Removing Dust from CMB Observations with Diffusion Models

David Heurtel-Depeiges *
Flatiron Institute
Ecole Polytechnique, Institut Polytechnique de Paris
david.heurtel-depeiges@polytechnique.edu

Blakesley Burkhart
Rutgers University
Flatiron Institute
bburkhart@flatironinstitute.org

Ruben Ohana†
Flatiron Institute
rohana@flatironinstitute.org

Bruno Régaldo-Saint Blancard†
Flatiron Institute
bregaldo@flatironinstitute.org

Abstract

In cosmology, the quest for primordial B-modes in cosmic microwave background (CMB) observations has highlighted the critical need for a refined model of the Galactic dust foreground. We investigate diffusion-based modeling of the dust foreground and its interest for component separation. Under the assumption of a Gaussian CMB with known cosmology (or covariance matrix), we show that diffusion models can be trained on examples of dust emission maps such that their sampling process directly coincides with posterior sampling in the context of component separation. We illustrate this on simulated mixtures of dust emission and CMB. We show that common summary statistics (power spectrum, Minkowski functionals) of the components are well recovered by this process. We also introduce a model conditioned by the CMB cosmology that outperforms models trained using a single cosmology on component separation. Such a model will be used in future work for diffusion-based cosmological inference.

1 Introduction

The cosmic microwave background (CMB) is a key cosmological observable to constrain models describing the dynamical evolution of the Universe over its nearly 14 billion years of history [19]. However, CMB observations suffer from the contamination of various "foreground" signals (of astrophysical origin) and instrumental noise. For CMB analysis, these contaminants need to be removed, thus requiring precise component/source separation methods (e.g., [20, 12]). In particular, the active search for primordial B-modes in CMB polarization observations [13] has shed light on one of the main hurdles: the refined modeling of the thermal emission of interstellar dust grains radiating on top of the CMB, also known as the "dust foreground" [5].

Over the past decades, various models of the dust foreground have been developed (e.g., [1, 10, 18, 30, 25, 24, 23]). A recent series of works has notably leveraged the dazzling progresses of deep generative models for dust modeling [3, 29, 15]. In particular, [17] made a first application of diffusion models for dust modeling and data denoising. We now take a new step in that direction by leveraging diffusion-based dust models for dust/CMB component separation.

Recently, diffusion models (DMs) have achieved state-of-the-art performance in generating high-quality natural images (e.g., [11, 27, 22, 26]). These models are trained on a denoising objective, with the goal to remove injected noise from images sampled from the target distribution. This training

*Work done during an internship at the Flatiron Institute. † Project supervisors.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

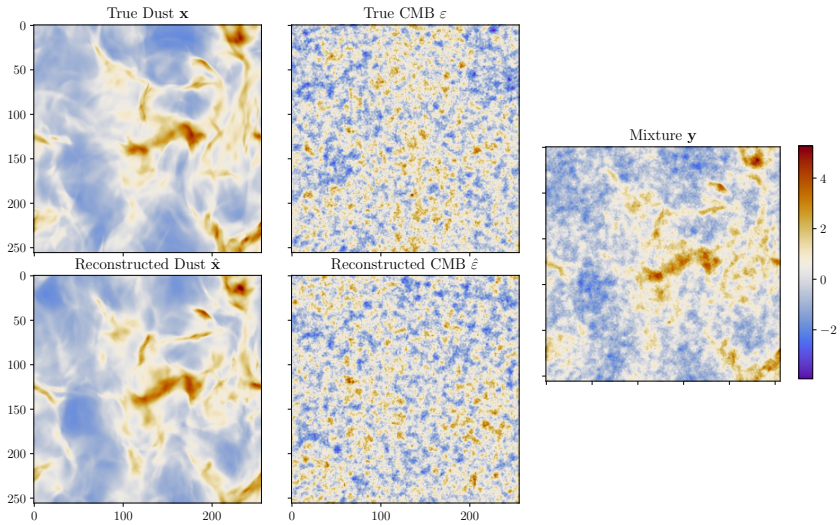

Figure 1: Observed and reconstructed maps. The true dust $\mathbf{x}$ and CMB ε maps compose the observed mixture $\mathbf{y}$. We reconstruct the dust $\hat{\mathbf{x}}$ and CMB $\hat{\varepsilon}$ with our diffusion model. The CMB maps are consistently rescaled for better visualization. The global unit is arbitrary.

framework can be formalized in terms of diffusion processes that gradually add noise over time [28]. In this work, we demonstrate that this formalism naturally applies to CMB component separation provided that the CMB can be assumed Gaussian.

