano dataset originally proposed in GANO Rahman et al. (2022), and UNO as the base architecture. The volcano InSAR dataset consists of 4096 data points of spatial resolution 128×128, derived from raw interferograms produced from satellites covering the Long Valley Caldera near Mammoth Lakes, California, United States. Since the dataset consists of relatively few examples, we employ a light amount of data augmentation during training in the form of random horizontal and vertical flips. We present key elements of our loss function and architecture below and provide further details about these experiments in Appendix J.3, e.g. learning and hyperparameter details.

Instead of manually comparing histograms of these evaluation metrics to their respective statistics computed on the training set, here we quantitatively measure how close their histograms are by measuring the 1D Wasserstein distance between them. That is, we define:

$$w_{\text{var}} = W_1(\text{var}(\theta(\mathbf{u})), \text{var}(\theta(\tilde{\mathbf{u}}))) \tag{24}$$

$$w_{\text{skew}} = W_1(\text{skew}(\theta(\mathbf{u})), \text{skew}(\theta(\tilde{\mathbf{u}}))) \tag{25}$$

$$w_{\text{total}} = w_{\text{var}} + w_{\text{skew}} \tag{26}$$

where $\mathbf{u} = \{\mathbf{u}_i\}_{i=1}^N$ denotes the training set and conversely $\tilde{\mathbf{u}} = \{\tilde{\mathbf{u}}_j\}_{j=1}^M$ generated samples from the diffusion model. We set $M = 256$ for fast metric tracking since it is computationally expensive to generate many samples. During training, we periodically evaluate both evaluation metrics and keep track of checkpoints corresponding to the smallest values seen so far, then when we perform a final evaluation of the model we use the checkpoint corresponding to the smallest w_{total} seen so far.

For these experiments we sample from a 2D GRF based on the RBF kernel (Section J.3), and therefore an important hyperparameter to tune is γ, the smoothness of the noise.

When the best model has been selected, we re-compute the Wasserstein metrics using $M = 1024$ instead. While $M = 256$ was used during training, we find it did not show significant changes in the computed statistics.
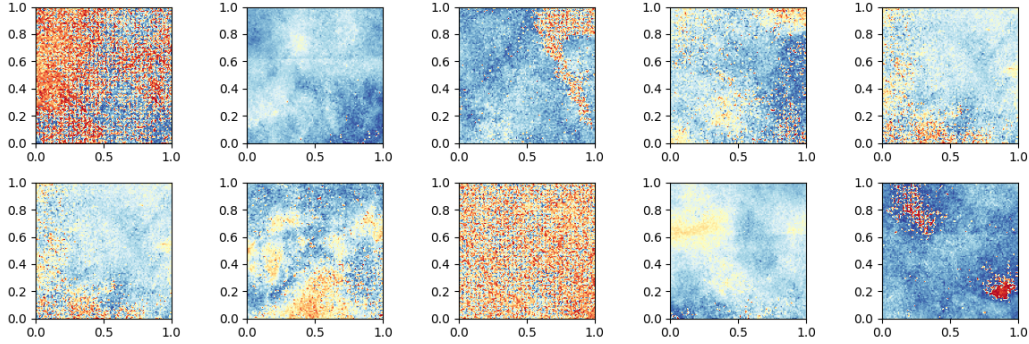
(a) Random samples from the best performing diffusion model.
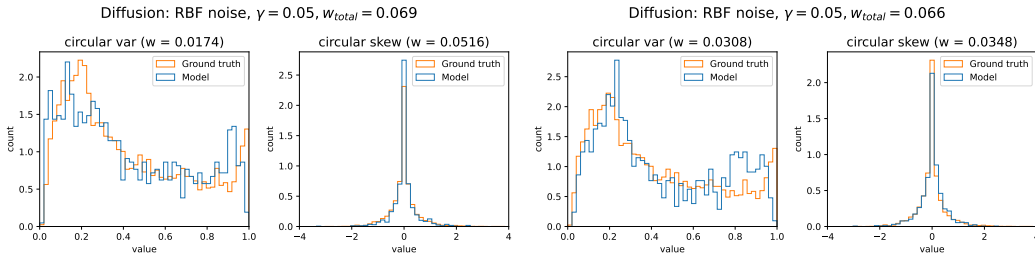
(b) Best performing diffusion model   (c) Best performing GANO model

Figure 4: **Results of Volcano dataset experiments** (Section 5.3): Samples from the best performing FNO diffusion model with an RBF scale of $\gamma = 0.05$. For both histograms, $M = 1024$ generated samples were used to compute skew and variance.

Results In Figure 4 we demonstrate samples and histograms produced by our best performing diffusion model, with smoothness parameter $\lambda = 0.05$. These are shown in Figures 4a and 4b respectively, and a reference GANO model is also shown in 4c. We can see that our model is able to accurately learn the ground truth function, as indicated by the histograms shown in Figure 4b. Due to the noisiness of this dataset we found that the best results were achieved with an RBF scale parameter of $\gamma = 0.05$ (see Section J.3), which corresponds to very rough levels of noise.

At generation time, in order to generate at twice the resolution we construct a meshgrid that is twice as granular as that used in training. For example, if s^2 is the original resolution then we compute a meshgrid $x \in [0,1]^{(s^2)^2 \times 2}$ and use that to sample RBF noise at twice the resolution. Concretely, we train DDO on a downsampled version of Volcano at 60×60 resolution and double the resolution at generation time to 120×120. In order to quantitatively evaluate this task we still compute skew and variance metrics as previously described (Equation 24), but now these are computed between the super-resolution samples and the original 120×120 resolution dataset. These results are shown in Figure 14, and we demonstrate results comparing independent Gaussian noise to different values of RBF smoothness γ used during training. As expected, independent noise performs abysmally (Figure 14a), and we also found that $\gamma = 0.05$ did not perform well (Figure 14b). However, smoother levels of RBF noise performed well, and the best results were achieved with $\gamma = 0.2$ (Figure 14d). This demonstrates the ability of our model to query the sampled function.

5.4 MNIST-SDF Dataset

Next, to demonstrate the efficacy of the proposed method on function generation in conjunction with images, we conduct experiments on MNIST-SDF (Sitzmann et al., 2020) and compare the proposed method to GANO (Rahman et al., 2022) and MultilevelDiff (Hagemann et al., 2023). GANO is an adversarial training-based function space generative model. As compared to other concurrent works on diffusion models designed for function spaces (Kerrigan et al., 2023; Pidstrigach et al., 2024; Baldassari et al., 2023), we selected MultilevelDiff since it was already tested on similar two-dimensional datasets, while other models have been limited to one-dimensional datasets. In this setting, (Bond-Taylor and Willcocks, 2024) is another a function space diffusion model which aims to model 2D image datasets. However, this model relies on frameworks tailored for high-fidelity image modeling, which includes performing diffusion in latent spaces, thereby making direct comparisons unfair. An indirect comparison with this model is discussed in Appendix K.

MNIST-SDF is a collection of 2D signed distance functions (SDFs), each of which is extracted by applying a distance transform to every image in the MNIST dataset. As compared to resizing finite-dimensional datasets for various resolutions, this conversion makes the dataset defined on function space, and allows us to consistently compute evaluation statistics like FID (Heusel et al., 2017) and precision-recall (Kynkäänniemi et al., 2019) metrics across different resolutions. Examples of the 2D SDFs are shown in Figure 5(a).

We aim to train models on 32×32-resolutions and evaluate the FID and precision-recall at different resolutions. Specifically for image datasets, we choose to upsample 32×32 resolution to a 64×64 resolution, since this ensures we can select the number of Fourier modes to represent the data or noise to be higher than the discretization. We will discuss the necessity

of the upsampling in the following section. The other experimental details, including the architecture and training procedure, are described in the Appendix J.4

As we find that the classifiers pre-trained for the evaluation metrics are more suitable for the original MNIST-like binary digits than 2D SDFs, we evaluate the metrics after generating binary masks by thresholding the sample SDFs to larger than 0. The masked 2D SDFs images are illustrated in Figure 5.(b). We follow the styleGAN3's evaluation protocol (Karras et al., 2021) for the FID and precision-recall.

(a) MNIST-SDF   (b) Masked
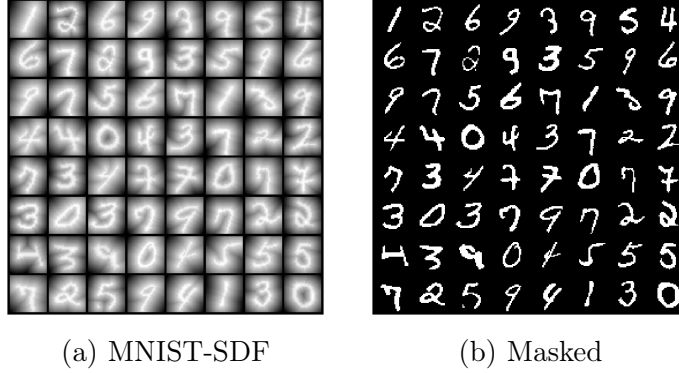
Figure 5: **MNIST-SDF Dataset.** (a) Samples of 2D SDFs in the datasets. A 2D SDF is generated by applying a distance transform to each image in the MNIST dataset. (b) Binary masks extracted by thresholding the sample SDFs where the value is larger than 0. Every generated samples from the models will be masked before running FID and precision-recall metrics.

Upsampling the Finite-dimensional Observations   Our theory suggests that any sample function in the target data distribution should be smoother than the samples of the noise distribution to satisfy $\mu(H_{\mu_0}) = 1$. This assumption implies that when we represent the data and noise in Fourier space, the Fourier bases required to describe all noise samples include the basis set of the data. This requirement leaves us some implementation constraints when dealing with the finite observations of the data or noises, especially when the number of bases representing a noise sample is larger than the discretization size. This is often the case since useful distributions like Gaussian measures are often obtained in d-dimensional observations while its basis set size is much larger than d. In this section, we will discuss how to address the constraints efficiently.

For most experimental scenarios, we assume that we can only access d-dimensional observations of the data. This means that we can treat the size of the data's basis set to be d. While the true size could be larger than d, we won't be able to model anything other than observations on the d basis. However, we can learn such discrete observations on the Fourier bases by discrete Fourier transforms and generate arbitrary discretization from them.

On the contrary to data, for some useful distributions in the infinite-dimensional space, the size of basis sets is often larger than the observation size d; for example, Gaussian measures. For Gaussian, there exists a positive integer $M < \infty$ such that the M-size of the basis set can represent all its samples. When we select M to be higher than d to model the d-dimensional data, the components on M-d number of bases won't be observed at the given

discretization. Thus, any model may fail to generate proper images at any resolution larger than d, as the unseen noise components will be introduced.

To address this, for a given d-dimensional observations, we propose upsampling to d' such that d' is large enough to M. For the 32×32-resolution observation of the MNIST-SDF, we choose d' to 64×64. For the upsampling of the finite observations, this paper follows the filtered upsampling implementation discussed in Karras et al. (2021). While the models are trained on 64×64-resolution, we use only 32 modes in the spectral convolutions (at the lowest level) of the architectures used in DDO, MultilevelDiff, and GANO, which enables them to generate images at the 32×32-resolution as well.

Note that one can truncate the modes of Gaussians up to d so that the unseen noise components won't be introduced. However, we find that such truncations often generate artifacts in super-resolution tasks; periodic waves are drawn, which are supposed to be straight lines.

	64×64			128×128			256×256		
	FID	Prec*	Rec†	FID	Prec	Rec	FID	Prec	Rec
GANO	3.41	**0.75**	0.63	13.05	0.68	0.50	23.89	0.60	0.32
MultilevelDiff	35.09	0.03	0.06	201.08	0.00	0.00	365.90	0.00	0.00
DDO (Ours)	**2.74**	0.73	**0.68**	**7.96**	**0.71**	**0.60**	**17.76**	**0.65**	**0.39**

*Precision. †Recall.

Table 1: Results of MNIST-SDF experiments.

Results Table 1 shows the FID and precision-recall metrics evaluated from learned models in the MNIST-SDF experiments. DDO outperforms both GANO and MultilevelDiff baselines, except for the precision at the training resolution (64×64). Moreover, DDO demonstrates higher recall at all resolutions, which is coherent with the general property of diffusion-based models, whose objectives are to minimize the KL divergence between the data distribution and the model. Note that such connections to the DDO's objective are briefly discussed in Lemma 10. Interestingly, MultilevelDiff exhibits a notable performance lag in comparison to the other two. The discussion related to this performance gap will be presented while examining the generated samples.

Figure 6 illustrates the generated samples at various resolutions produced by our DDO model and the baselines[3]. Visually, all models appear to achieve high-quality generation of all digits across all resolutions. However, in the case of MultilevelDiff, the variation in digit shapes and styles is noticeably lower. This tendency becomes more pronounced at higher resolutions (see Figure 16), which only leads to generating a few digits. This characteristic explains the poor FID scores and precision-recall metrics observed for MultilevelDiff above. In contrast, both DDO and GANO not only achieve high-quality generation across all resolutions but also maintain the variation in styles and digits observed at the training resolution.

On the other hand, the generated samples from both DDO and GANO show curved boundaries, which the MNIST-SDF dataset doesn't have. This artifact originates from the spectral convolution, as it cut off higher frequency components than its parameters'

---

3. Figure 16 illustrates additional samples at various resolutions produced by all models.

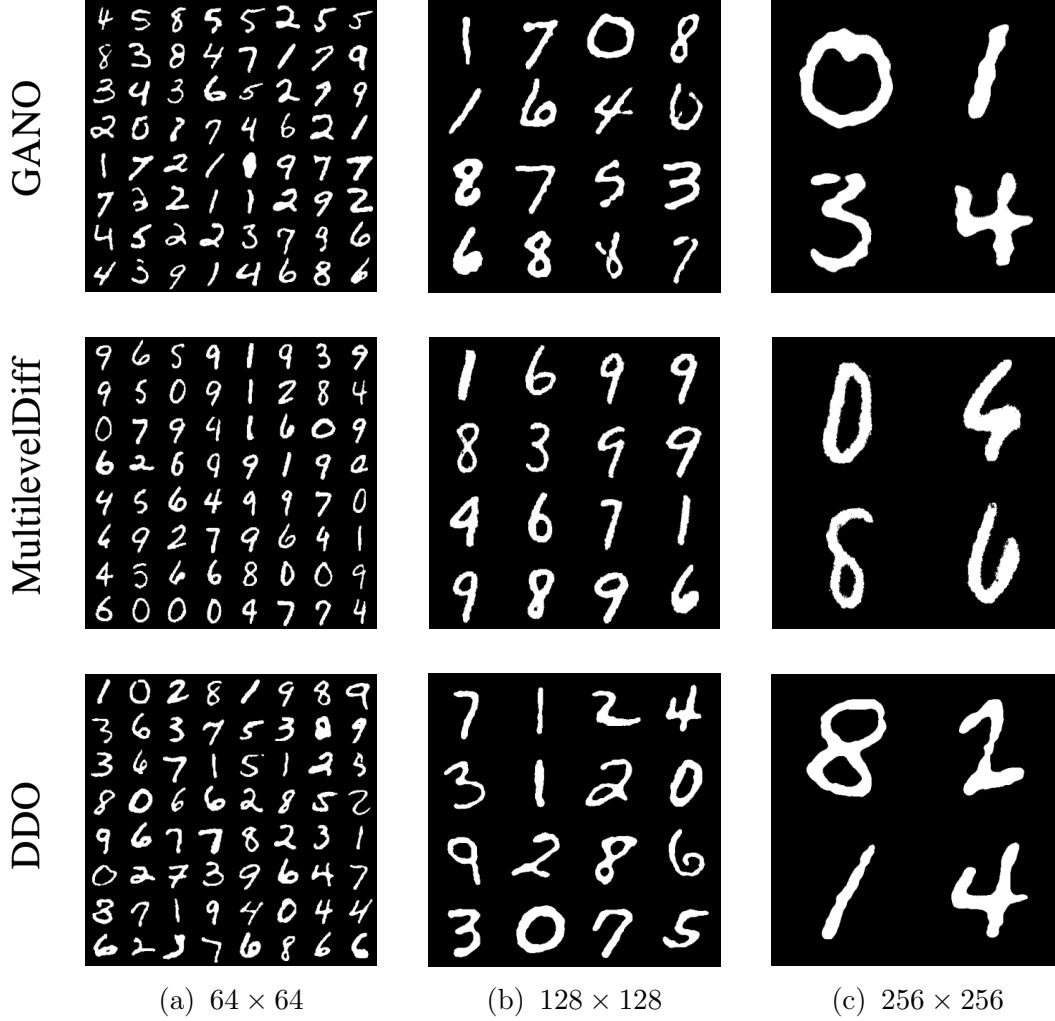(a) 64×64 (b) 128×128 (c) 256×256

Figure 6: **Generated Samples** (Section 5.4): Generated samples (masked) of the learned GANO, MultilevelDiff, and DDO models at various resolutions; (a) 64×64, (b) 128×128, and (c) 256×256-resolutions. All images are plotted according to their relative resolutions. All models are trained on 64×64-resolution images, which are upsampled from 32×32-resolution observations of 2D SDFs.

highest mode. Again, this results in the loss of some high-frequency components which would be necessary to represent arbitrary curved lines with no artifacts. This observation also emphasizes the importance of upsampling during training instead of truncating the noise. We leave addressing such artifacts for future works.

5.5 Darcy Flow Bayesian Inverse Problem

We apply our method to the geophysical inverse problem of recovering a subsurface permeability field from pointwise observations of the pressure at the surface, which is also known

as the Darcy flow inverse problem. We refer to Section 4.6 for the abstract formulation of such problems and how our framework can be applied to solve them. In this setting, the forward model is defined as the solution of the following elliptic partial differential equation (PDE) on the domain $\mathcal{D} = (0, 1)^2$ given by

$$-\nabla(a(s)\nabla p(s)) = 1, \quad s \in \mathcal{D} \tag{27a}$$
$$p(s) = 0, \quad s \in \partial\mathcal{D}, \tag{27b}$$

where $a(s) \in \mathbb{R}_+$ represents the strictly positive permeability and $p(s)$ is the pressure. We consider the parameter $u = \log(a)$ defining the log-permeability, which we recover from 64 observations on an 8×8 grid, i.e., $y(s_i) = p(s_i)$ for locations s_i in the interior of domain $\mathcal{D}$ on a regular grid. To invert for the solution, we consider a log-normal prior distribution for the log-permeability $u = \log(a) \sim \mathcal{N}(0, (-\Delta + \tau I)^{-2})$, where Δ is the Laplacian operator and we set $\tau = 9$. We assume the observations are corrupted with Gaussian observational noise $\eta \sim \mathcal{N}(0, \Sigma_m)$ where Σ_m is a diagonal matrix with entries $\mathbb{E}[p(s_i)^2]/5$, implying a signal-to-noise ratio of 5.

