```
\documentclass{article}
\usepackage{graphicx} % Required for inserting images

\title{18 Aug}
\author{collin5229 }
\date{August 2025}

\begin{document}

\maketitle

\section{Introduction}
```

Diffusion Models for Emulating Neutral Hydrogen Maps: A Novel Approach to 21cm Cosmology

1. Introduction

The detection and analysis of neutral hydrogen (HI) through its characteristic 21cm emission line represents one of the most promising avenues for understanding cosmic evolution during the Epoch of Reionisation and beyond. However, the computational demands of high-resolution cosmological simulations and observational limitations pose significant challenges for comprehensive parameter estimation and model validation. This dissertation explores the application of diffusion models—a class of generative machine learning architectures—to emulate HI intensity maps, offering a computationally efficient pathway for parameter recovery and cosmological inference.

The fundamental premise underlying this research stems from recognising that traditional numerical simulations, while physically accurate, require substantial computational resources that limit their practical application in parameter space exploration. Diffusion models, originally developed for image generation tasks, demonstrate remarkable capability in capturing complex spatial correlations and statistical properties inherent in cosmological fields. By training these models on existing simulation datasets, we can generate synthetic HI maps that preserve essential statistical characteristics while dramatically reducing computational overhead.

This approach presents several methodological advantages over conventional techniques. Unlike traditional emulators that rely on simplified analytical approximations, diffusion models can capture non-linear relationships and higher-order correlations present in the underlying cosmological processes. Furthermore, the probabilistic nature of these models enables uncertainty quantification and ensemble generation, providing robust statistical frameworks for parameter estimation.

2. Background on Neutral Hydrogen Maps

Neutral hydrogen represents the most abundant element in the universe, serving as a fundamental tracer of cosmic structure formation and evolution. The 21cm hyperfine transition of neutral hydrogen provides a unique observational window into epochs ranging from the Dark Ages through reionisation to the present day. This emission mechanism arises from the spin-flip transition between the parallel and antiparallel spin states of the electron-proton system, producing photons with a rest frequency of 1420.4 MHz.

The intensity mapping technique exploits the collective 21cm emission from large volumes of neutral hydrogen, mapping the three-dimensional distribution of matter without requiring individual source detection. This methodology offers several observational advantages, including the ability to probe large cosmological volumes simultaneously and access redshift information directly through frequency measurement. The brightness temperature of 21cm emission depends on both the neutral hydrogen density and the local spin temperature, creating a complex relationship with underlying cosmological parameters.

Current theoretical frameworks predict that HI intensity maps contain rich cosmological information encoded in their statistical properties. The power spectrum of 21cm fluctuations carries signatures of primordial density perturbations, baryon acoustic oscillations, and redshift-space distortions. These features enable constraints on fundamental cosmological parameters, including the matter density, dark energy equation of state, and primordial non-Gaussianity.

However, the interpretation of 21cm observations requires sophisticated modelling of astrophysical processes including star formation, feedback mechanisms, and reionisation topology. These complex interdependencies necessitate detailed numerical simulations that can accurately capture the relevant physics across multiple scales, from individual galaxies to cosmological volumes.

3. Limitations in Detecting HI Maps

Contemporary efforts to detect and characterise neutral hydrogen through 21cm observations face numerous technical and systematic challenges that significantly impact data quality and scientific interpretation. Foreground contamination represents the most substantial observational obstacle, with galactic synchrotron emission, free-free radiation, and extragalactic point sources producing signals orders of magnitude stronger than the cosmological 21cm signal.

Radio frequency interference (RFI) from terrestrial and satellite sources creates additional complications, particularly in the frequency bands corresponding to cosmologically interesting redshifts. The dynamic nature of RFI and its potential temporal correlation with observational schedules demands sophisticated mitigation strategies that may introduce systematic biases in the recovered signal.

Instrumental systematics pose equally significant challenges, including frequency-dependent beam patterns, calibration uncertainties, and bandpass instabilities. These effects can mimic cosmological signals, creating spurious correlations that contaminate scientific measurements. The polarisation leakage from intense foreground sources represents a particularly insidious systematic effect, as it can produce frequency-dependent structures that resemble cosmological signals.

The ionospheric effects introduce additional complexity, particularly for low-frequency observations targeting high-redshift epochs. Ionospheric refraction and phase delays vary temporally and spatially, creating systematic errors that correlate with source positions and observational conditions. These atmospheric effects become increasingly problematic for precision cosmology applications requiring percent-level accuracy in power spectrum measurements.

Furthermore, the thermal noise limitations of current radio telescopes restrict the sensitivity achievable within reasonable observational timeframes. The inherent trade-off between angular resolution, frequency resolution, and sensitivity constrains the accessible parameter space and limits the precision of cosmological parameter recovery.