We formalize our problem as follows. We observe an image $\mathbf{y} \in \mathbb{R}^{N\times N}$ that is a additive mixture of a dust signal $\mathbf{x}$ and CMB ε (with $\mathbf{x}, \varepsilon \in \mathbb{R}^{N\times N}$):

$$\mathbf{y} = \mathbf{x} + \varepsilon. \tag{1}$$

We assume that $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{\Sigma}_\phi)$, where $\mathbf{\Sigma}_\phi$ is the normalized covariance of the CMB parametrized by cosmological parameters ϕ, and σ is the CMB amplitude. In that context, given a cosmology ϕ and a prior distribution on dust $p(\mathbf{x})$, the goal of component separation is to sample $p(\mathbf{x}|\mathbf{y}, \phi)$.[2]

In this paper, we show that, given a few hundred examples of dust maps, DMs can be trained so that their sampling process coincides with the sampling of $p(\mathbf{x}|\mathbf{y}, \phi)$. We also show that usual summary statistics (power spectrum, Minkowski functionals) are recovered by this process.

Organization of the paper: we describe our data in Sect. 2 and present our method in Sect. 3. We then show our results in Sect. 4 and discuss perspectives in Sect. 5.

2 Description of the Data

Dust Emission Maps. We construct simulated dust emission maps in total intensity from a turbulent hydrodynamic simulation of the diffuse interstellar medium taken from the CATS database [6]. We assume here that dust emission is proportional to the gas density of the simulation. Simulated maps then simply correspond to gas column density maps. We provide details on the hydrodynamic simulation in App. A.1. Our dataset consists of 991 dust emission maps of size 256×256 (for an example, see Fig. 1 top left panel).

CMB maps. If we neglect secondary anisotropies, CMB anisotropies are extremely well described by an isotropic Gaussian random field on the sphere [19], entirely characterized by its covariance matrix (or power spectrum). As a consequence, we can obtain CMB maps by sampling Gaussian draws with covariance $\mathbf{\Sigma}_\phi$ parametrized by cosmological parameters ϕ (for an example, see Fig. 1 top middle panel). In this work, we consider a standard cosmological model. We only vary the cosmological parameters H_0 and ω_b and choose for the remaining ones fiducial values consistent with *Planck* 2018

[2]Note that, since $\mathbf{y} = \mathbf{x} + \varepsilon$, a sample of $p(\mathbf{x}|\mathbf{y}, \phi)$ trivially yields a sample of $p(\varepsilon|\mathbf{y}, \phi)$.

analysis [19]. We work with simulated maps covering an effective angular surface of approximately $34 \times 34\,\mathrm{deg}^2$. We refer to App. A.2 for further technical details.

3 Method

The contaminated CMB observation $\mathbf{y}$ can be viewed as the result of a forward diffusion process mapping the dust signal $\mathbf{x}$ to $\mathbf{y}$. We leverage DMs to learn the corresponding reverse diffusion process, thus enabling component separation.

3.1 Diffusion Models: Formalism

Given a forward diffusion process $(\mathbf{z}_t)_{t\in[0,1]}$ mapping the target distribution p_0 to a simple Gaussian distribution p_1, DMs are trained to reverse this process (i.e. to map p_1 to p_0). In this work, we consider the following forward diffusion process in $\mathbb{R}^d$ with $d = N^2$:

$$\mathrm{d}\mathbf{z}_t = -\frac{1}{2}\beta(t)\mathbf{z}_t\,\mathrm{d}t + \sqrt{\beta(t)}\boldsymbol{\Sigma}_\phi^{\frac{1}{2}}\,\mathrm{d}\mathbf{w}_t, \tag{2}$$

with $\beta : [0,1] \to \mathbb{R}_+$ a scalar positive function, $\boldsymbol{\Sigma}_\phi \in \mathbb{R}^{d\times d}$ the covariance matrix of the CMB, and $\mathbf{w}_t$ a standard d-dimensional Wiener process. We use a linearly increasing β adjusted so that, approximately, $p_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\phi)$. The reverse process is also a diffusion process [28]:

$$\mathrm{d}\mathbf{z}_t = \left[-\frac{1}{2}\beta(t)\mathbf{z}_t - \beta(t)\boldsymbol{\Sigma}_\phi\nabla_{\mathbf{z}_t}\log p_t(\mathbf{z}_t)\right]\mathrm{d}t + \sqrt{\beta(t)}\boldsymbol{\Sigma}_\phi^{\frac{1}{2}}\,\mathrm{d}\overline{\mathbf{w}}_t, \tag{3}$$

where $\overline{\mathbf{w}}_t$ denotes a standard backward-time Wiener process. One can then generate samples of p_0 by solving this reverse SDE starting from $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\phi)$ (see App. B.3 for further details).

3.2 Posterior Sampling and Reverse Process

We use the forward process in Eq. (2) to map the dust prior distribution $p_0 = p_{\mathrm{dust}}$ to the CMB distribution $p_1 = p_{\mathrm{CMB}|\phi}$, and train a diffusion model on mixtures of dust and CMB maps to learn the inverse mapping. This enables sampling of dust maps conditioned on CMB seeds.

DMs are commonly trained using an identity covariance matrix, $\boldsymbol{\Sigma}_\phi = \mathbf{I}_d$, resulting in a standard Gaussian distribution for p_1. By instead incorporating the true statistics of the CMB through $\boldsymbol{\Sigma}_\phi$ in the SDE, our approach is tailored to extracting the dust component from the observed mixture.