To generate data at a resolution of 64×64 for training, we solve the PDE on a regular grid of resolution 1024×1024 for each realization of the permeability field $a^i = \exp(u^i)$ for $u^i \sim \mu(\cdot)$ to obtain the observations $y^i = \mathcal{F}(u^i) + \eta^i$. This process yields pairs of observations (u^i, y^i) drawn from the joint measure for the parameter and observation. Finally, we downsample the high-resolution observation to a resolution of 64×64 for the training data.

To quantify the performance of the learned posteriors, we evaluate the relative errors of the sample mean and variance relative to those of the posterior obtained through a Markov chain Monte Carlo (MCMC) simulation based on a pre-conditioned Crank Nicolson method that is consistent in function space (Cotter et al., 2013); see the first row of Figure 8 for the MCMC results. The relative errors are defined by

$$\mathcal{E}_{\text{mean}} := \frac{\|m_{\text{MCMC}} - m_{\text{model}}\|_{L^2}^2}{\|m_{\text{MCMC}}\|_{L^2}^2} \quad \text{and} \quad \mathcal{E}_{\text{var}} := \frac{\|\sigma_{\text{MCMC}}^2 - \sigma_{\text{model}}^2\|_{L^2}^2}{\|\sigma_{\text{MCMC}}^2\|_{L^2}^2},$$

where $(m_{\text{MCMC}}, \sigma_{\text{MCMC}}^2)$ and $(m_{\text{model}}, \sigma_{\text{model}}^2)$ are the sample mean and variance pairs for MCMC and the model, respectively. To compute the sample means and variances, we use 10,000 samples generated by the model and 10,000 from MCMC. Similar to the MNIST-SDF experiments, we compare the proposed method to GANO and MultilevelDiff.

	64×64		128×128		256×256	
	$\mathcal{E}_{\text{mean}}^*$	$\mathcal{E}_{\text{var}}^\dagger$	$\mathcal{E}_{\text{mean}}$	$\mathcal{E}_{\text{var}}$	$\mathcal{E}_{\text{mean}}$	$\mathcal{E}_{\text{var}}$
GANO	0.28	0.75	0.32	0.76	0.34	0.77
MultilevelDiff	0.26	0.52	0.23	0.61	0.23	0.72
DDO (Ours)	0.26	0.69	0.27	0.78	0.28	0.79

*Scaled error of mean. †Scaled error of variance.

Table 2: Results of Darcy flow Bayesian inverse problem.

Results   Table 2 presents the relative errors of the sample mean and variance relative to those of the posterior computed through the Markov chain Monte Carlo (MCMC) simulation. In general, DDO outperforms the GANO baseline, showcasing better alignment with the true posterior. In particular, our model maintains consistently lower relative mean errors at all resolutions in contrast to GANO, highlighting DDO's consistency for sampling at higher resolutions. However, unlike the results from the MNIST-SDF experiments (Section 5.4), the MultilevelDiff baseline achieves the best performance among all methods. Notably, this method also demonstrates better mean and variance errors compared to the other two methods. We will revisit the performance analysis of MultilevelDiff in contrast to MNIST-SDF experiments later in this section.

Figure 7 illustrates the sample mean and variance at various resolutions generated by our DDO model and the baselines, including MCMC. These results help to identify the trends observed in the relative errors. Notably, as shown in Figure 7 (b), GANO's sample variance is significantly lower than MCMC, potentially indicating that mitigating mode collapse in GANO remains challenging. Although the other two models perform better than GANO, neither our DDO nor the MultilevelDiff models the sample variations accurately in comparison to the MCMC's statistics. In addition, DDO and MultilevelDiff also exhibit a tendency for decreased variation as the resolution increases. To further analyze these statistical differences, we examine the generated samples in more detail.

Figure 8 depicts the generated samples of the trained models at the training resolution, while the generated samples at higher resolutions are shown in Figures 17 and 18 (in Appendix J.5). For MCMC samples, the reverse C-shaped valley in the center retains low values, with strong variation at the edges across different samples. In contrast, GANO fails to maintain sufficiently low values in the reverse C-shaped region, exhibiting only minor local variations. DDO demonstrates a pattern similar to the MCMC samples but does not manage to achieve sufficiently low values in the center of the field, which appears to contribute to its error.

As we discussed above, MultilevelDiff demonstrates improved performance on this dataset unlike with MNIST-SDF. This appears to result from its noise design and the inherent characteristics of the Darcy flow solutions. While GANO and DDO rely on Gaussian-based trace-class noise, the distribution of the Darcy flow solutions exhibits heavier tails, which decay more slowly than a Gaussian distribution. In contrast, MultilevelDiff utilizes a combination of a spectral convolution-based kernel and a fixed kernel, where the spectral convolution-based noise facilitates the generation of high-frequency noise more effectively.

Interestingly, the MNIST-SDF experiment results also support the hypothesis that MultilevelDiff's noise design has an advantage in modeling high-frequency components. Despite its notable underperformance (see Figure 6), MultilevelDiff avoids producing wavy aliasing artifacts, even when it is underfit on the MNIST-SDF dataset. This observation underscores the importance of trace-class noise for diffusion-based models on function space. This suggests that further refinements to noise design could be a promising direction for future research.
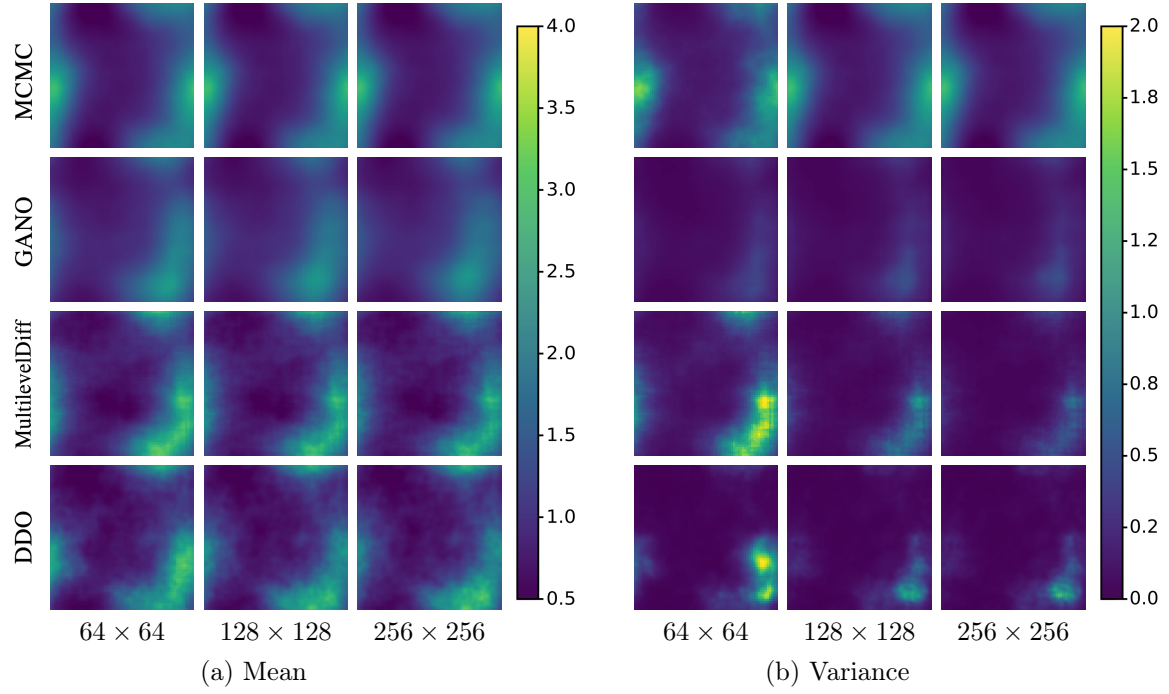
22

Figure 7: **Posterior Sample Statistics** (Section 5.5): The sample (a) mean and (b) variance of posterior samples of the MCMC as well as the learned GANO, MultilevelDiff, and DDO models at various resolutions. 10,000 samples are used.
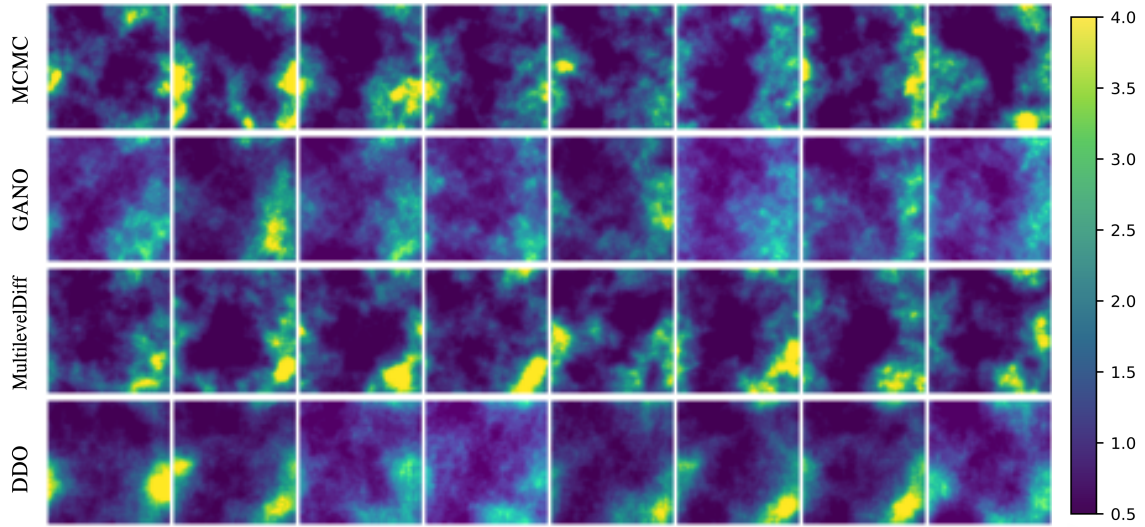

Figure 8: **Posterior Samples at 64×64 resolution** (Section 5.5): The samples of the MCMC as well as the learned GANO, MultilevelDiff, and DDO models at the training resolution (64×64).

5.6 Additional Results

Finally, we analyze the trade-offs inherent in different neural operator approaches, including spectral and continuous convolution methods, and demonstrate their practical implications through comparative experiments. This is because, unlike finite-dimensional models, function space models require parametric designs that not only enhance expressivity but also satisfy discretization invariance. Consequently, in function-valued deep generative models, the design of the neural operator is a critical factor in determining overall model performance. In this context, in Appendix K, we provide a comprehensive analysis of various neural operator designs introduced in function-valued generative models, examining their respective strengths and weaknesses, such as aliasing issues in spectral convolution and overfitting tendencies in continuous convolution methods. Readers interested in understanding the impact of these design choices on model performance are encouraged to explore this section.

6. Discussion and Conclusions

We propose DDOs, the first theoretical framework and numerical demonstration of resolution invariant diffusion generative models on function space. Our approach generalizes denoising score matching for trace-class noise corruptions that live in the Hilbert space of the data, and considers a discrete-time diffusion model for sampling using infinite-dimensional Langevin dynamics. Future work will connect this framework with noise scales that depend continuously on time (as in Appendix C) to the forward and backward SDEs in Song et al. (2020). Defining the backward SDE will require satisfying conditions that guarantee time reversibility of infinite-dimensional diffusions; see Föllmer and Wakolbinger (1986) for examples of these conditions. Adapting the covariance of the reference noise process based on the data distribution may also be helpful for generative modeling with other functional datasets, and to extend this framework to solve inverse problems with infinite-dimensional parameters Stuart (2010). Lastly, rigorous error analysis (e.g. using an approximate score) will be important to understand the class of data distributions that can be accurately characterized with infinite-dimensional diffusion models.

Acknowledgments

References

R.A. Adams and J.J.F. Fournier. *Sobolev Spaces.* ISSN. Elsevier Science, 2003. ISBN 9780080541297.

Robert J Adler. *The geometry of random fields.* SIAM, 2010.

Ivan Anokhin, Kirill Demochkin, Taras Khakhulin, Gleb Sterkin, Victor Lempitsky, and Denis Korzhenkov. Image generators with conditionally-independent pixel synthesis. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14278–14287, 2021.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv:1701.04862*, 2017.

Lorenzo Baldassari, Ali Siahkoohi, Josselin Garnier, Knut Solna, and Maarten V de Hoop. Conditional score-based diffusion models for bayesian inference in infinite dimensions. *Advances in Neural Information Processing Systems*, 36:24262–24290, 2023.

Francesca Bartolucci, Emmanuel de Bézenac, Bogdan Raonic, Roberto Molinaro, Siddhartha Mishra, and Rima Alaifari. Representation equivalent neural operators: a framework for alias-free operator learning. *Neural Information Processing Systems (NeurIPS)*, 2024.

Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3D scene generation. *arXiv:2207.13751*, 2022.

Hugo Berard, Gauthier Gidel, Amjad Almahairi, Pascal Vincent, and Simon Lacoste-Julien. A closer look at the optimization landscapes of generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2020.

V.I. Bogachev. *Gaussian Measures*. Mathematical Surveys and Monographs. American Mathematical Society, 2015. ISBN 9781470418694.

Sam Bond-Taylor and Chris G Willcocks. ∞-diff: Infinite resolution diffusion with subsampled mollified states. In *International Conference on Learning Representations (ICLR)*, 2024.

Wessel P Bruinsma, James Requeima, Andrew YK Foong, Jonathan Gordon, and Richard E Turner. The Gaussian neural process. *arXiv preprint arXiv:2101.03606*, 2021.

Gary J. Chandler and Rich R. Kerswell. Invariant recurrent solutions embedded in a turbulent two-dimensional Kolmogorov flow. *Journal of Fluid Mechanics*, 722:554–595, 2013.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8628–8638, 2021.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

Gene Chou, Yuval Bahat, and Felix Heide. Diffusion-SDF: Conditional generative modeling of signed distance functions. In *The IEEE International Conference on Computer Vision (ICCV)*, 2023.

Simon L Cotter, Gareth O Roberts, Andrew M Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3): 424–446, 2013.

KJ Craswell. Density estimation in a topological group. *The Annals of Mathematical Statistics*, 36(3):1047–1048, 1965.

G. Da Prato, G. De Prato, and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1992. ISBN 9780521385299.

Giuseppe Da Prato. *An introduction to infinite-dimensional analysis*. Springer Science & Business Media, 2006.

Sophie Dabo-Niang. Kernel density estimator in an infinite-dimensional space with a rate of convergence in the case of diffusion process. *Applied mathematics letters*, 17(4):381–386, 2004.

Masoumeh Dashti and Andrew M Stuart. The Bayesian approach to inverse problems. In *Handbook of uncertainty quantification*, pages 311–428. Springer, 2017.

Maarten De Hoop, Daniel Zhengyu Huang, Elizabeth Qian, and Andrew M Stuart. The cost-accuracy trade-off in operator learning with neural networks. *Journal of Machine Learning*, 1(3):299–341, 2022.

Léa Demeule, Mahtab Sandhu, and Glen Berseth. Adaptive resolution residual networks — generalizing across resolutions easily and efficiently. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Emilien Dupont, Hyunjik Kim, S. M. Ali Eslami, Danilo Jimenez Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you can treat it like one. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 5694–5725. PMLR, 17–23 Jul 2022a.

Emilien Dupont, Yee Whye Teh, and Arnaud Doucet. Generative models as distributions of functions. In *International Conference on Artificial Intelligence and Statistics*, pages 2989–3015. PMLR, 2022b.

Vincent Dutordoir, Alan Saul, Zoubin Ghahramani, and Fergus Simpson. Neural diffusion processes. In *International Conference on Machine Learning*, pages 8990–9012. PMLR, 2023.

Alex Dytso, Ronit Bustin, H Vincent Poor, and Shlomo Shamai. Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, 5 (1):1–40, 2018.

Lawrence C. Evans. *Partial differential equations*. American Mathematical Society, 2010.

Vladimir Sergeevich Fanaskov and Ivan V Oseledets. Spectral neural operators. In *Doklady Mathematics*, volume 108, pages S226–S232. Springer, 2023.

H Föllmer and A Wakolbinger. Time reversal of infinite-dimensional diffusions. *Stochastic processes and their applications*, 22(1):59–77, 1986.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. In *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*, 2018.

Matthias Gelbrich. On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990.

S. Ghosal and A. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. ISBN 9780521878265.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Florentin Guth, Simon Coste, Valentin De Bortoli, and Stephane Mallat. Wavelet score-based generative modeling. *Advances in neural information processing systems*, 35:478–491, 2022.

Paul Hagemann, Lars Ruthotto, Gabriele Steidl, and Nicole Tianjiao Yang. Multi-level diffusion: Infinite dimensional score-based diffusion models for image generation. *arXiv:2303.04772*, 2023.

P.R. Halmos. *Measure Theory*. Graduate Texts in Mathematics. Springer New York, 1976. ISBN 9780387900889.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415*, 2016.

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. *Neural Information Processing Systems (NeurIPS)*, 2017.

Dave Higdon. Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, pages 37–56. Springer, 2002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems (NeurIPS)*, 2020.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. Neural wavelet-domain diffusion for 3D shape generation. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

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

Bowen Jing, Gabriele Corso, Renato Berlinghieri, and Tommi Jaakkola. Subspace diffusion generative models. In *European conference on computer vision*, pages 274–289. Springer, 2022.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

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 4396–4405, 2019.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Neural Information Processing Systems (NeurIPS)*, 2021.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 9538–9563. PMLR, 2023.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pages 11201–11228. PMLR, 2022.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. In *International Conference on Learning Representations*, 2019.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Neural Information Processing Systems (NeurIPS)*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kiselev, Fedor Nazarov, and Roman Shterenberg. Blow up and regularity for fractal Burgers equation. *Dynamics of Partial Differential Equations*, 5(3):211–240, 2008.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Nikola Kovachki, Samuel Lanthaler, and Siddhartha Mishra. On universal approximation and error bounds for Fourier neural operators. *Journal of Machine Learning Research*, 22 (290):1–76, 2021.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Antti Kupiainen. Quantum fields and probability. *arXiv preprint arXiv:1611.05240*, 2016.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Neural Information Processing Systems (NeurIPS)*, 2019.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 4328–4343, 2022a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2020a.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv:2003.03485*, 2020b.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Kovachki, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning dissipative dynamics in chaotic systems. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 16768–16781, 2022b.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Gabriel Lord, Catherine Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge University Press, 08 2014. ISBN 978-0521728522. doi:10.1017/CBO9781139017329.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Nicholas H Nelsen and Andrew M Stuart. The random feature model for input-output maps between Banach spaces. *SIAM Journal on Scientific Computing*, 43(5):A3212–A3243, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning (ICML)*, 2022.

Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv:2202.11214*, 2022.

Angus Phillips, Thomas Seror, Michael Hutchinson, Valentin De Bortoli, Arnaud Doucet, and Emile Mathieu. Spectral diffusion processes. *arXiv:2209.14125*, 2022.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3D using 2D diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.

Md Ashiqur Rahman, Manuel A Florez, Anima Anandkumar, Zachary E Ross, and Kamyar Azizzadenesheli. Generative adversarial neural operators. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856.

Md Ashiqur Rahman, Zachary E Ross, and Kamyar Azizzadenesheli. U-NO: U-shaped neural operators. *Transactions on Machine Learning Research*, 2023.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv:1710.05941*, 2017.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2004.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations*, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Paul A Rosen, Eric Gurrola, Gian Franco Sacco, and Howard Zebker. The InSAR scientific computing environment. In *EUSAR 2012; 9th European conference on synthetic aperture radar*, pages 730–733. VDE, 2012.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494, 2022.

Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. MetaSDF: Meta-learning signed distance functions. In *Advances in Neural Information Processing Systems*, volume 33, pages 10136–10147, 2020.

Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. Adversarial generation of continuous images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10753–10764, 2021.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Neural Information Processing Systems (NeurIPS)*, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2020.

Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Roger Temam. *Infinite-dimensional dynamical systems in mechanics and physics.* Applied mathematical sciences. Springer-Verlag, New York, 1988.

Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Neural Information Processing Systems (NeurIPS)*, 2021.

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

Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *Neural Information Processing Systems (NeurIPS)*, 2022a.

Vikram Voleti, Christopher Pal, and Adam M Oberman. Score-based denoising diffusion with non-isotropic Gaussian noise models. In *NeurIPS 2022 Workshop on Score-Based Methods*, 2022b.

Gege Wen, Zongyi Li, Qirui Long, Kamyar Azizzadenesheli, Anima Anandkumar, and Sally M Benson. Real-time high-resolution co 2 geological storage prediction using nested Fourier neural operators. *Energy & Environmental Science*, 16(4):1732–1741, 2023.

P. Whittle. On stationary processes in the plane. *Biometrika*, 41(3/4):434–449, 1954.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1):1059, 2024.

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.

Yan Yang, Angela F Gao, Jorge C Castellanos, Zachary E Ross, Kamyar Azizzadenesheli, and Robert W Clayton. Seismic wave propagation and inversion with neural operators. *The Seismic Record*, 1(3):126–134, 2021.

Linqi Zhou, Yilun Du, and Jiajun Wu. 3D shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021.

Appendix A. Notation

We denote by $\mathbb{R}$ the real numbers and by $\mathbb{R}^n$ their n-fold Cartesian product and write $\| \cdot \|_2$ for Euclidean norm. We write $\mathbb{N}$ for the set of natural numbers. We denote by H a real, separable, Hilbert space and by $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ its inner-product and norm respectively. We write $\mathcal{B}(H)$ for the Borel sets of H generated by the open sets induced from the norm topology. For probability measures μ, ν on $\big(H, \mathcal{B}(H)\big)$, we say μ is absolutely continuous with respect to ν and denote it $\mu \ll \nu$ if, for any $B \in \mathcal{B}(H)$, $\nu(B) = 0$ implies $\mu(B) = 0$. If $\mu \ll \nu$ and $\nu \ll \mu$ hold then we μ and ν are equivalent and denote it $\mu \sim \nu$. If neither $\mu \ll \nu$ or $\nu \ll \mu$ hold then we say μ and ν are mutually singular and denote it $\mu \perp \nu$. We say u is a random variable distributed according to μ and denote it $u \sim \mu$ if the law of u is μ. Given two random variables u, v, we write $u \perp v$ if they are independent. For any mapping, $f : H \to \mathbb{R}$, we denote by $\mathbb{E}_{u \sim \mu}[f(u)]$ the expected value of f under μ.

For any bounded operator $C : H \to H$, we say C is self-adjoint if $\langle Cu, v \rangle = \langle u, Cv \rangle$ for all $u, v \in H$. We say an operator is positive if $\langle Cu, u \rangle > 0$ for all $u \in H \setminus \{0\}$ (equivalently non-negative if $\langle Cu, u \rangle \geq 0$). We say C is trace-class, or nuclear, if for any orthonormal basis $\{\phi_j\}_{j=1}^{\infty}$ of H, we have $\text{Tr}(C) = \sum_{j=1}^{\infty} \langle C\phi_j, \phi_j \rangle < \infty$. For any self-adjoint, non-negative, trace-class operator, we denote by $C^{1/2}$ the unique operator such that $C = C^{1/2}C^{1/2}$. We denote by H^* the topological (continuous) dual H which is itself a separable Hilbert space consisting of all bounded linear functionals $l : H \to \mathbb{R}$ with an inner-product induced by the Reisz map. Since it follows by the Riesz representation theorem that for any $l \in H^*$, there exists a unique element $v \in H$ such that $l(u) = \langle u, v \rangle$ for any $u \in H$, we define the Riesz map $R : H^* \to H$ by $l \mapsto v$.

Appendix B. Proofs of Theorem

B.1 Convolution of measures

The following results holds more generally for Radon Gaussian measures on locally convex spaces. We show them here in the Hilbert space setting to avoid introducing extra notation but refer the reader to Bogachev (2015) for a thorough overview of the more general setting.

Table 3: **Score-based diffusion models:** from finite to infinite dimension.

Setting	Finite dimension	Infinite dimension
Data space	Euclidean spaces	Function spaces
Base measures	Lebesgue measure	Gaussian random fields
Noise in diffusion	Multivariate random variables	Gaussian random fields
Score	Score function	Score operator
Process	Langevin process in finite dimensions y	Langevin process in function spaces
Learning loss	Euclidean norm	Norm on function spaces (discretization invariant)
Controls	Variance	Length scale, variance, energy, etc.
Base model	Neural networks	Neural operators

Let $(H, \langle \cdot, \cdot \rangle)$ be a real, separable, Hilbert space and μ and ν be two probability measures on the Borel σ-algebra $\mathcal{B}(H)$. Then the product measure $\mu \otimes \nu$ is defined on $\mathcal{B}(H) \otimes \mathcal{B}(H) = \mathcal{B}(H \times H)$. Define the mapping $T : H \times H \to H$ by $T(u, v) = u + v$. Then the pushforward of $\mu \otimes \nu$ under T is called the convolution of μ and ν and is denoted $\mu * \nu$. In particular, given two independent random variables $u \sim \mu$ and $v \sim \nu$, the random variable $u + v$ is distributed according to $\mu * \nu$. It can be shown that, for any $B \in \mathcal{B}(H)$, we have

$$(\mu * \nu)(B) = \int_H \mu(B - v) \, d\nu(v) = \int_H \nu(B - u) \, d\mu(u) = (\nu * \mu)(B), \qquad (28)$$

for example, see Appendix A.3 in Bogachev (2015) and references therein. The following result shows that if ν is a centered Gaussian and μ charges its Cameron-Martin space, the convolution is equivalent, in the sense of measures, to ν.

Theorem 6 *Let μ, ν be two probability measures on $(H, \mathcal{B}(H))$ with $\nu = N(0, C)$ for some $C : H \to H$ self-adjoint, positive, and trace-class. If $\mu(C^{1/2}(H)) = 1$, then $\nu_u \sim \nu$ where ν_u is the conditional for $v|u$ and $\nu * \mu \sim \nu$.*

Proof For any $B \in \mathcal{B}(H)$, we have by Equation 28,

$$(\nu * \mu)(B) = \int_H \nu(B - u) \, d\mu(u).$$

Therefore $(\nu * \mu)(B) = 0$ if and only if $\nu(B - u) = 0$ for μ-almost any $u \in H$ since ν is non-negative. For any $u \in H$, define the measures

$$\nu_u(B) = \nu(B - u), \qquad \forall B \in \mathcal{B}(H)$$

which are Gaussian $\nu_u = N(u, C)$. By the Cameron-Martin Theorem, given as Proposition 2.26 in Da Prato et al. (1992), $\nu_u \ll \nu$ for any $u \in C^{1/2}(H)$. Let $B \in \mathcal{B}(H)$ be such that $\nu(B) = 0$. Since, $\nu_u \ll \nu$ for any $u \in C^{1/2}(H)$, we have that $\nu(B - u) = 0$. Since $\mu(C^{1/2}(H)) = 1$, $\nu(B - u) = 0$ for μ-almost any $u \in H$ and therefore $\nu * \mu \ll \nu$.

Now let $B \in \mathcal{B}(H)$ be such that $(\nu * \mu)(B) = 0$ then $\nu_u(B) = 0$ for μ-almost any $u \in H$. Since $\mu(C^{1/2}(H)) = 1$, again by the Cameron-Martin Theorem, $\nu_u \ll \nu$. But, by Theorem 2.25 in Da Prato et al. (1992), Gaussians are either equivalent or mutually singular , therefore

$\nu \ll \nu_u$ and thus $\nu(B) = 0$ hence the result follows. ∎

Let $A : H \to H$ be a linear operator. If $u \sim \mu$, then from definition, the random variable Au is distributed according to the measure $\mu \circ A^{-1}$ where A^{-1} denotes the pre-image of A. In particular, for any $B \in \mathcal{B}(H)$,

$$(\mu \circ A^{-1})(B) = \mu(\{u \in H : Au \in B\}). \tag{29}$$

The following corollary of Theorem 6 addresses random variables of the form $Au + v$ where $u \sim \mu$ and $v \sim \nu$ are independent.

Corollary 7 *Let μ, ν be two probability measures on $(H, \mathcal{B}(H))$ with $\nu = N(0, C)$ for some $C : H \to H$ self-adjoint, positive, and trace-class. Let $A : H \to H$ be a linear operator such that $A(H) \subseteq C^{1/2}(H)$ then $\nu * (\mu \circ A^{-1}) \sim \nu$.*

Proof From equation Equation 29 and the assumption that $A(H) \subseteq C^{1/2}(H)$, we have

$$\begin{aligned}
(\mu \circ A^{-1})(C^{1/2}(H)) &= \mu(\{u \in H : Au \in C^{1/2}(H)\}) \\
&= \mu(H) \\
&= 1.
\end{aligned}$$

Thus the result follows by Theorem 6. ∎

B.2 Conditional scores

Let H be a real, separable, Hilbert space and denote by $\mathcal{B}(H)$ its Borel σ-algebra. Let γ be a probability measure on $(H \times H, \mathcal{B}(H) \otimes \mathcal{B}(H))$. We introduce the coordinates $(u, v) \sim \gamma$. Denote by μ marginal of u, by ν the marginal of v, and by γ^u the conditional $v|u$ for μ-almost any $u \in H$. Let μ_0 be a probability measure on $(H, \mathcal{B}(H))$ and suppose that $\nu \ll \mu_0$ and $\gamma^u \ll \mu_0$ for μ-almost any $u \in H$.

Lemma 8 *The Radon–Nikodym derivatives of ν and γ^u with respect to μ_0 satisfy*

$$\frac{d\nu}{d\mu_0}(v) = \mathbb{E}_{u \sim \mu} \frac{d\gamma^u}{d\mu_0}(v), \qquad \text{for } \mu_0\text{-almost any } v \in H.$$

Proof Let $B \in \mathcal{B}(H)$ then by definition of a conditional measure and Fubini's Theorem,

$$\begin{aligned}
\nu(B) &= \int_H \int_B d\gamma^u(v) d\mu(u) \\
&= \int_H \int_B \frac{d\gamma^u}{d\mu_0}(v) d\mu_0(v) d\mu(u) \\
&= \int_B \left(\int_H \frac{d\gamma^u}{d\mu_0}(v) d\mu(u) \right) d\mu_0(v).
\end{aligned}$$

34

We also have,

$$\nu(B) = \int_B \frac{d\nu}{d\mu_0}(v)d\mu_0(v).$$

Therefore

$$\int_B \frac{d\nu}{d\mu_0}(v)d\mu_0(v) = \int_B \left(\int_H \frac{d\gamma^u}{d\mu_0}(v)d\mu(u) \right) d\mu_0(v).$$

Since B is arbitrary, we must have that

$$\frac{d\nu}{d\mu_0}(v) = \int_H \frac{d\gamma^u}{d\mu_0}(v)d\mu(u)$$

for μ_0-almost any $v \in H$ which is the desired result. ∎

Let $E \subseteq H$ be a Hilbert space continuously embedded in H and denote by D_E the Frechet differential operator on H in the direction of E. Suppose that $\nu \sim \mu_0$ and $\gamma^u \sim \mu_0$ so that all respective Radon–Nikodym derivatives exist and are positive. Define,

$$\Phi(v) := \log \frac{d\nu}{d\mu_0}(v), \quad \Psi(v; u) := \log \frac{d\gamma^u}{d\mu_0}(v)$$

for μ_0-almost ant $v \in H$ and μ-almost any $u \in H$. Suppose that Φ and $\Psi(\cdot; u)$ are once D_E-continuously differentiable . Furthermore assume

$$\mathbb{E}_{v \sim \nu} \|D_E \Phi(v)\|_{E^*}^2 < \infty, \quad \mathbb{E}_{u \sim \mu} \mathbb{E}_{v \sim \gamma^u} \|D_E \Psi(v; u)\|_{E^*}^2 < \infty$$

where E^* denotes the topological dual of E. Let $G_\theta : H \to E^*$ be a parametric mapping with parameters $\theta \in \mathbb{R}^p$. Assume that, for all $\theta \in \mathbb{R}^p$,

$$\mathbb{E}_{v \sim \nu} \|G_\theta(v)\|_{E^*}^2 < \infty.$$

Define the functionals,

$$F(\theta) := \mathbb{E}_{v \sim \nu} \|D_E \Phi(v) - G_\theta(v)\|_{E^*}^2$$
$$J(\theta) := \mathbb{E}_{u \sim \mu} \mathbb{E}_{v \sim \gamma^u} \|D_E \Psi(v; u) - G_\theta(v)\|_{E^*}^2.$$

Theorem 9 *There exists a constant $C < \infty$ independent of $\theta \in \mathbb{R}^p$ such that*

$$F(\theta) = J(\theta) + C, \qquad \forall\, \theta \in \mathbb{R}^p.$$

Proof We have

$$F(\theta) = \mathbb{E}_{v \sim \nu} \left[\|G_\theta(v)\|_{E^*}^2 - 2\langle D_E \Phi(v), G_\theta(v) \rangle_{E^*} \right] + C_1$$

where $C_1 = \mathbb{E}_{v \sim \nu} \|D_E \Phi(v)\|_{E^*}^2 < \infty$ by assumption. Similarly,

$$J(\theta) = \mathbb{E}_{u \sim \mu} \mathbb{E}_{v \sim \gamma^u} \left[\|G_\theta(v)\|_{E^*}^2 - 2\langle D_E \Psi(v; u), G_\theta(v) \rangle_{E^*} \right] + C_2$$

where $C_2 = \mathbb{E}_{u \sim \mu} \mathbb{E}_{v \sim \gamma^u} \|D_E \Psi(v; u)\|_{E^*}^2 < \infty$ by assumption. By definition of a conditional measure,

$$\mathbb{E}_{v \sim \nu} \|G_\theta(v)\|_{E^*}^2 = \mathbb{E}_{u \sim \mu} \mathbb{E}_{v \sim \gamma^u} \|G_\theta(v)\|_{E^*}^2$$

for any $\theta \in \mathbb{R}^p$. Using Lemma 8, the Leibniz integral rule, and Fubini's Theorem, we find

$$
\begin{aligned}
\mathbb{E}_{v\sim\nu}\langle D_E\Phi(v),G_\theta(v)\rangle_{E^*} &= \mathbb{E}_{v\sim\nu}\langle D_E\log\frac{d\nu}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{v\sim\nu}\langle\frac{d\mu_0}{d\nu}(v)D_E\frac{d\nu}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{v\sim\mu_0}\langle D_E\frac{d\nu}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{v\sim\mu_0}\langle D_E\mathbb{E}_{u\sim\mu}\frac{d\gamma^u}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{v\sim\mu_0}\langle\mathbb{E}_{u\sim\mu}D_E\frac{d\gamma^u}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{v\sim\mu_0}\langle\mathbb{E}_{u\sim\mu}\frac{d\gamma^u}{d\mu_0}(v)D_E\log\frac{d\gamma^u}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{v\sim\mu_0}\mathbb{E}_{u\sim\mu}\frac{d\gamma^u}{d\mu_0}(v)\langle D_E\log\frac{d\gamma^u}{d\mu_0}(v),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{u\sim\mu}\mathbb{E}_{v\sim\mu_0}\frac{d\gamma^u}{d\mu_0}(v)\langle D_E\Psi(v;u),G_\theta(v)\rangle_{E^*} \\
&= \mathbb{E}_{u\sim\mu}\mathbb{E}_{v\sim\gamma^u}\langle D_E\Psi(v;u),G_\theta(v)\rangle_{E^*}.
\end{aligned}
$$

Setting $C = C_1 - C_2$ completes the proof. $\blacksquare$

It remains to show that approximating the score of the convolved measure ensues we are close to the true measures. The following lemma relates the Wasserstein distance of the two.

