

SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area — PhD thesis in
Theoretical and Scientific Data Science

Generating the Universe: Field-Level Diffusion Models for Emulation, Reconstruction, and Posterior Inference

Candidate:
Satvik Mishra

Advisors:
Prof. Roberto Trotta
Prof. Matteo Viel

Academic Year 2025–2026



Abstract

The distribution of neutral hydrogen and its connection to the underlying dark matter field is a powerful cosmological probe, but predicting the observable signal is hindered by the large computational cost of hydrodynamical simulations, particularly on the small, non-linear scales that upcoming intensity-mapping and imaging surveys are most sensitive to. This thesis develops a suite of machine learning methods, built around conditional generative diffusion models, to replace or augment these simulations directly at the field level, across the chain from field generation to field-level inference.

We first introduce **LODI**, a pipeline combining an attention-based ResUNet (**HALOgen**) with a conditional variational diffusion model to generate high-fidelity 21 cm brightness temperature maps from dark matter-only simulations, recovering the power spectrum to within 10% for $k \leq 10 h \text{ Mpc}^{-1}$.

We then present **Cosmo-FOLD**, which removes **LODI**'s explicit sub-volume overlap prescription in favor of a sliding window diffusion trajectory, generating large, spatially coherent 3D fields from small trained sub-volumes without the boundary artifacts, or the added overlap compute of tiling-based approaches. Trained on $25^3 h^{-1} \text{ Mpc}$ **CAMELS** volumes, **Cosmo-FOLD** generalizes to the $205 h^{-1} \text{ Mpc}$ IllustrisTNG box with the power spectrum accurate to within 10% up to $k \lesssim 5 h \text{ Mpc}^{-1}$, and a positional encoding scheme further improves the recovered bispectrum across several configurations. We also adapt **Cosmo-FOLD** to generate field level dark matter posteriors conditioned on sparse stellar tracers and an external Planck prior on Ω_m and σ_8 , injected at inference time via Monte Carlo marginalization with no retraining. This is presented as a first, single field proof of concept demonstration rather than a fully validated result.

Next, we develop **FOREST**, that rests on the framework of Diffusion Posterior Sampling (DPS), using a latent Diffusion model as a generative prior, for application on the field level line of sight reconstruction of the full physical state of the intergalactic medium from Lyman- α forest sightlines, using a differentiable physics based forward model, and

only a single observation of Lyman- α flux. Finally, we develop a fast neural emulator mapping dark matter fields to HI and galaxy fields, conditioned on six astrophysical and cosmological parameters, and embed it within a simulation-based inference pipeline. Moving from power-spectrum to full 3D field-level summaries increases the median figure of merit on Ω_{m} and σ_8 by a factor of ~ 3 , and combining both tracers rather than either alone increases it by a further factor of 2–7 depending on configuration, with relative parameter bias staying below 1.5% throughout.

Together, these results show that diffusion based generative models can reproduce cosmological and astrophysical fields with high fidelity at otherwise computationally prohibitive volumes, and, in their most developed form, be turned into practical posterior samplers, recovering cosmological information that is otherwise discarded when a field is compressed to a summary statistic.

Acknowledgements

This thesis would not exist without the guidance and honest feedback of Prof. Roberto Trotta and Prof. Matteo Viel throughout these 4 years, who supervised this work throughout, and the wider AstroML group at SISSA, whose discussions repeatedly caught problems worth solving. Parts of Chapter 6 were developed jointly with Dr. Giulio Scelfo and fellow comrade Mauro Rigo; their contributions are gratefully acknowledged and disclosed in that chapter. I also thank the members of my thesis committee for their time and careful reading this work.

Contents

Abstract	i
Acknowledgements	iii
Acknowledgements	iii
1 Introduction	1
1.1 The Cosmological Context	2
1.1.1 The expanding Universe	3
1.1.2 Growth of structure	4
1.1.3 The power spectrum and Gaussianity	4
1.1.4 The nonlinear scale	5
1.1.5 The Jeans scale	5
1.1.6 Beyond linear theory: a brief note on perturbation theory	6
1.1.7 Large scale structure for cosmological parameter inference . . .	8
1.1.8 Thesis Outline	9
1.2 Bayesian Inference Basics	11
1.3 Machine Learning for Field Level Cosmology	12
1.3.1 Convolutional neural networks and the U-Net	13
1.3.2 Attention	17
1.3.3 Normalizing flows	18
1.3.4 Simulation-based inference and neural posterior estimation . . .	20
1.3.5 Variational autoencoders	21
1.3.6 An introduction to diffusion models	22
1.3.7 Diffusion posterior sampling	25
1.4 Cosmological Simulations	27
1.4.1 N-body simulations	28

1.4.2	Hydrodynamical simulations and subgrid physics	29
1.4.3	The IllustrisTNG simulations	30
1.4.4	CAMELS	31
1.4.5	The Lyman-alpha forest	31
1.5	21 cm Cosmology and Intensity Mapping	32
1.5.1	The hyperfine transition and spin temperature	32
1.5.2	Self shielding and optical depth	33
1.5.3	Intensity mapping	34
1.5.4	Modeling the HI field	34
1.6	Galaxy Surveys and 21 cm Experiments	35
1.6.1	Galaxy Surveys	36
1.6.2	21 cm Experiments	37
1.7	Addressing the simulation bottleneck	39
1.8	Statement of Contributions	40
2	LODI: Large, Fast and Accurate Neutral Hydrogen Intensity Maps with Latent Overlap Diffusion	42
2.1	Problem statement	42
2.2	LODI pipeline	43
2.3	Data and simulations	43
2.4	Methodology	44
2.4.1	HALOgen: Extracting halos from DM only sim	44
2.4.2	LODI: Generating 21 cm brightness temperature fields with Latent overlap diffusion	51
2.4.3	Generating larger volumes with latent overlap	60
2.5	Results	64
2.5.1	Performance of HALOgen	66
2.5.2	Performance of LODI	67
2.5.3	Computational cost	68
2.6	Conclusions	69
3	Cosmo-FOLD: Large-Volume Cosmological Field Generation	73
3.1	Problem statement	73
3.2	A plug-in solution	74
3.3	Methodology	74

3.3.1	Diffusion model architecture	74
3.3.2	Training dataset	76
3.3.3	Cosmo-FOLD: Field Overlap Latent Diffusion	78
3.3.4	Position encodings for higher order correlation	81
3.4	Results	83
3.4.1	Coherent gas temperature generation	84
3.4.2	Upscaling to IllustrisTNG	85
3.4.3	Dark matter from stellar density fields with upscaling	87
3.4.4	Improving Bispectrum via positional encoding	89
3.4.5	Computational cost	93
3.5	Conclusions	93
4	Posterior Predictive Fields under Planck Cosmology	97
4.1	Introduction	97
4.2	Methodology	99
4.2.1	CAMELS Multifield Dataset	99
4.2.2	Training and validation split	99
4.2.3	Network modifications for training	100
4.3	Sampling from dark matter posterior field	102
4.4	Results	103
4.5	Discussion	105
5	FOREST: Diffusion Posterior Sampling for Field Level Posterior Reconstruction	108
5.1	Introduction	108
5.2	Novelty: Generative prior	109
5.3	Data Preparation	111
5.3.1	Hydrodynamical simulations and line-of-sight fields	111
5.3.2	Sightline data format	112
5.4	The Variational Diffusion Prior	112
5.4.1	Multi-field representation	113
5.4.2	Latent diffusion and Anisotropic geometry	113
5.4.3	Training the FOREST pipeline	115
5.5	Theory of the Lyman-alpha Optical Depth	117
5.5.1	Physical picture	117

5.5.2	The Voigt function and the Tepper-Garcia approximation	119
5.5.3	Discretized optical depth along a sightline	119
5.6	Diffusion Posterior Sampling	120
5.6.1	Specializing the general formulation to Lyman-alpha tomography	120
5.6.2	The Lyman-alpha likelihood	121
5.6.3	The guidance gradient and sampling algorithm	122
5.7	Results	122
5.7.1	Metrics	124
5.7.2	1D LoS reconstruction	125
5.7.3	3D field	128
5.8	Discussion	129
5.8.1	What the results show	129
5.8.2	Limitations of the current results	130
5.8.3	Future work	131
6	Field-Level Emulation for Multi-Tracer Simulation-Based Inference	135
6.1	Introduction	135
6.2	Pipeline overview	137
6.3	Dark matter maps and target observables	137
6.4	The HGL emulator: architecture	139
6.5	Emulator training	142
6.6	Downstream use: simulation-based inference	143
6.7	Results	144
6.7.1	Emulator validation	144
6.7.2	Impact on cosmological constraints	145
6.8	Discussion	154
7	Conclusion, Caveats and Future Work	156
7.1	The common thread	156
7.2	What has been demonstrated	157
7.3	Some critical points	158
7.3.1	Training data volume and generalization	158
7.3.2	Computational cost	159
7.3.3	Which loss function?	159
7.3.4	Limitations and caveats	160

7.4	Future work	161
7.4.1	Full box, multi simulation IGM reconstruction	161
7.4.2	Lightcones: stitching multi redshift emulator outputs	162
7.4.3	Simulation-independent (JEPA-style) representations	162
7.4.4	Diffusion based likelihood for simulation discrimination	163
7.4.5	Extending DPS to survey-realistic foreground and RFI removal	164
7.4.6	Structured (large-scale-structure) initialization of the reverse dif- fusion process	164
7.5	Closing remark	165

List of Figures

1.1	Canonical CNN architecture and a convolutional kernel	14
1.2	Architecture of a standard U-Net	15
1.3	Visualizing the attention mechanism	19
1.4	Forward and reverse processes of a diffusion model	23
2.1	Overview of the full HALOgen + LODI generative pipeline	45
2.2	HALOgen U-Net architecture for halo assignment	46
2.3	Halo density channel masking prescription	50
2.4	Denoising U-Net architecture of LODI	56
2.5	Architecture of a single residual block	58
2.6	Comparison of overlap methods at a box boundary	61
2.7	Latent overlap denoising process on a 2D slice	63
2.8	Assembly of the full volume via staged latent overlap	65
2.9	Pipeline output: dark matter input to generated 21 cm map, compared against independent tiling	66
2.10	HALOgen predicted vs. target halo mass density maps	70
2.11	Generated vs. ground-truth brightness temperature map and the power spectrum	71
2.12	Residual power spectrum across 7 test simulations	72
3.1	Denoising network architecture of Cosmo-FOLD	75
3.2	Visualizing Cosmo-FOLD 's sliding-window generation at different diffusion steps	81
3.3	Generated vs. ground-truth gas temperature map, for different stitching schemes	84
3.4	Upscaling to larger volumes with Cosmo-FOLD	86
3.5	Voxel statistics for Cosmo-FOLD -generated IllustrisTNG gas temperature	87

3.6	Voxel statistics for Cosmo-FOLD-generated ρ_{DM} of IllustrisTNG	88
3.7	Cosmo-FOLD trained on CAMELS, applied to generate IllustrisTNG volumes	90
3.8	Dark matter power spectrum: Cosmo-FOLD CAMELS-to-IllustrisTNG up- scaling	91
3.9	Bispectrum comparison: Cosmo-FOLD vs. Cosmo-FOLD+pos	92
3.10	Power spectrum and voxel histogram, positional encoding variant . . .	93
4.1	Sampled field-level dark matter posterior mean and variance	104
4.2	Posterior residual histograms, under- vs. over-dense voxels	105
4.3	Sampled posterior distributions for three voxels situated at different densities.	106
5.1	Convolutional structure of the VAE used in FOREST	115
5.2	1D line of sight power spectrum at SNR= 5 with FOREST	126
5.3	1D line of sight power spectrum at SNR= 50 with FOREST	126
5.4	Reconstructed sightline profiles for the baryonic IGM fields	127
5.5	Correlation coefficients vs. observed sightline coverage	128
5.6	Reconstructed sub-volume slices at SNR= 50 for different smoothing, with FOREST	133
5.7	Reconstructed sub volume slices at SNR= 5 for different smoothing . .	134
6.1	Overview of the full HGL emulator pipeline	137
6.2	Modified FiLM-conditioned Residual block for the HGL emulator	140
6.3	HGL emulator architecture	142
6.4	Emulator performance at $25 h^{-1}$ Mpc: galaxy tracer	145
6.5	Emulator performance at $25 h^{-1}$ Mpc: HI tracer	146
6.6	Emulator performance at $50 h^{-1}$ Mpc: galaxy tracer	147
6.7	Emulator performance at $50 h^{-1}$ Mpc: HI tracer	148
6.8	Visualizing the Figure of Merit table	149
6.9	Posteriors on Ω_m, σ_8 : fixed astrophysics	150
6.10	Posteriors on Ω_m, σ_8 : free astrophysics	151
6.11	Posteriors: single- vs. multi-tracer comparison	153

List of Tables

3.1	Computational cost of Cosmo-FOLD use cases	94
4.1	CAMELS Latin Hypercube parameter ranges	100
5.1	Binary layout of a sightline data file	112
6.1	HGL emulator exact architecture specifics	141
6.2	Median Figure of Merit by summary vs. tracer and combination	147
6.3	Relative bias on Ω_m and σ_8 prediction	148

Chapter 1

Introduction

Modern cosmology is, in large parts, an observationally driven science. But for much of its early history, it was necessarily mostly speculative pursuit, constrained by relatively little observational evidence. That began to change when Friedmann’s and Lemaître’s solutions to Einstein’s field equations showed that an expanding universe (non static) was a mathematically consistent possibility, years before Hubble’s measurements of galaxy redshifts and distances established the empirical evidence for cosmic expansion, culminating in Hubble’s 1929 demonstration of the distance velocity relation, moving cosmology decisively into a quantitative, testable science within a single generation. In the century since, that same expanding, evolving picture has been refined into the tightly constrained Λ CDM model underlying everything in this thesis: a Universe whose structure, namely the galaxies, filaments and voids observed today, grew from tiny primordial density fluctuations under the relentless pull of gravity. What has changed most dramatically since is not this underlying physical picture, largely in place by the 1980s, but the tools available to test it. It went from a handful of measured redshifts and pen and paper estimates, to satellite grade CMB maps, million galaxy spectroscopic surveys, and, as this Chapter reviews, cosmological simulations and machine learning models capable of extracting information from these datasets at a level of detail unimaginable just a couple of decades ago. Now, as we welcome the new era of data intensive cosmology, we race against time and look for modern techniques that will help us further the cause of precision cosmology by enabling us to use the data in the fastest and most efficient way possible. Having said this, let us start by first defining what sort of data we actually deal with, and more importantly, why.

1.1 The Cosmological Context

Neutral (atomic) hydrogen (HI) plays an important role in cosmology and structure formation processes. Its distribution follows the underlying matter density field, and for most of the cosmic history it constitutes the content of baryons to fuel star formation. This makes it a powerful and novel tracer of the large scale structure (LSS) of the Universe (e.g. [Pritchard & Loeb 2012](#); [Ansari et al. 2012](#); [Santos et al. 2015](#); [Villaescusa-Navarro et al. 2014](#)). While Cosmic Microwave Background (CMB) observations ([Hinshaw et al. 2013](#); [Planck Collaboration VI 2020](#); [Mallaby-Kay et al. 2021](#)) and galaxy redshift surveys (e.g. [Alam et al. 2017](#)) have significantly constrained the parameters of the standard Λ CDM cosmological model, some key questions are still unresolved. In particular, the fundamental nature of dark matter and dark energy is still unknown, and persistent tensions between different cosmological measurements have yet to be fully understood, including the recent evidence for evolving dark energy (e.g. [Riess et al. 2019](#); [Verde et al. 2019](#); [Wong et al. 2020](#); [DESI Collaboration et al. 2025](#)). Mapping the large scale distribution of HI and tracking its evolution over cosmic time offer a complementary approach to traditional galaxy surveys, by probing large volumes at high redshifts, and thus potentially providing stress tests for many cosmological models in new regimes (e.g. [Bull et al. 2016](#); [Villaescusa-Navarro et al. 2015](#); [Obuljen et al. 2018](#); [Berti et al. 2022](#); [Villaescusa-Navarro et al. 2017](#)).

The 21 cm line, which results from the forbidden spin flip transition within the hyperfine structure of the ground state of neutral hydrogen (see [Furlanetto et al. \(2006\)](#)), is redshifted by the cosmological expansion and can be observed at radio wavelengths. A rapidly growing set of experiments, interferometric arrays and single dish instruments alike (reviewed in Section 1.6.2), aims to detect this signal using intensity mapping (IM) techniques ([Bharadwaj et al. 2001](#); [Battye et al. 2004](#); [McQuinn et al. 2006](#); [Chang et al. 2008](#); [Seo et al. 2010](#); [Battye et al. 2013](#); [Kovetz et al. 2017](#)), and several have already achieved detections through cross correlation with optical galaxy surveys ([Chang et al. 2010](#); [Masui et al. 2013](#); [Anderson et al. 2018](#); [Wolz et al. 2022](#); [Cunnington 2022](#); [Paul et al. 2023](#)). Radio cosmology is also a flagship science driver for the SKA Observatory (SKAO),¹ which, operated in single dish mode, is expected to conduct 21 cm IM surveys across cosmologically relevant scales out to redshift $z \sim 3$ ([Santos et al. 2015](#); [SKA Cosmology SWG 2020](#)); its precursor MeerKAT is already contributing through the MeerKLASS survey ([Santos et al. 2017](#)), with encouraging early results ([Wang et al.](#)

¹<https://www.skao.int/>

2021; Irfan et al. 2022) including a first cross correlation detection with WiggleZ galaxy data (Cunnington et al. 2022).

Alongside several technical efforts mainly focused on foreground modeling and mitigation, refining the forecast capabilities of 21 cm IM, both as a standalone probe and in combination with other cosmological observables, is crucial (e.g. (Carucci et al. 2017; Berti et al. 2023; Berti et al. 2024)). These forecasts are essential not only for strengthening the scientific case for 21 cm IM radio cosmology, but also in order to optimize the design and strategy of upcoming surveys.

1.1.1 The expanding Universe

Cosmology describes the origin, evolution, and large scale structure (LSS) of the Universe as a whole. On the largest scales, the Universe is observed to be homogeneous and isotropic, an assumption known as the cosmological principle. Together with the assumption of a spatially flat geometry, strongly supported by measurements of the cosmic microwave background (CMB) (Planck Collaboration VI 2020), this fixes the spacetime metric to the Friedmann, Robertson, Walker (FRW) form,

$$ds^2 = -c^2 dt^2 + a^2(t) d\mathbf{r}^2, \quad (1.1)$$

where $a(t)$ is the scale factor describing the relative ratio of sizes of the Universe as a function of cosmic time t , conventionally normalized to $a = 1$ today. Inserting this metric into the Einstein field equations relates the expansion rate to the energy content of the Universe, giving the Friedmann equation,

$$H(t) \equiv \frac{1}{a} \frac{da}{dt} = H_0 \sqrt{\Omega_m a^{-3} + \Omega_\Lambda}, \quad (1.2)$$

where $H_0 \equiv H(a = 1)$ is the Hubble constant, Ω_m is the matter density parameter (dark matter plus baryons), and Ω_Λ is the dark energy density parameter, with $\Omega_m + \Omega_\Lambda = 1$ for a flat Universe (radiation is a negligible contribution at the late times relevant to this thesis, and is dropped here for clarity). At early times, when $\Omega_m a^{-3} \gg \Omega_\Lambda$, the Universe is matter dominated and $a(t) \propto t^{2/3}$; at late times, as Ω_Λ comes to dominate, the expansion instead accelerates towards $a(t) \propto e^{H_0 t}$. Current measurements give $\Omega_m = 0.315 \pm 0.007$ (Planck Collaboration VI 2020), with late time probes broadly consistent with this value, though not without some persistent tension.

1.1.2 Growth of structure

The Universe is not perfectly homogeneous: small primordial density perturbations, seeded during an early inflationary epoch, grow under gravitational instability into the large scale structure observed today. It is convenient to describe these perturbations through the fractional matter overdensity,

$$\delta(\mathbf{r}, t) \equiv \frac{\rho(\mathbf{r}, t) - \langle \rho(\mathbf{r}, t) \rangle}{\langle \rho(\mathbf{r}, t) \rangle}, \quad (1.3)$$

where $\langle \dots \rangle$ denotes an ensemble average. At times well after recombination and on sub horizon scales, δ obeys, in Fourier space,

$$\frac{d}{dt} \left(a^2(t) \frac{d}{dt} \delta(\mathbf{k}, t) \right) = 4\pi G a^2(t) \rho(t) \delta(\mathbf{k}, t), \quad (1.4)$$

sourced by the Poisson equation relating the gravitational potential Φ to δ . Because this equation is linear in δ , its solution factorizes into a part depending only on time and a part depending only on scale,

$$\delta(\mathbf{k}, a) = \frac{2k^2}{5\Omega_m H_0^2} \mathcal{R}(\mathbf{k}) T(k) D_+(a), \quad (1.5)$$

where $\mathcal{R}(\mathbf{k})$ encodes the primordial (inflationary) fluctuations, $T(k)$ is the transfer function describing how different scales evolve differently around the epoch of matter-radiation equality, and $D_+(a)$ is the linear growth factor. The growth factor satisfies $D_+(a) \propto a$ during matter domination, and is suppressed at late times as dark energy domination takes over and accelerated expansion begins to suppress further collapse;

1.1.3 The power spectrum and Gaussianity

The statistical properties of the density field are most commonly characterized through the (linear) matter power spectrum $P(k)$, defined via the Fourier transform of two point correlation function,

$$(2\pi)^3 P(k) \delta_D^3(\mathbf{k} + \mathbf{k}') = \langle \delta(\mathbf{k}) \delta(\mathbf{k}') \rangle, \quad (1.6)$$

with δ_D the Dirac delta function. The factorized form of $\delta(\mathbf{k}, a)$ above implies $P_{\text{Lin}}(k, a) \propto T^2(k) D_+^2(a)$. A key property of the *linear* density field is that, to very high precision, it is Gaussian: this is a direct consequence of the (approximately Gaussian) initial conditions

from inflation, preserved under linear gravitational evolution. For a Gaussian field, the power spectrum is a *complete* statistical description, every higher order correlator $\langle \delta(\mathbf{k}_1) \cdots \delta(\mathbf{k}_n) \rangle$ either vanishes (for odd n) or reduces to a sum of products of two point functions (for even n). The overall amplitude of the (linear) power spectrum is conventionally summarized by the parameter σ_8 , the root mean square overdensity in spheres of radius $8 h^{-1} \text{ Mpc}$,

$$\sigma_8^2(a) = \frac{1}{2\pi^2} \int_0^\infty dk k^2 P_{\text{Lin}}(k, a) W^2(kR), \quad R = 8 h^{-1} \text{ Mpc}, \quad (1.7)$$

with W a top-hat filter in Fourier space. Together, Ω_m and σ_8 are two of the most fundamental parameters describing the amount and clustering strength of matter in the Universe, and much of observational cosmology, including the probes discussed later in this Chapter, is aimed at constraining them.

1.1.4 The nonlinear scale

The Gaussian, linear picture above is only valid while density fluctuations remain small, $|\delta| \ll 1$. This breaks down once fluctuations become of order unity, defining a nonlinear wavenumber k_{NL} through the dimensionless power spectrum,

$$\Delta^2(k_{\text{NL}}) \equiv \frac{k_{\text{NL}}^3}{2\pi^2} P_{\text{Lin}}(k_{\text{NL}}) \sim 1, \quad (1.8)$$

which for present-day ΛCDM parameters occurs at $k_{\text{NL}} \sim 0.2 h/\text{Mpc}$ ([Takahashi et al. 2012a](#)). Beyond this scale, gravitational collapse drives the density field away from Gaussianity, producing the cosmic web of halos, filaments, and voids observed today; this is precisely the regime that the field level machine learning methods developed in this thesis are designed to model.

1.1.5 The Jeans scale

It is worth being careful to distinguish k_{NL} , which marks where *gravitational* clustering of matter becomes nonlinear, from a second, physically distinct scale relevant once baryons (ordinary gas) are considered. This is the Jeans scale, which marks where gas pressure, rather than gravity, decides whether a region can collapse: baryonic structure and compact-object formation is a tug of war between the thermal/kinetic pressure of the gas and the gravitational pull compressing it.

A simple, order of magnitude way to see where this scale comes from is to compare two timescales for a gas cloud of size λ and density ρ : the time for it to collapse under its own gravity, and the time for a pressure (sound) wave to cross it and counteract that collapse. The gravitational free fall time scales as

$$t_{\text{gravity}} \sim \frac{1}{\sqrt{G\rho}}, \quad (1.9)$$

while the sound-crossing time is

$$t_{\text{sound}} \sim \frac{\lambda}{c_s}, \quad (1.10)$$

where c_s is the sound speed of the gas. Collapse can proceed only if gravity acts faster than pressure can respond, i.e. $t_{\text{gravity}} \lesssim t_{\text{sound}}$. Setting the two timescales equal defines the Jeans length,

$$\lambda_J \sim \frac{c_s}{\sqrt{G\rho}}, \quad (1.11)$$

or, as a wavenumber, $k_J \sim 2\pi/\lambda_J$. (A more careful derivation, linearizing the fluid equations with pressure included, gives the same scaling with an $\mathcal{O}(1)$ prefactor (Peacock 1998).) Below this scale, thermal pressure support resists collapse and washes out matter structure; above it, self gravity wins and the gas can collapse along with the dark matter.

The Jeans scale and the nonlinear scale k_{NL} therefore answer different questions. One is about gravitational clustering of matter in general (relevant for dark matter, which has no pressure support), the other about whether *gas* specifically can collapse into that clustered structure, and both are relevant once modeling baryonic tracers such as neutral hydrogen. This is one of the reasons, elaborated in Section 1.4, that hydrodynamical simulations (rather than dark-matter-only N -body simulations) are needed to model such tracers realistically.

1.1.6 Beyond linear theory: a brief note on perturbation theory

Extending the description of structure formation beyond the linear regime is, in principle, possible order by order in what is known as cosmological perturbation theory, which introduces higher order corrections to the power spectrum (e.g. at “1-loop” order) and to new statistics such as the bispectrum,

$$(2\pi)^3 B(k_1, k_2, k_3) \delta_{\text{D}}^3(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) = \langle \delta(\mathbf{k}_1) \delta(\mathbf{k}_2) \delta(\mathbf{k}_3) \rangle, \quad (1.12)$$

which vanishes identically for a Gaussian field but carries additional cosmological information once non linearities become significant (Sefusatti et al. 2006). In practice, however, this perturbative approach becomes rapidly more complex at higher orders, and even at next to leading order it already departs from full N -body simulation results by an order of magnitude on the mildly non linear scales relevant to this work (Chudaykin et al. 2020). The large scale structure traced by all of these observables is, in short, the end product of a well understood but analytically intractable evolutionary history: a Gaussian random field of primordial perturbations, evolving under gravity into the highly non Gaussian cosmic web of halos, filaments, and voids described above (e.g. Obuljen et al. 2023; Villaescusa-Navarro et al. 2018a). Once gravitational (and, for baryonic tracers, astrophysical) non linearities become important, the power spectrum alone is no longer a complete statistical description: information leaks into higher order correlations, precisely on the mildly non linear and non linear scales that are often the most information rich, and standard perturbative techniques break down exactly there. Restricting analyses to the two point function alone therefore discards a substantial fraction of the cosmological information encoded in the field, particularly on the small scales probed by upcoming high resolution 21 cm IM and galaxy surveys. This breakdown of analytic methods, rather than any detailed treatment of perturbation theory itself, which is not the focus of this thesis, is the key motivation for turning to simulations, and ultimately to the machine learning based, field level approaches developed in later Chapters.

This has motivated a shift, over the last several years, toward *field level* inference and emulation. Rather than compressing a cosmological field into a small set of hand chosen summary statistics before performing inference, one instead works directly with the field itself (or a learned representation of it), in principle retaining the full non Gaussian information content (e.g. Jamieson et al. 2023; Angulo & Hahn 2022). An example in this direction is the work of Rouhiainen (2024), which explicitly frames the non Gaussianity of the late-time, small scale LSS as the central obstacle to analytic modeling, and addresses it by learning the field level probability density of the matter distribution directly with normalizing flows, explicitly analogous to how gravitational evolution transforms an initially Gaussian field into today’s complex non Gaussian one, before turning to denoising diffusion models for volumetric, 3D super resolution tasks.

1.1.7 Large scale structure for cosmological parameter inference

The physics above directly shapes how the LSS is used in practice to constrain cosmological parameters such as Ω_m and σ_8 . The traditional approach compresses an observed field (a galaxy catalog, a weak lensing map, an intensity map) into a small set of hand-chosen summary statistics, most commonly the power spectrum, and increasingly the bispectrum, and performs likelihood based (or Gaussian likelihood) inference on the resulting data vector, using perturbation theory or fitting functions such as HaloFit (Takahashi et al. 2012b) to model the observable as a function of $\phi = (\Omega_m, \sigma_8, \dots)$. In practice, this traditional likelihood is not evaluated on a grid but explored with Markov Chain Monte Carlo (MCMC) methods, most commonly a Metropolis-Hastings sampler (Metropolis et al. 1953; Hastings 1970): propose a step in parameter space, accept or reject it with a probability set by the ratio of posterior densities, and repeat, so that the resulting chain of accepted points is, after enough steps, distributed according to the posterior itself. This has been the golden standard of cosmological parameter estimation for decades, and remains the standard tool whenever the likelihood is tractable.

Even before the deep learning era, however, cosmologists occasionally faced likelihoods too complex to write down explicitly, for instance when systematics or selection effects have no simple analytic form. *Approximate Bayesian Computation* (ABC) was the pre deep learning answer to exactly this problem: rather than evaluating a likelihood at all, one simulates mock data for many candidate parameter values and retains those simulations close enough to the actual observation under some chosen distance metric, building an approximate posterior from the accepted simulations alone (Weyant et al. 2013; Akeret et al. 2015). This is, in essence, simulation based inference without a neural network. This is the same idea of replacing an explicit likelihood with simulations, but relying on nearest neighbor acceptance/rejection rather than a learned density estimator. This becomes rapidly inefficient as the dimensionality of the data grows, precisely the bottleneck that neural simulation-based methods were developed to overcome.

These theoretical and analytical approaches to model LSS are well understood and computationally cheap, but, as argued above, it is information lossy by construction on the mildly non linear and non linear scales where perturbation theory itself begins to fail (e.g. Chudaykin et al. 2020).

Two developments over roughly the last decade have pushed beyond this. First, *simulation based inference* (SBI, also called likelihood-free inference) sidesteps the

need for an analytic likelihood altogether. Cosmological simulations are run over a grid of parameters ϕ , and a learned surrogate (e.g. a neural density estimator) is trained on the resulting $(\mathbf{x}, \phi)$ pairs to approximate the posterior $p(\phi|\mathbf{x})$ directly (e.g. Jeffrey et al. 2020; Angulo & Hahn 2022). Large public simulation suites such as CAMELS (Villaescusa-Navarro et al. 2021b) were built explicitly to enable this kind of approach (see Section 1.4), and its associated Multifield Dataset (Villaescusa-Navarro et al. 2022) has become a standard benchmark for exactly this task.

Second, and building on SBI, *field level inference* removes the summary-statistic compression step entirely, training models to infer ϕ (or a full field level posterior) directly from the observed field itself, in principle saturating the available cosmological information rather than only what survives compression into a power spectrum, bispectrum, or learned low dimensional summary (e.g. Jamieson et al. 2023).

Crucially, these methods are no longer confined to simulated forecasts. Field level and simulation based inference have already been applied directly to real observational data: a field level SBI analysis of the actual BOSS CMASS galaxy sample constrains σ_8 to a precision approximately 2.65 times tighter than an equivalent power-spectrum-based analysis of the same data (Lemos et al. 2023), while a simulation-based, neural compression analysis of real Dark Energy Survey Year 3 weak lensing data constrains the dark energy equation of state parameter w (under a w CDM model) more than twice as precisely as an equivalent power-spectrum-only analysis (Jeffrey et al. 2024). This demonstrates that simulation based and field level inference have moved from a promising forecasting tool to a method already delivering competitive cosmological constraints from real surveys.

1.1.8 Thesis Outline

This thesis adopts the same field level philosophy, but applied to a different pair of problems. Rather than only performing parameter inference from an observed field, we are also concerned with learning the field level *mapping between cosmological tracers* (dark matter, halos, 21 cm brightness temperature, baryonic fields) directly from simulations. The two goals are complementary: an accurate field level emulator of the kind developed in this thesis is itself a natural forward simulator for the SBI/field level inference pipelines described above, since it can generate the large number of realistic, high-resolution training pairs that such methods require far more cheaply than full hydrodynamical simulations. Taken together, these contributions, spanning diffusion-based generation

and reconstruction, and emulation-based modeling of cosmological and astrophysical fields, are built around the central idea of extracting the information that is lost when a field is compressed into a summary statistic, using generative models trained directly on simulations.

We first lay out the foundation of the background that the thesis is built around, starting from the basics of Bayesian inference in Section 1.2. Since the thesis uses machine learning in all the chapters, we introduce some concepts on convolution based machine learning in section 1.3. Next, we briefly go over cosmological simulations in section 1.4, as they are used to train our models. Section 1.5, lays out briefly the theoretical idea behind the 21 cm intensity mapping and how the HI field is modeled in different ways. This is followed by a overview of cosmological surveys and experiments, in section 1.6, where we go through both galaxy based surveys and 21 cm mapping experiments.

Chapter 2 (LODI) develops a novel, physically motivated pipeline for predicting the 21 cm signal at the field level. A U-Net first maps dark matter to halos, then a custom diffusion model paints the 21 cm brightness temperature field on top, down to small non linear scales, with a latent-overlap scheme to scale the generation up to volumes larger than the network was trained on. This is the first demonstration of a diffusion based, halo conditioned approach to 21 cm IM, complementary to the forward modeling methods reviewed in Section 1.5.4, predicting halos and the 21 cm signal jointly is what enables the cross correlation studies with other LSS tracers. Last but not the least, we motivate the existence of this thesis by explaining how, for parameter inference, a substantial bottleneck is the expensive simulations, and that we can go around them to achieve our purpose.

Chapter 3 (Cosmo-FOLD “Cosmological Field Overlap Latent Diffusion”) extends the same diffusion framework with a novel general stitching algorithm, to generate large volume fields, tackling the challenge of scaling generative models to high fidelity cosmological volumes under GPU-memory limits and the intrinsic non locality of gravitational effects. It generates gas temperature fields from dark matter, and dark matter fields from sparse stellar density, on the IllustrisTNG TNG300 volume, together with a positional encoding scheme aimed at improving the recovered bispectrum. Trained on $\leq 1\%$ (and validated on $\approx 0.5\%$) of the volume, it reproduces the 3-point statistics well, naturally enforces periodic boundary conditions, and eliminates the edge artifacts

of previous overlapping patch approaches even under highly sparse stellar input.

Chapter 4 showcases a proof of concept implementation of **Cosmo-FOLD** for field level posterior sampling of dark matter fields using constraints from data. Whereas Chapter 5 takes a different route and develops **FOREST**, a general framework of diffusion posterior sampling for inverse problems and applies it for reconstructing dark matter and baryonic fields from noisy or incomplete data along a line of sight from an observed Lyman α forest spectrum.

Chapter 6 moves from a single tracer to a multi tracer setting, using a conditional U Net emulator to generate galaxy and HI fields from dark matter, and combining them in a simulation based inference pipeline to show what is gained, in terms of constraining power on Ω_m and σ_8 , by moving from power spectrum based inference to full 3D field level inference, and from a single tracer to two combined.

We conclude in Chapter 7 with a discussion of the common thread running through these chapters and promising avenues for future work, including applications to cross correlation studies between 21 cm IM and other LSS tracers, and to more realistic, survey like data.

1.2 Bayesian Inference Basics

Cosmology is, in a specific sense, harder than most experimental sciences: we only ever get to observe one Universe, at one set of times, from one vantage point. There is no way to re run the experiment with a different Ω_m or σ_8 and see what comes out (hence simulations). This makes the Bayesian interpretation of probability, as a degree of belief, updated with respect to data, rather than a long run frequency of repeated trials, the natural framework for cosmological inference, and it is the language used throughout the following Chapters.

Bayes' theorem itself follows from a basic property of joint probabilities, $p(\theta, \mathbf{d}) = p(\mathbf{d}|\theta) p(\theta) = p(\theta|\mathbf{d}) p(\mathbf{d})$, rearranged to read

$$p(\theta | \mathbf{d}) = \frac{p(\mathbf{d} | \theta) p(\theta)}{p(\mathbf{d})}, \quad (1.13)$$

where θ denotes the parameters of interest (e.g. Ω_m, σ_8) and $\mathbf{d}$ the observed data. Each

term has a name and a role:

- $p(\theta)$, the *prior*, encodes what is known (or assumed) about θ before seeing the data at hand, e.g. a previous CMB constraint, or simply a broad, uninformative range.
- $p(\mathbf{d}|\theta)$, the *likelihood*, is the forward model: given a value of θ , how probable is it to observe this particular data $\mathbf{d}$? This is where a physical model of the observable enters, whether that model is an explicit Gaussian likelihood on a summary statistic, or (as in Chapter 6) an implicit one defined only through forward simulations.
- $p(\theta|\mathbf{d})$, the *posterior*, is the object of interest: the updated belief about θ once the data $\mathbf{d}$ has been taken into account, combining prior knowledge with what the data itself says.
- $p(\mathbf{d})$, the *evidence*, is a normalization constant (the probability of the data, marginalized over all possible θ) that does not need to be computed for parameter inference, but becomes important when comparing two different models rather than two values of the same parameter.

The whole point of an inference pipeline, in this language, is to get from $p(\theta)$ to $p(\theta|\mathbf{d})$: to turn a prior belief into a posterior one, using the data through the likelihood.

1.3 Machine Learning for Field Level Cosmology

In recent years, machine learning techniques, and in particular generative models, have emerged as powerful emulators to reproduce cosmological and astrophysical fields (Peraudin et al. 2019; Li et al. 2021; Jamieson et al. 2023; Rigo et al. 2025; Hassan et al. 2022; Andrianomena et al. 2022; Sharma et al. 2025; Bernardini et al. 2025; Andrianomena et al. 2024). Among them, *diffusion models* have emerged as the current state of the art for high fidelity generative modeling, and are the central tool used throughout this thesis. Given their importance to everything that follows, we introduce them here in some detail, together with the surrounding building blocks, convolutional networks, attention, normalizing flows, and simulation-based inference, that the rest of the thesis assumes some familiarity with.

1.3.1 Convolutional neural networks and the U-Net

Most of the fields this thesis works with, some of which are dark matter density, HI brightness temperature, gas temperature and HI ionization fraction, are naturally represented as 2D or 3D grids of voxel values, and the mappings we want to learn (dark matter to halos, dark matter to gas temperature, and so on) is local in a specific sense: the value of the target field at a given voxel to a good approximation depends on the source field in a neighborhood around that voxel, not on the field far away. A *convolutional neural network* (CNN) is built around exactly this assumption. Rather than connecting every input voxel to every output voxel with an independent weight, as a generic fully-connected network would, a convolutional layer instead slides a small, learnable filter, or *kernel*, typically a few voxels across, over the input field, computing a weighted sum of the local neighborhood at every position (see the bottom panel of Fig. 1.1). Formally, for a single channel input x and a $K \times K$ kernel w with bias b , the value of the output feature map at position (i, j) is

$$y_{i,j} = \sigma \left(\sum_{m=1}^K \sum_{n=1}^K w_{m,n} x_{i+m,j+n} + b \right), \quad (1.14)$$

where σ is a non linear activation function, discussed below. For an input of spatial size $N \times N$, kernel size K , stride s , and zero-padding p , the resulting output has size

$$N_{\text{out}} = \left\lfloor \frac{N - K + 2p}{s} \right\rfloor + 1, \quad (1.15)$$

so that, e.g., a 7×7 input convolved with a 3×3 kernel at stride 1 and no padding yields a 5×5 output, as shown in Fig. 1.1. The same filter weights $w_{m,n}$ are reused at every position (“weight sharing”), which both drastically reduces the number of parameters to learn and builds in translation invariance: a feature the network learns to detect (e.g. a halo-scale overdensity) is detected the same way wherever it appears in the box.

The non linear activation σ applied after each convolution is what allows the network to represent non linear relationships; throughout this thesis it is almost always either a rectified linear unit (ReLU),

$$\text{ReLU}(x) = \max(0, x), \quad (1.16)$$

or a Sigmoid Linear Unit (SiLU)

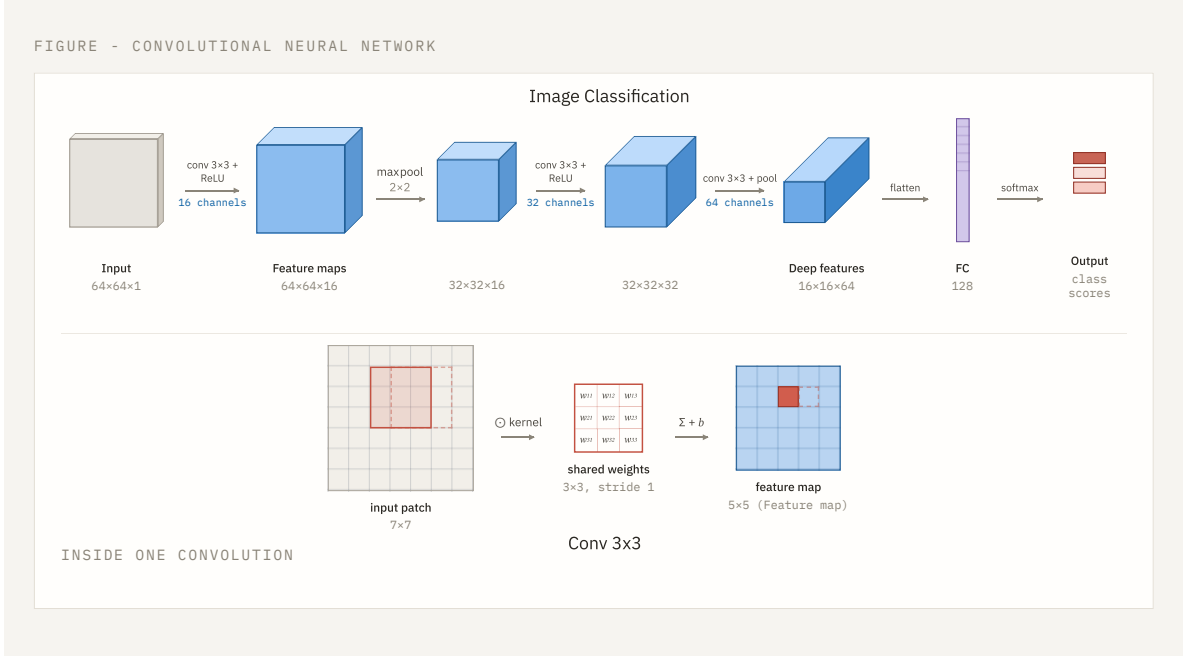


Figure 1.1: Top: a canonical convolutional network, where the input is progressively transformed through convolution, non linear activation, and pooling layers into a compact set of deep features, before being combined by a final fully-connected layer, here shown for an image classification task, though the same building blocks apply equally to the field to field regression networks used throughout this thesis. Bottom: detail of a single 3×3 convolutional kernel W , acting on an input of shape 7×7 to produce a 5×5 output (dashed vs. solid outline indicating adjacent kernel positions at stride 1).

$$\text{SiLU}(x) = x \cdot \sigma(x) = \frac{x}{1 + e^{-x}}, \quad (1.17)$$

Stacking several convolutional layers, usually interleaved with such an activation and occasionally a downsampling (pooling) step, lets the network build up increasingly large effective receptive fields and increasingly abstract features from small scale texture in the first layers to large-scale structure in the deeper ones. For max-pooling with a 2×2 window and stride 2, as shown in the top panel of Fig. 1.1,

$$y_{i,j} = \max_{(m,n) \in \{0,1\}^2} x_{2i+m, 2j+n}, \quad (1.18)$$

which halves the spatial resolution while retaining the largest value in each 2×2 block. In a classification setting (for example, identifying handwritten digits from 0 to 9) the final feature maps would be flattened and passed through a fully connected layer

and a softmax² to assign class probabilities, as in Fig. 1.1. However the field to field networks used in this thesis instead keep a spatial output at every stage, as described next.

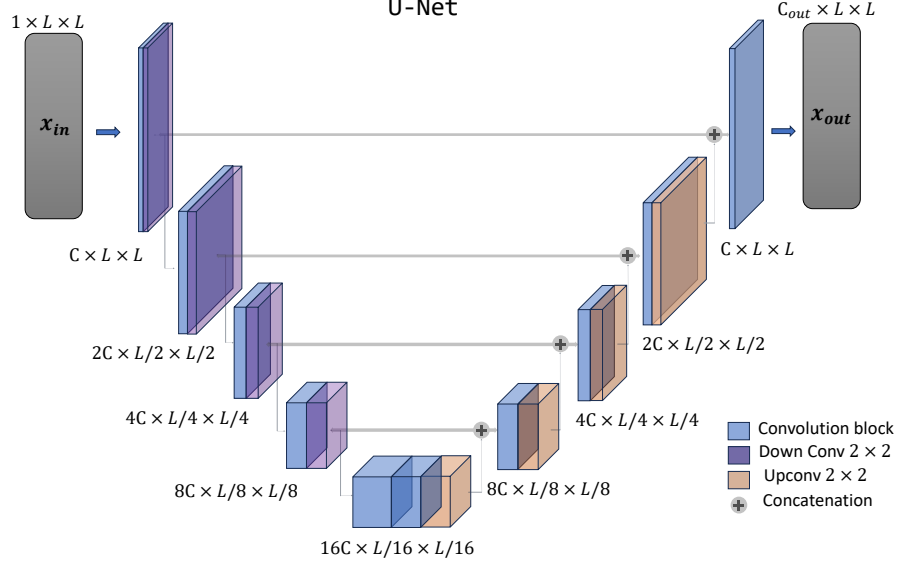


Figure 1.2: The network architecture of a standard U-net. This particular Unet works with 2D images. The dimension of the latent space at each level is mentioned, with C being the number of feature channels learnt by the Unet.

A *U-Net* (Ronneberger et al. 2015) (See Fig 1.2) is a specific, and by now very standard, way of arranging convolutional layers for tasks where the input and output are both fields of the same kind (an “image-to image”, or here “field to field”, translation task): an *encoder* path of convolutional layers progressively downsamples the input into a coarser, more abstract representation (the left arm of the network), a *decoder* path then progressively upsamples this representation back up to the original resolution, and *skip connections* directly link each encoder layer to its corresponding decoder layer at the same resolution. In the specific image we show, the Unet is designed to work with 2

²The softmax function $\sigma(\mathbf{z})_i = e^{z_i} / \sum_j e^{z_j}$ maps a vector of real valued outputs to a normalized probability distribution over classes, where z_i is the un normalized value for class i .

dimensional images. Each convolution block generally consists of 2 convolution layers, with a convolution layer depicted in the bottom image of Fig 1.1. Each subsequent layer reduces the dimension of the data while adding more channels to them. Each channel represents an extracted feature of the data, that is learnt by the network. The skip connections are what make the U-Net well suited to field level cosmology specifically: without them, the fine, small scale detail present in the input (which the encoder discards on its way to the coarse bottleneck representation) would simply be lost, and the decoder would have to reconstruct it from scratch; with them, that detail is carried directly across the network and only needs to be combined with the new information the network has learned, rather than regenerated. Every network used to map between fields in this thesis, the dark matter to halo network of Chapter 2, the denoising networks underlying the diffusion models of Chapters 2 and 3, and the galaxy/HI emulator of Chapter 6 is a variant of this same U-Net design.

Within each stage of a more modern U-Net, and more generally, in most sufficiently deep convolutional networks, individual convolutional layers are typically grouped into *residual blocks* (He et al. 2015) rather than stacked directly. A residual block (Fig. 2.5) computes some transformation $g(\mathbf{x})$ of its input (e.g. two convolutional layers with normalization and a non linearity in between) and then adds the original input back, $\mathbf{x} + g(\mathbf{x})$, rather than outputting $g(\mathbf{x})$ alone. This is a different kind of skip connection from the encoder to decoder ones described above, a much shorter range one, spanning only a single block rather than the whole network, but it serves a related purpose. It lets the network default to the identity mapping if a given block turns out not to be useful, which in practice makes much deeper networks trainable, since gradients can flow backward through the addition even if the learned transformation g itself is poorly conditioned. Every U-Net used in this thesis is built from stacks of such residual blocks, not from bare convolutional layers.

These networks are also, almost without exception, not unconditional: they need to be told about additional information beyond the input field itself, a cosmological parameter, an astrophysical feedback strength parameter, a diffusion timestep, or several of these at once. The simplest way to do this is to concatenate the conditioning information as extra input channels, which is what the diffusion models of Chapters 2 and 3 do (Section 1.3.6). A more flexible alternative, used for the multi parameter conditioning of Chapter 6, is *feature-wise linear modulation* (FiLM) (Perez et al. 2017): rather than adding the conditioning information as an input, a small auxiliary network maps it onto a pair of vectors $\boldsymbol{\gamma}, \boldsymbol{\beta}$, one value per feature channel, which then rescale

and shift every channel of an intermediate feature map,

$$\text{FiLM}(\mathbf{x}, \mathbf{c}) = \gamma(\mathbf{c}) \odot \mathbf{x} + \beta(\mathbf{c}). \quad (1.19)$$

Because this modulation is applied at every layer rather than only at the input, it gives the conditioning information many more opportunities to influence the network’s computation, which matters when, as in Chapter 6, the network needs to generalize across a continuous range of several conditioning parameters at once rather than being retrained for each one. More details and a diagram can be seen in section 6.4

This kind of convolutional field to parameter mapping has already been benchmarked directly on the CAMELS Multifield Dataset introduced in Section 1.1.7: convolutional networks trained on CAMELS multifield maps can already extract Ω_m and σ_8 while marginalizing over four independent supernova and AGN feedback parameters (Villaescusa-Navarro et al. 2021a), and graph neural networks operating directly on galaxy or halo point clouds achieve comparable results, inferring Ω_m to within roughly 12% from as few as ~ 1000 galaxy positions in a single small simulation volume (Villanueva-Domingo & Villaescusa-Navarro 2022).

1.3.2 Attention

A convolutional layer, by construction, only looks at a small local neighborhood at a time; a feature at one corner of the box can only influence a feature at the opposite corner once information has passed through enough stacked convolutional layers to connect the two, which can require many layers and does not always happen efficiently in practice. This is a real limitation for cosmological fields, where long range effects, such as large scale gravitational tides, or feedback from a single active galactic nucleus displacing gas well beyond its host halo cannot be captured by a network that can only combine nearby voxels.

Self attention (Vaswani et al. 2023) addresses this by letting every position in a feature map directly interact with every other position, regardless of distance (Fig. 1.3). Concretely, given a set of n input token (voxel) embeddings stacked into a matrix $X \in \mathbb{R}^{n \times d}$, three learned linear projections map X onto a set of *queries*, *keys*, and *values*,

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V, \quad (1.20)$$

with $Q, K \in \mathbb{R}^{n \times d_k}$ and $V \in \mathbb{R}^{n \times d_v}$. The similarity between every pair of positions

is then obtained from the dot product of their query and key vectors, $QK^\top \in \mathbb{R}^{n \times n}$, rescaled by $\sqrt{d_k}$ to keep the resulting values in a numerically well-behaved range, and passed through a row-wise softmax so that the weights attributed by any one position to all others sum to one,

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \quad (1.21)$$

giving an $n \times n$ matrix of *attention weights*. The final, context-aware output is a weighted sum of all positions’ value vectors, using these weights,

$$Z = AV = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (1.22)$$

with $Z \in \mathbb{R}^{n \times d_v}$. In effect, every voxel gets to “ask” the rest of the field, through its query, which other voxels’ keys are most relevant to it, and then assembles its output from exactly those voxels’ values, rather than being restricted to a fixed local window, as a convolution is. This mechanism is also the backbone of large Language Models (LLMs).

This comes at a computational cost though, that grows quadratically with the number of voxels, since forming $QK^\top$ requires comparing every position with every other one: for n voxels the attention matrix A has n^2 entries, so doubling the resolution of a 3D field increases the cost of a single attention layer by a factor of 64. This is why, throughout this thesis, self attention is used only at the coarsest resolution inside a U-Net’s bottleneck, where there are relatively few voxels left to compare, rather than at full field resolution.

1.3.3 Normalizing flows

Several places in this thesis need a way to represent a complicated, high dimensional probability distribution, the posterior over cosmological parameters given some data, for instance, in a form that is both flexible enough to capture non Gaussian structure and simple enough to sample from and evaluate a likelihood for. A *normalizing flow* is one way to build exactly such an object: start from a simple, tractable base distribution $p_0(\mathbf{z})$ (typically a standard multivariate Gaussian), and pass a sample $\mathbf{z} \sim p_0$ through an invertible, differentiable transformation f_θ to obtain a sample $\mathbf{x} = f_\theta(\mathbf{z})$ from the target distribution. Because f_θ is invertible, the change of variables formula gives an

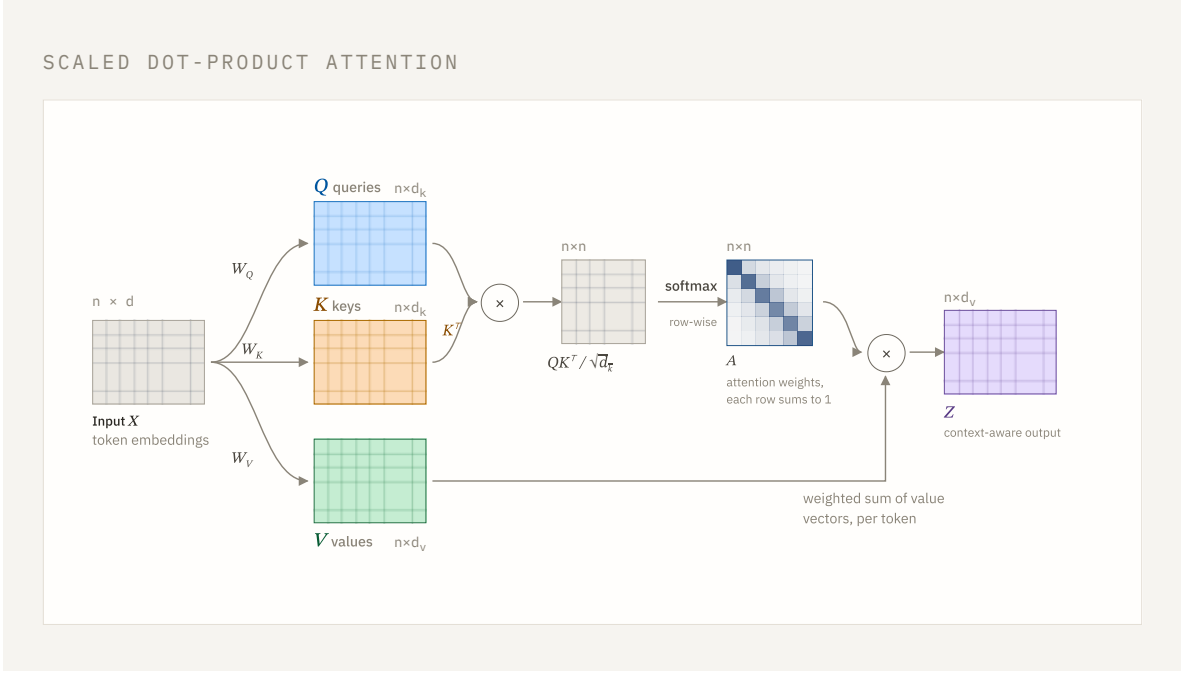


Figure 1.3: Scaled dot-product self attention. The input token (voxel) embeddings X are linearly projected onto queries Q , keys K , and values V (Eq. 1.20). The similarity QK^T between every pair of positions is rescaled by $\sqrt{d_k}$ and normalized row-wise with a softmax to give the attention weights A , each row of which sums to one (Eq. 1.21). The output Z is obtained by using A to take a weighted sum of the value vectors for every position (Eq. 1.22).

exact expression for the density of $\mathbf{x}$,

$$p_{\theta}(\mathbf{x}) = p_0(f_{\theta}^{-1}(\mathbf{x})) \left| \det \frac{\partial f_{\theta}^{-1}}{\partial \mathbf{x}} \right|, \quad (1.23)$$

in terms of the (tractable) base density and the Jacobian determinant of the transformation. A single transformation of this kind is rarely expressive enough on its own, so in practice several are chained together, $\mathbf{x} = f_K \circ \dots \circ f_1(\mathbf{z})$, each one designed so that its Jacobian determinant remains cheap to compute despite the added flexibility. Training then amounts to maximizing the log likelihood of the resulting density on a set of samples $\{\mathbf{x}_i\}$ directly. A *Masked Autoregressive Flow* (MAF) (Papamakarios et al. 2018), used in Chapter 6 to model a posterior over cosmological parameters, is one widely used instance of this construction, in which each transform is autoregressive: every output dimension is a function of the base variable at that dimension and of all previously generated dimensions, which keeps the Jacobian triangular and hence trivial

to invert and to compute a determinant for.