Indeed, assuming cosmology ϕ known, observation $\mathbf{y}$ can be viewed as a (rescaled) sample of $\mathbf{z}_{t^\star}$ with $t^\star \in [0,1]$ resulting from the forward process of Eq. (2). Sampling the posterior $p(\mathbf{x}|\mathbf{y},\phi)$ can be achieved by solving the reverse SDE of Eq. (3), starting at time $t^\star$ and initialized with $\mathbf{z}_{t^\star} = \mathbf{y}$ [see 2, Sect. 5]. In this work, we assume that $t^\star$ is given[3]. Future work will address the problem of inferring it from the data.

3.3 Training

While we could have trained a DM based on Eqs. (2) and (3) for a fixed cosmology ϕ, physical applications involve unknown CMB parameters. Our long-term goal being to infer ϕ from observations $\mathbf{y}$, we choose here to train a DM conditioned by the cosmology ϕ. As a proof of concept, we only consider a dependency on cosmological parameters H_0 and ω_b. With $\phi = (H_0, \omega_b)$, we choose a broad prior $p(\phi) \sim \mathcal{U}([50, 90] \times [0.0075, 0.0567])$.

Our DM is a three-level UNet with ResBlocks, trained by minimizing the denoising score-matching loss [28]. We encountered substantial training instabilities due to the large ($\sim 10^4$) condition number of $\boldsymbol{\Sigma}_\phi$. By renormalizing the loss and model outputs (details in App. B), we were able to mitigate these issues. Additional training specifics are provided in the Appendix.

[3] It is a closed form function of σ.

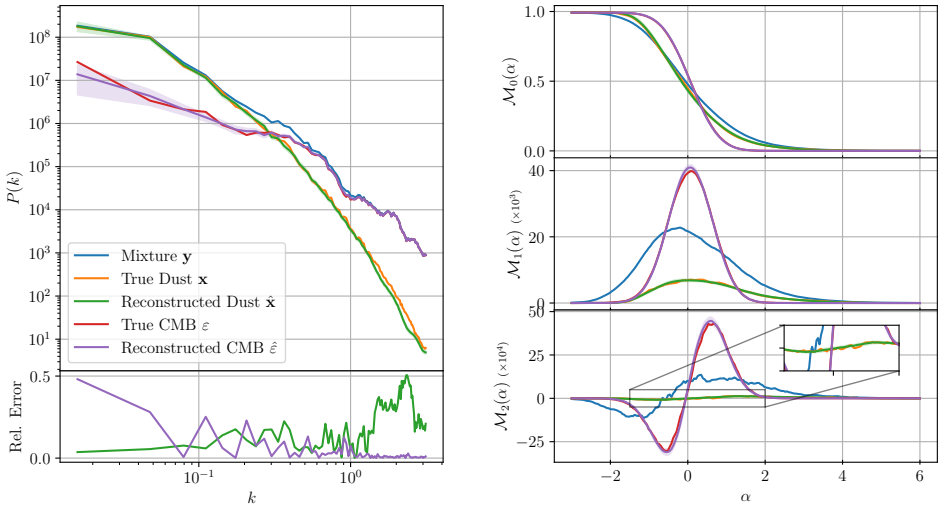

Figure 2: Reconstruction of summary statistics: power spectrum (left) and Minkowski functionals (right) of the mixture and the true and reconstructed dust and CMB maps.

4 Results

4.1 Generative Modeling

As a first validation of our diffusion model, we investigate in App. C its performances for the generation of dust emission maps. This first analysis shows that our model generalizes very well beyond its training set by producing visually and quantitatively realistic maps.

4.2 Component Separation

We evaluate our model's component separation performance on an observation $\mathbf{y}$ produced for a CMB map with $H_0 = 70$, $\omega_b = 0.0321$ (i.e. center of prior range $p(\phi)$) and amplitude $\sigma = 0.6\,\sigma_{\mathbf{x}}$.

Fig. 1 shows the reconstructed CMB and dust maps. The large-scale features of $\hat{\mathbf{x}}$ clearly match those of $\mathbf{x}$. However, at small scales, where the CMB is dominant, we observe significant discrepancies. This was expected and underlies the probabilistic aspect of the reconstruction.

We also show in Fig. 2 the reconstructed power spectra and Minkowski functionals. These match the statistics of the true components to a very good extent. The CMB spectrum is very well recovered, even at scales where the dust dominates. For dust, the power spectrum is overall well recovered. However, the relative error at small scales presents wavy patterns that exhibit a flaw of the reconstruction: these correspond to residual imprints of CMB power spectrum bumps that were not entirely smoothed out during reconstruction. Overall, the Minkowski functionals agree to a very good extent, underlying the success of our model in reconstructing non-Gaussian dust features.

We have verified that the previous results are not specific to the choice of cosmology. However, we report slightly worse reconstructions for cosmologies at the edge of the prior range $p(\phi)$.

