The Astrophysical Journal, 978:64 (14pp), 2025 January 01 https://doi.org/10.3847/1538-4357/ad8bc3

Diffusion-HMC: Parameter Inference with Diffusion-model-driven Hamiltonian Monte Carlo

Nayantara Mudur[1,2,3], Carolina Cuesta-Lazaro[2,3,4], and Douglas P. Finkbeiner[1,2,3]

[1] Department of Physics, Harvard University, 17 Oxford Street, Cambridge, MA 02138, USA; nmudur@g.harvard.edu
[2] Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA
[3] The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4] Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Received 2024 June 10; revised 2024 October 21; accepted 2024 October 23; published 2024 December 24

Abstract

Diffusion generative models have excelled at diverse image generation and reconstruction tasks across fields. A less explored avenue is their application to discriminative tasks involving regression or classification problems. The cornerstone of modern cosmology is the ability to generate predictions for observed astrophysical fields from theory and constrain physical models from observations using these predictions. This work uses a single diffusion generative model to address these interlinked objectives—as a surrogate model or emulator for cold dark matter density fields conditional on input cosmological parameters, and as a parameter inference model that solves the inverse problem of constraining the cosmological parameters of an input field. The model is able to emulate fields with summary statistics consistent with those of the simulated target distribution. We then leverage the approximate likelihood of the diffusion generative model to derive tight constraints on cosmology by using the Hamiltonian Monte Carlo method to sample the posterior on cosmological parameters for a given test image. Finally, we demonstrate that this parameter inference approach is more robust to small perturbations of noise to the field than baseline parameter inference networks.

Unified Astronomy Thesaurus concepts: Astrostatistics (1882); Cosmological parameters (339); Astronomical simulations (1857); Cosmology (343); Dark matter (353); Neural networks (1933)

1. Introduction

Ongoing and upcoming missions, such as the Dark Energy Spectroscopic Instrument (DESI),[5] the Vera C. Rubin Observatory's Legacy Survey of Space and Time,[6] and the Nancy Grace Roman Space Telescope[7] will map the cosmos at unprecedented resolution and volume. This has created a proportionate demand for simulations that can generate predictions from theory. Cosmological simulations, however, are expensive to run, and can only be generated for a limited set of initial conditions and points in parameter space.

The canonical summary statistic used for parameter inference is the two-point correlation function, or the power spectrum Pk at large, linear scales where perturbation theory holds. At smaller scales, however, gravitational collapse induces non-Gaussianity in the fields. This means the information content of cosmological fields is not fully captured by the power spectrum at the large scales alone. For example, recent work (C. Hahn et al. 2023; N.-M. Nguyen et al. 2024) derived much stronger constraints on σ_8 by going beyond the two-point correlation function at linear scales and analyzing nonlinear modes at smaller scales ($k > 0.25\,h\,\mathrm{Mpc}^{-1}$) or by extracting information at the field level. A multitude of other statistics—such as the marked power spectrum, the bispectrum, the wavelet scattering transform, and void probability functions—have also been devised (N. Hamaus et al. 2016; B. Régaldo-Saint Blancard et al. 2023; G. Valogiannis & C. Dvorkin 2022; E. Paillas et al. 2023) in an effort to capture higher-order correlations in non-Gaussian fields. Previous work (K. Heitmann et al. 2009; M. Mustafa et al. 2019; E. Paillas et al. 2023; G. Valogiannis et al. 2024; D. Sharma et al. 2024b) has addressed the prohibitive cost of simulations by creating emulators or surrogate models that learn to interpolate between predictions of a specific summary statistic between training points using formalisms such as Gaussian processes. More recently, simulation-based inference at the field level has also been used, where a likelihood model parameterized by a neural network is learned on the fields (C. Cuesta-Lazaro & S. Mishra-Sharma 2024; B. Dai & U. Seljak 2024).

Generative models are a class of machine learning approaches that enable one to simulate the ability to draw samples from a complicated target probability density and include variational autoencoders, normalizing flows, and generative adversarial networks (GANs). Diffusion generative models (J. Sohl-Dickstein et al. 2015; Y. Song et al. 2021) involve a forward diffusion (noising) process that transforms samples from the target distribution to those from the standard normal. In the denoising diffusion probabilistic model (DDPM; J. Ho et al. 2020), the noising process consists of a variance schedule β_t over a fixed number of time steps, T, that determines the incremental noise added to the image. Since the diffusion process can be formulated as a stochastic differential equation (SDE), the DDPM variance schedule corresponds to the discretization of this SDE. In the generative direction, a neural network or score model is used to parameterize the reverse transformation.

Diffusion models are alternatively referred to as score-based generative models since parameterizing the reverse diffusion

[5] www.desi.lbl.gov/the-desi-survey/

[6] www.lsst.org/

[7] roman.gsfc.nasa.gov/

process is equivalent to learning the "score" or $\nabla \log p_t(x)$ of the data (B. D. O. Anderson 1982; Y. Song et al. 2021). Since in high dimensions the target distribution invariably lies on a thin manifold, the incremental addition of the random noise blurs the distribution and makes the score progressively easier to learn. Diffusion models are the underlying mathematical framework that have given rise to the photorealistic image generation successes of DALL.E[8] and Stable Diffusion (R. Rombach et al. 2021), and have been shown to mitigate mode collapse, a phenomenon often encountered with GANs in which a generative model fails to generate multiple modes in a distribution. In scientific applications, they have been used for problems involving protein folding and ligand prediction and medical imaging reconstruction (Y. Song et al. 2022; G. Corso et al. 2023). They have been applied in astrophysics to reconstruction problems involving dust (N. Mudur & D. P. Finkbeiner 2022; D. Heurtel-Depeiges et al. 2023), cosmological simulations and initial conditions reconstruction (C. Cuesta-Lazaro & S. Mishra-Sharma 2024; N. Mudur et al. 2023; A. Rouhiainen et al. 2024; R. Legin et al. 2024; V. Ono et al. 2024), and strong lensing problems (Y. Jagvaral et al. 2022; B. Remy et al. 2022).

In this work, we apply diffusion generative models to emulate cold dark matter density fields conditional on cosmological parameters, and demonstrate that the trained model can also be used to derive tight and robust constraints on cosmological parameters. In Section 3, we examine the ability of the model to appropriately capture the statistics of the distribution of fields corresponding to different parameters, and further compare the effect of modulating a single parameter at a time on the statistics of the resulting fields in the true and the generated set. We then quantify the ability of the model to capture the full range of cosmic variance for a single parameter, as a means to assess the extent of mode collapse. In Section 4, we examine how the diffusion model's approximate likelihood can be used to solve the inverse problem of constraining the cosmological parameters of a given input field. We then use the Hamiltonian Monte Carlo (HMC; S. Duane et al. 1987; R. M. Neal 2011; M. Betancourt 2017) method to draw samples from the estimated posterior on the cosmological parameters given an input field, and compare our estimates with a power spectrum baseline. A novel contribution of this work is our use of an HMC to sample a posterior consisting of an approximation to the diffusion model conditional likelihood to solve a downstream inference task. Finally, we demonstrate that the Diffusion-HMC-based parameter inference estimates are more robust to perturbations composed of uncorrelated noise relative to the estimates from a discriminative neural network directly trained to estimate parameters.

2. Data Sets, Architecture, and Training

Data sets. We work with cold dark matter density fields at $z = 0$ from the IllustrisTNG (A. Pillepich et al. 2018; D. Nelson et al. 2019) suite from the CAMELS Multifield Dataset (F. Villaescusa-Navarro et al. 2021a, 2022). The diffusion model is trained to generate the minmax transform applied to the log (base 10) of these dark matter fields. The minmax transform is pegged to the minimum and maximum of the log of the entire data set, [9.42, 15.44]. The data set contains 1000 simulations for 1000 different cosmologies with 15 two-dimensional fields per simulation. The "cosmology" is parameterized by a parameter vector with two cosmological (Ω_m and σ_8) and four astrophysical parameters. Each simulation tracked the evolution of 256^3 dark matter particles and 256^3 fluid elements and took around 6000 CPU hours to generate (see F. Villaescusa-Navarro et al. 2021a, 2022 for more details). The fields span 25 Mpc h^{-1} on each side. We train on 70% of the parameters in the Latin hypercube (LH) set, i.e., 700 parameters or 10,500 fields, and condition the model only on the cosmological parameters Ω_m and σ_8. Since we use dark matter density fields in this study, we do not expect them to constrain or contain much information about the astrophysical parameters. Example generated dark matter fields are shown in Figure 2.

Diffusion model setup. We follow the DDPM formalism, in which a target image x_0 is transformed to a sample from $x_T \sim \mathcal{N}(0, \mathbb{I})$ over the course of $T = 1000$ time steps. The forward diffusion process follows an incremental noise schedule $\{\beta_t\}$, and the noise is added in a variance-preserving way:

$$\begin{aligned} &\text{For t} \in [0, T-1],\ q(x_{t+1}|x_t) = \mathcal{N}(\sqrt{1-\beta_t}\,x_t,\ \beta_t\mathbb{I}) \\ &\text{and}\ \ q(x_{t+1}|x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\,x_0,\ (1-\bar{\alpha}_t)\mathbb{I}) \\ &\text{where}\ \bar{\alpha}_t = \prod_{t'=0}^{t} 1-\beta_{t'}. \end{aligned} \tag{1}$$

à

The score/noise-predictor model is a U-Net (O. Ronneberger et al. 2015) similar to that used in J. Ho et al. (2020). There are four downsampling layers, with each layer consisting of two ResNet blocks (S. Zagoruyko & N. Komodakis 2016), group normalization (Y. Wu & K. He 2019), and attention (A. Vaswani et al. 2017; Z. Shen et al. 2021). We use circular convolutions in the downsampling layers since the input fields have periodic boundary conditions. Each parameter is normalized to lie between [0, 1] with respect to its range, $\Omega_m \in [0.1, 0.5]$, $\sigma_8 \in [0.6, 1.0]$. A multilayer perceptron (MLP) transforms the cosmology vector into a space with the same dimension as the time embedding, and each ResNet block additionally has an MLP conditional on cosmology. The variance schedule β_t is nonlinear with smaller steps at smaller t and larger steps for larger values of t (see Figure 8). We train the model to generate the log of the fields, and randomly rotate and flip the image to account for these invariances. During training, for each batch of images $\boldsymbol{x}_0$, a batch of time steps is sampled uniformly along with a noise pattern $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$. The loss function minimized is $\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\phi(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon},\ t,\ \theta))\|^2$ for each set of $\{\boldsymbol{x}_0, t, \theta, \epsilon\}$, where θ is the parameter vector. Our implementation minimizes the Huber loss, which behaves as an L1 loss for values of the loss greater than 1 and a mean squared error (MSE) loss otherwise. From the training curves, the loss during training is less than 1 throughout, so the loss being minimized is in effect the MSE loss. We used the Weights and Biases framework to track experiments.[9] The model has 31.2 million parameters.