Beyond density estimation for parameter inference, normalizing flows have also been used to directly model the field level probability density of cosmological fields themselves: Rouhiainen et al. (2021) learns an explicit, tractable representation of the field level probability density, conditional on cosmological parameters, from which both parameter posteriors and new field realizations can be sampled, while a wavelet-based multiscale flow similarly models the full field level likelihood of weak-lensing maps and can, as a byproduct, flag distribution shifts caused by unmodeled effects such as baryonic feedback (Dai & Seljak 2024).

1.3.4 Simulation-based inference and neural posterior estimation

Section 1.1.7 already introduced the basic idea of simulation-based inference (SBI): rather than writing down an explicit likelihood $p(\mathbf{d}|\theta)$, generate a large number of simulated (parameter, data) pairs $\{(\theta_i, \mathbf{d}_i)\}$ and learn the relationship between the two directly. *Neural posterior estimation* (NPE) is the specific version of this idea used in Chapter 6: a normalizing flow $q_\phi(\theta | \mathbf{d})$ (Section 1.3.3), conditioned on the data $\mathbf{d}$, is trained to directly approximate the posterior $p(\theta|\mathbf{d})$, by maximizing the log probability the flow assigns to the true parameters θ_i given the corresponding simulated data $\mathbf{d}_i$, averaged over the simulated training set,

$$\max_{\phi} \frac{1}{N} \sum_{i=1}^N \log q_\phi(\theta_i | \mathbf{d}_i). \quad (1.24)$$

Because the flow is conditioned on $\mathbf{d}$ rather than trained for one fixed dataset, the same trained network can be evaluated on any new data $\mathbf{d}$ (simulated or, in principle, real) to obtain an approximate posterior $q_\phi(\theta|\mathbf{d}) \approx p(\theta|\mathbf{d})$ in a single forward pass, with no additional sampling or optimization required at inference time, unlike traditional likelihood-based approaches, which typically need to re-run an MCMC (or similar) sampler for every new dataset. This amortized cost is what makes NPE practical for the kind of large validation and test sets (hundreds of independent realizations) used in Chapter 6, where evaluating the posterior for each one individually via classical sampling would be far more expensive.

NPE is only one of several simulation-efficient strategies for building the neural density estimator itself, and a substantial methodological literature exists comparing them in

a cosmological context. Sequential neural posterior estimation trains a conditional density estimator (typically a normalizing flow) directly on the posterior, adaptively concentrating simulations in the region of parameter space near the observed data rather than sampling the full prior uniformly (e.g. [Greenberg et al. 2019](#)); neural ratio estimation instead trains a classifier to distinguish jointly from independently sampled parameter-data pairs, which in its truncated, marginal form estimates only the low dimensional marginal posteriors actually needed and has been shown to scale efficiently to the many nuisance parameter regime typical of CMB and other cosmological analyses, with a simulation budget that is close to independent of the number of nuisance parameters marginalized over ([Miller et al. 2021](#); [Cole et al. 2022](#)). This idea of truncated marginal neural ratio estimation has also been used in supernova cosmology and parameter inference [Karchev et al. \(2022\)](#). Software packages such as `pydelfi` ([Alsing et al. 2019](#)) combine this kind of density estimation with active learning, adaptively proposing new simulations where the current posterior estimate is least certain, and have demonstrated full posterior inference on real cosmological data (Type Ia supernovae, tomographic cosmic shear) with as few as $\mathcal{O}(10^3, 10^4)$ simulations.

A complementary strategy, often combined with the above rather than competing with it, is to first *compress* the data down to a small number of informative summary statistics before any density estimation is attempted, reducing the dimensionality that the density estimator itself has to handle. *Information Maximizing Neural Networks* (IMNNs) ([Charnock et al. 2018](#); [Makinen et al. 2022](#)) train a network to find the non linear summary of the data that maximizes the retained Fisher information, requiring only simulations generated around a single fiducial parameter value rather than across the full prior range.

1.3.5 Variational autoencoders

A *variational autoencoder* (VAE) ([Kingma & Welling 2022](#)) is a generative latent-variable model that learns to squeeze a high dimensional field into a much smaller *latent* representation and then reconstruct it. It is built from two networks trained jointly: an *encoder*, which maps an input field $\mathbf{x}_0$ to a low dimensional latent $\mathbf{z}$, and a *decoder*, which maps $\mathbf{z}$ back to a reconstruction $\hat{\mathbf{x}}_0$. What makes it *variational*, rather than an ordinary autoencoder, is that the encoder does not output a single latent vector but the parameters of a distribution: for each input it predicts a mean $\boldsymbol{\mu}$ and a standard deviation $\boldsymbol{\sigma}$, and the latent is then drawn from the corresponding Gaussian. To keep

this sampling step differentiable, so the network can still be trained by backpropagation, the latent is written through the *reparameterization trick*,

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (1.25)$$

which pushes all the randomness into $\boldsymbol{\epsilon}$ and leaves $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ on a smooth, differentiable path. Training minimizes two terms together: a reconstruction loss that forces $\hat{\mathbf{x}}_0$ to match $\mathbf{x}_0$, and a *Kullback, Leibler* (KL) regularization term that pulls the encoded latent distribution towards a simple standard-Gaussian prior, keeping the latent space smooth and well behaved. The KL divergence is a standard, if asymmetric, measure of how far one distribution is from another,

$$D_{\text{KL}}(p \parallel q) = \int p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \geq 0, \quad (1.26)$$

vanishing if and only if $p = q$ everywhere. This is related to the *evidence lower bound* (ELBO). For any latent-variable model with an (approximate) encoder $q(\mathbf{z}|\mathbf{x})$, the log marginal likelihood splits exactly as

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}) = \underbrace{\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z}))}_{\text{ELBO}} + D_{\text{KL}}(q(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})), \quad (1.27)$$

where the last term, the gap between the bound and the true log likelihood, is itself a KL divergence and so is never negative. The two pieces of the ELBO are exactly the reconstruction and prior regularization terms above, so training a VAE by minimizing reconstruction error plus the prior KL is precisely maximizing this lower bound on $\log p_{\boldsymbol{\theta}}(\mathbf{x})$. This same ELBO is what the diffusion models introduced next are trained on.

VAEs enter this thesis in **FOREST** of Chapter 5, which uses a VAE of this kind to *compress* the high resolution Lyman- α fields by a factor of ~ 20 before running the diffusion model in that compact latent space rather than on the full voxel grid, an approach known as *latent diffusion*.

1.3.6 An introduction to diffusion models

Diffusion models (Sohl-Dickstein et al. 2015; Ho et al. 2020) are a class of generative model built around a simple idea: it is easy to destroy structure in data by gradually adding noise, and if a neural network can learn to reverse this process step by step,

Diffusion Models

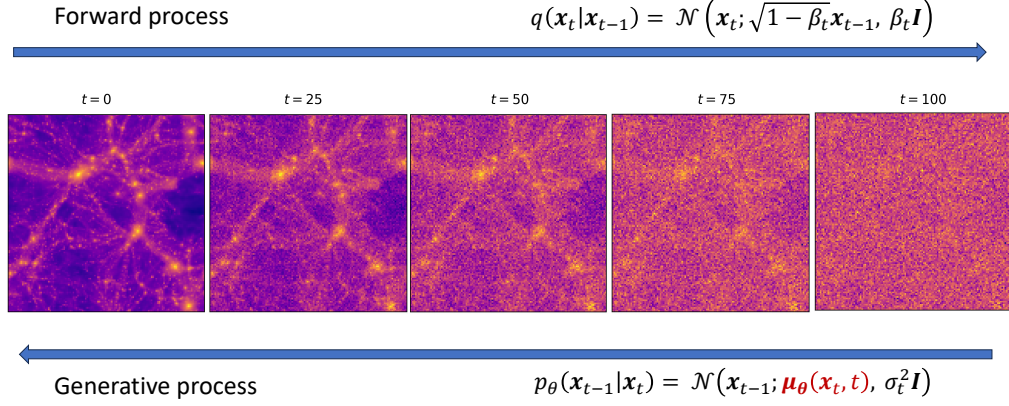


Figure 1.4: A visual depiction of a denoising diffusion model. The top arrow pointing right shows the forward corruption process, while the bottom arrow shows the generative process, with the $\boldsymbol{\mu}_\theta$ in red being the mean of the generative distribution, to be learned and predicted by a denoising network.

it can turn pure noise back into a realistic sample. Concretely, given a data sample $\mathbf{x}_0$ (in our case, a cosmological field such as a dark matter density map), the *forward process* (Figure 1.4) defines a sequence of increasingly noisy versions $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$ by repeatedly mixing in a small amount of Gaussian noise,

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}), \quad (1.28)$$

where $\beta_t \in (0, 1)$ is a (small, pre specified) noise-schedule parameter at step t , and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Because the composition of Gaussians is itself Gaussian, this process has a convenient closed form that lets us jump directly from the clean data $\mathbf{x}_0$ to any noise level t without simulating every intermediate step,

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}), \quad \bar{\alpha}_t \equiv \prod_{s=1}^t (1 - \beta_s). \quad (1.29)$$

For a sufficiently large number of steps T and an appropriately chosen schedule $\{\beta_t\}$, $\bar{\alpha}_T \approx 0$, and $\mathbf{x}_T$ is (to excellent approximation) indistinguishable from pure Gaussian noise, regardless of the original field $\mathbf{x}_0$, all of the original cosmological information has been destroyed.

The generative model is obtained by learning to *reverse* this process: starting from pure noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, a neural network is trained to gradually remove noise, step by step, until a clean sample $\mathbf{x}_0$ emerges. Each reverse step is itself modelled as a Gaussian,

$$p_{\theta}(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_{\theta}(\mathbf{x}_t, t), \sigma_t^2 \mathbf{I}), \quad (1.30)$$

where the mean $\boldsymbol{\mu}_{\theta}$ is predicted by a neural network with parameters θ (in practice, a U-Net, see Chapter 2, that takes the noisy field $\mathbf{x}_t$ and the current noise level t as input).³ Rather than predicting $\boldsymbol{\mu}_{\theta}$ directly, it is simpler to instead train the network to predict the *noise* ϵ that was added to produce $\mathbf{x}_t$ from $\mathbf{x}_0$ in Eq. 1.29, i.e. a network $\epsilon_{\theta}(\mathbf{x}_t, t)$ such that $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$, with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This turns training into a straightforward regression problem: sample a training field $\mathbf{x}_0$, sample a random timestep t and noise ϵ , form the corresponding noisy field $\mathbf{x}_t$, and minimize the simple loss

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\left\| \epsilon - \epsilon_{\theta}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|^2 \right]. \quad (1.31)$$

This loss can be derived rigorously as the variational lower bound on the data log likelihood (Sohl-Dickstein et al. 2015; Ho et al. 2020; Kingma et al. 2023), i.e. diffusion models are, under the hood, latent-variable models trained by maximizing the ELBO, with $\mathbf{x}_1, \dots, \mathbf{x}_T$ playing the role of a sequence of latent variables.

As with the VAE above (Eq. 1.27), maximizing this ELBO amounts to maximizing a lower bound on the data log likelihood, with the (intractable) gap to the true value set by the non-negative KL divergence between the model’s approximate posterior over the latents and the true one. What is special to a diffusion model is that its ELBO decomposes neatly into a sum of KL divergences between the (Gaussian) forward and reverse distributions at each timestep t , which is what the simplified ϵ -prediction loss of Eq. 1.31 is, up to a t -dependent reweighting, actually minimizing.

An equivalent perspective (Song et al. 2021) models the entire forward/reverse process as a stochastic differential equation (SDE), in which the trained network ϵ_{θ} is directly related to the *score* of the noised data distribution, $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t)$; this score-based SDE

³In practice, σ_t^2 is often set to the forward process variance or left as a learnable parameter. For our work, we use σ_t^2 is learned during training (see Section 2.4.2).

viewpoint unifies diffusion models with earlier score-matching approaches.

Once trained, new samples are generated by starting from Gaussian noise $\mathbf{x}_T$ and iteratively applying the learned reverse step for $t = T, T - 1, \dots, 1$, at each step using the network’s noise prediction $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$ to estimate and subtract off the noise, recovering progressively cleaner fields until $\mathbf{x}_0$ is reached. Conditional generation, central to this thesis, e.g. generating a 21 cm brightness temperature field conditional on a dark matter field, or a high resolution field conditional on a low resolution one, is achieved by additionally passing the conditioning field into the network at every step, $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{x}_{\text{cond}})$, so that the reverse process learns to denoise *towards* samples consistent with the given condition rather than towards the unconditional data distribution. More information on how this is achieved is laid out in Chapter 2.

Diffusion models have rapidly become the dominant approach for high fidelity image and video generation (Dhariwal & Nichol 2021), owing to their training stability (compared to, e.g., generative adversarial networks) and their ability to produce diverse, high quality samples. Their extension to volumetric (3D) scientific data, and specifically to cosmological fields, is comparatively recent and is the subject of Chapters 2 and 3 of this thesis.

1.3.7 Diffusion posterior sampling

Everything above describes *generating* a field, either unconditionally or conditioned on another field known exactly. A related but distinct problem, addressed in Chapter 5, is that of *inverse* problems: given a corrupted, incomplete, or otherwise indirect observation $\mathbf{y}$ of an underlying field $\mathbf{x}_0$, related to it through some known (or approximately known) forward process, e.g. noise, masking, a survey selection function, or a full physical forward model, how does one recover the posterior $p(\mathbf{x}_0|\mathbf{y})$ over the fields that could have produced it?

Diffusion posterior sampling (DPS) (Chung et al. 2023) answers this without training a new, problem-specific network for every choice of observation. The starting point is that the reverse process of Section 1.3.6 is driven by the *score* of the noised data distribution, $\nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t)$, which the trained network supplies through its noise prediction $\boldsymbol{\epsilon}_\theta$. That is the score of the *prior*, exactly right for unconditional generation but not for sampling conditioned on an observation the network never saw. What we want instead, at every reverse step, is the score of the noised *posterior*, which Bayes’ theorem (Section 1.2)

splits into two additive pieces,

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{y}) = \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)}_{\text{prior score (the trained network)}} + \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t)}_{\text{likelihood guidance}}. \quad (1.32)$$

The first term is exactly what the unconditional, pretrained diffusion model already provides, with no retraining. The second, the *guidance* term, is the only place the observation $\mathbf{y}$ and the forward model enter, and it is what turns a single unconditional prior into a sampler for a specific inverse problem.

The obstacle is that this guidance term is not directly available. A forward model relates $\mathbf{y}$ to the *clean* field $\mathbf{x}_0$, not to the noisy intermediate $\mathbf{x}_t$ that the reverse process actually operates on. The exact link follows from marginalizing over all clean fields consistent with the current noisy state,

$$p(\mathbf{y} | \mathbf{x}_t) = \int p(\mathbf{y} | \mathbf{x}_0) p(\mathbf{x}_0 | \mathbf{x}_t) d\mathbf{x}_0, \quad (1.33)$$

but the second factor $p(\mathbf{x}_0 | \mathbf{x}_t)$ is intractable, since it encodes the full data distribution learned by the network. DPS makes it tractable with a single approximation, collapsing $p(\mathbf{x}_0 | \mathbf{x}_t)$ to a point mass at its mean,

$$p(\mathbf{x}_0 | \mathbf{x}_t) \approx \delta(\mathbf{x}_0 - \hat{\mathbf{x}}_0(\mathbf{x}_t)), \quad (1.34)$$

where $\hat{\mathbf{x}}_0(\mathbf{x}_t) \approx \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$ is the denoiser’s own estimate of the clean field, read straight off the noise prediction through Tweedie’s identity, an identity that states that one can retrieve the Bayes optimal posterior mean using the noise estimate.⁴

$$\hat{\mathbf{x}}_0(\mathbf{x}_t) = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\bar{\alpha}_t}}. \quad (1.35)$$

The marginalization integral then collapses to a single evaluation at this estimate, giving the workable approximation

$$p(\mathbf{y} | \mathbf{x}_t) \approx p(\mathbf{y} | \hat{\mathbf{x}}_0(\mathbf{x}_t)), \quad (1.36)$$

⁴The Tweedie formula is a classical result in estimation theory (Efron 2011) stating that the posterior mean of a signal observed under additive Gaussian noise can be expressed as $\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t] = \mathbf{x}_t + \sigma_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. In the diffusion model context, this identity connects the learned denoiser output to the score $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$, and provides a closed form estimate of the clean data $\hat{\mathbf{x}}_0$ at any point during the reverse process.

which a known forward model *can* evaluate, since the right-hand side depends on $\mathbf{x}_t$ only through the deterministic clean-field estimate $\hat{\mathbf{x}}_0(\mathbf{x}_t)$. The approximation is crude at early, noisy steps, where $\hat{\mathbf{x}}_0$ is still very uncertain, and tightens as $t \rightarrow 0$ and the noisy state approaches the clean field.

Equations 1.32-1.36 are the entire general recipe. Reuse a single unconditionally pretrained prior, and at every reverse step add a guidance term built from how well the current clean field estimate reproduces the data. This turns one pretrained generative model into a flexible Bayesian prior that can be used across many inverse problems without retraining. Everything that remains is problem specific, the explicit form of the likelihood $p(\mathbf{y} | \hat{\mathbf{x}}_0)$ set by the forward model, the resulting guidance gradient, its (empirically scheduled) strength, and the full sampling algorithm, and is developed where it is used, in Chapter 5’s FOREST, to reconstruct the dark matter and Lyman- α fields from an observed forest spectrum through a differentiable optical depth forward model.

Note that this is different from the idea of a conditional diffusion model in the sense that in the latter, we learn a stochastic mapping to generate data explicitly conditioned on a conditioning field. This means that the conditioning field is baked in the training itself, resulting in arguably better performance on conditioning fields, when they are similar to that of the training fields, but completely failing when the conditioning field does not sit in distribution with the ones used to train the model on. Where as in the DPS case, although we lose some accuracy based on our approximations, it can work for any arbitrary conditioning data as long we have a physics informed forward model. Another important difference is that a conditional diffusion model need a lot of conditioning fields to train on. Where as the DPS method can work even with a single conditioning field, and can be deployed for most problems where the physics is known.

1.4 Cosmological Simulations

Before it was possible to run even a single realistic simulation, structure formation could only be studied analytically. Press & Schechter (1974) showed that the abundance of collapsed dark matter halos could be predicted directly from the linear power spectrum, without simulating a single particle, by modeling gravitational collapse as the first crossing of a density threshold by a Gaussian random field. Extensions of this approach

remained the primary quantitative tool for halo abundances well into the era when genuine N -body simulations first became feasible. The first simulations, run on the computers of the late 1970s and 1980s, evolved a few thousand to a few tens of thousands of particles (Efstathiou et al. 1985), fewer than the particle count of a single one of the 256^3 -voxel sub-volumes used routinely in this thesis. What has driven the field since then has been mostly improving computational approaches rather than conceptual. The TreePM and moving-mesh algorithms described below were, in their essentials, already understood decades ago, and it is the coming decades of exponential growth in particle counts and force resolution, not a change in the underlying physical modeling, that has turned simulations into the training data generators the field level machine learning methods of this thesis rely on.

Modeling the universe, using analytic and semi analytic approaches, only works so far and has its own limitations, some of which were very briefly mentioned in the previous sections. Understanding the evolution and structure formation due to gravitational interactions and baryonic feedback processes becomes increasingly difficult the smaller the physical scales probed become (for example, below the Jeans scale), and also the higher the density of the fields we wish to study. Numerically solving the coupled equations of gravity (and, where relevant, hydrodynamics) directly is the standard way around this limitation, and the simulations produced this way are also the training data for the machine learning models developed later in this thesis.

1.4.1 N-body simulations

The simplest class of cosmological simulation, an N -body simulation, discretizes the (collisionless) dark matter distribution into a large number of massive particles and evolves their positions and velocities forward in time under their mutual gravity, within a periodic comoving box that expands according to the background cosmology. Computing the gravitational force from every particle on every other particle exactly scales as $\mathcal{O}(N^2)$, which is prohibitive for the particle counts used in modern simulations. In practice, codes instead combine a hierarchical tree algorithm (accurate at short range) with a particle mesh (PM) method solved via fast Fourier transforms (efficient at long range), an approach known as TreePM. GADGET-4 (Springel et al. 2021) is a widely used, publicly available example of such a TreePM code, and (together with its predecessors) underlies many of the dark matter only simulations referenced throughout this thesis. A faster, approximate alternative used in Chapter 6 to generate the large ensembles of dark-

matter-only initial conditions its emulator is trained on is **FastPM** (Feng et al. 2016), a particle mesh only scheme that trades short range force accuracy for speed. The initial conditions for such a simulation are generated directly from the linear power spectrum and transfer function introduced in Section 1.1, displaced onto the particle grid using a Lagrangian perturbation theory scheme. We won’t go into further details of these topics since its not the main focus of this thesis.

1.4.2 Hydrodynamical simulations and subgrid physics

Modeling baryonic tracers such as neutral hydrogen, rather than dark matter alone, requires simulating the gas dynamics as well as gravity. Two broad numerical strategies exist for this: smoothed-particle hydrodynamics (SPH), which represents the gas as a set of particles whose properties are smoothed over a kernel, and the moving mesh approach, which instead discretizes the gas onto an unstructured, quasi Lagrangian Voronoi mesh that moves with the flow. **Arepo** (Weinberger et al. 2020) is the moving mesh code used by the IllustrisTNG simulations discussed below, and was developed in part to combine the adaptivity of SPH with the more accurate shock capturing and reduced numerical noise of a fixed mesh (Eulerian) hydrodynamics solver. For the purposes of this thesis, **GADGET** and **Arepo** can both simply be thought of as the underlying numerical engines used to produce the simulations analyzed in later Chapters, differing in how they discretize the fluid.

Even at the resolution of state of the art cosmological simulations, several physical processes central to galaxy formation, star formation, stellar and active galactic nucleus (AGN) feedback, gas cooling and heating and chemical enrichment, occur on scales far smaller than a single resolution element, and cannot be simulated from first principles. These processes are instead captured through effective *subgrid* models. These are simplified prescriptions, calibrated against observations, that approximate the net effect of the unresolved physics on the resolved gas. This is directly relevant to this thesis’s own results: the suppression of small scale power around $k \sim 1 \, h \, \text{Mpc}^{-1}$, caused by feedback, and the enhancement at $k \gg 1 \, h \, \text{Mpc}^{-1}$ from stars forming in the innermost parts of dark matter haloes, seen in hydrodynamical simulations relative to dark matter only ones, originate precisely from these subgrid feedback processes (Chisari et al. 2019; Schneider & Teyssier 2015; van Daalen et al. 2020). The exact amount of suppression and enhancement varies between simulation suites, but the shape of the effect is a common

prediction across most models, which makes baryonic fields such as gas temperature and stellar density, sensitive to exactly this physics, natural targets and test cases for the emulation methods developed in this thesis.

Star formation is a concrete example of how such a subgrid model works in practice. Following the prescription of [Springel & Hernquist \(2003\)](#) (the basis for the star formation model in IllustrisTNG), gas above a threshold density is assumed to reside in a two phase interstellar medium, consisting of cold, dense clouds embedded in a hot ambient phase in local pressure equilibrium. The cold clouds are the sites of star formation. Rather than converting gas smoothly, individual gas elements above the density threshold are converted into collisionless star particles *stochastically*, with a conversion probability, each timestep, set by a star formation timescale tuned so that the resulting population reproduces the observed relation between gas surface density and star formation rate surface density. Feedback, cooling, and chemical enrichment are treated with conceptually similar, but distinct, subgrid recipes, which are not detailed further here. The broader point is that not all small scale physics can be modeled with perfect physics based complexity, that would be computationally prohibitive. This introduces a fundamental limitation: subgrid approximations introduce systematic uncertainties and biases into the simulated fields. As a result, any machine learning model trained on these simulations inherits these limitations as an upper bound on reconstruction accuracy. No matter how well our inference, reconstruction or generation method performs, it cannot surpass the fidelity of the ground truth simulation data on which it was trained. As we move on, it is important that we remember this fact when we claim the quality and validity of our results.

1.4.3 The IllustrisTNG simulations

IllustrisTNG (TNG) is a suite of large volume, high resolution magnetohydrodynamical simulations run with **Arepo**, incorporating a comprehensive galaxy formation model built on subgrid prescriptions of the kind described above ([Nelson et al. 2021](#); [Pillepich et al. 2017](#)). Multiple box sizes and resolutions are publicly available (TNG50, TNG100, TNG300), trading volume against resolution; the TNG300 and TNG100 volumes, whose computational cost is discussed in Section 1.7, are the ones directly used in this thesis. What distinguishes TNG from a dark matter only simulation of the same volume is precisely the addition of gas dynamics and the calibrated subgrid physics described

above, which is what allows IllustrisTNG to produce realistic galaxy populations, and, of direct relevance to this thesis, physically motivated baryonic fields (gas temperature, HI content, stellar density) alongside the dark matter distribution.

1.4.4 CAMELS

The Cosmology and Astrophysics with Machine Learning Simulations (CAMELS) project (Villaescusa-Navarro et al. 2021b) takes a complementary approach. Rather than a small number of very large, fixed-cosmology volumes, CAMELS runs thousands of smaller simulations, varying *both* cosmological parameters (e.g. Ω_m , σ_8) *and* astrophysical/subgrid feedback parameters across a broad range, using multiple independent simulation codes and galaxy-formation models. This was explicitly designed to enable machine learning and simulation based inference applications that need to be trained and tested for robustness across the full space of plausible feedback scenarios, not just cosmological parameters, since real feedback physics is itself uncertain. This robustness to astrophysics concern is directly relevant to any machine learning emulator built on a single simulation suite (such as the TNG and Sherwood-based methods in this thesis), and is revisited in the discussion of future work.

1.4.5 The Lyman-alpha forest

A different, complementary probe of the diffuse intergalactic medium (IGM), as opposed to the highly dense collapsed halos traced by HI intensity mapping, or the luminous galaxies traced by the surveys discussed in Section 1.6.1, is provided by the Lyman- α forest, which is a dense series of absorption features imprinted on the spectra of high redshift background quasars by intervening neutral hydrogen along the line of sight. Even for a given absorption frequency, there is a series of absorption at different observed frequencies because of the redshifting of light that travels all the way from the quasar to us, the observer.

The standard way to model this signal in a simulation is to extract one dimensional *line of sight (LOS) skewers* of the baryonic field properties through the simulated volume, mimicking the path a real quasar’s light would take through the intervening IGM. Each skewer is then converted into a mock absorption (transmitted flux) spectrum via an integral of the local optical depth along the line of sight. Crucially, the resulting flux spectrum is not a simple tracer of the dark matter density alone. It depends on the

baryonic gas density, on its thermal state, and on the line of sight peculiar velocity field, which distorts the mapping between real space and the observed (velocity/redshift) space of the spectrum. This sensitivity to baryonic physics, and not just to the underlying gravitational clustering, is precisely why hydrodynamical, rather than dark matter only, simulations are required to model the Lyman- α forest realistically.

The Lyman- α analysis of **FOREST** in this thesis (Chapter 5) is built on the Sherwood-Relics suite of hydrodynamical simulations (Bolton et al. 2017; Puchwein et al. 2023), run with **GADGET-3** at $z = 2.5$ in periodic boxes of comoving side $40 h^{-1}\text{Mpc}$. This is the same simulation suite used by the variational autoencoder (VAE)-based DeepCHART reconstruction method (Maitra et al. 2026) (Section 1.3.5), a recent work on mapping dark matter to Lyman alpha forest signal.

1.5 21 cm Cosmology and Intensity Mapping

1.5.1 The hyperfine transition and spin temperature

Neutral hydrogen (HI) undergoes a forbidden ground state hyperfine splitting, caused by the interaction between the magnetic moments of the proton and electron, whose transition corresponds to a rest frame wavelength of 21 cm (frequency $\nu_{21} = 1420.4\text{ MHz}$). The relative population of the two hyperfine levels is described by the spin temperature T_S , defined through the Boltzmann relation between the level populations,

$$\frac{n_1}{n_0} = 3 e^{-T_\star/T_S} \quad (1.37)$$

where n_1 and n_0 are the populations of the upper and lower hyperfine levels, the factor of 3 is the ratio of their statistical weights, and $T_\star \equiv h\nu_{21}/k_B \approx 0.068\text{ K}$ is the temperature corresponding to the transition energy.

The spin temperature is set by a competition between three effects: absorption and(or) stimulated emission of CMB photons, which drives T_S towards the CMB temperature T_{CMB} , collisions between hydrogen atoms (and, at high redshift, with free electrons), which drive T_S towards the kinetic temperature of the gas T_K , and scattering of Lyman- α photons, which does the same via the Wouthuysen Field effect (Furlanetto et al. 2006). In the regimes relevant to this thesis, at the redshifts and densities probed by post reionization 21 cm intensity mapping, collisions dominate and are efficient enough that T_S closely tracks T_K , decoupling it from T_{CMB} .

The observable 21 cm signal is the differential brightness temperature of a patch of sky relative to the CMB, which can be written schematically as

$$\delta T_b(z) \approx \bar{T}_b(z) \left(1 + \delta_{\text{HI}}\right) \left(1 - \frac{T_{\text{CMB}}}{T_S}\right), \quad (1.38)$$

where δ_{HI} is the HI density contrast and $\bar{T}_b(z)$ is a redshift-dependent normalization set by the mean HI abundance and the background cosmology (Furlanetto et al. 2006). In the post-reionization regime relevant to the intensity mapping experiments discussed in Section 1.6.2, $T_S \gg T_{\text{CMB}}$ to good approximation, so the factor in the right parentheses tends to unity and the 21 cm signal is seen in emission, tracing $(1 + \delta_{\text{HI}})$ directly.

1.5.2 Self shielding and optical depth

Not all neutral hydrogen contributes equally to the observable signal. In the presence of an ionizing ultraviolet background, gas at low density is highly ionized, while sufficiently dense gas self shields against this background and remains predominantly neutral. This transition is governed by the local optical depth to ionizing photons, and is commonly modeled in simulations via the fitting formula of Rahmati et al. (2013), which prescribes the local photoionization rate as a function of gas density:

$$\frac{\Gamma_{\text{phot}}}{\Gamma_{\text{UVB}}} = (1 - f) \left[1 + \left(\frac{n_{\text{H}}}{n_0}\right)^\beta\right]^{\alpha_1} + f \left[1 + \frac{n_{\text{H}}}{n_0}\right]^{\alpha_2}, \quad (1.39)$$

where Γ_{UVB} is the uniform UV background rate, n_0 is a characteristic self shielding density threshold that depends on redshift and Γ_{UVB} , and $\alpha_1, \alpha_2, \beta, f$ are parameters calibrated against full radiative transfer simulations. At low densities ($n_{\text{H}} \ll n_0$), the gas sees the full background and is highly ionized; at high densities ($n_{\text{H}} \gg n_0$), Γ_{phot} is strongly suppressed and the gas remains predominantly neutral. The neutral fraction itself is then obtained by solving the ionization equilibrium between photoionization, collisional ionization, and recombination at each resolution element. This self shielding prescription is what allows a simulation, which by default tracks total gas (and dark matter) density, to be converted into a physically meaningful HI field for comparison with 21 cm observations, rather than simply assuming all gas is neutral.

1.5.3 Intensity mapping

Rather than individually detecting and cataloging galaxies, intensity mapping (IM) measures the aggregate, unresolved 21 cm emission from all of the HI within a given angular pixel and frequency channel, effectively treating the sky as a continuous 3D field rather than a catalog of discrete sources. This has two main advantages over traditional galaxy redshift surveys: it does not require individually resolving faint sources, allowing much larger comoving volumes and higher redshifts to be surveyed in a given observing time (Villaescusa-Navarro et al. 2018b). And the observed frequency directly encodes the redshift of the emitting gas via the cosmological red-shifting of the 21 cm line, providing three dimensional maps without the need for spectroscopic follow up of individual objects (e.g. Bharadwaj et al. 2001; Battye et al. 2004; Chang et al. 2008). The price paid for this efficiency is that the signal is dominated by bright astrophysical and instrumental foregrounds (Galactic synchrotron emission, extragalactic point sources, and radio frequency interference) that must be modeled and removed, and that individual objects can no longer be identified, only the aggregate field. Cross correlating IM maps with an independent, overlapping galaxy survey is a standard and powerful strategy for both validating the recovered HI signal and mitigating foreground contamination, since foregrounds are not expected to correlate with the (independently observed) galaxy positions.

1.5.4 Modeling the HI field

Translating between the underlying matter distribution and the observable HI field that intensity mapping probes requires a model connecting the two, and three broad approaches exist.

Halo models: Halo model approaches assign HI to dark matter halos according to a prescribed HI-halo mass relation $M_{\text{HI}}(M)$ (Padmanabhan et al. 2023), often calibrated against observations of damped Lyman- α systems, both from the halo model side (Padmanabhan & Refregier 2016) and observationally (Dev et al. 2023; Obuljen et al. 2019). This approach is computationally cheap and can be applied to very large volumes, but relies on a simplified, effective description of a process that is, in reality, governed by detailed gas physics within and around individual halos.

Perturbation theory based modeling. As discussed in Section 1.1, standard perturbation theory can in principle be extended to describe biased tracers such as HI (Obuljen et al. 2023), but is subject to the same breakdown on mildly non linear and non linear scales as the dark matter perturbation theory discussed there, and is not treated in further analytic detail in this thesis.

Hydrodynamical simulations. Full hydrodynamical simulations solve for the gas dynamics directly (subject to the self shielding and other subgrid prescriptions discussed in Section 1.4), capturing small scale astrophysical processes such as star formation, feedback, radiative transfer, that the two approaches above must instead approximate (Villaescusa-Navarro et al. 2018a). This comes at substantially higher computational cost, as quantified in Section 1.7, and is the primary motivation throughout this thesis for developing fast machine learning based surrogates that retain most, if not all of the hydrodynamical simulation fidelity at a fraction of the cost.

Data driven and generative approaches. Beyond these three established methods, a growing body of work learns the HI matter mapping directly from simulations in a data driven way, without prescribing a specific functional form for $M_{\text{HI}}(M)$ in advance (e.g. Hitz et al. 2025; Li et al. 2022; Spinelli et al. 2020). These forward modelling approaches extend predictions to large volumes by applying calibrated prescriptions for the HI halo connection, but still fall short of learning the full field level mapping from simulations. In the specific context of 21 cm intensity mapping, several recent efforts predict the brightness temperature signal at the field level using generative models, including normalising flows (Hassan et al. 2022), neural networks (Wadekar et al. 2021), and generative adversarial networks (Andrianomena et al. 2022). Diffusion models have so far found very limited application in this domain, and none specifically within 21 cm IM, the gap addressed by the diffusion based pipeline of Chapter 2.

1.6 Galaxy Surveys and 21 cm Experiments

Beyond the theoretical treatment of the 21 cm signal in Section 1.5, two further observational routes to the large scale structure are worth briefly reviewing: galaxy surveys, one of the most established probes of large scale structure, and the current and upcoming generation of 21 cm intensity mapping experiments that are putting the theoretical

picture above to the test. Together with HI intensity mapping itself, they cross correlate for extracting and inferring cosmological parameters, a theme of Chapter 6 of this thesis.

1.6.1 Galaxy Surveys

We first delve into the two dominant survey approaches, namely the spectroscopic redshift surveys and the weak gravitational lensing.

Redshift surveys and baryon acoustic oscillations

Spectroscopic redshift surveys measure the three dimensional positions of large numbers of galaxies by combining their angular position on the sky with a redshift obtained from an optical spectrum, directly mapping out the galaxy distribution as a tracer of the underlying matter field. The Sloan Digital Sky Survey (SDSS) and its successors the Baryon Oscillation Spectroscopic Survey (BOSS) ([Dawson 2013](#); [Alam et al. 2017](#)) and its extension eBOSS have produced some of the most precise galaxy clustering measurements to date, and the ongoing Dark Energy Spectroscopic Instrument (DESI) survey is extending this approach to substantially larger volumes and higher redshifts ([DESI Collaboration et al. 2025](#)).

A key feature exploited by these surveys is the baryon acoustic oscillation (BAO) signal. Acoustic waves in the pre recombination photon-baryon plasma imprint a characteristic scale on the late-time galaxy distribution, corresponding to the sound horizon at recombination. Because this physical scale is known precisely from CMB physics, its apparent size in a galaxy survey acts as a “standard ruler”, providing a geometric measurement of the expansion history that is largely insensitive to the details of non linear structure formation, in contrast to the shape of the full power spectrum, which is sensitive to precisely the non linear physics discussed in Section 1.1. A second, complementary signal available in the same data is redshift-space distortions (RSD): peculiar velocities along the line of sight distort the observed (redshift-space) galaxy positions relative to their true, real space positions, providing a measurement of the growth rate of structure and thus an additional handle on gravity and the growth factor $D_+(a)$ introduced in Section 1.1 (e.g. [Ross et al. 2015](#)).

Weak gravitational lensing

A complementary approach, sensitive to the total matter distribution (dark and baryonic) rather than only to the biased tracer traced by galaxy positions, is weak gravitational lensing. Light from distant background galaxies is very slightly deflected and sheared by the intervening large scale structure along the line of sight. While this distortion is far too small to detect in any single galaxy image, it can be measured statistically from the coherent alignment of many background galaxy shapes, a signal known as cosmic shear. Major ongoing and recent lensing surveys include the Dark Energy Survey (DES) (Abbott et al. 2019) and the Kilo Degree Survey (KiDS) (Hildebrandt et al. 2017), with substantially larger and deeper ongoing surveys, that include Euclid (Blanchard et al. 2020; Euclid Collaboration 2020; Scaramella et al. 2022) and the Vera C. Rubin Observatory’s Legacy Survey of Space and Time (LSST) (Ivezić et al. 2019).

Because weak lensing probes the total matter distribution directly, rather than a specific luminous or 21 cm tracer, cross correlating a lensing survey with 21 cm intensity maps provides a way to isolate the cosmological signal from other systematics and astrophysical uncertainties that do not correlate between independent probes, enabling foreground mitigation. Cross correlations between weak lensing and post reionization neutral hydrogen probes have already been explored as a route to constraining dark energy (Dash & Sarkar 2021), and this kind of multi tracer cross correlation, extended to the field level using the machine learning tools developed later in this thesis, is one of the main reasons for motivating the joint dark matter to galaxy, and dark matter to HI fields emulation approaches presented in chapter 6.

1.6.2 21 cm Experiments

As mentioned, a complementary observable to galaxy surveys are 21 cm intensity mapping experiments, ranging from currently operating, to next generation facilities still under construction. These experiments probe different redshift ranges, enabling tomographic mapping of HI across vast and distinct cosmic time.

21 cm intensity mapping instruments

The Canadian Hydrogen Intensity Mapping Experiment (CHIME) is a transit interferometer without moving parts, covering the frequency range corresponding to $0.8 \lesssim z \lesssim 2.5$ (Bandura et al. 2014). CHIME has already reported a direct detection of

cosmological 21 cm emission (CHIME Collaboration et al. 2022) between $0.78 \lesssim z \lesssim 1.43$, a major milestone for interferometric 21 cm IM. Related interferometric instruments include CHORD (Collaboration et al. 2026), a next generation Canadian array currently under construction and designed to substantially improve on CHIME’s sensitivity and frequency coverage, and HIRAX (Hydrogen Intensity and Real time Analysis eXperiment), a planned array of 6m dishes to be sited in South Africa, targeting baryon acoustic oscillations over a wide sky area and redshift range $0.8 \lesssim z \lesssim 2.5$ (Newburgh et al. 2016). On the single dish side, the Green Bank Telescope (GBT) was used for some of the earliest cross correlation detections of the 21 cm signal (Masui et al. 2013; Wolz et al. 2022), and the Five hundred meter Aperture Spherical Telescope (FAST) in China, the largest single dish radio telescope in the world, is forecast to deliver competitive intensity mapping constraints once dedicated IM surveys are carried out (Hu et al. 2020).

MeerKAT and MeerKLASS

MeerKAT, a 64 dish interferometer in South Africa and an SKA precursor, is one of the current facilities for single dish 21 cm intensity mapping through its MeerKAT Large Area Synoptic Survey (MeerKLASS) program (Santos et al. 2017; MeerKLASS Collaboration 2025). A lot of work has gone into the calibration and foreground mitigation pipeline needed to extract a cosmological signal from single dish autocorrelation data (Wang et al. 2021; Irfan et al. 2022), resulting in a first cross correlation detection of the 21 cm signal with WiggleZ galaxies at 7.7σ significance (Cunnington et al. 2022). More recently, deeper MeerKLASS observations, cross correlated with an independent spectroscopic galaxy sample, have pushed this kind of measurement to improve statistical significance and to new redshift ranges, representing the deepest single dish HI intensity maps obtained to date. Taken together, these results mark a milestone for the field, demonstrating that the current 21 cm intensity mapping can deliver robust cosmological cross correlation detections ahead of the full SKA era.

The Square Kilometer Array Observatory

Radio cosmology is one of the flagship science drivers of the Square Kilometer Array Observatory (SKAO),⁵ which will ultimately comprise two complementary telescopes: SKA Low, located in Western Australia and optimized for lower radio frequencies (and

⁵<https://www.skao.int/>

correspondingly higher redshifts, probing the epoch of reionization and beyond), and SKA Mid, located in South Africa (at the same site as MeerKAT, which will be incorporated into the array), covering higher frequencies corresponding to the lower redshift, post reionization Universe. Operated in single dish (autocorrelation) mode, SKA Mid is expected to conduct 21 cm intensity mapping surveys across cosmologically relevant volumes out to $z \sim 3$ (Santos et al. 2015; SKA Cosmology SWG 2020), directly extending the redshift range already being explored by its MeerKAT/MeerKLASS precursors. The results already obtained by CHIME and MeerKAT/MeerKLASS, discussed above, serve as important pathfinders and proofs of concept for the much larger surveys the SKAO is ultimately designed to carry out.

1.7 Addressing the simulation bottleneck

A common thread running through the ideas laid out above is that in every case the most physically complete description, full hydrodynamical simulations resolving gravity, gas dynamics, and the subgrid astrophysics of star formation and feedback, to match and simulate the real observations is also the most computationally expensive one, precisely in the small scale, non linear regime ($k \gtrsim k_{\text{NL}}$) where the new instruments of Section 1.6.2 are beginning to deliver data, and where perturbative and halo based approaches remain tractable only by trading away the very physical content (baryonic feedback, self shielded HI, small scale non linear clustering) those surveys are most sensitive to. This tension, between the fidelity of hydrodynamical simulations and their cost at the scale required for modern inference, is the subject of this section.

As introduced in Section 1.4, cosmological simulations are essential tools for understanding the formation and evolution of structure in the Universe, letting us explore the complex interplay of gravitational, hydrodynamic, and astrophysical processes shaping the distribution of matter from large down to highly non linear galactic and sub galactic scales. However, high fidelity hydrodynamical simulations are computationally very expensive, limiting their applicability when the parameter space is large in inference studies. For example, the IllustrisTNG simulations (Nelson et al. 2021; Pillepich et al. 2017) required approximately 35 million CPU hours for the largest IllustrisTNG run and about 18 million CPU hours for TNG100, while EAGLE (Schaye et al. 2014) consumed roughly 4.5 million CPU hours.

Such computational demands become particularly prohibitive for cosmological or

astrophysical field level inference and simulation based inference (SBI) frameworks, which require large ensembles of conditional forward simulations to explore the parameter spaces. Recent works have shown promising applications, including the cosmological context (Angulo & Hahn 2022; Jeffrey et al. 2020; Bairagi et al. 2025; Zeghal et al. 2025; Karchev et al. 2023; Saxena et al. 2024; Tucci & Schmidt 2024). However, their applicability remains limited by the requirement of generating sufficient high fidelity simulations. This computational bottleneck underlines the need for fast, accurate surrogate models capable of reproducing and generating cosmological fields at a fraction of the cost, not limited to dark matter, but crucially including baryonic observables that can be directly constrained with data.

This same bottleneck constrains the theoretical modeling of the 21 cm signal itself. Of the three approaches introduced in Section 1.5.4, the full hydrodynamical route is the most limited by cost, since modeling the IM signal requires simultaneously large volumes and the high resolution needed to resolve the small mass ($\sim 10^{10} M_{\odot}/h$) dark matter halos that host HI. A method capable of learning the small scale physics from state of the art simulations while at the same time reaching large scales and volumes would therefore be of great importance, this is the computational and modeling gap this thesis addresses.

1.8 Statement of Contributions

Several chapters of this thesis are based, in part, on work carried out in collaboration with others and published or submitted as journal articles or preprints. This section lists these chapters, the corresponding publication, and the specific contribution to each.

Chapter 2 (LODI) is based on the paper Mishra et al. (2025) named **Large, fast and accurate HI intensity maps with latent overlap diffusion**(S. Mishra, R. Trotta, M. Viel), on which the author of this thesis is the first author. As first author, the candidate led the conceptual development, implementation, training, and analysis presented in this chapter, under the supervision of the co-authors.

Chapter 3 (Cosmo-FOLD) which is based on the preprint Mishra et al. (2026) and currently under review, called **Cosmo-FOLD: Fast generation and upscaling of field level cosmological maps with overlap latent diffusion**(S. Mishra, R. Trotta, M. Viel), on which the author of this thesis is the first author. The

author’s specific contribution was developing and implementing the **Cosmo-FOLD** and positional encoding techniques, training and validating the machine learning methods, under the supervision of the co-authors.

Chapter 4 (Posterior Predictive Fields under Planck Cosmology) presents an *original but unpublished work, developed for this thesis*, and is a direct extension to the work of [Mishra et al. \(2026\)](#). The modification, implementation and testing of the machine learning methodology was carried out by the author of this thesis.

Chapter 5 (FOREST) or Field level pOsterior REconstruction via Sightline Tomography, *presents an original and ongoing, unpublished for now*, that has been developed almost entirely by the author of this thesis, with useful insights coming from the co supervisors.

Chapter 6 is based on the preprint [Scelfo et al. \(2026\)](#) with the title: **Field level multi tracers simulation based inference of cosmological parameters from 3D maps** (G. Scelfo, S. Mishra, M. Rigo, R. Trotta, M. Viel) on which the author is the second author. This article is currently under review. The author’s contribution to that work was the design, implementation, and validation of the HGL field level emulator (Sections [6.4](#), [6.5](#), [6.7.1](#)). The simulation based inference (SBI) pipeline built on top of the emulator (Section [6.6](#)) and the resulting cosmological constraints (Section [6.7.2](#)) were developed jointly with the co-authors.

Chapter 2

LODI: Large, Fast and Accurate Neutral Hydrogen Intensity Maps with Latent Overlap Diffusion

2.1 Problem statement

Section 1.7 of the introduction established the core computational bottleneck this thesis addresses: full hydrodynamical simulations are the most physically complete way of predicting baryonic observables such as the 21 cm brightness temperature field, but are far too expensive to run, both in numbers and at the volumes, required by field level and simulation based inference. For 21 cm intensity mapping specifically, section 1.5.4 reviewed the three established ways around this cost. Namely, halo model prescriptions, which are cheap and scale to arbitrarily large volumes but rely on a simplified, effective $M_{\text{HI}}(M)$ relation calibrated against a limited set of observations, perturbation theory, which is analytically principled but breaks down on the same mildly non linear and non linear scales where most of the small scale HI signal resides, and full hydrodynamical simulations themselves, which capture this small scale physics directly but are, again, the expensive option this thesis is trying to move away from. None of these three routes is simultaneously fast, accurate on non linear scales, and scalable to the large volumes intensity mapping surveys aim to probe, more so in the future.

Machine learning offers a fourth route, and section 1.5.4 also reviewed the data driven efforts already applied specifically to 21 cm intensity mapping. Normalizing flows (Hassan et al. 2022), convolutional neural networks (Wadekar et al. 2021), and

generative adversarial networks (Andrianomena et al. 2022) have all been used to predict the HI brightness temperature field directly from the underlying matter distribution. Diffusion models, reviewed more generally in section 1.3.6 for their comparatively high fidelity at capturing small scale, non Gaussian structure, have by contrast found very limited application in cosmology and, to our knowledge, none specifically within 21 cm intensity mapping. This is the specific gap LODI targets: a diffusion based, field level model for the HI brightness temperature field, conditioned directly on the dark matter distribution.

2.2 LODI pipeline

LODI closes this gap in two stages. A conditional U-Net (HALOgen, section 2.4.1) first learns the mapping from a fast, approximate dark matter only simulation to the underlying halo distribution, removing the need for a full N-body run just to identify halos. A diffusion model then paints the 21 cm brightness temperature field onto that halo distribution (section 2.4.2), capturing small scale, non linear structure that halo model or perturbative approaches have a hard time approaching. A latent overlap generation scheme, described later in this chapter, lets both networks, trained on a single, computationally tractable box size, generate volumes substantially larger than that training box, without the seams or edge artifacts of naive patch-based tiling. This is, to our knowledge, the first demonstration of a diffusion-based, halo-conditioned approach to 21 cm intensity mapping.

2.3 Data and simulations

For this work, we use the CAMELS suite of cosmological simulations (Villaescusa-Navarro et al. 2022), specifically the IllustrisTNG ones, which are created by varying cosmological parameters (Ω_m and σ_8) and astrophysical parameters (supernova and AGN feedback) across different realizations. Specifically, A_{SN1} and A_{SN2} are normalization factors for the energy and speed of galactic wind feedback, respectively; A_{AGN1} and A_{AGN2} control the energy output and of AGN feedback. These factors are with respect to the ones that are used in the big IllustrisTNG runs, and so a value set to 1 for these astrophysics parameters means that they have the same magnitude as the IllustrisTNG counterpart. These simulations are run using the moving-mesh hydrodynamics code

AREPO (Weinberger et al. 2020). Stellar feedback follows the IllustrisTNG prescriptions, and particles are evolved from $z = 127$.

For our work, we train on the Cosmic Variance (CV) set of IllustrisTNG realizations, comprising 27 simulations with identical cosmology and astrophysics but different initial conditions. Each simulation box has a comoving side length of $25 h^{-1}$ Mpc with 256^3 gas and dark matter particles. We use the fiducial IllustrisTNG model: $\Omega_m = 0.3089$, $\sigma_8 = 0.8159$, and all four astrophysical parameters (collectively denoted as $\mathbf{A}$) set to their standard values of 1.0. This choice allows us to test generalization across different initial seeds while keeping cosmology and astrophysics fixed. Here, Ω_m denotes the present-day matter density parameter, and σ_8 is the root-mean-square amplitude of linear matter fluctuations on scales of $8 h^{-1}$ Mpc.

2.4 Methodology

In this section, we explain the machine learning technique, the mock datasets used, and the training methods implemented.

Our strategy is the following: starting from a dark matter only map, we use a U-Net architecture to predict the halo map, as described in section 2.4.1.

We then use the halo map as input to a variational diffusion model trained to inpaint the HI brightness temperature, as described in section 2.4.2.

An overview of the full pipeline is presented in Fig. 2.1.

2.4.1 HALOgen: Extracting halos from DM only sim

Network architecture

A U-Net, as discussed in the introduction section 1.3.1, is a deep hierarchical convolutional neural network with encoder and decoder layers, used to perform image to image translation tasks, an ideal architecture for our aim of predicting halo maps from ρ_{DM} maps.

We use ResNet blocks (He et al. 2015) (Fig 2.5) as the fundamental blocks for the U-Net architecture, replacing the simple convolution blocks mentioned in section 1.3.1 as well as an attention layer (See sec. 1.3.2) at the bottleneck motivated by Petit et al. (2021). The attention block enhances the ability of the model to capture and highlight complex structure and long-range dependencies within the data. The architecture,

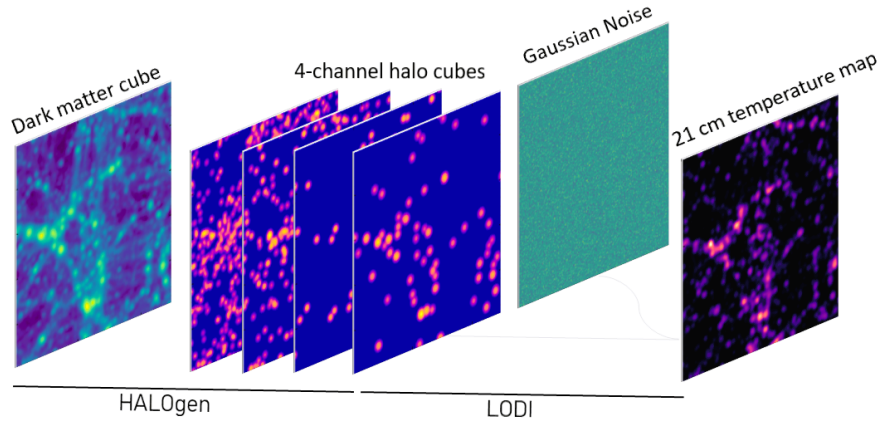


Figure 2.1: An overview of our generative pipeline, starting from dark matter particle distribution as produced by an N-body code to the final 21cm intensity map: in the first step, a ResNet with attention bottleneck called **HALOgen** is used to predict the dark matter halo mass density in four mass channels; subsequently, a variational diffusion model with latent overlap, or **LODI** generates the 21 cm brightness temperature map.

depicted in Fig. 2.2, consists of 3 encoding blocks with grouped convolutions, which process a 3D ρ_{DM} input of size 64^3 voxels. The input is progressively downsampled through three stages, reaching a bottleneck representation of size of 8^3 voxels, with $2 \times 2 \times 2$ kernel max-pooling layers employed for downsampling to reduce spatial resolution while preserving dominant features. For the up-sampling leg of the architecture, we use trilinear interpolation instead of the more commonly used transposed convolutions to avoid checkered board artifacts and grid like patterns that can arise when upsampling images. Trilinear interpolation estimates voxel values by computing a weighted average of the eight nearest neighbors, ensuring a smoother reconstruction. To enhance high-level feature preservation, we incorporate skip connections (represented as horizontal grey lines in Fig. 2.2) after upsampling, whose aim is to transfer directly feature maps from corresponding encoder layers to decoder layers, thus bypassing the bottleneck.