Lemma 10 *Let $\eta \sim \mu_\sigma$ be a noise random variable with finite p-moment, and let $v = u+\eta \sim \mu * \mu_\sigma$. Then, the Wasserstein-p distance for $p \geq 1$ satisfies $W_p(\mu * \mu_\sigma, \mu) \leq \|\eta\|_{L^p(\mu_\sigma)}$.*

Proof Let (u,η) follow the product coupling $(\mu \otimes \mu_\sigma)(du, d\eta)$. Then, let the coupling (v,η) where $v = u + \eta$ be drawn according to $(u+\eta, u)_\sharp(\mu \otimes \mu_\sigma)(du, d\eta)$. Choosing this coupling (v,η) to upper bound the Wasserstein-p distance, we have

$$
W_p(\mu * \mu_\sigma, \mu)^p = \inf_{\gamma\in\Pi(\mu*\mu_\sigma,\mu)}\int|v-u|^p d\gamma(v,u) \leq \int|(u+\eta)-u|^p(\mu \otimes \mu_\sigma)(du,d\eta).
$$

Given that the integrand is independent of u, we have $W_p(\mu*\mu_\sigma,\mu)^p \leq \int|\eta|^p d\mu_\sigma = \|\eta\|_{L^p(\mu_\sigma)}^p$, and the result follows. $\blacksquare$

Remark 11 *For a Gaussian measure μ_σ, the p-th moment is finite by the Fernique Theorem. Moreover, by Theorem 6.6 in Stuart (2010) there is a constant $C_p > 0$ so that $\|\eta\|_{L^p} \leq C_p(Tr(C_\sigma))$. This can be used to establish the convergence rate of $W_p(\mu * \mu_\sigma, \mu) \to 0$ as $\sigma \to 0$.*

Lemma 12 *The Fréchet derivative $D_{H_{\mu_0}}\Psi(w;u)$ as defined in Equation 10 is in $H_{\mu_0}^*$.*

Proof Notice that since $u \in H_{\mu_0}$, μ-almost surely, we can find $g \in H$ such that $u = C^{1/2}g$. For any $w \in H_{\mu_0}$, we can similarly find $f \in H$ such that $w = C^{1/2}f$. We can write both g and f in the othronormal basis $\{\varphi_j\}_{j=1}^{\infty}$,

$$g = \sum_{j=1}^{\infty} \langle g, \varphi_j \rangle \varphi_j, \qquad f = \sum_{j=1}^{\infty} \langle f, \varphi_j \rangle \varphi_j$$

with both series converging in H. Orthonormality implies

$$\langle u, \varphi_j \rangle = \lambda_j^{1/2} \langle g, \varphi_j \rangle, \qquad \langle w, \varphi_j \rangle = \lambda_j^{1/2} \langle f, \varphi_j \rangle$$

for any $j \in \mathbb{N}$. Therefore equation Equation 11 becomes

$$\begin{aligned}
D_{H_{\mu_0}} \Psi(v; u)w &= \sum_{j=1}^{\infty} \langle g, \varphi_j \rangle \langle f, \varphi_j \rangle \\
&\leq \left(\sum_{j=1}^{\infty} |\langle g, \varphi_j \rangle|^2 \right)^{1/2} \left(\sum_{j=1}^{\infty} |\langle f, \varphi_j \rangle|^2 \right)^{1/2} \\
&= \|g\| \|f\|
\end{aligned}$$

which is finite hence the result follows. ∎

Appendix C. Multiple Noise Scales

To satisfy our absolute continuity condition for the perturbed data measure in Sections 4.2 and 4.5, we need $A_t(H) \subseteq \text{Im}(C_t^{1/2})$ for all t. The following lemmas show that this condition holds for different time-dependent scalar weightings of the forward and noise covariance operators.

Lemma 13 *Let $\eta_t \sim \mu_t = \mathcal{N}(0, C_t)$ where $C_t = g(t)C$ and $v_t = A_t u + \eta_t$ where $A_t = f(t)u$ for all $t \in I$ for $u \in H$. Assuming the mappings $f, g : \mathcal{I} \to \mathbb{R}$ satisfy $0 < M_1 \leq f(t), g(t) \leq M_2$ for all $t \in \mathcal{I}$ and $\mu\big(C^{1/2}(H)\big) = 1$, then $A_t(H) \subseteq C_t^{1/2}(H)$ for all $t \in \mathcal{I}$*

Proof Let $L_t = g(t)/f(t)C$. We will first show that $L_t^{1/2}(H) = C^{1/2}(H)$. By Lemma 6.15 in Stuart (2010), the image of the two positive-definite, and self-adjoint linear operators on a Hilbert space H are equal if and only if there exists constants $K_1, K_2 > 0$ such that $K_1 \langle u, Cu \rangle \leq \langle u, L_t u \rangle \leq K_2 \langle u, Cu \rangle$ for all $u \in H$.

Under the conditions on f, g, for any $u \in H$, we have

$$\langle u, Cu \rangle = \langle u, \frac{f(t)g(t)}{f(t)g(t)} Cu \rangle \leq \frac{M_2}{M_1} \langle u, \frac{g(t)}{f(t)} Cu \rangle = K_1 \langle u, L_t u \rangle$$

$$= K_1 \langle u, \frac{g(t)}{f(t)} Cu \rangle \leq K_1 \frac{M_2}{M_1} \langle u, Cu \rangle = K_2 \langle u, Cu \rangle,$$

where $K_1 = M_2/M_1$ and $K_2 = K_1^2$. Then, for $\mu\big(C^{1/2}(H)\big) = 1$ we have $u \in C^{1/2}(H)$ for $u \in H$. From the image equivalence, we have $u \in L_t^{1/2}(H)$ and so $A_t u = f(t)u \in C_t^{1/2}(H)$. ∎

Lemma 14 *Let $C_t = f(t)AA^*$ where $A\colon H \to H$ is a linear operator and $f\colon [0,T] \to \mathbb{R}$ is a function satisfiying $c = \sup_{t \in [0,T]} 1/f(t) < +\infty$. Then, for $K = AA^*$ we have that $Im(K^{1/2}) \subseteq Im(C_t^{1/2})$ for all $t \in [0,T]$.*

Proof The image of A is equivalent to image of the $K^{1/2}$ where $K = AA^*$ is a positive-definite, and self-adjoint operator. For any $u \in H$, we have

$$\langle u, Ku \rangle = \langle u, AA^*f(t)/f(t)u \rangle \leq \sup_{t \in [0,T]} \frac{1}{f(t)} \langle u, AA^*f(t)u \rangle = c\langle u, C_t u \rangle.$$

The result on the image spaces follows by Lemma 6.15 in Stuart (2010). ∎

Example 2 *The function $f(t) = e^{\gamma t}$ for $\gamma > 0$ satisfies the condition in the lemma above with $c = 1$. This choice motivates the following study.*

Alternatively, we can define the forward process for data corruption with multiple noise scales using a stochastic differential equation (SDE), as in Song et al. (2020). Let us consider the linear SDE $du_t = -Lu_t dt + dW_t$ for $u_t \in H$ where W_t is a Q-Wiener process and $L\colon H \to H$ is a linear and positive-definite operator where its eigenvectors form an orthonormal basis for H. The solution of this SDE for any $t > 0$ is given by

$$u(t) = e^{-Lt}u(0) + \int_0^t e^{L(s-t)}dW_s.$$

Letting $u(0) = u$ and $A_t := e^{-Lt}$, we can treat $u \mapsto A_t u$ as the forward model and the second term $\eta_t := \int_0^t e^{L(s-t)}dW_s$ as the additive noise process, which is drawn independently of u. The following abridged theorem from Da Prato et al. (1992) describes the statistical properties of the noise process.

Theorem 15 *Assuming $\int_0^T Tr[A_r Q A_r^*]dr < \infty$, then (i) η_t is Gaussian, (ii) has continuous paths, and (iii) its covariance is given by*

$$C_t := Cov(\eta_t) = \int_0^t A_r Q A_r^* dr, \quad t \in [0,T].$$

To satisfy the absolute continuity conditions on the perturbed data measure for $u(t)$ as before, we need to show that for each t, $A_t(H) \subseteq Im(C_t^{1/2})$. As shown in Corollary B.7 of Da Prato et al. (1992), the image of $C_t^{1/2}$ for a covariance of the form above is equivalent to the image of the linear operator $B_t\colon H \to H$ defined as

$$B_t u := \int_0^t L_{t-s}Qu\,ds.$$

38

B_t and A_t are both linear and self-adjoint operators, so the condition $\mathrm{Im}(A_t) \subseteq \mathrm{Im}(B_t)$ holds if and only if there exists a constant $K > 0$ so that $\langle u, A_t u \rangle \leq K \langle u, B_t u \rangle$ for all $u \in H$. Using the decomposition of $u = \sum_{j=1}^{\infty} \langle u, \psi_j \rangle \psi_j$ where $\{\psi_j\}_{j=1}^{\infty}$ are eigenvectors and $\{\lambda_j\}_{j=1}^{\infty}$ are eigenvalues of L, we have

$$A_t u = \sum_j e^{-Lt} \langle u, \psi_j \rangle \psi_j = \sum_j e^{-\lambda_j t} \langle u, \psi_j \rangle \psi_j.$$

Choosing the noise covariance to be $Q = L^{\gamma}$ for some scalar γ such that draws remain in H, we have

$$B_t u = \int_0^t \sum_j e^{-L(t-s)} \langle u, \psi_j \rangle Q \psi_j ds$$

$$= \int_0^t \sum_j e^{-\lambda_j(t-s)} \langle u, \psi_j \rangle \lambda_j^{\gamma} \psi_j ds$$

$$= \sum_j \langle u, \psi_j \rangle \lambda_j^{\gamma} \psi_j \int_0^t e^{-\lambda_j(t-s)} ds$$

$$= \sum_j \langle u, \psi_j \rangle \lambda_j^{\gamma-1}(1 - e^{-\lambda_j t}) \psi_j.$$

We can now compare the images of the operators. For $u \in H$ we have

$$\langle u, A_t u \rangle = \sum_j e^{-2\lambda_j t} |\langle u, \psi_j \rangle|^2, \qquad \langle u, B_t u \rangle = \sum_j \lambda_j^{2(\gamma-1)}(1 - e^{-2\lambda_j t})^2 |\langle u, \psi_j \rangle|^2.$$

For each $\lambda_j > 0$, there exists a time t_j such that $e^{-2\lambda_j t} \leq \lambda_j^{2(\gamma-1)}(1 - e^{-2\lambda_j t})^2$ for all $t > t_j$. For these times, we satisfy the condition required for our theory. Generalizing these results is an important direction for future work.

Appendix D. Crank–Nicolson Discretization

In Sections 4.4 and 4.5, we introduced our method based on the Euler-Maruyama discretization. Here, we will show the Crank–Nicolson discretization sampling method and relate it to existing methods in the literature. In particular, we will work in the setting of multiple noise scale as introduced in Section 4.5.

Let $F : H \times \mathcal{I} \to H$ be defined as $F(u, t) = -u + R D_{H_{\mu_t}} \Phi(u, t)$. For a fixed $t \in \mathcal{I}$, we apply the Crank–Nicolson method to the linear part of the drift in Equation 16 to obtain

$$(2 + h_t)u_{n+1} = (2 - h_t)u_n + 2h_t G(u_n, t) + \sqrt{8h_t}\eta_n^{(t)} \tag{30}$$

where we define $G : H \times \mathcal{I} \to H$ by $G(u, t) = R D_{H_{\mu_t}} \Phi(u, t)$ and $h_t > 0$. For any $h_t \in (0, 2)$, Equation 30 can be written as

$$u_{n+1} = \alpha_t u_n + (1 - \alpha_t)G(u_n, t) + \beta_t \eta_n^{(t)} \tag{31}$$

with the transformation $\beta_t^2 = 8h_t/(2 + h_t)^2$ where $\beta_t \in (0, 1)$ and we define $\alpha_t = \sqrt{1 - \beta_t^2}$. Equation 31 is a type of Metropolis-adjusted Langevin proposal in the function space setting

and is related to the celebrated pre-conditioned Crank–Nicolson MCMC method Cotter et al. (2013). We remark that Equation 31 resembles the exact, single-step, Gaussian approximation sampling method of Ho et al. (2020). We leave the design and analysis of algorithms based on this approach for future work.

Appendix E. Denoising Diffusion Probabilistic Models

In Section 4.5 we showed that for a particular choice of data and noise scaling, we may recover the forward process of the DDPM framework proposed in Ho et al. (2020). Let us recall the noise process in DDPM: for some sequence $0 < \beta_1 \leq \cdots \leq \beta_T < 1$, let $\alpha_t = \prod_{s=1}^t (1 - \beta_s)$. Then, for $u_0 \sim \mu$ we define

$$u_t = \sqrt{\alpha_t} u_0 + \sqrt{1 - \alpha_t}\eta, \qquad \eta \sim \mu_0 = N(0, C) \tag{32}$$

for $t \geq 1$ with $\eta \perp u_0$. Under the assumption that $\mu\big(C^{1/2}(H)\big) = 1$, we have that the measure ν_t defined as the law of u_t, is equivalent to the measure $\mu_t = N(0, (1 - \alpha_t)C)$ for any t. Furthermore the law of the conditional $u_t | u_0$ is equivalent to μ_t, μ-almost surely. We may therefore apply the theory presented in Section 4 to obtain a sequence of tractable score-matching problems. Once solved, we obtain a sequence of approximate scores which can be used within an annealed Langevin algorithm similar to Algorithm 1 to obtain samples. This procedure, however, is not equivalent to the sampling procedure in Ho et al. (2020) which compares the backwards conditionals $u_{t-1} | u_t, u_0$ to Gaussian parameterizations and therefore an exact backwards sampling method is derived. We show now how a similar scheme may be derived in infinite dimensions.

We may compute directly that the law of $u_{t-1} | u_t, u_0$, denoted by π_{t-1}, is the Gaussian $N(u_{t-1}; m_t, c_t C)$ where

$$m_t = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1 - \alpha_t} u_0 + \frac{\sqrt{1 - \beta_t}(1 - \alpha_{t-1})}{1 - \alpha_t} u_t, \qquad c_t = \frac{(1 - \alpha_{t-1})\beta_t}{1 - \alpha_t}.$$

We can therefore consider the parametric Gaussian measure $\rho_{t-1}(u_{t-1}; u_t, t) = N(u_{t-1}; G_\theta(u_t, t), c_t C)$ for some $G_\theta \colon H \times \mathcal{I} \to H$. If we are to compare π_{t-1} and ρ_{t-1} using the Kullback-Liebler (KL) divergence, as done in Ho et al. (2020), we need that π_{t-1} and ρ_{t-1} are equivalent otherwise their KL divergence is infinite. Since π_{t-1} and ρ_{t-1} have the same covariance, by the Feldman-Hájek theorem we need only that $m_t - G_\theta(u_t, t) \in C^{1/2}(H)$ for the measures to be equivalent. Using the forward process in Equation 32, we can also write

$$m_t = \frac{\sqrt{\alpha_{t-1}}\beta_t + \sqrt{(1 - \beta_t)\alpha_t}(1 - \alpha_{t-1})}{1 - \alpha_t} u_0 + \frac{\sqrt{1 - \beta_t}(1 - \alpha_{t-1})}{\sqrt{1 - \alpha_t}}\eta,$$

and even with the assumption $\mu\big(C^{1/2}(H)\big) = 1$, we have that $m_t \notin C^{1/2}(H)$, $\mu \otimes \mu_0$-almost surely because $\eta \notin C^{1/2}(H)$, μ_0-almost surely; see Appendix H. It is therefore not enough to constrain the range of G_θ to $C^{1/2}(H)$; we need instead that, for every realization of the data and noise, G_θ yields from u_t precisely a direction so that $m_t - G_\theta(u_t, t) \in C^{1/2}(H)$. We may accomplish this with the following re-parameterization,

$$G_\theta(u_t, t) = \frac{\sqrt{1 - \beta_t}(1 - \alpha_{t-1})}{1 - \alpha_t} u_t + F_\theta(u_t, t)$$

for some $F_\theta : H \times \mathcal{I} \to C^{1/2}(H)$. Then

$$m_t - G_\theta(u_t, t) = \frac{\sqrt{\alpha_{t-1}}\beta_t}{1 - \alpha_t} u_0 - F_\theta(u_t, t)$$

which is an element of $C^{1/2}(H)$, μ-almost surely. It now follows that π_{t-1} and ρ_{t-1} are equivalent measures and we may therefore compute their KL divergence, in particular,

$$
\begin{aligned}
D_{\mathrm{KL}}(\pi_{t-1}, \rho_{t-1}) &= \| C_t^{-1/2} \big(m_t - G_\theta(u_t, t) \big) \|^2 \\
&= \frac{1 - \alpha_t}{(1 - \alpha_{t-1})\beta_t} \left\| C^{-1/2} \left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1 - \alpha_t} u_0 - F_\theta(u_t, t) \right) \right\|^2 .
\end{aligned}
$$

Moreover, we may optimize the following joint objective that minimizes the KL divergence at all times $t > 1$,

$$\sum_{t>1} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta \sim \mu_0} \frac{1 - \alpha_t}{(1 - \alpha_{t-1})\beta_t} \left\| C^{-1/2} \left(\frac{\sqrt{\alpha_{t-1}}\beta_t}{1 - \alpha_t} u - F_\theta\big(\sqrt{\alpha_t}u + \sqrt{1 - \alpha_t}\eta, t\big) \right) \right\|^2$$

We note that F_θ cannot be further re-parameterized so that it learns the noise from the signal, even in the case C is positive so that $C^{-1/2}$ can be dropped from the objective, because re-parameterizing it this way will violate its range condition which is crucial for obtaining measure equivalence. Once F_θ is learned, sampling from the approximate backwards conditional $u_{t-1}|u_t$ amounts to evaluating

$$u_{t-1} = \frac{\sqrt{1 - \beta_t}(1 - \alpha_{t-1})}{1 - \alpha_t} u_t + F_\theta(u_t, t) + \sqrt{c_t}\eta, \qquad \eta \sim \mu_0.$$

An alternative approach is to not require measure equivalence and instead compare π_{t-1} and ρ_{t-1} in a metric which is finite when comparing singular measures. One example of such a metric is the Wasserstein-p distance. Since we do not require equivalence, we may parameterize ρ_{t-1} more generally by allowing a different scaling for the covariance or even learning a different covariance operator. We will not pursue this here for the sake of simplicity in exposition and will consider π_{t-1} and ρ_{t-1} to have the same covariance. In this case, the Wasserstein-2 distance Gelbrich (1990) is

$$W_2^2(\pi_{t-1}, \rho_{t-1}) = \| m_t - G_\theta(u_t, t) \|^2.$$

Similarly to before, we may use the forward process in Equation 32 and write

$$m_t = \frac{1}{\sqrt{1 - \beta_t}} \left(u_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}}\eta \right).$$

Therefore, by re-parameterizing,

$$G_\theta(u_t, t) = \frac{1}{\sqrt{1 - \beta_t}} \left(u_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} F_\theta(u_t, t) \right)$$

for some $F_\theta : H \times \mathcal{I} \to H$ yields

$$m_t - G_\theta(u_t, t) = \frac{\beta_t}{\sqrt{(1 - \beta_t)(1 - \alpha_t)}} \Big(F_\theta(u_t, t) - \eta \Big).$$

41

Moreover, we may optimize the following joint objective for all times $t > 1$,

$$\sum_{t>1} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta \sim \mu_0} \frac{\beta_t^2}{(1 - \beta_t)(1 - \alpha_t)} \| F_\theta (\sqrt{\alpha_t} u + \sqrt{1 - \alpha_t} \eta, t) - \eta \|^2. \tag{33}$$

Once F_θ is learned, sampling from the approximate backwards conditional $u_{t-1}|u_t$ amounts to evaluating

$$u_{t-1} = \frac{1}{\sqrt{1 - \beta_t}} \left(u_t - \frac{\beta_t}{\sqrt{1 - \alpha_t}} F_\theta(u_t, t) \right) + \sqrt{c_t} \eta, \qquad \eta \sim \mu_0. \tag{34}$$

The above derivation precisely yields the framework in Ho et al. (2020) with the only difference that μ_0 does not have an identity covariance. Note that we did not even require the assumption $\mu(C^{1/2}(H)) = 1$ because we allowed ourselves to compare mutually singular measures. It is unclear whether such a formulation is preferred in practical applications or if it may eventually yields instabilities in the algorithm or lack of convergence. Furthermore, we note that both derivations worked directly with conditionals $u_{t-1}|u_t, u_0$ instead of the true backwards conditionals $u_{t-1}|u_t$ and did not establish equivalence between their optimization objectives. Empirical comparisons of these two objectives for diffusion modeling is an important direction to explore in future work.