4.3 Interest of a Cosmology-Dependent Model

We have worked with a cosmology-dependent model mainly for (long-term) purposes of diffusion-based cosmological inference. Nevertheless, we observed that, while seemingly more intricate, it was not harder to train than a DM trained for a single cosmology. In particular, it did not require more training epochs. Moreover, our cosmology-dependent model proved to be on par or perform better than single-cosmology models. We have notably observed experimentally that it yields better power spectrum reconstruction for dust generative modeling. This suggests that our model has learned a more robust representation of dust, and illustrates the interest of multi-task learning for this problem.

5 Conclusion and Perspectives

In this work, we demonstrated that diffusion models can be naturally applied to separate cosmic microwave background (CMB) signals from dust contamination, under the assumption that the CMB follows a Gaussian distribution. We trained a model conditioned on cosmological parameters to map from a prior distribution of simulated dust maps to the known CMB distribution. We showed that such a model allows for posterior dust and CMB sampling given an observed mixture and assumed cosmology. In particular, we showed that key statistics of the individual components are accurately recovered through this process. While based on simulations, this proof-of-concept establishes diffusion models as a promising approach for analyzing CMB observational data using arbitrary dust priors and less than a thousand dust examples. Future directions include developing a diffusion-based pipeline for inferring cosmological parameters ϕ from contaminated CMB observations. Additional work is needed to extend this approach to polarization data and account for other contaminants like instrumental noise. Another challenge for the rightful application of our method to observational data will be to generalize this work to the case of weakly non-Gaussian CMB signal.

Acknowledgments and Disclosure of Funding

We would like to thank Fiona McCarthy for her help with the generation of the CMB simulations, as well as Jiequn Han, Ronan Legin, Charles Margossian, Chirag Modi, Loucas Pillaud-Vivien, and Yuling Yao for valuable discussions.

References

[1] E. Allys, F. Levrier, S. Zhang, C. Colling, B. Regaldo-Saint Blancard, F. Boulanger, P. Hennebelle, and S. Mallat. The RWST, a comprehensive statistical description of the non-Gaussian structures in the ISM. *Astronomy & Astrophysics*, 629:A115, Sept. 2019.

[2] B. D. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[3] K. Aylor, M. Haq, L. Knox, et al. Cleaning our own dust: simulating and separating galactic dust foregrounds with neural networks. *Monthly Notices of the Royal Astronomical Society*, 500(3):3889–3897, Jan. 2021.

[4] S. Bialy, B. Burkhart, and A. Sternberg. The H I-to-H_2 Transition in a Turbulent Medium. *The Astrophysical Journal*, 843(2):92, July 2017.

[5] BICEP2/Keck Array and Planck Collaborations. Joint Analysis of BICEP2/Keck Array and Planck Data. *Physical Review Letters*, 114(10):101301, Mar. 2015.

[6] B. Burkhart, S. Appel, S. Bialy, J. Cho, A. Christensen, D. Collins, C. Federrath, D. Fielding, D. Finkbeiner, A. Hill, et al. The catalogue for astrophysical turbulence simulations (cats). *The Astrophysical Journal*, 905(1):14, 2020.

[7] J. Cho and A. Lazarian. Compressible magnetohydrodynamic turbulence: mode coupling, scaling relations, anisotropy, viscosity-damped regime and astrophysical implications. *Monthly Notices of the Royal Astronomical Society*, 345(12):325–339, Oct. 2003.

[8] C. Correia, B. Burkhart, A. Lazarian, V. Ossenkopf, J. Stutzki, J. Kainulainen, G. Kowal, and J. R. de Medeiros. Opacity Broadening of ^{13}CO Linewidths and its Effect on the Variance-Sonic Mach Number Relation. *The Astrophysical Journal Letters*, 785(1):L1, Apr. 2014.

[9] C. Correia, J. R. D. Medeiros, A. Lazarian, B. Burkhart, and E.-m. c. Pogosyan, D. Principal component analysis studies of turbulence in optically thick gas. *The Astrophysical Journal*, 818(2), 2 2016.

[10] D. P. Finkbeiner, M. Davis, and D. J. Schlegel. Extrapolation of Galactic Dust Emission at 100 Microns to Cosmic Microwave Background Radiation Frequencies Using FIRAS. *The Astrophysical Journal*, 524:867–886, Oct. 1999.

[11] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[12] N. Jeffrey, F. Boulanger, B. D. Wandelt, B. Regaldo-Saint Blancard, E. Allys, and F. Levrier. Single frequency CMB B-mode inference with realistic foregrounds from a single training image. *Monthly Notices of the Royal Astronomical Society: Letters*, 510(1):L1–L6, 11 2021.

[13] M. Kamionkowski and E. D. Kovetz. The quest for b modes from inflationary gravitational waves. *Annual Review of Astronomy and Astrophysics*, 54(1):227–269, 2016.

[14] G. Kowal, A. Lazarian, and A. Beresnyak. Density Fluctuations in MHD Turbulence: Spectra, Intermittency, and Topology. *The Astrophysical Journal*, 658(1):423–445, Mar. 2007.