The last decoder block leads to a set of grouped convolution layers that transform the output into 4 channels, each corresponding to a halo map density field $\rho_{\text{h}}^{(i)}$, where $i = 1, \dots, 4$ is the channel number. Each channel describes halos within a top-hat mass range (a bin). Attributing halos to different mass bins is necessary to reduce the

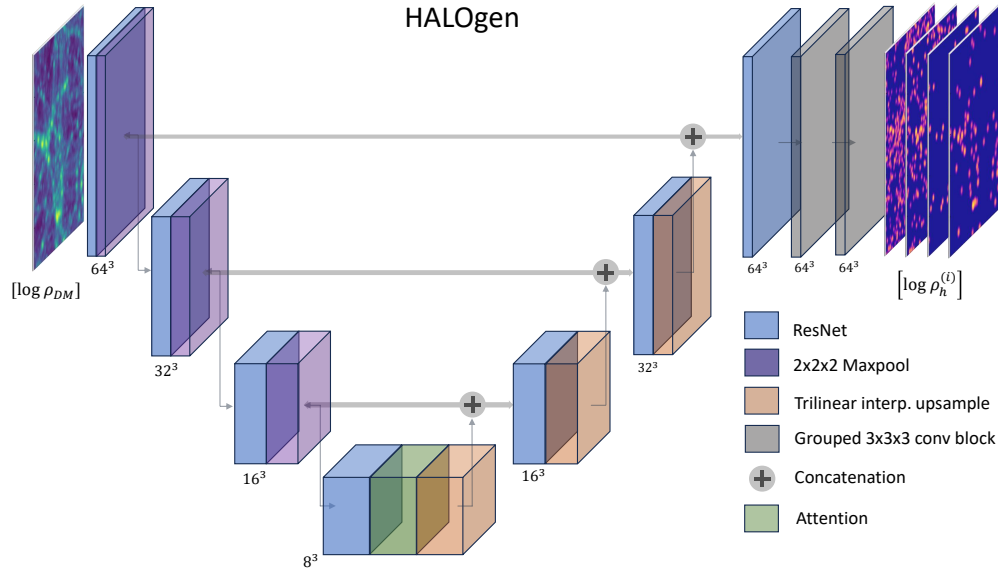


Figure 2.2: Overview of the **HALOgen** (Halo Assignment and Generation with U-Net) architecture used for halo assignment from an input DM-only map, employing grouped 3D convolutions (gray blocks) and an attention mechanism at the bottleneck. Upsampling is done via trilinear interpolation to avoid artifacts in the upsampled map. Horizontal lines represent skip connections. The input data has shape of 64^3 voxels while the bottleneck has a spatial shape of 8^3 .

dynamic range of the relationship between halo mass and HI mass, and to account for its strong dependence on halo mass: halos with mass $M_{\text{halo}} < 3 \times 10^{10} M_{\odot}$ contribute $\approx 95\%$ of the total number count of halos, but their combined HI mass is only $\approx 5\%$ of the total HI in the simulation. At the other end of the mass distribution, halos of mass $\geq 10^{12} M_{\odot}$ populate only about $\approx 0.5\%$ of the simulation but account for more than 25% of the HI mass. Our binning both reduces the dynamic range spanned by the M_{HI} to M_{halo} relationship (thus facilitating learning) and also accounts for the halo-mass dependency of the relation. The HI mass function, as determined by [Ma et al. \(2025\)](#); [Jones et al. \(2018\)](#), features a knee HI mass of $\approx 7.5 \times 10^9 M_{\odot}$, with the power-law slope for low mass being $\alpha = -1.3$, indicating that contribution from HI regions smaller than the knee HI mass is non-negligible. Following [Villaescusa-Navarro et al. \(2018a\)](#), the HI cutoff mass for our case is $\approx 10^8 M_{\odot}$, which is almost two orders of magnitude smaller than the knee mass determined observationally. This provides reassurance that we are accounting for most of the HI regions.

The (fixed) boundaries of the top-hat mass ranges are chosen to obtain similar HI mass within each bin¹, and are given by:

$$\left[3 \times 10^{10}, 10^{11}, 5 \times 10^{11}, 10^{12}, \max(M_{\text{halo}}) \right] M_{\odot}. \quad (2.1)$$

where the maximum value of M_{halo} is $\approx 1.7 \times 10^{14} M_{\odot}$. We use group convolution and group normalization layers to enforce that output channels be independent; experimentation with mixing the information between different groups (4 or multiples of 4) has shown that this degrades performance. The output of the network has a voxel dimension of 4×64^3 . Compared to larger simulations such as TNG300, which host halos up to $\approx 10^{15} M_{\odot}$, the smaller CAMELS volumes do not sample such large systems. While this restricts the direct modeling of the rarest, most massive clusters, CAMELS provides a more machine-learning-friendly baseline, and our approach can in the future be extended to larger simulations or large cluster regions.

To create the full 256^3 voxels map, since this is a deterministic convolutional model, we pass the full 256^3 input to the trained U-net to get the final output. Note that this is possible in this case not just because of the convolutional nature, but also because the trained model is small enough that the computations on the full 256^3 volume does

¹This remains true, up to a few percent, for all maps in the cross-validation set of the simulation, which share the same astrophysical parameters. Generalizing this approach across different parameter values will be the focus of future work.

not overload the memory of the GPU we use.

Training dark matter and halo mass density maps

To obtain the input DM mass density map, we take the positions for each DM particle from the DM only simulations with side $25 h^{-1}$ Mpc and convert them to 3 dimensional ρ_{DM} fields of size 256^3 voxels, by using the cloud in cell (CIC) algorithm. The (CIC) algorithm assigns particle masses to a regular grid by distributing each particle’s mass to the eight nearest grid points, weighted linearly by the particle’s relative distance from each grid point. This produces a continuous density field while conserving total mass. Our field resolution is thus $\approx 0.1 h^{-1}$ Mpc. We use for our model training the CAMELS dataset from [Villaescusa-Navarro et al. \(2022\)](#), a suite of hydrodynamical and N-body cosmological simulations.

To determine the halo mass density field ρ_{h} , we use the SUBFIND algorithm ([Springel et al. 2000](#)), available in the simulation suite to extract the halo center of mass positions and then use the same CIC algorithm, weighted by the mass of the halos. We manually create the four halo channels using the bins given in Eq. (2.1), and apply the CIC algorithm on each bin separately.

Maps are smoothed with a Gaussian kernel of fixed radius $R = 0.2 h^{-1}$ Mpc to increase the signal to noise ratio (SNR) and remove irrelevant small-scale features to help training.

HALOgen loss function

The loss function needs to deal with the sparsity of the $\rho_{\text{h}}^{(i)}$ maps. Furthermore, a given DM voxel may be associated with multiple halos of different masses (and falling into different halo mass bins), making it challenging for the model to disentangle their respective contributions. We call this phenomenon ‘channel mixing’, i.e., the fact that halo information from one channel leaks into adjacent channels. Moreover, since the binning scheme is not explicitly encoded within the neural network architecture, the model must learn to distinguish halos across adjacent bins without misclassifications. Ensuring proper bin differentiation is crucial to maintaining the physical consistency of the model predictions.

To address the above issues, we create a custom-weighted mask, m_w^i ($i = 1, \dots, 4$), associated to each halo channel i . The purpose of the mask is to give higher weight to pixels with halos in channel i , while enforcing a negligible contribution to the loss from

halos in neighboring channels, thus helping both with sparsity and channel separation.

For a given halo channel i , with halo mass density field $\rho_h^{(i)}$, we define its weight as:

$$w^i = \frac{N_{\text{mesh}}^3}{\sum_{j=\max(1,i-1)}^{\min(4,i+1)} \sum_{k=1}^{N_{\text{mesh}}} \theta(\rho_h^{(j,k)} - T_c^{(j)})} \geq 1, \quad (2.2)$$

where $\rho_h^{(j,k)}$ is the halo field density for channel j at pixel k and $\theta(x)$ is the Heaviside function. The threshold $T_c^{(j)}$ is determined from the empirical CDF of $\rho_h^{(j,k)}$, in such a way that

$$P(\rho_h^{(j,k)} > T_c^{(j)}) = 0.05. \quad (2.3)$$

In other words, the threshold eliminates pixels that contain low-density values, and only retains the highest 5% of the log-density distribution in each channel. Note that for a given channel j , the pixels used for the computation are taken not only from the channel j , but also from the immediate neighboring channels, $j+1$ and $j-1$. For the boundary channels (i.e., $j=1$ and $j=4$), we include only the next or previous channel, respectively. Now we define a pixel- and channel-specific mask $m_w^{(i,k)}$ by incorporating these weights:

$$m_w^{i,k} = \begin{cases} w^i & \text{if } \rho_h^{(i,k)} > T_c^{(i)} \\ 1 & \text{if } \rho_h^{(i,k)} \leq T_c^{(i)}. \end{cases} \quad (2.4)$$

We normalize the mask such that $\sum_{i,k} m_w^{i,k} = 64$, or each training mask of shape 4×64^3 adds to 1 on average, since each simulation is made of 64 sub-volumes $(256/64)^3 = 64$.

An illustration of the construction of $m_w^{(i,k)}$ is shown in figure 2.3. By construction, pixels corresponding to halo sites for channel j (beige) have high mask values; the mask also has large values for adjacent channel's halo positions (golden), in order to enforce the absence of halos there in the particular channel under consideration, since for these regions, the target voxel value corresponds to the small background values.

The masked loss function is then written as:

$$\mathcal{L} = \sum_{j=1}^4 \left(\alpha L_2^j \odot m_w^j \beta^j + (1 - \alpha) L_q^j \right) + \lambda_w L_1^{\text{reg}} \quad (2.5)$$

with $\alpha, \beta^{j=1,\dots,4}, \lambda_w$ being hyper-parameters of the model. The loss incorporates both the mean squared error L_2^j and the quantile loss L_q^j between the pixel-level target density,

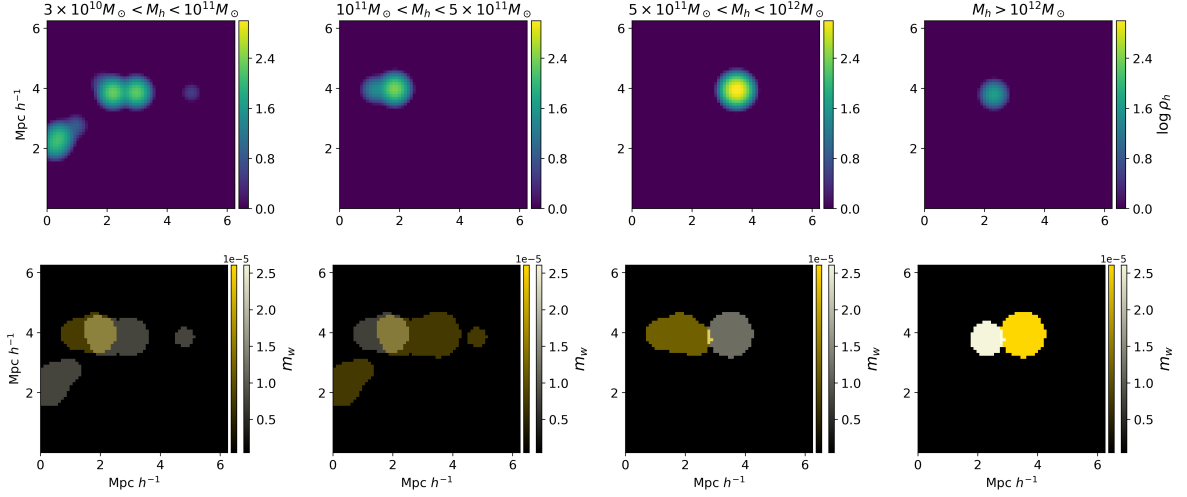


Figure 2.3: Illustration of the halo density channels masking prescription. The first row of 4 images shows a 2D slice of the training halo maps $\rho_h^{(j)}$ for each halo mass channel j . The second row shows the corresponding 2D weighted mask maps, $m_w^{j,k}$. In each masking map, the contribution from the halo j is shown in beige, and the contribution from the neighboring channel's halo location is shown in golden.

$\rho_h^{(j,k)}$, and its predicted value, $\rho_{\text{pred}}^{(j,k)}$, weighed by the mask for the L_2 loss, with the direct product running over pixel values k :

$$L_q^j = \frac{1}{N_{\text{mesh}}} \sum_{k=1}^{N_{\text{mesh}}} \max \left(q \left(\rho_h^{(j,k)} - \rho_{\text{pred}}^{(j,k)} \right), (q-1) \left(\rho_h^{(j,k)} - \rho_{\text{pred}}^{(j,k)} \right) \right), \quad (2.6)$$

$$L_2^j = \frac{1}{N_{\text{mesh}}} \sum_{k=1}^{N_{\text{mesh}}} \left(\rho_{\text{pred}}^{(j,k)} - \rho_h^{(j,k)} \right)^2.$$

Here $q \in (0, 1)$ is the chosen quantile. The weights β^j set the relative scaling of the contribution of a particular channel to the loss. To further deal with sparsity, we add the L_1^{reg} regularization term, defined as:

$$L_1^{\text{reg}} = \sum_{i=1}^{N_{\text{params}}} |w_\theta^i| \quad (2.7)$$

where $|w_\theta^i|$ are the neural network weights, and i runs over the network's parameters.

Training of HALOgen

To train HALOgen, we use one single simulation from the CV set and test on other simulations in the same set, with different initial seeds, but same simulation parameters. Both dark matter and halo maps are divided into 64^3 sub-volumes, rotated, and augmented to get rotationally equivariant training.

We input the 64^3 voxels ρ_{DM} field in batches of 16 and apply the loss function in equation (2.5) to the predicted output, using the weighted mask in equation (2.2). The optimizer used for training is **RMSprop** (Root Mean Square Propagation), an adaptive gradient-based optimization algorithm that scales the learning rate for each parameter by a running average of its recent squared gradients, a standard technique in deep learning architectures. The learning rate schedule is set as follows: the first 10 epochs are run with a learning rate of 2×10^{-4} ; after that, we use Cosine Annealing Warm Restarts with the maximum and minimum values set to 10^{-4} and 3×10^{-5} , respectively, with a repeat cycle of 50 epochs. The cosine annealing warm restarts changes the learning rate between the maximum and minimum value, with a cosine decay, while repeating this cycle every 50 epochs in our case. We run the model for 260 epochs, and store the model’s weights at the point in which the validation loss is lowest. We subdivide the data into training and validation set, with a 75% and 25% split. In equation (2.5), the coefficient λ_w of the regularization term L_1^{reg} is set to 10^{-6} . For the results in the paper, we use $\alpha = 0.6$, $q = 0.7$ and $\beta^j = [1, 2, 5, 8]$. The above hyper-parameters have been chosen to achieve the best predictions, after extensive testing.

2.4.2 LODI: Generating 21 cm brightness temperature fields with Latent overlap diffusion

VDM: Variational Diffusion Model

A brief introduction to generative diffusion models is given in Sec. 1.3.6. Hereafter, any reference to diffusion models refers specifically to the Variational Diffusion Model (VDM; Kingma et al. 2023). In what follows, we describe the key features that distinguish the VDM from the standard Denoising Diffusion Probabilistic Model (DDPM; Ho et al. 2020), namely the continuous time formulation, the signal to noise ratio parameterization of the noise schedule, and the variational objective that enables all components including the noise schedule itself to be learned end to end.

From discrete steps to continuous time Recall from Sec. 1.3.6 that a DDPM defines a forward (noising) process over T discrete steps with a fixed variance schedule $\{\beta_t\}_{t=1}^T$, and that the closed-form forward marginal (Eq. 1.29) is

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}\right), \quad \bar{\alpha}_t \equiv \prod_{s=1}^t (1 - \beta_s). \quad (2.8)$$

The VDM generalises this by promoting the discrete index $t \in \{1, \dots, T\}$ to a continuous variable $t \in [0, 1]$. The forward process remains Gaussian:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}\right), \quad (2.9)$$

where α_t and σ_t are the continuous-time analogues of $\sqrt{\bar{\alpha}_t}$ and $\sqrt{1 - \bar{\alpha}_t}$, respectively, satisfying the variance-preserving constraint $\alpha_t^2 + \sigma_t^2 = 1$.²

Rather than specifying the noise level through a list of β values, the VDM encodes the entire schedule in a single scalar function: the *log signal-to-noise ratio* (log-SNR),

$$\gamma(t) = \log \frac{\alpha_t^2}{\sigma_t^2}, \quad (2.10)$$

which is a monotonically *decreasing* function of t : at $t = 0$ the signal dominates ($\gamma \rightarrow +\infty$), while at $t = 1$ noise dominates ($\gamma \rightarrow -\infty$). Because $\alpha_t^2 + \sigma_t^2 = 1$, the signal and noise coefficients can be recovered from $\gamma(t)$ alone:

$$\alpha_t^2 = \text{sigmoid}(\gamma(t)), \quad \sigma_t^2 = \text{sigmoid}(-\gamma(t)). \quad (2.11)$$

The entire noise schedule is therefore encoded in the single scalar function $\gamma(t)$, which replaces $\{\beta_t\}$ of the DDPM.

The variational objective The name *Variational* Diffusion Model reflects the fact that the training objective is derived rigorously as an Evidence Lower Bound (ELBO) on the log-likelihood of the data, rather than as a heuristic reweighting of a denoising loss. Recall from Sec. 1.3.6 that DDPM training uses a “simple” loss (Eq. 1.31) that drops the SNR-dependent prefactors from the true variational bound. The VDM instead retains the full ELBO structure.

²We switch from $\mathbf{x}_t$ to $\mathbf{z}_t$ for the noisy variable in the VDM to distinguish the continuous time latents from the discrete DDPM chain, and use α_t in place of $\sqrt{\bar{\alpha}_t}$ to avoid confusion with the per-step β_t notation of Sec. 1.3.6.

In the continuous-time limit ($T \rightarrow \infty$), [Kingma et al. \(2023\)](#) show that the negative ELBO decomposes into three terms:

$$-\text{ELBO} = \underbrace{\mathcal{L}_{\text{diffusion}}}_{\text{denoising quality}} + \underbrace{\mathcal{L}_{\text{latent}}}_{\text{prior matching}} + \underbrace{\mathcal{L}_{\text{recon}}}_{\text{reconstruction}}. \quad (2.12)$$

We discuss each in turn.

Diffusion loss. The diffusion loss is an integral over the continuous time variable,

$$\mathcal{L}_{\text{diffusion}} = \frac{1}{2} \int_0^1 \gamma'(t) \mathbb{E}_{q(\mathbf{z}_t | \mathbf{x}_0)} \left[\left\| \mathbf{x}_0 - \hat{\mathbf{x}}_{\theta}(\mathbf{z}_t, t) \right\|^2 \right] dt, \quad (2.13)$$

where $\hat{\mathbf{x}}_{\theta}(\mathbf{z}_t, t)$ is the model’s estimate of the clean data from the noisy input $\mathbf{z}_t$, and $\gamma'(t) = d\gamma/dt$ is the derivative of the log-SNR with respect to time. During training, the integral is estimated by sampling $t \sim \mathcal{U}(0, 1)$ and evaluating the integrand at the sampled time.

The factor $\gamma'(t)$ plays a crucial role: it naturally *weights* different noise levels according to the rate at which the SNR changes. If the schedule varies rapidly at some time t_* , meaning the model transitions quickly between signal and noise dominated regimes, the loss receives more weight there. This is fundamentally different from the DDPM, where the “simple” loss of [Ho et al. \(2020\)](#) drops all SNR-dependent prefactors, treating every noise level equally.

An equivalent and widely used form of Eq. (2.13) is the *noise prediction* parameterisation. If the network predicts the noise $\hat{\epsilon}_{\theta}(\mathbf{z}_t, t)$ instead of the clean signal, the two are related by the identity,

$$\hat{\mathbf{x}}_{\theta} = \frac{\mathbf{z}_t - \sigma_t \hat{\epsilon}_{\theta}}{\alpha_t}, \quad (2.14)$$

This identity also forms the backbone of our work in chapter 5, as it connects the estimated signal $\hat{\mathbf{x}}_{\theta}$ to the predicted noise $\hat{\epsilon}_{\theta}$ by the model.

Latent loss. The latent loss measures how close the fully noised distribution $q(\mathbf{z}_1 | \mathbf{x}_0)$ is to the standard Gaussian prior $p(\mathbf{z}_1) = \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathcal{L}_{\text{latent}} = D_{\text{KL}} \left(q(\mathbf{z}_1 | \mathbf{x}_0) \parallel p(\mathbf{z}_1) \right). \quad (2.15)$$

Since both distributions are Gaussian, the KL divergence is available in closed form. For a well-behaved schedule with $\gamma(1) \ll 0$ (i.e. the endpoint SNR is very low), $q(\mathbf{z}_1|\mathbf{x}_0)$ is already close to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and this term is small.

Reconstruction loss. The reconstruction loss accounts for the final decoding step from $\mathbf{z}_0$ to the data $\mathbf{x}_0$:

$$\mathcal{L}_{\text{recon}} = -\mathbb{E}_{q(\mathbf{z}_0|\mathbf{x}_0)} \left[\log p_{\theta}(\mathbf{x}_0 | \mathbf{z}_0) \right]. \quad (2.16)$$

In practice, this takes the form of a Gaussian negative log-likelihood evaluated at $\mathbf{z}_0$, which is a very lightly noised version of $\mathbf{x}_0$ (since $\gamma(0) \gg 0$).

Total loss. The total training loss is a weighted sum,

$$\mathcal{L} = \lambda_{\text{diff}} \mathcal{L}_{\text{diffusion}} + \lambda_{\text{lat}} \mathcal{L}_{\text{latent}} + \lambda_{\text{rec}} \mathcal{L}_{\text{recon}}, \quad (2.17)$$

where $\lambda_{\text{diff}}, \lambda_{\text{lat}}, \lambda_{\text{rec}}$ are scalar coefficients. Setting all three to unity recovers the exact negative ELBO, providing a principled lower bound on $\log p(\mathbf{x}_0)$.

Learned noise schedule A key advantage of the VDM framework is that the noise schedule $\gamma(t)$ can be *learned* jointly with the denoising network $\hat{\mathbf{x}}_{\theta}$. Since $\gamma'(t)$ appears as a weighting factor in the diffusion loss (Eq. 2.13), the shape of the schedule controls which noise levels receive the most training signal, while the endpoint values $\gamma(0)$ and $\gamma(1)$ set the SNR range over which the model operates. For this chapter, we adopt a *learned linear schedule*:

$$\gamma(t) = b + |w|t, \quad (2.18)$$

where b and w are learnable scalar parameters, and $|w|$ enforces monotonicity.³ This is to be contrasted with the DDPM, where the schedule $\{\beta_t\}$ is a hyperparameter (typically linear or cosine) that must be manually tuned.

Sampling Sampling from the VDM proceeds by discretising the continuous reverse process into N_{steps} steps. Starting from $\mathbf{z}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, we iterate backwards through a

³Since γ must decrease with t , the sign convention is chosen such that $w < 0$ and $|w|$ ensures the slope is negative. Equivalently, some implementations write $\gamma(t) = \gamma_{\text{max}} + (\gamma_{\text{min}} - \gamma_{\text{max}})t$ with $\gamma_{\text{max}} > \gamma_{\text{min}}$ by construction.

set of times $1 = t_N > t_{N-1} > \dots > t_0 = 0$:

$$\mathbf{z}_{t_{i-1}} = \frac{\alpha_{t_{i-1}}}{\alpha_{t_i}} \left(\mathbf{z}_{t_i} + \sigma_{t_i}^2 \nabla_{\mathbf{z}} \log p(\mathbf{z}_{t_i}) \right) + \tilde{\sigma}_i \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2.19)$$

where $\nabla_{\mathbf{z}} \log p(\mathbf{z}_{t_i})$ is the score function, estimated from the network’s noise prediction via $\hat{\boldsymbol{\epsilon}}_{\theta}$, and $\tilde{\sigma}_i$ is the posterior standard deviation determined by the SNR values at t_i and t_{i-1} .

Crucially, the number of sampling steps N_{steps} is a *sampling-time* hyperparameter that is independent of training. One can train the model once and then adjust N_{steps} at inference to trade off sample quality for computational cost. This decoupling is a direct consequence of the continuous time formulation: the model learns a denoiser valid for all $t \in [0, 1]$, and the sampling procedure merely queries it at a discrete subset of times.

Connection to score matching and the forward KL The diffusion loss (Eq. 2.13) is closely related to *denoising score matching*. The score of the noisy marginal, $\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t)$, can be expressed in terms of the noise as

$$\nabla_{\mathbf{z}_t} \log q(\mathbf{z}_t) = -\frac{\boldsymbol{\epsilon}}{\sigma_t}, \quad (2.20)$$

so training the network to predict $\boldsymbol{\epsilon}$ is equivalent to training it to estimate the score. Since the score can be rewritten in terms of the noise via Eq. 2.20, and assuming the network provides a good approximation $\hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{z}_t, t) \approx \boldsymbol{\epsilon}$, the reverse sampling step (Eq. 2.19) takes the form

$$\mathbf{z}_s = \frac{\alpha_s}{\alpha_t} (\mathbf{z}_t - \sigma_t c \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{z}_t, t)) + \sqrt{(1 - \alpha_s^2) c} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (2.21)$$

where $c = \exp(\gamma(s) - \gamma(t) - 1)$ and $0 < s < t \leq 1$. This is the form we use in practice for ancestral sampling from the trained VDM.

Finally, it is worth noting that since the forward diffusion process $q(\mathbf{z}_t | \mathbf{x}_0)$ is fixed and data defined while the reverse process $p_{\theta}(\mathbf{z}_{t-1} | \mathbf{z}_t)$ is learned, the ELBO objective corresponds to minimizing the *forward* KL divergence $D_{\text{KL}}(q \| p_{\theta})$ the model is trained to cover all modes of the data distribution.

Network architecture

Denoising network architecture The denoising network $\hat{\epsilon}_\theta(\mathbf{z}_t, \boldsymbol{\tau}^t, \mathbf{x}_{\text{cond}})$ is a 3D U-Net, adapted from the 2D architecture of Ono et al. (2024) by replacing all 2D grouped convolutions with their 3D counterparts and adding an extra convolutional layer to each residual block to increase the network’s capacity for volumetric data.

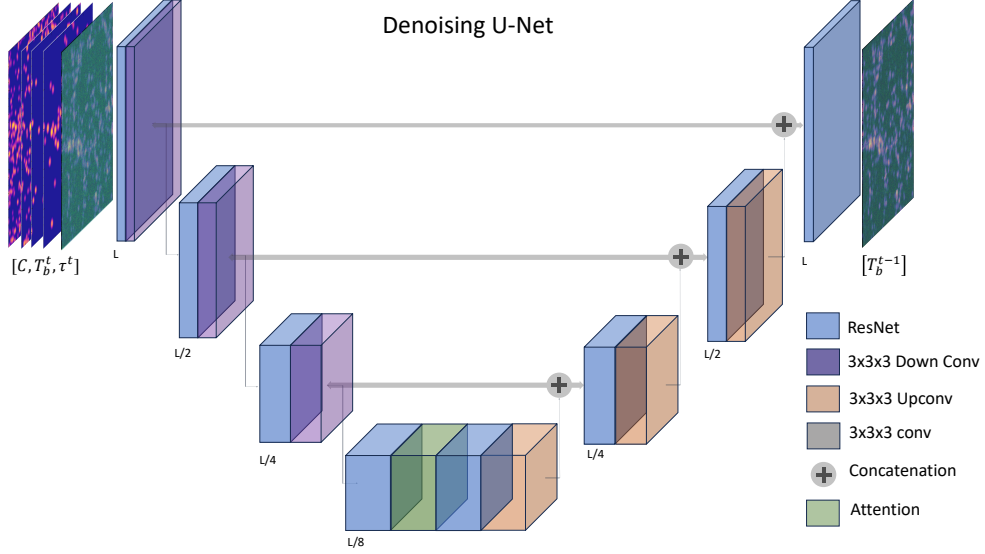


Figure 2.4: The architecture of the denoising model: a 3D U-net with residual network blocks and an attention layer at the bottleneck. The input to the network is the noisy temperature map at time step t , the time step embedding and the conditional halo maps. The output is a less noisy version of the temperature map.

The encoder decoder structure consists of 4 downsampling and 4 upsampling stages, each built from residual convolutional blocks with skip connections between corresponding encoder and decoder levels. At the bottleneck, we include a self-attention block, which has been shown to facilitate faster convergence and capture both global and short-range correlations in the data (Ono et al. 2024). The full architecture is illustrated in Fig. 2.4, while the layout of a residual block is depicted in Fig. 2.5. As mentioned, each convolution layer is a 3 dimensional convolution with learnable parameters, and is followed by a SiLU non linearity (See eq: 1.17).

The `net1` first extracts features from input vector $\mathbf{x}$, to which the time step embedding $\boldsymbol{\tau}^t$ is added via a learned linear projection so that the dimension of $\boldsymbol{\tau}^t$ matches with the extracted features. After this, a zero initialized convolution layer mixes this quantity. The final vector $\mathbf{h}$ produced is added to the input $\mathbf{x}$ and passed on to a downsampling

block that reduces the spatial dimensionality of the output by 2 in each of the three axes.

Note that we use group normalization instead of the more commonly used batch normalization. The reason is because our batches (number of training samples the gradient descent algorithm averages over per step) are quite small, and batch normalization can get quite noisy with low batch sizes.

The network receives three inputs at every reverse step: the noised field $\mathbf{z}_t$ (in this specific case it is $\mathbf{T}_b^t$, the noisy brightness temperature map), the conditioning field $\mathbf{x}_{\text{cond}}$ (concatenated channel wise with $\mathbf{z}_t$), and a sinusoidal time-step embedding $\boldsymbol{\tau}^t$ that informs the network of the current noise level. The embedding is defined as

$$\begin{aligned}\boldsymbol{\tau}^t &= (\sin(\tilde{t}\boldsymbol{\omega}), \cos(\tilde{t}\boldsymbol{\omega})), \quad \tilde{t} = 1000t, \quad t \in [0, 1], \\ \boldsymbol{\omega} &= (\omega_k)_{k=0}^{d/2-1}, \quad \omega_k = \exp\left(-\log(\tau_{\max}) \frac{k}{d/2-1}\right),\end{aligned}\tag{2.22}$$

where $\tau_{\max} = 10^4$ and d is the embedding dimension. The factor of 1000 stretches the unit interval $t \in [0, 1]$ to a range where the sinusoidal frequencies are well-separated. We use $d = 64$ throughout this work. The time embedding is injected into every residual block, allowing the network to modulate its behaviour as a function of the noise level.

Building the 21 cm temperature training map

To produce the 21 cm temperature map for training, we proceed as follows: we obtain the neutral hydrogen fraction $x_{\text{HI},i}$ and the gas particle mass $m_{g,i} [\text{M}_{\odot}]$, for each gas particle i in the hydrodynamic simulation, from which we first compute the HI mass per particle, $\text{M}_{\text{HI},i}$, as:

$$\text{M}_{\text{HI},i} = m_{g,i} x_{\text{HI},i} X, \tag{2.23}$$

where X is the hydrogen mass fraction, which in this CAMELS simulations is set to 0.76. In addition, since the simulations (as detailed below) do not include self-shielding effects for star-forming particles, we manually set $x_{\text{HI},i} = 1$ for particles with star formation rate larger than zero. The reason for having a separate star formation rate parameter on top of the star particles is that in the underlying multiphase ISM model, self-shielded gas is represented as a gas element until stochastically converted to a star particle, where the conversion probability depends on the star formation timescale (Springel & Hernquist 2003). This simple and effective prescription, also employed in the post-

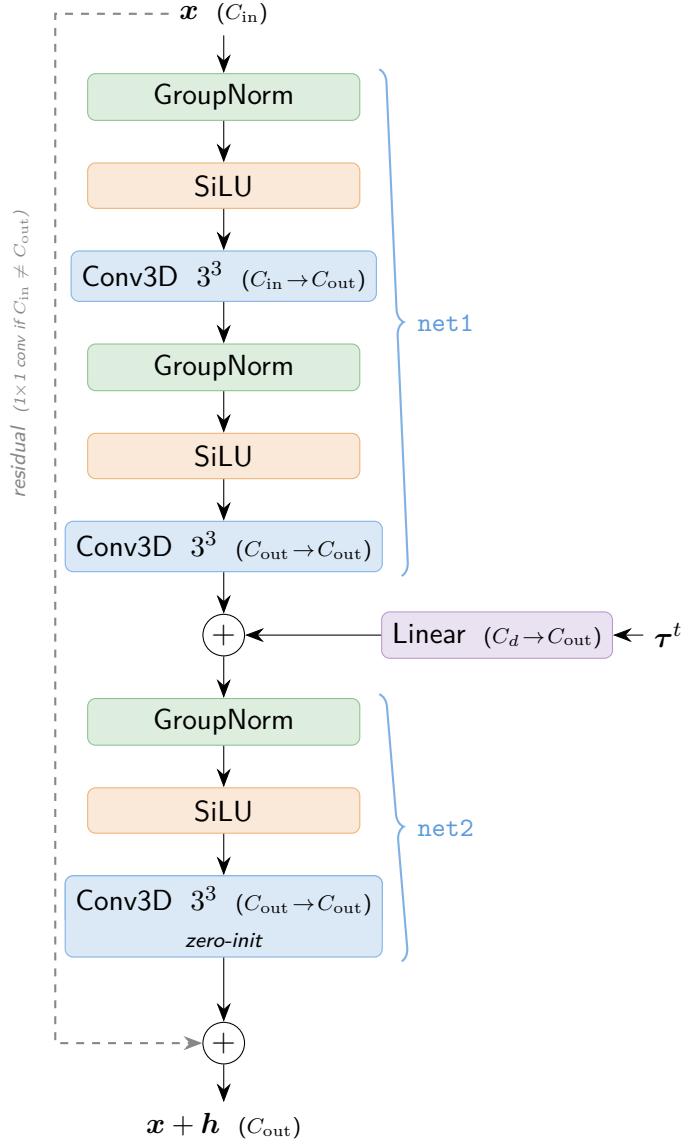


Figure 2.5: Architecture of a single residual block. The block consists of two sub-networks (**net1** and **net2**), with the time-step embedding τ^t injected additively between them via a learned linear projection. A residual skip connection (dashed) bypasses the main path; when the input and output channel dimensions differ, a 1×1 convolution is applied along the skip path. The final convolution in **net2** is zero-initialized so that each block acts as an identity at the start of training.

processing of the CAMELS simulations (Villaescusa-Navarro et al. 2022), assumes that all the star-forming particles are fully self-shielded against ionizing radiation. As shown in Villaescusa-Navarro et al. (2018a), this assumption results in a good match to observational data sets like abundance of HI and column density distribution of Damped

Lyman- α systems.

We then use the same CIC procedure described above to interpolate $M_{\text{HI},i}$ to a continuous mass field, M_{HI} , after which we divide each voxel by its volume to obtain the density field ρ_{HI} . The 21 cm brightness temperature in the post-reionization universe is given to a good approximation by the following expression in [Villaescusa-Navarro et al. \(2018a\)](#):

$$T_b(\mathbf{x}) = 189 h \left(\frac{H_0(1+z)^2}{H(z)} \right) \frac{\rho_{\text{HI}}(\mathbf{x})}{\rho_c} [\text{mK}], \quad (2.24)$$

where H_0 is the Hubble Lemaître constant today, $\rho_{\text{HI}}(\mathbf{x})$ is the interpolated neutral hydrogen density at cell location $\mathbf{x}$, z is the redshift, $H(z)$ is the Hubble function at redshift z , T_b is expressed in mK, and $H_0 = 100 h \text{ km/s/Mpc}$. The quantity $\rho_c = \frac{3H_0^2}{8\pi G}$ denotes the critical density today.

We will work in the plane parallel approximation and assume the redshift of the box as a constant (in future work we plan to extend this to redshift varying light cones), given the fact that redshift evolution within the simulated volume would require a realistic set of lightcones mocks and this is beyond the scope of the paper. Once the maps are created, we smooth them with a Gaussian kernel of fixed radius $R = 0.2 \text{ Mpc } h^{-1}$. This choice balances two considerations. On the one hand, smoothing helps to stabilize training by reducing sharp fluctuations and noise in the target fields; on the other hand, excessive smoothing would wash out the small scale structures that are physically important for our task. Given that the voxel resolution is $\approx 0.1 h^{-1} \text{ Mpc}$, we chose a slightly larger smoothing radius that facilitates improved training of the network while preserving most of the relevant small scale information.

Loss function

When training LODI, we sample a 21 cm brightness-temperature map $\mathbf{T}_b$ from the training set and apply the forward diffusion process (Eq. 2.9), setting $\mathbf{x}_0 = \mathbf{T}_b$ and $\mathbf{z}_t = \mathbf{T}_b^t$. We use $T = 50$ timesteps.

During training, the denoising network predicts the noise by minimizing Eq. 2.12. For example, following this same convention, the diffusion loss 2.13 gets modified to:

$$\mathcal{L}(\mathbf{T}_b) = \frac{T}{2} \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), t \sim \mathcal{U}[1, T]} \left[\left(\exp(\gamma(t-1) - \gamma(t)) - 1 \right) \left\| \epsilon - \hat{\epsilon}_\theta(\mathbf{T}_b^t, t) \right\|_2^2 \right], \quad (2.25)$$

where the expectation is over the noise distribution and the uniform discrete time distribution.

During generation, we sample from $\mathbf{T}_b^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and iteratively denoise using the learned distribution $p_\theta(\mathbf{T}_b^{t-1} \mid \mathbf{T}_b^t, \boldsymbol{\rho}_{\text{DM}}, t)$ to produce samples $\mathbf{T}_b^0 \sim p(\mathbf{T}_b \mid \boldsymbol{\rho}_{\text{DM}})$. The ancestral sampling (Eq. 2.21) therefore turns into :

$$\mathbf{T}_b^s = \frac{\alpha_s}{\alpha_t} \left(\mathbf{T}_b^t - \sigma_t c \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{T}_b^t, t) \right) + \sqrt{(1 - \alpha_s^2)} c \boldsymbol{\epsilon}, \quad (2.26)$$

Training of LODI We train LODI on 5 simulations of the CV set from CAMELS, each with a different initial seed. We process the 5 simulations as explained above, and divide them into 32^3 voxel sub-volumes. To enforce rotational equivariance, we augment each simulation cube with all symmetry-equivalent flips and 90° rotations. Training on more simulations is necessary to account for cosmic variance in the large scales comparable to the box size and also to train on more halos of large masses, especially $M_{\text{halo}} > 10^{12} M_\odot$. To compress the dynamic range of the sparse HI field, we train the network in logarithmic space of the fields. This data augmentation expands the raw training set substantially and would correspondingly increase training time. We therefore apply a targeted sub sampling strategy: because HI emission is concentrated around dark matter halos particularly massive ones and thus sparse, we preferentially draw sub-volumes that contain objects in the higher mass halo channels. We define important regions here as regions where $\log \rho_h \geq 10$, since this density value is about 10^3 times smaller than the minimum density value in our dataset, prior to gaussian smoothing. Values smaller than that are small enough that they do not contribute to any meaningful HI content. To maintain diversity we also include 50% of sub-volumes centered on the more numerous lower-mass halos, the first 2 mass bins in this case. This balanced selection reduces the total training data sets from $\approx 410\,000$ to $\approx 250\,000$, almost halving the training time without degrading the representativeness of the data, while it increases the relative importance of the larger mass halos as it is seen more frequently by the neural network. We finally create a training and validation split of 75% and 25%, respectively.

2.4.3 Generating larger volumes with latent overlap

It is not possible to input the entire $25^3 (h^{-1} \text{ Mpc})^3$ simulation box, corresponding to 256^3 voxels to our network, because of memory constraints. Such data cube would require several terabytes of GPU memory to train. Instead, we divide the domain into 512 smaller 32^3 voxels sub-volumes, with the view of generating for each its 21 cm field

with the trained the diffusion model, and then combine them at the end to generate the full volume.

Latent overlap method

The naive solution would be to generate the brightness temperature field in each 32^3 sub-volume, then tile them to create the target, 256^3 volume we call this approach ‘tiling’. However, tiling introduces artificial, periodic discontinuities at the boundaries of each 32^3 voxel cube, arising because the brightness temperature field $\mathbf{T}_b$ is sampled from the conditional distribution $\mathbf{T}_b \sim p(\mathbf{T}_b|\boldsymbol{\rho}_h)$, which depends solely on the 32^3 halo field $\boldsymbol{\rho}_h$, with no reference to adjacent boxes.

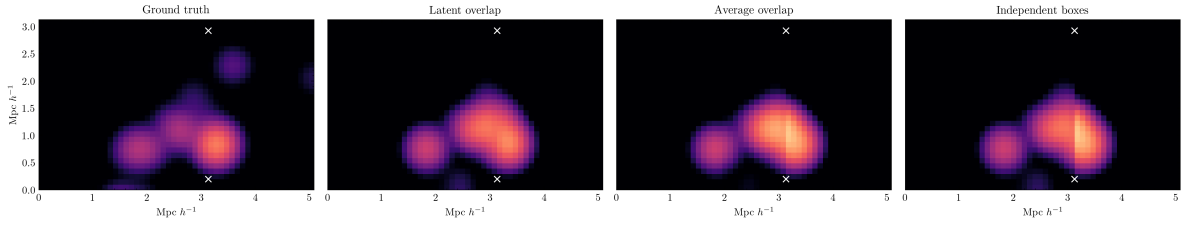


Figure 2.6: A comparison of different overlapping methods are shown, where the line of discontinuity is vertical and passes through the 2 marked crosses in the images. The left image is the ground truth. The latent overlap method is shown in the left center, compared with two other cases. The center right is the case where for the overlap region, we take the average pixel values of the two boxes and the right image shows the other case when we generate 2 separate boxes and just concatenate them, while keeping the the overlapping pixel region from the first box. The images shown are averaged in the third axis over a distance of about $0.9 h^{-1} \text{ Mpc}$

To address this, we adapt the inpainting method of [Lugmayr et al. \(2022\)](#), which in our work we call ‘latent overlap’ approach. The first box, B_1 , is generated by starting with a Gaussian noise ϵ_1 , to obtain a sample $\mathbf{T}_{B_1}^0$. While generating this box in $T = 50$ steps, a set of T intermediate latent maps are produced using Eq. (2.26): $[\mathbf{T}_{B_1}^1, \mathbf{T}_{B_1}^2, \dots, \mathbf{T}_{B_1}^T]$. These intermediate latent images can be interpreted as discrete snapshots of the evolving $\mathbf{T}_b$ distribution under the diffusion process, forming a trajectory from random noise to the final sampled field.

When we next consider an adjacent box, B_2 , we want to enforce that its generation process follows a ‘similar’ trajectory as B_1 . For this, we partially overlap boxes B_2 onto B_1 such that

$$\text{Vol}(B_1 \cap B_2) = V_{\text{overlap}}$$

where V_{overlap} is the overlapping volume and is defined as the fraction of one box (specified below). For our work, the overlap volume or fraction is specified in section 2.4.2. Now, at every t -th denoising step of B_2 , we replace $\mathbf{T}_{B_2}^t$ with the previously-generated $\mathbf{T}_{B_1}^t$ in the region V_{overlap} , thus obtaining $\mathbf{T}_{B_2}^{t*}$, which is identical to $\mathbf{T}_{B_1}^t$ in the overlap volume. This new $\mathbf{T}_{B_2}^{t*}$ is then denoised using the denoising U-net to obtain $\mathbf{T}_{B_2}^{t-1}$. Mathematically, the updated state of B_2 at the t -th denoising step is given by:

$$\mathbf{T}_{B_2}^{t*} = M_{\text{overlap}} \mathbf{T}_{B_2}^t + (1 - M_{\text{overlap}}) \mathbf{T}_{B_1}^t,$$

where M_{overlap} is the overlap mask defined as:

$$M_{\text{overlap}}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in V_{\text{overlap}} \\ 0, & \text{otherwise} \end{cases}$$

and

$$\mathbf{T}_{B_1}^t = \alpha_t \mathbf{T}_{B_1}^0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where α_i and σ_i are the noise schedule parameters and are learnt by the diffusion model as explained in section 2.4.2, and $\mathbf{T}_{B_1}^0$ was produced by the diffusion model in a previous generative pass.

Performing this latent overlap step ensures that the entire image $\mathbf{T}_{B_2}^{i*}$ is denoised harmoniously, thereby eliminating boundary artifacts, while producing a non-overlapping region that fits seamlessly with the overlapping region. We carry out this procedure T times, and at each step B_2 is pushed closer to the distribution of B_1 by forcing it to align with the overlapping region of B_1 while moving along the trajectory of $\mathbf{T}_{B_1}^i$. This latent overlap method between boxes B_1 and B_2 is shown in algorithm 1, while figure 2.6 provides a comparison of the results obtained with latent overlap against simple tiling of independent boxes. In principle, as T is increased, the alignment should improve. Figure 2.7 shows the diffusion process implementing this algorithm on a 2D patch of overlapping boxes. In this case, the left box has already been produced and does not change throughout the entire process⁴. We call this approach LODI, for Latent Overlap Diffusion for Intensity mapping.

⁴Note that left half of the image does not change during the process, despite the misleading impression given by the colorbar boundaries being adjusted from left to right to accommodate the different dynamic range of pixel values.

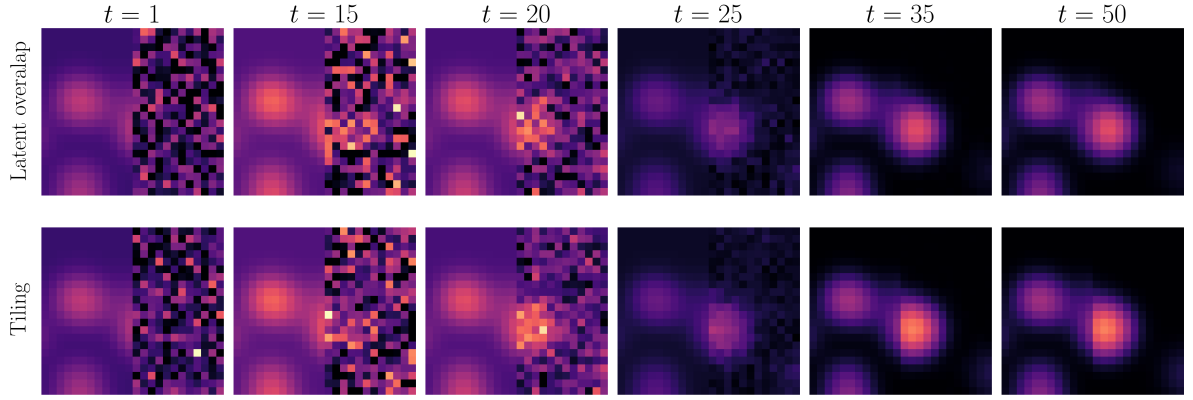


Figure 2.7: Illustration of LODI, using the latent overlap method: output of the diffusion process at different steps in the denoising process (top row, left to right) when using the latent overlap method on a 2D illustrative slice. At each time step, the left half of the field has already been generated. The bottom row shows the same time steps but with simple tiling and no latent overlap: the discontinuity at the boundary is apparent.

Algorithm 1 Latent overlap method

```

1:  $\mathbf{T}_{B_2}^T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2:  $\boldsymbol{\epsilon}_{B_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
3: Calculate  $M_{\text{overlap}}$ 
4: for  $i = T - 1, \dots, 0$  do
5:    $s(i) = (i)$  and  $t(i) = i + 1$ 
6:    $\mathbf{T}_{B_1}^t = \alpha^t \mathbf{T}_{B_1}^0 + \sigma^t \boldsymbol{\epsilon}_{B_1}$ 
7:    $\mathbf{T}_{B_2}^{t*} = M_{\text{overlap}} \mathbf{T}_{B_2}^t + (1 - M_{\text{overlap}}) \mathbf{T}_{B_1}^t$ 
8:    $\boldsymbol{\epsilon}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
9:    $\mathbf{T}_{B_2}^s = \frac{\alpha_s}{\alpha_t} \left( \mathbf{T}_{B_2}^{t*} - \sigma_t c \hat{\boldsymbol{\epsilon}}_\theta(\mathbf{T}_{B_2}^{t*}; t) \right) + \sqrt{(1 - \alpha_s^2) c} \boldsymbol{\epsilon}_2$ 
10: end for
11: return  $\mathbf{T}_{B_2}^0$ 

```

Patching 21 cm fields

To generate the entire $25^3 (h^{-1} \text{ Mpc})^3$ volume, we need to define the sequence in which the sub-volumes B_i cubes are processed. Ideally, we aim for an algorithm that enables parallelization of the process, for faster generation, while using the largest possible number of starting sub-volumes (i.e., the ones that are generated conditionally on the learnt diffusion process, but without reference to adjacent regions), because these are the ones that are truly in-distribution w.r.t. to the diffusion generative process.

To execute this, we create a mesh of 125 sub-volumes B_i of size 32^3 voxels, equally spaced from each other, arranged in a 3D grid (Stage 1, top panels in Fig. 2.8, showing

from left to right the three orthogonal planes, $X - Y$, $X - Z$, $Y - Z$; maps are averaged along the third axis). In Stage 2, 3 and 4, we generate nearest neighbor boxes along each of the 3 cartesian direction using the latent generation method with an overlap of $4 \times 32 \times 32$ voxels. This greatly reduces the out of distribution error, and helps with parallelizing the generation process since most of the boxes can be generated independently of each other. Consequently, the configuration in Stage 4 exhibits an alternating pattern where every other B_i is present along each axis. In each of these 3 stages, 100 new B_i s are generated per axis, amounting to 300 new generations in total.

In the subsequent Stages (5, 6 and 7), additional sub-volumes are generated adjacent to the ones generated in stages 2,3 and 4 along the cartesian axes, overlapping with the original sub-volume with the same overlap fraction of $4 \times 32 \times 32$. This procedure is performed sequentially for the 3 cartesian directions, requiring 80 cubes per axis (240 in total). Finally, the remaining 64 sub-volumes are generated at stage 8, each of which overlaps with the neighboring 6 voxels with the same overlap volume of $4 \times 32 \times 32$ to complete the configuration, resulting in a final assembly of 729 B_i s arranged in a $9 \times 9 \times 9$ grid resulting in a total volume of $25 (h^{-1} \text{ Mpc})^3$. This boosts the speed of generation by a factor of ≈ 6 , compared to generating sub-volumes sequentially. In principle we could further parallelize the generation process by combining the different stages into one, but this is limited by the available GPU memory.

The overall output of our generative pipeline is shown in Fig. 2.9. We start with an unseen dark matter density field, shown in the top left panel, run it through HALOgen (whose intermediate halo mass density maps are shown in the second column of figure 2.10), and then produce the brightness temperature map with LODI (bottom right panel of Fig. 2.9). Using the naive Tiling approach (bottom left panel of Fig. 2.9), the discontinuity between training sub-volumes is evident (corners of boundaries are indicated by green crosses for better visualization).

2.5 Results

In this section, we demonstrate the performance of the entire generative pipeline, producing HI intensity maps from previously unseen dark matter simulation maps. We validate our results using the CV set of the IllustrisTNG simulations at $z = 0$, with the same parameters for the simulations as the training maps, but with different initial seeds. We postpone the study of generalization capabilities to other cosmologies to a future, dedicated work.

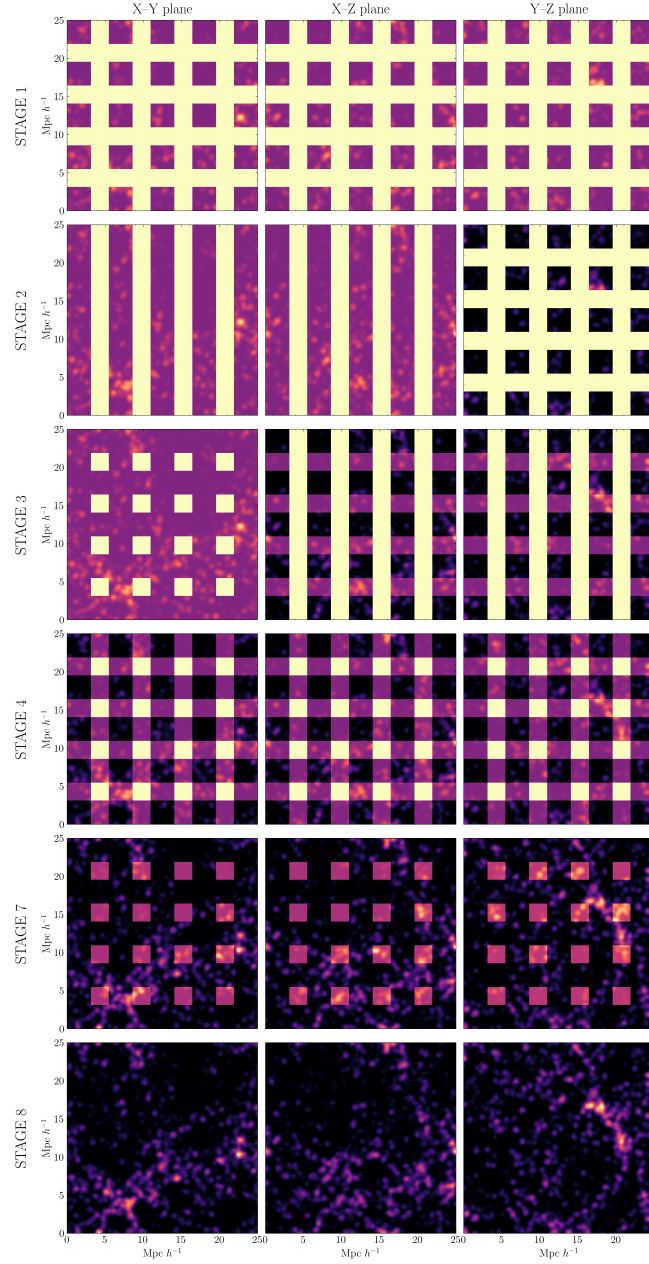


Figure 2.8: Assembly of the full volume cube of voxel size 256^3 from $9 \times 9 \times 9$ sub-volumes using the latent overlap method, for a simulation from the validation set. Each row corresponds to a different stage (top to bottom) of the latent generation process and each column displays one of the three orthogonal 3D cartesian planes of the simulation volume (the maps are shown by projecting along the third spatial axis across the full $25 h^{-1}$ Mpc extent of the simulation box). Although the figure displays six rows, the complete process consists of eight sequential steps. The fifth row represents the net result of three consecutive sub-steps, which have been aggregated for clarity.

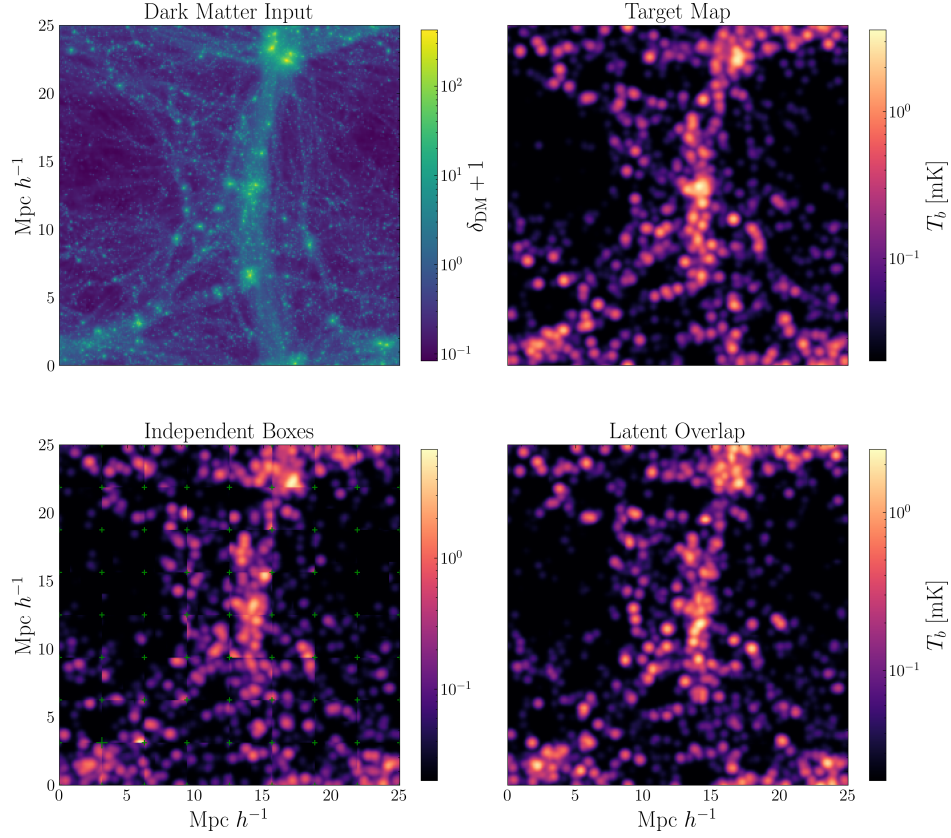


Figure 2.9: Top left: The dark matter density map, not seen during training, used as input for the generative pipeline. Top right: the ground truth (target) HI brightness temperature map from the simulation. Bottom left: The generated output using the Tiling method, with green markers highlighting the sub-volumes boundaries where discontinuities arise. Bottom right: brightness temperature map generated using the latent overlap method of LODI, which significantly reduces the discontinuities seen in the tiling approach. The brightness temperature map is conditioned on the output provided by HALOgen. All the maps are shown by projecting along the third spatial axis across the full $25 h^{-1}$ Mpc extent of the simulation box.