4. Machine Learning Applications in 21cm Cosmology

The application of machine learning techniques to 21cm cosmology has emerged as a transformative approach for addressing computational challenges and extracting maximal scientific information from observational data. Traditional analysis methods, while theoretically wellmotivated, often struggle with the high-dimensional nature of cosmological datasets and the complex non-linear relationships between observables and underlying parameters.

Neural network architectures have demonstrated particular efficacy in cosmological parameter estimation tasks, leveraging their capacity to identify subtle patterns in high-dimensional data spaces. Convolutional neural networks excel at recognising spatial features in intensity maps, while recurrent architectures can capture temporal correlations in time-series data. These approaches have successfully extracted cosmological parameters from simulated datasets with precision exceeding traditional methods.

Deep learning techniques have also revolutionised foreground removal strategies, with algorithms capable of separating cosmological signals from contaminating emissions based on their distinct spectral and spatial characteristics. These methods exploit the inherent differences between smooth foreground spectra and the structured frequency dependence of cosmological signals, achieving separation levels previously unattainable through conventional techniques.

Machine learning approaches have proven particularly valuable for handling systematic uncertainties and calibration challenges. Algorithms can learn complex mappings between instrumental responses and true sky signals, enabling more accurate calibration procedures and systematic error mitigation. These techniques demonstrate remarkable adaptability to varying observational conditions and instrumental configurations.

The computational efficiency gains offered by machine learning methods enable extensive parameter space exploration and uncertainty quantification studies that would be prohibitively expensive using traditional simulation-based approaches. This capability opens new possibilities for robust cosmological inference and model comparison studies.

5. Discriminative Models in Cosmological Analysis

Discriminative models represent a class of machine learning architectures designed to learn the conditional probability distribution $P(y|x)$, mapping input data to specific output categories or parameter values. In cosmological contexts, these models excel at parameter estimation tasks, where the objective involves inferring cosmological parameters from observed or simulated intensity maps.

The fundamental strength of discriminative approaches lies in their direct optimisation for the specific predictive task, enabling efficient parameter recovery without requiring explicit modelling of the full data distribution. Supervised learning frameworks train these models on labeled datasets, where input intensity maps are paired with known cosmological parameters derived from the simulation specifications.

Convolutional neural networks represent the most widely adopted discriminative architecture for cosmological applications, exploiting the spatial structure inherent in intensity maps through hierarchical feature extraction. These models learn increasingly complex spatial patterns through multiple convolutional layers, culminating in dense layers that map spatial features to parameter estimates.

Recent advances in attention mechanisms have enhanced the capability of discriminative models to identify relevant spatial regions and scales for parameter estimation. Transformer architectures, originally developed for natural language processing, demonstrate remarkable performance in identifying long-range correlations and multi-scale features in cosmological data.

However, discriminative models face inherent limitations in uncertainty quantification and robustness assessment. These architectures provide point estimates without comprehensive uncertainty characterisation, potentially masking systematic biases or overfitting issues. Furthermore, discriminative models require extensive training datasets spanning the relevant parameter space, creating computational demands that may limit their practical applicability.

6. Power Spectrum Analysis for Parameter Recovery

The power spectrum represents a fundamental statistical descriptor of cosmological fields, encoding information about the underlying matter distribution and cosmological parameters through its amplitude, shape, and scale dependence. For 21cm intensity mapping, the power spectrum provides a natural bridge between theoretical predictions and observational data, enabling quantitative comparison between models and measurements.

The monopole power spectrum captures isotropic clustering information, while higher-order multipoles encode anisotropic features arising from redshift-space distortions and observational effects. The quadrupole and hexadecapole components contain valuable information about the growth rate of structure and the Alcock-Paczynski effect, enabling constraints on dark energy properties and gravitational physics.

Machine learning approaches to power spectrum analysis offer several advantages over traditional methods. Neural networks can learn complex mappings between power spectrum features and cosmological parameters, capturing non-linear relationships that analytical models may miss. These approaches can simultaneously analyse multiple multipole moments and frequency bins, extracting maximal information from the available data.

The integration of power spectrum analysis with generative models presents novel opportunities for parameter recovery and uncertainty quantification. By generating ensembles of synthetic maps consistent with observed power spectra, these approaches enable comprehensive exploration of parameter degeneracies and systematic uncertainties.

However, power spectrum-based methods face limitations when dealing with non-Gaussian features and higher-order correlations present in cosmological fields. The compression of spatial information into statistical summaries may discard valuable information contained in the full field distribution, motivating the development of alternative analysis approaches.

7. Large-Scale Field Reconstruction and Analysis

Large-scale cosmological fields contain hierarchical structure spanning multiple decades in scale, from individual galaxies to the cosmic web's largest filaments and voids. Understanding these multi-scale correlations requires sophisticated analysis techniques capable of capturing both local features and global statistical properties.

Wavelet-based decomposition methods have proven particularly effective for multi-scale analysis, enabling the separation of cosmological signals across different spatial frequencies. These techniques exploit the localised nature of wavelet functions to identify scale-dependent features while preserving spatial information, offering advantages over traditional Fourier methods for non-stationary fields.