[15] N. Krachmalnicoff and G. Puglisi. ForSE: A GAN-based Algorithm for Extending CMB Foreground Models to Subdegree Angular Scales. *The Astrophysical Journal*, 911(1):42, Apr. 2021.

[16] A. Lewis, A. Challinor, and A. Lasenby. Efficient computation of cosmic microwave background anisotropies in closed friedmann-robertson-walker models. *The Astrophysical Journal*, 538(2):473, aug 2000.

[17] N. Mudur and D. P. Finkbeiner. Can denoising diffusion probabilistic models generate realistic astrophysical fields? *ArXiv*, abs/2211.12444, 2022.

[18] Planck Collaboration XI. *Planck* 2013 results. XI. All-sky model of thermal dust emission. *Astronomy & Astrophysics*, 571:A11, 2014.

[19] Planck Collaboration I. *Planck* 2018 results. I. Overview, and the cosmological legacy of *Planck*. *Astronomy & Astrophysics*, 641:A1, 2020.

[20] Planck Collaboration IV. *Planck* 2018 results. IV. Diffuse component separation. *Astronomy & Astrophysics*, 641:A4, 2020.

[21] S. K. N. Portillo, Z. Slepian, B. Burkhart, S. Kahraman, and D. P. Finkbeiner. Developing the 3-point Correlation Function for the Turbulent Interstellar Medium. *The Astrophysical Journal*, 862(2):119, Aug. 2018.

[22] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[23] B. Régaldo-Saint Blancard, E. Allys, C. Auclair, F. Boulanger, M. Eickenberg, F. Levrier, L. Vacher, and S. Zhang. Generative Models of Multichannel Data from a Single Example-Application to Dust Emission. *The Astrophysical Journal*, 943(1):9, Jan. 2023.

[24] B. Regaldo-Saint Blancard, E. Allys, F. Boulanger, F. Levrier, and N. Jeffrey. A new approach for the statistical denoising of Planck interstellar dust polarization data. *Astronomy & Astrophysics*, 649:L18, May 2021.

[25] B. Regaldo-Saint Blancard, F. Levrier, E. Allys, E. Bellomi, and F. Boulanger. Statistical description of dust polarized emission from the diffuse interstellar medium. A RWST approach. *Astronomy & Astrophysics*, 642:A217, Oct. 2020.

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[28] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[29] B. Thorne, L. Knox, and K. Prabhu. A generative model of galactic dust emission using variational autoencoders. *Monthly Notices of the Royal Astronomical Society*, 504(2):2603–2613, June 2021.

[30] F. Vansyngel, F. Boulanger, T. Ghosh, et al. Statistical simulations of the dust foreground to cosmic microwave background polarization. *Astronomy & Astrophysics*, 603:A62, July 2017.

A Complements on Data

A.1 Hydrodynamic Simulation

The simulation uses a third-order-accurate hybrid essentially non-oscillatory scheme [7] to solve the fluid hydrodynamic equations. It has periodic boundary conditions and an isothermal equation of state $p = c_s^2 \rho$, with c_s the isothermal sound speed. Turbulent motions are generated with a random large-scale solenoidal driving at a wave number $k \approx 2.5$ (i.e. 1/2.5 the box size) and the driving is continuous to prevent the decay of turbulence in one turn over time. The simulation has 256^3 resolution elements. Related simulations have been described and used in many previous works [7, 14, 8, 9, 4, 21]. They are part of the Catalog for Astrophysical Turbulence Simulations [6].

The primary control parameters of the simulations are the dimensionless sonic Mach number, $\mathcal{M}_s \equiv |\mathbf{v}|/c_s$, which in our case is 7. $\mathbf{v}$ is the velocity, c_s and v_A are respectively the isothermal sound speed and Alfvén speed.

A.2 CMB Simulations

Given a set of cosmological parameters ϕ, CMB power spectra are computed using `CAMB` [16]. Gaussian draws on the sphere are then projected on 256×256 patches using `pixell` [4] with a pixel size of $8'$. We only vary the cosmological parameters H_0 and ω_b in this work and use $\Omega_K = 0$, $\omega_c = 0.12$, $\tau = 0.0544$, $n_s = 0.9649$, $\ln(10^{10} A_s) = 3.044$, $m_\nu = 0.08$.

B Models and Training Details

We withheld 10% of the 991 images to validate the quality of generation/denoising of our diffusion model trained on the remaining 90%.

B.1 Architecture

Backbone Architecture. Our score network is a three-level UNet with ResBlocks and a bottleneck of size 32×32 or 16×16. Each level has one ResBlock with three convolutions, whether descending or ascending and use GroupNorm as normalization with SiLU for activation. In each block, we rescale the skip-connections. Skip-connections across the bottleneck are concatenated and not added, contrary to internal skip-connections. We also enforce periodic boundary conditions by using 'circular' padding in our convolutions.

Time t is also as seen an input, as in [11, 28], and a sine/cosine time embedding is used. After being transformed by a MLP, this embedding is added at each block. (whose structure differs if the SDE is discretized a priori or if we train in continuous time) and an MLP-transformed version of this embedding is then added at the beggining of each block.