Appendix F. Examples of $\mu(H_{\mu_0}) = 1$

In this section, we give examples of data distributions μ that satisfies the condition $\mu(H_{\mu_0}) = 1$, where H_{μ_0} is a Cameron-Martin space of a perturbation noise. These will help us gaining intuition on the condition.

F.1 Gaussian

When we can expect the assumption $\mu(H_{\mu_0}) = 1$ to hold, one of the simplest examples is when the data measure is Gaussian. Consider the space

$$H = \dot{L}^2(\mathbb{T}^d; \mathbb{R}) := \left\{ u \in L^2(\mathbb{T}^d; \mathbb{R}) : \int_{\mathbb{T}^d} u \, dx = 0 \right\},$$

where $\mathbb{T}^d$ is the d-dimensional unit torus. We denote by $\dot{H}^s(\mathbb{T}^d; \mathbb{R})$ for any $s > 0$ as the corresponding periodic, mean-zero Sobolev spaces (Adams and Fournier, 2003). Let $\mu = N(0, C_1)$ where

$$C_1 = \sigma_1^2(-\Delta + \tau_1^2 I)^{-\alpha_1}. \tag{35}$$

Here $-\Delta$ is the negative Laplacian with periodic boundary conditions, I is the identity operator, and $\sigma_1, \tau_1, \alpha_1$ are positive scalars. Covariances of the type Equation 35 are said to be of the Matérn-type because Gaussian processes defined by Matérn kernels are the only stationary solutions to certain SPDEs with differential operator C_1^{-1} (Whittle, 1954; Lindgren et al., 2011). We make extensive use of such covariances throughout the rest of this work as the Gaussian measures defined by them are amenable to analysis and efficient sampling. When $\alpha_1 > d/2$, Lemma 6.27 in Stuart (2010) implies that $\mu(\dot{H}^s(\mathbb{T}^d; \mathbb{R})) = 1$ for any $s \in [0, \alpha_1 - d/2)$. We will assume that $\alpha_1 > d$. Let $\mu_0 = N(0, C_2)$ where

$$C_2 = \sigma_2^2(-\Delta + \tau_2^2 I)^{-\alpha_2} \tag{36}$$

with $\alpha_2 > d/2$ so that C_2 is trace-class. It is easy to compute that $C_2^{1/2}(H) = \dot{H}^{\alpha_2}(\mathbb{T}^d; \mathbb{R})$. Therefore, the assumption $\mu(H_{\mu_0}) = 1$ is satisfied for any $\alpha_2 \in (d/2, \alpha_1 - d/2)$.

The above analysis reveals that there is a gap of size $d/2$ between the regularity of the data and the noise. In particular, $\mu(\dot{H}^s(\mathbb{T}^d; \mathbb{R})) = 1$ for $s \in [0, \alpha_1 - d/2)$ while $\mu_0(\dot{H}^m(\mathbb{T}^d; \mathbb{R})) = 1$ for $m \in [0, \alpha_1 - d)$. Therefore, in order to consider perturbations with Gaussians of the form Equation 4, the noise must be at least $d/2$ "less smooth" than the data, in a Sobolev sense. Furthermore since we want to consider noise with a trace-class covariance so that it is amenable to approximation, we have a fundamental limit on the regularity of the data. That is, the data must live in $\dot{H}^s(\mathbb{T}^d; \mathbb{R})$ for some $s > d/2$. This assumption can be satisfied, for example, when the data measure is defined as the pushforward of some PDE solution operator; we show explicit examples below (in Appendix F).

F.2 Gaussian Mixture

Let $D \subset \mathbb{R}^d$ be a bounded, open set with Lipschitz boundary and consider $H = L^2(D; \mathbb{R})$. We will consider the covariances Equation 35 and Equation 36 where $-\Delta$ is instead the negative Laplacian with zero Dirichlet boundary conditions on D. Suppose $\alpha_1 - d/2 > s$ for some $s \geq d$ and let $f_1, f_2 \in H_0^1(D; \mathbb{R}) \cap H^s(D; \mathbb{R})$. Define μ so that, if $u \sim \mu$, then

$$\begin{cases} u \sim N(f_1, C_1) & \text{w.p. } p, \\ u \sim N(f_2, C_1) & \text{w.p. } 1 - p, \end{cases}$$

for some $0 \leq p \leq 1$. By Lemma 6.27 in Stuart (2010), $u \in H_0^1(D; \mathbb{R}) \cap H^s(D; \mathbb{R})$ μ-almost surely. Therefore $\mu_0 = N(0, C_2)$ with $\alpha_2 \in (1/2, s]$ implies $\mu(H_{\mu_0}) = 1$.

F.3 Pushforwards

F.3.1 NAVIER-STOKES

Consider the vorticity form of the two-dimensional Navier-Stokes equations on the unit torus,

$$\begin{aligned} \partial_t u + \nabla^\perp \phi \cdot u - \epsilon \Delta u &= f, & \text{in } \mathbb{T}^2 \times (0, \infty), \\ -\Delta \phi &= u, & \text{in } \mathbb{T}^2 \times (0, \infty), \end{aligned} \tag{37}$$

with initial condition $u(\cdot, 0) = u_0$ for some $u_0, f \in \dot{L}^2(\mathbb{T}^2; \mathbb{R})$ and $\epsilon > 0$. It is shown in Temam (1988) that for any $\epsilon > 0$, Equation 37 has a unique weak solution such that $u(\cdot, t) \in \dot{H}^s(\mathbb{T}^2; \mathbb{R})$ for any $s > 0$ and $t > 0$. We may thus define the flow map $Q : \dot{L}^2(\mathbb{T}^2; \mathbb{R}) \times (0, \infty) \to \dot{H}^s(\mathbb{T}^2; \mathbb{R})$ for any $s > 0$ by $(u_0, t) \mapsto u(\cdot, t)$. Let $\rho = N(0, C_1)$ with C_1 given by Equation 35 for any $\alpha_1 > 1$. Let $\mu = Q(\cdot, T)_\sharp \rho$ for some $T > 0$. Then, for $u \sim \mu$, we have $u \in \dot{H}^s(\mathbb{T}^2; \mathbb{R})$, for any $s > 0$, μ-almost surely. Therefore $\mu_0 = N(0, C_2)$ with any $\alpha_2 > 1$ implies $\mu(H_{\mu_0}) = 1$ where C_2 is given by Equation 36.

F.3.2 BURGERS' EQUATION

Consider the one-dimensional Burgers' equation on the unit torus,

$$\begin{aligned} \partial_t u + \frac{1}{2} \partial_x (u^2) - \epsilon \partial_{xx}^2 u &= f, & \text{in } \mathbb{T} \times (0, \infty), \\ u(\cdot, 0) &= u_0, & \text{in } \mathbb{T}, \end{aligned} \tag{38}$$

43

for some $u_0, f \in \dot{L}^2(\mathbb{T}; \mathbb{R})$ and $\epsilon > 0$. By Theorem 1.1 in Kiselev et al. (2008), the solution $u(\cdot, t)$ is real analytic for all times $t > 0$, so we may define the flow map $Q : \dot{L}^2(\mathbb{T}; \mathbb{R}) \times (0, \infty) \to \dot{H}^s(\mathbb{T}; \mathbb{R})$ for any $s > 0$ by $(u_0, t) \mapsto u(\cdot, t)$. Let $\rho = N(0, C_1)$ with C_1 given by Equation 35 for any $\alpha_1 > 1/2$. Let $\mu = Q(\cdot, T)_\sharp \rho$ for some $T > 0$. Then, for $u \sim \mu$, we have $u \in \dot{H}^s(\mathbb{T}; \mathbb{R})$, for any $s > 0$, μ-almost surely. Therefore $\mu_0 = N(0, C_2)$ with any $\alpha_2 > 1/2$ implies $\mu(H_{\mu_0}) = 1$ where C_2 is given by Equation 36.

F.3.3 DARCY FLOW

Let $D \subset \mathbb{R}^d$ be a bounded, open set with Lipschitz boundary and consider the steady-state of the Darcy flow equation,

$$\begin{aligned} \nabla \cdot (a\nabla u) &= f, & \text{in } D, \\ u &= 0, & \text{in } \partial D, \end{aligned} \tag{39}$$

for some $a \in L^\infty(D; \mathbb{R}_+)$ and $f \in L^2(D; \mathbb{R})$. It is shown in Evans (2010) that Equation 39 has a unique weak solution $u \in H_0^1(D; \mathbb{R})$ and thus we can define the mapping $Q : L^\infty(D; \mathbb{R}_+) \to H_0^1(D; \mathbb{R})$ by $a \mapsto u$. Let $\rho = N(0, C_1)$ with C_1 given by Equation 35 where $-\Delta$ is instead the negative Laplacian with zero Neumann boundary conditions on D. Assume that $\alpha_1 > d/2$. For some $0 < c_- < c_+ < \infty$, define $T : \mathbb{R} \to \mathbb{R}_+$ by

$$T(x) = \begin{cases} c_-, & x < 0, \\ c_+, & x \geq 0. \end{cases}$$

We may view $T : L^2(D; \mathbb{R}) \to L^\infty(D; \mathbb{R}_+)$ as a Nemistkii operator, that is,

$$(Tf)(x) = T(f(x)), \qquad \forall\, f \in L^2(D; \mathbb{R}).$$

Let $\mu = (Q \circ T)_\sharp \rho$ then, for $u \sim \mu$, we have $u \in H_0^1(D; \mathbb{R})$ μ-almost surely. Therefore $\mu_0 = N(0, C_2)$ with $\alpha_2 \in (1/2, 1]$ implies $\mu(H_{\mu_0}) = 1$ where C_2 is given by Equation 36 and $-\Delta$ is instead the negative Laplacian with zero Dirichlet boundary conditions on D. Notice that will this condition on α_2, C_2 is trace-class only when $d = 1$.

Appendix G. The Karhunen-Loéve Expansion

Let C be a self-adjoint, positive, semi-definite operator in a Hilbert space H with an orthonormal set of eigenvectors (functions) $\phi_j \in H$ and corresponding eigenvalues λ_j in a decreasing order, i.e., $\lambda_1 \geq \lambda_2 \geq \dots$. The Karhunen-Loéve (KL) expansion represents a Gaussian random variable $u \sim \mathcal{N}(m, C)$ with mean $m \in H$ and covariance operator C as

$$u = m + \sum_{j=1}^{\infty} \sqrt{\lambda_j} \phi_j \xi_j,$$

where $\{\xi_j\}_{j=1}^{\infty}$ is an i.i.d. sequence of $\mathcal{N}(0, 1)$ random variables. This construction allows us to sample a Gaussian measure on a Hilbert space and, for certain domains and boundary conditions, can be implemented with fast FFT-based methods Lord et al. (2014). For more details on the KL expansion, we refer the reader to Da Prato (2006); Adler (2010).

Appendix H. Noise Regularity

Consider the Gaussian μ_0 as defined in Section 4. It follows by Lemma 6.10 in Stuart (2010) that $\mu_0(H_{\mu_0}) = 0$. In particular, any random variable $u \sim \mu_0$ is not contained in H_{μ_0} with probability one. This makes quantities of the form

$$\|C^{-1/2}(u - g)\| = \infty, \qquad \mu_0\text{-almost surely}$$

for any fixed $g \in H$ since $u - g \notin H_{\mu_0}$ with probability one. To see this, consider the following formal calculation. Let

$$C\phi_j = \lambda_j \phi_j, \quad \|\phi_j\| = 1$$

by an eigendecomposition of C. By the spectral theorem, $\{\phi_j\}_{j=1}^{\infty}$ forms a complete orthonormal basis for H and for positive-definite C we have $\lambda_j > 0$. Suppose $u \sim N(0, C)$. Then, by the Karhunen-Loeve expansion (see Appendix G) we have

$$u = \sum_{j=1}^{\infty} \sqrt{\lambda_j} \xi_j \phi_j, \quad \xi_j \sim N(0, 1).$$

From $C^{-\frac{1}{2}} \phi_j = \frac{1}{\sqrt{\lambda_j}} \phi_j$, we have that $C^{-\frac{1}{2}} u = \sum_{j=1}^{\infty} \xi_j \phi_j$, and thus

$$\|C^{-\frac{1}{2}} u\|^2 = \sum_{j=1}^{\infty} |\xi_j|^2 = \infty, \qquad \text{a.s.}$$

The same result holds if we subtract some element $g \in H$ from u. Writing $g = \sum_{j=1}^{\infty} g_j \phi_j$, where $g_j = \langle g, \phi_j \rangle$, then

$$\|C^{-\frac{1}{2}}(u - g)\|^2 = \sum_{j=1}^{\infty} \left| \xi_j - \frac{g_j}{\sqrt{\lambda_j}} \right|^2 = \infty, \qquad \text{a.s.}$$

And the situation is not improved even if $g \in H_{\mu_0}$ so that $g_j/\sqrt{\lambda_j} \to 0$ and the series $\sum_{j=1}^{\infty}(g_j/\sqrt{\lambda_j})\phi_j$ converges in H. This is a fundamental difficulty of the infinite-dimensional setting. We refer the reader to Section 3.5 in Stuart (2010) for a further discussion.

We perform a simple numerical experiment to demonstrate this. We fix the data measure $\mu = N(0, C_1)$ where C_1 has form Equation 35 with $\alpha_1 = 3$, $\sigma_1 = 4$, and $\tau_1 = 1$. We set $\mu_0 = N(0, C_2)$ where C_2 has the form Equation 36 with $\alpha_2 = 2$, $\sigma_2 = 0.2$, and $\tau_2 = 1$. We fix a FNO architecture wich retains 32 modes and has a width of 64 and re-train it at different resolutions of the data and noise so as to minimize either

$$\mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta \sim \mu_0} \|\eta - G_\theta(u + \eta)\|^2$$

or

$$\mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta \sim \mu_0} \|C^{-1/2}(\eta - G_\theta(u + \eta))\|^2.$$

We then compare the test errors, simply defined as the same quantity as the training loss but evaluated on new draws from the data and noise distributions. The results are shown in Figure 9. We see that the blue curve stays constant, confirming that the FNO can learn to represent noise from the Sobolev space $\dot{H}^{3/2}(\mathbb{T}; \mathbb{R})$ in a discretization invariant way. On the other hand, when training with the pre-conditioner $C^{-1/2}$, we see the test error grow as we increase the resolution. This demonstrates the effect of the infinity in the loss function.
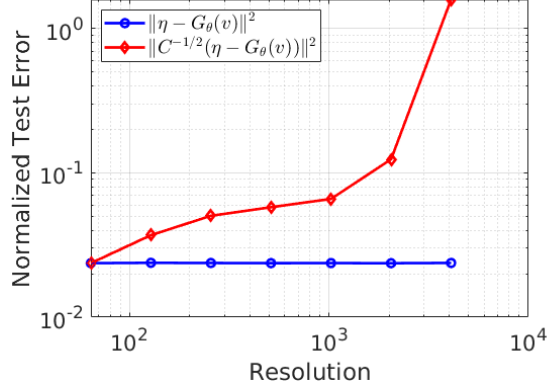
Figure 9: **Noise Regularity** Test error when training with two different loss functions across different resolutions. Red curve is re-scaled so that it matches the error of the blue curve at the lowest resolution for the sake of visualization.

Appendix I. Smoothing Operators

To illustrate the effect of the choice of the Camerion-Martin space (CM-space) of a (forward) noise process, we train an FNO architecture with varying training conditions. Specifically, we simulate that samples from data distribution, $u \sim \mu$, which may or may not be in the noise's CM-space.

Similar to Section F.2, we consider a mixture of two Gaussians as data distribution μ where $d = 1$, $D = (0, 2\pi)$, $f_1 = -10/6x + 5$, $f_2 = -f_1$, and $p = 0.5$. For its covariance C_1, we choose $\alpha_1 = 1.5$, $\sigma_1 = 3$, and $\tau_1 = 3$.

We compare four training conditions by varying the noise covariance C_2; (a) noise process uses white noise, (b) all data samples lie on the CM-space of the noise covariance, (c) there exist samples $u \notin C_2^{1/2}(H)$, and (d) apply a smoothing operator A, while the same covariance as in (c). For (b), we choose $\alpha_1 = 1$, $\sigma_1 = 1.73$, and $\tau_1 = 3$, and for (c) and (d), we use $\alpha_1 = 2$, $\sigma_1 = 10$, and $\tau_1 = 3$. For smoothing operator A in (d), we use a Gaussian blur such that $A(H) \subseteq C_2^{1/2}(H)$.