Machine learning approaches to large-scale field analysis leverage the capacity of neural networks to identify complex spatial patterns and correlations. Graph neural networks show particular promise for cosmic web analysis, treating galaxies and halos as nodes in a cosmic network and learning the relationships between local properties and large-scale environment.

The development of field-level inference techniques represents a significant advance in cosmological analysis, enabling parameter recovery from complete spatial information rather than compressed statistics. These methods maintain the full information content of cosmological fields while providing robust parameter estimation frameworks.

Generative models offer unique capabilities for large-scale field reconstruction, enabling the synthesis of complete cosmological volumes conditioned on partial observational data. These approaches can fill in missing information due to observational limitations while preserving the statistical properties of the underlying cosmological fields.

8. Generative Models in Cosmological Applications

Generative models represent a paradigm shift in cosmological simulation and analysis, offering the capability to synthesise realistic cosmological fields without explicit numerical integration of the underlying differential equations. These models learn the statistical properties of cosmological datasets and generate new realisations that preserve essential features while dramatically reducing computational requirements.

Generative Adversarial Networks (GANs) have demonstrated remarkable success in cosmological applications, learning to generate synthetic dark matter distributions, galaxy catalogues, and intensity maps that are statistically indistinguishable from simulation data. The adversarial training framework ensures that generated samples capture both global statistical properties and local structural features.

Variational Autoencoders (VAEs) offer alternative generative frameworks with explicit probabilistic formulations, enabling uncertainty quantification and latent space exploration. These models learn compressed representations of cosmological fields, facilitating parameter space exploration and interpolation between different cosmological models.

Diffusion models represent the most recent advancement in generative architectures, demonstrating superior performance in image synthesis tasks through iterative denoising processes. These models offer several advantages over GAN-based approaches, including training stability, mode coverage, and explicit likelihood computation capabilities.

The application of generative models to cosmological parameter estimation presents novel opportunities for inverse problems and uncertainty quantification. By learning the mapping between parameters and observational data, these models enable direct sampling from posterior distributions and comprehensive uncertainty characterisation.

9. Aims and Objectives

This research aims to develop and validate a novel framework for cosmological parameter estimation using diffusion models trained on neutral hydrogen intensity maps. The primary objective involves demonstrating that diffusion models can generate synthetic HI maps that preserve the statistical properties necessary for accurate parameter recovery while providing substantial computational advantages over traditional simulation methods.

Primary Objectives:

- Develop a diffusion model architecture optimised for cosmological intensity map generation, incorporating relevant physical constraints and boundary conditions
- Train the model on comprehensive datasets spanning realistic parameter ranges for contemporary cosmological models
- Validate the generated maps through comparison with power spectrum statistics, bispectrum analysis, and morphological descriptors
- Demonstrate parameter recovery capabilities using both discriminative and likelihood-based inference approaches
- Quantify computational efficiency gains compared to traditional numerical simulation methods

Secondary Objectives:

- Investigate the impact of training dataset size and parameter space coverage on model performance and generalisation capabilities
- Explore conditioning mechanisms for targeted parameter space exploration and interpolation
- Develop uncertainty quantification frameworks for parameter estimation using ensemble generation techniques
- Assess the robustness of the approach to systematic effects and observational limitations
- Compare performance against existing emulation techniques and traditional analysis methods

Methodological Innovations:

The research will explore several novel methodological approaches, including the integration of physical constraints into the diffusion process through custom loss functions, the development of hierarchical models for multi-scale feature preservation, and the implementation of conditional generation frameworks for targeted parameter exploration.

10. Overview and Structure

This dissertation presents a comprehensive investigation into the application of diffusion models for neutral hydrogen map emulation and cosmological parameter estimation. The research methodology combines theoretical development, numerical implementation, and empirical validation to establish a robust framework for next-generation cosmological analysis.

Chapter Organisation:

Chapter 2 provides detailed theoretical background on diffusion models, covering the mathematical foundations, training procedures, and relevant architectural considerations for cosmological applications. This chapter establishes the theoretical framework underlying the proposed methodology.

Chapter 3 describes the simulation datasets and data preprocessing procedures used for model training and validation. This includes discussion of cosmological parameter ranges, simulation specifications, and data augmentation strategies employed to enhance model robustness.

Chapter 4 presents the model architecture and training methodology, including network design choices, loss function formulation, and optimisation procedures. Particular attention is given to incorporating physical constraints and ensuring stable training dynamics.

Chapter 5 contains the core results, demonstrating model performance through comprehensive validation studies including power spectrum preservation, parameter recovery accuracy, and computational efficiency analysis. Comparison studies against existing methods provide context for the achieved performance levels.

Chapter 6 discusses the implications of the results for cosmological parameter estimation and 21cm cosmology more broadly. This includes analysis of current limitations, potential systematic effects, and future development directions.