Training. We first downsample the images by a factor of 4 and train the conditional diffusion model on these 64×64 images for 60,000 iterations. Since the U-Net is formulated in terms of relative downsampling, the same architecture can be applied to images with different resolution. To train the model

[8] openai.com/dall-e-2

[9] http://wandb.ai/site/

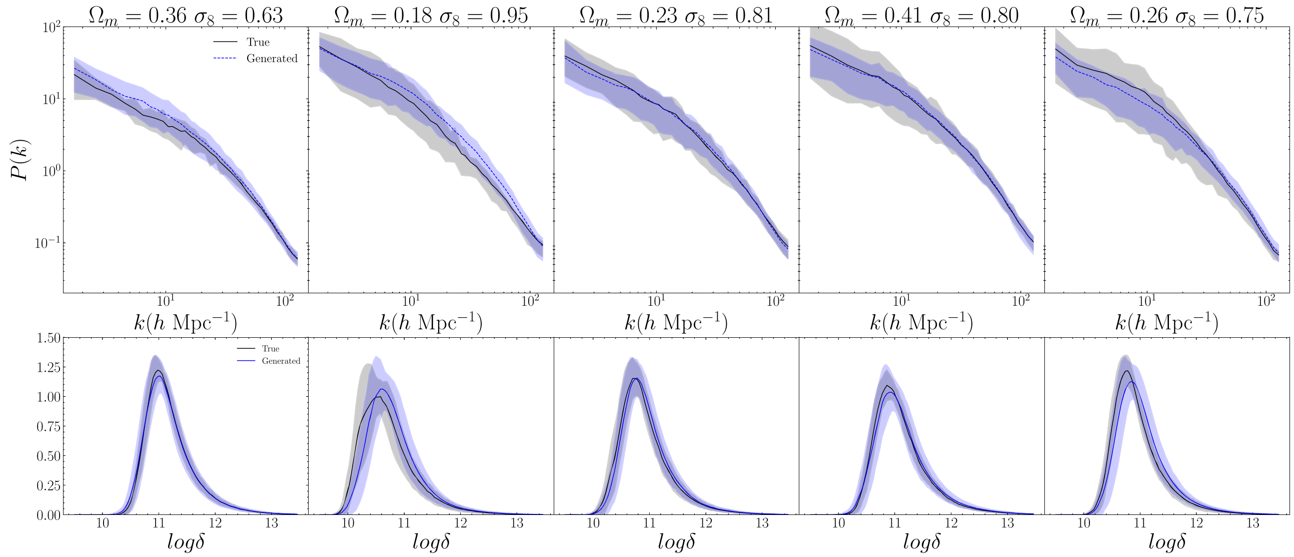

Figure 1. Generated fields at different cosmologies. Upper row: power spectrum of the unlogged fields for five different validation parameters. The lines depict the mean power spectrum and the envelope indicates the 16th and 84th percentiles of the distribution. True simulations are shown in black, generated in blue. Lower row: mean and standard deviation envelopes for the density histograms of the log fields.

to emulate 256 × 256 fields, we initialize the checkpoint with the weights of the 64 × 64-trained model after 60,000 iterations and trained for over 400,000 iterations. We found that initializing the 256 × 256 model with the weights of the 64 × 64 model and then training the model led to faster convergence.

The noise prediction loss does not fully capture sample quality and convergence, and we need an alternative metric to assess the quality of the generated samples (L. Theis et al. 2015). We sampled 500 fields (with 50 fields for 10 different validation parameters) for the checkpoints after 200,000, 220,000, 240,000, 260,000, 280,000, 300,000, 320,000, and 340,000 iterations. Sampling a batch of 50 256 × 256 fields from our model takes 310 s (6.2 s per field). We computed the reduced chi-squared statistic (Equation (B3)) of the power spectrum of each generated field, s, with respect to the reference distribution comprised of the 15 true fields for that parameter. We then compute the mean and standard error of these values across all parameters and sampled fields.

While the diffusion generative model is trained to generate the fields in log space, the power spectra we compute here and in Figures 1–3 correspond to the overdensity power spectra of the "linear" ($10^{\mathrm{GeneratedFields}}$) fields. The checkpoint corresponding to the 260,000th iteration had the lowest value for the chi-squared statistic of the power spectra of the linear generated fields relative to the distribution of true fields, corresponding to 2.29 ± 0.49 (Figure 12). To put this number in perspective, we can examine the effect of cosmic variance on this metric using a leave-one-out cross-validation approach, by computing the reduced chi-squared statistic of each sample of a true field, using the 14 other true fields corresponding to the same parameter as the reference distribution. The mean of this value across the 10 parameters is 1.70 ± 0.36. We use the 260,000 checkpoint for our analysis. We plot these values, along with the reduced chi-squared statistic of the power spectra of the log fields and the p-values of the mean intensity in Appendix B.1.

3. Summary Statistics

In this section, we examine the consistency of our generated fields relative to the true fields using three sets of simulations under the IllustrisTNG CAMELS suite. The LH suite varies the largest range of cosmological and astrophysical parameters but has a limited number (15) of fields for each parameter, with all parameters varying randomly. The one-parameter (1P) suite consists of 15 fields with the same seed, and one-dimensional variations of each parameter modulated systematically, while the others are fixed to the fiducial value. The cosmic variance (CV) set consists of 405 fields for the fiducial parameter value.

Varying cosmology in a LH. We examine the consistency of the summary statistics of the distribution of true and generated fields for a given validation parameter from the LH set in Figure 1. We have 15 true fields and 50 generated fields for each parameter. We derived the boundaries of the envelope using the estimates of the 16th and 84th percentiles, while the solid line demarcates the mean of the distribution of power spectra in 35 log-spaced k bins. The lower panel depicts the density histograms of the log fields. The envelopes are again derived using the percentiles, while the solid lines indicate the means of the histograms for the true and generated fields. The distribution of the power spectra and the density histograms of the true and the generated fields are in good agreement with each other.

Varying cosmological parameters one at a time (1P). In Figure 2, we generate "1P" sets and examine whether the effect of modulating a single parameter, while keeping the others constant, is the same as is observed in the 1P CAMELS suite. We sample 15 fields corresponding to 15 different seeds for each of the parameters. The fields corresponding to the same seed across parameters have the same position and orientation of their seeded structures as visible in Figure 2. For each seed, we then compute the ratio of the power spectrum of a field at a different parameter to the power spectrum of the fiducial parameter value, and compute the average and the percentile-based standard deviation envelopes across all seeds as depicted in the right column of Figure 2. The dashed (solid) line and

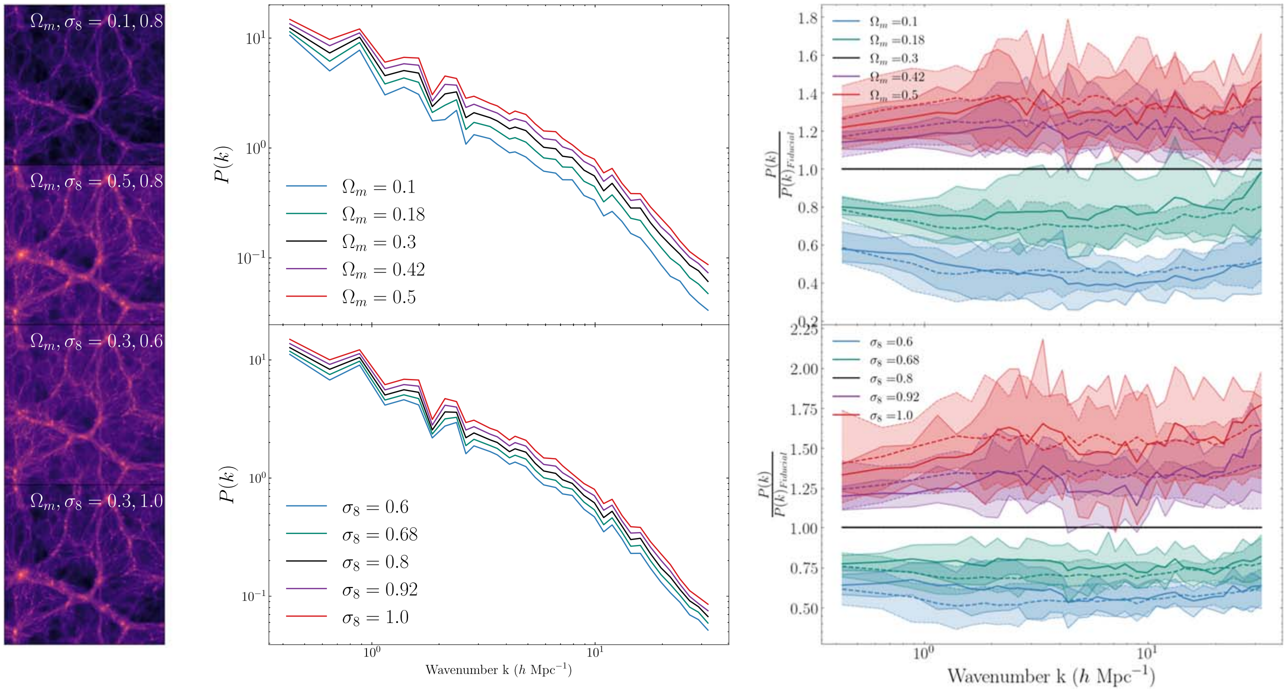

Figure 2. Generated “1P” fields. Left column: generated fields corresponding to the extreme values of each parameter for a single seed, with the other value held fixed at the fiducial value (0.3 for Ω_m and 0.8 for σ_8). Middle column: power spectra of the generated fields for the same seed, for different values of each parameter, holding the other fixed. Right column: mean and standard deviation for the ratio of the power spectra at the modified parameter value to the power spectra for the field at the fiducial parameter value (black) for 15 slices from the CAMELS data set (solid) and 15 seeds for the generated fields from the diffusion model (dashed). The effect of modulating a parameter on the generated fields’ power spectra is consistent with that of the true fields.