We train models at a resolution of 512 for $5,000$ iterations. Unlike Section F.2, we only trained in the resolution 512 and sample with varying resolutions (See Figure 10).

From (a), we can observe that the trained model successfully generates the samples in resolution 512, the same resolution during training. However, when the trained model tries to generate higher-resolution samples, its samples collapse into modes. While the parametric score operator is discretization-invariant, due to independent Gaussian noise, the induced distribution from the model is not in function-valued space.

On the contrary, when $C_2^{1/2}(H)$ is sufficiently large enough to include all samples from the data distribution, the proposed method learns the data distribution. Moreover, the model successfully generates samples in higher dimensions, as the model distribution is a measure in a function-valued space.

If samples are not in $C_2^{1/2}(H)$, the model fails to learn the data distribution as in Figure 10 (c). As we discussed in Section 4.2, however, one can apply a smoothing operator A so that

46

$A(H) \subseteq C_2^{1/2}(H)$. This results in losing some information about the data, depending on the choice of the smoothing operator. Here, high-frequency noises are cut out as we use a Gaussian blur.

The results demonstrate that the proposed method will learn the distributions in function space; thus, it is discretization invariant. Moreover, the result further implies that the choice of noise process and smoothing operator will determine which perspectives of data distributions the models will learn. Furthermore, the experiment led to several open questions for choosing noising processes most suitable for applications.
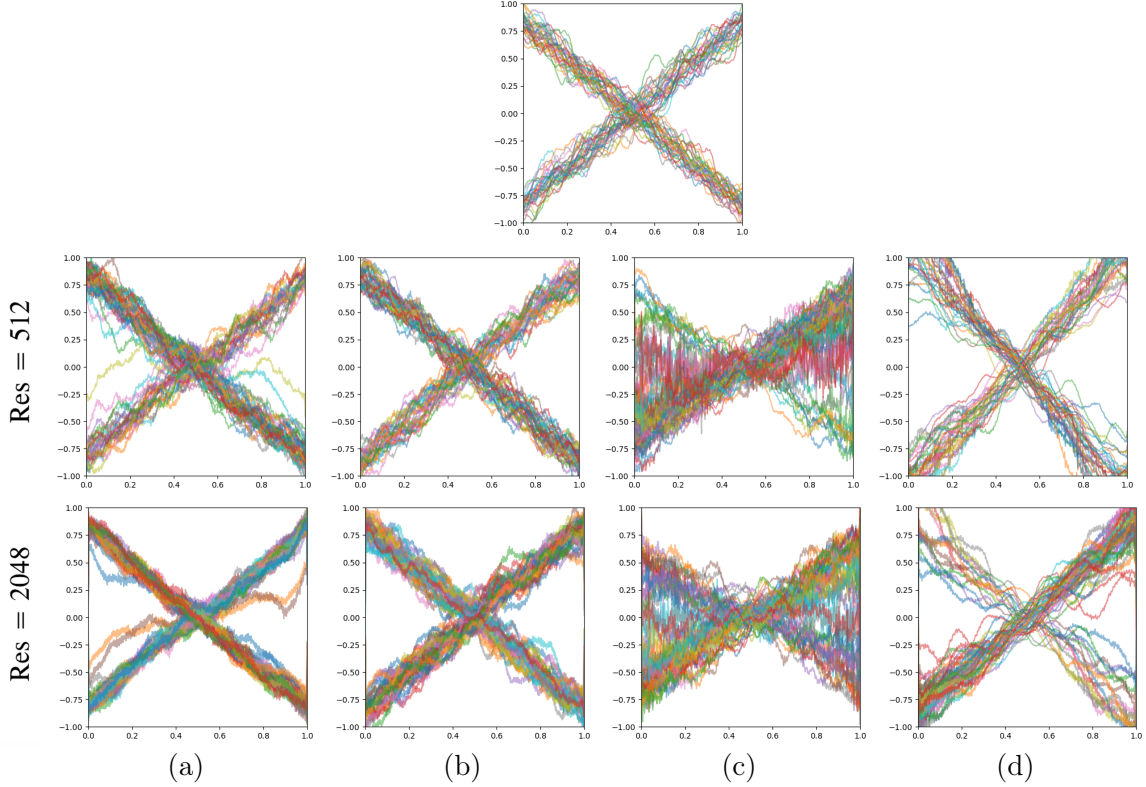
Figure 10: **Smoothing Operators (Section 4.2)** First row illustrates 128 sample paths from the data distribution μ, a Mixture of two Gaussians. The second and third row illustrate generated samples from trained models. We train a FNO architecture with varying choice of the noise' covariance C_2; (a) independent Gaussian noise process, (b) for all $u \in \mu$, $u \in C_2^{1/2}(H)$, (c) $\exists u \notin C_2^{1/2}(H)$, and (d) apply a smoothing operator while the same covariance as in (c). For smoothing operator A in (d), we use a Gaussian blur such that $A(H) \subseteq C_2^{1/2}(H)$. The models are trained in a 512 resolution, and generate with varying resolutions, such as 512 (Second row) and 2048 (Third row).

Appendix J. Experimental Details

In all examples we train by picking $\mathcal{I} = [10]$ and sample with Algorithm 1 by fixing $M = 200$ and $\epsilon = 2 \times 10^{-5}$. We choose $\sigma_1 = 1.0$ and $\sigma_{10} = 0.01$ with all other σ parameters defined
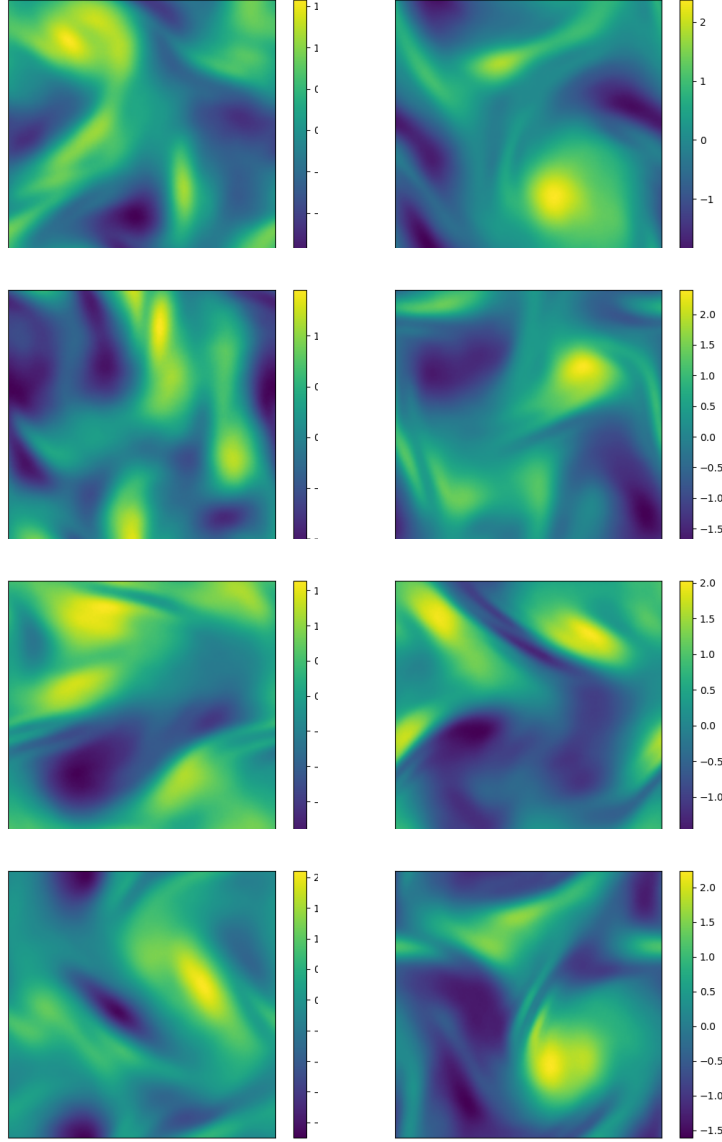
Figure 11: **Navier-Stokes.** Uncurated samples at the resolution 1024×1024 from our diffusion model trained on a dataset at the resolution 128×128.

by a geometric sequence. We train with a combined loss defined by Equation 18 where we re-scale the the noise by σ_t^{-1} and the score by σ_t, following Song and Ermon (2019). In particular, our model learns to approximate $v \mapsto \sigma_t^{-2}\left(RD_{H_{\mu_t}}\Phi(u,t) - v\right)$. Note that the σ_t^{-2} term is canceled by the adaptive time-step in Algorithm 1, however, as in Song and Ermon (2019), we find that this re-scaling significantly improves performance for all models. We
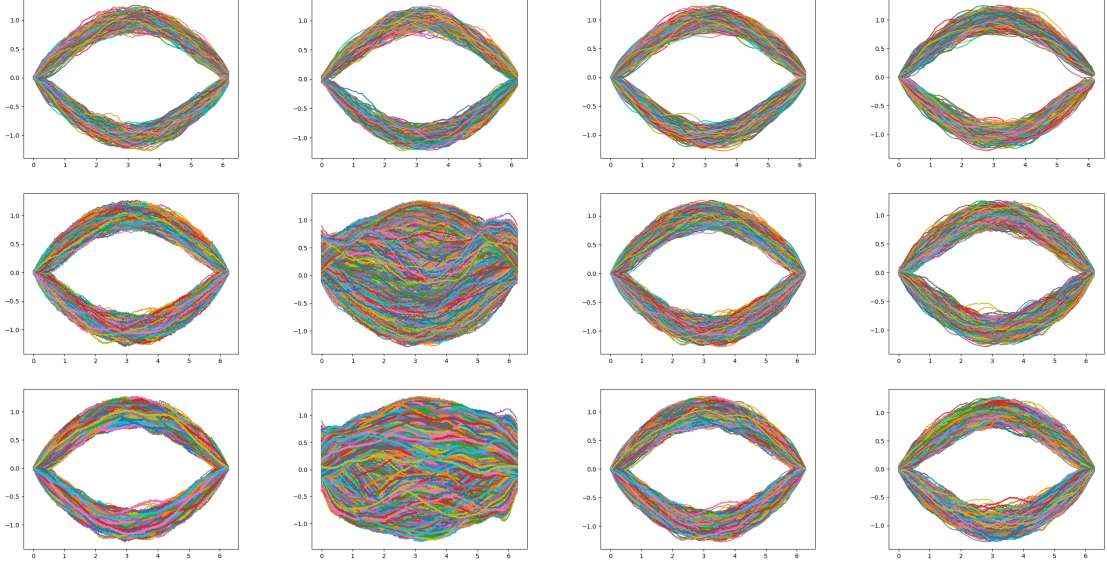
Figure 12: **Guassian mixture.** Each row represents the resolutions $64, 512, 2048$ respectively. The first and third columns show samples from ν_T (the perturbed data measure with the smallest amount of noise added) for the white noise and trace-class noise respectively. The second and fourth columns show samples generated from the trained model using white noise and trace-class noise respectively.

leave a theoretical analysis of this for future work. To be explicit, our loss function is

$$\min_{\theta} \frac{1}{|\mathcal{I}|} \sum_{t \in \mathcal{I}} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta_t \sim \mu_t} \left\| \frac{\eta_t}{\sigma_t} + \sigma_t F_\theta(u + \eta_t, t) \right\|^2$$

where $\| \cdot \|$ is the $L^2(D; \mathbb{R})$ norm and D is problem-dependent. We train with the Adam optimizer for a total of 300 epochs and an initial learning rate 10^{-3}, which is decayed by half every 50 epochs.

J.1 Gaussian Mixture

We consider the problem setting of Section F.2 with $d = 1$ and $D = (0, 2\pi)$. We pick $f_1 = \sin(x/2)$ and $f_2 = -f_1$ as well as $p = 0.5$. We generate $N = 10,000$ samples for training at a resolution of 2048 and subsample these to obtain all other datasets. For the data covariance C_1, we choose $\alpha_1 = 3$, $\sigma_1 = 3$, and $\tau_1 = 3$, and, for the noise covariance C_2, we choose $\alpha_1 = 0.6$, $\sigma_1 = 0.5$, and $\tau_1 = 0.1$. We train with a FNO architecture wich retains 48 modes and has a width of 128. We re-train the model at each separate resolution. In Figure 12 we visualize samples from each of the models trained with white noise as well as the trace-class noise with covariance C_2 at the resolutions $64, 512$, and 2048. We point out that the models trained at high resolutions with white noise completely fail to capture the right distribution.

49

J.2 Navier-Stokes

We consider a problem setting similar to Section F.3.1. We fix the initial condition $u(\cdot, 0) = 0$ and instead generate random forcings f from the Gaussian ρ. The same theory as in Section F.3.1 still applies and the pushforward is non-Gaussian due to the non-linearity of the PDE. We solve it up to the final time $T = 5$ with the pseudo-spectral scheme of Chandler and Kerswell (2013) with $\epsilon = 1/500$ and $N = 10,000$ samples for training. We pick $\alpha_1 = 4$, $\sigma_1 = 3\sqrt{3}$ and $\tau_1 = 3$ for the reference Gaussian ρ, following (De Hoop et al., 2022). All data is generated with a 1024×1024 resolution and the 128×128 is created from it by sub-sampling. With train our mode with a FNO architecture retaining 80×80 modes with a width of 64. We pick the parameters $\alpha_2 = 1.5$, $\sigma_2 = 4$, and $\tau_2 = 5$ for our noise covariance. In Figure 11, we show more samples generated by the model, performing zero-shot super-resolution.

J.3 Volcano Dataset

For the volcano experiments, we use the loss formulation from Equation (18) but employ a noise schedule $\{\sigma_1, \ldots, \sigma_L\}$ (as is standard with SBGMs). Inspired by Song and Ermon (2020) we optimize a weighted variant of it where F_θ is preconditioned with σ_i:

$$\min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\eta_i} \| \eta_i + \sigma_i F_\theta(u + \eta_i, \sigma_i) \|^2 \tag{40a}$$

$$= \min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \mathbb{E}_{i \sim U(1, L)} \| \mathbf{L}_i \epsilon + \sigma_i F_\theta(u + \mathbf{L}_i \epsilon, \sigma_i) \|^2 \tag{40b}$$

$$= \min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \mathbb{E}_{i \sim U(1, L)} \| \sigma_i \mathbf{L} \epsilon + \sigma_i F_\theta(u + \sigma_i \mathbf{L} \epsilon, \sigma_i) \|^2, \tag{40c}$$

$$= \min_{\theta \in \mathbb{R}^p} \mathbb{E}_{u \sim \mu} \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{I})} \mathbb{E}_{i \sim U(1, L)} \| \sigma_i \big(\mathbf{L} \epsilon + F_\theta(u + \sigma_i \mathbf{L} \epsilon, \sigma_i) \big) \|^2 \tag{40d}$$

where $\eta_i \sim \mathcal{N}(0, \mathbf{C}_i^2) = \mathcal{N}(0, \sigma_i^2 \mathbf{C}^2)$, and via the reparameterisation trick this can be re-written as $\eta_i = \sigma_i \mathbf{L} \epsilon$, where $\mathbf{L} = \text{chol}(\mathbf{C})$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Here, we use a covariance computed using the RBF kernel over a 2D meshgrid representing the coordinates of the image (function):

$$\mathbf{C}(x)_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\gamma^2} \right), \quad i, j \in \{1, \ldots, s^2\} \tag{41}$$

where $x \in [0, 1]^{s^2 \times 2}$ for a spatial resolution of s, and $\mathbf{C} \in \mathbb{R}^{s^2 \times s^2}$. The hyperparameter γ controls the smoothness of the noise, with larger values indicating higher levels of smoothness. Example sample noises are illustrated in Figure 13.
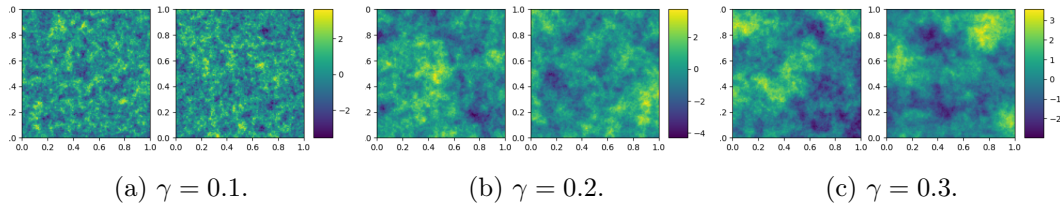

(a) $\gamma = 0.1$.   (b) $\gamma = 0.2$.   (c) $\gamma = 0.3$.

Figure 13: **Samples with RBF kernels**: Sample noises for varying smoothness parameters γ of RBF kernels.

As suggested in Song and Ermon (2020), we redefine $F_\theta(\cdot, \sigma) = F_\theta(\cdot)/\sigma$ since the authors noted that this makes the noise prediction task more robust to a wide range of noise scales.

Architecture   The architecture we use is a U-shaped neural operator (UNO). This architecture consists of a series of Fourier neural operator blocks (FNOBlocks) which progressively downsample or upsample the input. We use the implementation of FNOBlocks from the Neural Operators library Li et al. (2020a); Kovachki et al. (2023). Concretely, for the Volcano dataset the example u has a spatial dimension of 120×120 with two channels, and is lifted from $2 \rightarrow 128$ channels with a preprocessing convolutional layer, which also pads the 120 pixel image (function) to 128 pixels. Afterwards, the lifted input is run through four FNOBlocks which progressively lift the channel dimension and spatial dimensions to $(128 \rightarrow 256 \rightarrow 512 \rightarrow 512)$ and $(128 \rightarrow 96 \rightarrow 64 \rightarrow 32)$, respectively. A similar set of blocks is used in the decoder block along with skip connections.

For FNOBlock, we use Tucker factorisation with a rank of 0.1. In order to ensure the number of learnable parameters does not explode, we also constrain the number of Fourier modes in each FNOBlock to be 50% of the number of input channel dimensions for each block. The resulting U-shaped neural operator contains a total of 142M learnable parameters.