Cosmology-Dependent Model. In the case of a parametrized family of SDE (Eq. (2) and corresponding paragraph) we need to train a diffusion model to reverse a continuum of forward SDEs parametrized by ϕ. We adapt the above architecture by adding a dependency to ϕ that is parallel to the dependency on time t. We embed ϕ with a single linear layer with activation and then, for each ResBlock, this initial embedding is further transformed by another MLP (see Fig. B.1).

B.2 Loss Functions

Single Cosmology. Our model is trained by minimizing the denoising score-matching loss which is:

$$\mathcal{L}(\theta) = \mathbb{E}_t \left[\lambda(t) \mathbb{E}_{\mathbf{z}_0, \mathbf{z}_t | \mathbf{z}_0} \left[\left\| \mathbf{s}_\theta(\mathbf{z}_t, t) + \mathbf{\Sigma}^{-1} \left(\mathbf{z}_t - \sqrt{\bar{\alpha}(t)} \mathbf{z}_0 \right) / \left(1 - \bar{\alpha}(t)\right) \right\|_2^2 \right] \right], \tag{4}$$

with $\bar{\alpha}(t) = \exp(-\int_0^t \beta(u) \mathrm{d}u)$. When generating samples, we then multiply the output of $\mathbf{s}_\theta$ by $\mathbf{\Sigma}$. This can induce numerical errors that are difficult to counteract if the matrix $\mathbf{\Sigma}$ is ill-conditioned (which is our case). Another downside is the fact that training seems extremely unstable and difficult to monitor with a loss that doesn't decrease.

[4] https://github.com/simonsobs/pixell

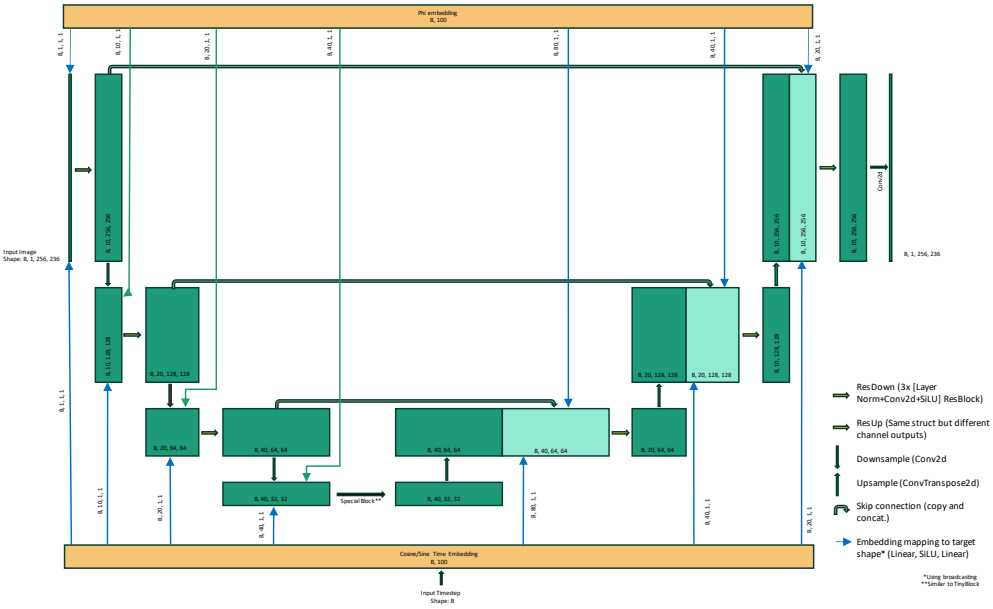

Figure B.1: Architecture of our diffusion model for a parametrized family of SDE.

We found that re-normalizing what we learn could help dealing with both of issues: instead of $\mathbf{s}_\theta$, we learn $\mathbf{m}_\theta = -\sqrt{1-\bar{\alpha}(t)}\boldsymbol{\Sigma}\,\mathbf{s}_\theta$.

The square root part of this renormalization is standard, the -1 is for convenience but the $\boldsymbol{\Sigma}$ is new and its role is different than that of the square root. In DDPMs and score-based models, square root renormalization (or its equivalent in the case of different SDE) is implemented so that models try to learn a multivariate standard Gaussian instead of a quantity with widely varying variance.

Multiple Cosmologies. The generalized training loss is a straightforward extension of Eq. (4):

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\phi}\left\{\lambda(t)\mathbb{E}_{\mathbf{z}_0,\,\mathbf{z}_t|(\mathbf{z}_0,\phi)}\left[\left\|\mathbf{m}_\theta(\mathbf{z}_t,t,\phi) - \frac{1}{\sqrt{1-\bar{\alpha}(t)}}\left(\mathbf{z}_t - \sqrt{\bar{\alpha}(t)}\mathbf{z}_0\right)\right\|_2^2\right]\right\} \tag{5}$$

Training examples are generated by combining dust samples and CMB realizations for cosmological parameters randomly drawn from the prior $p(\phi)$. Since the computation of a single covariance matrix $\boldsymbol{\Sigma}$ takes a few seconds, we train neural emulator to approximate $\phi \to \boldsymbol{\Sigma}_\phi$ as described in Appendix B.6 for faster generation of training samples.