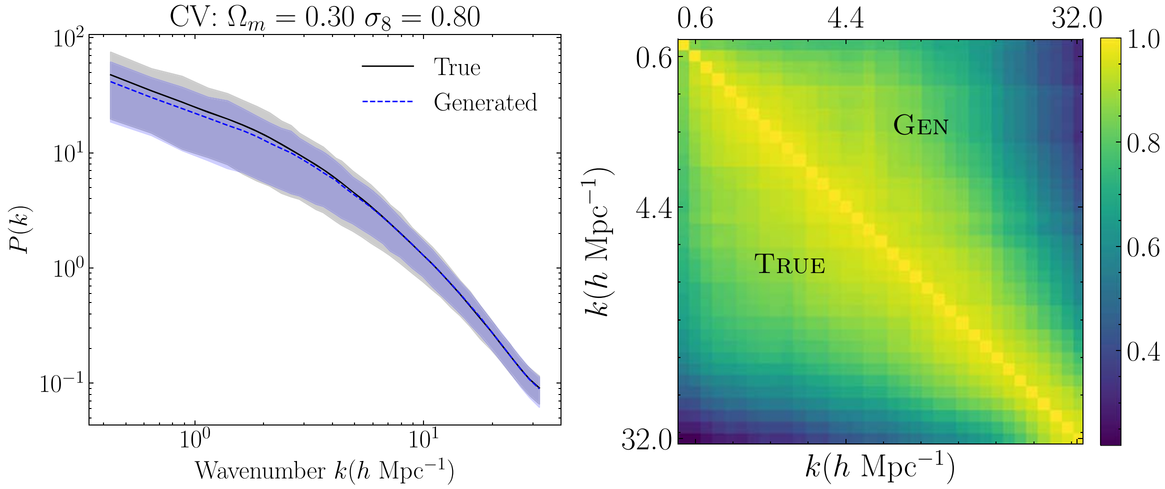

Figure 3. Generated “CV” fields. Left: power spectra of the 405 true and generated fields, with the mean and 16th to 84th percentiles. Right: correlation matrix of the power spectra of the true and generated fields. The lower triangular matrix corresponds to the correlation matrix of the true fields while the upper triangular matrix corresponds to the correlation matrix of the generated fields.

envelope correspond to the generated (true) ratios. The ratios for the generated “1P” set are in good agreement with those of the true “1P” set, since the dashed lines are within the solid envelopes.

Reproducing cosmic variance (CV). The CV set has 405 (27 × 15) fields for the fiducial parameter value of [0.3, 0.8] and varying initial conditions, designed to quantify the effect of cosmic variance. The CV set allows us to quantify the consistency between the second moments of the true and the generated distributions. In particular, it allows us to test the ability of the model to generate a diverse set of samples for the same cosmological parameters such that it reproduces the true underlying distribution at fixed cosmology.

We generate 405 samples from our trained diffusion model, compute their power spectra, and examine the standard deviation and the correlation matrices of different k modes of the power spectrum in Figures 3 and 4. The correlation between the modes of the power spectra is largely consistent between the true and the generated samples, although the generated samples appear to have a slight excess correlation around $5\,h\,\mathrm{Mpc}^{-1}$.

The standard deviations of the distribution are also consistent, although the standard deviations of the generated

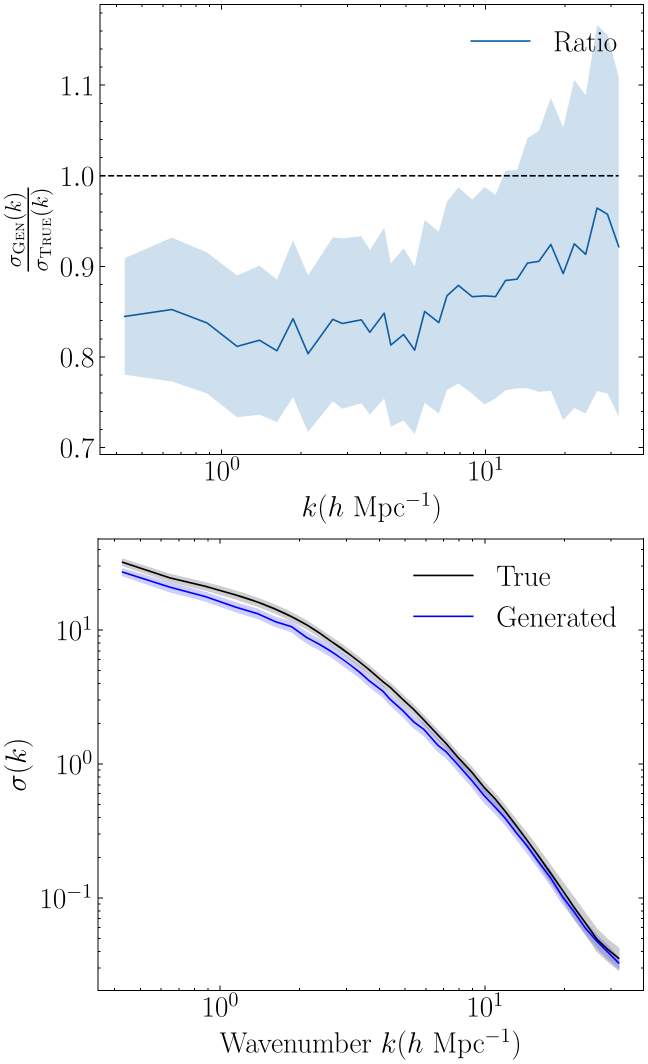

Figure 4. Generated "CV" fields. Upper: ratio of the standard deviations of the generated fields to that of the true fields. Lower: standard deviations of the power spectra in each k bin for the generated and true fields.

power spectra are slightly underpredicted relative to the true power spectra at the largest length scales ($k < 5\, h\, \mathrm{Mpc}^{-1}$). To construct the standard deviation estimator, we use the jackknife approach to generate a distribution of 405 estimates of the standard deviation for each set, where the ith estimate corresponds to the standard deviation computed using all samples excluding that of the ith field. We then compute the mean and the standard deviation for these sets in order to capture the mean and the standard deviation on the estimate of the standard deviation in the upper-right panel of Figure 3. We use the ratios of the jackknife-estimated means and errors in the lower-right panel.

The ability to capture the full diversity of k modes for a single parameter may be in tension with the ability to distinguish between and appropriately modulate the power spectrum for different parameters. This tension is enhanced for our assessment of mode collapse in spectral space, compared to canonical machine learning data sets involving discrete classes, since the cosmological parameters that we condition the diffusion generative model on lie on a continuum. Thus, for a given conditioning parameter, a generated field with too much or too little power on certain scales relative to the mean of the statistic is more likely to wander into the typical set of a field with a different cosmological parameter, since modulating the parameters also modulates the power spectrum (as in Figure 2). In the context of natural images, one often deals with categorical descriptors, where the boundaries between whether an object qualifies as one class or another are typically more clearly delineated, and the ability to capture the full diversity of cats is unlikely to cause the model to "wander" into islands of image space corresponding to a dog or an airplane. Thus, the slightly lower standard deviation can be partly attributed to the possibility that the model chooses to compromise on the diversity of samples for a single parameter, in order to be able to accurately generate fields that look different for different parameters.

We compute the covariance between the modes of the power spectrum. With the full covariance, we can now statistically quantify the consistency of the generated samples relative to that of the true samples using the multidimensional reduced chi-squared statistic. We use 350 samples of the true fields to set up the reference distribution used to compute the covariance. We compute the inverse covariance and adjust for the Hartlap factor (Equation (B1); J. Hartlap et al. 2007). We then compute the multidimensional reduced chi-squared statistic of the entire sample of the true and generated fields, and compute the means of the 405 chi-squared statistics for each distribution following Equation (B2). For the true fields, the mean of the chi-squared distribution is 33.4, while that of the generated is 36.6. Since our power spectra consist of 35 log-spaced bins, this is consistent with the expected mean for a chi-squared distribution with 35 degrees of freedom, i.e., 35.

4. Parameter Inference

A trained diffusion model can be used to estimate the variational lower bound (VLB) of the log likelihood (C. Cuesta-Lazaro & S. Mishra-Sharma 2024; D. P. Kingma & R. Gao 2023). In the case of a conditional diffusion model, this likelihood estimate is also conditional, i.e.,

$$\begin{aligned}\mathbb{E}_q[-\log p_\phi(x_0|\theta)] &\leqslant \mathbb{E}_q[-\log p_\phi(x_0|x_1, \theta)\\ &+ \sum_{t\geqslant 1} D_{KL}[q(x_t|x_{t+1}, x_0)||p_\phi(x_t|x_{t+1}, \theta)]\\ &+ D_{KL}[q(x_T|x_0)||p(x_T)]] \simeq L_{\mathrm{VLB}}, \end{aligned}\tag{2}$$

$$L_{\mathrm{VLB}} = L_0 + L_1 \ldots L_{T-1} + L_T, \tag{3}$$

$$\hat{L}_{\mathrm{VLB}} = \sum_{t<T_{\mathrm{MAX}}} L_t, \tag{4}$$

where ϕ denotes the diffusion model architecture, θ is the conditioning cosmology (in our case, a vector with Ω_m and σ_8), p_ϕ are the reverse (learned) distributions, and q are the forward (analytical) distributions. During training, the diffusion model's noise prediction loss terms are equivalent to terms of the reweighted VLB (D. P. Kingma & R. Gao 2023). Since the predicted noise is conditional on cosmology, the terms of the VLB thus encode dependencies on the cosmological parameters. The contrast between the VLB evaluated at one parameter θ_1 relative to another parameter θ_2 for a fixed field x_0 can thus be used to find the region of parameter space that maximizes the conditional likelihood for the field.