The research represents a significant contribution to computational cosmology, offering a novel approach that combines cutting-edge machine learning techniques with fundamental cosmological theory. The proposed methodology has potential applications beyond 21cm cosmology, providing a general framework for cosmological field emulation and parameter estimation across multiple observational probes.

CHAPTER 2

\section{Chapter 2: Fundamentals of Machine Learning}

\subsection{2.1 Introduction to Machine Learning}

The convergence of computational power and statistical methodologies has fundamentally transformed astronomical research paradigms. Machine learning, as a computational framework, provides systematic approaches for extracting meaningful patterns from complex observational datasets. Within the context of astrophysical research, these methodologies enable sophisticated analysis of phenomena that traditional analytical approaches cannot adequately address.

Contemporary astronomical observations generate unprecedented volumes of data, necessitating computational frameworks capable of identifying subtle correlations and predictive relationships. The application of machine learning techniques to astronomical datasets represents a paradigm shift from hypothesis-driven analysis toward data-driven discovery methodologies.

\subsection{2.2 Machine Learning Methods: Theoretical Foundations}

\subsubsection{2.2.1 Supervised Learning Frameworks}

Supervised learning methodologies operate through the systematic analysis of input-output relationships within labeled datasets. These approaches enable the development of predictive models capable of inferring complex mappings between observational parameters and physical phenomena. Within astronomical contexts, supervised learning facilitates the classification of celestial objects and the prediction of evolutionary trajectories based on empirical observations.

\subsubsection{2.2.2 Unsupervised Learning Paradigms}

Unsupervised learning techniques focus on identifying latent structures within unlabeled datasets. These methodologies prove particularly valuable for exploratory analysis of astronomical phenomena, enabling the discovery of previously unrecognized correlations and structural patterns. Clustering algorithms and dimensionality reduction techniques exemplify unsupervised approaches that reveal hidden organizational principles within complex datasets.

\subsection{2.3 Deep Learning: Architectural Considerations}

\subsubsection{2.3.1 Neural Network Fundamentals}

Neural networks constitute computational frameworks inspired by biological information processing mechanisms. These systems utilize interconnected processing units to approximate complex functional relationships through iterative parameter optimization. The fundamental architecture consists of layers of artificial neurons that transform input signals through weighted connections and nonlinear activation functions.

The mathematical foundation of neural networks relies on the universal approximation theorem, which establishes that sufficiently complex networks can approximate any continuous function to arbitrary precision. This theoretical foundation provides the basis for the remarkable flexibility observed in deep learning applications across diverse scientific domains.

\subsubsection{2.3.2 Activation Functions: Nonlinear Transformation Mechanisms}

Activation functions introduce essential nonlinearity into neural network architectures, enabling the approximation of complex, non-monotonic relationships. The selection of appropriate activation functions significantly influences network performance and convergence characteristics.

The Rectified Linear Unit (ReLU) activation function has emerged as a particularly effective choice due to its computational efficiency and favorable gradient propagation properties. Alternative activation functions, including the Exponential Linear Unit (ELU) and Swish functions, offer specialized advantages for specific applications requiring smoother gradient behavior or enhanced robustness to input variations.

\subsubsection{2.3.3 Loss Function Formulations}

Loss functions quantify the discrepancy between predicted and observed outcomes, providing optimization targets for neural network training processes. The selection of appropriate loss functions depends critically on the specific characteristics of the target domain and the desired model behavior.

\paragraph{Regression Loss Functions}

Mean Squared Error (MSE) loss functions provide intuitive measures of prediction accuracy for continuous target variables. These functions exhibit favorable mathematical properties, including differentiability and convexity, which facilitate efficient optimization processes. However, MSE loss functions demonstrate sensitivity to outliers, potentially compromising model robustness in the presence of anomalous observations.

Alternative regression loss formulations, such as Mean Absolute Error (MAE) and Huber loss, offer enhanced robustness characteristics while maintaining computational tractability. The Huber loss function, in particular, combines the computational advantages of MSE for small errors with the robustness of MAE for large deviations.

\paragraph{Classification Loss Functions}

Cross-entropy loss functions provide theoretically grounded approaches for multi-class classification problems. These functions maximize the likelihood of correct predictions while penalizing confident incorrect classifications, resulting in well-calibrated probability estimates.

Focal loss represents an innovative extension of cross-entropy loss designed to address class imbalance challenges prevalent in astronomical datasets. This formulation dynamically adjusts the loss contribution of well-classified examples, focusing optimization efforts on difficult-to-classify instances.

\subsubsection{2.3.4 Residual Block Architectures}

Residual connections address the degradation problem observed in deep neural networks, enabling the training of significantly deeper architectures without performance deterioration. These connections facilitate direct information flow between non-adjacent layers, preserving gradient information throughout the optimization process.

The mathematical formulation of residual blocks introduces skip connections that add the input directly to the output of transformation layers. This architectural innovation enables the learning of residual mappings rather than complete transformations, simplifying the optimization landscape and improving convergence characteristics.