Training Details   We train the models using ADAM Kingma and Ba (2014) with default moving average hyperparameters $\beta = (0.9, 0.999)$, with a learning rate of $2e^{-4}$. For the noise schedule, we employ a geometric schedule using $(\sigma_1, \sigma_L) = (30, 0.01)$, for $L = 500$ time steps. The number of SGLD iterations per timestep is $T = 3$ and we use a step size of $6e^{-6}$. At generation time, we use annealed SGLD and run the Markov chain on $u_0 \sim \mathcal{N}(0, \mathbf{C})$.

Hyperparameters   We use circular skewness and circular variance, moments of circular variables which were originally proposed in GANO (Rahman et al., 2022) Given an image (function) $u \in \mathbb{R}^{s^2 \times 2}$ we can define its angle as $\theta = \text{atan2}(u_{:,1}, u_{:,2})$. If we define $R_p(\theta) = \frac{1}{s^2}\sqrt{z_p(\theta)}$ for $z_p(\theta) = (\sum_k^{s^2}(\cos(p\theta_k) + i\sin(p\theta_k))^2$ and $\varphi_p(\theta) = \arg(z_p(\theta))$ (for s^2 spatial dimensions) then:

$$w_{\text{var}}(\theta) = 1 - R_1(\theta), \quad w_{\text{skew}}(\theta) = \frac{R_2(\theta)\sin(\varphi_2(\theta) - 2\varphi_1(\theta))}{(1 - R_1(\theta))^{3/2}} \tag{42}$$

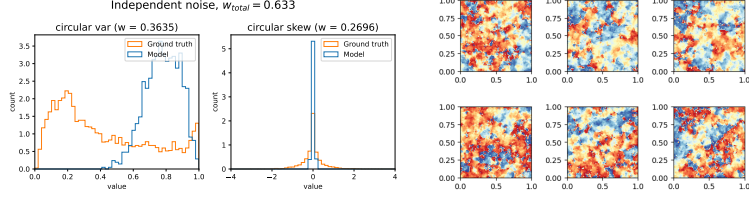where $i = \sqrt{-1}$ and k is a summation over the spatial dimensions s^2.
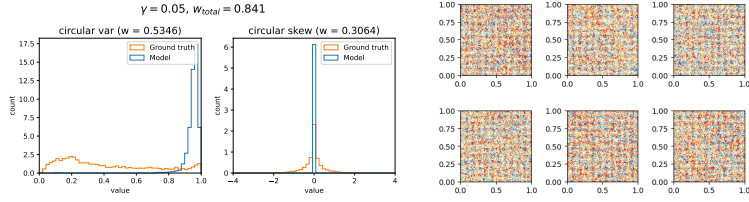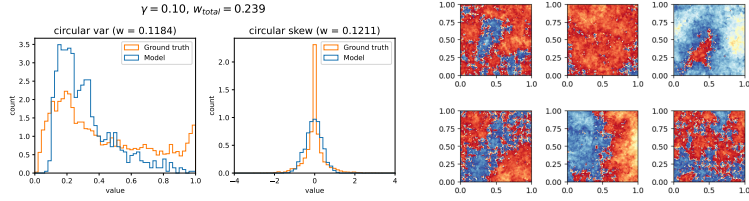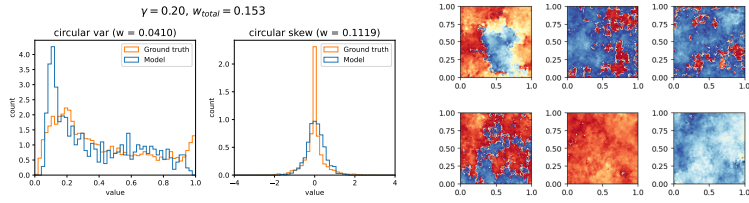
(a) Independent Gaussian noise, $w_{\text{total}} = 0.633$

(b) RBF noise, $\gamma = 0.05$, $w_{\text{total}} = 0.841$

(c) RBF noise, $\gamma = 0.1$, $w_{\text{total}} = 0.239$

(d) RBF noise, $\gamma = 0.2$, $w_{\text{total}} = 0.153$

Figure 14: Super-resolution experiments for DDO. Each subplot corresponds to a different level of RBF smoothness, which is denoted by γ (larger γ corresponds to smoother noise). DDO is trained on 60×60 resolution functions, and 120×120 functions are produced as per Section 5.3. Variance and skew statistics are computed from $M = 1024$ samples and the Wasserstein distance is computed between those samples and the corresponding statistics from the 120×120 resolution version of the training set (i.e. the original resolution). We can see that both the independent noise experiment and $\gamma = 0.05$ experiment fail to produce plausible examples.
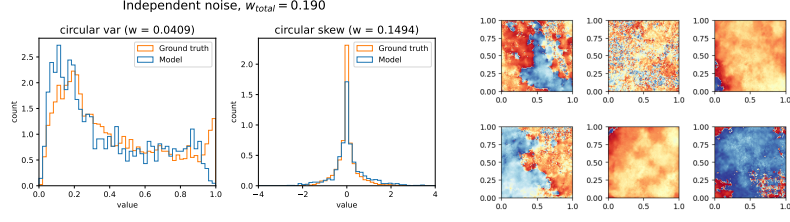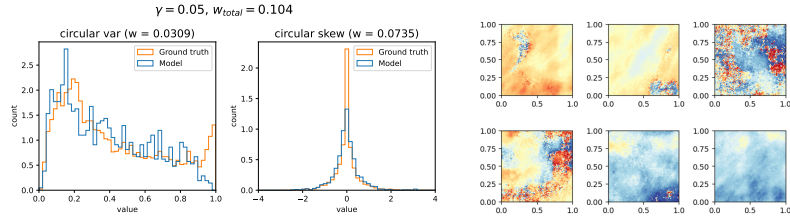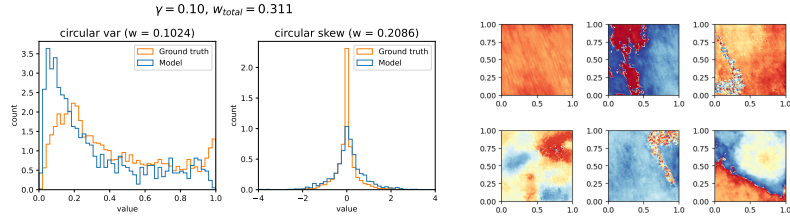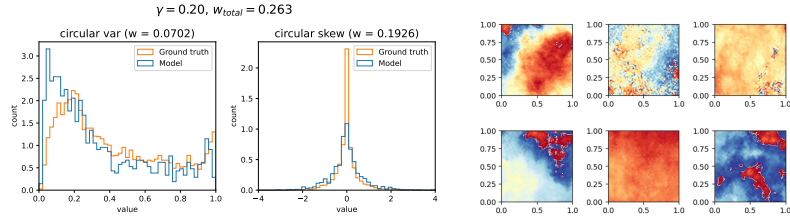
(a) Independent Gaussian noise, $w_{\text{total}} = 0.187$

(b) RBF noise, $\lambda = 0.05$, $w_{\text{total}} = 0.104$

(c) RBF noise, $\lambda = 0.1$, $w_{\text{total}} = 0.311$

(d) RBF noise, $\lambda = 0.2$, $w_{\text{total}} = 0.263$

Figure 15: Super-resolution experiments for GANO. Each subplot corresponds to a different level of RBF smoothness, which is denoted by γ (larger γ corresponds to smoother noise). GANO is trained on 60×60 resolution functions, and 120×120 resolution functions are produced as per Section 5.3. Variance and skew statistics are computed from $M = 1024$ samples and the Wasserstein distance is computed between those samples and the corresponding statistics from the 120×120 resolution version of the training set (i.e. the original resolution).

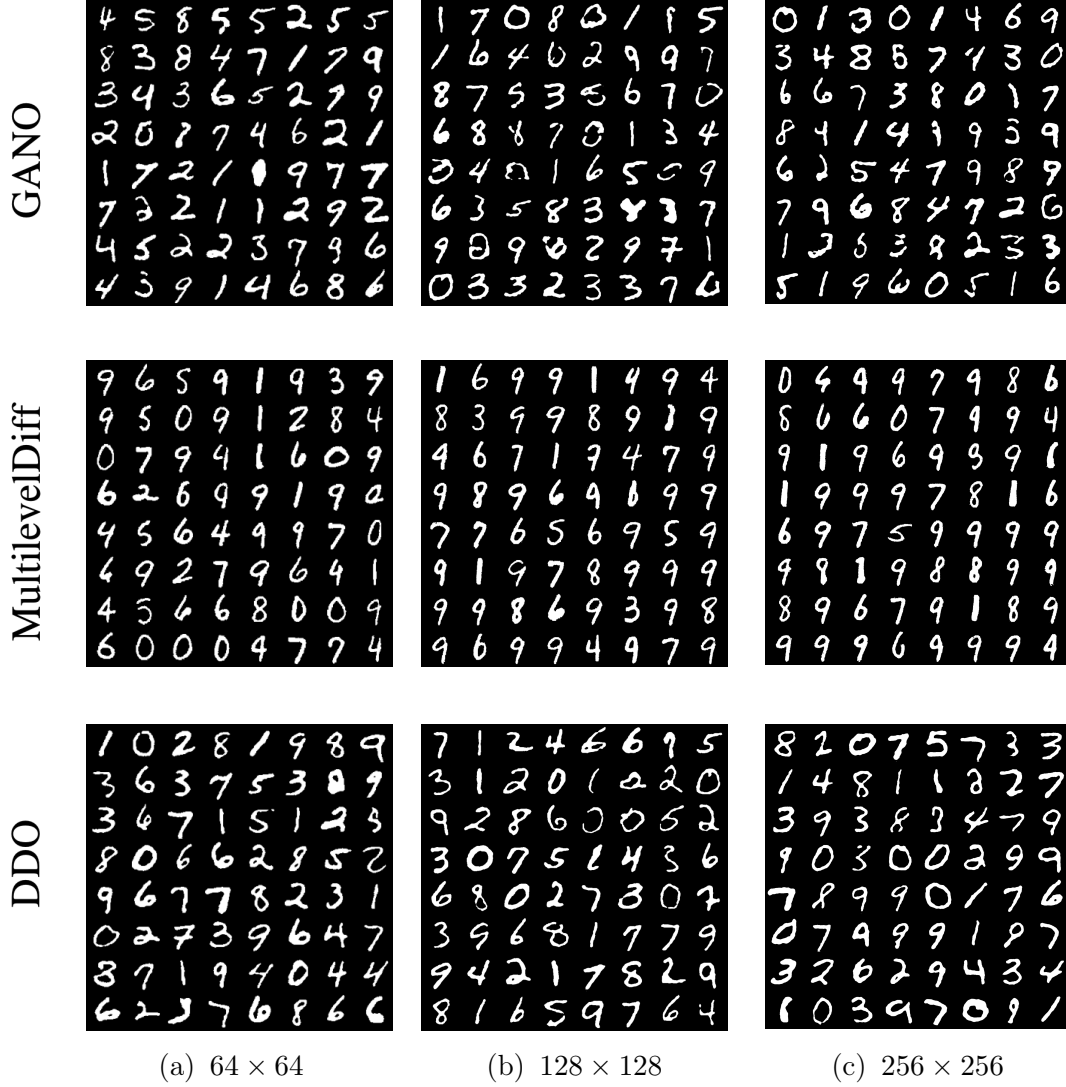(a) 64×64 (b) 128×128 (c) 256×256

Figure 16: **Additional generated Samples** (Section 5.4): Additional samples (masked) of the learned GANO, MultilevelDiff, and DDO models at various resolutions; (a) 64×64, (b) 128×128, and (c) 256×256-resolutions. 64 samples are plotted for all resolutions. All models are trained on 64×64-resolution images, which are upsampled from 32×32-resolution observations of 2D SDFs.

J.4 MNIST-SDF Dataset

For the noising processes in the MNIST-SDF experiments, we adopt Equation 32 in Appendix E. More specifically, we follow a more generalized formulation described in Kingma et al. (2021). For $t \in (0, 1]$, we first define u_ts as in,

$$u_t = \alpha_t u + \sigma_t \eta, \tag{43}$$

where $\eta \sim \mu_0 = N(0, C)$ and $u \sim \mu$. In this experiment, we use *variance-preserving* setting for the forward process; thus, α_t and σ_t will satisfy $\alpha_t^2 + \sigma_t^2 = 1$ for $t \in [0, 1]$. The approximate backwards conditional in Equation 34 will be re-written as $u_s | u_t$ for $\forall s, t$ s.t. $0 < s < t \leq 1$;

$$u_s = M_\theta(u_t, t) + \sigma_{\text{back},s,t}\eta, \tag{44}$$

where $\eta \sim \mu_0$. M_θ and $\sigma_{\text{back},s,t}$ are written as

$$M_\theta(u_t, t) = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}u_t - \frac{\alpha_{s|t}\sigma_{t|s}^2}{\sigma_t^2}\left(u_t - \sigma_t F_\theta(u_t, t)\right) \tag{45}$$

$$\sigma_{\text{back},s,t} = \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}, \tag{46}$$

where $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s} = \sigma_t^2 - \alpha_{t|s}^2\sigma_s^2$.

The aforementioned formulations provide us several benefits. Most importantly, we can choose arbitrary sequence of discrete time steps in order to generate the learned distribution. This implies that we can train F_θ at all $t \in (0, 1]$ and choose the sampling sequence later after training. As a result, instead of the training loss Equation 33, we obtimize the following objective,

$$\mathbb{E}_{u\sim\mu, \eta\sim\mu_0, t\sim U(0,1)}\|F_\theta(\alpha_t u + \sigma_t \eta, t) - \eta\|^2. \tag{47}$$

In the following sections, we will continue on the trace-class noises used in the experiments and other details.

GRFs with p-th Order Exponential Covariance Functions We observe that the models often generate artifacts with the blurring diffusions when the noise process is non-periodic Gaussian measures (Section F.1), as the Gaussian smoothing operator is periodic. Thus, we use periodic Gaussian measures with p-th order exponential covariance function, whose kernel function is written as,

$$k(x, x') = \sigma \exp\left(-\left|\frac{x - x'}{l}\right|^p\right), \tag{48}$$

where l is a length scale, p is order, and σ is a magnitude. Note that the RBF kernel used in the volcano experiments is a special case of Equation 48

As the data is assumed to be sampled on a lattice in $\mathbb{R}^d$, we employ the convolution construction of the Gaussian measures for efficient sampling (Higdon, 2002). For a given exponential kernel k, the corresponding convolution kernel c is a function of the absolute difference x and is written as

$$c(x) = \mathcal{F}^{-1}\left[\sqrt{\mathcal{F}[k]}\right](x)$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ are Fourier and inverse Fourier transform, respectively. We denote the closed form $\mathcal{F}(k)(\omega)$ for p-th order exponential kernel as $\phi_p(\omega)$, and ϕ_p is written as

$$\phi_p(\omega) = \begin{cases} \frac{2}{1+\omega^2}, & p = 1, \\ 2\pi c_p \frac{p|\omega|^{\frac{1}{1-p}}}{2|p-1|} \int_0^1 U_p(x) \exp\left(-|\omega|^{\frac{p}{p-1}} U_p(x)\right) dx, & p \in (0, 2] \setminus \{1\}, \end{cases}$$

55

where c_p and U_p are defined as

$$c_p = \frac{p}{2^{\frac{p+1}{p}} \Gamma\left(\frac{1}{p}\right)} \quad \text{and} \quad U_p(x) = \left(\frac{\sin\left(\frac{\pi x p}{2}\right)}{\cos\left(\frac{\pi x}{2}\right)}\right)^{\frac{p}{1-p}} \frac{\cos\left(\frac{\pi x(p-1)}{2}\right)}{\cos\left(\frac{\pi x}{2}\right)},$$

and Γ is the Gamma function. The detailed information about ϕ_p, including its derivation, is described in Dytso et al. (2018).

Architecture and Training Details   This section provides a detailed description of the experimental configuration used in our study. Before elaborating on the hyperparameters, we introduce a neural operator in the image experiments. We modify the UNet network architectures (Ronneberger et al., 2015), which has been widely adopted in the context of diffusion-based generative models (Ho et al., 2020; Song et al., 2020; Nichol and Dhariwal, 2021) and elaborated with regards to neural operators (Rahman et al., 2022, 2023). We adopt the Improved Denoising Diffusion Probabilistic Models (IDDPM, Nichol and Dhariwal 2021), whose time-conditional modulation uses elementwise affine transformations (shift and scale) instead of shifting-only modulation introduced in Ho et al. (2020). To do that, we introduce four major modifications. First, we replace all regular convolutional layers with spectral convolutions. Second, we apply group normalizations on Fourier spaces instead of Euclidean spaces. Here, the normalization coefficients will be computed using the first k modes in Fourier spaces, but the coefficient will be applied to all modes; thus, the normalization works in a resolution-invariant manner. Third, we deprecate the self-attention layers and dropouts. Moreover, for the downsample and upsample operations, which are one of the key components of the UNet-like hierarchical network architectures, we adopt filtered downsample and upsample algorithms discussed in Karras et al. (2021). Therefore, we conclude that the resulting deep learning architecture is a valid neural operator as it is a combination of convolutional operators, normalizations defined in Fourier spaces, and point-wise transformations.

Specifically for the MNIST-SDF experiment, we set the number of base channels to 64 and use the three-stage model. The channel multipliers for each stage are set to 1, 2, and 2, respectively. For each stage, four residual blocks are used. At the first stage, all spectral convolution layers comprise 32 modes, and we halve the modes as the stage increases.

We chose the cosine noise scheduling following IDDPM with a variance-preserving form for the noise schedule. We follow the sine schedule introduced in Hoogeboom and Salimans (2022) for the blurring schedule. We perform early stopping based on the FID of 5000 generated samples relative to a subset of the training data. We use the exponential moving average (EMA) technique for the evaluation and test; we set the EMA rate to be 0.999. We describe the hyperparameter details in the Table 4.