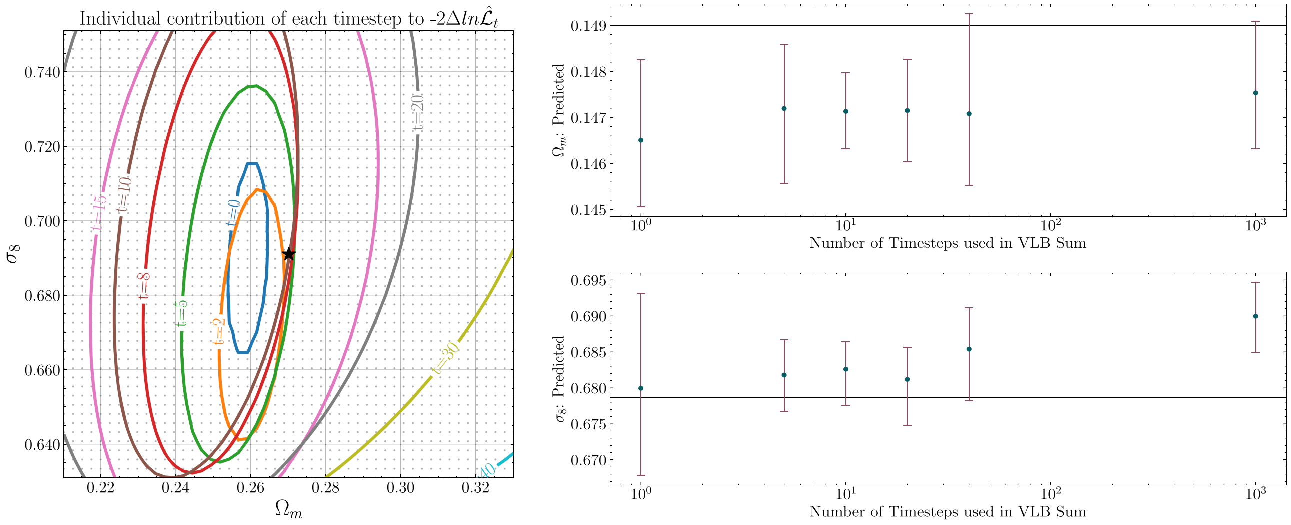

Figure 5. Investigating the contribution of the time steps used in the VLB sum. Left: 1σ contours of each $-2\Delta\ln L_t$'s individual contribution for different time steps. The contours are all centered near the true parameter (black star) and become wider as t increases. Right: mean and 1σ predictions for a single field as a function of the number of time steps used in the VLB sum optimized by the HMC. Reducing the number of time steps used to compute the VLB does not significantly affect the diffusion model predictions.

4.1. Examining the Influence of Different Time Steps

Since using all the terms L_t that contribute to the VLB is computationally expensive, we first investigate how sensitive each of the contributing terms L_t is to changes in cosmological parameters over a grid in Figure 5(a). For an input field x_0 and a single seed, we evaluate each of the $L_t(x_0|\boldsymbol{\theta}_{\mathrm{EVAL}})$ terms over a 50×50 grid in $[\Omega_m, \sigma_8]$, centered on the value of the true field $\boldsymbol{\theta}_{\mathrm{TRUE}}$ and extending to ± 0.06 about the true parameter. To disentangle the individual contribution of each term, we subtract the minimum value of each $L_t(x_0|\boldsymbol{\theta}_{\mathrm{EVAL}})$ on the grid, and multiply it by 2 to yield $-2\Delta\ln\hat{\mathcal{L}}_t$. We plot the contour corresponding to 2.30, or the 1σ contour for a chi-squared distribution with 2 degrees of freedom. Two observations are apparent: all contours are minimized in the vicinity of the true parameter, and the curvature of the contours decreases with increasing time. Increasing time involves increasing the amount of noise added to the input image, which can explain the increased uncertainty in the true value of the parameter. Thus, dropping the terms corresponding to the higher time steps in Equation (2) is unlikely to result in weaker constraints on cosmology.

4.2. HMC-based Parameter Inference

To compute the parameter estimates for fields, we draw samples from the posterior on the parameter using the HMC method. The HMC, or hybrid Monte Carlo method (S. Duane et al. 1987; R. M. Neal 2011; M. Betancourt 2017), is a Markov Chain Monte Carlo approach that draws samples from a probability distribution $\pi(\theta)$ via the introduction of an auxiliary momentum variable $\boldsymbol{p}$ and solves the equations of Hamiltonian dynamics in order to update the momentum and position (θ). HMC enables a more efficient exploration of high-dimensional probability distributions. In our case, using an HMC also helps us circumvent the problem of having to redefine the extents and granularity of the parameter grid depending on how confident the constraints for a given T_{MAX} are. The Hamiltonian governing the dynamics in the HMC chain is

$$H = U + K \text{ where } U = -\log\pi(\theta) \text{ and } K = \frac{\boldsymbol{p}^{-1}\boldsymbol{M}^{-1}\boldsymbol{p}}{2}, \tag{5}$$

$$\begin{aligned}\log\pi(\theta) &= \log p_\phi(x_0|\theta) + \log p_{\mathrm{PRIOR}}(\theta)\\ &\simeq -\hat{L}_{\mathrm{VLB}} + \log p_{\mathrm{PRIOR}}\\ &= -\sum_{t<T_{\mathrm{MAX}}} L_t + \log p_{\mathrm{PRIOR}}.\end{aligned} \tag{6}$$

We used the `Hamiltorch` (A. D. Cobb et al. 2019) package for the HMC. We explain more details about the HMC implementation in Appendix A. The prior is chosen to be a flat prior over $\Omega_m \in [0.1, 0.5]$ and $\sigma_8 \in [0.6, 1.0]$. The initial parameter value is always the fiducial value of [0.3, 0.8], and we designate the first 100 samples as burn-in samples that are discarded. We reduce the averages over random noise in each L_t term in Equation (2) to a single stochastic estimate of L_t in Equation (4).

We now examine the effect of truncating terms in Equation (4) using the HMC-based parameter inference. Truncating terms allows us to perform inference faster and can allow us to explore the trade-off between dropping terms for speed and higher precision with more time steps. In the right column of Figure 5, for a single field we plot the mean predictions and the 15.9th to 84.1st (1σ) percentiles computed using 200 samples with the approximate $-\log p_\phi(x_0|\theta)$ using Equation (4) as a function of T_{MAX} on the x-axis for the same field. The true value of the cosmological parameter for this field is demarcated by the solid black lines. For this field and parameter, using more terms asymptotically removes the bias on Ω_m while increasing the bias on σ_8. However, using the first 20 time steps only changes the mean prediction for Ω_m and σ_8 by -0.26% and -1.27% of the prediction using all 1000 time steps, respectively.

We now turn our attention to the performance of our parameter inference approach relative to a power spectrum baseline. For subsequent HMC-based parameter inference in this

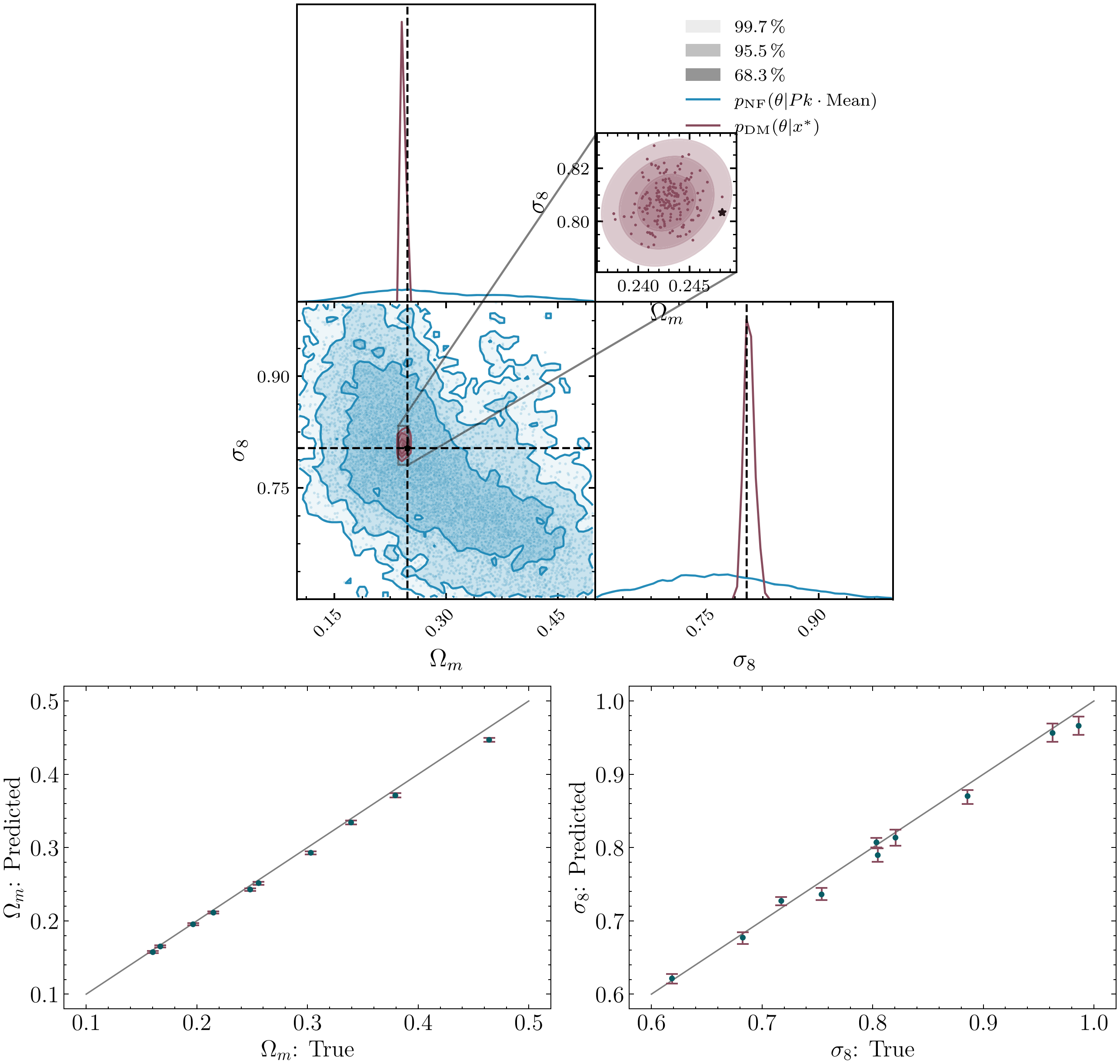

Figure 6. Top: comparison of the parameter estimates for a single field with the power-spectrum-based estimator and with the diffusion model likelihood. The star demarcates the true parameter corresponding to the input fields. The shaded contours demarcate the 68.27%, 95.45% and 99.73% confidence intervals. Lower panel: predicted parameter and ground-truth parameter for 10 different fields. The error bars correspond to the 15.9th to 84.1st percentiles for the marginal probability distributions for each parameter for each input field.

section, we use Equation (4) with $T_{\rm MAX} = 20$ to approximate the conditional negative log likelihood. Drawing 500 samples with $T_{\rm MAX} = 20$ takes ∼32 minutes for a single field.

To assess the constraining power of the power spectrum, we use the neural posterior estimation approach and train a normalizing flow to represent the posterior of $p_{\rm NF}(\theta|Pk \cdot {\rm Mean})$ using the `Lampe` (F. Rozet et al. 2021) package. Since computing the overdensity power spectrum involves dividing by the mean of the field, we further concatenate the mean of the log field as an additional feature, since the prediction for Ω_m for a small box size of 25 Mpc h^{-1} is very sensitive to the mean of the fields. For a single field, we compare the posteriors obtained by drawing 10,000 samples from the power-spectrum-based estimator and 400 samples from the diffusion model +HMC in Figure 6. Other details about our implementation are given in the Appendix A.2.