2.5.1 Performance of HALOgen

The output of HALOgen is shown in figure 2.10, split in the four mass channels (top to bottom), with the input dark matter map shown on the top left of figure 2.9. We reiterate that this input map was unseen during training, as it was generated with a different seed (but same cosmological and astrophysical parameters as the training map).

Visually, the predicted pixel values for the halo mass density (left column) look very close to the target maps (center column), except for the smallest mass range (first row). This can be more precisely quantified by comparing the distribution of log-density

values between the target and predicted $25^3 (h^{-1} \text{Mpc})^3$ box (right column). While we observe a consistent under-prediction in the mean value of the smallest halo mass range (albeit still within the 16 – 84 percentile band), the agreement for the other mass ranges is excellent. We trace back the relatively poorer performance for the smallest mass halo density to our definition of masked loss: as discussed above, a lower weight is given to pixels of the smallest halos, since they are the most abundant ones. Therefore, we can expect that the loss function puts greater attention to faithfully reconstructing the density of larger mass halos. We plot the median instead of the mean since in the post-processing of the output halo maps, we set pixel values below a threshold of $\log \rho_h = 9.5$ to 0, which in the pixel comparison plot skews the mean value of the pixel distribution. The median values are agnostic to these outlier near zero values. The reason for setting this cutoff is because such low predicted values are not possible given the training data and lie below the lower density limit (given voxel of side of $\approx 0.1 h^{-1} \text{Mpc}$ this threshold would correspond to a mass lower limit of $10^{6.5} M_\odot$, much smaller than what we are working with, and so zero for all practical purpose).

2.5.2 Performance of LODI

We define the dimensionless brightness-temperature field by normalizing each spatial point by its global mean:

$$\hat{T}_b(\mathbf{x}) = \frac{T_b(\mathbf{x})}{\langle T_b \rangle}. \quad (2.27)$$

The dimensionless power spectrum is accordingly defined as:

$$P_{\hat{T}_b}(k) = \langle \hat{T}_b(\mathbf{k}) \cdot \hat{T}_b^*(\mathbf{k}) \rangle, \quad (2.28)$$

where $\hat{T}_b(\mathbf{k})$ is the Fourier transform of $\hat{T}_b(\mathbf{x})$ and $k = |\mathbf{k}|$. This dimensionless formulation allows direct comparison of small-scale fluctuations across different simulations.

To assess the performance of LODI, we compute the 21 cm power spectrum produced by our pipeline and compare it to the ground truth via the residual:

$$\Delta P_{\hat{T}_b}(k) = P_{\hat{T}_b}^{\text{gen}}(k) - P_{\hat{T}_b}^{\text{true}}(k), \quad (2.29)$$

where $P_{\hat{T}_b}^{\text{gen}}(k)$ and $P_{\hat{T}_b}^{\text{true}}(k)$ denote the LODI-generated and ground-truth simulation power spectra, respectively. Since the diffusion model was trained on Gaussian smoothed maps, the comparison plots and power spectra calculations are also done by Gaussian

smoothing the ground truth maps with the same filter radius of $R = 0.2 \ h^{-1} \text{ Mpc}$. Note that when testing our model, we compare against maps that include HI from all the halos *across the full halo-HI mass range*. This allows us to verify that the specific halo mass threshold used during training does not significantly affect the recovered 21 cm signal.

Figure 2.12 illustrates the end to end performance by showing the residual power spectrum obtained by applying the whole pipeline on 7 previously unseen dark matter density fields (i.e., with different initial seeds). For each of the 7 dark matter fields, we first produce the halo maps for each using **HALOgen**, and then generate 5 realizations of brightness temperature maps.

In Figure 2.11, we present an example T_b map generated using the full pipeline, alongside its corresponding ground truth map, unseen before during training, each averaged along the third axis. We also plot the dimensional power spectrum $P_{\hat{T}_b}$, computed from 20 LODI realizations conditioned on the same halo maps produced by **HALOgen**. Visually, the generated and true maps show considerable matching, and the power spectra remain in good agreement up to $k \simeq 10 \ h \text{ Mpc}^{-1}$, as can also be seen in the residual values shown in the bottom panel. This result showcases the intrinsic variance of the generative model LODI, as seen in the spread of the power spectra residuals. At high wavenumbers the predicted power spectrum exhibits excess power relative to the truth. This arises because the generated 21 cm maps display more compact radial profiles than the true maps, which are comparatively more extended and diffuse. Since LODI is conditioned on halo maps only, the network tends to concentrate 21 cm emission within halo regions, producing higher central intensities and a steeper radial decline. This enhanced contrast amplifies small scale variance and leads to the observed excess power at higher wave modes.

2.5.3 Computational cost

To generate these results, **HALOgen** was trained for about 21 hours on a single A100 GPU. The training of LODI took 600,000 iterations, using a batch size of 16. The total training time was around 30 hours, using a single A100 GPU. When generating, the total time taken by the pipeline to produce one temperature map for the whole 256^3 volume starting from the dark matter field is 106 seconds (16 seconds for **HALOgen**, 90 seconds for LODI).

2.6 Conclusions

In this work, we presented a two-steps generative pipeline to generate and inpaint large, 3 dimensional 21 cm intensity maps, starting from dark matter only simulations. Although this work applied the method to $25^3 (h^{-1} \text{ Mpc})^3$ simulations, it can be extended to inpaint simulations of any size, provided the resolution is the same as the training set simulation. In the first step, the **HALOgen** algorithm uses a U-Net to produce high-fidelity halo maps, subdivided in mass channels (bins); in the second step, **LODI** uses a conditional variational diffusion model coupled with a latent overlap method to paint the 21 cm signal on top of the halo field. We also developed a method to parallelize the generation of sub-volumes of intensity maps coherently into a larger volume simulation, without having to modify the diffusion model. Thus, this method is agnostic to the type of diffusion model used. A more powerful and general method is being investigated and will be demonstrated in a dedicated future work.

To create the halo maps, we use an attention ResUnet along with a weighted masking technique, to enforce halo separation based on halo mass. For this, we create 4 halo maps from a single dark matter density field, which we then use as a conditional in our diffusion model, a variational diffusion model that learns to denoise a Gaussian field, conditioned on the 4 halo maps, to produce the 21 cm intensity maps at redshift $z = 0$.

The pipeline produces statistically unbiased estimates of the 21 cm power spectrum, and achieves a $\leq 10\%$ accuracy on average, up to $k \simeq 10 h \text{ Mpc}^{-1}$, as can be seen in figure 2.12. We test the model on multiple unseen simulations, with different initial seeds, showcasing the robustness of the model to capture and reproduce features well into the non-linear small-scale regime.

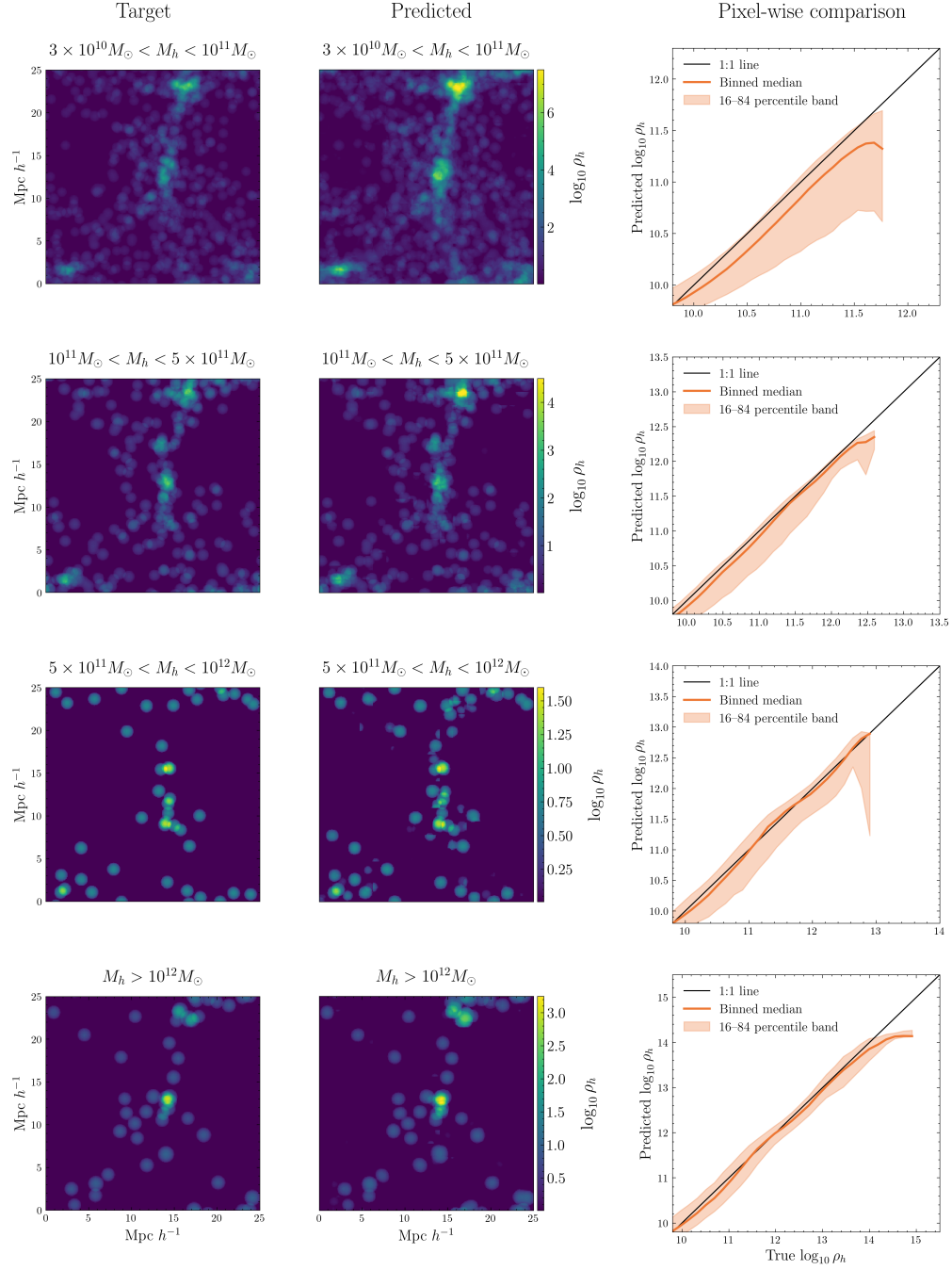


Figure 2.10: Output of HALOgen from a previously unseen dark matter simulation. The rows show the comparison between the target (left) and predicted (center) halo mass density maps for each of the four mass channels (top to bottom). The halo maps are visualized by projecting along the third spatial axis across the entire $25 h^{-1}$ Mpc of the simulation. The right column gives a pixel-wise comparison between the distribution of log-densities for the target and predicted map, with the mean and 1σ band computed from the 256^3 voxels.

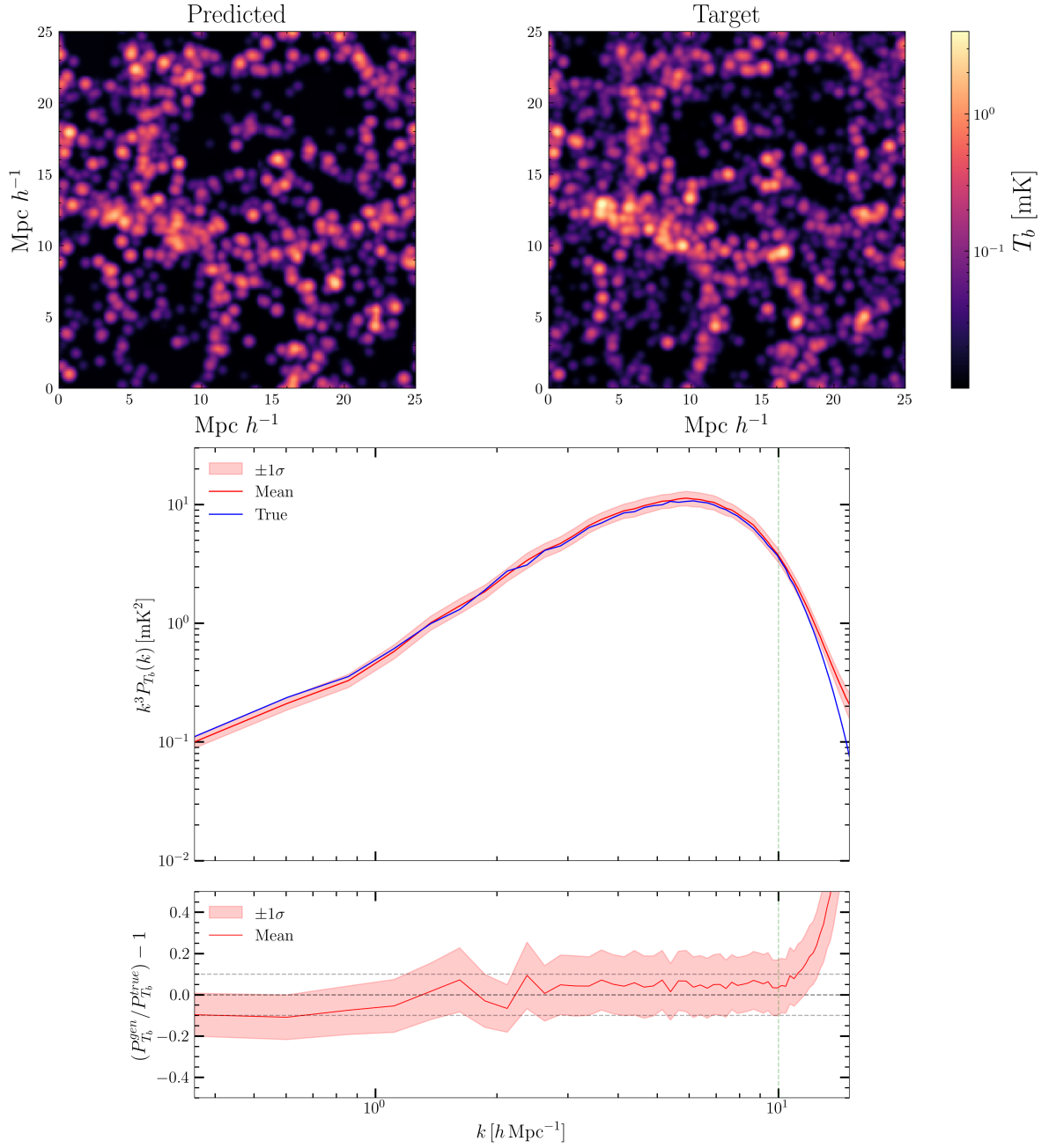


Figure 2.11: Generation of brightness temperature maps with LODI: one sample of diffusion-generated map (top left) compared to the unseen ground truth (top right) visualized by averaging over the third spatial axis, along the entire $25 h^{-1}$ Mpc length of the simulation box. The bottom panel compares the mean power spectrum and its standard deviation obtained from 20 diffusion realizations, each produced with a different seed for the diffusion model, showing good agreement well into the non-linear regime, up to scales $k \simeq 10 h \text{ Mpc}^{-1}$, also visible in the residual spectra shown in the bottom panel. The maps are smoothed with a Gaussian kernel of radius of $0.2 h^{-1}$ Mpc.

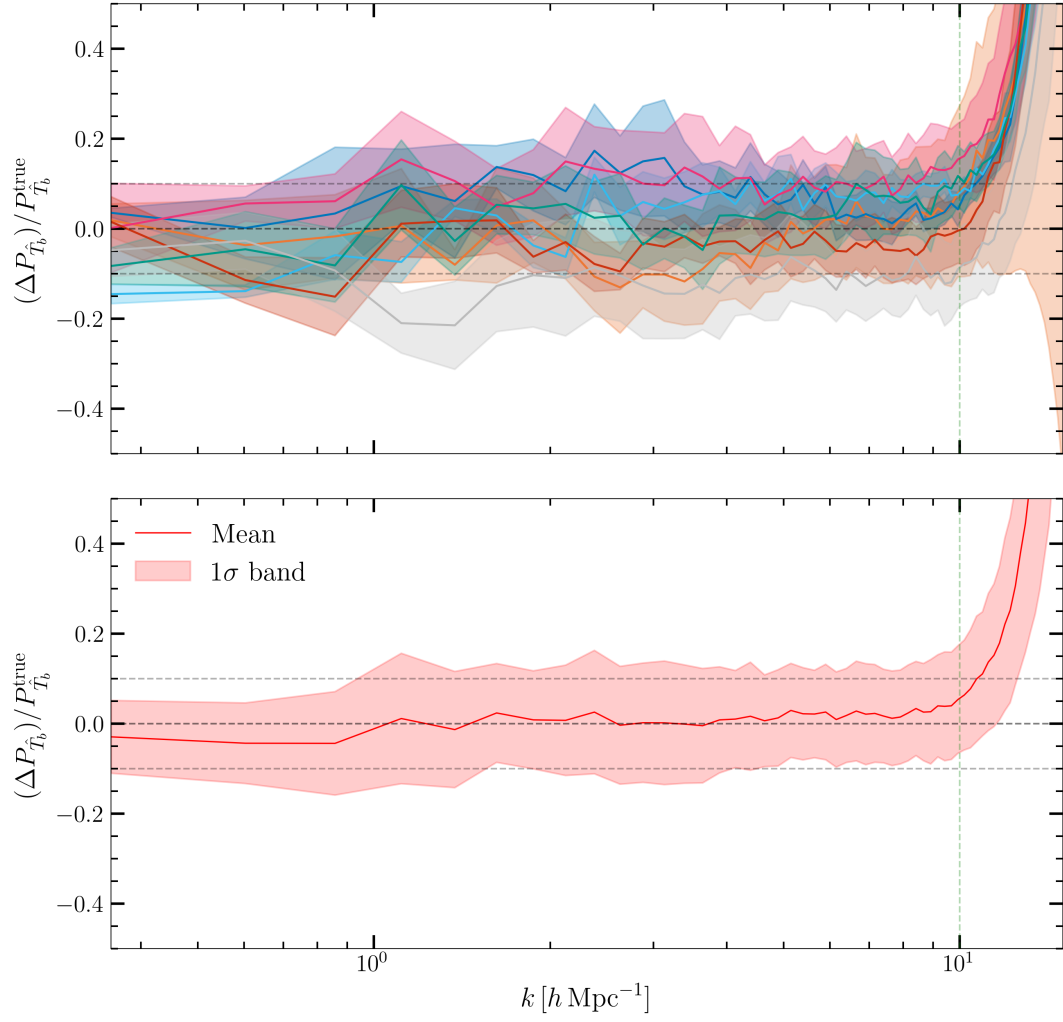


Figure 2.12: Top: mean (lines) and standard deviation (shaded bands) of the residual of $P_{\hat{T}_b}(k)$ power spectrum for 7 dark matter simulations (each color a different seed), with 5 diffusion realizations each. The bottom plot shows the result when combining all the realizations.

Chapter 3

Cosmo-FOLD: Large-Volume Cosmological Field Generation

3.1 Problem statement

High-fidelity hydrodynamical simulations are far too expensive to run in the numbers required by field-level and simulation based inference, a bottleneck already established in Section 1.7 of the Introduction, together with the case for fast, learned surrogates as the way around it.

Beyond LODI and the generative-modeling approaches already surveyed in Section 1.5.4, diffusion models specifically have already been used to emulate baryonic processes affecting galaxy properties (Chadayammuri et al. 2023), to augment simulation resolution from a smaller set of high-resolution runs (Kodi Ramanah et al. 2020), and more broadly to model cosmological fields (Rouhiainen 2024).

What none of these address is the specific limitation this chapter targets: LODI’s own overlapping sub-volume scheme for generating fields larger than its training box. This approach works well for localized structures, such as bounded halos, but fails for extended structures, unless LODI is explicitly retrained to generate larger volumes conditioned on surrounding patches via overlap. Dark matter filaments and diffuse gas structures are two examples that fall into this second category. The overlap requirement also inflates the number of sub-volumes that must be generated by as much as 40% (e.g. from 512 to 729 for a 256^3 -voxel target), with a corresponding increase in compute time (Section 3.3.3 gives the full mechanism and numbers). Moreover, as we shall see, higher order correlations are not guaranteed to be predicted perfectly with LODI, under some

specific cases of field generation.

3.2 A plug-in solution

To address this, we introduce **Cosmo-FOLD** (“Cosmological Field Overlap Latent Diffusion”), which replaces LODI’s overlapping sub-volumes with a sliding window generation scheme that maintains coherence implicitly across the whole diffusion trajectory, respecting periodic boundary conditions of simulations, and without needing an overlap prescription at all. This chapter is structured as follows: Section 3.3 describes the methodology, the network modifications, and the datasets used; Section 3.4 presents results on unseen test simulations and on volumes up to 100 times larger than **Cosmo-FOLD** was trained on; we conclude in Section 3.5.

3.3 Methodology

In this section, we give an overview of the datasets used for our work and how exactly are they processed for training. We also present modifications to the LODI diffusion model that we presented in the previous Chapter.

3.3.1 Diffusion model architecture

We adopt the variational diffusion model of Kingma et al. (2023), which was implemented in our previous work, (Mishra et al. 2025), detailed in the previous chapter (Ch: 2), in an algorithm that we called LODI. LODI trains a variational diffusion model to sample a target 3D field, η , conditioned on a field ξ , by sampling from the conditional distribution $p_\phi(\eta | \xi)$. For this paper, we use the same denoising neural network as LODI: a U-Net with an encoder decoder convolutional neural network with residual connections as the convolutional blocks and an attention block at the bottleneck to capture both global and short range correlations in the data. The encoder has 4 downsampling and upsampling residual blocks, and it is shown in Fig. 3.1. The inputs to the network are the conditioning field ξ , the noised target η^t at time step t , the sinusoidal time step embedding τ^t , defined in Eq. 2.22.

This work introduces improvement to the LODI approach in the form of a different overlap method (**Cosmo-FOLD**) for the generation of volumes larger than the training box, while maintaining long range coherence and fidelity, which is explained in section 3.3.3.

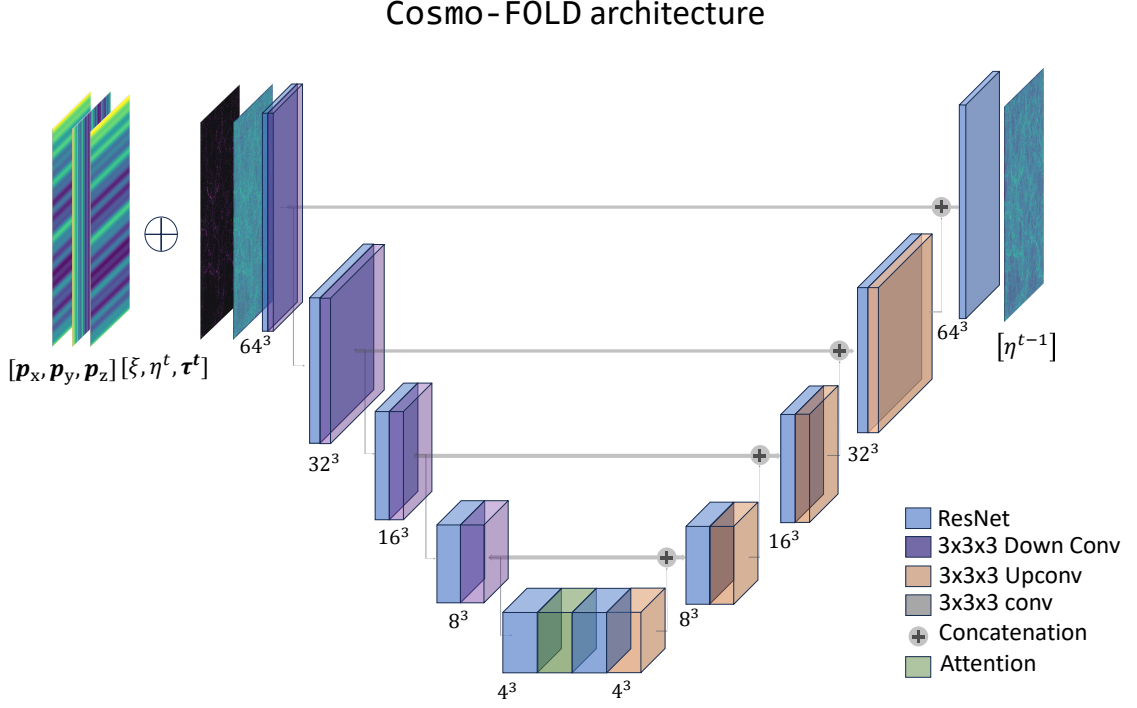


Figure 3.1: Architecture of the denoising network of **Cosmo-FOLD**: a 3D U-net with residual network blocks and a self-attention bottleneck. The inputs to the network are the conditioning field ξ , the noised target η^t at time step t , the sinusoidal time step embedding τ^t corresponding to the diffusion time step t , and the optional positional embedding $\mathbf{p}(\mathbf{x})$ for each of the coordinate axes. The output η^{t-1} is the target field denoised by one extra step.

Another improvement is the introduction of optional positional encoding $\mathbf{p}$, defined in Eq. 3.4, that are concatenated channel wise to the inputs of the diffusion model, thereby adding 3 extra channels, one for each coordinate axis.

3.3.2 Training dataset

For this work, we use two separate simulated datasets. The CAMELS IllustrisTNG subset (henceforth **CAMELS**) was introduced in Chapter 2 and used for training the conditional field generation model. Here, we use the same CV set with fiducial cosmological and astrophysical parameters to train **Cosmo-FOLD**.

The second dataset is IllustrisTNG300-2 (henceforth IllustrisTNG) (Nelson et al. 2021), a full hydrodynamical simulation in a comoving volume of side $205 h^{-1}$ Mpc with 1250^3 gas and dark matter particles. This larger-volume simulation shares the same cosmological parameters as **CAMELS** and fiducial feedback prescription ($\mathbf{A} = 1$). We use IllustrisTNG to demonstrate upscaling: the ability of **Cosmo-FOLD** to generate target fields for volumes substantially larger than those in the training set.

All 3D fields are extracted at redshift $z = 0$, the maximally non-linear regime. From each simulation, we extract the stellar mass density ρ_{star} , cold dark matter density ρ_{DM} , and mass-weighted mean gas temperature $\mathbf{T}_g$. For **CAMELS**, fields are voxelized to 256^3 for $\mathbf{T}_g$ and 128^3 for ρ_{DM} (resolutions ≈ 0.1 and $0.2 h^{-1}$ Mpc). For IllustrisTNG all three fields are discretized to 1024^3 voxels ($\approx 0.2 h^{-1}$ Mpc resolution). Table 3.1 summarizes these configurations.

Generating gas temperature maps from dark matter maps

Our first use case is to predict $\eta = \mathbf{T}_g$ from the $\xi = \rho_{\text{DM}}$ fields, in two different modalities: first, we show with **CAMELS** the performance of the model when the box size is the same between training and generation; second, we demonstrate with IllustrisTNG the ability of our method to upscale to a 100 times larger volume the generative capability learnt from a small simulation box.

For **CAMELS**, we train on three CV simulations, each of side $25 h^{-1}$ Mpc (all with the same cosmological and astrophysical parameters) with different random seeds, each voxelized at 256^3 . For IllustrisTNG, we train on a 256^3 voxels sub cube, corresponding to a box of side $51.25 h^{-1}$ Mpc, taken from the full IllustrisTNG run (Note again the difference in resolution between **CAMELS** and IllustrisTNG). In both cases, we subdivide this 256^3 training volume into 64 sub-volumes, each encompassing 64^3 voxels, which are used as individual samples for training. We note that subdividing into 64^3 voxels for training and generating is carried out due to GPU memory limitations, as larger training volumes would not fit into the available GPU memory. Due to the upscaling capabilities of our method, this is however not a fundamental limitation, and actually

makes the learning more efficient. Data augmentation is applied by random rotations and reflections to enforce rotational equivariance. This gives us 24 different orientations per training sample. Since their density spans over five orders of magnitude, all fields are log-scaled prior to training. This results in $\sim 24,600$ training samples for CAMELS and $\sim 8,200$ for IllustrisTNG.

Generating dark matter maps from stellar density maps

The second use case is to generate dark matter density fields $\eta = \rho_{\text{DM}}$ conditioned on $\xi = \rho_{\text{star}}$. We therefore use for training a 256^3 voxel sub-cube of the full IllustrisTNG run, and further sub-divide into 64^3 voxel sub-volumes with the same data augmentation strategy as in Section 3.3.2, yielding $\approx 8,200$ training samples.

In addition to this baseline setting, we introduce a new test for the generative capabilities across different simulations, in which the model, trained on the small CAMELS boxes, is applied to the full IllustrisTNG volume to infer ρ_{DM} from ρ_{star} . This CAMELS $\rightarrow$ IllustrisTNG transfer experiment allows us to assess the robustness of the conditional mapping across simulation suites. Specifically, we take four of the small CAMELS simulations, each discretized into 128^3 voxels, subdivide them into 64^3 voxel sub-volumes, and apply the same data augmentation procedure to obtain $\approx 2,600$ individual samples.

In all cases, the data are split into 70% for training and 30% for testing.

Diffusion model training and validation

The training procedure is identical for both applications described in sections 3.3.2.

At each iteration, the target field η , either $\mathbf{T}_g$ or ρ_{DM} , is sampled from the training set and corrupted with a Gaussian according to Eq. (2.9). The optimizer employed for training is AdamW (Loshchilov & Hutter 2019), with a batch size of 8. The learning rate schedule adopted is Cosine Annealing Warm Restarts, with the maximum and minimum values for the learning rate set to 3×10^{-4} and 10^{-6} , respectively, with repeat cycles of 20 epochs. The model is trained for 100,000 steps and the model checkpoint that corresponds to the lowest loss is stored.

For the CAMELS simulations we set $T = 400$ for both ρ_{DM} and $\mathbf{T}_g$ generation, while for the case of IllustrisTNG, we choose $T = 200$ for $\mathbf{T}_g$ generation. For the case of IllustrisTNG from CAMELS a timestep of $T = 100$ for ρ_{DM} generation is chosen, as a

compromise between generation accuracy (which requires larger T) and computational effort (which increases with larger T).

3.3.3 Cosmo-FOLD: Field Overlap Latent Diffusion

As explained above, the diffusion model is trained on small sub-volumes, which for the case of IllustrisTNG are 1.5% of the volume one wishes to generate the target field for. The central problem is how to then combine the generated fields in such a way to obtain an optimal generation for the much larger target volume, that was never seen in during training.

Simply generating independent sub-volumes each with the same size of the training data volume and then assembling them to generate the target field leads to discontinuities at the cube boundaries, as illustrated in the bottom left panel of Figure 3.3 for the case $\xi = \rho_{\text{DM}}$ and $\eta = \mathbf{T}_g$. The discontinuities arise because $\mathbf{T}_g$ is sampled from the conditional distribution $\mathbf{T}_g \sim p(\mathbf{T}_g | \rho_{\text{DM}})$ which depends solely on the local voxels of the ρ_{DM} sub-volume, without considering the surrounding context. Consequently, the edges of adjacent cubes do not perfectly align. However, the boundary artifacts do not impede the model’s primary objective of capturing the small-scale relationship between dark matter and the temperature field.

In LODI, we address this issue by using overlapping sub-volumes at the generation stage. Each cube being generated shared a boundary region with its neighbors so that latent representations were correlated across overlaps. This ensured that the boundary conditions between sub-volumes in the overlapping region were aligned. Although LODI was shown to work well when the overlapping regions feature localized structures such as dark matter halos, the approach fails for extended structures that exceed the size of a sub-volume, such as dark matter filaments or $\mathbf{T}_g$, as demonstrated in the bottom central panel of Figure 3.3. The second issue of LODI is the increased runtime: for example, to generate a 256^3 voxel simulation with LODI, we would need to tile a minimum of $512 \cdot 32^3$ sub-volumes (this being the training sub-volume size chosen for LODI). However, due to the requirement to overlap sub-volumes, the number of forward generation sub-volumes needed increases by 40% to 729, with a corresponding increase of compute time. Increasing the size of the training volume would in principle help to mitigate this, but the attention block’s computational cost scales quadratically with input dimensions, thus making generation slower rather than faster for larger sizes. Second, because the attention module was not trained on larger spatial volumes, it would need to extrapolate

beyond its learned regime, potentially degrading correlation accuracy.

To address all of the above issue, we introduce here the novel **Cosmo-FOLD** method (summarized in Algorithm 2), a generation scheme that builds over and replaces the overlapping sub-volume strategy while overcoming its key limitations. **Cosmo-FOLD** is able to produce large-scale, spatially coherent cosmological fields with higher fidelity and lower computational cost than LODI, thus enabling efficient synthesis of full-volume realizations without sacrificing small-scale structure detail.

Unlike other approaches that denoise overlapping chunks with some prescription to enforce continuity, **Cosmo-FOLD** employs a differential sliding window strategy that maintains coherence implicitly throughout the diffusion trajectory. Crucially, this does not use any overlap prescription. This design reduces generation time by $\approx 40\%$ relative to LODI when applied to the IllustrisTNG suite of **CAMELS** (assuming an 8-voxel overlap per side for 64^3 chunks). Furthermore, **Cosmo-FOLD** is explicitly designed to respect periodic boundary conditions, which were not enforced in LODI.

Cosmo-FOLD generation scheme

To illustrate the **Cosmo-FOLD** method (See fig. 3.2), we consider generating a target η field, conditioned on ξ , with a volume of 256^3 voxels. We first initialize a noise map of the same shape as the target, which corresponds to $\eta^{t=T}$, where the upper index denotes the step in the denoising process. Next, this cube is partitioned into a 4^3 grid of non overlapping chunks, each with 64^3 voxels, matching the size of the training samples. We then apply one step of the diffusion denoising model ($t \rightarrow t - 1$) to each chunk. Each 64^3 sub-volume is denoised independently for a single time step according to the learned conditional distribution:

$$\tilde{\eta}_j^{t-1} \sim p_\phi \left(\eta_j^{t-1} \mid \eta_j^t, \xi_j^t \right), \quad (3.1)$$

where $j = 1, \dots, 4^3$ indexes the sub-volume position and ξ_j^t is the sub-volume j of the field being conditioned upon, ξ . While the conditioning field is independent of diffusion time, we use the superscript t to denote the voxel shifting corresponding to time t (see below). The one step denoised sub-volumes $\tilde{\eta}_j^{t-1}$ are first arranged side by side to obtain the intermediate large volume field $\tilde{\eta}^{t-1}$. We then shift this intermediate field by one voxel along each of the three spatial axes to obtain the updated field $\eta^{t-1} = \text{Shift}(\tilde{\eta}^{t-1})$. The conditioning field must be shifted in an analogous manner to remain consistently aligned with the updated tiling. We therefore denote by ξ^t the conditioning field at the

grid orientation used at diffusion step t , and obtain ξ^{t-1} by applying the same one-voxel shift to ξ^t , i.e. $\xi^{t-1} = \text{Shift}(\xi^t)$.

The pair (η^{t-1}, ξ^{t-1}) then serves as input to the next denoising step. In this way, at each iteration the entire cube is denoised again using the most recent latent state, but with an offset grid so that different voxel neighborhoods are processed together over successive steps.

This reconstruction and grid shift occurs a total of T times, one for each time step, ensuring that information propagates smoothly across sub-volume boundaries. Since the simulation fields satisfy periodic boundary conditions, the grids also wrap around the edges of the simulation box when shifted. In Fig. 3.2, for the use case in the first line of Table 3.1: the image on the left is the full CAMELS volumes generated at time $t = 310$ (about a quarter through the denoising process, as here $T = 400$), and the partition of the sub-volumes follows the green grid lines. Some structure begins to emerge from the initial pure noise. After another 80 time steps (central panel), the new grid structure is shown with solid lines, and the old grid is dashed. At $t = 140$ step (right image), the grid has further shifted and the noise level has been reduced.

Algorithm 2 Cosmo-FOLD

```

1:  $\eta^T \sim \mathcal{N}(\mathbf{0}, \mathbb{I})$ 
2: for  $t = T - 1, \dots, 0$  do
3:   Partition in sub-volumes
4:   for cube  $j \in \text{sub-volumes}$  do
5:      $\tilde{\eta}_j^{t-1} \sim p_\phi(\eta_j^{t-1} \mid \eta_j^t, \xi_j^t)$ 
6:   end for
7:   Tile  $\tilde{\eta}_j^{t-1}, \xi_j^t$  to get  $\tilde{\eta}^{t-1}, \xi^t$   # recreate the original volume size at time  $t - 1$ 
8:   shift tiled  $\tilde{\eta}^{t-1}, \xi^t$  by 1 voxel along 3 axes with periodic boundary conditions to
   get  $\eta^{t-1}$  and new shifted  $\xi^{t-1}$ 
9: end for
10: return  $\eta^0$ 

```

The success of Cosmo-FOLD follows from the fact that the algorithm does not privilege any fixed grid orientation, thereby preventing systematic edge artifacts. After each shift, every new sub-volume overlaps with several one voxel wide planes that were denoised in previous time steps. These boundary planes are initially slightly out of distribution relative to the interior voxels, but the subsequent denoising pass naturally smooths away such small discontinuities, and adds spatial correlations in those voxels, ensuring global

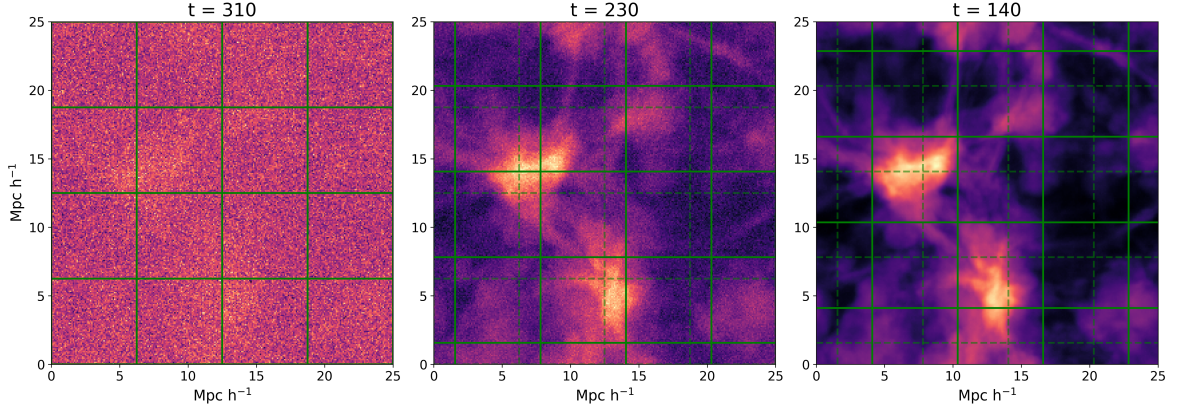


Figure 3.2: Illustration of the **Cosmo-FOLD** overlap latent diffusion method trained on **CAMELS** to generate a gas temperature map conditioned on dark matter: example generated gas temperature cube at $t = 310$ (left panel), with its sub-volume grid shown in green solid lines. At a later diffusion timestep, $t = 230$ (middle panel), the gas field has been denoised, with solid lines showing the current grid and dashed lines the grid location at $t = 310$. The process continues until $t = 0$ is reached, with the right panel showing another snapshot at $t = 140$.

coherence. Provided that the noise increment per time step is sufficiently small, the diffusion model perceives no abrupt transitions, and the denoiser refines the entire field as a continuous, spatially consistent structure across iterations.

3.3.4 Position encodings for higher order correlation

The two point correlation function or its Fourier counterpart, the power spectrum, is the most widely used statistic to characterize the statistical properties of a cosmological field. However, the power spectrum fully describes the field only if the underlying density fluctuations are a Gaussian random field, which is not the case at later cosmological times (Bernardeau et al. 2002; Sefusatti et al. 2006). To capture deviations from Gaussianity, higher-order statistics are required, the lowest order of which is the bispectrum, the Fourier transform of the three point correlation function.

For a given density field ρ , the bispectrum $B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$ is defined as:

$$\langle \delta(\mathbf{k}_1) \delta(\mathbf{k}_2) \delta(\mathbf{k}_3) \rangle = (2\pi)^3 \delta^D(\mathbf{k}_1 + \mathbf{k}_2 + \mathbf{k}_3) B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3), \quad (3.2)$$

where $\delta(\mathbf{k})$ is the Fourier transform of the overdensity field $\delta(\mathbf{x})$ and δ^D is the Dirac

delta function. Since a triangle configuration in $\mathbf{k}$ space can also be represented with two wave modes and the angle θ between them, we will express the bispectrum as $B(\mathbf{k}_1, \mathbf{k}_2, \theta)$.

The bispectrum is an important source of cosmological information: cosmological parameters (like the amplitude of fluctuations, bias of galaxies, neutrino mass, etc.) can be degenerate in terms of the power spectrum, and the bispectrum helps break these degeneracies, especially in the context of galaxy clustering. In addition, several inflationary scenarios predict the presence of primordial non Gaussianity (Bartolo et al. 2004). It is therefore important that the bispectrum of dark matter is predicted accurately by the diffusion process.

In practice, it is also often useful to consider the *reduced bispectrum*, which isolates the shape dependence of the three point function by factoring out the overall amplitude set by the power spectrum. The reduced bispectrum $Q(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)$ is defined as

$$Q(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3) = \frac{B(\mathbf{k}_1, \mathbf{k}_2, \mathbf{k}_3)}{P(k_1)P(k_2) + P(k_2)P(k_3) + P(k_3)P(k_1)}, \quad (3.3)$$

where $P(k)$ denotes the power spectrum of the overdensity field.

Since the bispectrum and the reduced bispectrum are a function of both the separation between the points and the angle between the wave modes, there exist infinitely many unique triangle configurations. It is therefore impossible to train a model to learn the bispectrum for all possible configurations individually. To address this issue, and to enable the model to capture the scale and angular dependence of the bispectrum across all configurations, we supplement the Cosmo-FOLD input with periodic positional encodings. These provide global phase information that is otherwise inaccessible to purely local convolutional filters of the diffusion model. For each spatial coordinate axis, we construct a positional feature as a sum of sinusoidal functions at multiple frequencies:

$$\mathbf{p}(\mathbf{x}) = \sum_{i=0}^n \alpha_i [\sin(2\pi \mathbf{x} f_i) + \beta \cos(2\pi \mathbf{x} f_i)], \quad (3.4)$$

where $\mathbf{x} = (x, y, z)$, and where we work with normalized coordinates, so $x, y, z \in [0, 1]$. $\{f_i\}_{i=0}^n = 2^n$ are the frequencies, each dealing with a different scale, thus spanning a hierarchy of scales within the box. α_i are the fixed amplitude weights per frequency bands and β is chosen to add a relative amplitude between the sine and cosine signals,

which has shown in our case to improve results. The positional encodings $\mathbf{p}(\mathbf{x})$ are concatenated to the conditioning field ξ (in our case, this is ρ_{DM}), yielding an array of shape $3 \times d^3$, where d are the total number of voxels per side. We test the model for different values of the hyperparameter n . In principle, the larger the n , the more phase information the model has access to. However, it can also confuse the convolutional network since we are adding all frequencies for one coordinate into one channel, making it hard for the network to disentangle the phase information at different scales. For the results of this paper, we use $n = 4$, while $\alpha_i = [0.33, 0.29, 0.25, 0.21, 0.19]$ and $\beta = 1.5$. These values were chosen heuristically based on a limited set of exploratory experiments, in which we varied the relative weighting of low and high frequency components and monitored reconstruction quality. Although we have not systematically explored the full hyperparameter space, these values were found to yield satisfactory results. The coefficients α_i also control the relative strength of the positional encoding signal with respect to the conditioning field ξ . If the signal is too weak (small α_i), the model is unable to effectively utilize the positional information. Conversely, excessively large values of α_i inject spurious structure at the corresponding Fourier modes, which is particularly detrimental given the sparsity of ξ (this sparsity argument mainly holds for stellar fields). The chosen values therefore represent a compromise that allows the network to exploit positional information without overwhelming the physical signal. We apply this modification of the method on the IllustrisTNG simulation, where we condition on ρ_{star} and generate ρ_{DM} , and verify the quality of the bispectrum of the generated dark matter map for a few different triangle configurations.

3.4 Results

We evaluate the performance of **Cosmo-FOLD** using both unseen test simulations and large-scale generative experiments. We first test the model’s performance on an independent unseen realization from the Cosmic Variance subset of the **CAMELS** simulation suite, which has simulations with the same cosmology but different initial conditions, with a box size of $25 h^{-1} \text{ Mpc}$ per side, discretized into 256^3 voxels. This assesses the model’s generalization to new initial conditions. To demonstrate its capability to scale to larger domains, we then apply **Cosmo-FOLD** to the IllustrisTNG simulation at $z = 0$ in two distinct settings: *(i)* generating the gas temperature field conditioned on the dark matter density, and *(ii)* generating the dark matter density conditioned on a sparse stellar

density field. The IllustrisTNG volume spans $205 h^{-1}$ Mpc per side and is discretized into 1024^3 voxels, while the model is trained on only $\approx 1.5\%$ of its total volume. For this second experiment, we demonstrate the model performance both when trained on IllustrisTNG and when trained on CAMELS, thus showcasing its scale generalization capabilities, beyond the simulation volume it has been trained on.

3.4.1 Coherent gas temperature generation

Generating with CAMELS

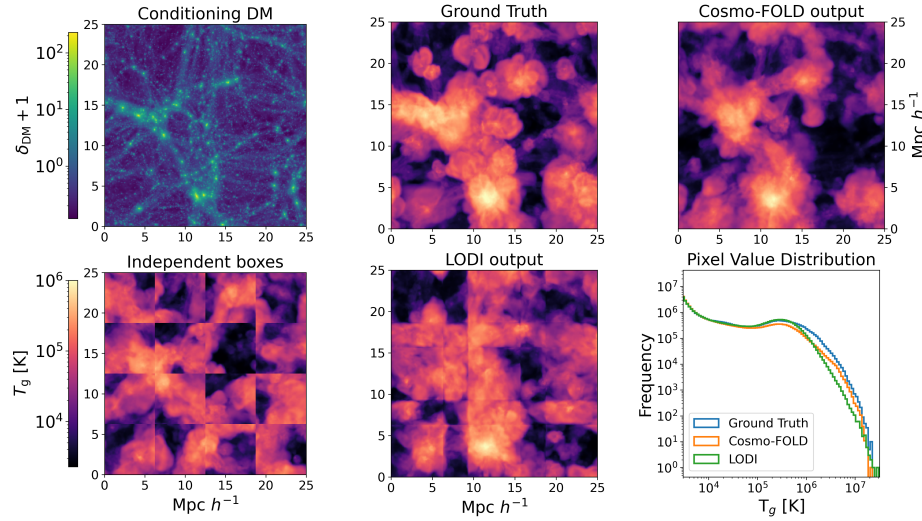


Figure 3.3: Generated gas temperature map (top right), conditioned on the ρ_{DM} field (top left) for the CAMELS simulation, with the target box size equal to the training box size. The true (unseen) temperature map is shown in the top middle panel. For comparison, the bottom left panel shows the output obtained by simply tiling independent sub-boxes, while the bottom middle panel is the output of our previous LODI method. The histograms of the distribution of pixel values are shown in the bottom right, averaged over 5 realizations of the diffusion models outputs.

We first evaluate Cosmo-FOLD by generating a gas temperature field T_g conditioned on the dark matter density field ρ_{DM} , coming from an unseen CAMELS simulation with the same cosmology and astrophysics as the training simulation, but with different initial conditions. The simulation volume is a cube of side $25 h^{-1}$ Mpc, that is discretized into 256^3 voxels. The true (unseen) and Cosmo-FOLD-generated gas temperature maps are shown in Fig. 3.3 in the top middle and top right panels, respectively. An example

generated gas map with simple tiling of independent boxes is shown in the bottom left, exhibiting macroscopic discontinuities at the boundaries of the training boxes, while the output of our previous LODI method is shown for comparison in the bottom middle panel. LODI produces visible edge artifacts at patch boundaries that are absent in the Cosmo-FOLD method, which delivers seamless, spatially coherent generations and looks faithful to the ground truth. The bottom right panel shows a comparison of the distribution of voxel values, averaged over 5 generations for the same conditional dark matter field. For temperatures below $10^6 K$, the voxel distribution obtained with LODI is marginally better than the Cosmo-FOLD, whereas for larger values Cosmo-FOLD exhibits a slightly better agreement. This is because in LODI some regions are generated independently and without overlaps, which produces more locally accurate results. However, accuracy and overall fidelity is lost due to the subsequent edge effects, which is not the case in our novel approach.

3.4.2 Upscaling to IllustrisTNG

Generating IllustrisTNG gas temperature field

Next, to evaluate performance on upscaling to larger domains, we apply Cosmo-FOLD to the IllustrisTNG simulation (Figure 3.4). Here, the diffusion model is trained and validated on a 256^3 voxels subset of the data, or $\approx 1.5\%$ of the simulation (70% of which is for training and rest for validating). The ground truth gas temperature field $\mathbf{T}_g$ is shown in the top right panel, to be compared with two different generation routes obtained with Cosmo-FOLD. In the first case (top row), the generated $\mathbf{T}_g$ field is conditioned on the *true* ρ_{DM} field of the IllustrisTNG simulation, while in the second case (middle and bottom row), the conditional ρ_{DM} field has also been generated using the Cosmo-FOLD algorithm from the underlying stellar field, as described in section 3.4.3.

We compare the statistical distributions of voxel values using both 1-dimensional histograms and 2-dimensional $\rho_{\text{DM}} - \mathbf{T}_g$ joint distributions in Fig. 3.5. The 1D histogram shows that for intermediate $\mathbf{T}_g$ values, ranging from $1\text{K} \leq \mathbf{T}_g \leq 10^{11.5}\text{K}$, the abundance of voxels is accurately reproduced for both (true and generated) conditioning dark matter maps. At the highest temperatures the model under-predicts rare, extreme-valued voxels, that represent a fraction in the range $10^{-6} - 10^{-8}$ of the total voxel population, very difficult to capture due to their rarity among the training examples. The 2D contour plot, displaying the 68% and 95% joint probability mass of dark matter density voxel values and $\mathbf{T}_g$ values, shows close agreement between the target and generated contours

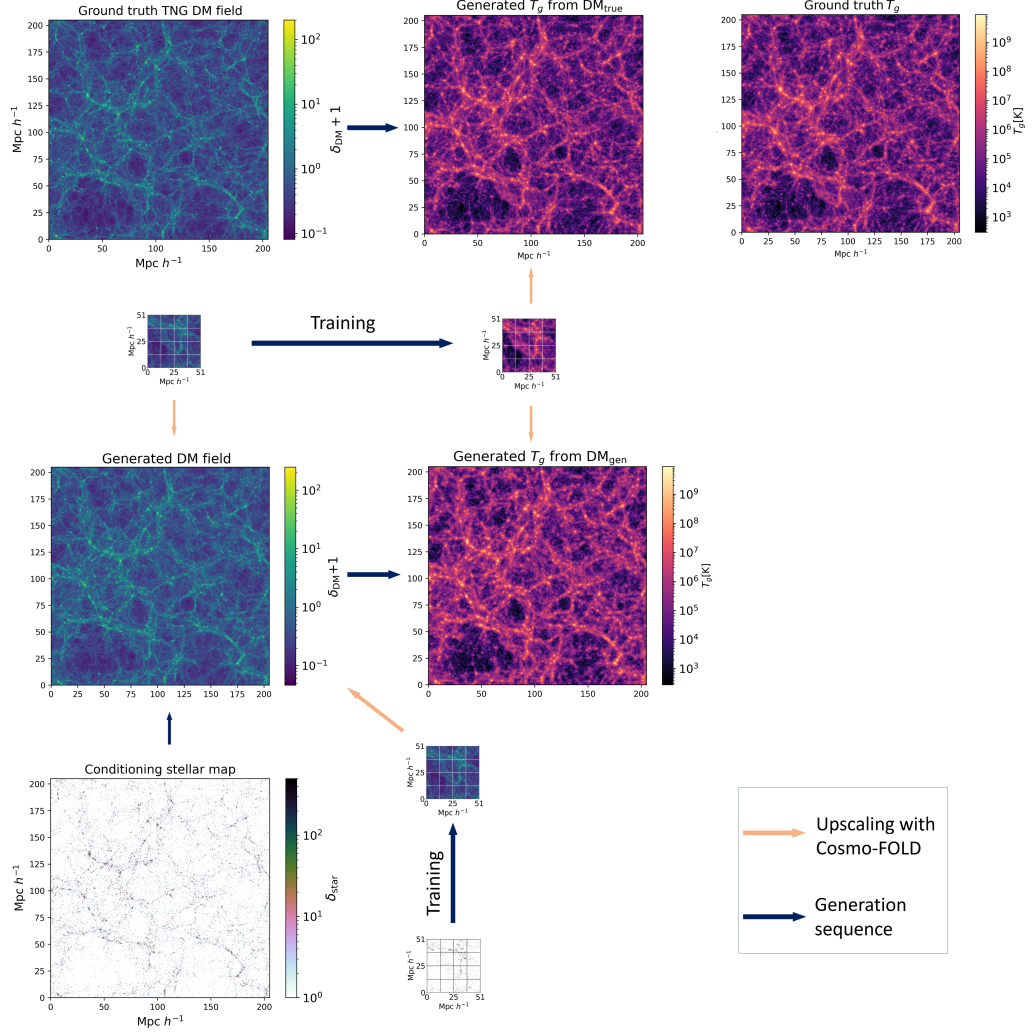


Figure 3.4: Upscaling to larger volumes via the *Cosmo-FOLD* method, applied to generate the target gas temperature field T_g of IllustrisTNG (top right), conditioned on the true ρ_{DM} (top left) or to the ρ_{DM} field (middle left) generated from the stellar density field (bottom left). Orange arrows depict upscaling, i.e., large volume generation after training on a much smaller volume. The dark blue arrows show the generation direction, pointing from the conditional field to the generated field.

when the true dark matter map is used for conditioning, except where large T_g values correspond to low density dark matter regions, where *Cosmo-FOLD* under-predicts the T_g values. This is expected because the conditional dark matter density carries limited

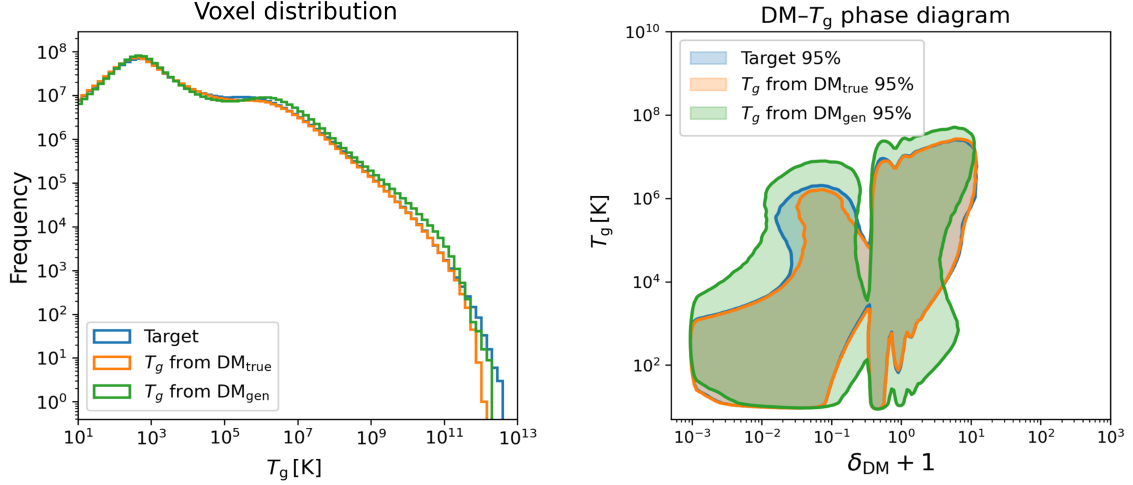


Figure 3.5: Voxel statistics, comparing the quality of the generated T_g fields of IllustrisTNG simulations using the Cosmo-FOLD pipeline. Left: voxel distribution of the gas temperature field, generated using both the ground-truth ρ_{DM} field and the Cosmo-FOLD-generated ρ_{DM} field. Right: 95% contour levels of the joint $\rho_{\text{DM}} - T_g$ distribution for the same two generated cases, compared against the ground truth.

information about the diffuse feedback heated gas that populate these voxels. The agreement with the ground truth is somewhat reduced when the conditioning map has been itself generated from the stellar map, which is again expected because of the compounding of errors.