\subsubsection{2.3.5 Optimization Strategies}

\paragraph{Gradient-Based Optimization}

Stochastic Gradient Descent (SGD) provides the fundamental optimization framework for neural network training. Adaptive optimization algorithms, including Adam and AdamW, incorporate momentum and adaptive learning rate mechanisms to enhance convergence stability and efficiency.

The Adam optimizer combines the advantages of adaptive gradient algorithms with momentum-based approaches, maintaining separate learning rates for individual parameters while incorporating exponentially decaying averages of past gradients and squared gradients.

\paragraph{Overfitting and Underfitting Phenomena}

Overfitting occurs when models exhibit excessive complexity relative to the available training data, resulting in poor generalization to unseen observations. Regularization techniques, including dropout, weight decay, and early stopping, provide systematic approaches for controlling model complexity and improving generalization performance.

Underfitting represents the complementary problem of insufficient model complexity, resulting in inadequate representation of underlying patterns. The identification of optimal model complexity requires careful consideration of the bias-variance tradeoff and systematic evaluation using validation datasets.

\subsection{2.4 Convolutional Neural Networks: Spatial Pattern Recognition}

Convolutional neural networks represent specialized architectures designed for processing spatially structured data. These networks utilize convolution operations to identify local patterns while maintaining spatial relationships through hierarchical feature extraction processes.

The convolution operation applies learned filters to input data, producing feature maps that highlight relevant spatial patterns. Pooling operations provide translation invariance and dimensionality reduction, enabling efficient processing of high-resolution inputs while preserving essential structural information.

Within astronomical contexts, convolutional networks demonstrate particular effectiveness for image analysis tasks, including object detection, morphological classification, and the identification of transient phenomena in observational datasets.

\subsection{2.5 Generative Models: Probabilistic Data Synthesis}

\subsubsection{2.5.1 Autoencoder Architectures}

Autoencoders constitute unsupervised learning frameworks designed to learn efficient data representations through reconstruction objectives. These architectures consist of encoder networks that compress input data into compact latent representations, followed by decoder networks that reconstruct the original input from the compressed representation.

Variational autoencoders (VAEs) extend the basic autoencoder framework by incorporating probabilistic latent variable models. These approaches enable principled generation of new samples by sampling from learned latent distributions, providing valuable tools for data augmentation and uncertainty quantification.

\subsubsection{2.5.2 U-Net Architectures for Spatial Reconstruction}

U-Net architectures represent specialized convolutional networks designed for dense prediction tasks requiring precise spatial localization. These networks utilize skip connections between encoder and decoder pathways to preserve fine-grained spatial information throughout the reconstruction process.

The symmetric encoder-decoder structure enables efficient processing of high-resolution inputs while maintaining computational tractability. Skip connections facilitate the integration of multi-scale features, resulting in accurate reconstructions that preserve both global structure and local detail.

\subsubsection{2.5.3 Diffusion Models: Probabilistic Generation Frameworks}

Diffusion models represent a novel class of generative models based on the systematic corruption and reconstruction of data through learned denoising processes. These approaches model the generation process as a reverse diffusion procedure, gradually transforming random noise into structured outputs through iterative refinement steps.

The theoretical foundation of diffusion models relies on stochastic differential equations that describe the forward corruption process and the corresponding reverse generation procedure. Training involves learning denoising networks capable of predicting the noise component at each corruption level, enabling high-quality sample generation through sequential denoising operations.

Recent advances in diffusion model architectures demonstrate remarkable capabilities for generating high-fidelity samples across diverse domains. The application of these methodologies to astronomical data synthesis presents opportunities for addressing data scarcity challenges and enabling detailed analysis of rare phenomena through synthetic data generation.

Within the context of HI mapping applications, diffusion models offer promising approaches for generating realistic neutral hydrogen distributions while preserving the complex spatial correlations observed in astronomical surveys. The stochastic nature of the generation process enables the exploration of plausible alternative configurations, providing valuable insights into the range of possible evolutionary outcomes and observational scenarios.

\subsection{2.6 Research Context and Methodological Foundations}

The integration of machine learning methodologies into astrophysical research requires careful consideration of domain-specific constraints and observational limitations. The unique characteristics of astronomical datasets, including sparse sampling, heteroscedastic noise, and complex selection effects, necessitate specialized approaches that account for these observational realities.

The development of machine learning frameworks for HI mapping applications represents a convergence of computational innovation and astrophysical insight. These methodologies enable the exploration of previously inaccessible parameter spaces while providing systematic approaches for quantifying uncertainties and validating theoretical predictions against empirical observations.
\\

\section{Chapter 3: Accelerating Cosmological Astrophysics with Machine Learning Simulations (CAMELS)}

\subsection{Introduction}

The computational demands of modern cosmological simulations present increasingly formidable challenges to astrophysical research. Traditional N-body and hydrodynamic simulations, while providing unprecedented detail in cosmic structure formation, require substantial computational resources that often limit parameter space exploration and statistical analysis. The Cosmological and Astrophysical Machine Learning Simulations (CAMELS) framework emerges as a transformative approach, leveraging machine learning architectures to accelerate simulation workflows while maintaining scientific fidelity.