The diffusion model has significantly narrower constraints for the parameters relative to the power spectrum baseline. Note that, as shown in Figure 2, the cosmological parameters Ω_m and σ_8 are strongly correlated at the level of the power spectrum on the small scales probed by our simulations, since they are both modulating the amplitude of the power spectrum. This result is consistent with P. Villanueva-Domingo & F. Villaescusa-Navarro (2022), who also found that the galaxy power spectrum in conjunction with a MLP yielded weak constraints on Ω_m and σ_8.

In the second row, we plot the predicted cosmological parameters relative to the truth for 10 different fields across different parameters in the validation set. The dots annotate the mean of the 400 samples and the error bars correspond to the 15.9th to 84.1st percentiles of the samples. The mean of the samples is close to the true value of $\boldsymbol{\theta}_{\rm TRUE}$ over a broad range

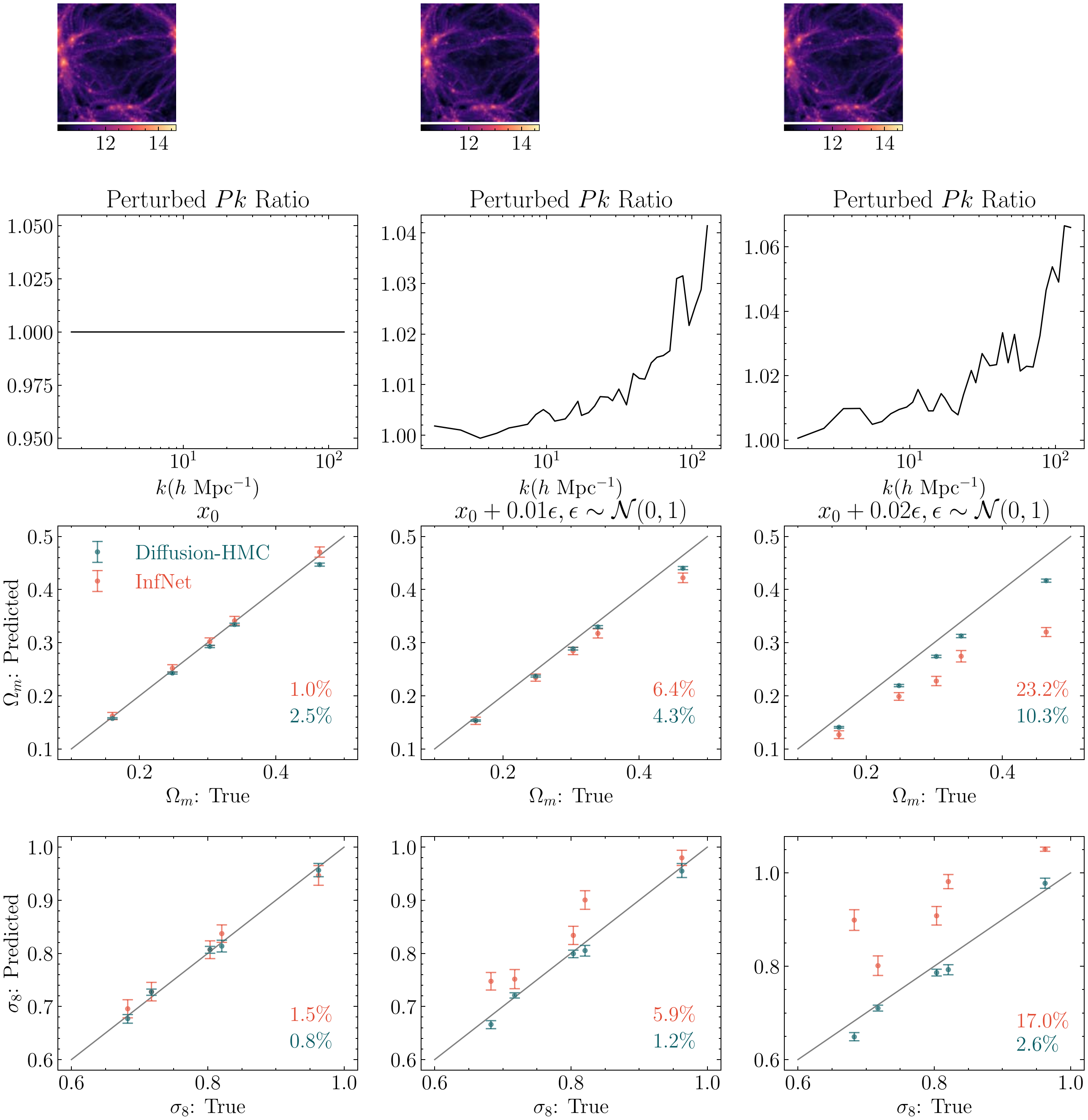

Figure 7. First row: an example field corresponding to $\Omega_m, \sigma_8 = 0.33, 0.96$ with its perturbed counterparts, along with the ratio of the perturbed linear field's power spectrum to that of the original field $\frac{Pk_{\mathrm{Noisy}}}{Pk_{\mathrm{Original}}}$. Second and third rows: mean and 15.8th to 84.1st percentile predictions using the Diffusion-HMC estimates (green) and the mean and sigma predictions obtained from the parameter inference network in F. Villaescusa-Navarro et al. (2021a), denoted by InfNet. The mean absolute percentage bias (of the true prediction) for each parameter is also indicated. The Diffusion-HMC constraints are more robust to perturbations to the input image.

of parameters. The Ω_m predictions are biased by -0.006 (2% of the true prediction) on average over the range of parameters, while the mean absolute bias for σ_8 is 0.01 (1.26% of the true prediction). The σ_8 prediction uncertainties (mean z score $|\bar{z}| = 1.14$) are better calibrated relative to the uncertainties for Ω_m ($|\bar{z}| = 2.75$), but overall we find the error bars to be slightly underpredicted. We attribute the miscalibration to the truncation of terms. It is possible that including more time steps, averaging over more seeds, and examining alternate choices of the variance schedule could yield better calibrated uncertainties; we leave this exploration for future work.

4.3. Robustness

Although our constraints are much tighter than those derived from the two-point correlation function, robustness to noise and observational systematics is an important consideration guiding the use of parameter inference methods on survey data. Since the terms of the VLB (and the noise prediction loss terms) include terms where the original image has been noised, we hypothesize that the parameter estimates learned by the model are naturally more robust to perturbations involving the addition of scaled white noise to the field.

To test this, we examine the extent to which our parameter estimates change relative to the predictions of the parameter inference network in F. Villaescusa-Navarro et al. (2021b). The neural network in F. Villaescusa-Navarro et al. (2021b) is trained to predict the mean and the standard deviation of the parameters given an input dark matter field. In the leftmost column of Figure 7, we compare sample predictions obtained from the diffusion model (Diffusion-HMC) relative to those obtained from the parameter inference network (InfNet) on five fields (x_0). In the next two columns, we add increasing levels of $\mathcal{N}(0, \sigma)$ noise with $\sigma \in [0.01, 0.02]$, and examine the estimates for both approaches. We perform this experiment in a noise-agnostic setting, i.e., we assume that we do not know the level of noise a priori and do not modify either the Diffusion-HMC-based approach or the parameter-inference-network-based approach. We use the same $T_{\mathrm{MAX}} = 20$ and use 200 samples for the Diffusion-HMC estimates for all fields. Although the noise perturbations to the field are visually imperceptible, and correspond to a change of less than ∼4%–6% to the power spectrum at the smallest scales, the predictions of the parameter inference network are significantly (∼6%–23%) disturbed. Indeed, we find that the Diffusion-HMC-based constraints are perturbed far less than the InfNet-based constraints. The Diffusion-HMC σ_8 constraints are also more robust than its Ω_m constraints.

5. Conclusion

In this work, we used a diffusion generative model to emulate dark matter density fields conditional on cosmological parameters. The model learns to reproduce the modulation of summary statistics, such as the power spectra, for changes in cosmological parameters. We additionally assess the extent of mode collapse through the lens of cosmic variance by examining the diversity of the power spectra at the fiducial cosmology.

We then directed our attention to the inverse problem of parameter inference and disentangled the contribution of each term L_t in the expression for the VLB to constraints on cosmological parameters. Our findings reveal that the strength of the constraints decreases as t increases. K. Clark & P. Jaini (2024) and A. C. Li et al. (2023) explored this question in the context of using the diffusion model's conditional VLB terms for classification. K. Clark & P. Jaini (2024) also found a negative exponential weighting of the terms to be optimum, while A. C. Li et al. (2023) found that intermediate time steps had the highest accuracy when only a single time step is used. The difference in which terms contain the most information about the conditioning attribute (cosmology/class) is interesting, and could be partly attributed to the difference in their formulation and weighting of the VLB terms and in how changes in the conditioning vector affect an image in different settings. Modulating cosmology modifies global attributes of the fields, such as their intensity and power spectra, as seen in Figure 2, while the information that distinguishes different breeds of pets from each other tends to be relatively localized in the bounding box containing the animal.

If it is indeed always the case that the time steps nearest to the image manifold contain most cosmological information, one could in future swap out our discrete time architecture with continuous time diffusion models (Y. Song et al. 2021), where t is a continuous variable with $t \in [0, 1]$, and prioritize steps that lie near the image manifold or $t = 0$.

These insights motivated us to truncate the diffusion model conditional VLB-based approximation for $p_\phi(x_0|\theta)$ by subsampling the terms to only use the first T_{MAX} terms (Equation (4)). This approximation allows us to backpropagate its gradient and plug this estimate for $p_\phi(x_0|\theta)$ into an HMC and sample the posterior on $p_\phi(\theta|x_0)$. The Diffusion-HMC approach yields tight constraints on cosmological parameters, competitive with a bespoke parameter inference network F. Villaescusa-Navarro et al. (2021b) trained only to infer cosmology given a field.

In our experiments, we use only a single seed in each HMC step and a $T_{\mathrm{MAX}} = 20$ to speed up inference. In our case, the use of an HMC allowed us to circumvent the requirement of choosing a grid with the appropriate resolution required to resolve the constraints, and eliminates the dependence on the number of points in the grid. However, the Diffusion-VLB estimates could also be used in other settings that do not entail the use of an HMC. While our approach scaled with the number of time steps we use in the sum, in constrast A. C. Li et al. (2023) scaled with the number of classes in the classification data set.