3.4.3 Dark matter from stellar density fields with upscaling

We next apply Cosmo-FOLD to the inverse problem of reconstructing the dark matter overdensity field ρ_{DM} , from the corresponding stellar field ρ_{star} . We evaluate the model capabilities to transfer from CAMELS to IllustrisTNG.

IllustrisTNG from IllustrisTNG

First, we apply the model to the IllustrisTNG simulation, generating the ρ_{DM} field (middle left panel in Fig. 3.4) from the stellar density field (bottom left panel) voxelized into 1024^3 cells, and covering a cube of side $205 h^{-1}$ Mpc. The 256^3 voxels sub-volume used for training and validation ($\approx 1.5\%$ of the total) is shown by the smaller boxes on

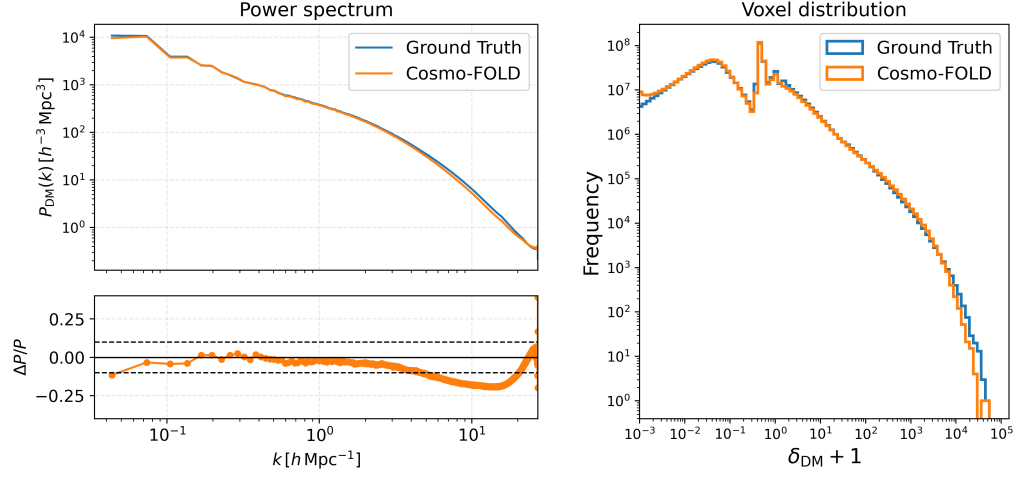


Figure 3.6: Voxel statistics comparing the quality of the ρ_{DM} field of the IllustrisTNG simulation generated from a IllustrisTNG stellar field with upscaling. Left: generated dark matter power spectrum together with the relative deviation from the ground truth. Horizontal dashed lines delimit the 10% level deviation. Right: voxel distribution of the dark matter field generated by Cosmo-FOLD when conditioned on the stellar density field.

the bottom right (the grid shows the further subdivision in 64^3 training volumes). To assess the quality of the generation, we compare the 1 point statistic in the form of a histogram of voxel values and the 2 point correlation statistics in the form of the power spectrum $P_{\text{DM}}(k)$ of the generated and the ground truth dark matter fields in Fig. 3.6, where we also show power spectrum residuals. The power spectrum matches to better than 10% well into the non-linear regime, increasing only beyond $k \approx 5 h \text{ Mpc}^{-1}$. The reason for the loss of power at the smallest scales is their extreme non-linear nature, which is harder to predict. For the largest scales, the error is a combination of cosmic variance and the approximate correlation for voxel with separations much larger than the training volume. However, this effect is mild since in our case the wave mode corresponding to the sub-volume scale is $k \approx 0.5 h \text{ Mpc}^{-1}$, but the power spectrum does not lose power till $k \approx 0.1 h \text{ Mpc}^{-1}$. We note that the small scale performance can potentially be improved by increasing the total denoising steps T , at the cost of increased generation time.

IllustrisTNG from CAMELS

Here we revert to the baseline **Cosmo-FOLD** model without positional encodings and assess its ability to generate IllustrisTNG volumes when trained exclusively on small **CAMELS** simulations. We train and validate the model on four independent **CAMELS** simulations with a comoving box size of side $25 h^{-1}$ Mpc, generating dark matter overdensity fields conditioned on stellar density fields. Each simulation is partitioned into 64^3 voxel sub-volumes for training. Figure 3.7 illustrates the training and generation pipeline, with the four **CAMELS** simulations shown in the bottom row. To quantitatively assess the quality of the generated IllustrisTNG dark matter field, we compare its power spectrum to that of the ground truth. We note that visually, the generated field looks very similar to the ground truth field, recreating most of the filaments and voids, even if during training, the filaments and structures seen by **Cosmo-FOLD** were very different from IllustrisTNG, and much smaller in size.

The resulting power spectrum exhibits two distinct regimes, as shown in Fig 3.8. On scales smaller than the **CAMELS** box size, the generated field reproduces the true power spectrum to within 10% accuracy, extending into the non-linear regime up to $k \leq 6 h \text{Mpc}^{-1}$. In contrast, for scales approaching and exceeding the **CAMELS** box size, the model exhibits a systematic loss of power, with residuals exceeding the 10% level. This behavior is expected, as **Cosmo-FOLD** cannot learn or reconstruct large-scale modes that are absent from the training volumes. As a result, power on scales larger than those present in the training data cannot be faithfully recovered.

3.4.4 Improving Bispectrum via positional encoding

Beyond one- and two-point statistics, an important test of generative models for cosmological fields is their ability to reproduce higher order correlations that encode non-Gaussian structure arising from non-linear gravitational collapse at late times. In particular, the bispectrum of the dark matter density field provides a sensitive probe of mode coupling induced by non-linear gravitational evolution. To assess whether our model captures this information, we repeat the generation of the dark matter field from the stellar density field while augmenting the conditioning input with sinusoidal positional encodings $\mathbf{p}(\mathbf{x})$, as described in Section 3.3.4. These encodings provide the network with explicit information about spatial location within the simulation volume, which can help to break translational degeneracies inherent to patch-based training and improve the modeling of large scale phase correlations.

CAMELS $\rightarrow$ TNG GENERALIZATION

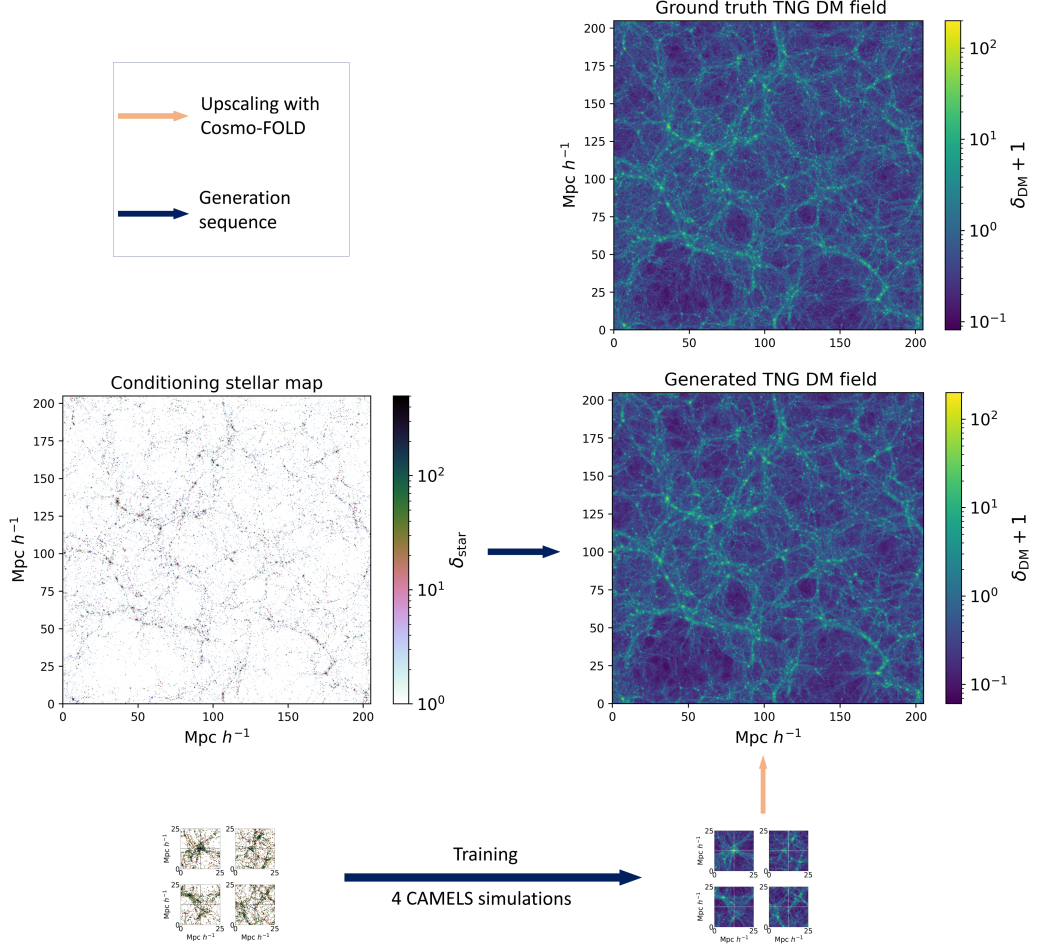


Figure 3.7: Capability of **Cosmo-FOLD** to generate IllustrisTNG simulations, ρ_{DM} (bottom right) while being trained on small **CAMELS** simulations (bottom four small squares).

The resulting bispectra and reduced bispectra of the generated dark matter fields are shown in Fig. 3.9, where we compare them to the ground-truth simulation across a range of triangle configurations in Fourier space. In all panels, the blue curves correspond to **Cosmo-FOLD** augmented with positional encodings, the orange curves to the baseline **Cosmo-FOLD** model without positional encodings, and the black curves to the ground truth.

We present results for four representative triangle configurations (see Section 3.3.4 for

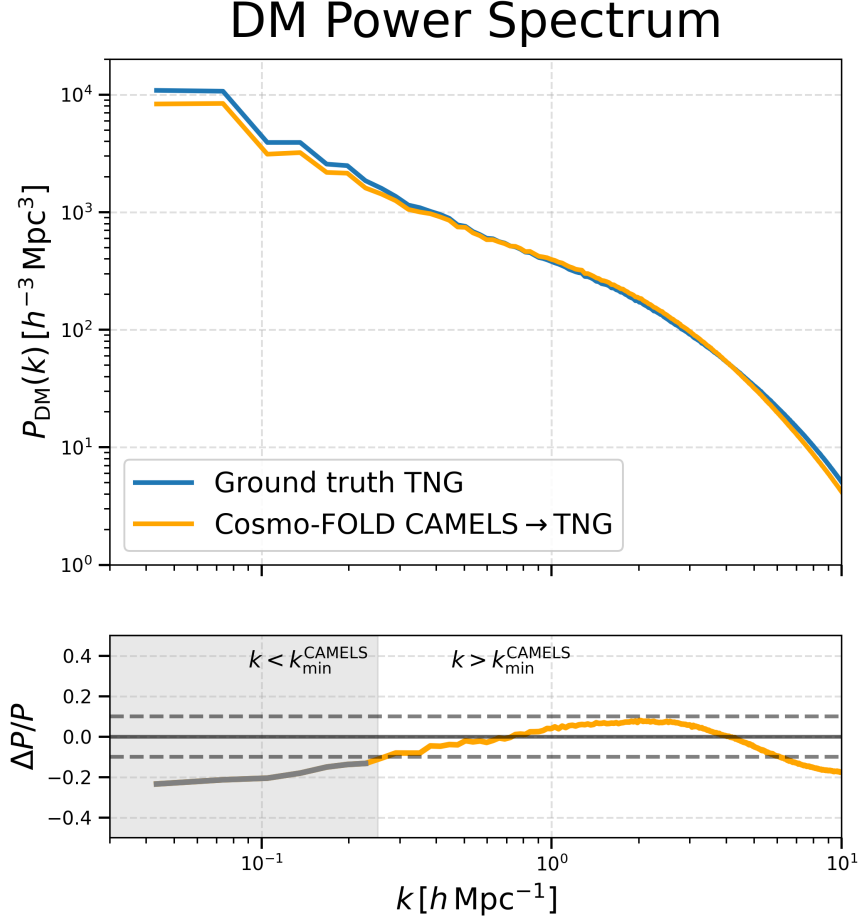


Figure 3.8: The dark matter power spectrum generated by the *Cosmo-FOLD* method, when trained on *CAMELS* simulations and upscaled to produce the IllustrisTNG map, as illustrated in Fig.3.7. The bottom plot shows the relative difference in the power spectrum when compared to the IllustrisTNG ground truth dark matter spectrum. The gray shaded areas show the scales which contain wavemodes larger than those present in the *CAMELS* simulation data.

details). The leftmost column corresponds to equilateral configurations with wavevectors $k_1 = k_2$ and the angle θ between them as $\pi/3$. The middle-left column shows squeezed configurations with $k_1 = k_2$, but with $\theta = 0$. The middle-right and rightmost columns correspond to fixed-scale configurations with $(k_1, k_2) = (0.1, 1.0) h \text{ Mpc}^{-1}$ and $(k_1, k_2) = (0.3, 0.1) h \text{ Mpc}^{-1}$. Across all configurations, we find that the inclusion of positional encodings significantly improves agreement with the true bispectrum going from $\approx 100\%$ in relative error to $\leq 10\%$ on mid to large scales. This demonstrates that the model is

able to recover non-Gaussian features of the dark matter field that are not constrained by the power spectrum alone.

Adding positional encodings does not appreciably reduce the quality of the lower-order statistics. As shown in Fig. 3.10, both the one-point distribution and the power spectrum of the generated fields remain in close agreement with the ground truth when positional encodings are added, with only a mild degradation compared to the case without positional encodings (Fig. 3.6). We note the presence of additional power at intermediate scales, $k \simeq 0.3 - 1.0 h \text{ Mpc}^{-1}$, which can be attributed to spurious structure introduced by directly concatenating positional encodings to the sparse stellar field at the voxel level. This additional signal can partially confuse the network and enhance power at these scales. Conversely, power on the largest scales, previously suppressed in the absence of positional encodings, is more accurately recovered.

Overall, these results indicate that the inclusion of positional information substantially improves the modeling of higher order correlations, while preserving the accuracy of one and two-point statistics to a high degree.

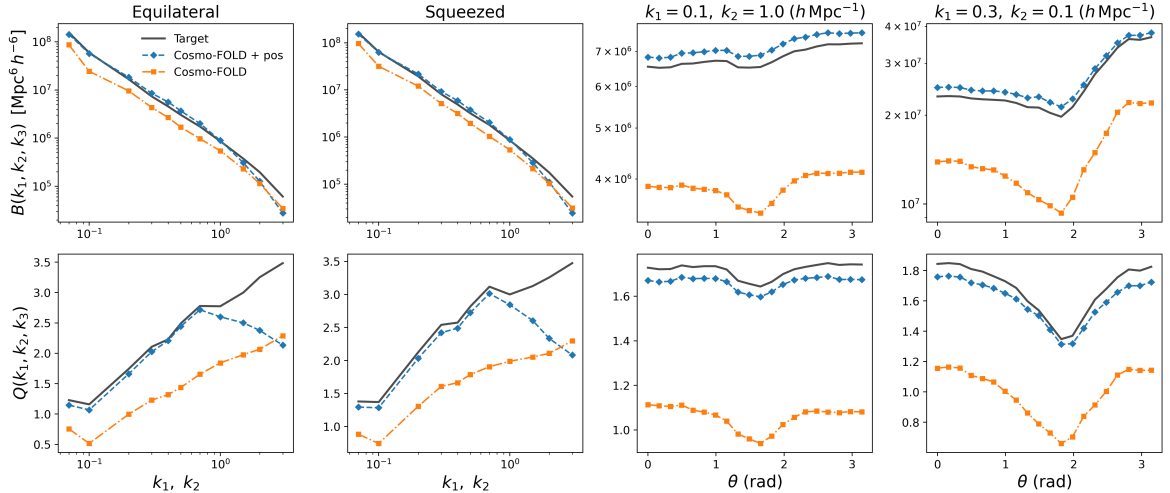


Figure 3.9: Comparison of four bispectrum configurations (top row) and their corresponding reduced bispectrum (bottom row) between the baseline **Cosmo-FOLD** algorithm and the variant augmented with positional embeddings, denoted as **Cosmo-FOLD+pos**. The equilateral (left column) and squeezed (left-center column) triangle cases are shown as functions of wavenumber, obtained by varying the k -values along the corresponding triangle configurations. For the remaining two columns, the wavenumbers are fixed at the values indicated in the panel titles, and the bispectrum is evaluated as a function of the angle θ between k_1 and k_2 .

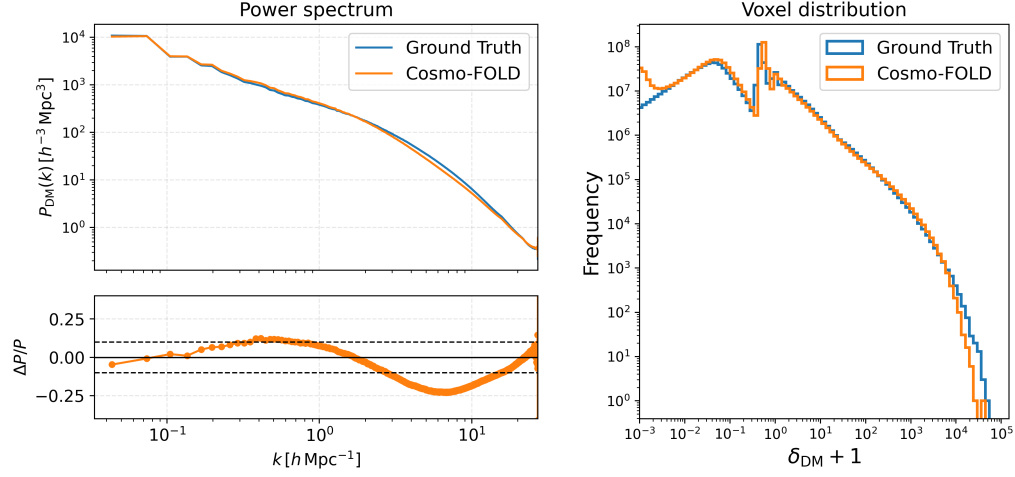


Figure 3.10: Power spectrum and a histogram plot comparing the 2 point correlation and voxel value distribution between the target and generated ρ_{DM} field, conditioned on both the ρ_{star} and positional encodings.

3.4.5 Computational cost

A tabulated summary of the computational time required to produce the various results presented in this paper is provided in Table 3.1. All results were generated on a single NVIDIA A100 GPU with 40 GB of memory. We further note that the generation time scales approximately linearly with the number of denoising steps T and can therefore be reduced by decreasing T , at the cost of degraded performance on the smallest scales.

3.5 Conclusions

We have presented **Cosmo-FOLD** (Field Overlap Latent Diffusion), a novel framework for generating conditional, arbitrarily large cosmological fields by coherently stitching together outputs from a diffusion model trained on smaller volumes. Building upon the variational diffusion architecture of **LODI**, **Cosmo-FOLD** enables seamless reconstruction of large scale structure while preserving local physical fidelity, does not require extra generation time, and works as a plug in and play technique.

We demonstrated the performance of **Cosmo-FOLD** through two representative and complementary test cases. In the first case, we generated gas temperature fields conditioned on the dark matter density field. This choice highlights both the ability of the diffusion model to synthesize complex baryonic observables and the effectiveness

Table 3.1: Computational cost associated with the Cosmo-FOLD use cases presented in this chapter. All results were obtained on a single NVIDIA A100 GPU with 40 GB of memory. Reported times correspond to wall-clock generation time. For each generated field, we indicate its physical content, its box size and the number of voxels. We also indicate the method that was used to generate the said volume along with the number of denoising steps used.

Use case	Generated volume	Method	Denoising steps T	Wall-clock time
CAMELS				
T_g from ρ_{DM}	T_g , 25 Mpc h^{-1} , 256^3	Independent boxes	400	~ 6 min
T_g from ρ_{DM}	T_g , 25 Mpc h^{-1} , 256^3	Cosmo-FOLD	400	~ 6 min
T_g from ρ_{DM}	T_g , 25 Mpc h^{-1} , 256^3	LODI	400	~ 11 min
IllustrisTNG				
T_g from ρ_{DM} (upsampling)	T_g , 205 Mpc h^{-1} , 1024^3	Cosmo-FOLD	200	~ 10 h
ρ_{DM} from ρ_{star} (upsampling)	ρ_{DM} , 205 Mpc h^{-1} , 1024^3	Cosmo-FOLD	100	~ 5 h
Same as above + $\mathbf{p}(\mathbf{x})$	ρ , 205 Mpc h^{-1} , 1024^3	Cosmo-FOLD+pos	100	~ 5 h
CAMELS $\rightarrow$ IllustrisTNG transfer				
ρ_{DM} from ρ_{star} (upsampling)	ρ_{DM} , 205 Mpc h^{-1} , 1024^3	Cosmo-FOLD	100	~ 5 h

of **Cosmo-FOLD** in patching together highly structured fields without degrading the generative performance of the base model. To further study these results, we compared the generation of a 256^3 voxels, $25 h^{-1}$ Mpc a side **CAMELS** simulation using different patching strategies, providing a direct comparison with simpler inpainting approaches (Fig. 3.3).

In the second case, we considered the generation of the IllustrisTNG dark matter density field conditioned on the stellar density field. We first trained and validated the model on a small subset of the simulation volume (approximately 1.5% of the full IllustrisTNG box) and demonstrated that **Cosmo-FOLD** is able to reconstruct the full dark matter field, including filamentary large-scale structures, while recovering the power spectrum to within 10% accuracy up to $k \leq 5 h \text{ Mpc}^{-1}$. By augmenting **Cosmo-FOLD** with positional encodings, we further showed that 3-point statistics, such as the bispectrum and reduced bispectrum, are reproduced accurately across a range of triangular configurations, with clear improvements over the case without positional information (Fig. 3.10). Finally, we demonstrated the transferability of the approach by generating the same IllustrisTNG dark matter field using a model trained on small box of $25 h^{-1}$ Mpc a side **CAMELS** simulations, as shown in (Fig. 3.7).

We emphasize that the positional encoding strategy explored in this pipeline should be regarded as a proof of principle, demonstrating that the inclusion of explicit spatial information can substantially improve the recovery of higher-order statistics. The mild degradation observed in the power spectrum at intermediate scales arises from directly concatenating positional encodings at the voxel level, which can introduce spurious structure. This effect is not fundamental to the approach and may be mitigated through more structured methods of incorporating positional information that avoid injecting artificial power.

We also stress that these two test cases are particularly challenging, since the way in which the sub-grid small scale physical processes are coupling small scales to larger environment will be radically different. For example, it is envisaged that galactic feedback is likely to generate large temperature fluctuations and eject hot/warm baryons, along path of least resistance and pollute cosmic voids. On the contrary, the physics of star formation will be sensitive to gas properties well inside virialized haloes, and could be regarded as a more local physical effect. The fact that our method manages to capture both scenarios with reasonable accuracy could be regarded as a great success.

Moreover, the generation times reported in Table 3.1 can be substantially reduced by decreasing the number of denoising steps T and/or by lowering the target resolution,

depending on the requirements of a given application. As an illustrative example, reducing the number of denoising steps to $T = 50$ and the output resolution from 1024^3 to 512^3 voxels leads to only a modest degradation of small scale features, while reducing the total generation time by approximately a factor of 16. In this configuration, a full field can be generated in approximately 19 minutes, while going from the maximum theoretical resolved wave mode from $k_{\text{max}} \simeq 31 \, h \, \text{Mpc}^{-1}$ to a $k_{\text{max}} \simeq 16 \, h \, \text{Mpc}^{-1}$, which remains higher than the smallest scales that the model can reliably predict.

Lastly, although an extension to apply **Cosmo-FOLD** technique has to be explored in future works, a proof of concept application that generalizes to different cosmology has been presented in the next chapter.

Chapter 4

Posterior Predictive Fields under Planck Cosmology

4.1 Introduction

The relationship between the dark matter density field and the observable baryonic tracers such as galaxies, stars, or diffuse gas is fundamentally uncertain, shaped by astrophysical processes (supernova and AGN feedback, star formation) that are only approximately captured by sub-grid models in modern hydrodynamical simulations (Villaescusa-Navarro et al. 2022). Recovering the underlying dark matter distribution from these biased tracers is therefore an inverse problem: for a given observed stellar or galaxy field, many plausible dark matter configurations are consistent with the data, and the astrophysical parameters governing the tracer to matter connection are themselves unknown. A principled solution requires sampling from a full posterior distribution over dark matter field configurations, rather than producing a single point estimate.

Field-level inference in cosmology. As discussed in Chapter 1 (Section 1.1.7), compressing the 3D density field into low order summary statistics discards a substantial fraction of the cosmological information encoded in the non Gaussian, non-linear regime of structure formation (Leclercq & Heavens 2021; Jasche & Wandelt 2013). *Field-level inference* instead aims to extract information directly at the level of the 3D field itself. Early approaches to this problem relied on explicit forward models of gravitational dynamics coupled with Hamiltonian Monte Carlo or related samplers to explore the resulting high-dimensional posterior, most notably the *Bayesian Origin Reconstruction*

from *Galaxies* (BORG) algorithm (Jasche & Wandelt 2013). The BORG program has since been applied directly to real survey data, recovering the three dimensional matter density field underlying the SDSS-III BOSS galaxy sample (Lavaux et al. 2019), and, most recently, accelerated by replacing its expensive N -body forward model with a machine learned, field-level emulator (Doeser et al. 2024), a direct precedent for combining a learned model with Bayesian field-level posterior sampling in the way this chapter also pursues. Closely related, but methodologically distinct, extensions target efficient sampling at scale (Bayer et al. 2023), robustness to model misspecification (Shao et al. 2023), and, building on the same BORG framework, joint constraints on primordial non-Gaussianity (Andrews et al. 2023). These methods are powerful but computationally demanding, as each posterior draw typically requires an expensive forward simulation or gradient evaluation through a differentiable simulator, in contrast to the amortized, single-network approach pursued below.

Generative field-level inference. Early machine learning approaches to the dark matter galaxy connection used neural networks to learn point estimates of this mapping (Modi et al. 2018). More recently, diffusion based generative models have been used to go beyond point estimates and directly sample the posterior distribution of dark matter fields conditioned on biased tracers. Closely related to the present work, Ono et al. (2024) trained a diffusion model on the CAMELS suite to sample dark matter density fields conditioned on stellar mass fields, explicitly marginalizing over the uncertain cosmological and astrophysical parameters of the simulation suite. Diffusion priors have also been successfully deployed as posterior samplers for weak-lensing mass mapping in three dimensions (Zhao et al. 2026). Our thesis also follows up on the general framework of diffusion posterior sampling for inverse problems (Chung et al. 2023) that we present in FOREST, in Chapter 5.

This work. This short chapter presents a possible and useful application and extension of the diffusion model and the Cosmo-FOLD method developed in the previous chapters. This can be thought of as a working example with a machine learning model that can work with a slightly more general cosmology case, of course while being in the realm of simulations for the time being.

In this chapter, we extend the Cosmo-FOLD architecture introduced in Chapter 3 to the problem of field level posterior sampling. Rather than using the variational diffusion model purely as a generator, we use the same conditional architecture to

draw samples from the approximate posterior $p_\phi(\boldsymbol{\rho}_{\text{DM}} \mid \boldsymbol{\rho}_{\text{star}}, \Omega_m, \sigma_8)$, trained on the CAMELS Multifield Dataset (Villaescusa-Navarro et al. 2022) Latin hypercube suite so that the model implicitly marginalizes over the astrophysical feedback parameters at training time. Building on this, we introduce a scheme to incorporate external observational constraints on cosmology, specifically, a Planck-derived posterior on Ω_m and σ_8 directly into the generative sampling procedure via Bayes’ theorem, allowing the resulting dark matter posterior samples to be consistent with both the observed tracer field and external cosmological data. We validate the resulting posterior samples at the voxel level and discuss the implications for reconstructing the dark matter distribution from observational tracers.

4.2 Methodology

In this section, we discuss both the dataset used and the methodology implemented.

4.2.1 CAMELS Multifield Dataset

The CAMELS Multifield Dataset (CMD) is a collection of preprocessed 2D maps and 3D grids, containing many different properties of gas, dark matter, and stars from 2,000 distinct simulated universes at several cosmic times. These simulations contain both gravity only N-body simulation suite using GADGET-3, as well as several magnetohydrodynamical simulation suites; the IllustrisTNG CMD suites were produced using the AREPO code to solve the same sub-grid physics models as the original IllustrisTNG simulations. This is the one that we use for our following work. Specifically, we work with the Latin Hypercube (LH) subset of the CMD. It contains 1,000 simulations sampling a six-dimensional parameter space with a Latin hypercube design, ensuring efficient, space-filling coverage without the combinatorial cost of a full grid. The parameters varied are:

4.2.2 Training and validation split

We use 700 simulations from the CMD set of Latin Hypercube (LH) to train and validate, and the remaining 300 to test the results on. Each simulation field of 128^3 voxels is chunked into 64^3 voxels. A total of 24 different flips and rotations are applied to the data and augmented as well. In the end this gives us a total of 134,400 data points, out

Table 4.1: Cosmological and astrophysical parameters varied in the CAMELS Latin Hypercube (LH) set. The fiducial values ($\Omega_m = 0.3$, $\sigma_8 = 0.8$, $A_i = 1$ for all astrophysical parameters) correspond to the original IllustrisTNG calibration and sit at the center of the sampled range.

Parameter	Range	Sampling	Physical meaning
Ω_m	[0.1, 0.5]	linear	Present-day matter density parameter
σ_8	[0.6, 1.0]	linear	Amplitude of linear fluctuations at $8 h^{-1}$ Mpc
A_{SN1}	[0.25, 4.0]	logarithmic	Norm of galactic wind energy (SN feedback)
A_{SN2}	[0.5, 2.0]	logarithmic	Norm of galactic wind speed (SN feedback)
A_{AGN1}	[0.25, 4.0]	logarithmic	Norm of AGN feedback energy
A_{AGN2}	[0.5, 2.0]	logarithmic	Norm of AGN feedback burstiness

of which 75% is used for training and 25% for validating.

4.2.3 Network modifications for training

To get around the problem of cosmic variance that comes from the random initial condition used to generate the cosmic initial density field, we modify the previously mentioned **Cosmo-FOLD** by conditioning the generative model on the large scale modes of the simulation field. This is done by taking the full 128^3 voxels, $25^3(h^{-1} \text{ Mpc})^3$ volume, and then average via pooling every block of $8 \times 8 \times 8$ voxels into a new voxel, to get a resulting coarse grained downsampled 16^3 voxels field ρ_{star}^L . This field is then concatenated via convolutional layers to the denoising Unet of the diffusion model of **Cosmo-FOLD**. Note that this means that there is a unique ρ_{star}^L field corresponding to all the 24×8 sub training chunks coming from each 128^3 ρ_{star} field (factor of 24 from the rotations for augmentations in the training data). In the end, the **Cosmo-FOLD** is conditioned on $\{\rho_{\text{star}}^L, \rho_{\text{star}}, \Omega_m, \sigma_8\}$, to produce ρ_{DM} fields. Note that from now on, in this work, whenever we condition on ρ_{star} , we are jointly conditioning over $\{\rho_{\text{star}}^L, \rho_{\text{star}}\}$. The criteria for saving the checkpoint stays the same as the previous work of **Cosmo-FOLD**. We save those weights for which the ELBO (Evidence Lower Bound) of the data is maximized, similar to the case of our **Cosmo-FOLD** work.

Conditioning mechanism

All three sources of global conditioning information, the diffusion time step, the cosmological parameters, and the large scale context field, are injected into the denoising network through a single, shared *additive* conditioning vector, rather than through

separate pathways. This conditioning vector is constructed as follows.

The time-step embedding $\boldsymbol{\tau}^t$ is first passed through a small MLP to obtain a time-conditioning vector,

$$\mathbf{c}_t = \text{MLP}_t(\boldsymbol{\tau}^t) \in \mathbb{R}^{4d}, \quad (4.1)$$

where d is the base embedding dimension of the network (for our work, $d = 64$). This has already existed in the previous implementations of **Cosmo-FOLD** and **LODI**. As a new modification, the cosmological parameters $\boldsymbol{\theta}_c = \{\Omega_m, \sigma_8\}$ are separately embedded through a two-layer MLP,

$$\mathbf{c}_\theta = W_2 \text{SiLU}(W_1 \boldsymbol{\theta}_c) \in \mathbb{R}^{4d}, \quad (4.2)$$

where $W_1 \in \mathbb{R}^{4d \times 2}$ and $W_2 \in \mathbb{R}^{4d \times 4d}$ are learnable weight matrices. The large scale context field $\boldsymbol{\rho}_{\text{star}}^L \in \mathbb{R}^{1 \times 16 \times 16 \times 16}$ is encoded by two 3-dimensional convolutional layers followed by global average pooling over the spatial dimensions,

$$\mathbf{c}_{\text{ctx}} = W_{\text{ctx}} \left\langle \text{Conv3D} \left(\text{SiLU} \left(\text{Conv3D}(\boldsymbol{\rho}_{\text{star}}^L) \right) \right) \right\rangle_{xyz} \in \mathbb{R}^{4d}, \quad (4.3)$$

where $\langle \cdot \rangle_{xyz}$ denotes averaging over the three spatial axes and W_{ctx} is a learnable, zero-initialized linear projection.

The three embeddings are then combined by simple addition into a single conditioning vector,

$$\mathbf{c} = \mathbf{c}_t + \mathbf{c}_\theta + \mathbf{c}_{\text{ctx}} \in \mathbb{R}^{4d}, \quad (4.4)$$

which is passed identically to every residual block of the U-Net (Sec. 2.4.2), where it is projected to the appropriate channel dimension and injected additively, as illustrated in Fig. 2.5. Because the cosmological embedding $\mathbf{c}_\theta$ enters through the same additive pathway as the time-step and context embeddings, at inference time Ω_m and σ_8 can be freely varied per sample without retraining, which is what enables the Monte Carlo marginalization over the Planck posterior described in Sec. 4.3.

The zero-initialization of W_{ctx} (and of the final convolution in every residual block, Sec. 2.4.2) ensures that at the start of training, the additional context and cosmology conditioning pathways contribute no signal, so the network initially behaves as an unconditional (or minimally conditioned) denoiser and gradually learns to make use of the extra conditioning information as training progresses.

4.3 Sampling from dark matter posterior field

After training our diffusion model, we can draw samples from the learned approximate posterior distribution $p_\phi(\boldsymbol{\rho}_{\text{DM}}|\boldsymbol{\rho}_{\text{star}}, \Omega_m, \sigma_8)$, because in our training, the model learns to implicitly marginalize over the astrophysical parameters:

$$p_\phi(\boldsymbol{\rho}_{\text{DM}}|\boldsymbol{\rho}_{\text{star}}, \Omega_m, \sigma_8) \approx \int p_\phi(\boldsymbol{\rho}_{\text{DM}}|\boldsymbol{\rho}_{\text{star}}, \Omega_m, \sigma_8, \boldsymbol{\theta}_a) p(\boldsymbol{\theta}_a) d\boldsymbol{\theta}_a \quad (4.5)$$

where $\boldsymbol{\theta}_a$ are the astrophysical parameters and $p(\boldsymbol{\theta}_a)$ is the training prior over the astrophysics parameters. In our LH set, we have a uniform prior over the linearly or log sampled $\boldsymbol{\theta}_a$ space, which can be seen in table 4.1. In practice, the above expression is an approximation whose accuracy depends on how well the network captures the conditional distribution across the full training range of $\boldsymbol{\theta}_a$. We also note that we are implicitly treating the training distribution over $\boldsymbol{\theta}_a$ as a proxy for the true prior over astrophysical feedback strength, an assumption we make on the CAMELS LH set.

Now, to sample the conditional posterior $p(\boldsymbol{\rho}_{\text{DM}}|\boldsymbol{\rho}_{\text{star}})$, conditioned on an unseen $\boldsymbol{\rho}_{\text{star}}$, while incorporating external known cosmological constraints, we proceed as follows. We consider the case where these cosmological parameters Ω_m and σ_8 are themselves uncertain, constrained only through external data $\mathcal{D}_{\text{planck}}$:

$$p_\phi(\boldsymbol{\rho}_{\text{DM}}|\boldsymbol{\rho}_{\text{star}}, \mathcal{D}_{\text{planck}}) = \int p_\phi(\boldsymbol{\rho}_{\text{DM}}|\boldsymbol{\rho}_{\text{star}}, \boldsymbol{\theta}_c) p(\boldsymbol{\theta}_c|\mathcal{D}_{\text{planck}}) d\boldsymbol{\theta}_c \quad (4.6)$$

where we use the data driven posterior $p(\boldsymbol{\theta}_c|\mathcal{D}_{\text{planck}})$ from Planck observations to act as a prior that we sample from, to marginalize over the possible realizations of $\boldsymbol{\theta}_c$, where $\boldsymbol{\theta}_c = \{\Omega_m, \sigma_8\}$. Specifically, we use the `base_plikHM_TTTEEE_lowl_lowE_lensing` MCMC chain from the Planck Legacy Archive, corresponding to the baseline Λ CDM model constrained with the full *Planck* temperature, polarization (TT,TE,EE), low- ℓ likelihood, and CMB lensing data (Collaboration et al. 2020). This two-step approach is computationally advantageous: rather than retraining the network whenever external cosmological constraints are updated, we reuse a single model trained on the full LH parameter range and reweight at inference time via the expression above, analogous to a Bayesian update on a fixed likelihood.

We approximate this integral via Monte Carlo sampling. Cosmological parameters

θ_c^i are drawn from the Planck MCMC chain and for each θ_c^i we generate a ρ_{DM} field:

$$\begin{aligned}\theta_c^i &\sim p(\theta_c \mid \mathcal{D}_{\text{Planck}}) \\ \rho_{\text{DM}}^i &\sim p_\phi(\rho_{\text{DM}} \mid \rho_{\text{star}}, \theta_c^i)\end{aligned}\tag{4.7}$$

This yields a set of samples that approximate the posterior distribution of dark matter fields at the field level. We then compute voxel wise posterior statistics, including mean and variance, and estimate marginal distributions using histograms across samples. For our work, we use 1000 generated posterior samples for the same conditional stellar field. To generate each sample, we first generate 64^3 voxel sub-volume for each of the realization, and then patch 8 of those via **Cosmo-FOLD** to create the full 128^3 voxels realization. We note that the ρ_{star} used at inference, corresponds to underlying cosmological parameters that lie within the support of the Planck posterior, so that all generated samples remain within the domain of validity of the prior. Hence we *assume* for this case that the ρ_{star} field we observe belongs to the Planck cosmology, which for this work is proof of concept assumptions. Ideally, we would replace this with something we observe in reality.

4.4 Results

In this section, we present the results of our method, to show that we can sample high dimensional field level posteriors, given data constraints and conditioned on baryonic observables, which in this case is ρ_{star} . Note that for this work, we use ρ_{star} that corresponds to the Planck best fit Cosmology. It can be further extended to account for different cosmology cases. In figure 4.1, we showcase generated 3 dimensional simulation cubes. We use a previously unseen ρ_{star} , along with the global context ρ_{star}^L and sampled θ_c^i , to generate ρ_{DM}^i (see Eq: 4.7). One approximate sampled posterior map $p_\phi(\rho_{\text{DM}} \mid \rho_{\text{star}}, \theta_c)$ is shown on the top right side of the plot. For reference, the ground truth dark matter field is shown on the top plot of the middle row. After we repeatedly sample via Eq 4.7, we obtain the posterior distribution of voxels, from which we plot the posterior mean in the bottom of the middle row and the standard deviation in the bottom right. As can be seen, the posterior mean washes out the small scale variability at the voxel level and the local filaments, but preserves the large scale filamentary structures, while the standard deviation field is more noisy and captures the voxel wise variability. All the plots take the sparse stellar field, shown on the left, as a conditional.

To visualize single voxel posteriors, in Fig. 4.3 we plot sampled posterior distribution

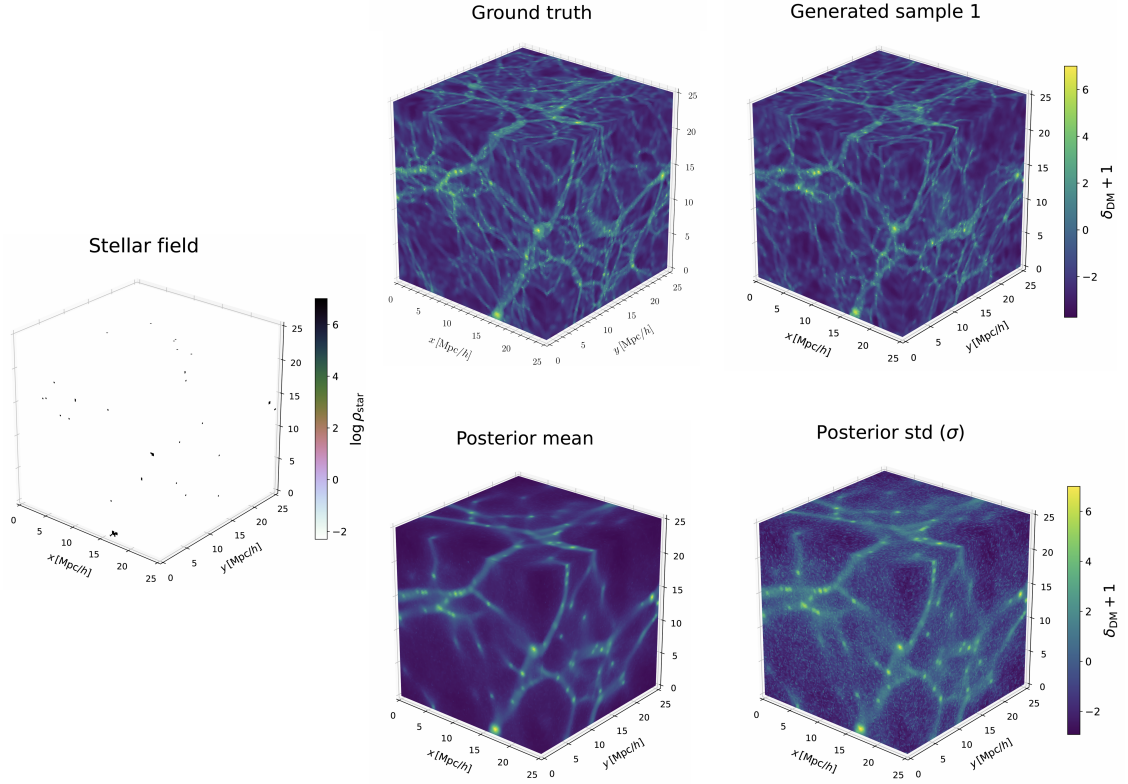


Figure 4.1: Depiction of the sampled field level dark matter posterior for an unseen test simulation. The conditioning stellar field is shown on the left side of the plot while the corresponding ground truth dark matter field is shown on the top of the middle row. A generated dark matter field with our diffusion model is shown on the top right, while the average and standard deviation of 1000 such samples is depicted in the lower plot of the middle and right row respectively.

for 3 voxels of different density, along with the ground truth values. We see that the predicted posterior mass sits well with the ground truth voxel density values. To better quantify the quality of the generated posteriors, we also plot standard residual σ_{res} per voxel as a histogram, where σ_{res} corresponds to:

$$\sigma_{\text{res}}^i = \frac{|\overline{\text{DM}}_{\text{pred}}^i - \text{DM}_{\text{true}}^i|}{\sigma_{\text{pred}}^i}, \quad (4.8)$$

where $\overline{\text{DM}}_{\text{pred}}^i$ is the predicted posterior mean for a voxel i and the $\text{DM}_{\text{true}}^i$ is the ground truth value for the same. Meanwhile σ_{pred}^i is the standard deviation of the predicted posterior of that voxel. We plot σ_{res} for two distinct cases in Fig. 4.2, where the orange case is for underdense dark matter regions while the blue is for the overdense regions.

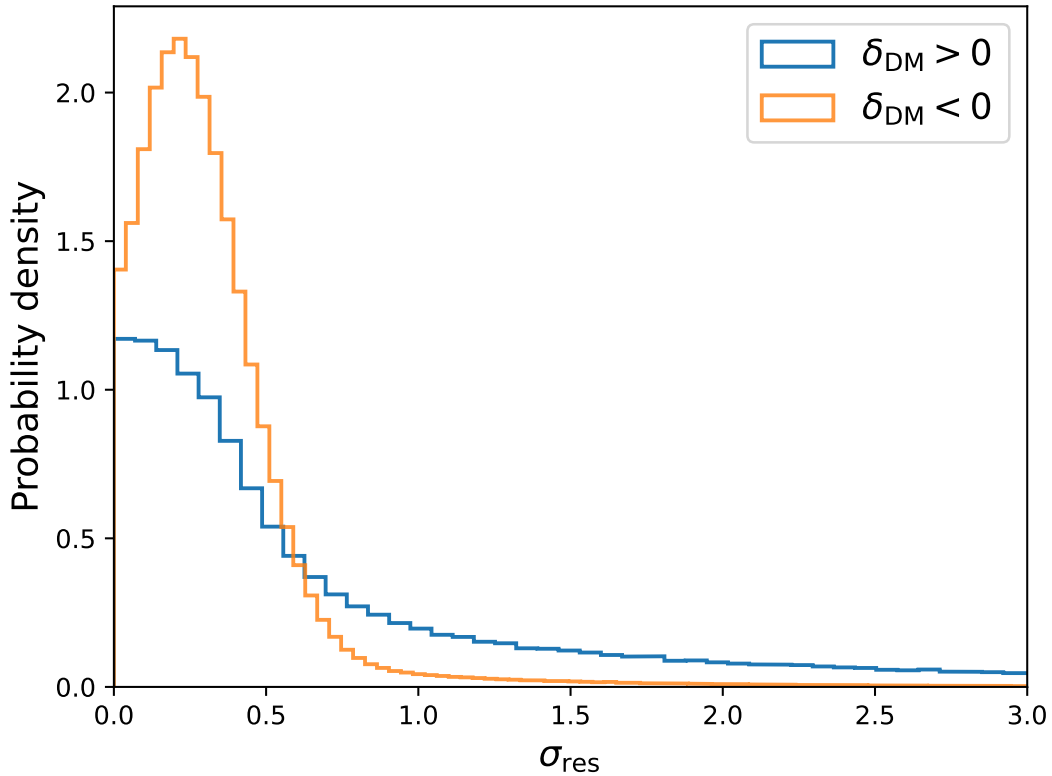


Figure 4.2: Posterior residual histograms(see Eq. 4.8) for under dense (orange) and over dense(blue) voxels, plotted for an aggregate of 1000 generated dark matter samples.

4.5 Discussion

In this chapter, we have demonstrated that a conditional variational diffusion model, trained on the CAMELS Latin Hypercube suite, can be used to sample field-level

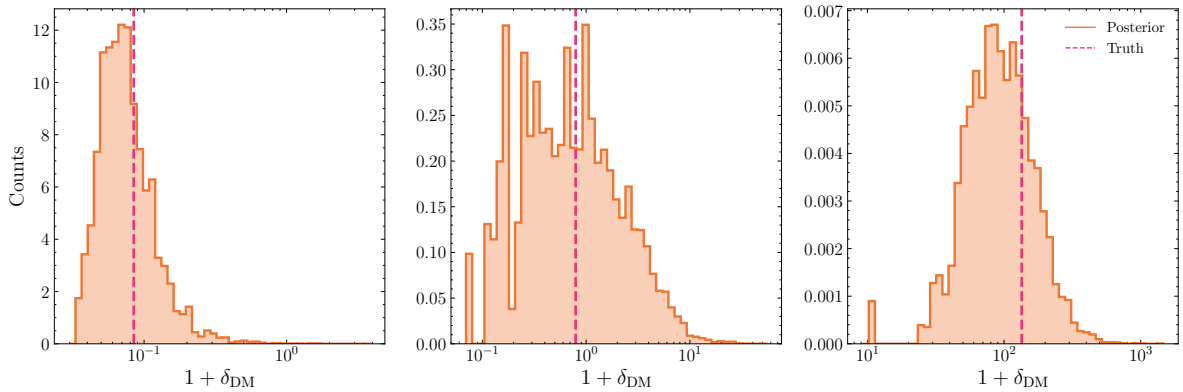


Figure 4.3: Sampled posterior with 1000 sampled points, generated via the diffusion model, for 3 different voxels. For each posterior, the ground truth value is shown as a vertical dashed red line.

posterior distributions of the dark matter density given an observed stellar mass field. Rather than producing a single point estimate of the underlying dark matter distribution, our approach yields a full ensemble of plausible three dimensional realizations, from which we can extract voxel wise posterior statistics such as the mean and standard deviation, as shown in Fig. 4.1. This is one of the first demonstrations of field level posterior sampling for the dark matter stellar connection using a variational diffusion model conditioned jointly on a local tracer field, its large scale context, and external cosmological information.

A key feature of our approach is that the astrophysical uncertainty inherent in the stellar to dark matter mapping that are only approximately captured by sub grid models is handled implicitly, by training across the full range of astrophysical parameters sampled by the CAMELS LH set. This allows the network to learn a distribution over plausible dark matter fields for a given stellar field, rather than assuming a single, fixed astrophysical model. At the same time, cosmological parameters are treated differently: rather than marginalizing over them at training time, we retain the ability to condition on them explicitly, which allows us to inject external observational constraints specifically, the Planck 2018 posterior on Ω_m and σ_8 , directly into the sampling procedure via Bayes’ theorem (Sec. 4.3) and Monte Carlo sampling.. This illustrates how a single trained generative model can be reused across different external priors without retraining.

The voxel level diagnostics presented in Sec. 4.4, the single-voxel posterior distributions (Fig. 4.3) and the residual histograms split by density regime (Fig. 4.2) indicate that the predicted posterior mass is generally consistent with the ground truth dark

matter values, in both under dense and over dense regions. We also observe, qualitatively, that the posterior mean field preserves large scale filamentary structure while smoothing over small-scale, voxel-level variability; this is an expected consequence of averaging over a stochastic mapping in which many small-scale configurations are equally consistent with a given large scale and stellar-field context, and is not indicative of a failure of the underlying model.

We stress that the results presented here are based on a single test field, and should be regarded as a first, encouraging demonstration of the method rather than a comprehensive validation. Unlike the BORG algorithm and its extensions, which have already been applied to real observational data such as the SDSS-III BOSS survey (Lavaux et al. 2019), the results presented here remain entirely simulation based. Extending the present approach to real observational tracer fields, with their attendant survey masks, shot noise and selection effects, is an important direction for future work. An important limitation here is that we are sampling for a $\boldsymbol{\rho}_{\text{star}}$ which already corresponds to the Planck Cosmology, and so using Planck MCMC chains are reasonable here. This is an easier test because the true cosmology sits at the Planck posterior mode. A general, more random $\boldsymbol{\rho}_{\text{star}}$ would require generating a conditional $\boldsymbol{\rho}_{\text{DM}}$ along with an efficient likelihood evaluation technique (ELBO score from the diffusion model, for example) for MCMC or some gradient based methods for efficient parameter sampling. Also, a more thorough assessment of the calibration of the generated posteriors including coverage tests is left for future work.

Chapter 5

FOREST: Diffusion Posterior Sampling for Field Level Posterior Reconstruction

5.1 Introduction

As mentioned in the general introduction of the thesis, the intergalactic medium (IGM) at high redshift holds the majority of the Universe’s baryonic content, and on megaparsec scales its distribution traces the underlying dark matter field shaped by gravitational instability (Meiksin 2009; McQuinn 2016; Bi & Davidsen 1997). The Lyman- α (Ly α) forest, the pattern of absorption imprinted by intervening neutral hydrogen on the spectra of background quasars and galaxies, remains the most direct observational window onto this diffuse structure (Gunn & Peterson 1965; Lynds 1971; Williger et al. 2000). For over two decades, sufficiently dense grids of background sightlines have made it possible to move beyond one dimensional flux statistics and reconstruct the IGM tomographically in three dimensions (Cucci et al. 2008; Lee et al. 2014, 2018), a program now being extended by surveys such as DESI Ly α QSO, WEAVE-QSO, and Subaru-PFS/CLAMATO/LATIS, and, in the coming decade, by ELT/MOSAIC, which will push sightline separations down toward $1 h^{-1}$ Mpc. At these finer separations, the reconstruction problem becomes increasingly sensitive to the baryonic physics, pressure smoothing, thermal broadening, and radiative feedback, that decouples the IGM from the pure dark matter field below $\sim 1 h^{-1}$ Mpc (Schaye et al. 2000; Viel et al. 2002; Nasir et al. 2017), motivating reconstruction methods that can go beyond the linear, large

scale regime.

Classical tomographic reconstruction has relied on Wiener filtering or Gaussian process interpolation of the sparse sightline data (Lee et al. 2014, 2018), which is efficient but smooths out precisely the small scale, non linear structure that upcoming surveys are designed to resolve. More recent semi-analytic approaches, notably TARDIS (Horowitz et al. 2019, 2021), improve on this by embedding a physically motivated prior, nonlinear structure evolved forward with FastPM (Feng et al. 2016), inside an iterative maximum-likelihood scheme, using the Fluctuating Gunn-Peterson Approximation (FGPA) as the forward model linking density to absorption. This is a more faithful treatment of gravitational nonlinearity, but it inherits the limitations of FGPA at small scales and is computationally expensive, since every posterior sample requires re-running the gravity solver.

The last few years have seen deep generative models offer a third route: rather than approximating the forward problem or the prior analytically, learn one or both directly from hydrodynamical simulations. Variational autoencoders (VAE), GANs, Transformers and diffusion models have all been used to emulate non-linear cosmological fields. Moriwaki et al. (2025) have applied generative Transformers to generate galaxy properties which enable fast generation of intensity maps and redshift survey catalogs. Horowitz et al. (2022) showed that a VAE trained on simulations can learn the mapping between dark matter and baryonic tracers directly, without any FGPA-like approximation. This is the paradigm adopted by DeepCHART (Maitra et al. 2026), developed and trained on the same $z = 2.5$, $(40 h^{-1}\text{cMpc})^3$ Sherwood hydrodynamical simulations, also used throughout this chapter. DeepCHART uses a 3D U-Net VAE to map sparse Ly α forest sightlines, optionally combined with coeval galaxy positions, directly onto the 3D dark matter density field in a single amortized forward pass, mapping dark matter fields at a transverse sightline separation of $d_{\perp} \simeq 2.4 h^{-1}\text{cMpc}$ matched to the Subaru-PFS Deep survey.

5.2 Novelty: Generative prior

DeepCHART and the approach developed in this chapter are fundamentally different. An amortized VAE such as DeepCHART is trained end to end on paired (sightline, δ_{DM}) data for a specific pairing of survey geometry and noise model. Once trained, inference is near instantaneous, but the learned mapping is strongly coupled with the forward model

used during training. So a new sightline density, or instrumental resolution requires retraining, and the network delivers a single deterministic point estimate of δ_{DM} per input.