B.3 Sampling Procedure

The standard sampling scheme for an SDE is Euler-Murayama (equivalent to Euler's method for ODE). Given a SDE:

$$\mathrm{d}\mathbf{z}_t = \mathbf{f}(\mathbf{z}_t,t)\mathrm{d}t + \mathbf{G}(\mathbf{z}_t,t)\mathrm{d}\mathbf{w}_t,$$

and an initial point $\mathbf{z}_0$, Euler-Murayama sampling for a timestep Δt follows this update rule:

$$\forall k,\;\; \mathbf{z}_{(k+1)\Delta t} = \mathbf{z}_{k\Delta t} + \Delta t\,\mathbf{f}(\mathbf{z}_{k\Delta t},k\Delta t) + \sqrt{\Delta t}\;\mathbf{G}(\mathbf{z}_{k\Delta t},k\Delta t)\varepsilon,$$

with $\varepsilon \sim \mathcal{N}(0,\mathbf{I}_d)$. Note that variants of Euler-Murayama exist with equivalent of the Runge-Kutta method or adaptive learning rate schemes. A very similar equation exists for sampling in reverse time.

For the majority of our experiments, we discretized SDEs before training and used therefore a fixed time schedule for SDE and ODE sampling. The rest of our implementation is close to that of [28] with inputs from different freely available repository.

B.4 Training heuristics

We performed a number of coarse grid search hyper-parameter optimization for different versions of the problem that we solve. In this section, we discuss failures on architectures tests.

Overfitting. In comparison with diffusion models trained under white noise, our framework is much more prone to overfitting. This is made worse by the small dataset size we have had to work with. Our understanding is that because the noise (CMB) power spectrum value at small scales is much less than that of white noise, small and medium scales are less affected by our forward diffusion process, making it easier for the network to recognize structures in the images, therefore overfitting.

We also tried our framework on choices of $\mathbf{\Sigma}$ not corresponding to CMB realizations. For some value of said covariance matrix that whose condition number is more extreme than that of CMB covariances, training was impossible because the model systematically overfitted.

For generating CMB realizations, we found there exists a range of model capacities that avoid overfitting. Specifically, UNet architectures with a large bottleneck size and a small number of channels tend to perform best.

Choice of hyperparameters. Apart from avoiding overfitting, the choice of hyper-parameters did not seem to impact much of model performance. We typically maxed out capacity without overfitting and models produced this way tended to have the best performance.

Learning rate. We used AdamW with an inverse square root schedule with a peak ranging from 4×10^{-4} to 10^{-3}

Number of discretization steps. Increasing this value in the SDE produced better results, as expected. Our understanding is that it is mainly due to small values of t where details are sampled. At small scales, a small time step allows us to mitigate (in part) what can be observed in the power spectrum of samples from our generative models: oscillations in power spectrum values at small scales correspond to those of the CMB power spectrum (see Fig. C.2).

Training. When training our new diffusion models, we observed the minimum loss reached was substantially higher (1-2 orders of magnitude) than that of white noise diffusion models, depending on the settings and dataset. However, as with white noise models, we found it was essential to continue training long after the loss plateaued, occupying 80-90% of total training time, since interrupting training too early resulted in poorer performance based on summary statistics and visual assessment.

B.5 Overfitting Checks

In order to verify that the model is not overfitting, we check for mode collapse and dataset copying with a specific L_2 distance taking into account the periodic boundary conditions. This was particularly important as our models seem more prone to overfitting than model trained under white noise. Further discussion of why can be found in App.B.4.

B.6 Emulator for the map $\phi \to \mathbf{\Sigma}_\phi$

We train an emulator to compute efficiently the power spectrum of the CMB for arbitrary $\phi = (H_0, \omega_b)$ parameters withing a prior range. We use an architecture based on a simple MLP with two hidden layers and 100 hidden neurons per hidden layer.

B.7 Hardware and training times.

All of our training was performed on NVIDIA A100 GPUs. One epoch takes $\sim$ 3 seconds for approximately 1,000 images. We would usually train for between 10,000 and 20,000 epochs or between 10 and 20 hours (taking into account time for sampling during training). At inference, sampling 128 images with 1,000 steps in Euler-Maruyama takes $\sim$ 1 minute. We found that for batch sizes larger than 16, sampling time is directly proportional to the number of images (as well as to the number of discretization steps).