Lastly, we demonstrate that the diffusion model likelihood confers the Diffusion-HMC constraints with greater robustness against the addition of small amounts of noise to the input image relative to the behavior of a parameter inference network. This echoes the behavior of diffusion models in other discriminative tasks such as in A. C. Li et al. (2023), M. Prabhudesai et al. (2023), and H. Chen et al. (2024), where diffusion-model-based classifiers have been shown to possess higher robustness to adversarial examples or perturbations. This is a pertinent finding since B. Horowitz & P. Melchior (2022) showed that the powerful constraints derived from neural-network-based parameter inference may often come at the cost of their susceptibility to slight perturbations that the canonical two-point correlation based analyses are impervious to. We find that our diffusion-model-based parameter inference approach enables more noise-robust field-level inference. It would be interesting to further explore the differences in inductive biases learned by discriminative networks (e.g., F. Villaescusa-Navarro et al. 2021b; D. Sharma et al. 2024a) relative to generative models repurposed for discriminative tasks.

The simulation volume of the data set we work with is still much smaller than the scales mapped by astrophysical surveys. Future work could focus on scaling up to more survey-realistic scenarios involving larger simulation volumes and directly observed tracers. C. Cuesta-Lazaro & S. Mishra-Sharma (2024) also showed that diffusion generative models that work with point clouds can allow one to emulate and perform cosmological parameter inference with point cloud data. Our exploration into robustness could inform applications to real data, and future investigation could focus on conferring and quantifying robustness against other survey-related and observational noise effects. Alternative formulations of the generative process (A. Bansal et al. 2024; J. Wildberger et al. 2024) could also be relevant to this exercise.

In this work, we demonstrated that a diffusion model can be trained not just to emulate fields, but that its likelihood estimate can be adapted to work with the HMC framework to derive tight constraints on cosmological parameters. This makes a step toward advancing the use of diffusion-model-based priors for a range of downstream tasks from image generation and restoration to inference problems.

6. Code

The checkpoint used for the results in this paper is available on Zenodo: doi:10.5281/zenodo.13993010. The code for this project is available at Diffusion-HMC,[10] with a current copy uploaded to the above Zenodo DOI. We acknowledge F. Villaescusa-Navarro et al. (2021b), Improved-Diffusion (A. Q. Nichol & P. Dhariwal 2021),[11] The Annotated Diffusion,[12] DDPM,[13] VDM (D. Kingma et al. 2021),[14] and Y. Song et al. (2021) for code snippets.

Packages: Hamiltorch (A. D. Cobb et al. 2019), Lampe (F. Rozet et al. 2021), GNU Parallel (O. Tange 2018).

Acknowledgments

This work was supported by the National Science Foundation under Cooperative Agreement PHY2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). We are grateful to Yueying Ni, Francisco Villaescusa-Navarro, Core Francisco Park, Andrew K. Saydjari, Justina R. Yang, Yilun Du, Ana Sofia Uzsoy, Shuchin Aeron, and many others for insightful conversations.

Appendix A
Parameter Inference

Figure 8 plots the standard deviations of the noised distributions as well as the signal-to-noise ratio as a function of time step. In Equation (5), we set $\boldsymbol{M}^{-1}$ to be diagonal with [1, 5]. Setting the inverse mass matrix in an HMC to be close to the covariance of the expected posterior distribution helps the chain explore the space better. In this case, we choose a step size of 5×10^{-4}, because of the steep gradient of the posterior distribution with respect to Ω_m. For parameter inference on a field whose true Ω_m value is 0.101, i.e., on the prior boundary, we need to further reduce the step size to 1×10^{-4} in order for the parameters to be accepted. We modified the Hamiltorch package in order to generate samples. For $T_{\mathrm{MAX}} \geqslant 20$, we compute the contributions to the VLB loss in batches of 10 time steps and accumulate the gradient contribution for each batch. This enables us to compute the gradients with 1000 time steps. The choice of mass matrix accelerates the chain's convergence to the correct region of parameter space for σ_8. We approximate $\log p_\phi(x_0|\theta)$ with the truncated VLB in Equation (4). While we do not compute the expectation over multiple seeds within a single evaluation of Equation (5), for speed every evaluation uses a different seed and noise pattern. The prior is chosen to be a flat prior over $\Omega_m \in [0.1, 0.5]$ and $\sigma_8 \in [0.6, 1.0]$:

$$\begin{aligned} L_{\mathrm{vlb}} &= L_0 + L_1 \ldots L_{T-1} + L_T \\ &= \mathbb{E}[-\log p_\phi(x_0|\theta)] \leqslant \mathbb{E}_q[\, D_{KL}[q(x_T|x_0)||p(x_T)] \\ &\quad + \sum_{t \geqslant 1} D_{KL}[q(x_t|x_{t+1}, x_0)||p_\phi(x_t|x_{t+1}, \theta)] \\ &\quad - \log p_\phi(x_0|x_1, \theta)]. \end{aligned} \quad \text{(A1)}$$

NOTATION:

1. x_0: Normalized input field.
2. ϕ: Noise model (neural network).
3. θ: Conditioning cosmology, i.e., a vector with Ω_m and $\sigma_{8.}$

COMPUTING VLB TERMS:

$$\begin{aligned} L_0 &= -\log p_\phi(x_0|x_1, \theta) = -\ln \mathcal{N}(x_0|\mu_0, \beta_0) \\ &= \sum_p \frac{(x_0 - \mu_0)^2}{2\beta_0} + 0.5 \ln|2\pi\beta_0| \\ &\quad \text{For } t \in [1, T-1], \\ L_t &= D_{KL}[q(x_t|x_{t+1}, x_0)||p_\phi(x_t|x_{t+1}, \theta)] \\ &\quad q(x_t|x_{t+1}, x_0) = \mathcal{N}(\tilde{\mu}_t(x_{t+1}, x_0), \tilde{\beta}_t) \\ &\quad \tilde{\mu}_t(x_{t+1}, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}\beta_t}}{1 - \bar{\alpha}_t} x_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} x_{t+1} \\ &\quad \text{and } \tilde{\beta}_t = \beta_t \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \\ &\quad p_\phi(x_t|x_{t+1}, \theta) = \mathcal{N}(\tilde{\mu}_t(x_{t+1}, \hat{x}_{0,\phi}), \tilde{\beta}_t) \\ &\quad \hat{x}_{0,\phi} = \frac{x_{t+1} - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\phi(x_{t+1}, t, \theta)}{\sqrt{\bar{\alpha}_t}}. \end{aligned}$$

A.1. HMC Convergence

For a single field, we examine the convergence of parameters, when the chain starts from different initial parameters, in Figure 9. The chains are indistinguishable beyond around 50 samples. We thus choose a burn-in of 100 samples. The $\hat{R}$ for both parameters computed using the samples in [100–300] is 0.997. A $\hat{R}$ of greater than 1.1 usually indicates that the chains have not converged and still retain some memory of their initialization. While the $\hat{R}$ is theoretically expected to be around 1 or slightly greater, some numerical variation about this expected value can result in values that are slightly less than 1. The $\hat{R}$ is a measure of the variance between chains divided by the variance within chains.

A.2. Parameter Estimation Baselines

Power spectrum NPE baseline. We use a masked autoregressive flow (G. Papamakarios et al. 2017) to implement the normalizing flow that predicts the two-dimensional parameter vector given the 129-dimensional feature vector for a single field (128 bins for the power spectrum +1 for the mean of the log fields). The power spectrum is the log of the overdensity power spectrum of the (unlogged) fields. In Figure 6, the power spectrum sample contours are smoothed by convolving with a Gaussian kernel with a scale of 0.8 and the Diffusion-HMC samples are smoothed by a kernel of 0.2. The ellipses in the inset figure are computed using the covariance of the 400 diffusion model samples and finding the ellipses corresponding to the 68.3%, 99.4%, and 99.7% confidence intervals, using the eigen decomposition of the covariance.

A.3. Additional Robustness Tests

Dropping the prior. In Figure 10, we explore robustness without the prior in the HMC setting, for the noise levels in Figure 7 as well as with the addition of more noise. For small amounts of noise ($\sigma = 0.01, 0.02$), removing the prior does not affect the bias of the Diffusion-HMC-inferred parameters since

[10] github.com/nmudur/diffusion-hmc
[11] github.com/openai/improved-diffusion
[12] huggingface.co/blog/annotated-diffusion
[13] github.com/hojonathanho/diffusion
[14] github.com/google-research/vdm

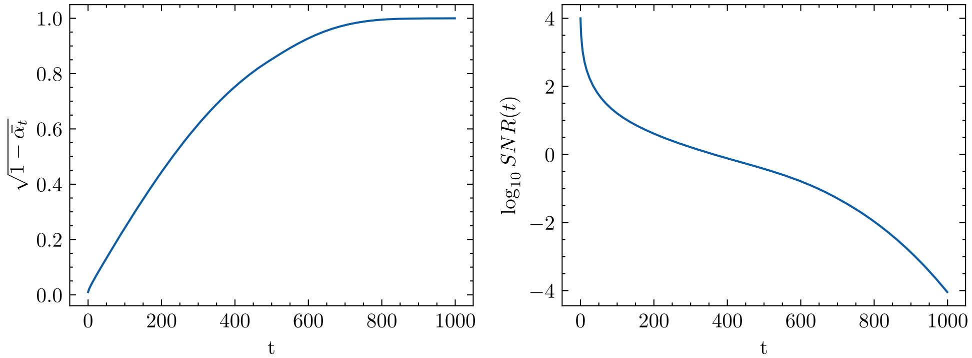

Figure 8. The scale/standard deviation of the cumulative noise in Equation (1) added to the image over different time steps.