FOREST (Field-level pOsterior REconstruction via Sightline Tomography) is a technique that builds on the diffusion posterior sampling (DPS) framework (Chung et al. 2023), in which a diffusion model learns an unconditional latent space generative prior over the baryonic and δ_{DM} fields themselves, trained once, without any paired flux data, and independently of any particular observational setup. At inference time this prior is combined with an explicit, differentiable forward model through a likelihood guidance term added to the reverse diffusion process, so that a single pretrained prior can be redeployed across different observational setups without retraining, yielding genuine posterior samples rather than point estimates. We train a latent version of the variational diffusion model (VDM) prior on a compressed latent space of the five physical fields, dark matter density (δ_{DM}), gas density (Δ_{gas}), neutral hydrogen fraction (x_{HI}), temperature (T), and peculiar velocity (v_{pec}), extracted from the same Sherwood hydrodynamical simulations underlying DeepCHART, and pair it with a fully differentiable, GPU-vectorized Voigt profile forward model that maps these fields onto synthetic Ly α forest spectra. Posterior samples of the joint 3D field are then drawn by guiding the reverse diffusion process with the gradient of the data likelihood at each denoising step, following the DPS formalism, with a scheduled guidance strength ζ added to stabilize sampling in three dimensions. A notable feature of this forward model is that δ_{DM} itself never enters the optical depth calculation directly, only Δ_{gas} , x_{HI} , T , and v_{pec} do, so any recovery of the dark matter field is mediated entirely by correlations the prior has learned between dark matter and the gas phase tracers, rather than by a direct likelihood constraint. This makes the dark matter reconstruction a genuine test of what the learned prior encodes about structure formation.

This chapter is organized as follows. Section 5.3 describes the construction of the 3D training dataset from the Sherwood simulations to train the **FOREST** pipeline. Section 5.4 details the architecture and training of the variational diffusion prior, together with the frozen VAE used in the **FOREST**’s latent formulation. Section 5.5 presents the differentiable Voigt forward model. Section 5.6 lays out the diffusion posterior sampling algorithm itself, the Bayesian decomposition into score and likelihood guidance terms, the Tweedie estimate, and memory-efficient gradient computation, as adapted to this

three-dimensional, multi-channel setting. Section 5.7 presents reconstruction results, voxel wise fidelity for all five fields, and recovery of the matter power spectrum using FOREST. These results are preliminary in the specific sense discussed in Section 5.8.2, since for the time being they are tested on a $40 \times 10 \times 10 (h^{-1} \text{ Mpc})^3$ sub-volume of the full simulation box rather than the complete box $40^3 (h^{-1} \text{ Mpc})^3$, with the full-box extension set out as next future work in Section 5.8.3. Lastly, Section 5.8 discusses the implications of these results, including what the free (likelihood blind) recovery of δ_{DM} implies about the prior, and directions for future work.

5.3 Data Preparation

5.3.1 Hydrodynamical simulations and line-of-sight fields

The training and validation data for this chapter are drawn from the same suite of Sherwood Relics hydrodynamical simulations (Bolton et al. 2017; Puchwein et al. 2023) used by Maitra et al. (2026), for direct comparability. These simulations are run with GADGET-3, self-consistently solving the coupled photoionization and thermal evolution equations of the gas under a specified ultraviolet background, within a periodic cube of comoving side $L_{\text{box}} = 40 h^{-1} \text{ Mpc}$ populated with 2×512^3 dark matter and gas particles. Eight random-seed realizations are used for training and validating the diffusion prior and two are reserved for testing.

Rather than gridding the raw particle data directly, physical quantities are first evaluated along one dimensional sightlines threading the box, in a configuration matching Ly α forest tomography surveys. At redshift $z = 2.5$, a sightline is discretized into n_{bins} velocity/position bins, and at each bin the local gas overdensity Δ_{gas} , temperature T , neutral hydrogen fraction x_{HI} , and line of sight peculiar velocity v_{pec} are obtained via standard SPH kernel interpolation from the neighboring gas particles,

$$f_i = \sum_j f_j \frac{m_j}{\rho_j} W(|\mathbf{r}_i - \mathbf{r}_j|, h_j), \quad (5.1)$$

where f_j , m_j , and ρ_j are the quantity, mass, and density carried by particle j , and W is the standard cubic-spline smoothing kernel (Monaghan 1992; Maitra et al. 2026). The neutral hydrogen fraction x_{HI} and temperature T carried by each gas particle are set on the fly during the simulation by GADGET-3’s coupled photoionization and thermal evolution solver under the ultraviolet background of Puchwein et al. (2023), including

self-shielding corrections that suppress photoionization in dense gas; these particle-level values are then interpolated onto the sightline grid via Eq. 5.1 exactly as Δ_{gas} and v_{pec} are. The dark matter overdensity δ_{DM} is obtained analogously from the dark matter particle distribution and gridded onto the same sightline geometry, but, as emphasized in Section 5.1, it is retained purely as a target field for the diffusion prior and never enters the optical depth calculation described in Section 5.5.

5.3.2 Sightline data format

Each simulation output is stored as a flat binary file containing one full set of sightlines through the box. The sightline file layout is summarized in Table 5.1.

Table 5.1: Binary layout of a single sightline data file. All floating-point values are `float64`. The bins/sightline counts are `int32`.

Field	Shape	Description
$z, \Omega_m, \Omega_\Lambda, \Omega_b, h, L_{\text{box}}, X_h$	7 scalars	Simulation header
$n_{\text{bins}}, n_{\text{los}}$	2 scalars	Pixels per sightline; number of sightlines
xlos, ylos, zlos	$(n_{\text{los}},)$	Sightline axis and transverse position
posaxis, velaxis	$(n_{\text{bins}},)$	Comoving position and Hubble velocity
$\Delta_{\text{gas}}, x_{\text{HI}}, T, v_{\text{pec}}$	$(n_{\text{los}}, n_{\text{bins}})$	SPH-interpolated fields (Eq. 5.1)

The reference optical depth τ_{true} is stored separately as an array of shape $(n_{\text{los}}, n_{\text{bins}})$. For our work, the values of $(n_{\text{los}}, n_{\text{bins}})$ for the full simulations are 16,384 (128×128) and 1,024 respectively. As explained in the later section, we train on smaller chunks of this data, and so the values during training are n_{los} for 1,024 (32×32), with n_{bins} being unchanged.

5.4 The Variational Diffusion Prior

Every other diffusion model in this thesis, **HALOgen** and **LODI** in Chapter 2, **Cosmo-FOLD** in Chapter 3, and its posterior predictive extension in Chapter 4, is a *conditional* model, trained to map one field onto another (dark matter to halos, dark matter to gas temperature, stellar density to dark matter). The prior used in **FOREST** is different in kind: it is *unconditional*, trained to model the joint distribution $p(\mathbf{x}_0)$ of the five-field state itself,

$$\mathbf{x}_0 = (\delta_{\text{DM}}, \Delta_{\text{gas}}, x_{\text{HI}}, T, v_{\text{pec}}), \quad (5.2)$$

extracted from the Sherwood-Relics simulations (Section 5.3.1), with no conditioning information supplied at training time at all. This is precisely what makes the same trained network reusable, at inference time, against different sightline geometries, noise levels, or observational setups (Section 5.6): conditioning on a specific observation only ever enters through the likelihood-guidance term added to the reverse process, never through the network’s own training.

5.4.1 Multi-field representation

The five physical fields of Eq. 5.2 are stacked as five channels of a single volumetric tensor, $\mathbf{x}_0 \in \mathbb{R}^{5 \times N_x \times N_y \times N_z}$, and modeled jointly by one network, rather than trained as five independent single-channel diffusion models. Treating them jointly helps FOREST with learning the implicit cross-field correlations during training, in particular between δ_{DM} and the four gas phase fields that actually enter the forward model of Section 5.5, that make the eventual likelihood-blind recovery of δ_{DM} possible at all (Section 5.6). Five independently trained priors would share no such correlations. The denoising network itself follows the same 3D U-Net with residual blocks and a self-attention bottleneck used throughout this thesis (Sections 1.3.1 and 1.3.2 of the Introduction; Section 2.4.2 of Chapter 2), adapted to use the correct dimensionality input and output.

5.4.2 Latent diffusion and Anisotropic geometry

Because the training volume of $40^3 (h^{-1} \text{ Mpc})^3$ is built from 1D sightlines threading the box (Section 5.3.1) rather than read directly off a native 3D grid, the natural tensor shape is highly anisotropic: The line of sight resolution is $\approx 0.039 h^{-1} \text{ Mpc}$ (Ideally this would be set by the spectral resolution a survey would deliver), against a coarser resolution of $\approx 0.31 h^{-1} \text{ Mpc}$ along the two transverse directions (set by the physical separation between neighbouring sightlines). We denote this native shape $N_{\parallel} \times N_{\perp} \times N_{\perp} = 1024 \times 128 \times 128$. To get multiple samples and also due to memory reasons, we chunk these simulations into $1024 \times 32 \times 32$ sub volumes. Note that we do not chunk the LOS direction since we want the diffusion model to see the full line of sight structure in its training (since we realized that quantities such as v_{pec} have strong long range correlations along LOS, and so the model needs to see them in a single pass to capture them). Running the reverse diffusion process directly on a tensor of this size is computationally demanding, especially if we want to use the attention mechanism

(Sec.1.3.2). We therefore compress the $5 \times 1024 \times 32 \times 32$ training data into a latent representation which has a dimension of $8 \times 128 \times 16 \times 16$, resulting in a 20 times compression of the data. We denote this latent space by $\mathbf{x}_0^L$. This said compression is done via training a separate variational autoencoder (VAE), a general background of which have been introduced in the introduction, in the section 1.3.5. A Variational Diffusion model (2.4.2) is then trained to learn the prior of this latent space $\mathbf{x}_0^L$. The frozen VAE decoder is used only once, at the very end, to map the final denoised latent sample back to the native five-field volume. This is the technique of latent diffusion that is used in FOREST for sightline reconstruction.

Encoding the line of sights

As mentioned, to perform diffusion on the full voxel space, we would need much more GPU memory than available, especially if we wanted to employ the attention mechanism. Additionally, compressing the data removes noisy small scale features from the data, which benefits us since a resolution of $0.039 h^{-1} \text{ Mpc}$ is much smaller than the Jeans scale ($\approx 1 h^{-1} \text{ Mpc}$), and sub-Jeans scales carry negligible correlated physical information that would be smoothed away in the analysis regardless, a practice also implemented in other related works(Maitra et al. (2026)). A third reason is that we found out that training on a compressed data point better preserves the long range correlations, especially in the case of v_{pec} . Last, since training the pipeline in the native resolution on the diffusion model takes days, compressing the data for the diffusion model drastically reduces training time. In our case, the training time went from about 50 hours to about 3 hours total.

Hence, for this work, we decided to do a moderate compression by a factor of 20, a sweet spot between very high compression, which could result in losing too much small scale information, and very light compression, which would result in no potential benefit of this added machinery. The architecture of the autoencoder is shown in Fig. 5.1. An important aspect of the architecture is that the downsampling performed in each encoder block is anisotropic, and is learned via strided convolutions. Across the three downsampling blocks, the transverse dimensions are reduced by a factor of 2 in each direction, while the line-of-sight dimension is reduced by a factor of 8. This choice preserves more of the spatial correlation structure in the transverse directions, which are initially sampled at a lower resolution compared to the line-of-sight direction.

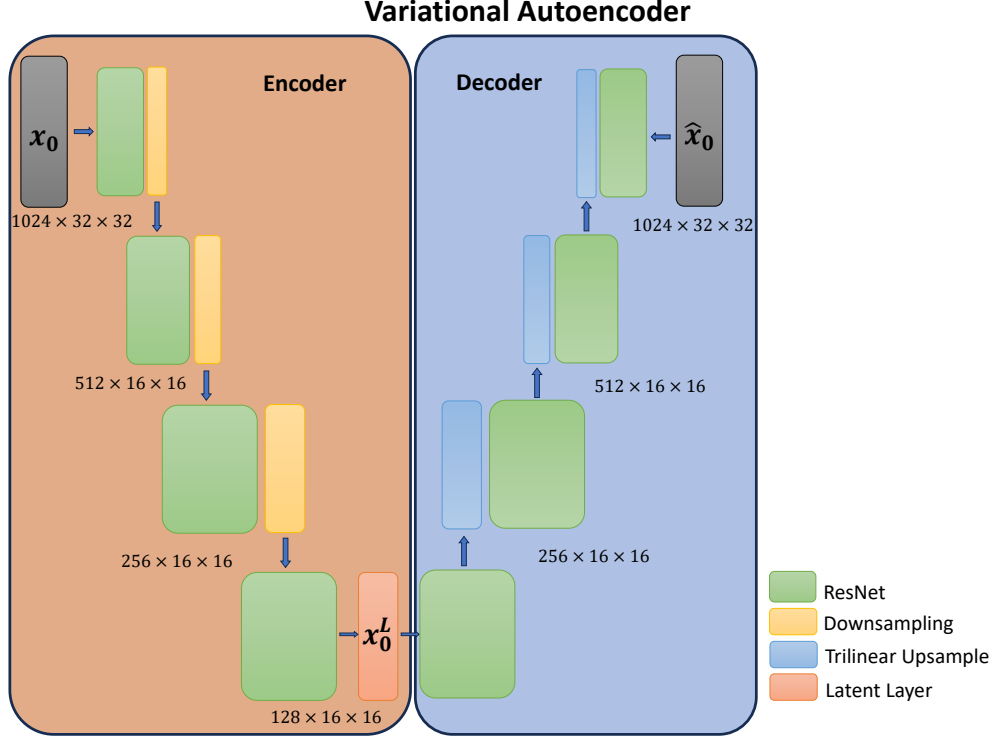


Figure 5.1: The convolution structure of the VAE described in sec. 5.4.2, for the FOREST technique. There are 3 downsampling blocks with the latent vector $\mathbf{x}_0^L$ at the bottleneck. The *spatial* dimensions are mentioned at each level.

The bottleneck of the VAE provides the compressed latent representation of the input data $\mathbf{x}_0$. Rather than encoding each input into a deterministic latent vector, the encoder parameterizes a Gaussian distribution in the latent space by predicting the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ for each latent component. Consequently, the final encoder output contains twice the number of channels required for the latent representation, corresponding to the parameters of this distribution. During training, the latent representation $\mathbf{z}$ is sampled using the reparameterization trick:

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (5.3)$$

5.4.3 Training the FOREST pipeline

The training happens in 2 stages. In the first stage, a VAE (Sec. 5.4.2) is trained on the native resolution $5 \times 1024 \times 32 \times 32$ dimensional $\mathbf{x}_0$ data. After training, the VAE is frozen and the encoder is then used to encode all the training $\mathbf{x}_0$ into $8 \times 128 \times 16 \times 16$

dimensional latent $\mathbf{x}_0^L$. The way the encoding happens is that we take the predicted $\boldsymbol{\mu}$ for each input $\mathbf{x}_0$ and set the corresponding $\boldsymbol{\sigma}$ to zero. This is done to ensure stable diffusion model training and reducing the variance coming from the VAE. From here on, this truncated $\boldsymbol{\mu}$ only latent space will be called $\mathbf{x}_0^L$. The diffusion prior is trained on this $\mathbf{x}_0^L$ with the same variational diffusion model (VDM) objective used throughout this thesis (Section 2.4.2 of Chapter 2): a learned, continuous-time noise schedule $\gamma(t)$ and the ELBO loss (Sec. 2.4.2).

Training the VAE

The VAE is trained in two stages for numerical stability. In Stage 1 (steps 0–2000), we minimize the L1 mean absolute error reconstruction loss and KL divergence (See Eq. 1.26) terms:

$$\mathcal{L}_{\text{stage 1}} = \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_1 + \beta \text{KL}(q_\phi(\mathbf{x}_0^L|\mathbf{x}_0) \| p(\mathbf{x}_0^L)), \quad (5.4)$$

where $\beta = 10^{-4}$ weights the KL term, $\hat{\mathbf{x}}_0$ is the reconstructed native field and $p(\mathbf{x}_0^L) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

In Stage 2 (steps 2,000–4,000), we introduce an additional spectral loss term with linear warmup to preserve fourier modes explicitly:

$$\mathcal{L}_{\text{stage 2}} = \|\mathbf{x}_0 - \hat{\mathbf{x}}_0\|_1 + \beta \text{KL}(q_\phi(\mathbf{x}_0^L|\mathbf{x}_0) \| p(\mathbf{x}_0^L)) + \lambda_{\text{spec}}(t) \mathcal{L}_{\text{spec}}, \quad (5.5)$$

The spectral loss that penalizes mismatch in the 1D power spectrum along the line of sight:

$$\mathcal{L}_{\text{spec}} = \sum_{c=1}^5 w_c \left\langle \left[\log P_{\text{1D,rec}}^{(c)}(k) - \log P_{\text{1D,GT}}^{(c)}(k) \right]^2 \right\rangle, \quad (5.6)$$

where the warmup schedule ramps the spectral weight from 0 to $\lambda_{\text{spec}} = 0.1$ over 500 steps. This warmup ramping helped stabilize training.

$P_{\text{1D,rec}}^{(c)}(k)$ and $P_{\text{1D,GT}}^{(c)}(k)$ are the 1D power spectra along the LoS axis for reconstructed and ground-truth fields respectively (computed via FFT and averaged over the transverse plane), the weights are $\mathbf{w} = [1.0, 1.0, 3.0, 1.0, 1.5]$ for channels $[\delta_{\text{DM}}, \Delta_{\text{gas}}, x_{\text{HI}}, T, v_{\text{pec}}]$. The heavier weight on x_{HI} ($w_2 = 3.0$) ensures faithful reconstruction of the ionization structure which directly couples to the observed Ly α absorption.

For training, we use 8 3D simulations from the Sherwood simulations of side $40(h^{-1} \text{ Mpc})^3$, gridded at $(1024, 128, 128)$ for each field, run with the same astrophysics and cosmology, but with different initial condition seed. Each of the 5 channels, corresponding to $[\delta_{\text{DM}}, \Delta_{\text{gas}}, x_{\text{HI}}, T, v_{\text{pec}}]$, are sub chunked into training samples $\mathbf{x}_0$ arrays of shape $(5, 1024, 32, 32)$. To maximize the number of training data points, we create these chunks with overlapping volumes to get 625 chunks per simulation. This gives us a total of 5,000 sims for training and validating. We do a 70%/30% training/validation split.

Training the latent VDM

Now to train the VDM, we first freeze the encoder of the autoencoder and produce the 5,000 $\mathbf{x}_0^L$ data points. At this point it must be mentioned that we are in fact reusing the same data used for training and validating the VAE, for this setup, since we are limited by the training sample size. Our VDM architecture closely follows Fig. 2.4 and the exact same machinery as described in Sec 2.4.2 and 2.4.2. The only crucial changes are that instead of isotropically pooling the spatial dimensions at every downsampling block, we follow the same idea used in the VAE and perform anisotropic downsampling. There are 3 downsampling blocks with the following compression factors: $(2, 2, 2), (2, 2, 2), (2, 1, 1)$. So the first 2 blocks are isotropically downsampling by a factor of 2 for each axis, but the third block only downsamples the line of sight axis. Besides this, there are 8 input channels to the diffusion model. The training happens for a total of 450 epochs, with a 70%/30% training/validation split. The model weights with the lowest validation loss are stored.

To get back the generated $\mathbf{x}_0$, we use the frozen decoder of the VAE on the diffusion model outputs.

5.5 Theory of the Lyman-alpha Optical Depth

5.5.1 Physical picture

A quasar or star-forming galaxy at redshift z_s illuminates the intervening intergalactic medium along the line of sight. Each parcel of neutral hydrogen between source and observer resonantly scatters photons out of the line of sight near the rest-frame Ly α wavelength $\lambda_\alpha = 1215.6701 \text{ \AA}$. Gas at comoving position x imprints absorption at a

velocity (equivalently, wavelength) offset

$$u(x) = v_{\text{H}}(x) + v_{\text{pec}}(x), \quad (5.7)$$

the sum of the local Hubble velocity $v_{\text{H}}(x) = H(z) x_{\text{phys}}$ and the gas's peculiar velocity along the line of sight, $v_{\text{pec}}(x)$. This redshift-space mapping is what allows v_{pec} to distort the apparent position of absorption features relative to the real-space density peaks that sourced them.

Two independent physical effects broaden the absorption profile of a parcel of gas away from a single frequency:

- **Thermal Doppler broadening.** Hydrogen atoms at temperature T have a Maxwell Boltzmann distribution of line-of-sight velocities, so photons resonant with atoms of different thermal velocity are absorbed at correspondingly shifted frequencies. This produces a Gaussian absorption profile of characteristic width (the Doppler parameter)

$$b = \sqrt{\frac{2k_B T}{m_H}}. \quad (5.8)$$

- **Natural (radiative) damping.** The finite lifetime of the excited $2p$ state, set by the Einstein spontaneous-decay rate $\Gamma_{\text{Ly}\alpha} = 6.265 \times 10^8 \text{ s}^{-1}$, gives the transition an intrinsic Lorentzian linewidth via Heisenberg's energy time uncertainty relation. This produces extended power law wings that fall off as Δv^{-2} rather than exponentially, controlled by the dimensionless damping parameter

$$a = \frac{\Gamma_{\text{Ly}\alpha} \lambda_{\alpha}}{4\pi b}. \quad (5.9)$$

The full line profile is the convolution of the Gaussian Doppler core with the Lorentzian damping wings, the *Voigt profile*. Because $a \sim 10^{-3}$ for Ly α at IGM temperatures, the profile is very nearly Gaussian near line center, with the damping term contributing only a small correction except deep in the wings.

A key consequence of this convolution is that the optical depth observed at a given pixel i is *not* a local function of the density at the corresponding position alone. Gas at every other pixel j along the line of sight contributes to τ at pixel i , weighted by how far j 's absorption profile extends in velocity space to reach i . Sharp density peaks in real space are therefore smeared in velocity space, and peculiar velocities shift each

absorption feature away from the real space location of its sourcing density peak due to Doppler effect. This is the origin of “redshift-space distortions” in the Ly α forest and is fully captured by treating τ as a discretized convolution rather than a pointwise map.

5.5.2 The Voigt function and the Tepper-Garcia approximation

The normalized Voigt line profile can be written in terms of the dimensionless velocity offset $x = (v - u)/b$ and damping parameter a via the Voigt function $H(a, x)$, defined so that $H(0, x) = e^{-x^2}$ recovers the pure-Doppler Gaussian core. Direct numerical evaluation of $H(a, x)$ is expensive inside a differentiable, GPU-vectorized pipeline, so we use the closed-form approximation of [Tepper-García \(2006\)](#),

$$H(a, x) \approx e^{-x^2} - \frac{a}{\sqrt{\pi} x^2} \left[e^{-2x^2} \left(4x^4 + 7x^2 + 4 + \frac{3}{2x^2} \right) - \frac{3}{2x^2} - 1 \right], \quad (5.10)$$

which reduces smoothly to the Gaussian limit $H(a, x) \rightarrow e^{-x^2}$ as $x \rightarrow 0$ and is accurate to better than 10^{-4} over the range of a and x relevant to the IGM.

5.5.3 Discretized optical depth along a sightline

Combining the physical picture of Section 5.5.1 with the Tepper-Garcia profile, the Ly α optical depth at velocity v_i is the sum, over every pixel j along the sightline, of that pixel’s contribution to the absorption at v_i :

$$\tau(v_i) = \sum_j \frac{c \sigma_{\text{Ly}\alpha} \Delta x_{\text{phys}} \bar{n}_{\text{H}}(z)}{\sqrt{\pi}} \frac{\Delta_{\text{gas},j} x_{\text{HI},j}}{b_j} H\left(a_j, \frac{v_i - u_j}{b_j}\right), \quad (5.11)$$

where $\sigma_{\text{Ly}\alpha} = \sqrt{3\pi\sigma_T/8} \lambda_\alpha f_{\text{osc}}$ is the frequency-integrated Ly α absorption cross-section written in terms of the Thomson cross-section σ_T and oscillator strength $f_{\text{osc}} = 0.4164$, Δx_{phys} is the physical width of one sightline pixel at redshift z , and $\bar{n}_{\text{H}}(z)$ is the mean physical hydrogen number density obtained from the cosmological parameters,

$$\bar{n}_{\text{H}}(z) = \frac{3H_0^2}{8\pi G} \frac{\Omega_b X_h}{(1+z)^3 m_H}, \quad H_0 = 100 h \text{ km s}^{-1} \text{ Mpc}^{-1}. \quad (5.12)$$

Equation 5.11 is precisely the standard Gunn,Peterson-type optical depth expression, evaluated *directly* from the hydrodynamical simulation’s Δ_{gas} , x_{HI} , and T fields rather than from an FGPA-style relating x_{HI} and T to Δ_{gas} alone through a fixed power-law

equation of state. This is the same distinction drawn in Section 5.1 between our forward model and the FGPA-based forward model used in TARDIS (Horowitz et al. 2019): by construction, none of the small-scale baryonic physics (pressure smoothing, shock heating, patchy reionization) needs to be re-derived analytically, since it is already encoded in the simulation fields that Eq. 5.11 consumes directly.

Because the simulation box is periodic, the velocity separation $|v_i - u_j|$ appearing inside H is wrapped onto the interval $[0, v_{\max}/2]$, where

$$v_{\max} = H(z) L_{\text{box,phys}} = H(z) \frac{L_{\text{box}} (1+z)^{-1}}{h} \quad (5.13)$$

is the total Hubble velocity spanned by one period of the box; separations larger than $v_{\max}/2$ are replaced by $v_{\max}$ minus themselves before evaluating Eq. 5.10.

Equation 5.11 is evaluated as a dense $n_{\text{bins}} \times n_{\text{bins}}$ pairwise sum per sightline, the discretized analogue of the convolution integral $\tau(v) = \int dx n_{\text{HI}}(x) \sigma_{\text{Ly}\alpha} \phi(v - u(x))$, implemented as a fully vectorized, GPU-based and automatically differentiable operation, a requirement for its later use as the likelihood term in diffusion posterior sampling (Section 5.6).

5.6 Diffusion Posterior Sampling

We now turn to another one of the chapter’s methodological contribution: how to combine the unconditional prior of Section 5.4 with the differentiable forward model of Section 5.5 to sample from the posterior over the joint field, $p(\mathbf{x}_0 | \mathbf{y})$, given an observed (noisy) Ly α flux spectrum $\mathbf{y}$ along one or more sightlines, without ever training a network that has seen $\mathbf{y}$ or any similar observation, or the FOREST technique.

5.6.1 Specializing the general formulation to Lyman-alpha tomography

The general diffusion posterior sampling recipe was set out in Section 1.3.7 of the Introduction. In brief: at each reverse step the score of the noised posterior decomposes, by Bayes’ theorem, into the pretrained prior score plus a likelihood-guidance term (Eq. 1.32), and the guidance term is made tractable by replacing the intractable $p(\mathbf{x}_0 | \mathbf{z}_t)$ with a point mass at the Tweedie estimate of the clean field (Eqs. 1.34–1.36), so that $p(\mathbf{y} | \mathbf{z}_t) \approx p(\mathbf{y} | \hat{\mathbf{x}}_0(\mathbf{z}_t))$. Here the prior is the unconditional five-field variational

diffusion model of Section 5.4, and the guidance term is the only place the observed Ly α flux $\mathbf{y}$, the sightline geometry, and the forward model of Section 5.5 enter the sampling procedure at all. The remainder of this section builds and explains the two problem specific pieces the general recipe leaves open. That is the explicit likelihood $p(\mathbf{y} \mid \hat{\mathbf{x}}_0)$ for this forward model, and the resulting guidance gradient and sampling algorithm.

5.6.2 The Lyman-alpha likelihood

In this chapter’s variational-diffusion notation, the clean-field estimate entering the likelihood approximation is the identity already introduced in Chapter 2 (Eq. 2.14), and generally as Eq. 1.35,

$$\hat{\mathbf{x}}_0(\mathbf{z}_t) = \frac{\mathbf{z}_t - \sigma_t \hat{\boldsymbol{\epsilon}}_{\theta}(\mathbf{z}_t, t)}{\alpha_t}. \quad (5.14)$$

This is exactly the Tweedie estimate written generally in the Introduction as Eq. 1.35, only in the variational-diffusion notation of Chapter 2. The VDM writes the forward process as $\mathbf{z}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}$, so its signal and noise coefficients (α_t, σ_t) are the continuous time, variance preserving counterparts of the DDPM factors $(\sqrt{\bar{\alpha}_t}, \sqrt{1 - \bar{\alpha}_t})$ used there, $\alpha_t = \sqrt{\bar{\alpha}_t}$ and $\sigma_t = \sqrt{1 - \bar{\alpha}_t}$. The clean field estimate is therefore the same object in both.

Modeling the observational noise on the flux as Gaussian with variance σ_y^2 , set by the survey’s spectral signal to noise ratio, turns the general approximation $p(\mathbf{y} \mid \mathbf{z}_t) \approx p(\mathbf{y} \mid \hat{\mathbf{x}}_0(\mathbf{z}_t))$ (Eq. 1.36) into an explicit, differentiable form,

$$p(\mathbf{y} \mid \hat{\mathbf{x}}_0(\mathbf{z}_t)) \propto \exp\left(-\frac{1}{2\sigma_y^2} \|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{z}_t))\|^2\right), \quad (5.15)$$

where $\mathcal{A}(\cdot)$ denotes the full forward operator of Section 5.5: extracting the four gas-phase channels $(\Delta_{\text{gas}}, x_{\text{HI}}, T, v_{\text{pec}})$ from the joint field estimate (recall that δ_{DM} never enters $\mathcal{A}$ directly; Section 5.5.1), evaluating the discretised optical depth of Eq. 5.11, and converting to transmitted flux via $F = e^{-\tau}$. We note that for this work, although for now we have used a Gaussian Likelihood with a fixed noise scale, this will be improved as this fixed noise prescription affects the reconstruction of highly absorbing regions, which are regions of higher density. More explanation in the discussion.

5.6.3 The guidance gradient and sampling algorithm

Combining Eqs. 1.36, 5.15, the guidance term of Eq. 1.32 becomes

$$\nabla_{\mathbf{z}_t} \log p(\mathbf{y} \mid \mathbf{z}_t) \approx -\zeta_t \nabla_{\mathbf{z}_t} \left\| \mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{z}_t)) \right\|^2, \quad (5.16)$$

where the gradient is taken with respect to $\mathbf{z}_t$ by backpropagating through both the Tweedie estimate $\hat{\mathbf{x}}_0(\mathbf{z}_t)$ (i.e. through the denoising network itself) and the differentiable forward model $\mathcal{A}$ (Section 5.5.3 was built explicitly to make this possible). We absorb $1/\sigma_y^2$ and the normalization of the Tweedie estimate into a single scalar *guidance strength* ζ_t , tuned and scheduled empirically rather than fixed analytically, the Gaussian likelihood derivation above is exact only in the $t \rightarrow 0$ limit, and ζ_t in practice compensates for the approximation error at earlier, noisier steps. For our work, ζ_t follows a schedule given by:

$$\zeta_t = \zeta \frac{1 - \cos(\pi(1 - t))}{2}. \quad (5.17)$$

This ensures that when $t \rightarrow 1$, ζ_t is close to zero, since the estimate of $\hat{\mathbf{x}}_0(\mathbf{z}_t)$ is very noisy and we do not want it to push the guidance in the wrong direction, hence the weak guidance strength. On the other hand, for $t \rightarrow 0$, ζ_t reaches the maximum value of ζ , which for this work is $\zeta = 10$. It was chosen by empirical testing. Too high of a guidance strength pushes the diffusion model out of its learned prior region and causes the outputs to become unphysical (deviate from the prior correlation structure). A weak guidance, on the other hand, does not get guided enough by the observation $\mathbf{y}$, and produces outputs that weakly correlate with it. Each reverse step then combines the ordinary unconditional update of Eq. 2.19 with this additional gradient, pulling the reverse process, at every step, towards states consistent with both the learned prior and the observed flux:

$$\mathbf{z}_{t-1} = \left[\text{unconditional reverse step, Eq. 2.19} \right] - \zeta_t \nabla_{\mathbf{z}_t} \left\| \mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{z}_t)) \right\|^2. \quad (5.18)$$

5.7 Results

To showcase the preliminary results of FOREST, we present the reconstruction of the baryonic and dark matter fields, having observed only a single realization of the Lyman- α

Algorithm 3 FOREST: Diffusion posterior sampling for joint field reconstruction

```

1:  $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T, \dots, 1$  do
3:    $\hat{\mathbf{x}}_0(\mathbf{z}_t) \leftarrow (\mathbf{z}_t - \sigma_t \hat{\epsilon}_\theta(\mathbf{z}_t, t)) / \alpha_t$  ▷ Tweedie estimate, Eq. 5.14
4:    $\mathbf{z}'_{t-1} \leftarrow$  unconditional reverse step from  $\mathbf{z}_t$  ▷ Eq. 2.19
5:    $g \leftarrow \nabla_{\mathbf{z}_t} \|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}}_0(\mathbf{z}_t))\|^2$  ▷ backprop through network and forward model
6:    $\mathbf{z}_{t-1} \leftarrow \mathbf{z}'_{t-1} - \zeta_t g$  ▷ Eq. 5.18
7: end for
8: return  $\mathbf{z}_0$  ▷ one posterior sample of the joint five-field state

```

flux. We test two different scenarios:

- We vary the observed sightline sampling density.
- We vary the signal-to-noise ratio (SNR) of the flux observations.

All these tests are done on an unseen test simulation. The test simulation has the same cosmology and astrophysics as the test simulations but was run with a different initial condition seed. Note that, as mentioned, we do not need to retrain **FOREST** to perform these various tests, since the observed flux never enters the training pipeline. Since these are preliminary results, we only show the outputs for a sub-volume of the full simulation of shape $40 h^{-1} \text{Mpc} \times 10 h^{-1} \text{Mpc} \times 10 h^{-1} \text{Mpc}$. For each combination, we generate 10 posterior samples of shape $(5, 1024, 32, 32)$, matching our training data resolution and spanning a single chunk of the full simulation volume.

Our analysis proceeds in two stages. First, we perform 1D Gaussian smoothing along the line of sight only, isolating the quality of individual sightline reconstructions without transverse mixing. This reveals how well the DPS recovers velocity structure and density perturbations on individual sightlines. Second, we apply full 3D smoothing to mix information across transverse neighbors, examining whether the large-scale 3D structure and correlations are faithfully reconstructed.

Noise model: We add a simple, flux-dependent noise to the synthetic observations. Rather than assuming constant noise, we adopt a spatially-varying noise model where the noise standard deviation at each pixel is inversely proportional to the local flux,

$$\sigma_{\text{noise}}(\mathbf{r}) = \frac{F(\mathbf{r})}{\text{SNR}}, \quad (5.19)$$

with a minimum floor of $\sigma_{\text{min}} = 10^{-4}$ to avoid numerical issues in deep absorption

features. This ensures that absorption troughs and high-transmission regions have comparable effective SNR, rather than a constant noise everywhere.

5.7.1 Metrics

Before presenting the results, we define the summary statistics used throughout this section.

1D power spectrum. For a field f defined on the LoS grid, we compute the 1D power spectrum along the line of sight, averaged over all transverse sightlines within the sub-volume,

$$P_{1D}(k) = \frac{L_{\text{LoS}}}{N_{\text{LoS}}^2} \left\langle |\widetilde{\delta f}(k)|^2 \right\rangle_{\perp}, \quad (5.20)$$

where $\widetilde{\delta f}(k)$ is the discrete Fourier transform of the overdensity field along the LoS axis, $L_{\text{LoS}} = 40 h^{-1} \text{ Mpc}$ is the box length along the line of sight, $N_{\text{LoS}} = 1024$ is the number of LoS pixels, and $\langle \cdot \rangle_{\perp}$ denotes averaging over the transverse (32×32) sightlines. This statistic quantifies how power is distributed across scales along the line of sight, and lets us check whether the reconstruction preserves the correct clustering amplitude at each k . A diffusion trained prior should be able to learn the multiscale power of the fields in principle.

Pearson correlation coefficient. To quantify how linearly related the reconstructed and ground-truth fields are, we use the Pearson correlation coefficient,

$$r_P = \frac{\sum_i (g_i - \bar{g})(l_i - \bar{l})}{\sqrt{\sum_i (g_i - \bar{g})^2} \sqrt{\sum_i (l_i - \bar{l})^2}}, \quad (5.21)$$

where g_i and l_i are the ground-truth and **FOREST** field values at voxel i , and $\bar{g}$, $\bar{l}$ are their respective means. r_P ranges from -1 to 1 , with $r_P = 1$ indicating a perfect linear relationship. Because r_P is sensitive to the absolute scale and outliers of the field, we compute it after applying the appropriate transform for each channel (linear for δ_{DM} and v_{pec} , $\log_{10}$ for Δ_{gas} , x_{HI} , and T , so that the statistic is not dominated by the multiplicative dynamic range of the underlying physical quantities.

Spearman correlation coefficient. As a complementary, rank-based statistic, we also compute the Spearman correlation coefficient,

$$r_S = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}, \quad (5.22)$$

where d_i is the difference between the ranks of g_i and l_i within their respective distributions, and n is the number of voxels compared. Unlike r_P , r_S only requires that the reconstruction preserve the correct *ordering* of field values, and is therefore more robust to monotonic but non-linear biases. For example, if 2 fields fluctuate the same way but their amplitudes are different, Pearson correlation would be smaller than the corresponding Spearman correlation coefficient. Comparing r_P and r_S side by side lets us disentangle whether a drop in correlation is due to genuine structural mismatch or simply an amplitude/normalization offset.

Both coefficients are computed per posterior sample and then aggregated across the 10 samples.

5.7.2 1D LoS reconstruction

For the first test, we study the average 1D flux power spectrum along with the smoothed, reconstructed line of sight, based on a single realization of the observed flux, for varying noise level and sightline density. We also report the Pearson and Spearman correlation coefficients for each field.

Figures 5.2 and 5.3 show the 1D power spectrum (Eq. 5.20) of the reconstructed fields with FOREST, averaged over all sightlines in the sub-volume (1024 in total), together with the residual ratio with respect to the ground truth, for six representative cases: SNR= 50 and SNR= 5, with 100% sightline coverage, 25% coverage, and 6.25% coverage. For example: a 6.25% coverage means a total of 64 observed lines, which for this test are uniformly spread across the transverse plane.

As can be seen, all fields stay very close to the true power spectrum, throughout almost all scales, except the very non linear scales above $20 - 40 h \text{ Mpc}^{-1}$. Since these spectra are calculated without smoothing the fields, this means that FOREST is at least able to reconstruct the 2 point statistics quite well.

Figure 5.4 shows a single, 1D LoS-smoothed sightline (smoothing scale $1 h^{-1} \text{ Mpc}$, along the line of sight only, i.e. without transverse mixing, for the same three cases, illustrating how the individual reconstructed profiles compare against the ground truth as the sightline density and noise level are varied.

Visually, the reconstructed field mean and the 16th to 84th percentile bands covers the true line of sight quantities quite well. However, as the observational support is reduced, for example the middle and right columns, the prediction of high density δ_{DM} peaks gets affected. This can be attributed to the fact that if the observation is

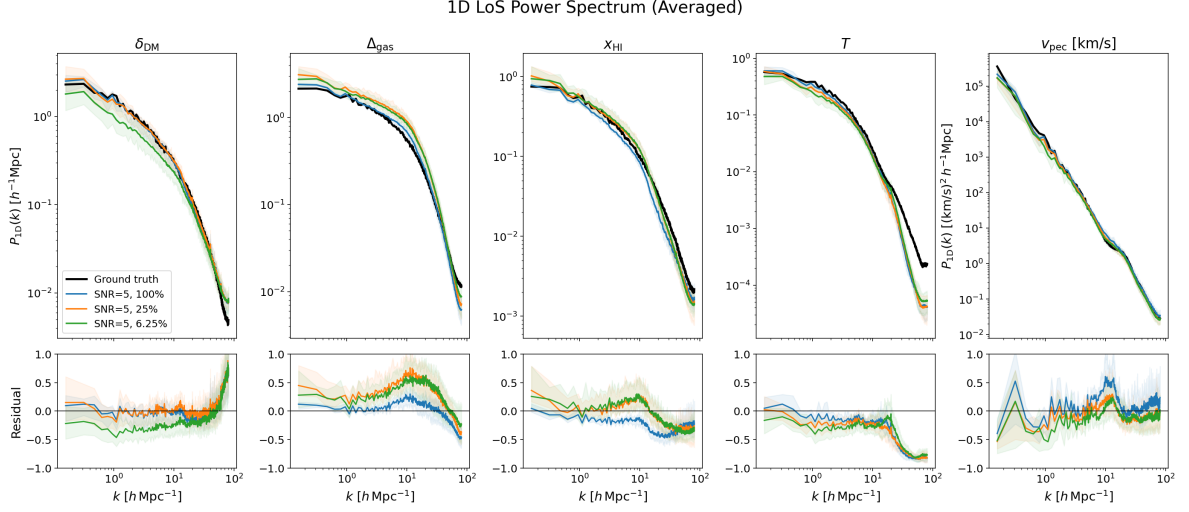


Figure 5.2: 1D line of sight power spectrum (top row) and residual ratio with respect to the ground truth (bottom row), for the five reconstructed fields, averaged over all sightlines in the sub-volume, with a observed SNR= 5, generated with the FOREST technique. Shaded bands show the minimum - maximum range across the 10 posterior samples. Three representative cases are shown: high SNR with full coverage, high SNR with sparse coverage, and low SNR with sparse coverage.

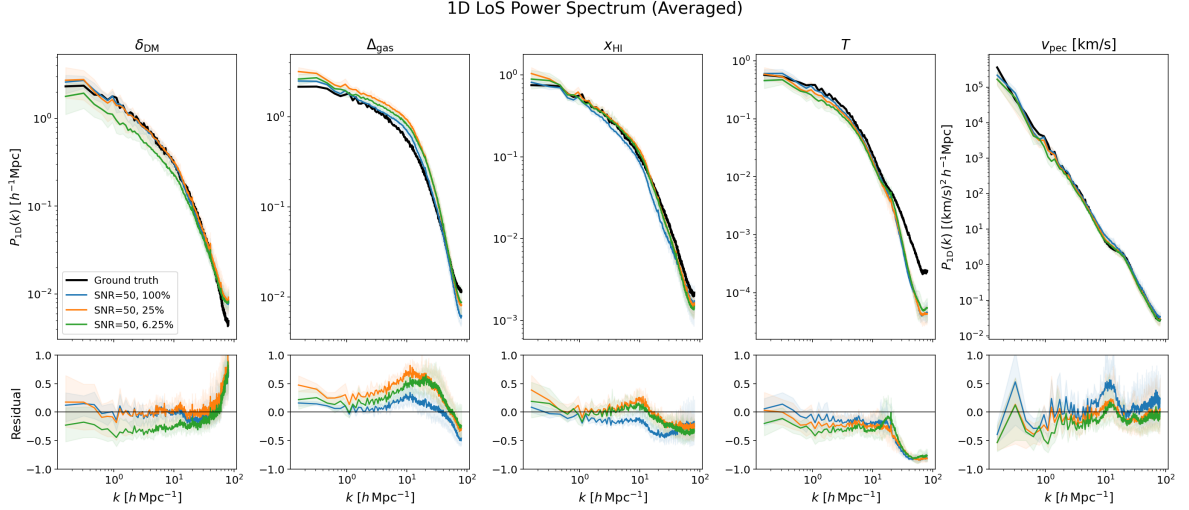


Figure 5.3: 1D line of sight power spectrum (top row) and residual ratio with respect to the ground truth (bottom row), for the five reconstructed fields, averaged over all sightlines in the sub-volume, with a observed SNR= 50, again with the FOREST technique. Shaded bands show the minimum - maximum range across the 10 posterior samples. Three representative cases are shown: high SNR with full coverage, high SNR with sparse coverage, and low SNR with sparse coverage.

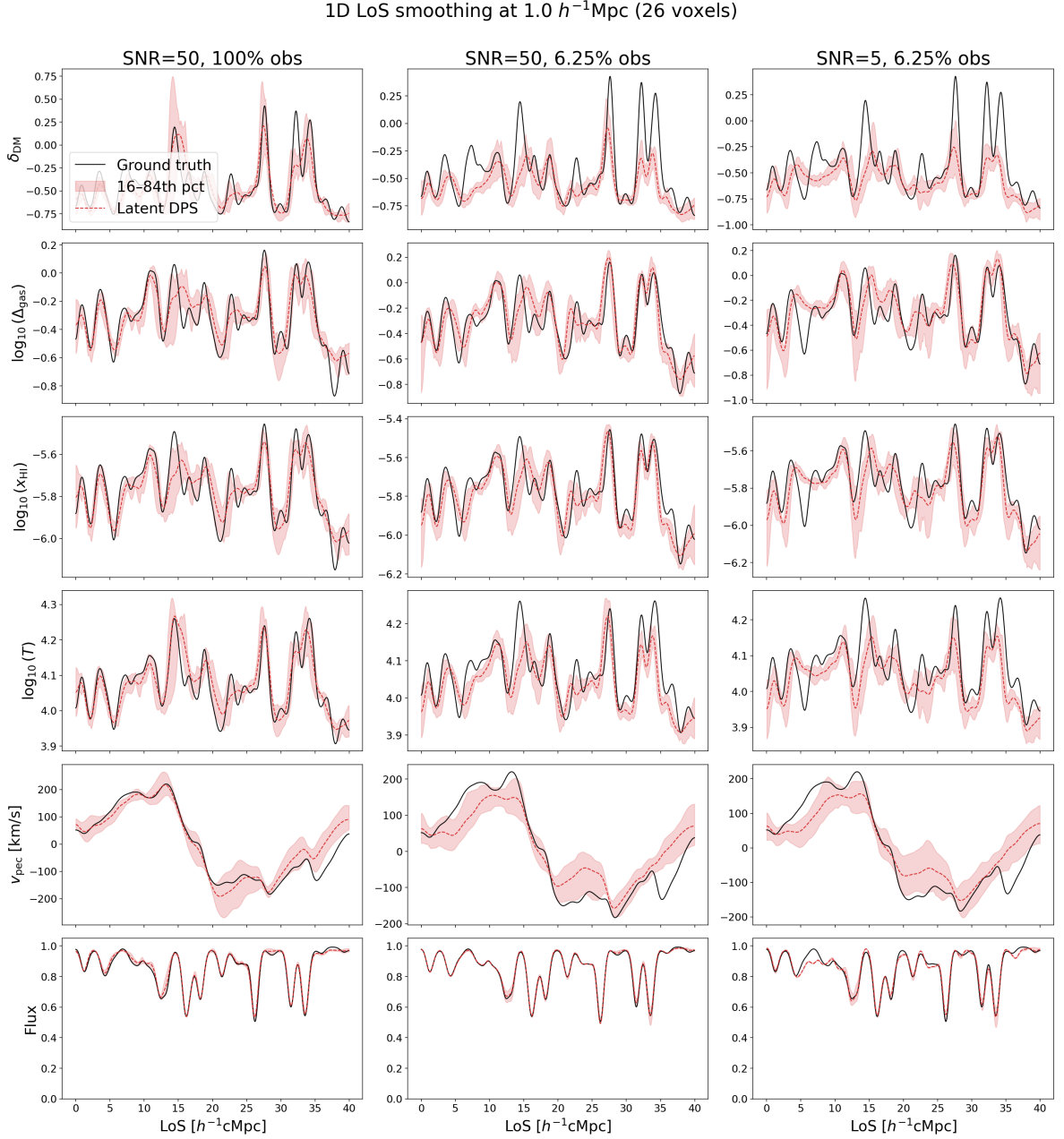


Figure 5.4: Reconstructed sightline profiles after 1D smoothing along the line of sight ($1 h^{-1} \text{Mpc}$ scale), for an observed sightline, for three representative cases out of the 6. The cases are mentioned at the top of the respective columns. Shaded bands show the 16th to 84th percentile range across posterior samples.

sparse, the diffusion model tries to fill those regions with the learned prior instead of a proper reconstructed field, while also being synchronous with the observed flux around it. Another major reason is an inefficient choice of Likelihood. We use a constant noise

level, which corresponds to a very high SNR for low F . These low F correspond to high dense region, and so the model gets confused especially in those dense regions. Note that the flux values here don't seem to reach close to 0. This is purely a smoothing artifact. In the native resolution, they do attain those values indeed.

5.7.3 3D field

Figure 5.5 summarizes the reconstruction quality across all six SNR $\times$ sightline-density combinations, with FOREST, showing the Pearson and Spearman correlation coefficients (Eqs. 5.21, 5.22) evaluated on the 3D fields after $2 h^{-1}$ Mpc smoothing, but this time done on the full 3D field chunk, as a function of the fraction of observed sightlines.

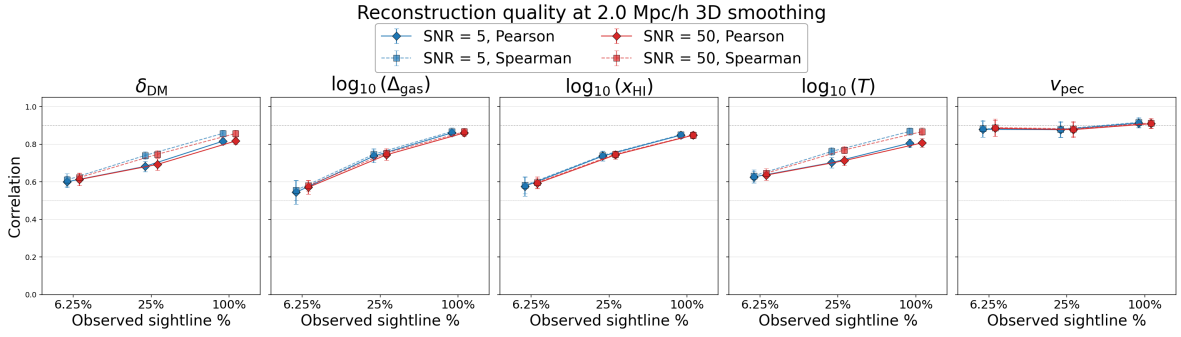


Figure 5.5: Pearson (diamonds, solid lines) and Spearman (squares, dashed lines) correlation coefficients between the reconstructed and ground-truth fields, evaluated after $2 h^{-1}$ Mpc 3D smoothing, as a function of the observed sightline fraction. Blue and red curves correspond to SNR= 5 and SNR= 50 respectively. Error bars show the 1σ spread across the 10 posterior samples.

To visualize the full 3D structure of the reconstruction, Figures 5.6 and 5.7 show a longitudinal slice through the reconstructed sub-volume, for SNR= 50 and SNR= 5 respectively, at each of the three sightline densities tested, both without smoothing and after $2 h^{-1}$ Mpc 3D Gaussian smoothing. The slices are in the line of sight and one of the transverse, and shows 32 lines stacked on top of one another to reconstruct the field in that sub volume.

As can be clearly seen in both the Figures 5.6 and 5.7, and also in the correlation plot (Figure 5.5 for 100% observed sight lines, the reconstruction for both the SNR cases are pretty much identical and close to the ground truth. As the observed sightline densities go down, the reconstruction becomes worse and further away from the truth. This is what is also expected since the guidance strength goes down with the number of

observed lines of sight.

Another important observation is that the reconstruction of the δ_{DM} fields and its correlations are slightly worse as compared to the other fields. This is expected since the forward modeling constraints the other 4 fields directly and so they are strongly regulated by the observed lines. Dark matter, however, is reconstructed purely based on the learnt joint prior. So it is more sensitive to the training uncertainty combined with the noisy reconstruction of the other 4 fields. This is also why very high peaks in δ_{DM} are not reconstructed since in the training data, such high values are quite sparse, and so the model has a hard time reconstructing them.

5.8 Discussion

5.8.1 What the results show

The results of Section 5.7 demonstrate that **FOREST**, a single, unconditionally trained diffusion prior over the joint five-field state can be turned, at inference time alone, into a working posterior sampler for Lyman- α tomography, across a range of sightline densities and noise levels, with no retraining between cases, even working with just one set of observations. The 1D power spectra of Figures 5.2 and 5.3 show that two-point statistics are recovered accurately down to the non-linear scales where the forward model itself becomes most sensitive to the small-scale gas physics, and the correlation coefficients of Figure 5.5 confirm this holds voxel-by-voxel, not only in a statistically averaged sense. As expected, both figures also show where the method is pushed hardest: fidelity degrades smoothly as the observed sightline fraction drops, most visibly in the high-density δ_{DM} peaks of Figure 5.4, where the guidance term has the least direct observational support to draw on and the reverse process falls back increasingly on the learned prior.

It is worth stating again that of the five reconstructed fields, δ_{DM} is the only one that never enters the forward model $\mathcal{A}$ of Section 5.5 at all. Its recovery is therefore mediated entirely by whatever correlations the diffusion prior learned, during unconditional training on the Sherwood-Relics simulations, between the dark matter field and the four gas-phase fields the likelihood actually constrains. That the correlation coefficients for δ_{DM} in Figure 5.5 track those of the directly-constrained fields as closely as they do is not a statement about the forward model at all, it is a direct, quantitative statement about how much of the dark matter, gas coupling learned by the network reflects genuine structure formation physics (gravitational in fall of gas onto the same potential wells that

host the dark matter overdensity) rather than a artifact of training. This the chapter’s most novel result, since it is not achievable by any forward model only or FGPA-based method such as TARDIS (Horowitz et al. 2019, 2021), which has no mechanism for inferring a field that is absent from its own likelihood.

5.8.2 Limitations of the current results

The results presented here are, as stated at the top of the chapter, preliminary in three specific ways that motivate the future work below. First, they cover a single $40 \times 10 \times 10 (h^{-1} \text{ Mpc})^3$ sub-volume, one $1024 \times 32 \times 32$ training-resolution chunk, which is 1/16 of the total volume of the simulation, rather than the full $40 \times 40 \times 40 (h^{-1} \text{ Mpc})^3$ simulation box. Second, the sightline density sweep of Section 5.7 tests robustness to *missing* observations by masking sightlines out of the same dense, fixed 32×32 training grid, not robustness to a genuinely different *native* sightline geometry of the kind a sparser real survey would deliver. Third, both the VAE (Section 5.4.3) and the latent diffusion prior (Section 5.4.3) are trained on the same 5,000 chunks drawn from only 8 independent simulations, differing only in initial condition seed rather than in astrophysics or cosmology. This however, is not a major issue since for testing we do use an unseen simulation, so any biases generated during training would be visible at test time. In particular, none of these limitations is expected to undermine the dark matter recovery result of Section 5.8.1: that result rests on correlations the prior learned between dark matter and the gas-phase fields, which more independent training simulations should reinforce rather than weaken, so the qualitative findings are robust even though the quoted fidelity numbers are not yet final.

Likelihood However, there is a catch. For this work, as explained before, we use a simple Gaussian Likelihood term with a global noise level σ . This seems fine as a first step but has a few obvious problems associated with it. First of all, this is incompatible with the way noise is added at the observation level, which depends on the F values. So regions with less absorption or $F \approx 1$ have a high value of absolute noise and dense regions with $F \approx 0$ have low values. But our global likelihood noise associates same noise to all the F values. This results in low absorption regions to have a high confidence for the model since flux values are higher than σ , but then for small F values, the model is under confident, for the exact opposite reason, hence under or sometimes over predicting those values. This, combined with the fact that there are few high density

peaks in the training, the model suffers from predicting high density regions.

5.8.3 Future work

Generating the full simulation box. The most immediate extension for **FOREST** is to move from the single sub-volume shown here to the full $40 \times 40 \times 40 (h^{-1} \text{ Mpc})^3$ box: 16 tiles of the current sub-volume’s transverse footprint (a factor of 4 in each transverse direction) would be needed to cover it. This is exactly the problem that **Cosmo-FOLD**’s field-overlap latent diffusion (Chapter 3) were built to solve for this thesis’s other diffusion models, and the same overlap and blend strategy is a natural candidate here: generate overlapping transverse tiles independently under DPS guidance, then blend them in the overlap region to enforce large scale coherence across tile boundaries, rather than treating each tile as an independent posterior draw.