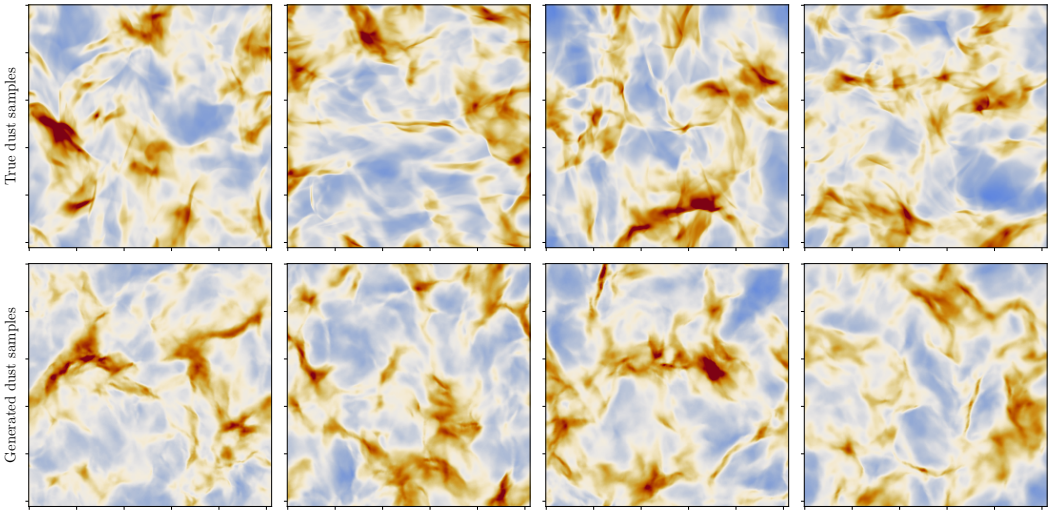

Figure C.1: Examples of original (top row) and generated (bottom row) dust emission maps.

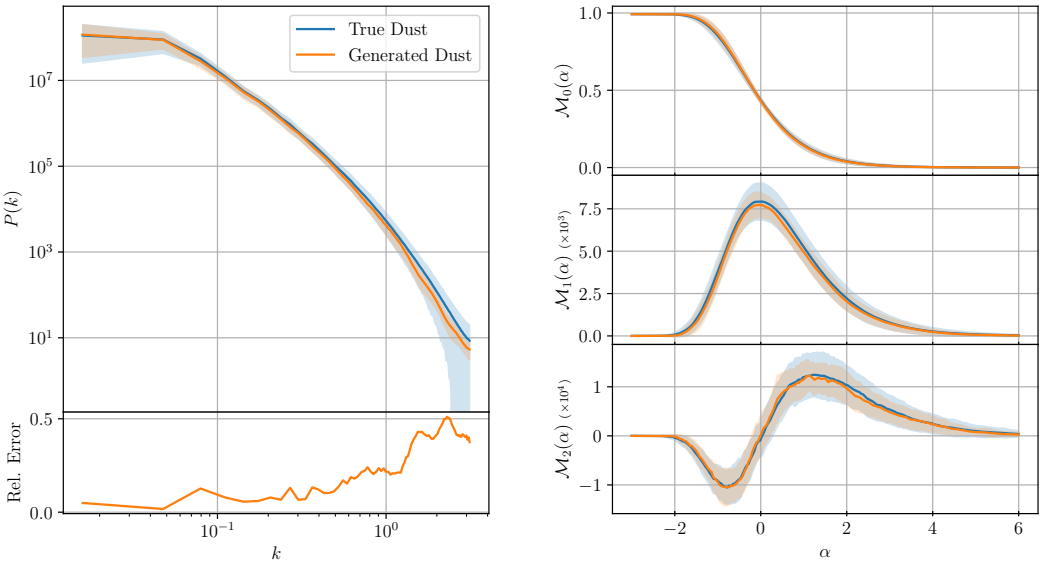

Figure C.2: Comparison of the power spectrum statistics (left) and Minkowski functionals (right) between dust samples from our diffusion model and training dust data.

C Model Performances for Dust Generative Modeling

We assess the abilities of our model in terms of generative modeling by generating dust samples and comparing them to training examples using the following:

Visual assessment. We compare in Fig. C.1 examples of generated dust samples (bottom row) to a random selection of original samples (top row). Maps are visually indistinguishable, which is a demonstration of the quality of the learned generative model. We rule out any potential trivial overfitting concerns using a custom metric defined in B.5, corroborating the ability of our model to generalize beyond its training set.

Quantitative assessment. In Fig. C.2 we compare generated and original samples using the empirical power spectrum (left) and Minkowski functionals (right). The power spectrum captures Gaussian

features, while Minkowski functionals quantify non-Gaussian information. Means and standard deviations are computed over 64 samples.

Overall, summary statistics match to a very good extent. Minkowski functionals are very well recovered both in terms of mean and standard deviation. The mean power spectrum is also well recovered especially at larger scales. However, we notice a small lack of power for the generated samples at small scales as well as lower standard deviations. We attribute this to three phenomena: UNet averages at small scales, a lack of data, and a finite number of discretization steps as we approach $t = 0$. We emphasize that for applications to source separation, errors made on the reconstruction of the dust power spectrum could directly impact the power spectrum of the reconstructed CMB. However, because these errors are mainly concentrated at small scales, where the CMB is dominant, the reconstruction error on the CMB power spectrum is negligible at small and medium scales (see Fig.2). We keep in mind that this could potentially introduce bias for cosmological inference applications, and we will address this issue in future work.