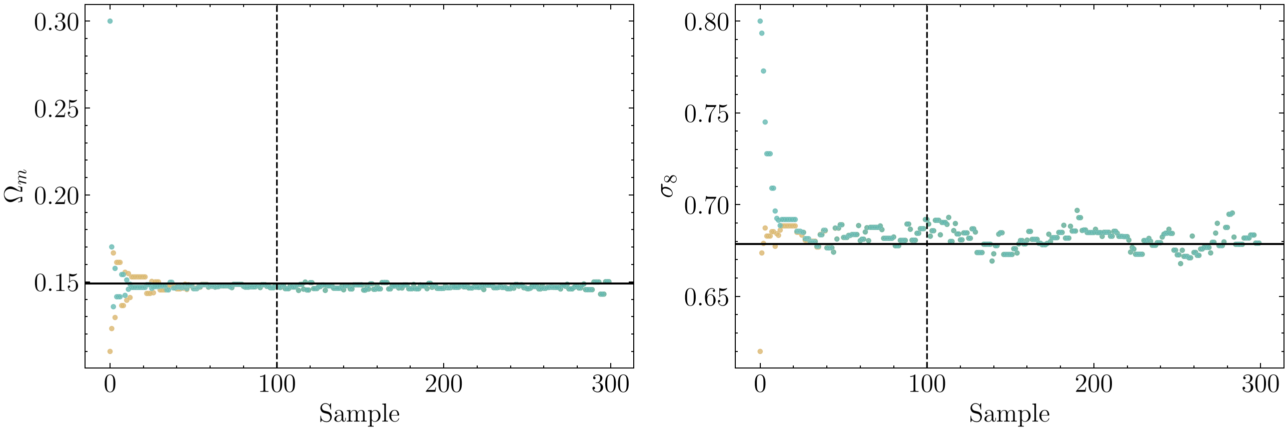

Figure 9. HMC chains for 300 samples, for the same field, starting at two different initial parameters: [0.11, 0.62] (beige) and [0.3, 0.8] (teal). The chains are well mixed by the cutoff we designate as our burn-in (100 samples), denoted by the dashed line.

the numbers in Figure 10 are identical/close to those in Figure 7. For noise with a scale of 0.05, the Diffusion-HMC constraints are more perturbed for Ω_m but more robust for σ_8 relative to the neural-network baseline. A standard deviation of 0.05 is equivalent to the diffusion noising time steps between the 1st and the 2nd time steps, while $\{0.01, 0.02\}$ are less than the first time step. Note, we plot the power spectra of the "true" (linear) mass density field in these figures, to be consistent with the power spectra in Figures 1 and 2. However, the noise is added to the log of the field. Thus, while the effect on the linear power spectrum is mild (around 5% at the smallest scales), for $\sigma = 0.05$ the effect on the power spectrum of the log field is around 35% at the smallest scales.

Dropping initial time steps. In the left panel of Figure 5, we observed that the isolated effect of the smallest time steps also had the strongest constraints. We thus ask the question: If we perform parameter inference in a setting where there is some knowledge of the amount of noise added, can dropping time steps confer additional robustness? In Figure 11, we add a comparison for the noising amounts corresponding to 0.02 and 0.05 and drop the first two time steps. The rest of the setting is the same as in Figure 10, i.e., we also drop the prior. We find that dropping these time steps confers greater robustness on the estimates, without noticeably reducing confidence. While we defer a more rigorous investigation to future work, the prospect of using the knowledge of the amount of noise added to strategically drop time steps could be of interest, reminiscent of scale-dependent analyses and scale cuts in other cosmological analysis methods (E. Krause et al. 2017; B. Dai & U. Seljak 2024).

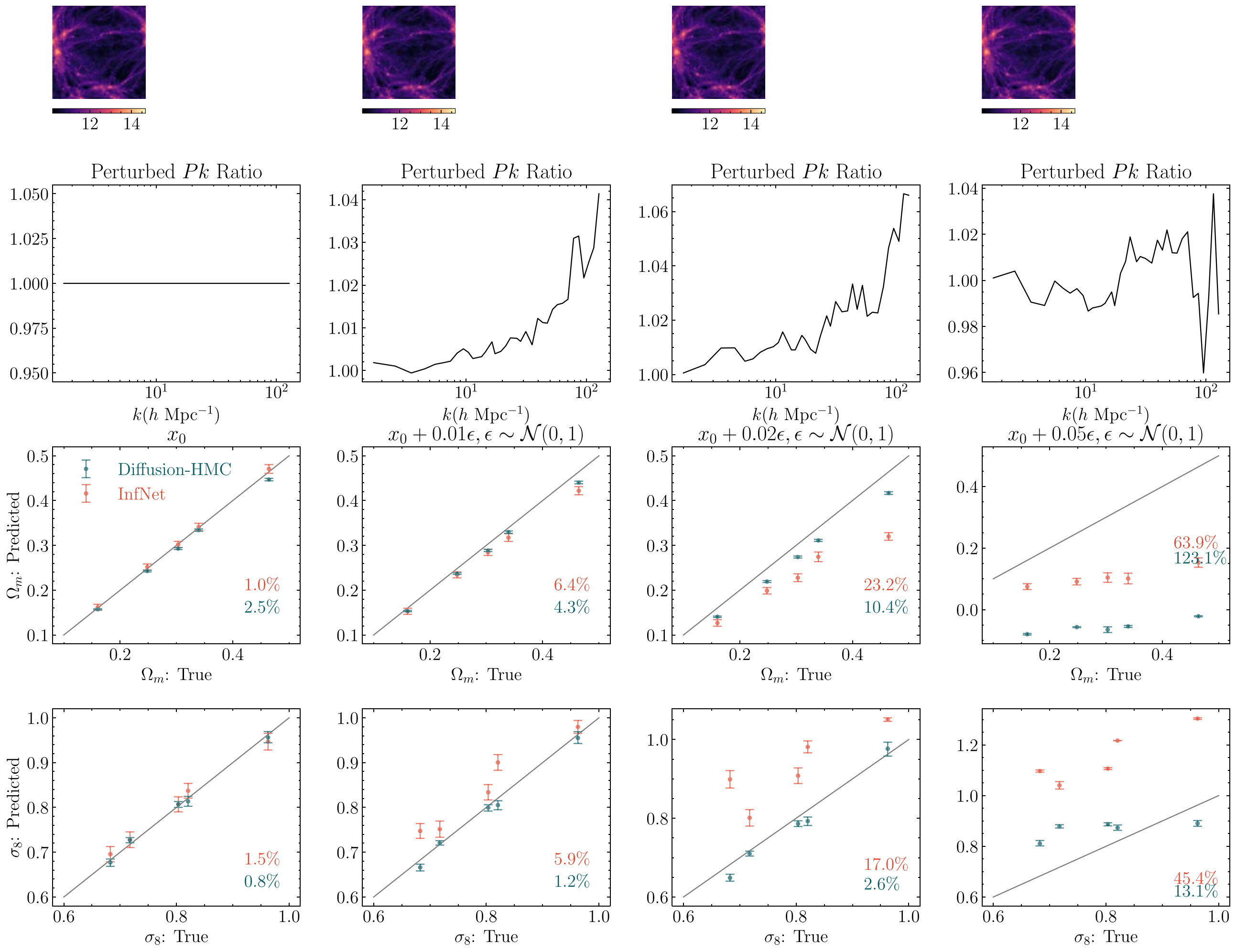

Figure 10. Robustness comparison without a prior applied in the HMC case.

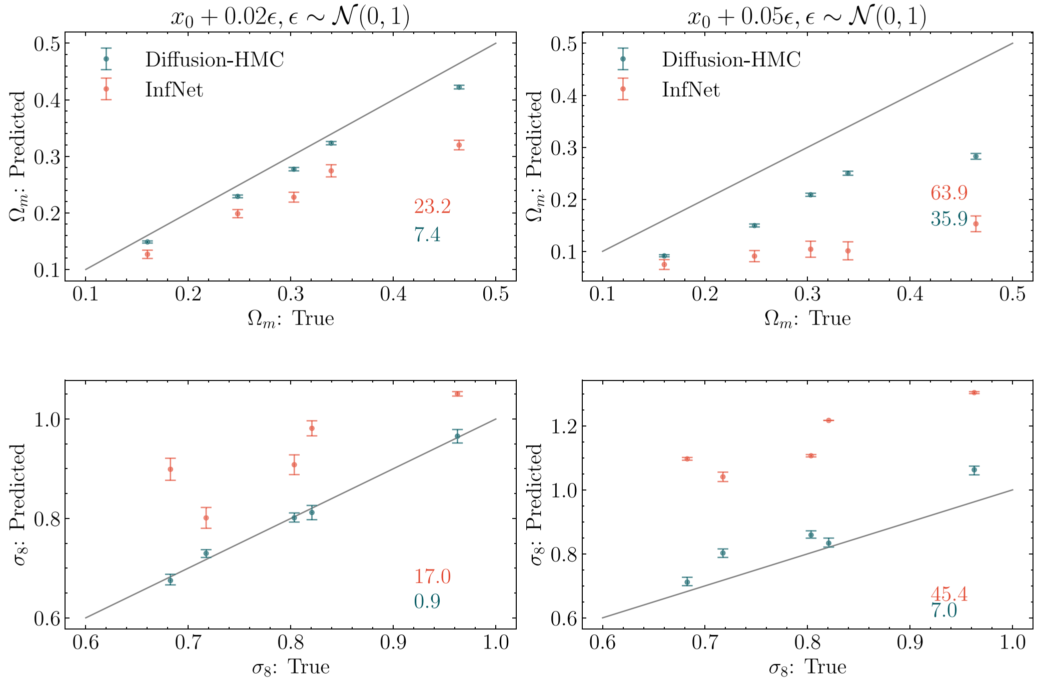

Figure 11. Robustness comparison with the first two noising time steps dropped and without a prior applied in the HMC case.

Appendix B
Summary Statistics

B.1. Reduced Chi-squared Statistics

For the CV fields, where we have 450 samples of the true and generated fields for a single parameter, we compute the reduced chi-squared statistic using an estimate of the covariance between different k bins. The number of k bins here is 35:

$$\boldsymbol{\mu} = \langle Pk_{\text{Ref}} \rangle$$

$$C = \text{Cov}[Pk_{\text{Ref}}] \qquad \hat{C}^{-1} = C^{-1}\frac{N-p-2}{N-1} \qquad \text{(B1)}$$

$$\chi_r^2(Pk_{\text{Test}}) = \sum_k (Pk_{\text{Test}} - \boldsymbol{\mu}) \cdot (\hat{C}^{-1}(Pk_{\text{Test}} - \boldsymbol{\mu})^T)^T. \qquad \text{(B2)}$$

For the LH fields during model selection, since we just have 15 fields in the true data set we cannot reliably estimate a covariance. We use the following formula instead. The number of k bins here is 128:

$$\chi_r^2(s) = \frac{1}{|k|}\sum_k \frac{(P(k)_s - \langle P(k)_{\text{TRUE}} \rangle)^2}{\sigma[P(k)_{\text{TRUE}}]^2}. \qquad \text{(B3)}$$

B.2. Across Checkpoints

For the eight checkpoints we generated 500 fields, with 50 fields for each of the 10 validation parameters. We then examined the reduced chi-squared statistic of the power spectra of the log fields, the linear fields, and the p-values of the means of the distributions of true and generated fields in Figure 12. A value of less than 0.05 would indicate that the two distributions of the means are statistically different. The p-values are above 0.05 for all eight checkpoints. Since these comparisons are limited by the number of samples in the true set for each parameter (15), we additionally plot the reduced chi-squared statistic derived by using each true field as the test and the other 14 fields as the reference. While there is some oscillation across checkpoints for each of these three statistics, the variation appears random.