This chapter explores the application of diffusion models as generative frameworks for emulating neutral hydrogen (HI) maps within the CAMELS paradigm. The methodology represents a departure from conventional interpolation techniques, instead employing probabilistic generative processes that capture the intrinsic stochasticity of astrophysical phenomena. Through careful analysis of the underlying physics and implementation of sophisticated neural architectures, we demonstrate how machine learning can fundamentally reshape computational astrophysics.

\subsection{Background on CAMELS}

The CAMELS project represents a systematic effort to bridge the gap between traditional cosmological simulations and modern machine learning methodologies. Initiated to address the computational bottlenecks inherent in parameter space exploration, CAMELS provides a comprehensive suite of hydrodynamic and N-body simulations spanning diverse cosmological and astrophysical parameters.

The framework encompasses simulations conducted with multiple codes, including AREPO and GADGET, ensuring robustness across different numerical implementations. Each simulation captures the evolution of cosmic structures from redshift $z \approx 15$ to the present epoch, incorporating critical physical processes including star formation, stellar feedback, and supermassive black hole growth. The parameter space exploration covers variations in:

\begin{itemize}
    \item Cosmological parameters: Ω_m (matter density), σ_8 (matter fluctuation amplitude), h (Hubble constant)

\item Astrophysical parameters: Stellar feedback efficiency (A_SN1, A_SN2), black hole feedback parameters (A_AGN1, A_AGN2)
\end{itemize}
The mathematical foundation underlying CAMELS simulations relies on the gravitational N-body problem coupled with hydrodynamic equations:

CHAPTER 4

\section{Chapter 4: Methodology - Deep Generative Models for Neutral Hydrogen Emulation}

\subsection{4.1 Introduction to Diffusion-Based Generative Modeling}

The computational modeling of neutral hydrogen (HI) maps represents a fundamental challenge in modern astrophysical research, where traditional N-body simulations demand substantial computational resources while often failing to capture the full complexity of baryon physics at galactic scales. This investigation employs diffusion probabilistic models as a novel approach to emulate HI distribution patterns, leveraging recent advances in generative machine learning to address computational limitations inherent in conventional simulation methodologies.

The methodology presented in this chapter establishes a framework for learning complex probability distributions underlying HI spatial configurations through a process of iterative noise addition and subsequent denoising. Unlike deterministic simulation approaches, this probabilistic framework captures the inherent stochasticity observed in real astronomical data while maintaining computational efficiency suitable for large-scale cosmological studies.

\subsection{4.2 Theoretical Framework: Diffusion Process Formulation}

The diffusion model implementation follows the mathematical formulation established by Ho et al. (2020), adapted specifically for astrophysical data characteristics. The forward diffusion process systematically corrupts clean HI maps through controlled Gaussian noise addition across T timesteps, mathematically expressed as:

\$$\textbackslash{}frac\{d\textbackslash{}mathbf\{v\}\textit{i\}{dt\} = -\textbackslash{}frac\{GM\}{r\^2\}\hat\{\mathbf\{r\}} + \textbackslash{}mathbf\{a\}}\{hydro\} + \textbackslash{}mathbf\{a\}_\{feedback\}$\$

where \$\mathbf\{a\}\textit{\{hydro\}$ represents hydrodynamic accelerations and \$\mathbf\{a\}}\{feedback\}$ incorporates stellar and AGN feedback mechanisms.

\subsection{HI Map Simulations}

Neutral hydrogen observations provide crucial insights into the cosmic web's structure and galaxy formation processes. The 21-cm line emission, corresponding to the hyperfine transition in hydrogen atoms, offers a unique probe of both local and high-redshift universe properties. However, generating synthetic HI maps from cosmological simulations presents significant computational challenges.

The relationship between simulated gas properties and observable HI emission involves complex radiative transfer calculations. The HI fraction in cosmic gas depends on the balance between photoionization and recombination processes:

\$$\textbackslash{}frac\{dn_{HI\}}\{dt\} = \textbackslash{}alpha_{rec\} n_p n_e - \textbackslash{}Gamma_{ion\} n_{HI\}$\$

where \$\alpha_{rec\}$ represents the recombination coefficient, \$\Gamma_{ion\}$ denotes the photoionization rate, and \$n_p\$, \$n_e\$, \$n_{HI\}$ are proton, electron, and neutral hydrogen number densities respectively.

Traditional approaches require post-processing of simulation snapshots through radiative transfer codes, consuming substantial computational resources and limiting the scope of parameter studies. The brightness temperature of 21-cm emission can be expressed as:

\$$T_b(\textbackslash{}mathbf\{r\}) = T_s \textbackslash{}left(1 - e\^{-\textbackslash{}tau_{21\}(\textbackslash{}mathbf\{r\})\}\right) \textbackslash{}approx T_s \textbackslash{}tau_{21\}(\textbackslash{}mathbf\{r\})\$$

where \$T_s\$ is the spin temperature and \$\tau_{21\}$ represents the optical depth of the 21-cm transition.