Training at native lower sightline densities. The sparsity tests in Section 5.7 mask sightlines out of a fixed, dense 32×32 grid at inference time, they test what happens when observational *coverage* of an otherwise well-resolved grid is incomplete, not what happens when the underlying survey’s native sightline separation is itself coarser than $0.31 h \text{ Mpc}^{-1}$. Real tomographic surveys such as DESI Ly α QSO or WEAVE-QSO (Section 5.1) have substantially sparser mean sightline separations than Subaru-PFS/CLAMATO-class surveys, and it is not guaranteed that a prior and forward model built around the current training grid’s resolution generalize to that regime by masking alone. The VAE’s anisotropic downsampling (Section 5.4.2) and the diffusion prior’s own learned correlations were both shaped by what the training chunks actually look like at 32×32 transverse resolution. Training a variant of the prior directly on chunks built from a coarser native transverse grid, and comparing its reconstruction fidelity against the current model evaluated on the same coarse case via masking alone, would directly test whether retraining at the target survey’s actual sightline density is necessary, or whether the current no-retraining claim (Section 5.1) genuinely holds across geometries and not only across coverage fractions of one fixed geometry.

Training with more data. The current prior is trained on 5,000 overlapping chunks drawn from only 8 independent simulations sharing one astrophysical and cosmological model. The overlapping-window chunking scheme (Section 5.4.3) means many of these 5,000 samples are not independent draws, so the effective diversity seen by the network

during training is smaller than the raw count suggests. This can very much lead to overfitting. An extension follows directly: more independent random-seed realizations at fixed cosmology and astrophysics, to test whether current fidelity is limited by training-set diversity rather than model capacity.

Optimizing the latent space. The $20\times$ compression factor used throughout this chapter (Section 5.4.2) was chosen qualitatively, as a "sweet spot" between over-compressing away small-scale information and under-compressing enough to make no computational difference, rather than from a systematic ablation. An important next step is to quantify the fidelity, compute trade-off directly. Train VAEs at a range of compression factors and anisotropic downsampling schedules, and evaluate each against the same reconstruction metrics already used in Section 5.7.1 (power spectrum residuals, Pearson/Spearman correlation), rather than relying on the qualitative reasoning of Section 5.4.2 alone. This would also clarify whether a single, fixed compression factor is even the right choice, or whether a non-uniform/adaptive latent representation, allocating more of the latent budget to the fields or scales that the guidance term relies on most directly (Δ_{gas} , x_{HI} given their role in the optical-depth forward model of Section 5.5), could improve reconstruction fidelity at the same overall compression.

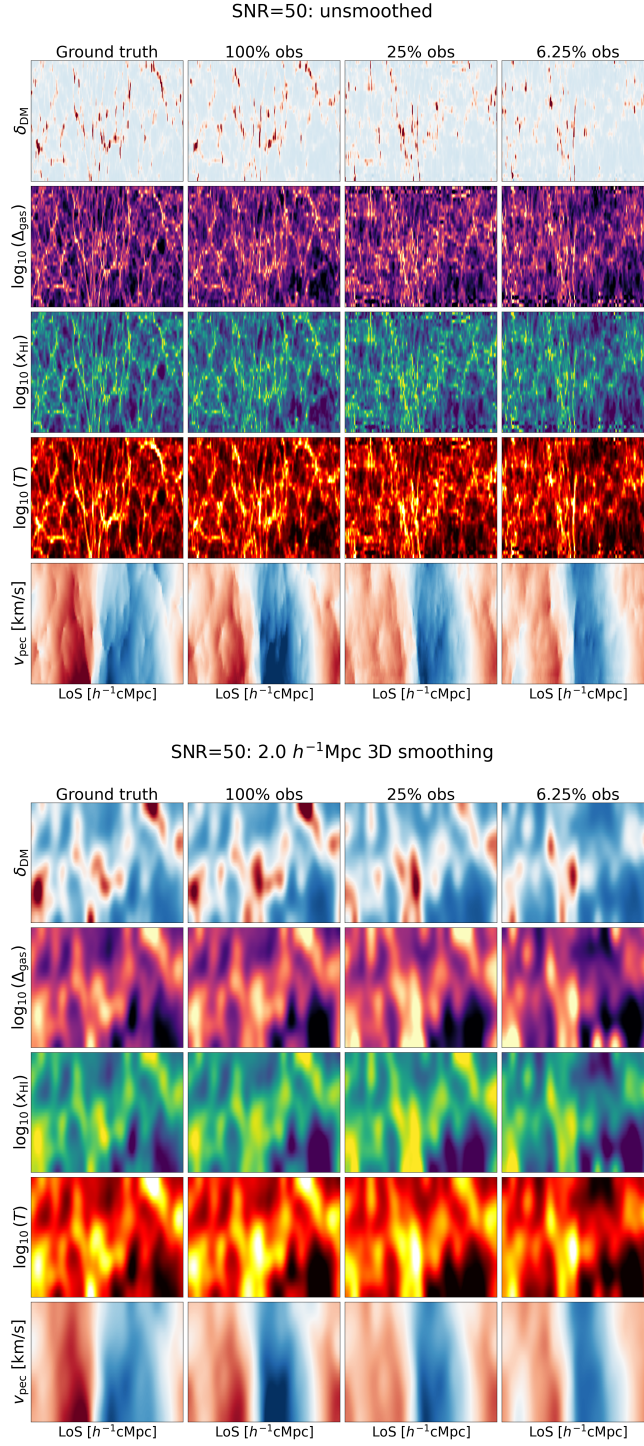


Figure 5.6: Longitudinal slices through the reconstructed sub-volume for SNR= 50, comparing the ground truth (leftmost column) against our FOREST method at 100%, 25%, and 6.25% sightline coverage. Top: unsmoothed fields. Bottom: after $2 h^{-1}$ Mpc 3D smoothing.

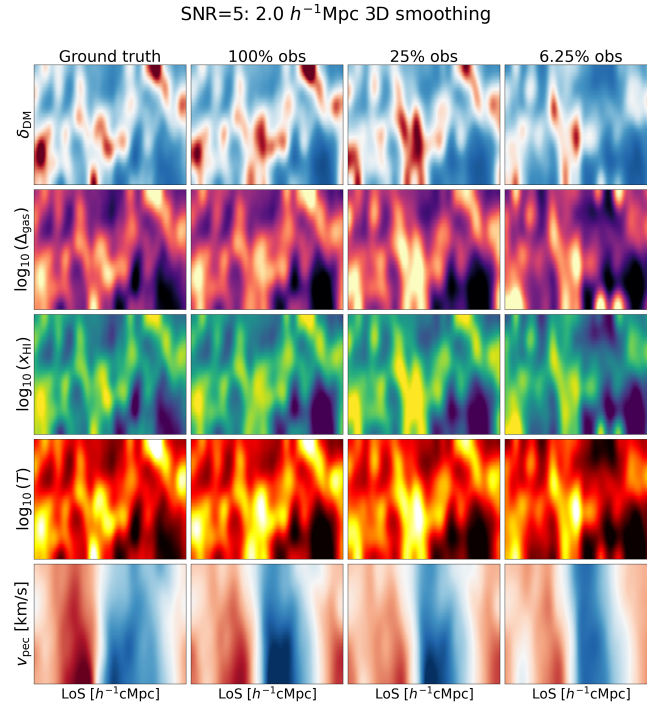
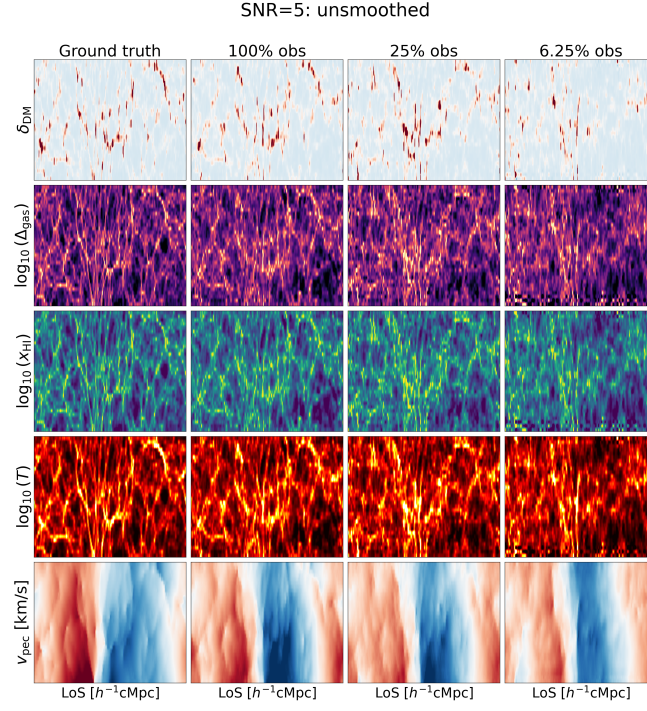


Figure 5.7: Same as Figure 5.6, but for SNR= 5.

Chapter 6

Field-Level Emulation for Multi-Tracer Simulation-Based Inference

6.1 Introduction

Upcoming large scale structure (LSS) surveys will map galaxy positions and neutral hydrogen (HI) intensity over volumes large enough that statistical noise ceases to be the limiting factor in cosmological parameter inference; instead, the accuracy of the forward model connecting cosmological and astrophysical parameters to the observed tracer fields becomes the bottleneck. That mapping is governed by non-linear gravitational collapse and by baryonic feedback processes (supernova and AGN feedback in particular) that can only be captured through full hydrodynamical simulations (Villaescusa-Navarro et al. 2022). As discussed in Section 1.1.7, traditional power-spectrum-based inference is tractable but information lossy on exactly these non linear, feedback sensitive scales.

Simulation-based inference (SBI) removes the need for an explicit likelihood by training a neural density estimator directly on simulation parameter pairs, and has become an increasingly common tool in cosmology precisely because forward simulations, not tractable likelihoods, are what is naturally available. Its main practical limitation is that it is simulation-hungry: field-level SBI in particular requires simulation budgets that are difficult to reach with full hydrodynamical suites such as CAMELS (Villaescusa-Navarro et al. 2022), which sample the cosmological and astrophysical parameter space with only $\mathcal{O}(10^3)$ runs. *Emulators* address this bottleneck by learning the mapping

from cheap, approximate dark matter-only simulations (or directly from parameters) to the corresponding baryonic observables, replacing the expensive hydrodynamical step with a network evaluation that costs milliseconds rather than CPU-hours. This is the role the emulator plays in the present chapter: it is not itself the science goal, but the component that makes a full-field, multi-tracer SBI analysis computationally feasible at all.

It is worth situating this approach relative to the rest of the thesis. Chapter 5’s **FOREST** develops diffusion-based posterior sampling as a route to field-level inference, drawing samples from an implicit posterior over the dark matter field itself. The work presented here follows a complementary strategy: rather than sampling the DM field posterior directly, we use a deterministic, conditional emulator to generate tracer fields from DM inputs, and perform inference with normalizing-flow-based neural posterior estimation (NPE) on either summary statistics or a learned latent representation of those fields. Both approaches share the same underlying motivation, extracting the cosmological information lost when a field is compressed to a power spectrum, but they trade off differently between expressiveness, computational cost, and the kind of posterior they target (a full field posterior in one case, a marginal posterior on a handful of cosmological parameters in the other).

The proof of concept study summarized in this chapter combines three ingredients: (i) fast, approximate dark matter simulations from **FastPM**; (ii) the **HGL** emulator, which paints galaxy number-count and HI brightness-temperature fields onto those DM maps; and (iii) an NPE pipeline that infers Ω_m and σ_8 either from the two tracers’ power spectra or from a full 3D field-level representation, single-tracer or jointly. The rest of the chapter is organized as follows. Section 6.2 gives a high-level overview of the pipeline. Section 6.3 describes the DM simulations and the target galaxy/HI fields used to train the emulator. Sections 6.4 and 6.5 describe the **HGL** architecture and its training in detail. This is the core contribution of the author to this work. Section 6.6 summarizes, briefly, how the emulated fields are subsequently used for SBI. Section 6.7 reports results, first on the fidelity of the emulator itself (Section 6.7.1), then on the downstream cosmological constraints it enables (Section 6.7.2). Section 6.8 discusses limitations and future directions.

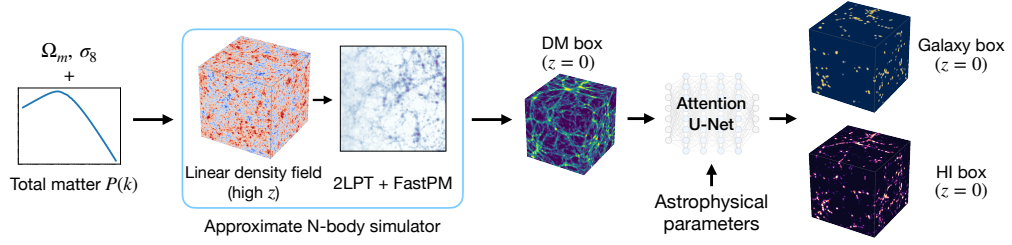


Figure 6.1: High-level overview of the pipeline described in this chapter. Approximate dark matter maps generated with **FastPM** are passed through **HGL**, the Attention U-Net emulator, to produce galaxy and HI fields conditioned on cosmological and astrophysical parameters. These emulated fields are then either compressed to a power spectrum or fed to a convolutional encoder, and in both cases used to train a normalizing-flow-based NPE model that infers the posterior on Ω_m and σ_8 .

6.2 Pipeline overview

Figure 6.1 sketches the full pipeline. Starting from a cosmology $\theta_C = \{\Omega_m, \sigma_8\}$, we generate an approximate DM density field at $z = 0$ with **FastPM**. The **HGL** emulator maps this DM field, together with a choice of astrophysical parameters θ_A , onto galaxy number-density and HI brightness-temperature fields (Sections 6.3 and 6.5). These emulated fields are the only place where the expensive, hydrodynamics-dependent physics enters the pipeline; everything downstream of them, the power spectrum extraction, the convolutional compression, and the normalizing flow, only ever sees emulator output (Section 6.6). All results in this chapter refer to observables at $z = 0$ in the plane-parallel approximation.

6.3 Dark matter maps and target observables

Approximate dark matter simulations. DM-only maps are generated with **FastPM** (Feng et al. 2016), a particle-mesh code that is several orders of magnitude cheaper than a full N-body run while still capturing the non-linear growth of structure relevant on the scales of interest here. We generate cubes at two box sizes, $L = 25 h^{-1} \text{ Mpc}$ (128^3 particles and mesh, saved at 64^3 resolution) and $L = 50 h^{-1} \text{ Mpc}$ (256^3 particles and mesh, saved at 128^3 resolution), so that both box sizes share the same voxel resolution. Initial conditions are set at $z = 9$ with 2LPT and evolved to $z = 0$ with 10 **FastPM** steps, at a cost of ~ 0.3 CPU-hours per realization. For the emulator training set, we reuse the same random seeds and linear power spectrum provided by the **CAMELS**

Latin Hypercube (LH) suite, since **FastPM**’s random number generator is compatible with the **N-GenIC** generator used by **CAMELS**; this lets us pair each **FastPM** DM cube with the corresponding hydrodynamical galaxy/HI target on a like-for-like basis. When emulating new maps outside the training set (i.e. for SBI), the linear power spectrum is instead obtained from the **CLASS** Boltzmann solver for arbitrary seeds and parameter draws.

CAMELS hydrodynamical targets. The emulator’s target fields are drawn from the IllustrisTNG LH suite of **CAMELS** (Nelson et al. 2021), which varies two cosmological parameters (Ω_m, σ_8) together with four astrophysical feedback-strength parameters ($A_{\text{SN1}}, A_{\text{AGN1}}, A_{\text{SN2}}, A_{\text{AGN2}}$) governing supernova and AGN feedback, over the same prior ranges already introduced in Table 4.1. Two tracers are extracted from each snapshot:

- **Galaxy number counts.** Sub-halos above $10^9 M_\odot/h$ in mass with non-zero stellar mass are painted onto a 64^3 grid with a cloud-in-cell (CIC) scheme, smoothed with a Gaussian kernel of radius $R = 0.2 h^{-1} \text{ Mpc}$, and low values are clipped to 10^{-4} to stabilize training.
- **HI brightness temperature.** For each gas particle i we combine the neutral hydrogen fraction x_{HI} and gas mass to obtain the HI density,

$$\rho_{\text{HI},i} = m_{g,i} X x_{\text{HI},i} \left(\frac{N_{\text{part}}}{aL} \right)^3 \left[\frac{\text{kg}}{\text{m}^3} \right], \quad (6.1)$$

with $X = 0.76$ the hydrogen mass fraction and a the scale factor, following the self shielding prescription of Villaescusa-Navarro et al. (2018a). After interpolating onto a 64^3 grid with CIC, this is converted to the post-reionization brightness temperature,

$$T_{\text{b}}(\mathbf{x}) = 189 h \left(\frac{H_0(1+z)^2}{H(z)} \right) \frac{\rho_{\text{HI}}(\mathbf{x})}{\rho_c} [\text{mK}], \quad (6.2)$$

and smoothed with the same $R = 0.2 h^{-1} \text{ Mpc}$ kernel as the galaxy field.

Why two box sizes. The $25 h^{-1} \text{ Mpc}$ box is what the **CAMELS** LH suite provides, and is therefore the only size at which paired DM/hydro training data exists. However, at this box size the power spectrum is noticeably sensitive to cosmic variance: realizations at fixed physical parameters but different random seeds can yield visibly different power spectra, which introduces a degeneracy between cosmology and realization noise. We

therefore adopt $50 h^{-1}$ Mpc boxes for the SBI stage to improve large scale mode sampling, while still training the emulator itself at $25 h^{-1}$ Mpc, since the LH suite is not available at the larger box size (the available alternative, the SB35 suite, varies 30 astrophysical parameters simultaneously instead of 4, making it unsuitable for this emulator). This choice, using an emulator trained at one box size to generate maps at another, is only valid because the two box sizes share the same voxel resolution. This of course is also the main limitation of the current pipeline, which is also discussed in Section 6.7.1.

6.4 The HGL emulator: architecture

The task of predicting a 3D galaxy or HI field from its DM counterpart is, structurally, an image-to-image translation problem, and we address it with a conditional 3D U-Net (Ronneberger et al. 2015), which we refer to as HGL (HALOgen-like). This builds directly on the HALOgen architecture introduced in the author’s earlier work (Mishra et al. 2025), but with a structural change that is the central methodological contribution described here: HALOgen was designed to emulate fields at a single, fixed cosmology, whereas HGL must generalize over both cosmological parameters $\theta_C = \{\Omega_m, \sigma_8\}$ and astrophysical parameters $\theta_A = \{A_{\text{SN1}}, A_{\text{AGN1}}, A_{\text{SN2}}, A_{\text{AGN2}}\}$ at inference time, rather than requiring a separately trained network per parameter point.

Backbone. The network follows a standard encoder - decoder U-Net layout with three downsampling and three upsampling stages, each with a ResNet block and either a down or upsampling layer (Table 6.1 and Figure 6.3). Each stage is a modified residual block (Fig 6.2) that consists of two 3D convolutions, each preceded by Group Normalization and a non linear activation. Downsampling is performed with max pooling and upsampling with trilinear interpolation. Skip connections concatenate encoder features into the corresponding decoder stage at matching resolution, which is what allows the network to recover high-frequency structure that would otherwise be lost during downsampling. At the bottleneck, a self attention layer lets the network model long-range, non-local correlations between voxels, which is particularly relevant for capturing large scale structure that a purely convolutional receptive field would only see indirectly through depth.

Conditioning via FiLM. The mechanism that lets a single network generalize across parameters is Feature-wise Linear Modulation (FiLM) (Perez et al. 2017). For a feature

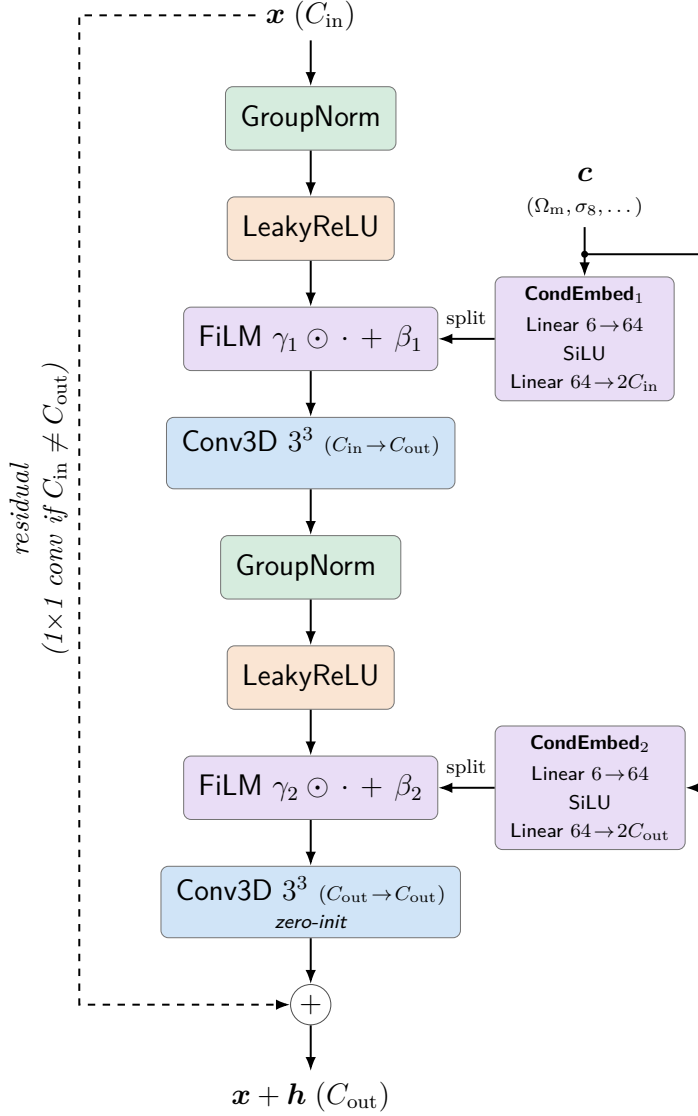


Figure 6.2: Residual block of the emulator U-Net with FiLM conditioning. Each of the two convolutional stages is preceded by GroupNorm and a Leaky ReLU activation. The cosmology and astrophysics parameter vector $\mathbf{c}$ is mapped by two *independent* two-layer MLPs to per-channel modulation coefficients (γ_i, β_i) , applied as $\gamma_i \odot \mathbf{h} + \beta_i$ immediately before each convolution. The second convolution is zero-initialised so that the block starts as the identity map. The skip path uses a $1 \times 1 \times 1$ convolution when the channel count changes.

layer $\mathbf{x} \in \mathbb{R}^{C \times D^3}$ inside the modified residual block, and a conditioning vector $\mathbf{c} \in \mathbb{R}^{d_c}$, FiLM applies a channel-wise affine transformation,

$$\text{FiLM}(\mathbf{x}, \mathbf{c}) = \gamma(\mathbf{c}) \odot \mathbf{x} + \beta(\mathbf{c}), \quad (6.3)$$

Table 6.1: HGL emulator architecture. L denotes the input map’s linear size in voxels. Every residual block is $x \rightarrow \text{GN} \rightarrow \text{FiLM} \rightarrow \text{Conv3D} \rightarrow \text{GN} \rightarrow \text{FiLM} \rightarrow \text{Conv3D} + x$.

Stage	Spatial dims	Channels (in→out)	Operation
Input	L^3	1	input field
Encoder	$(L/2)^3$	$1 \rightarrow 16$	ResNet block + MaxPool
	$(L/4)^3$	$16 \rightarrow 32$	ResNet block + MaxPool
	$(L/8)^3$	$32 \rightarrow 64$	ResNet block + MaxPool
	$(L/8)^3$	$64 \rightarrow 128$	ResNet block
Bottleneck	$(L/8)^3$	$128 \rightarrow 128$	self attention
Decoder	$(L/4)^3$	$128 \rightarrow 64$	upsample + ResNet block
	$(L/2)^3$	$64 \rightarrow 32$	upsample + ResNet block
	L^3	$32 \rightarrow 16$	upsample + ResNet block
Output	L^3	$16 \rightarrow 1$	ResNet block

where $\gamma, \beta \in \mathbb{R}^C$ are produced from $\mathbf{c}$ by a shared multi-layer perceptron, $[\gamma(\mathbf{c}), \beta(\mathbf{c})] = f_\theta(\mathbf{c})$, and $\odot$ is an element-wise product across channels. Two such transformations are applied per residual block (one per convolution), so the number of (γ, β) pairs at a given stage scales with that stage’s channel count (See figure 6.2). $\mathbf{c}$ here, is the concatenation of cosmological and astrophysical parameters, $\mathbf{c} = [\theta_C, \theta_A] \in \mathbb{R}^6$, so the same physical parameters that define a CAMELS simulation are what condition every feature layer of the network. Replacing the fixed affine parameters of a standard normalization layer with FiLM’s conditioning-dependent ones is what turns a single cosmology emulator such as the original HALOgen into one that can be queried at arbitrary points in parameter space without retraining.

What the emulator can and cannot vary. Given a DM map $\mathcal{F}_i^{\text{DM}}(\Omega_{m_i}, \sigma_{8_i})$ at cosmology i , the network produces the corresponding galaxy or HI field at that same cosmology, for any astrophysics j of choice,

$$\mathcal{F}_i^{\text{DM}}(\Omega_{m_i}, \sigma_{8_i}) \xrightarrow{\text{U-Net}} \mathcal{F}_i^{\text{gal/HI}}(\Omega_{m_i}, \sigma_{8_i}, A_{\text{SN1}_j}, A_{\text{AGN1}_j}, A_{\text{SN2}_j}, A_{\text{AGN2}_j}). \quad (6.4)$$

Astrophysics can be varied freely at inference time because it is injected purely through the FiLM conditioning; cosmology cannot, because it is already baked into the input DM field itself through the linear power spectrum used to generate it. In other words, the emulator paints astrophysics onto a fixed gravitational skeleton, it cannot repaint a DM field generated at one cosmology as if it belonged to another. This is a structural

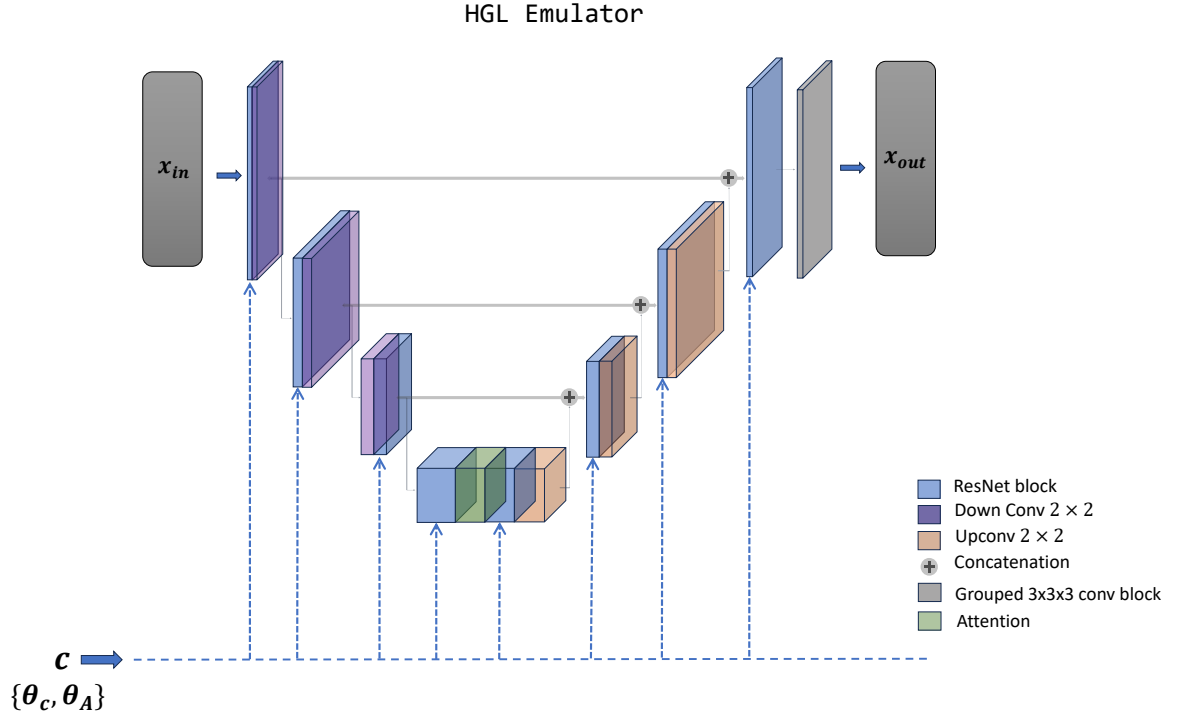


Figure 6.3: The modified U-Net for the HGL emulator. The Attention U-Net based architecture now has a modified ResNet block (Fig. 6.2), that takes as input the CAMELS parameters $\mathbf{c}$. This network has 4 residual blocks each in encode and decode layers and 3 down and upsampling layers.

constraint of the approach, not a training limitation, and it is what makes the DM-map generation step (Section 6.3) integral to the emulator rather than a separate pre-processing step.

6.5 Emulator training

Data splits. The 508 CAMELS LH seeds spanning the full cosmological and astrophysical prior range are split 75%/25% into training and validation sets, with an independent test set of 242 seeds held out entirely. To increase the effective size of the training set without introducing new simulations, each training/validation cube is augmented with all 24 possible axis-aligned rotations, giving $508 \times 24 = 12,192$ cube instances; the test

set is left unrotated at 242 cubes. Every map (input DM and target galaxy/HI) is standard-scaled using statistics fit on its own training split, and each network is trained on a single tracer channel at a time (i.e. separate networks for galaxies and for HI).

Loss function. Following the quantile-regression approach used in the original HALOgen model, the emulator is trained with a sum of three quantile losses,

$$\mathcal{L} = \sum_{i=1}^3 \text{QL}(y, \hat{y}; \tau_i), \quad \text{QL}(y, \hat{y}; \tau) = \frac{1}{N} \sum_{j=1}^N \max(\tau(y_j - \hat{y}_j), (\tau - 1)(y_j - \hat{y}_j)), \quad (6.5)$$

where y and $\hat{y}$ are the target and predicted voxel values. Rather than regressing a single point estimate, this trains the network to reproduce several quantiles of the conditional target distribution at once, which improves robustness on the sparse, highly non-Gaussian galaxy field in particular. We use $\tau \in \{0.45, 0.60, 0.75\}$ for HI and $\tau \in \{0.45, 0.60, 0.80\}$ for galaxies, chosen empirically. During training we checkpoint the model every time the loss improves, and select, a posteriori, both the checkpoint with the lowest validation loss (best at the pixel level) and the checkpoint with the smallest power-spectrum residual on held-out validation maps. For the HI emulator these two criteria pick the same checkpoint; for the galaxy emulator they do not, which is itself informative, it means the epoch at which the network best reproduces individual voxel values is not necessarily the epoch at which it best reproduces the two point statistics of the field, and the two should be tracked and reported separately rather than assuming pixel level accuracy implies summary-statistic accuracy (or vice versa). Once trained, generating a single 64^3 map takes $\sim 10^{-2}$ s on an NVIDIA A100 GPU, several orders of magnitude faster than the hydrodynamical simulation it replaces.

6.6 Downstream use: simulation-based inference

The remainder of the pipeline, developed jointly with the co-authors, is summarized here only to the extent needed to interpret the results in Section 6.7.2. Emulated galaxy and HI maps are used to train a neural posterior estimator (NPE) that approximates $p(\theta_{\mathcal{C}} \mid \mathbf{d})$ directly from simulation parameter pairs, modeling the posterior with a Masked Autoregressive Flow (MAF) (Papamakarios et al. 2018). Two data representations are compared: SBI on the power spectrum $P(k)$ of the emulated fields (case $\mathcal{P}$, a summary-statistics baseline), and SBI on a latent representation of the full 3D (or, for

comparison, axis-averaged 2D) field extracted by a convolutional encoder trained jointly with the flow (case $\mathcal{M}$). Both cases are run for each tracer individually and for the two tracers combined, either fixing the astrophysical parameters to their fiducial values or marginalizing over them. The purpose of this comparison, discussed further below, is to isolate how much cosmological information is retained (or lost) at each stage of compression, from a full field, down to a latent vector, down to a power spectrum, given input fields whose fidelity is itself set by the HGL emulator described above.

6.7 Results

6.7.1 Emulator validation

Figure 6.4 shows emulator performance at $25 h^{-1}$ Mpc, the box size it was trained on, for the galaxy tracers: power-spectrum residuals stay below $\sim 25\%$ across the scales probed, with well-centered averages, and the voxel value histograms of emulated and true maps agree closely.

Figure 6.5 shows the same test for the HI tracers, with equally good agreement. This is the regime the emulator was designed and validated for, and it performs as intended, on average, for both tracers.

Figure 6.6 shows the same test applied to $50 h^{-1}$ Mpc maps for the galaxy tracers, generated by the same $25 h^{-1}$ Mpc-trained network (the box-size argument in Section 6.3). The pixel value histograms remain well calibrated, but the power spectrum is visibly less accurately reproduced than at $25 h^{-1}$ Mpc.

Figure 6.7 shows the corresponding degradation for the HI tracers, similar to the galaxy case. The most likely explanation for both these $50 h^{-1}$ Mpc cubes is that larger boxes contain larger scale modes, and correspondingly more massive halos, than the network ever saw during training on smaller boxes; the emulator is being asked to extrapolate into a regime it was never exposed to, not merely to interpolate within its training distribution. We adopt this configuration for the SBI stage regardless, since it is internally consistent (both the SBI training and test sets are generated by the same emulator, so the network’s own biases are shared between them and do not, by construction, bias the SBI comparison), but this is the clearest limitation of the current emulator, and one that would need to be resolved, most directly by training on paired DM/hydro data at $50 h^{-1}$ Mpc once such a suite becomes available, before applying this pipeline to real survey data or to maps generated by a different DM-only simulator (see

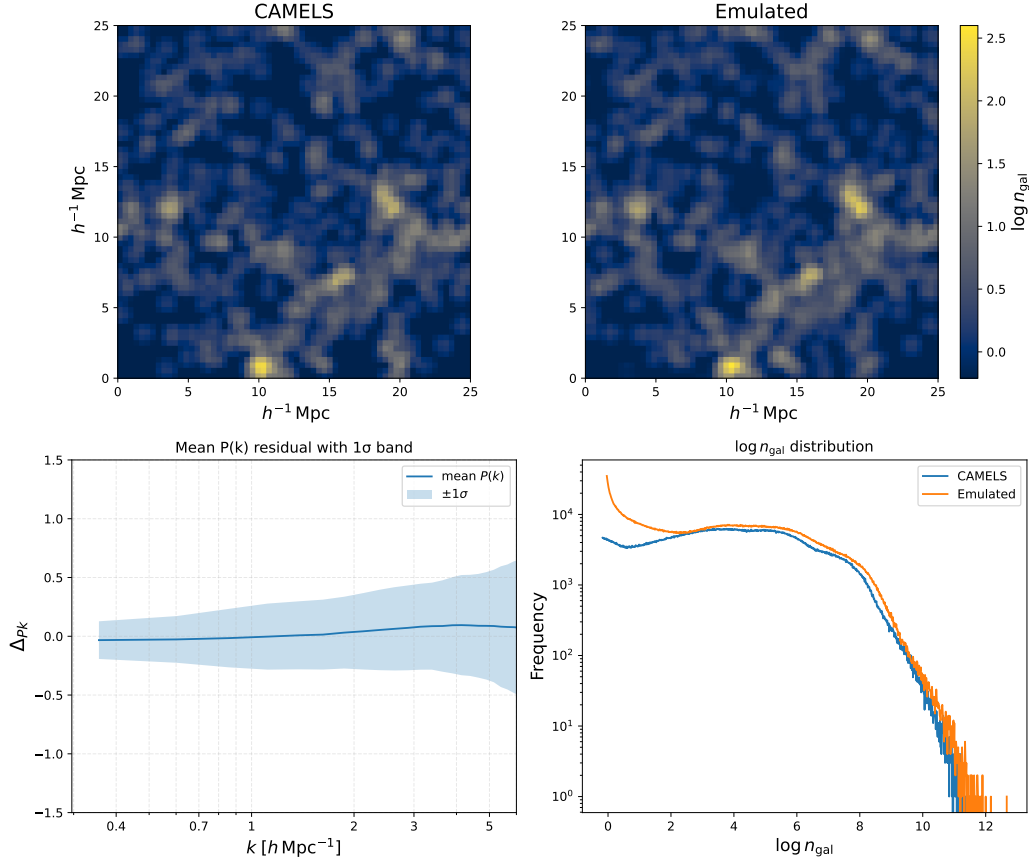


Figure 6.4: HGL emulator performance at $25 h^{-1} \text{ Mpc}$ on the held-out CAMELS LH test set (242 maps), for the galaxy tracers. A sample from CAMELS (left) vs. emulated map (right). Bottom left: mean and 1σ shaded band for the residual power spectrum $\Delta_{P(k)} = (P(k)^{\text{TRUE}} - P(k)^{\text{EMULATED}})/P(k)^{\text{TRUE}}$ for all the 242 test simulations. Bottom right: Histogram comparison for the same 242 simulations, true vs emulated.

Section 6.8).

6.7.2 Impact on cosmological constraints

The headline result of the full study, reported in Table 6.2, shows the decrease of the posterior volume when presented with full field data rather than just the summary statistics.

This is also visually evident in Figure 6.8, 6.9 and 6.10.

Moving from the power spectrum to a full 3D field-level representation increases the median FoM, defined as $\text{FoM} = 1/A_{68}$, where A_{68} is the area within the 68% joint credible region for σ_8 and Ω_m (marginalized over astrophysical parameters where

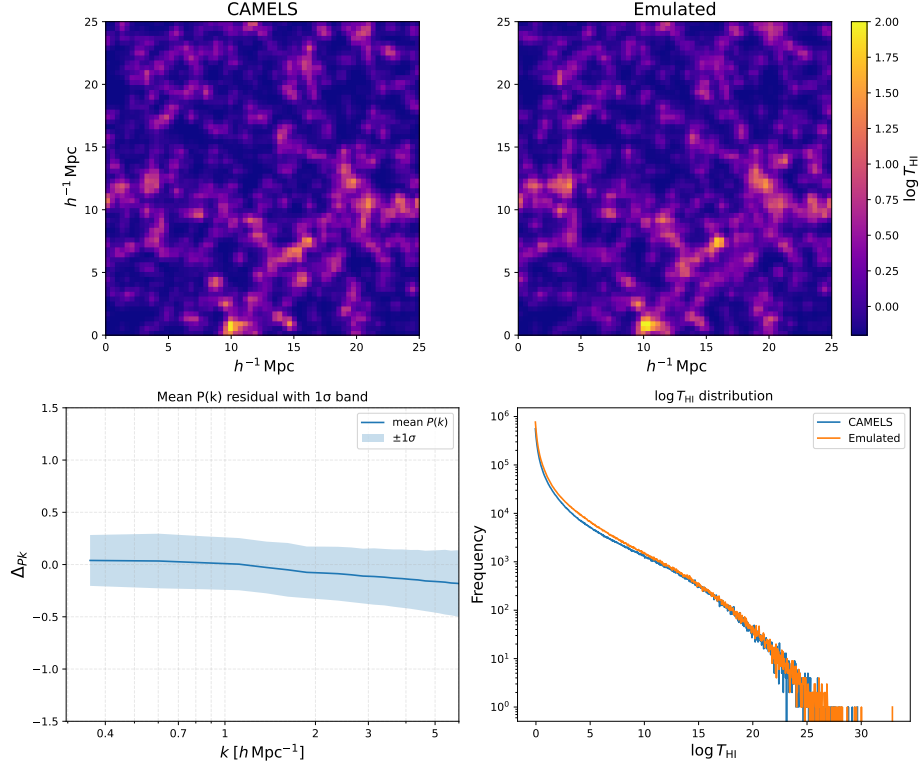


Figure 6.5: HGL emulator performance at $25 h^{-1} \text{ Mpc}$ on the held-out CAMELS LH test set (242 maps), for the HI tracers, very similar to Fig. 6.4. A sample from CAMELS (left) vs. emulated map (right). Bottom left: mean and 1σ shaded band for the residual power spectrum $\Delta_{P(k)} = (P(k)^{\text{TRUE}} - P(k)^{\text{EMULATED}})/P(k)^{\text{TRUE}}$ for all the 242 test simulations. Bottom right: Histogram comparison for the same 242 simulations, true vs emulated.

appropriate), by a factor of ~ 3 , and combining the two tracers rather than using either alone increases it further by a factor of $\sim 2\text{--}7$ depending on the case, as shown in Figures 6.9 and 6.10.

A flattened 2D field-level representation, by contrast, gives little to no improvement over the power spectrum. Table 6.3 shows that bias remains below $\sim 1.5\%$ in all cases, well within statistical uncertainties, and calibration (assessed with PP-plots in the full pipeline results) is generally good, with some degradation for the highest-information 3DM configuration when marginalizing over astrophysics.

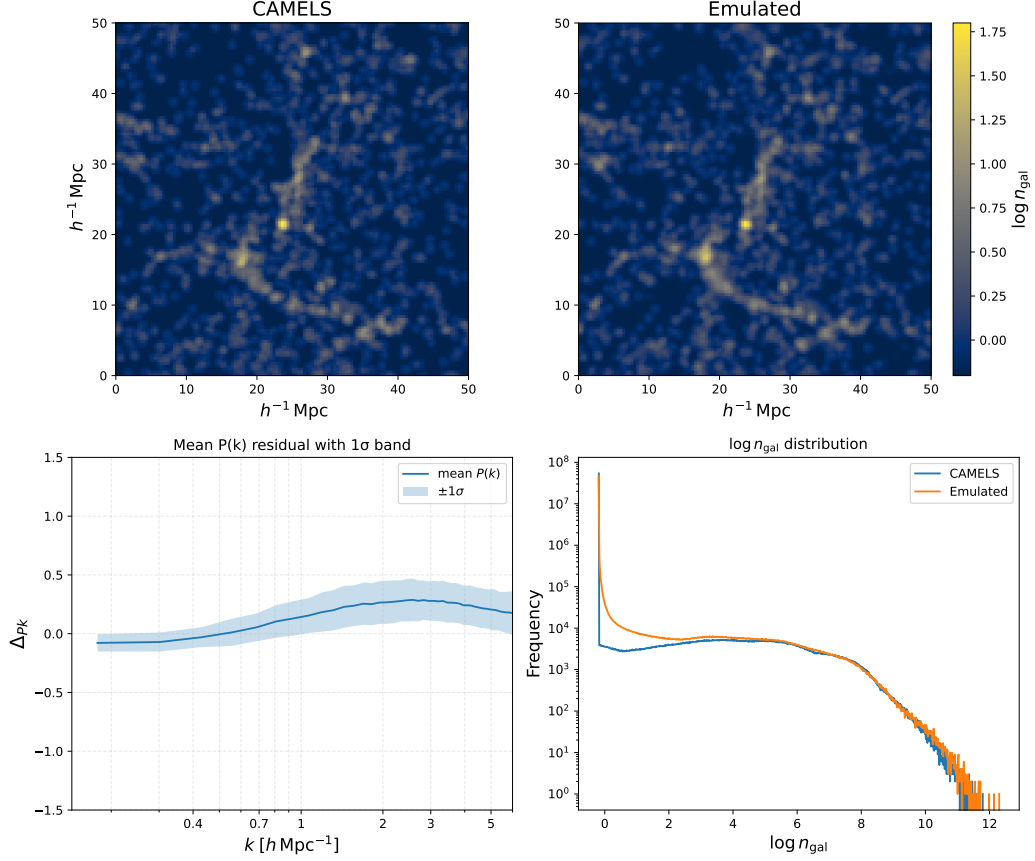


Figure 6.6: As Figure 6.4, but for 50 h^{-1} Mpc maps generated with the emulator trained on 25 h^{-1} Mpc cubes (Section 6.3). The test set here is the CAMELS CV suite (27 seeds at fixed fiducial cosmology), since no LH suite exists at this box size.

Table 6.2: Median Figure of Merit ($\text{FoM}_m = 1/A_{68}$, with A_{68} the 68% joint credible area on $\{\Omega_m, \sigma_8\}$) over the 200-realization ΛCDM -Planck test set, for each summary/tracer combination and astrophysics treatment. Adapted from the full pipeline results; see the associated paper for the complete table and posterior figures.

Summary	Tracer	FoM_m (fixed astro.)	FoM_m (free astro.)
$\mathcal{P}$	g	0.65	0.33
$\mathcal{P}$	HI	2.5	0.27
$\mathcal{P}$	$g \times \text{HI}$	4.7	1.0
2DM	$g \times \text{HI}$	5.2	1.2
3DM	g	2.4	0.47
3DM	HI	5.5	2.2
3DM	$g \times \text{HI}$	12	3.4

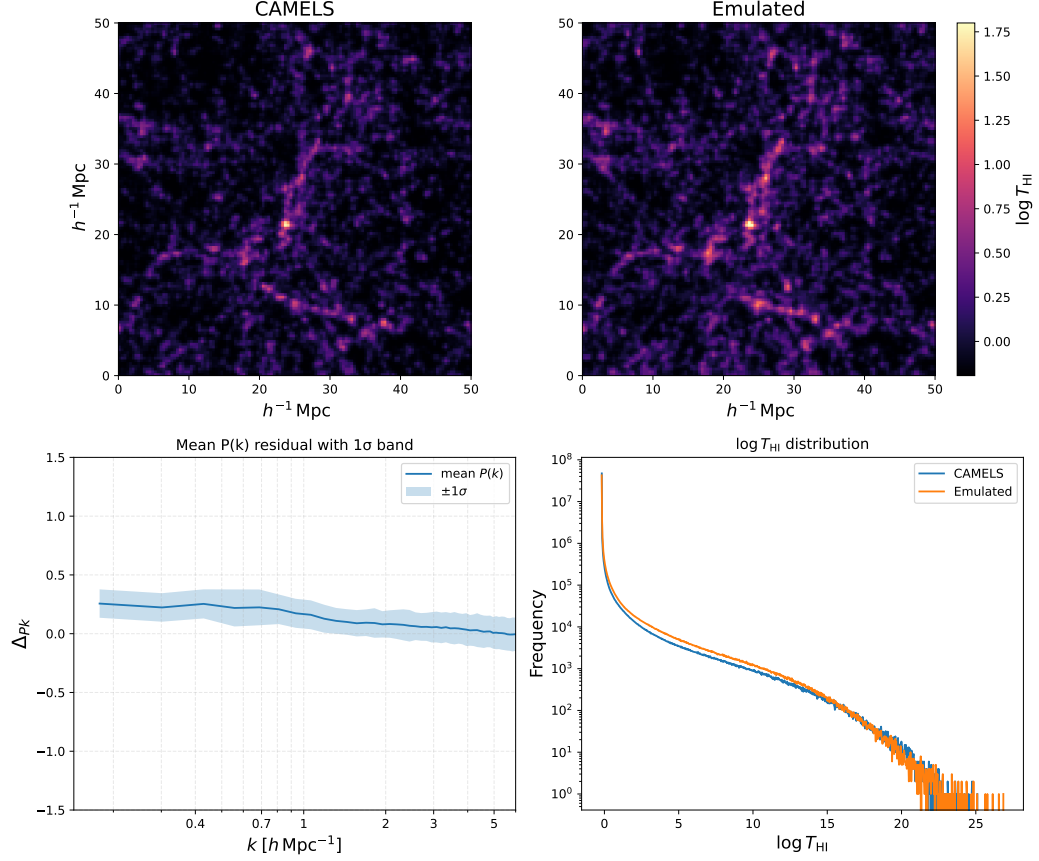


Figure 6.7: As Figure 6.5, but for 50 h^{-1} Mpc maps generated with the emulator trained on 25 h^{-1} Mpc cubes (Section 6.3). The test set here is the CAMELS CV suite (27 seeds at fixed fiducial cosmology), since no LH suite exists at this box size.

Table 6.3: Relative bias (median posterior value vs. ground truth, over the 200-realization Λ CDM-Planck test set) on Ω_m and σ_8 , in percent, for the fixed and free astrophysics cases. Adapted from the full pipeline results.

Summary	Tracer	Ω_m bias [%]		σ_8 bias [%]	
		fixed	free	fixed	free
$\mathcal{P}$	g	1.5	0.98	0.21	1.0
$\mathcal{P}$	HI	-0.23	1.0	0.64	-0.42
$\mathcal{P}$	g $\times$ HI	0.042	-0.19	1.0	0.23
2DM	g $\times$ HI	-0.72	0.75	0.73	-0.15
3DM	g	0.19	-1.1	-0.49	0.48
3DM	HI	-0.38	0.23	-0.25	-0.79
3DM	g $\times$ HI	0.084	-0.38	-0.076	0.25

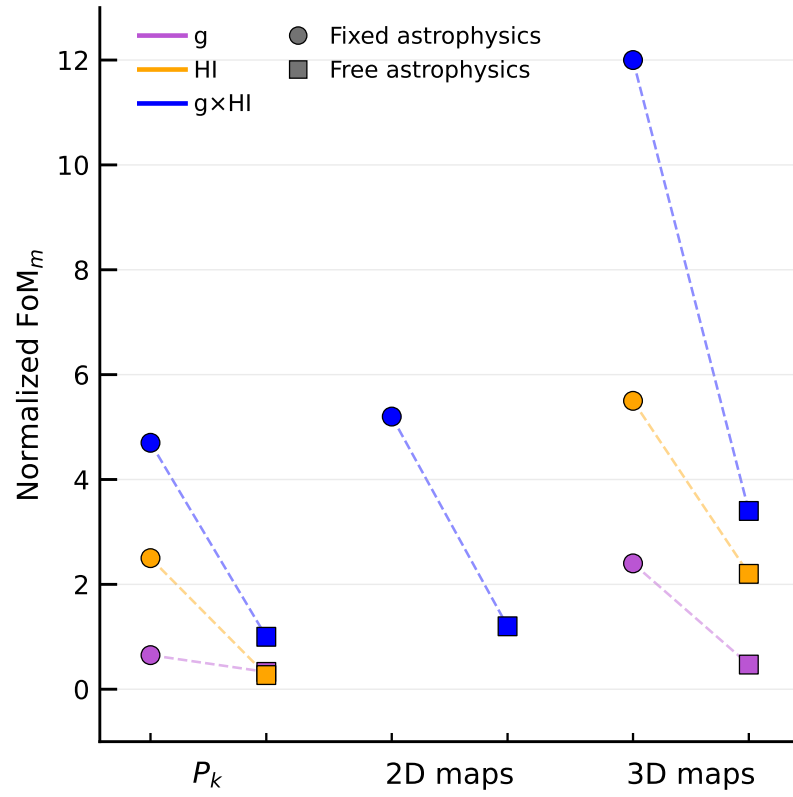


Figure 6.8: Visualizing the FOM table 6.2. The bottom axis is divided into 3 labeled parts and the y axis shows the FoM_m scores, with same colors representing same tracers.

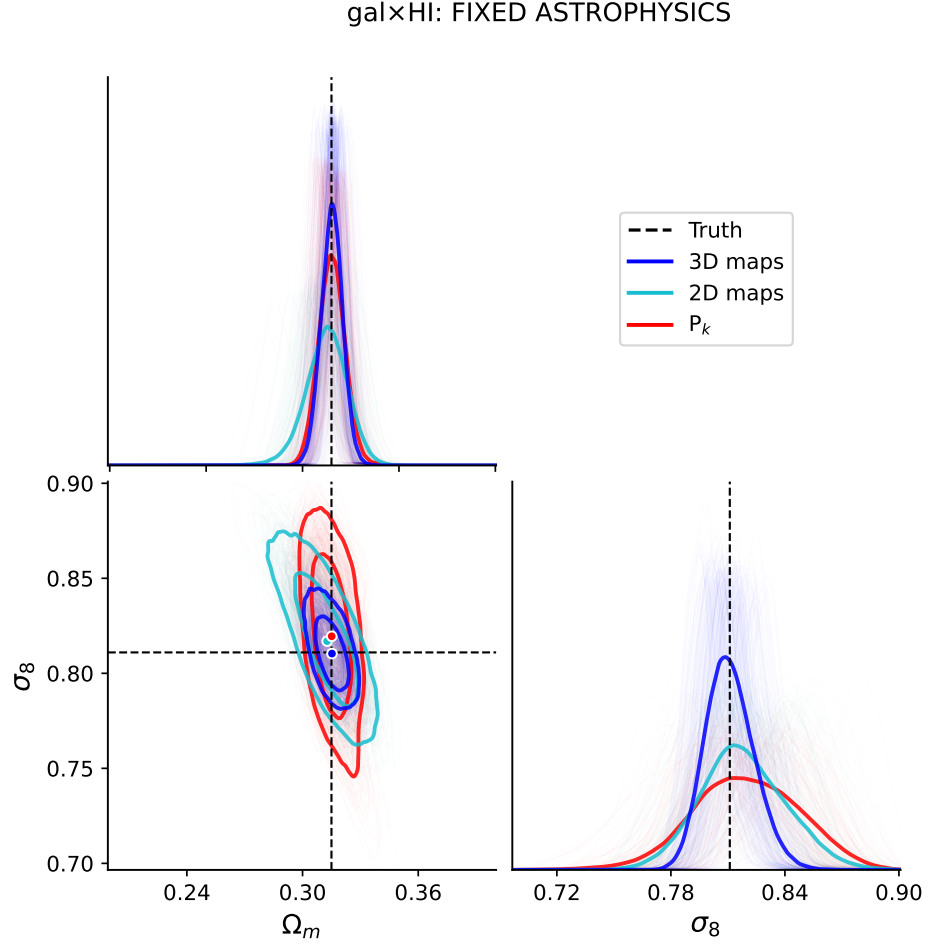


Figure 6.9: Posteriors on $\{\Omega_m, \sigma_8\}$ for $\mathcal{P}_{\text{fixedA}}^{\text{g}\times\text{HI}}$ (red) vs $\mathcal{M}_{\text{fixedA}}^{\text{g}\times\text{HI}}$ (2D, light blue, and 3D, blue) for 200 Λ CDM-Planck realizations test set (ground truth in black). Performing inference on the full 3D maps significantly improves precision. Dots in the lower left plot represent the median values computed from all the posteriors samples.

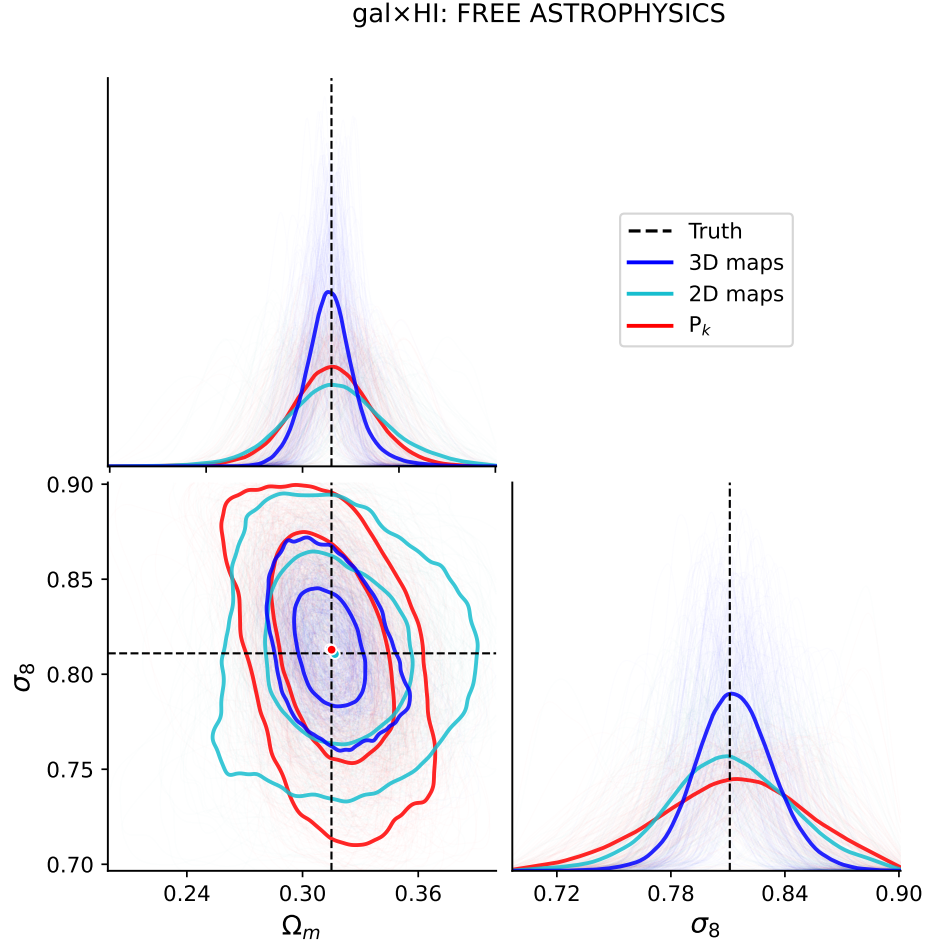


Figure 6.10: As in figure 6.9, but for the free astrophysics scenario, with astrophysical parameters marginalized over.