Baseline Models   For GANO (Rahman et al., 2022), we follow the methodology outlined in its paper and codebase,[4] with specific configurations adapted to the study. In particular, the input random field provided to the generator is the Gaussian Random Field (GRF) described in Equation 35, with parameter settings of $\sigma_1 = 1$, $\alpha_1 = 1.5$, and $\tau_1 = 1.0$. For the

---

4. https://github.com/neuraloperator/GANO

Table 4: Training details of DDO for the MNIST-SDF experiments

Architecture	Base channels	64
	# of ResBlocks per stage	4
	Channel multiplier	1,2,2
	# of modes per stage	32, 16, 8
	Activation function	GeLU[‡]
	# of params	258M
Diffusion	Noise schedule	Cosine
	$\log(\alpha_0^2/\sigma_0^2)$	10
	$\log(\alpha_1^2/\sigma_1^2)$	-10
Blurring	Blurring schedule	Sine
	d_0[§]	0.05
	d_1	0.25
Learning	Optimizer	Adam, β_1=0.9, β_2=0.999
	Learning rate	0.0001
	Batch size	32
	# of iterations	2M
Sampling	# of steps	250
GRFs	Length scale	0.05
	Magnitude	1
	Order	2

[§]d_0 and d_1 are frequency scalings at $t = 0$ and d_1, respectively.

[‡]Gaussian Error Linear Units function(Hendrycks and Gimpel, 2016).

generator utilizing the U-NO architecture (Rahman et al., 2023), the number of modes is set to 32 (with half-modes set to 16), and the number of channels is set to 64.

Training of GANO is conducted using a batch size of 32, with a gradient penalty parameter $\lambda = 10.0$, to stabilize the adversarial learning process. The learning rate is set to 0.0001, and optimization is performed using the default ADAM optimizer, consistent with the configuration in DDO. For evaluation, the Exponential Moving Average (EMA) technique from DDO's experimental framework is employed.

The experiments involving MultilevelDiff are conducted using the official codebase available at its repository.[5] For the score operator, the Fourier Neural Operator (FNO) is employed with a configuration of 32 modes (half-modes set to 16) and 256 channels. The noise perturbation is applied with the prior, which combines a spectral convolution-based noise as well as a random field with a fixed kernel. Further details about the hyperparameters used can be found in Table 1 of Hagemann et al. (2023).

The training process of MultilevelDiff follows the same approach as in DDO to ensure consistency with established methodologies. For evaluation, the Exponential Moving Average (EMA) technique is not used. Instead, evaluations are performed directly without applying EMA, yielding results that remain robust despite this omission.

---

5. https://github.com/PaulLyonel/multilevelDiff

Table 5: Training details of DDO for Bayesian Inverse Problems

Architecture	Base channels	32
	# of ResBlocks per stage	4
	Channel multiplier	1,2,2
	# of modes per stage	48, 32, 16
	Activation function	Swish[‡]
	# of params	139M
Diffusion	Noise schedule	Cosine
	$\log(\alpha_0^2/\sigma_0^2)$	10
	$\log(\alpha_1^2/\sigma_1^2)$	-10
Blurring	Blurring schedule	Sine
	d_0[§]	0.01
	d_1	0.05
Learning	Optimizer	Adam, β_1=0.9, β_2=0.999
	Learning rate	0.00001
	Batch size	16
	# of iterations	300,000
Sampling	# of steps	500
GRFs	Length scale	0.02
	Magnitude	1
	Order	2

[§]d_0 and d_1 are frequency scalings at $t = 0$ and d_1, respectively.

[‡]Swish function (Ramachandran et al., 2017).

J.5 Darcy Flow Bayesian Inverse Problem

In the Darcy Flow Bayesian Inverse Problem, we follow the methodology outlined in Appendix J.4, with adjustments made only to specific hyperparameters to suit the dataset better. Note that the data in this experiment is presented at a 64×64 resolution, while the MNIST-SDF experiment uses a 32×32 resolution for training. To condition on the 8×8-size observation using the neural operators, we upsample the observation to 64×64 and concatenate it with the original input to the operators.

In GANO, the input Gaussian Random Field (GRF) is configured with $\sigma_1 = 1$, $\alpha_1 = 1.0$, and $\tau_1 = 1.0$. Additionally, only modes up to 64 (with respect to DCT) are utilized, and thus the model will ignore any higher modes when it is asked to generate at higher resolutions. For the U-NO architecture in the generator, the number of modes is set to 64 (with half-modes set to 32), and the number of channels is configured to 64. Although experiments with increased channel numbers were tested, this configuration yields the best results.

For MultilevelDiff, the Fourier Neural Operator (FNO) is used with a configuration of 64 modes (half-modes set to 32) and 128 channels. For modifications to DDO, Table 5 provides a comprehensive overview of the changes applied.

As mentioned in Section 5.5, we also present the generated samples at 128×128 and 256×256 resolutions in Figures 17 and 18, respectively.
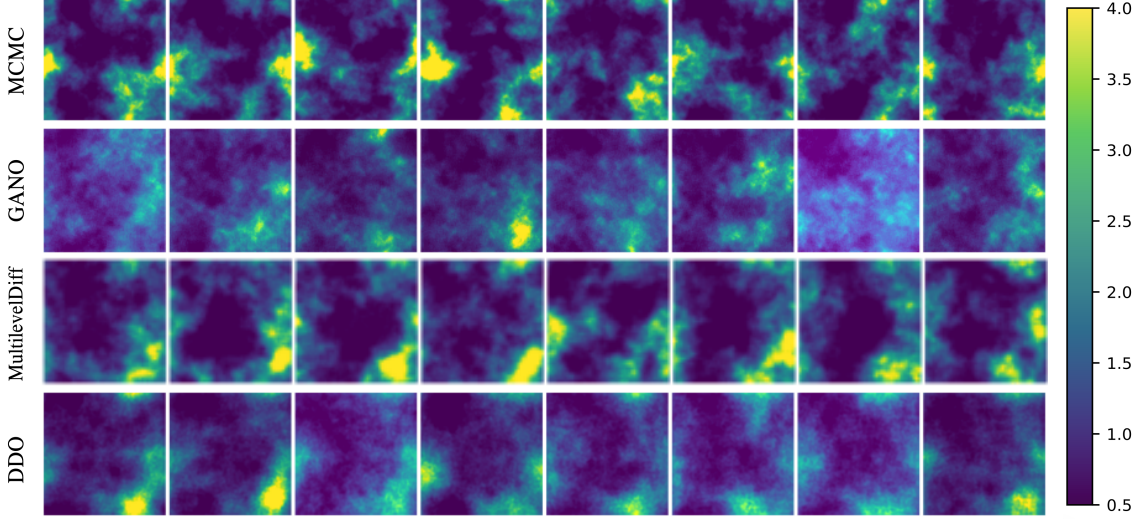
Figure 17: **Posterior Samples at 128×128 resolution** (Section 5.5): The samples from MCMC as well as the learned GANO, MultilevelDiff, and DDO models at 128×128 resolution. The models are trained at 64×64 resolution.
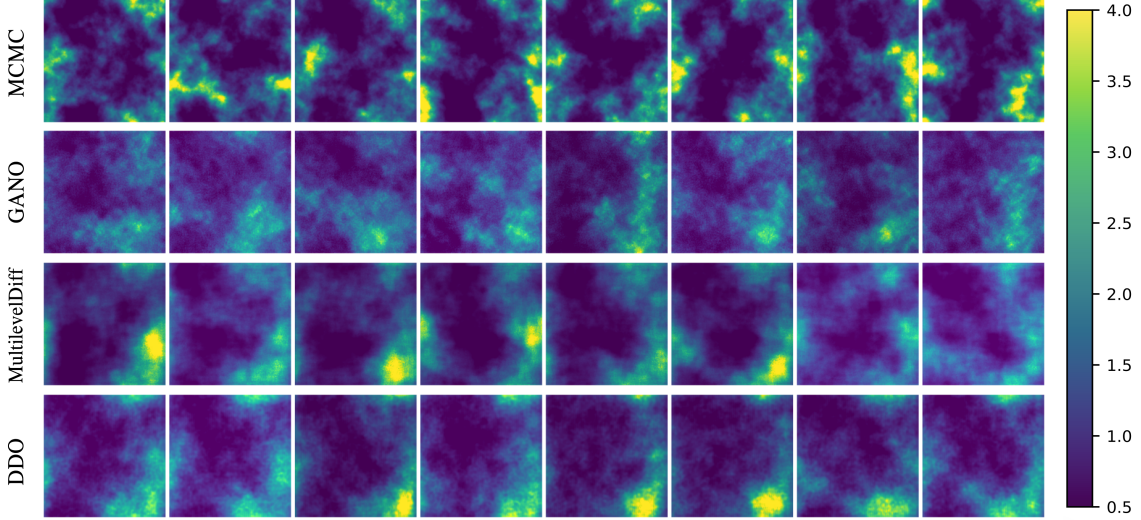

Figure 18: **Posterior Samples at 256×256 resolution** (Section 5.5): The samples from MCMC as well as the learned GANO, MultilevelDiff, and DDO models at 256×256 resolution. The models are trained at 64×64 resolution.

Appendix K. Neural Operators and Alias-free Models

This paper has conducted various experiments to evaluate the efficacy of DDO. In particular, GANO and MultilevelDiff were chosen as baselines. Except for a few cases, DDO mostly outperformed these methods, demonstrating the effectiveness of the proposed approach.

However, it is concerning that some notable methods, which have received significant attention in the field and propose function-valued diffusion models, were excluded from the baselines. For example, ∞-Diff (Bond-Taylor and Willcocks, 2024) successfully modeled high-fidelity image datasets such as CelebAHQ (Karras et al., 2018) and FFHQ (Karras et al., 2019), yet their method was left out of the baselines (see also Section 5.4). Briefly speaking, this method proposes a discrete-time latent diffusion model on function spaces, where the latent space is also a function space akin to the data space, and encoder-decoder structure maps between those two spaces; thus, function-valued diffusion models will model the latent distribution mapped from the data. Such a framework, including an encoder-decoder structure tailored for high-fidelity image modeling, makes direct comparison with DDO less fair.

To address these concerns and provide a meaningful comparison that includes DDO and related methods like ∞-Diff, we designed a simplified toy experiment. For this final experiment, we conducted a straightforward comparison to analyze the resulting differences in generated images from variations in neural operator design. Each model was given a set of two-dimensional coordinate pairs and trained to predict the RGB values of pixels at these coordinates by minimizing the mean squared error with respect to the true RGB values. As will be discussed below, the design of the neural operator plays a critical role in function-valued generative models. Therefore, this experimental setup abstracts away the influence of the application-specific frameworks, aiming to highlight the relative strengths and limitations of each neural operator.

Neural Operators in Function-valued Deep Generative Models   In deep generative modeling, defining the random variable for the generation and training of the parametric model is crucial. However, the design of the parametric model is also a very important factor. This is particularly true for function-valued deep generative models as compared to the finite-dimensional case because, in function space models, the parametric model—often referred to as a neural operator—must not only enhance expressivity but also satisfy a special property known as discretization invariance. The importance of discretization invariance lies in the fact that, while function-valued objects are theoretically handled, actual observations are ultimately discretizations of these functions.

Consequently, the design choices for the neural operator determine the performance characteristics of the parametric model on function space and the trade-offs involved. For example, spectral convolution, which leverages the Fourier series, theoretically and practically guarantees discretization invariance, but it is often susceptible to aliasing issues, such as ringing artifacts. On the other hand, continuous convolution-based methods may be more robust against aliasing but tend to overfit to specific discretizations, necessitating some form of regularization during training. Ultimately, the difference in the parametric model can have a more pronounced effect than the differences between generative models themselves.

For these reasons, conducting experiments to compare neural operators is both valuable and insightful. By doing so, we aim to provide a clearer understanding of the relative strengths and limitations of our proposed method in comparison to these baselines.

Experiment details   In the experiment, each model was learned to predict the RGB values of pixels at specified two-dimensional coordinates by minimizing the mean squared error relative to the true RGB values.

We used an image from the AFHQ animal faces dataset (Choi et al., 2020) for the experiment. All models were trained at a resolution of 256×256 and tested at various resolutions. For this, we compared the UNO used in GANO (Rahman et al., 2022), FNO used in Multileveldiff (Hagemann et al., 2023), Sparse Neural Operator (SparseNO) used in Bond-Taylor and Willcocks (2024), as well as a UNO variant used in our DDO model.

We observed that SparseNO requires significantly more computational memory compared to other models. Specifically, it was necessary to limit the size of a SparseNO model to fewer than 3 million parameters to fit within the memory constraints of a single NVIDIA A100 GPU. To ensure a fair comparison, we limited the size of all models to a similar scale (with the number of parameters kept below 3 million) and conducted all experiments using one NVIDIA A100 GPU. Additionally, we included a DDO model, whose size is 10 times larger for reference.

Results Figure 19 shows the predicted images at various resolutions generated by our DDO model and the baseline models. To highlight the characteristics of the neural operators, we enlarge the region within the red rectangle of the predicted images at 1024×1024 resolution in Figure 19 (d).

The FNO model performs well at 256×256 and 1024×1024 resolutions. However, it exhibits noticeable visual artifacts at 512×512, and ringing artifacts are observed at 1024×1024 due to its use of spectral convolution. Similarly, the UNO model performs well at 256×256 but displays consistent visual artifacts at higher resolutions, likely caused by its dependence on pointwise operations at every layer. The SparseNO model produces excellent results at the training resolution; however, artifacts appear at higher resolutions. These issues are likely due to the internal kernel resizing method, as SparseNO applies the bicubic interpolation to enlarge its fixed kernel to a higher resolution, which is not invariant to discretization and thus introduces errors.

In contrast, the DDO model produces consistent results across all resolutions. However, like FNO or UNO, it also shows some ringing artifacts at 1024×1024 due to its reliance on spectral convolution. Finally, the larger-scale DDO model demonstrates minimal ringing artifacts, with fine details remaining consistently preserved across resolutions. This better quality suggests that if the neural operator used in DDO can be successfully integrated into ∞-Diff instead of SparseNO, it may potentially lead to performance improvements.

In this section, we briefly introduce the characteristics of neural operators employed in function-valued diffusion models. However, discussions surrounding function-valued models extend far beyond these models. Extensive studies on neural operators (Li et al., 2020b,a; Kovachki et al., 2023; Fanaskov and Oseledets, 2023; Bartolucci et al., 2024) have explored the discretization invariance, leading to advancements in resolution-adaptive network architectures (Demeule et al., 2025). Additionally, alias-free network designs have been extensively discussed in works such as Karras et al. (2021). While these topics are highly relevant to the performance of function-valued generative models, a thorough investigation is beyond the scope of this paper and is left for future work.
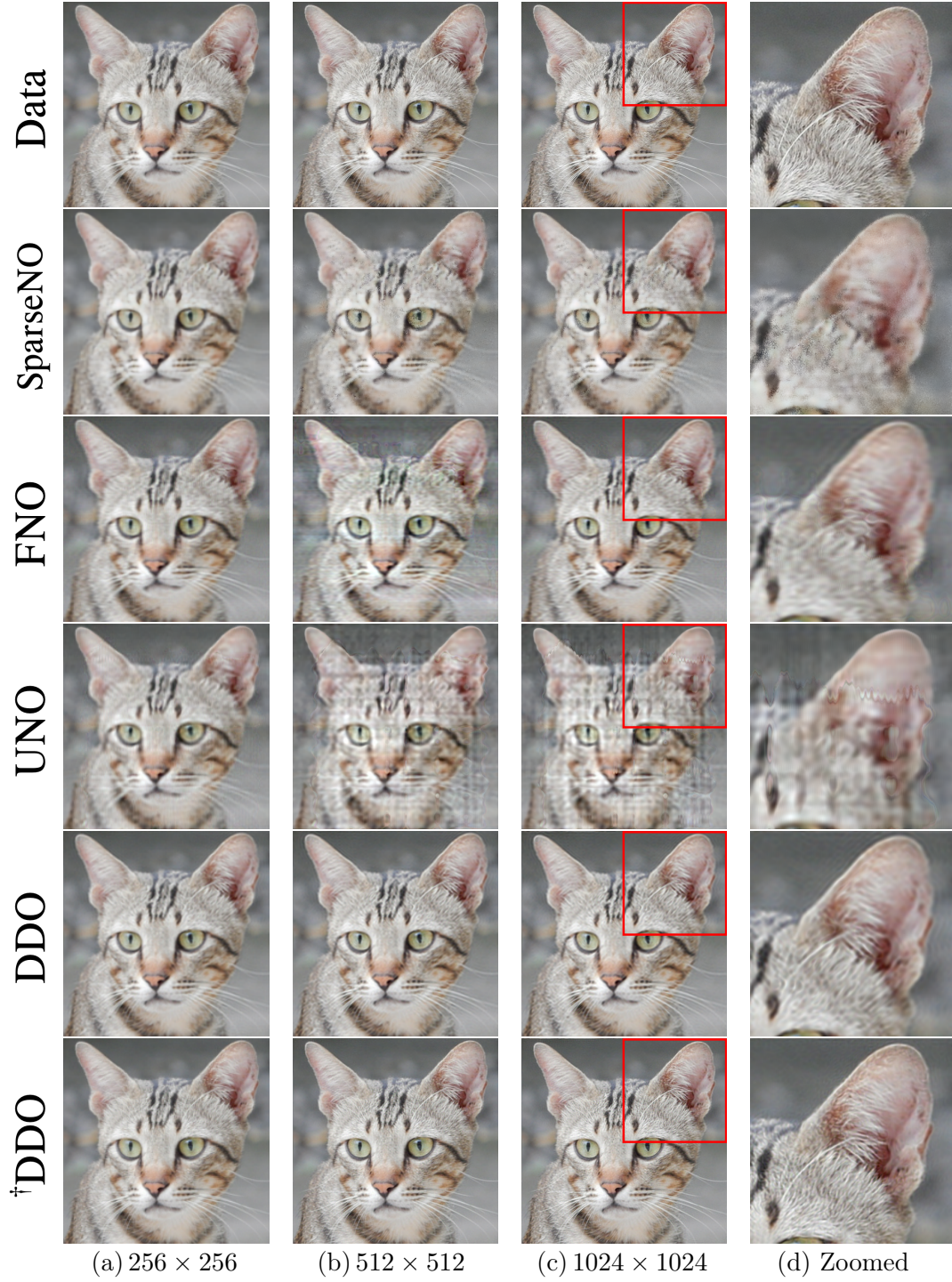
Figure 19: **Comparison of neural operators** (Section K): Each network architecture is trained to predict a image for given coordinate values. All models are trained at 256×256-resolution. Predicted images at various resolutions (a–c) are illustrated, and the red rectangular regions in (c) are enlarged and shown in (d). The number of parameters for all models is kept below 3 million (except for †DDO, which uses 10 times more parameters).