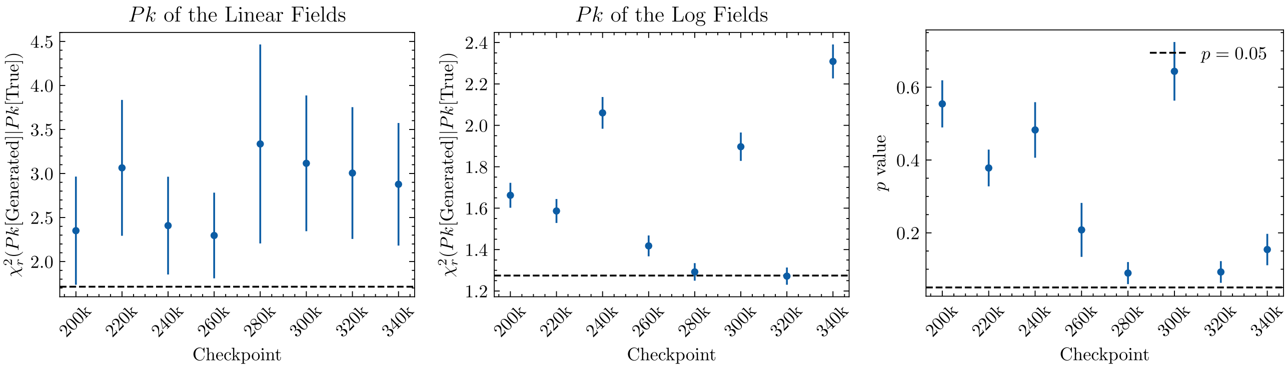

Figure 12. Mean and standard error on the mean for the reduced χ_r^2 statistic of the power spectra of the 50 linear (left) and log (center) fields for each parameter relative to the 15 true fields' power spectra for that parameter, across 10 different parameters for each of eight checkpoints. The dashed line demarcates the mean reduced χ_r^2 of the true fields using the other 14 true fields as the reference distribution (leave-one-out). Right: mean and standard error on the mean of the p-values of the distribution of the means of the 50 generated log fields relative to the distribution of the means of the 15 true fields for the same parameter. The p-values are above 0.05 for all of the eight checkpoints.

ORCID iDs

Nayantara Mudur https://orcid.org/0000-0001-5139-612X
Carolina Cuesta-Lazaro https://orcid.org/0000-0002-6069-2999
Douglas P. Finkbeiner https://orcid.org/0000-0003-2808-275X

References

Anderson, B. D. O. 1982, Stochastic Processes and Their Applications, 12, 313

Bansal, A., Borgnia, E., Chu, H.-M., et al. 2024, Advances in Neural Information Processing Systems, 36 ed. A. Oh, T. Naumann, A. Globerson et al. (New York: Curran Associates, Inc.), 41259

Betancourt, M. 2017, arXiv:1701.02434

Chen, H., Dong, Y., Shao, S., et al. 2024, Advances in Neural Information Processing Systems, 38 ed. A. Globerson & L. Mackey, (New York: Curran Associates, Inc.) arXiv:2402.02316

Clark, K., & Jaini, P. 2024, Advances in Neural Information Processing Systems, 36 ed. A. Oh, T. Naumann, & A. Globerson, (New York: Curran Associates, Inc.), 58921

Cobb, A. D., Baydin, A. G., Markham, A., & Roberts, S. J. 2019, arXiv:1910.06243

Corso, G., Stärk, H., Jing, B., Barzilay, R., & Jaakkola, T. 2023, Int. Conf. on Learning Representations

Cuesta-Lazaro, C., & Mishra-Sharma, S. 2024, PhRvD, 109, 123531

Dai, B., & Seljak, U. 2024, PNAS, 121, e2309624121

Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. 1987, PhLB, 195, 216

Hahn, C., Eickenberg, M., Ho, S., et al. 2023, PNAS, 120, e2218810120

Hamaus, N., Pisani, A., Sutter, P. M., et al. 2016, PhRvL, 117, 091302

Hartlap, J., Simon, P., & Schneider, P. 2007, A&A, 464, 399

Heitmann, K., Higdon, D., White, M., et al. 2009, ApJ, 705, 156

Heurtel-Depeiges, D., Burkhart, B., Ohana, R., & Blancard, B. R.-S. 2023, arXiv:2310.16285

Ho, J., Jain, A., & Abbeel, P. 2020, Advances in Neural Information Processing Systems, 33 ed. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin, (New York: Curran Associates, Inc.), 6840

Horowitz, B., & Melchior, P. 2022, arXiv:2211.14788

Jagvaral, Y., Mandelbaum, R., & Lanusse, F. 2022, arXiv:2212.05592

Kingma, D., Salimans, T., Poole, B., et al. 2021, Advances in Neural Information Processing Systems, 34 ed. M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, & J. Wortman Vaughan, (New York: Curran Associates, Inc.), 21696

Kingma, D. P., & Gao, R. 2023, Advances in Neural Information Processing Systems, 36 ed. A. Oh, T. Naumann, & A. Globerson, (New York: Curran Associates, Inc.), 65484, arXiv:2303.00848

Krause, E., Eifler, T., Zuntz, J., et al. 2017, arXiv:1706.09359

Legin, R., Ho, M., Lemos, P., et al. 2024, MNRAS: Letters, 527, L173

Li, A. C., Prabhudesai, M., Duggal, S., Brown, E., & Pathak, D. 2023, Proc. IEEE/CVF Int. Conf. on Computer Vision, ed. J. Kosecka et al. (Piscataway, NJ: IEEE), 2206

Mudur, N., Cuesta-Lazaro, C., & Finkbeiner, D. P. 2023, arXiv:2312.07534

Mudur, N., & Finkbeiner, D. P. 2022, arXiv:2211.12444

Mustafa, M., Bard, D., Bhimji, W., et al. 2019, ComAC, 6, 1

Neal, R. M. 2011, in Handbook of Markov Chain Monte Carlo, ed. S. Brooks et al. (Boca Raton FL: CRC Press), Ch 5

Nelson, D., Springel, V., Pillepich, A., et al. 2019, ComAC, 6, 1

Nguyen, N.-M., Schmidt, F., Tucci, B., Reinecke, M., & Kostić, A. 2024, arXiv:2403.03220

Nichol, A. Q., & Dhariwal, P. 2021, arXiv:2102.09672

Ono, V., Park, C. F., Mudur, N., et al. 2024, ApJ, 970, 174

Paillas, E., Cuesta-Lazaro, C., Percival, W. J., et al. 2023, MNRAS, 531, 898

Papamakarios, G., Pavlakou, T., & Murray, I. 2017, Advances in Neural Information Processing Systems, 30 ed. I. Guyon, U. von Luxburg, & S. Bengio, (New York: Curran Associates, Inc.)

Pillepich, A., Springel, V., Nelson, D., et al. 2018, MNRAS, 473, 4077

Prabhudesai, M., Ke, T.-W., Li, A. C., Pathak, D., & Fragkiadaki, K. 2023, Advances in Neural Information Processing Systems, 37 ed. A. Oh, T. Naumann, & A. Globerson, (New York: Curran Associates, Inc.), 17567

Régaldo-Saint Blancard, B., Allys, E., Auclair, C., et al. 2023, ApJ, 943, 9

Remy, B., Lanusse, F., Jeffrey, N., et al. 2022, 2201.05561A&A, 672, A51

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. 2021, arXiv:2112.10752

Ronneberger, O., Fischer, P., & Brox, T. 2015, in Int. Conf. on Medical Image Computing and Computer-Assisted Intervention, ed. N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Berlin: Springer), 234

Rouhiainen, A., Gira, M., Münchmeyer, M., Lee, K., & Shiu, G. 2024, PhRvD, 109, 123536

Rozet, F., Delaunoy, A., Miller, B., et al., 2021 LAMPE: Likelihood-free Amortized Posterior Estimation Version 0.8.2, Zenodo, doi:10.5281/zenodo.8405782

Sharma, D., Dai, B., & Seljak, U. 2024a, JCAP, 08, 010

Sharma, D., Dai, B., Villaescusa-Navarro, F., & Seljak, U. 2024b, arXiv:2401.15891

Shen, Z., Zhang, M., Zhao, H., Yi, S., & Li, H. 2021, in Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision, ed. G. Medioni & K. Bowyer (Piscataway, NJ: IEEE), 3530

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. 2015, PMLR, 37, 2256

Song, Y., Shen, L., Xing, L., & Ermon, S. 2022, Int. Conf. on Learning Representations

Song, Y., Sohl-Dickstein, J., Kingma, D. P., et al. 2021, Int. Conf. on Learning Representations https://openreview.net/forum?id=PxTIG12RRHS

Tange, O. 2018, GNU Parallel 2018 v1, Zenodo, doi:10.5281/zenodo.1146014

Theis, L., Oord, A. V. D., & Bethge, M. 2015, arXiv:1511.01844

Valogiannis, G., & Dvorkin, C. 2022, PhRvD, 106, 103509

Valogiannis, G., Yuan, S., & Dvorkin, C. 2024, PhRvD, 109, 103503

Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, Advances in Neural Information Processing Systems, 30 ed. I. Guyon, U. von Luxburg, & S. Bengio, (New York: Curran Associates, Inc.)

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021a, ApJ, 915, 71

Villaescusa-Navarro, F., Anglés-Alcázar, D., Genel, S., et al. 2021b, arXiv:2109.09747

Villaescusa-Navarro, F., Genel, S., Angles-Alcazar, D., et al. 2022, ApJS, 259, 61

Villanueva-Domingo, P., & Villaescusa-Navarro, F. 2022, ApJ, 937, 115

Wildberger, J., Dax, M., Buchholz, S., et al. 2024, Advances in Neural Information Processing Systems, 36 ed. A. Oh, T. Naumann, & A. Globerson, (New York: Curran Associates, Inc.), 16837

Wu, Y., & He, K. 2019, Int. J. Comput. Vis., 128, 742

Zagoruyko, S., & Komodakis, N. 2016, in British Machine Vision Conference (BMVC), ed. R. C. Wilson, E. R. Hancock, & W. A. P. Smith (London: BMVA Press), 87.1