It is worth spelling out why these particular gains appear, since the pattern is not simply “more information is always better.” The power spectrum is, by construction, a two point statistic: it is blind to any information encoded in higher-order correlations of the field. Non-linear gravitational evolution generates exactly this kind of higher-order, non-Gaussian structure, and it is particularly sensitive to σ_8 , which sets the amplitude of fluctuations and therefore the strength of mode coupling between different scales. The convolutional encoder feeding the $3\mathcal{DM}$ case is able to extract part of this non-Gaussian information; the power spectrum, by definition, cannot. That the $2\mathcal{DM}$ case fails to reproduce this gain is itself informative: flattening the cube along the line of sight destroys precisely the along-the-line structure that carries this extra information, so the improvement is not simply a property of “using a neural network” but specifically of preserving the field’s full three-dimensional structure.

The multi-tracer gain has a complementary explanation, illustrated in Figure 6.11. Galaxy number counts are a sparse, shot-noise-dominated tracer built from halos spanning a wide range of masses and bias factors, which mixes together objects that trace the underlying dark matter field in different ways; HI brightness temperature, by contrast, is a smoother, more continuous tracer with a more direct dependence on the underlying density.

Combining the two through their cross correlation lets the inference disentangle genuine cosmological signal from tracer specific effects (shot noise, bias) that do not correlate between the two tracers, which is precisely why $\mathcal{P}^{g \times \text{HI}}$ and $3\mathcal{DM}^{g \times \text{HI}}$ outperform either single tracer case on its own. None of this should be read as a criticism of galaxy clustering as an observational probe in general: real galaxy surveys additionally exploit physically motivated bias models, much larger survey volumes, and observables such as redshift space distortions that are absent from this idealized, box limited comparison.

SBI on 3D maps (FIXED ASTROPHYSICS)

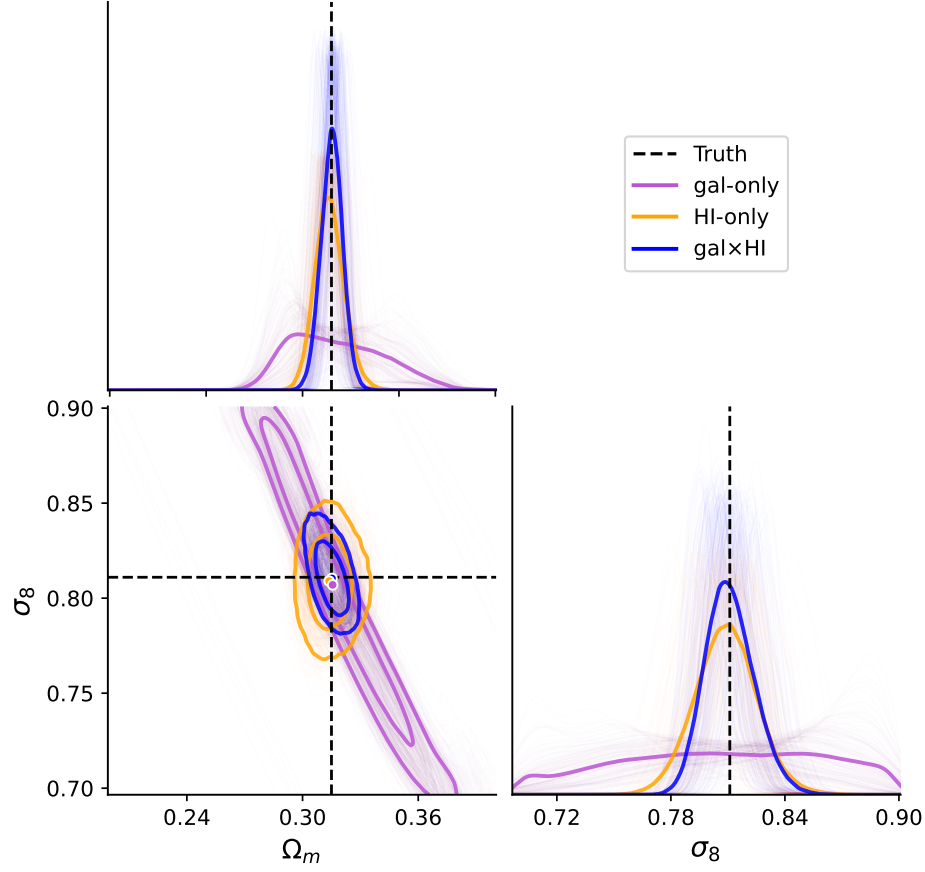


Figure 6.11: Posteriors on $\{\Omega_m, \sigma_8\}$ (fixed astrophysics) comparing single-tracers cases $3\mathcal{DM}_{\text{fixedA}}^{\text{g}}$ (purple) and $3\mathcal{DM}_{\text{fixedA}}^{\text{HI}}$ (orange) with multi-tracer case $3\mathcal{DM}_{\text{fixedA}}^{\text{g} \times \text{HI}}$ (blue).

From the perspective of this chapter, the relevant point is narrower: none of these downstream comparisons would be meaningful if the emulated fields feeding them were not themselves accurate, at both the pixel and summary statistic level. This is precisely what Section 6.7.1 was intended to establish, that the HGL emulator reproduces both the power spectrum and the pixel level statistics of the tracers it was trained on to good accuracy at $25 h^{-1} \text{ Mpc}$, with box-size extrapolation to $50 h^{-1} \text{ Mpc}$ identified as the dominant residual source of emulator-level systematic error. The downstream SBI comparison in Table 6.2 should therefore be read as a demonstration of what becomes possible once a sufficiently accurate emulator removes the need for a prohibitively large hydrodynamical simulation suite, rather than as an independent validation of the

emulator itself.

6.8 Discussion

The HGL emulator generalizes the single-cosmology HALOgen architecture (Mishra et al. 2025) to arbitrary cosmological and astrophysical parameters within the CAMELS LH prior, at a training and inference cost low enough to support a full SBI pipeline. Three limitations are worth highlighting, since they define the natural next steps for this line of work rather than being incidental to it.

First, and most importantly, generalization across box size is limited: the emulator trained at $25 h^{-1}$ Mpc shows a visibly larger power-spectrum residual when used to generate $50 h^{-1}$ Mpc maps (Section 6.7.1), most likely because the larger box exposes large scale modes and massive halos absent from the smaller training volumes. Because the current CAMELS LH suite provides paired DM/hydro simulations only at $25 h^{-1}$ Mpc, this cannot be fully resolved without new training data at larger box size; a supervised fine-tuning step, or a loss term that explicitly penalizes large scale power-spectrum residuals, would both be natural directions to explore in the meantime.

Second, the checkpoint selection issue noted in Section 6.5, that the best pixel level checkpoint and the best power-spectrum checkpoint do not coincide for the galaxy emulator, is a reminder that voxel wise loss and two point function accuracy are not interchangeable diagnostics for a field-level emulator, and that whichever one is optimized for should be chosen deliberately based on the downstream use case (map realism vs. summary-statistic accuracy) rather than by default.

Third, the entire pipeline, training, validation, and the SBI test set alike, is self consistent: everything is generated within the same emulator framework, and no attempt is made here to test transfer to independently generated maps (e.g. original CAMELS hydrodynamical maps not used in training, or maps from a different simulation suite altogether). Prior work on out of distribution sensitivity in field-level SBI has shown that even small numerical differences between simulation codes at fixed physical parameters can bias inferred posteriors (Bayer et al. 2025); since an emulator is, by construction, another source of such small-scale discrepancies relative to the simulations it approximates, a dedicated out of distribution study is a necessary step before this pipeline could be applied to real observational data, or even to hydrodynamical simulations outside the training suite.

Looking forward, the natural extensions are to retrain **HGL** once a larger paired training suite becomes available, to test transfer to independent simulation codes and to real survey-like maps (including instrumental noise, beams, and foregrounds for the HI tracer, and realistic selection functions for the galaxy tracer), and to extend the redshift range beyond the $z = 0$ snapshot considered here. More broadly, this combination of emulator and neural posterior estimation approach and the diffusion-based posterior sampling framework of Chapter 5 represent two complementary routes to the same underlying goal: recovering the cosmological information that summary statistics discard, in a computational regime where full hydrodynamical simulations, run at the scale a full posterior or a well-sampled training set requires, remain out of reach.

Chapter 7

Conclusion, Caveats and Future Work

7.1 The common thread

Every chapter of this thesis approaches the same problem from a different angle. Compressing a cosmological or astrophysical field down to a handful of summary statistics before doing inference on it throws away information, and increasingly that discarded information is exactly what upcoming, high resolution surveys are sensitive to (Section 1.1.7). The goal, throughout, has been to work with the field itself, or a learned representation of it, using deterministic emulators and generative models trained directly on hydrodynamical simulations rather than on hand calibrated prescriptions or perturbative expansions that break down on the scales of interest.

This is not five independent applications of the same idea: the chapters build on one another methodologically. LODI (Chapter 2) first demonstrates that a two stage, conditional generative pipeline can paint a full field onto a dark matter only input, and uses a latent overlap scheme, adapted from inpainting, to stitch small trained sub-volumes into a coherent large one. Cosmo-FOLD (Chapter 3) exists specifically because that overlap scheme has a real cost (generating redundant overlapping volumes) and a real failure mode on extended structures. Its sliding window generation removes the explicit overlap prescription entirely by folding the coherence requirement into the diffusion trajectory itself, and the added positional encoding(Cosmo-FOLD+pos) further improves and preserves the higher order correlations that would otherwise be lost when trying to generate large simulation volumes, especially with a sparse conditioning map.

Chapter 4 then takes the same architecture, generalizes it by making the generative model cosmology dependent, and asks a different question of it: not just a point estimate, but a posterior, with an external cosmological prior (Planck) injected through Bayes’ theorem at inference time, without retraining. Chapter 5 takes a parallel route, utilizing the diffusion model as a generative prior, to work with diffusion posterior sampling, derived from first principles, and applies it to a different inverse problem, Lyman- α tomography, where the forward model is not another neural network but a differentiable physical one (the Voigt-profile optical depth calculation). Chapter 6, finally, asks what all of this is for. It quantifies what is the quality of inference and how much is gained by moving from a power spectrum to a full 3D field, and from one tracer to two.

7.2 What has been demonstrated

This section recaps, briefly and chapter by chapter, only the results each chapter itself already establishes.

LODI (Chapter 2) showed that a conditional U-Net plus a conditional variational diffusion model, painting the 21 cm brightness temperature field onto a predicted halo distribution, recovers the 21 cm power spectrum to within 10% on average up to $k \approx 10 h \text{ Mpc}^{-1}$, and that the latent-overlap generation scheme gives a speed-up over naive sequential generation, at a full pipeline cost of about 106 seconds per generated 256^3 voxels volume on a single A100 GPU. Cosmo-FOLD (Chapter 3) showed that removing the explicit overlap prescription in favor of a sliding-window trajectory both fixes LODI’s failure on extended structure and reduces generation time by roughly 40% for a comparable case, while a model trained only on $25 h^{-1} \text{ Mpc}$ CAMELS boxes transfers to the $205 h^{-1} \text{ Mpc}$ TNG volume with the power spectrum accurate to within 10% up to $k \lesssim 5 h \text{ Mpc}^{-1}$, thereby showcasing a box size generalization result. The chapter also shows that positional encoding improves the higher order (bispectrum) statistics that would otherwise be lost at inference, especially with a sparse conditioning map. Chapter 4 showed that the same architecture can be turned into a posterior sampler, trained across the CAMELS Latin Hypercube so that astrophysical feedback is marginalized over implicitly, and that an external Planck posterior on (Ω_m, σ_8) can be folded into the generative process at inference time with no retraining. The chapter’s own discussion is explicit that its demonstration rests on a single test field and that a fuller calibration study is still needed. Chapter 5 showed that a single, unconditionally

trained diffusion prior over a joint field of five components of the IGM (four baryonic and one dark matter), combined with a differentiable forward physics model, recovers the Lyman- α forest tomography fields closely down to the non-linear scales, across a range of sightline densities and noise levels, without ever retraining between cases, most notably, the dark matter field, which never enters the forward model at all, is still recovered via correlations the prior learned purely from unconditional training. Chapter 6 showed that a similar emulating machinery, generalized across cosmological and astrophysical parameters via conditioning, generates a galaxy/HI field pair in about 10^{-2} seconds on a GPU, and that using it as the forward simulator inside a full 3D field-level SBI pipeline recovers Ω_m and σ_8 with bias below about 1.5% and several fold tighter constraints than an equivalent power spectrum analysis.

7.3 Some critical points

7.3.1 Training data volume and generalization

A pattern repeats across every chapter. That is, every machine learning model here is trained on a strikingly small number of independent simulations, made usable only through quite a bit of sub-volume chunking and augmentation. LODI trains its diffusion model on 5 simulations, whereas Cosmo-FOLD’s TNG-direct case uses 4. Chapter 4 uses 700 (plus 300 for testing) from the CAMELS Latin Hypercube and Chapter 5 uses 8 (plus 2 for testing); Chapter 6’s emulator is trained on 508 seeds, with again the same rotational augmentation to increase the effective dataset size.

On the other hand, in every case where an out of distribution test was actually attempted, some degradation was found. Cosmo-FOLD’s power spectrum visibly degrades beyond the training box’s own largest mode when applied to a volume $8\times$ larger. Chapter 6’s emulator shows a visibly less accurate power spectrum at $50 h^{-1}$ Mpc after training only at $25 h^{-1}$ Mpc, and its own discussion notes that the pipeline has never been tested against an independently generated simulation suite at all, citing the concern (Bayer et al. 2025) that even small inter code numerical differences can bias an inferred posterior. In short, these models interpolate well within the distribution they were trained on, without the ability to resolve, what happens outside it, or in other words, extrapolate, precisely the gap Section 7.4.3 below is aimed at.

7.3.2 Computational cost

The motivation stated in the Introduction (Section 1.7) is quantitative: the IllustrisTNG simulations cost on the order of 35 million and 18 million CPU-hours for their two largest runs, and EAGLE cost roughly 4.5 million. It is worth collecting, in one place, just how large a gap this thesis’s models open up against that baseline, since each chapter only ever quotes its own number. LODI’s full two-stage pipeline generates one 256^3 volume in about 106 seconds on a single GPU (with 50 diffusion time steps, which when reduced, can increase speed), after roughly 50 GPU-hours of one-time training. Cosmo-FOLD, takes about five hours in generation time to produce a full-volume generation of TNG-300 (see Table 3.1 for details). This speed, of course can be increased significantly by working in the latent space or by reducing the time steps. Chapter 6’s emulator, once trained, generates a galaxy/HI field pair in about 10^{-2} seconds. These are not incremental speed-ups over the simulations they stand in for, they are the difference between a computation that needs a dedicated allocation on a super computing facility and one that can be run, repeatedly, inside the inner loop of a simulation-based inference pipeline, which is precisely why Chapter 6’s SBI results (hundreds of thousands of forward simulations) are possible at all. This speed comes at a cost. When we compare emulated fields directly to simulated ones, the errors are not spread evenly across the field: they are concentrated at the smallest scales, in the densest regions, and in higher-order statistics.

7.3.3 Which loss function?

Chapter 6’s discussion surfaces another side of this story. The checkpoint that minimizes voxel wise reconstruction loss is not the checkpoint that best reproduces the power spectrum. This goes on to show that neither loss function is completely trustworthy for whether the model’s implicit physical knowledge is *optimal* for a given downstream question. Of course, in the limit of perfect training and infinite model expressive capacity, these loss functions would converge the model to the same result. However, in practice, with finite data and finite capacity, they do not, and which one is "right" is genuinely hit or miss rather than settled once for the whole thesis. Part of the reason is that a loss function is evaluated in a specific *space*, and as can be seen, this thesis’s chapters do not even agree on which space that should be. HALOgen’s masked loss (Section 2.4.1) combines an L_2 term with a quantile loss directly in voxel space, precisely because a single symmetric loss cannot handle the extreme sparsity of halo positions, whereas the variational diffusion prior used throughout LODI, Cosmo-FOLD, and FOREST uses a form

of the KL divergence and integrates its loss over a whole family of *noised* versions of the target field, weighted across noise levels by $\gamma'(t)$ (Eq. 2.13). Another modification comes when FOREST’s prior, unlike the others, computes that same diffusion loss in a VAE-compressed *latent* space rather than voxel space. So this means that FOREST’s total pipeline uses a voxel space MSE and fourier space loss, while the diffusion loss is calculated in latent space. Each of these loss terms are sensitive to different implicit biases by construction, and the emulator’s checkpoint dilemma is simply a clear way to show that.

7.3.4 Limitations and caveats

Each chapter comes with its own set of caveats and it is important to lay them out in one place to make it clear what improvements are necessary and required to be carried out in the future work. Beyond the per chapter limitations already stated, such as LODI’s untested cosmology/redshift generalization, or Chapter 6’s box-size and out-of-distribution gaps, three further limitations are visible only by looking across chapters.

Astrophysics marginalization and generative models. Chapter 4, is the only chapter that trains a generalized diffusion model that takes input the cosmology parameters and marginalizes over the astrophysics. However, if looked carefully, we notice that Chapter 6’s HGL goes a step further and trains across all of astrophysics available in the CAMELS dataset (the CAMELS Latin Hypercube). LODI, Cosmo-FOLD, and FOREST do not: all three are trained at a single, fixed astrophysical model. Although this is not a flaw in any one chapter, it raises a question. That is: why was the deterministic HGL emulator of the SBI work the only one to be trained on all of the astrophysics, but not the stochastic diffusion models. The answer lies in 2 points. First, although diffusion models are incredibly powerful generative models, they are slow due to their looped sampling procedure, which is slow for general SBI pipelines. Of course they can be sped up, but that requires further work and fine tuning. The second reason is that the diffusion model’s loss has a hard time working with sparse fields, which unfortunately both the galaxy and HI fields are. To generate HI and galaxy fields while simultaneously using the capabilities of a generative diffusion model, one would require to modify the loss functions, which although possible, is not trivial, and hence left for future work.

Everything is validated against simulation truth, not survey-realistic data.

No chapter yet runs its full pipeline against data carrying realistic instrumental noise, beam effects, masking, or a survey selection function end to end. **FOREST** comes close to working with semi realistic data, since it incorporates noise in its observed fields by varying the signal to noise ratio and also the sightline coverage synthetically. Chapter 6’s discussion names this explicitly as future work, but neither is the same as a real survey footprint. If we wish to see these methods applied to real survey data, this gap has to be bridged.

No chapter spans more than one redshift. **LODI**, **Cosmo-FOLD**, **FOREST** and **HGL** all operate at $z = 0$. This results in us not working in light cones, which cosmologists actually care about. Every result in this thesis is therefore a snapshot statement, not yet a statement about structure growth or evolution. This is in our opinion, the easiest work to move towards next, as it requires minimal modification of our presented work.

7.4 Future work

The extensions below range from short-term work, made possible by combining different works in this thesis, to more open, higher risk but higher rewarding directions.

7.4.1 Full box, multi simulation IGM reconstruction

As presented **FOREST** already proposes this (Section 5.8.3), and it is repeated here mainly because it is the single most direct, already half built extension in the thesis. The current DPS results cover one $40 \times 10 \times 10 (h^{-1} \text{ Mpc})^3$ sub-volume, one sixteenth of the full simulation box, while **Cosmo-FOLD** (Chapter 3) already solves exactly the coherent large volume generation problem this requires, via its sliding window diffusion trajectory. Applying **Cosmo-FOLD**’s generation scheme directly to Chapter 5’s guided reverse process, rather than treating each tile as an independent posterior draw, is the most concrete short deliverable to come out of combining two existing chapters of this thesis. However, there are some things to sort out before attempting this, such as constructing a better likelihood term, and hence this is left for the future (immediate future!).

7.4.2 Lightcones: stitching multi redshift emulator outputs

As mentioned right before this section, every generative model in this thesis is trained and evaluated at a single, fixed redshift (coeval boxes) (Section 7.3.4). Building a lightcone requires the opposite: a field, or a stack of fields, that varies smoothly with the observed frequency/redshift along the line of sight. The standard lightcone construction approach already used elsewhere in the field: train (or fine tune) the existing conditional architectures at a handful of discrete redshift snapshots and interpolate the conditioning between them, the same strategy used to stitch coeval boxes into 21 cm reionization lightcones, and, more directly analogous to this thesis’s own architecture, the strategy of training separate per-redshift generative models and stitching them together used for line intensity mapping lightcones by [Moriwaki et al. \(2025, CosmoGLINT\)](#). It would be interesting to see this idea extended to our **Cosmo-FOLD** method and generate large volume lightcones.

7.4.3 Simulation-independent (JEPA-style) representations

This is a direct response to the out of distribution gap named in Section 7.3.1. Every model in this thesis has been trained and validated within one simulation suite (except for **Cosmo-FOLD**, which has been tested on similar subgrid physics but different simulation sizes), and Chapter 6 names cross simulation suite transfer as an open, unresolved direction. There are recent works, for example [Ore et al. \(2025\)](#) (SKATR): a self-supervised vision transformer trained on 21 cm maps whose learned encoding is then reused, cheaply and without retraining the encoder itself, for downstream regression and generative inference of astrophysical and cosmological parameters. Its result shows a property this thesis is missing, which is that the authors show that SKATR’s representations generalize out of domain to differently simulated, differently noised, and higher resolution datasets than the ones the encoder was trained on. This is a direct match to **HGL**’s own unresolved question (transfer to an independently generated simulation, Section 6.8) and to **LODI/Cosmo-FOLD**’s 21 cm and intensity mapping setting specifically. A complementary, **CAMELS**-specific precedent for the same underlying goal is [Jo et al. \(2025\)](#), whose Model Insensitive Estimator is trained across the **CAMELS** TNG, SIMBA, Astrid, and EAGLE simulation suites (the same suites that exist in **CAMELS**) to extract only features common across all four while discarding suite specific ones, improving out of distribution Ω_m/σ_8 estimation accuracy by roughly 17%/38% relative to a suite specific baseline.

Both approaches, and the broader self supervised Joint Embedding Predictive Architecture family SKATR itself is built out of (Sobal et al. 2022; Assran et al. 2023), share the same mechanism: train an encoder to predict a representation, rather than raw pixels, so that simulation suite specific texture is discarded while the physical structure needed for the downstream task is retained, making the learned latent representation suite invariant, directly attacking the same failure mode already observed or mentioned in Chapter 3 and Chapter 6. Applying this to this thesis would naturally be the obvious next big step, departing from voxel space into a more physically meaningful latent space.

7.4.4 Diffusion based likelihood for simulation discrimination

The idea of using a trained diffusion model’s own (approximate) likelihood or score to distinguish between candidate simulation models, rather than training a separate classifier from scratch, is a natural extension of machinery this work already has in hand. The general goal, likelihood-free discrimination between competing simulation models, is already an active line of work in a closely related setting: Binnie et al. (2025) perform simulation-based Bayesian model selection between competing reionization scenarios using a 3D CNN summary statistic inside the `pydelfi` framework, successfully distinguishing the model used to generate mock 21 cm lightcone data in every case tested. What has not yet been done, to our knowledge, is the same kind of model discrimination using a diffusion model’s own score or evidence directly, rather than a separately-trained classifier network; training one unconditional prior per candidate simulation model (or per candidate subgrid/feedback prescription), and comparing an ELBO-based evidence estimate across priors, would be a natural, largely code-reuse extension of this thesis’s own machinery into a genuinely different scientific question (model comparison rather than parameter or field inference). Another extension could be to get an estimate as to how close a certain cosmological field is, given a certain set of cosmological and astrophysical parameters, using the same machinery mentioned above. Assuming that the diffusion model has been trained jointly taking as inputs both fields and corresponding parameters, at test time, we pass a test field and candidate parameters to the trained diffusion model, and ask the model to noise and denoise the said field under the candidate parameters, and store the calculated ELBO values. If the model was trained perfectly, then the ELBO (stochastic lower bound estimate of the true likelihood) would be highest for those candidate that were used to generate the test field, giving us an estimate of the un normalized posterior, or at least an idea of

what the ground truth parameters are, without ever using an explicit gaussian or other simple likelihoods.

7.4.5 Extending DPS to survey-realistic foreground and RFI removal

FOREST’s guided-sampling machinery is not specific to Lyman- α tomography. What it actually requires is an unconditional prior over the field of interest and a differentiable forward model connecting that field to a corrupted or partial observation. Radio-frequency interference flagging and foreground contamination in 21 cm intensity mapping (Section 1.5) is exactly this kind of problem, with the forward model replaced by a masking/foreground-mixing operator instead of the Voigt-profile optical depth calculation. Indeed, recent work in 21 cm cosmology has independently converged on the same differentiable, end to end philosophy: Kern (2025) present a fully differentiable Bayesian forward model that jointly samples the 21 cm signal, diffuse foregrounds, and the instrumental beam response at the field level, using automatic differentiation to enable Hessian-preconditioned HMC over an $\mathcal{O}(10^5)$ -dimensional parameter space. Their framework, however, represents the sky and beam directly in a harmonic (spherical stripe/cap) basis with an explicitly parametrized foreground and beam model, rather than learning an unconditional generative prior over the field and combining it with likelihood guidance at inference time; FOREST’s contribution would be complementary in exactly this sense, replacing their hand-parametrized priors with a learned VDM prior over the field, guided at sampling time by a differentiable masking/foreground operator in place of their forward model. This is not a hypothetical extension: the natural extension specific to this thesis is to train the prior on masks constructed from realistic SKA/MeerKAT flagging statistics (rather than idealized foreground contamination alone) and reuse Chapter 5’s guidance formalism essentially unchanged, which would also be a first, concrete step toward the survey-realism gap named in Section 7.3.4.

7.4.6 Structured (large-scale-structure) initialization of the reverse diffusion process

Lastly, there is something that is worth thinking about and hopefully improving upon. Every diffusion model in this thesis begins its reverse process from pure Gaussian noise, $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ (Section 1.3.6), and relies entirely on the learned network to build

large-scale structure back up over many denoising steps before adding small-scale detail, a generic property of diffusion models on natural images as much as on cosmological fields. The idea of replacing that flat-spectrum Gaussian seed with a seed that already carries realistic large-scale correlations, and letting the network’s job be only to add the non-linear, small-scale structure on top, has a direct link in [Randono \(2025\)](#), who propose seeding diffusion models with scale-invariant noise explicitly motivated by the analogous role of primordial density fluctuations in structure formation. In this thesis’s own setting, a cheap large-scale realization (a linear or 2LPT field, or the coarse conditioning field already used in Chapter 4) is available essentially for free, so rather than an analogy to cosmology motivating a general noise choice, an actual cosmological field could be the seed. This could in principle save a lot of diffusion steps that try to rebuild the large scale correlation from scratch, instead of using what already is provided. Moreover, this could be the eventual replacement for the positional embedding of **Cosmo-FOLD**, since the large scale info would be embedded in the diffusion model by default. This is an ambitious taking as it would require reformulating the mathematical idea of the diffusion model, and currently is beyond the scope of this work.

7.5 Closing remark

Taken as a whole, this thesis argues, empirically rather than by assertion, that the field itself, not the two-point function computed from it, is where the recoverable cosmological information actually lives, and that generative models are now capable, and fast enough, to generate high fidelity fields directly rather than with their compressed summaries. Each chapter contributes one piece of that argument: **LODI** shows a field can be painted at all; **Cosmo-FOLD** shows it can be done coherently and efficiently at scale; 4 shows a posterior, not just a point estimate, can be sampled, and makes a diffusion model generalized over different cosmology. **FOREST** shows that posterior-sampling machinery generalizes to a real inverse problem, guided with a physical forward model, and Chapter 6 shows the whole game of field level emulation pays for itself in tighter, less biased cosmological constraints. What remains, set out above, is mostly a matter of scale, speed and generality rather than of new ideas, namely larger volumes, more redshifts, more simulation suites, more astrophysical scenarios, and more realistic data, applied to machinery that this thesis has already shown, chapter by chapter, works.

Bibliography

- Abbott T. M. C., et al., 2019, [Phys. Rev. D](#), 99, 123505
- Akeret J., Refregier A., Amara A., Seehars S., Hasner C., 2015, [Journal of Cosmology and Astroparticle Physics](#), 2015, 043–043
- Alam S., et al., 2017, [Mon. Not. Roy. Astron. Soc.](#), 470, 2617
- Alsing J., Charnock T., Feeney S., Wandelt B., 2019, [Monthly Notices of the Royal Astronomical Society](#)
- Anderson C. J., et al., 2018, [MNRAS](#), 476, 3382
- Andrews A., Jasche J., Lavaux G., Schmidt F., 2023, [MNRAS](#), 520, 5746
- Andrianomena S., Villaescusa-Navarro F., Hassan S., 2022, Emulating cosmological multifields with generative adversarial networks, [doi:10.48550/arXiv.2211.05000](#), <https://ui.adsabs.harvard.edu/abs/2022arXiv221105000A>
- Andrianomena S., Hassan S., Villaescusa-Navarro F., 2024, Cosmological multifield emulator ([arXiv:2402.10997](#)), <https://arxiv.org/abs/2402.10997>
- Angulo R. E., Hahn O., 2022, [Living Reviews in Computational Astrophysics](#), 8, 1
- Ansari R., et al., 2012, [Astronomy & Astrophysics](#), 540, A129
- Assran M., Duval Q., Misra I., Bojanowski P., Vincent P., Rabbat M., LeCun Y., Ballas N., 2023, Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture ([arXiv:2301.08243](#)), <https://arxiv.org/abs/2301.08243>
- Bairagi A., Wandelt B., Villaescusa-Navarro F., 2025, How many simulations do we need for simulation-based inference in cosmology? ([arXiv:2503.13755](#)), <https://arxiv.org/abs/2503.13755>

- Bandura K., et al., 2014, [Proc. SPIE Int. Soc. Opt. Eng.](#), 9145, 22
- Bartolo N., Komatsu E., Matarrese S., Riotto A., 2004, [Physics Reports](#), 402, 103–266
- Battye R. A., Davies R. D., Weller J., 2004, [Mon. Not. Roy. Astron. Soc.](#), 355, 1339
- Battye R. A., Browne I. W. A., Dickinson C., Heron G., Maffei B., Pourtsidou A., 2013, [Mon. Not. Roy. Astron. Soc.](#), 434, 1239
- Bayer A. E., Seljak U., Modi C., 2023, in Proceedings of the 40th International Conference on Machine Learning. ([arXiv:2307.09504](#))
- Bayer A. E., et al., 2025, [The Astrophysical Journal](#), 989, 207
- Bernardeau F., Colombi S., Gaztañaga E., Scoccimarro R., 2002, [Physics Reports](#), 367, 1–248
- Bernardini M., et al., 2025, EMBER-2: Emulating baryons from dark matter across cosmic time with deep modulation networks ([arXiv:2502.15875](#))
- Berti M., Spinelli M., Haridasu B. S., Viel M., Silvestri A., 2022, [Journal of Cosmology and Astroparticle Physics](#), 2022, 018
- Berti M., Spinelli M., Viel M., 2023, 21cm Intensity Mapping cross-correlation with galaxy surveys: current and forecasted cosmological parameters estimation for the SKAO, [doi:10.48550/arXiv.2309.00710](#), <http://arxiv.org/abs/2309.00710>
- Berti M., Spinelli M., Viel M., 2024, [MNRAS](#), 529, 4803
- Bharadwaj S., Nath B. B., Nath B. B., Sethi S. K., 2001, [J. Astrophys. Astron.](#), 22, 21
- Bi H., Davidsen A. F., 1997, [The Astrophysical Journal](#), 479, 523–542
- Binnie T., Zhao X., Pritchard J. R., Mao Y., 2025, Likelihood-free Model Selection in Cosmic Reionization with Three-dimensional Tomographic 21 cm Lightcone Images ([arXiv:2502.08152](#)), <https://arxiv.org/abs/2502.08152>
- Blanchard A., et al., 2020, [Astron. Astrophys.](#), 642, A191
- Bolton J. S., Puchwein E., Sijacki D., Haehnelt M. G., Kim T.-S., Meiksin A., Regan J. A., Viel M., 2017, [Monthly Notices of the Royal Astronomical Society](#), 464, 897

- Bull P., et al., 2016, [Phys. Dark Univ.](#), 12, 56
- CHIME Collaboration et al., 2022, arXiv e-prints, p. [arXiv:2202.01242](#)
- Carucci I. P., Villaescusa-Navarro F., Viel M., 2017, The cross-correlation between 21cm intensity mapping maps and the Lyman-alpha forest in the post-reionization era, [doi:10.48550/arXiv.1611.07527](#), <http://arxiv.org/abs/1611.07527>
- Cauci S., Colombi S., Pichon C., Rollinde E., Petitjean P., Sousbie T., 2008, Recovering the topology of the IGM at $z \sim 2$ ([arXiv:0801.4335](#)), <https://arxiv.org/abs/0801.4335>
- Chadayammuri U., Ntampaka M., ZuHone J., Bogdán Á., Kraft R. P., 2023, [Monthly Notices of the Royal Astronomical Society](#), 526, 2812–2829
- Chang T.-C., Pen U.-L., Peterson J. B., McDonald P., 2008, [Phys. Rev. Lett.](#), 100, 091303
- Chang T.-C., Pen U.-L., Bandura K., Peterson J. B., 2010, [Nature](#), 466, 463
- Charnock T., Lavaux G., Wandelt B. D., 2018, [Physical Review D](#), 97
- Chisari N. E., et al., 2019, [The Open Journal of Astrophysics](#), 2, 4
- Chudaykin A., Ivanov M. M., Philcox O. H. E., Simonović M., 2020, [Phys. Rev. D](#), 102, 063533
- Chung H., Kim J., McCann M. T., Klasky M. L., Ye J. C., 2023, in International Conference on Learning Representations. ([arXiv:2209.14687](#))
- Cole A., Miller B. K., Witte S. J., Cai M. X., Grootes M. W., Nattino F., Weniger C., 2022, [Journal of Cosmology and Astroparticle Physics](#), 2022, 004
- Collaboration P., et al., 2020, [Astronomy & Astrophysics](#), 641, A6
- Collaboration T. C., et al., 2026, Overview of the Canadian Hydrogen Observatory and Radio Transient Detector (CHORD) Project ([arXiv:2607.09374](#)), <https://arxiv.org/abs/2607.09374>
- Cunnington S., 2022, [MNRAS](#), 512, 2408

- Cunnington S., et al., 2022, HI intensity mapping with MeerKAT: power spectrum detection in cross-correlation with WiggleZ galaxies ([arXiv:2206.01579](#))
- DESI Collaboration et al., 2025, [arXiv e-prints](#), p. [arXiv:2503.14738](#)
- Dai B., Seljak U., 2024, Multiscale Flow for Robust and Optimal Cosmological Analysis, [doi:10.48550/arXiv.2306.04689](#), <http://arxiv.org/abs/2306.04689>
- Dash C. B. V., Sarkar T. G., 2021, [JCAP](#), 02, 016
- Dawson et. al K. S., 2013, [The Astronomical Journal](#), 145, 10
- Dev A., et al., 2023, [Monthly Notices of the Royal Astronomical Society](#), 523, 2693–2709
- Dhariwal P., Nichol A., 2021, Diffusion Models Beat GANs on Image Synthesis, [doi:10.48550/arXiv.2105.05233](#), <http://arxiv.org/abs/2105.05233>
- Doeser L., Jamieson D., Stopyra S., Lavaux G., Leclercq F., Jasche J., 2024, [Monthly Notices of the Royal Astronomical Society](#), 535, 1258–1277
- Efron B., 2011, [Journal of the American Statistical Association](#), 106, 1602
- Efstathiou G., Davis M., White S. D. M., Frenk C. S., 1985, [ApJS](#), 57, 241
- Euclid Collaboration e. a., 2020, [Astronomy & Astrophysics](#), 642, A191
- Feng Y., Chu M.-Y., Seljak U., McDonald P., 2016, [Monthly Notices of the Royal Astronomical Society](#), 463, 2273
- Furlanetto S., Oh S. P., Briggs F., 2006, [Phys. Rept.](#), 433, 181
- Greenberg D. S., Nonnenmacher M., Macke J. H., 2019, Automatic Posterior Transformation for Likelihood-Free Inference ([arXiv:1905.07488](#)), <https://arxiv.org/abs/1905.07488>
- Gunn J. E., Peterson B. A., 1965, [ApJ](#), 142, 1633
- Hassan S., et al., 2022, [The Astrophysical Journal](#), 937, 83
- Hastings W. K., 1970, [Biometrika](#), 57, 97
- He K., Zhang X., Ren S., Sun J., 2015, Deep Residual Learning for Image Recognition, [doi:10.48550/arXiv.1512.03385](#), <http://arxiv.org/abs/1512.03385>

- Hildebrandt H., et al., 2017, *Mon. Not. Roy. Astron. Soc.*, 465, 1454
- Hinshaw G., et al., 2013, *ApJS*, 208, 19
- Hitz P., Berner P., Crichton D., Hennig J., Refregier A., 2025, Fast Simulation of Cosmological Neutral Hydrogen based on the Halo Model ([arXiv:2410.01694](https://arxiv.org/abs/2410.01694)), doi:<https://doi.org/10.1088/1475-7516/2025/04/003>, <https://arxiv.org/abs/2410.01694>
- Ho J., Jain A., Abbeel P., 2020, Denoising Diffusion Probabilistic Models, doi:10.48550/arXiv.2006.11239, <http://arxiv.org/abs/2006.11239>
- Horowitz B., Lee K.-G., White M., Krolewski A., Ata M., 2019, *The Astrophysical Journal*, 887, 61
- Horowitz B., Zhang B., Lee K.-G., Kooistra R., 2021, *The Astrophysical Journal*, 906, 110
- Horowitz B., Dornfest M., Lukić Z., Harrington P., 2022, *The Astrophysical Journal*, 941, 42
- Hu W., Wang X., Wu F., Wang Y., Zhang P., Chen X., 2020, *Mon. Not. Roy. Astron. Soc.*, 493, 5854
- Irfan M. O., et al., 2022, *MNRAS*, 509, 4923
- Ivezić v., et al., 2019, *The Astrophysical Journal*, 873, 111
- Jamieson D., Li Y., Oliveira R. A. d., Villaescusa-Navarro F., Ho S., Spergel D. N., 2023, *The Astrophysical Journal*, 952, 145
- Jasche J., Wandelt B. D., 2013, *MNRAS*, 432, 894
- Jeffrey N., Alsing J., Lanusse F., 2020, *Monthly Notices of the Royal Astronomical Society*, 501, 954
- Jeffrey N., et al., 2024, Dark Energy Survey Year 3 results: likelihood-free, simulation-based w CDM inference with neural compression of weak-lensing map statistics ([arXiv:2403.02314](https://arxiv.org/abs/2403.02314)), <https://arxiv.org/abs/2403.02314>

- Jo Y., Genel S., Sengupta A., Wandelt B., Somerville R., Villaescusa-Navarro F., 2025, Towards Robustness Across Cosmological Simulation Models TNG, SIMBA, ASTRID, and EAGLE, [doi:10.48550/arXiv.2502.13239](https://doi.org/10.48550/arXiv.2502.13239), <http://arxiv.org/abs/2502.13239>
- Jones M. G., Haynes M. P., Giovanelli R., Moorman C., 2018, *Monthly Notices of the Royal Astronomical Society*, 477, 2–17
- Karchev K., Trotta R., Weniger C., 2022, *Monthly Notices of the Royal Astronomical Society*, 520, 1056–1072
- Karchev K., Trotta R., Weniger C., 2023, SimSIMS: Simulation-based Supernova Ia Model Selection with thousands of latent variables ([arXiv:2311.15650](https://arxiv.org/abs/2311.15650)), <https://arxiv.org/abs/2311.15650>
- Kern N., 2025, *Mon. Not. Roy. Astron. Soc.*, 541, 687
- Kingma D. P., Welling M., 2022, Auto-Encoding Variational Bayes, [doi:10.48550/arXiv.1312.6114](https://doi.org/10.48550/arXiv.1312.6114), <http://arxiv.org/abs/1312.6114>
- Kingma D. P., Salimans T., Poole B., Ho J., 2023, Variational Diffusion Models, <http://arxiv.org/abs/2107.00630>
- Kodi Ramanah D., Charnock T., Villaescusa-Navarro F., Wandelt B. D., 2020, *MNRAS*, 495, 4227
- Kovetz E. D., et al., 2017, *arXiv e-prints*, p. [arXiv:1709.09066](https://arxiv.org/abs/1709.09066)
- Lavaux G., Jasche J., Leclercq F., 2019, Systematic-free inference of the cosmic matter density field from SDSS3-BOSS data ([arXiv:1909.06396](https://arxiv.org/abs/1909.06396)), <https://arxiv.org/abs/1909.06396>
- Leclercq F., Heavens A., 2021, *Monthly Notices of the Royal Astronomical Society: Letters*, 506, L85
- Lee K.-G., et al., 2014, *The Astrophysical Journal*, 795, L12
- Lee K.-G., et al., 2018, *The Astrophysical Journal Supplement Series*, 237, 31
- Lemos P., et al., 2023, SimBIG: Field-level Simulation-Based Inference of Galaxy Clustering ([arXiv:2310.15256](https://arxiv.org/abs/2310.15256)), <https://arxiv.org/abs/2310.15256>

- Li Y., Ni Y., Croft R. A. C., Di Matteo T., Bird S., Feng Y., 2021, [Proceedings of the National Academy of Sciences](#), 118
- Li Z., Guo H., Mao Y., 2022, Theoretical Models of the Atomic Hydrogen Content in Dark Matter Halos ([arXiv:2207.10414](#)), <https://arxiv.org/abs/2207.10414>
- Loshchilov I., Hutter F., 2019, Decoupled Weight Decay Regularization ([arXiv:1711.05101](#)), <https://arxiv.org/abs/1711.05101>
- Lugmayr A., Danelljan M., Romero A., Yu F., Timofte R., Van Gool L., 2022, RePaint: Inpainting using Denoising Diffusion Probabilistic Models, <http://arxiv.org/abs/2201.09865>
- Lynds R., 1971, [ApJL](#), 164, L73
- Ma W., et al., 2025, [Astronomy & Astrophysics](#), 695, A241
- Maitra S., Viel M., Kulkarni G., 2026, [Monthly Notices of the Royal Astronomical Society](#), 549
- Makinen T. L., Charnock T., Lemos P., Porqueres N., Heavens A. F., Wandelt B. D., 2022, [The Open Journal of Astrophysics](#), 5
- Mallaby-Kay M., et al., 2021, [ApJS](#), 255, 11
- Masui K. W., et al., 2013, [ApJL](#), 763, L20
- McQuinn M., 2016, [Annual Review of Astronomy and Astrophysics](#), 54, 313–362
- McQuinn M., Zahn O., Zaldarriaga M., Hernquist L., Furlanetto S. R., 2006, [Astrophys. J.](#), 653, 815
- MeerKLASS Collaboration 2025, Revealing cosmological fluctuations in 21cm intensity maps with MeerKLASS: from maps to power spectra, <https://arxiv.org/abs/2510.27549>
- Meiksin A. A., 2009, [Reviews of Modern Physics](#), 81, 1405–1469
- Metropolis N., Rosenbluth A. W., Rosenbluth M. N., Teller A. H., Teller E., 1953, [J. Chem. Phys.](#), 21, 1087

- Miller B. K., Cole A., Forré P., Louppe G., Weniger C., 2021, Truncated Marginal Neural Ratio Estimation - Data, doi:10.5281/ZENODO.5043706, <https://zenodo.org/record/5043706>
- Mishra S., Trotta R., Viel M., 2025, Large, fast and accurate HI intensity maps with latent overlap diffusion ([arXiv:2506.08086](https://arxiv.org/abs/2506.08086)), <https://arxiv.org/abs/2506.08086>
- Mishra S., Trotta R., Viel M., 2026, Cosmo-FOLD: Fast generation and upscaling of field-level cosmological maps with overlap latent diffusion ([arXiv:2601.14377](https://arxiv.org/abs/2601.14377)), <https://arxiv.org/abs/2601.14377>
- Modi C., Feng Y., Seljak U., 2018, JCAP, 2018, 028
- Monaghan J. J., 1992, [Annu. Rev. Astron. Astrophys.](#), 30, 543
- Moriwaki K., Jun R. L., Osato K., Yoshida N., 2025, [Monthly Notices of the Royal Astronomical Society](#), 545
- Nasir F., Bolton J. S., Viel M., Kim T.-S., Haehnelt M. G., Puchwein E., Sijacki D., 2017, [Monthly Notices of the Royal Astronomical Society](#), 471, 1056
- Nelson D., et al., 2021, The IllustrisTNG Simulations: Public Data Release ([arXiv:1812.05609](https://arxiv.org/abs/1812.05609)), <https://arxiv.org/abs/1812.05609>
- Newburgh L. B., et al., 2016, [Proc. SPIE Int. Soc. Opt. Eng.](#), 9906, 99065X
- Obuljen A., Castorina E., Villaescusa-Navarro F., Viel M., 2018, [Journal of Cosmology and Astroparticle Physics](#), 2018, 004
- Obuljen A., Alonso D., Villaescusa-Navarro F., Yoon I., Jones M., 2019, [Monthly Notices of the Royal Astronomical Society](#), 486, 5124–5138
- Obuljen A., Simonović M., Schneider A., Feldmann R., 2023, [Phys. Rev. D](#), 108, 083528
- Ono V., Park C. F., Mudur N., Ni Y., Cuesta-Lazaro C., Villaescusa-Navarro F., 2024, Debiasing with Diffusion: Probabilistic reconstruction of Dark Matter fields from galaxies with CAMELS, <http://arxiv.org/abs/2403.10648>
- Ore A., Heneka C., Plehn T., 2025, [SciPost Physics](#), 18
- Padmanabhan H., Refregier A., 2016, [Monthly Notices of the Royal Astronomical Society](#), 464, 4008–4017

- Padmanabhan H., Maartens R., Umeh O., Camera S., 2023, The HI intensity mapping power spectrum: insights from recent measurements, <http://arxiv.org/abs/2305.09720>
- Papamakarios G., Pavlakou T., Murray I., 2018, Masked Autoregressive Flow for Density Estimation, [doi:10.48550/arXiv.1705.07057](https://arxiv.org/abs/1705.07057), <http://arxiv.org/abs/1705.07057>
- Paul S., Santos M. G., Chen Z., Wolz L., 2023, A first detection of neutral hydrogen intensity mapping on Mpc scales at $z \approx 0.32$ and $z \approx 0.44$, [doi:10.48550/ARXIV.2301.11943](https://arxiv.org/abs/2301.11943), <https://arxiv.org/abs/2301.11943>
- Peacock J. A., 1998, Cosmological Physics. Cambridge University Press, Cambridge, [doi:10.1017/CBO9780511804533](https://www.cambridge.org/core/books/cosmological-physics/3EA0C68658270C9C3CC48E07A035C327), <https://www.cambridge.org/core/books/cosmological-physics/3EA0C68658270C9C3CC48E07A035C327>
- Perez E., Strub F., de Vries H., Dumoulin V., Courville A., 2017, FiLM: Visual Reasoning with a General Conditioning Layer ([arXiv:1709.07871](https://arxiv.org/abs/1709.07871)), <https://arxiv.org/abs/1709.07871>
- Perraudin N., Srivastava A., Lucchi A., Kacprzak T., Hofmann T., Réfrégier A., 2019, Cosmological N-body simulations: a challenge for scalable generative models ([arXiv:1908.05519](https://arxiv.org/abs/1908.05519)), <https://arxiv.org/abs/1908.05519>
- Petit O., Thome N., Rambour C., Soler L., 2021, U-Net Transformer: Self and Cross Attention for Medical Image Segmentation, <http://arxiv.org/abs/2103.06104>
- Pillepich A., et al., 2017, [Monthly Notices of the Royal Astronomical Society](#), 475, 648–675
- Planck Collaboration VI ., 2020, [A&A](#), 641, A6
- Press W. H., Schechter P., 1974, [ApJ](#), 187, 425
- Pritchard J. R., Loeb A., 2012, [Reports on Progress in Physics](#), 75, 086901
- Puchwein E., et al., 2023, [Monthly Notices of the Royal Astronomical Society](#), 519, 6162
- Rahmati A., Pawlik A. H., Raičević M., Schaye J., 2013, [Monthly Notices of the Royal Astronomical Society](#), 430, 2427–2445

- Randono A., 2025, Cloud Diffusion Part 1: Theory and Motivation ([arXiv:2507.05496](https://arxiv.org/abs/2507.05496)), <https://arxiv.org/abs/2507.05496>
- Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, *Astrophys. J.*, 876, 85
- Rigo M., Trotta R., Viel M., 2025, *Monthly Notices of the Royal Astronomical Society*, 541, 166–178
- Ronneberger O., Fischer P., Brox T., 2015, U-Net: Convolutional Networks for Biomedical Image Segmentation, <http://arxiv.org/abs/1505.04597>
- Ross A. J., Samushia L., Howlett C., Percival W. J., Burden A., Manera M., 2015, *Mon. Not. Roy. Astron. Soc.*, 449, 835
- Rouhiainen A., 2024, Cosmology at the Field Level with Probabilistic Machine Learning, <http://arxiv.org/abs/2402.07694>
- Rouhiainen A., Giri U., Münchmeyer M., 2021, Normalizing flows for random fields in cosmology ([arXiv:2105.12024](https://arxiv.org/abs/2105.12024)), <https://arxiv.org/abs/2105.12024>
- SKA Cosmology SWG ., 2020, *Publ. Astron. Soc. Austral.*, 37, e007
- Santos M. G., et al., 2015, *PoS*, AASKA14, 019
- Santos M. G., et al., 2017, in *MeerKAT Science: On the Pathway to the SKA.* ([arXiv:1709.06099](https://arxiv.org/abs/1709.06099))
- Saxena A., Meerburg P. D., Weniger C., Acedo E. d. L., Handley W., 2024, *RAS Techniques and Instruments*, 3, 724–736
- Scaramella R., et al., 2022, *Astronomy & Astrophysics*, 662, A112
- Scelfo G., Mishra S., Rigo M., Trotta R., Viel M., 2026, Field-level multi-tracers simulation-based inference of cosmological parameters from 3D maps ([arXiv:2605.26210](https://arxiv.org/abs/2605.26210)), <https://arxiv.org/abs/2605.26210>
- Schaye J., Theuns T., Rauch M., Efstathiou G., Sargent W. L., 2000, *Monthly Notices of the Royal Astronomical Society*, 318, 817
- Schaye J., et al., 2014, *Monthly Notices of the Royal Astronomical Society*, 446, 521–554

- Schneider A., Teyssier R., 2015, [JCAP](#), 2015, 049
- Sefusatti E., Crocce M., Pueblas S., Scoccimarro R., 2006, [Phys. Rev. D](#), 74, 023522
- Seo H.-J., Dodelson S., Marriner J., McGinnis D., Stebbins A., Stoughton C., Vallinotto A., 2010, [Astrophys. J.](#), 721, 164
- Shao H., Villaescusa-Navarro F., Villanueva-Domingo P., et al., 2023, [ApJ](#)
- Sharma D., Dai B., Villaescusa-Navarro F., Seljak U., 2025, [MNRAS](#), 538, 1415
- Sobal V., V J. S., Jalagam S., Carion N., Cho K., LeCun Y., 2022, Joint Embedding Predictive Architectures Focus on Slow Features ([arXiv:2211.10831](#)), <https://arxiv.org/abs/2211.10831>
- Sohl-Dickstein J., Weiss E. A., Maheswaranathan N., Ganguli S., 2015, Deep Unsupervised Learning using Nonequilibrium Thermodynamics, [doi:10.48550/arXiv.1503.03585](#), <http://arxiv.org/abs/1503.03585>
- Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., Poole B., 2021, Score-Based Generative Modeling through Stochastic Differential Equations, <http://arxiv.org/abs/2011.13456>
- Spinelli M., Zoldan A., De Lucia G., Xie L., Viel M., 2020, [Monthly Notices of the Royal Astronomical Society](#), 493, 5434
- Springel V., Hernquist L., 2003, [Monthly Notices of the Royal Astronomical Society](#), 339, 289–311
- Springel V., White S., Tormen G., Kauffmann G., 2000, Populating a cluster of galaxies - I. Results at $z=0$, [doi:10.48550/arXiv.astro-ph/0012055](#), <http://arxiv.org/abs/astro-ph/0012055>
- Springel V., Pakmor R., Zier O., Reinecke M., 2021, [Monthly Notices of the Royal Astronomical Society](#), 506, 2871
- Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012a, [Astrophys. J.](#), 761, 152
- Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012b, [The Astrophysical Journal](#), 761, 152

- Tepper-García T., 2006, *MNRAS*, 369, 2025
- Tucci B., Schmidt F., 2024, *JCAP*, 2024, 063
- Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I., 2023, Attention Is All You Need, [doi:10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762), <http://arxiv.org/abs/1706.03762>
- Verde L., Treu T., Riess A. G., 2019, *Nature Astron.*, 3, 891
- Viel M., Matarrese S., Mo H. J., Theuns T., Haehnelt M. G., 2002, *Monthly Notices of the Royal Astronomical Society*, 336, 685–698
- Villaescusa-Navarro F., Viel M., Datta K. K., Choudhury T. R., 2014, *JCAP*, 2014, 050
- Villaescusa-Navarro F., Bull P., Viel M., 2015, *ApJ*, 814, 146
- Villaescusa-Navarro F., Alonso D., Viel M., 2017, *Mon. Not. Roy. Astron. Soc.*, 466, 2736
- Villaescusa-Navarro F., et al., 2018a, *The Astrophysical Journal*, 866, 135
- Villaescusa-Navarro F., et al., 2018b, *Astrophys. J.*, 866, 135
- Villaescusa-Navarro F., et al., 2021a, Multifield Cosmology with Artificial Intelligence ([arXiv:2109.09747](https://arxiv.org/abs/2109.09747)), <https://arxiv.org/abs/2109.09747>
- Villaescusa-Navarro F., et al., 2021b, *The Astrophysical Journal*, 915, 71
- Villaescusa-Navarro F., et al., 2022, *The Astrophysical Journal Supplement Series*, 259, 61
- Villanueva-Domingo P., Villaescusa-Navarro F., 2022, *The Astrophysical Journal*, 937, 115
- Wadekar D., Villaescusa-Navarro F., Ho S., Perreault-Levasseur L., 2021, *The Astrophysical Journal*, 916, 42
- Wang J., et al., 2021, *Mon. Not. Roy. Astron. Soc.*, 505, 3698
- Weinberger R., Springel V., Pakmor R., 2020, *The Astrophysical Journal Supplement Series*, 248, 32

- Weyant A., Schafer C., Wood-Vasey W. M., 2013, [The Astrophysical Journal](#), 764, 116
- Williger G. M., Smette A., Hazard C., Baldwin J. A., McMahon R. G., 2000, [The Astrophysical Journal](#), 532, 77
- Wolz L., et al., 2022, [MNRAS](#), 510, 3495
- Wong K. C., et al., 2020, [Mon. Not. Roy. Astron. Soc.](#), 498, 1420
- Zeghal J., Lanzieri D., Lanusse F., Boucaud A., Louppe G., Aubourg E., Bayer A. E., 2025, [Astronomy & Astrophysics](#), 699, A327
- Zhao B., et al., 2026, arXiv preprint
- van Daalen M. P., McCarthy I. G., Schaye J., 2020, [MNRAS](#), 491, 2424