\subsection{Illustrating Simulations}

Contemporary astrophysical simulations capture phenomena across vast dynamical ranges, from individual star formation regions to cosmic web filaments spanning hundreds of megaparsecs. The multi-scale nature of these processes necessitates sophisticated numerical techniques and presents unique visualization challenges.

Modern simulation frameworks employ adaptive mesh refinement and smoothed particle hydrodynamics to resolve structures from kiloparsec to gigaparsec scales. The computational complexity scales approximately as O(N log N) for N-body calculations, with hydrodynamic components introducing additional computational overhead. Memory requirements typically scale linearly with particle number, often demanding high-performance computing resources.

The scientific value of simulations extends beyond individual realizations to encompass statistical ensemble properties. Parameter space exploration requires hundreds to thousands of simulation runs, creating an urgent need for acceleration techniques that maintain physical accuracy while reducing computational burden.

\subsection{Accelerating CAMELS with Machine Learning}

Machine learning applications in astrophysics have evolved from simple regression tasks to sophisticated generative modeling approaches. The integration of neural networks with cosmological simulations offers multiple pathways for computational acceleration, including surrogate modeling, dimensionality reduction, and direct emulation of complex physical processes.

The fundamental challenge lies in preserving the multi-scale correlations present in astrophysical data while achieving meaningful computational speedups. Traditional interpolation methods often fail to capture the non-linear relationships between input parameters and output fields, particularly in regimes where feedback processes dominate structure formation.

Neural network architectures provide natural frameworks for learning these complex mappings. The universal approximation theorem suggests that sufficiently wide neural networks can approximate arbitrary continuous functions:

\$$f(\textbackslash{}mathbf\{x\}) \textbackslash{}approx \textbackslash{}sum\_{i=1\}^\{N\} w\_i \textbackslash{}sigma(\textbackslash{}mathbf\{W\}_i \textbackslash{}cdot \textbackslash{}mathbf\{x\} + b\_i)\$$

where \$\sigma\$ represents the activation function, \$\mathbf\{W\}_i\$ are weight matrices, and \$b_i\$ denote bias terms.

However, astrophysical applications demand architectures that respect the underlying physics and maintain causality constraints. This requirement motivates the exploration of specialized neural network designs, particularly those incorporating translational invariance and hierarchical feature representation.

\subsection{Application of Diffusion Models}

Diffusion models represent a paradigm shift in generative modeling, offering superior sample quality and training stability compared to traditional generative adversarial networks. The core innovation lies in the gradual corruption and reconstruction of data through a forward diffusion process and learned reverse denoising.

The forward diffusion process can be mathematized as a Markov chain that gradually adds Gaussian noise to data:

\$$q(\textbackslash{}mathbf\{x\}\textit{t | \textbackslash{}mathbf\{x\}}\{t-1\}) = \textbackslash{}mathcal\{N\}(\textbackslash{}mathbf\{x\}\textit{t; \textbackslash{}sqrt\{1-\textbackslash{}beta\_t\}\mathbf\{x\}}\{t-1\}, \textbackslash{}beta\_t \textbackslash{}mathbf\{I\})\$$

where \$\beta_t\$ represents a variance schedule controlling noise addition rates. The reverse process involves learning the conditional distribution:

\$$p\_\theta(\textbackslash{}mathbf\{x\}\textit{\{t-1\} | \textbackslash{}mathbf\{x\}\textit{t) = \textbackslash{}mathcal\{N\}(\textbackslash{}mathbf\{x\}}\{t-1\}; \textbackslash{}boldsymbol\{\mu\}}\textbackslash{}theta(\textbackslash{}mathbf\{x\}\textit{t, t), \textbackslash{}boldsymbol\{\Sigma\}}\textbackslash{}theta(\textbackslash{}mathbf\{x\}_t, t))\$$

For astrophysical applications, this framework offers compelling advantages. The probabilistic nature naturally captures the uncertainty inherent in cosmic structure formation, while the gradual denoising process allows for fine-grained control over generated features. Unlike deterministic surrogate models, diffusion approaches can generate multiple plausible realizations for identical input parameters, reflecting the stochastic nature of astrophysical processes.

The application to HI map generation introduces domain-specific considerations. Astronomical images exhibit characteristic statistical properties, including power-law power spectra and non-Gaussian probability distributions. The diffusion framework must preserve these features while enabling conditional generation based on cosmological and astrophysical parameters.

\subsection{U-Net Model Architecture}

The U-Net architecture provides the foundational framework for implementing diffusion models in astrophysical contexts. Originally developed for medical image segmentation, the U-Net design incorporates skip connections that preserve multi-scale information throughout the encoding-decoding process.

\subsubsection{Residual Blocks}

Residual connections address the vanishing gradient problem that plagued early deep learning architectures. The mathematical formulation introduces identity mappings:

\$$\textbackslash{}mathbf\{y\} = \textbackslash{}mathcal\{F\}(\textbackslash{}mathbf\{x\}, \{W_i\}) + \textbackslash{}mathbf\{x\}$\$

where \$\mathcal\{F\}$ represents the learned residual function and \$\mathbf\{x\}$ denotes the identity mapping. This design enables training of substantially deeper networks while maintaining gradient flow throughout the architecture.

In astrophysical applications, residual blocks serve an additional purpose by preserving large-scale structure information that might otherwise be lost during the encoding process. The multiplicative nature of cosmic structure formation creates features spanning multiple orders of magnitude in spatial scale, demanding architectural components that maintain this hierarchical information.

\subsubsection{Downsampling Operations}

Downsampling layers reduce spatial resolution while increasing feature channel dimensions, enabling the network to capture increasingly abstract representations. Common implementations include strided convolutions and pooling operations:

\$$\textbackslash{}mathbf\{y\}\textit{\{i,j\} = \textbackslash{}max}\{(m,n) \textbackslash{}in R_{i,j\}} \textbackslash{}mathbf\{x\}_\{m,n\}$\$

for max pooling, where \$R_{i,j\}$ represents the pooling region. Alternative approaches employ learnable downsampling through convolutional layers with stride > 1:

\$$\textbackslash{}mathbf\{y\} = \textbackslash{}sigma(\textbackslash{}mathbf\{W\} * \textbackslash{}mathbf\{x\} + \textbackslash{}mathbf\{b\})\$$

where \$*\$ denotes the convolution operation with specified stride.

The choice of downsampling strategy significantly impacts the preservation of astrophysical features. Maximum pooling tends to preserve high-intensity regions, potentially emphasizing galaxy clusters and filaments. Average pooling provides smoother feature transitions but may attenuate important substructure information.

\subsubsection{Upsampling Operations}

Upsampling operations restore spatial resolution while reducing feature dimensions, enabling the generation of high-resolution outputs from compressed latent representations. Transposed convolutions provide learnable upsampling:

\$$\textbackslash{}mathbf\{y\}\textit{\{i,j\} = \textbackslash{}sum}\{m,n\} \textbackslash{}mathbf\{W\}\textit{\{m,n\} \textbackslash{}mathbf\{x\}}\{\lfloor i/s \textbackslash{}rfloor + m, \textbackslash{}lfloor j/s \textbackslash{}rfloor + n\}$\$

where \$s\$ represents the upsampling stride. This approach allows the network to learn optimal interpolation strategies tailored to the specific characteristics of astrophysical data.

Skip connections between corresponding downsampling and upsampling layers preserve fine-grained spatial information that might otherwise be lost during the encoding bottleneck. These connections are particularly crucial for astronomical applications, where small-scale features often carry significant physical meaning.

\subsubsection{Conditional Diffusion Implementation}

Conditional diffusion extends the basic framework to incorporate additional input parameters, enabling controlled generation based on cosmological and astrophysical variables. The conditioning mechanism typically involves concatenation or cross-attention:

\$$\textbackslash{}boldsymbol\{\epsilon\}_\textbackslash{}theta(\textbackslash{}mathbf\{x\}_t, t, \textbackslash{}mathbf\{c\}) = \textbackslash{}text\{UNet\}([\textbackslash{}mathbf\{x\}_t; \textbackslash{}mathbf\{c\}], t)\$$

where \$\mathbf\{c\}$ represents the conditioning vector containing cosmological parameters. Advanced implementations employ cross-attention mechanisms to achieve more sophisticated conditioning:

\$$\textbackslash{}text\{Attention\}(\textbackslash{}mathbf\{Q\}, \textbackslash{}mathbf\{K\}, \textbackslash{}mathbf\{V\}) = \textbackslash{}text\{softmax\}\left(\textbackslash{}frac\{\mathbf\{Q\}\mathbf\{K\}^T\}{\textbackslash{}sqrt\{d_k\}}\textbackslash{}right)\textbackslash{}mathbf\{V\}$\$

where \$\mathbf\{Q\}$ derives from the spatial features, while \$\mathbf\{K\}$ and \$\mathbf\{V\}$ incorporate conditioning information.

This architectural innovation enables precise control over generated HI maps, allowing researchers to explore parameter dependencies and generate synthetic observations for specific cosmological scenarios. The probabilistic nature of the generation process provides uncertainty quantification, crucial for robust scientific inference.

The synthesis of these architectural components creates a powerful framework for accelerating cosmological simulations while maintaining the complex statistical properties inherent in astrophysical phenomena. Through careful implementation and validation against traditional simulation methods, diffusion models represent a promising avenue for transforming computational astrophysics and enabling previously intractable parameter space explorations.

\end{document}