Informed Inversion:
Physics-Grounded Conditioning for Deep Imaging Inverse Problems

A dissertation presented

by

Linbo Tang

to

John A. Paulson School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Master of Science

in the subject of

Data Science

Harvard University

Cambridge, Massachusetts

May 2025

© 2025 Linbo Tang
All rights reserved.

Dissertation Advisor: Professor Yogesh Rathi Linbo Tang

Informed Inversion:
Physics-Grounded Conditioning for Deep Imaging Inverse Problems

Abstract

Deep learning solutions to imaging inverse problems typically treat observations as context-free inputs, ignoring the physical acquisition process that generated them. This thesis argues that explicitly conditioning on structured, acquisition-derived priors — rather than relying solely on learned implicit priors — yields consistent improvements in accuracy and cross-domain generalization. We term this approach *informed inversion* and demonstrate it on two canonical inverse problems in MRI.

First, we address MRI denoising in coil level. Rather than treating noise as a homogeneous nuisance, we condition the denoiser on a spatially resolved noise map measured directly from the scanner, enabling the model to adapt its restoration to the local noise statistics of each acquisition. This explicit prior eliminates the need for scanner-specific training data, and the method generalizes across platforms without fine-tuning, achieving up to 23% lower residual noise than the strongest baseline.

Second, we address accelerated MRI as a compressed sensing inverse problem. We show that encoding acquisition metadata — repetition time, flip angle, anatomy — via a generic text encoder fails to represent continuous numerical parameters. Our method introduces type-aware encoders paired with an out-of-distribution training strategy, yield a 20% SSIM improvement over the text-conditioned baseline and stronger cross-domain generalization.

Together, these contributions establish that physical priors must be encoded in a manner that respects their structure — a condition under which explicit acquisition-grounded conditioning substantially outperforms generic alternatives.

Contents

Abstract . . . iii
Acknowledgments . . . xi

Introduction **1**

1 Background **4**
1.1 Imaging Inverse Problems . . . 4
1.1.1 Denoising . . . 5
1.1.2 Compressed sensing reconstruction . . . 5
1.2 Deep Learning for Image Restoration . . . 6
1.2.1 Convolutional denoisers . . . 6
1.2.2 Blind denoising . . . 6
1.2.3 Transformer-based restoration . . . 6
1.2.4 Non-blind denoising with explicit priors . . . 7
1.2.5 Domain generalization in restoration . . . 7
1.3 Diffusion Models as Learned Priors . . . 8
1.3.1 Denoising Diffusion Probabilistic Models . . . 8
1.3.2 DDIM Sampling . . . 10
1.3.3 Conditional Diffusion Models . . . 11
1.3.4 Diffusion Models for Inverse Problems . . . 12
1.4 Conditioning Mechanisms in Deep Networks . . . 13
1.5 MRI Acquisition and Physical Conditioning Signals . . . 14
1.5.1 k-Space and the Fourier Relationship . . . 15
1.5.2 Parallel Imaging and Receiver Coil Arrays . . . 15
1.5.3 Noise in MRI: The Noise Scan . . . 16
1.5.4 Diffusion MRI . . . 16
1.5.5 Acquisition Metadata . . . 17
1.6 Evaluation Metrics . . . 18

2 Non-Blind Denoising with Explicit Noise Priors **20**
2.1 Introduction and Motivation . . . 20

2.1.1 The non-uniform nature of MRI noise . . . 20
2.1.2 The noise scan: a freely available physical prior . . . 21
2.1.3 From prior to generalization . . . 22
2.1.4 The coil-level processing advantage . . . 22
2.1.5 Training without MRI data . . . 23
2.2 Related Work . . . 23
2.2.1 Statistical dMRI Denoising . . . 23
2.2.2 Learning-Based Denoisers . . . 24
2.2.3 Blind and Non-Blind Noise Estimation . . . 24
2.2.4 Scanner Adaptability and Domain Generalization . . . 24
2.2.5 Coil-Level Processing in MRI . . . 25
2.3 Method . . . 25
2.3.1 Scanner-Specific Noise Prior Estimation . . . 25
2.3.2 Noise-Aware Transformer Architecture . . . 26
2.3.3 Coil Combination and SNR Analysis . . . 30
2.3.4 Synthetic Training Data . . . 30
2.4 Experiments . . . 31
2.4.1 Training Setup . . . 31
2.4.2 In-Vivo Multi-Scanner Evaluation . . . 32
2.4.3 Synthetic Noise Experiments . . . 35
2.5 Discussion . . . 39
2.5.1 Summary of Contributions . . . 39
2.5.2 Advantages of Explicit Noise Conditioning . . . 40
2.5.3 Training Without MRI Data . . . 40
2.5.4 Limitations . . . 41
2.5.5 Future Directions . . . 41

3 Structured Conditioning for Diffusion-Based Reconstruction 43
3.1 Introduction and Motivation . . . 43
3.2 Related Work . . . 45
3.2.1 Diffusion Models for Image Reconstruction . . . 45
3.2.2 Conditioning in Generative Models . . . 46
3.3 Method . . . 46
3.3.1 Data Curation . . . 46
3.3.2 Background: ContextMRI and its Limitations . . . 47
3.3.3 Type-Aware Metadata Embeddings . . . 48
3.3.4 Out-of-Distribution Training Strategy . . . 50
3.4 Experiments . . . 52

3.4.1 Experimental Setup . . . 52
3.4.2 Main Results . . . 53
3.4.3 Effect of Conditioning Strength . . . 53
3.4.4 Per-Field Metadata Ablations . . . 54
3.4.5 Qualitative Results . . . 56
3.4.6 OOD Noise Level Ablation . . . 57
3.4.7 Ensemble Reconstruction . . . 58
3.5 Discussion . . . 59
3.5.1 Summary . . . 59
3.5.2 Advantages of Type-Aware Conditioning . . . 60
3.5.3 Limitations . . . 60
3.5.4 Future Directions . . . 61

4 Discussion **62**
4.1 A Taxonomy of Acquisition-Context Conditioning . . . 62
4.2 Explicit Priors and Cross-Domain Generalization . . . 63
4.3 Future Directions . . . 63
4.3.1 Combining Noise and Metadata Conditioning . . . 63
4.3.2 Type-Aware Embeddings Beyond MRI . . . 64
4.3.3 Learned Noise Prior Estimation . . . 64
4.3.4 Extension to 3D and Temporal Imaging . . . 64
4.4 Conclusion . . . 65

References **66**

Appendix A Supplementary Results **70**
A.1 fastMRI Brain Metadata Distributions . . . 70
A.2 OOD Cardiology Metadata Distributions . . . 70
A.3 ContextMRI Cross-Attention Weight Analysis . . . 71

List of Tables

2.1 **Estimated residual noise levels** (σ_{MAD}) averaged across DWIs within each diffusion shell (excluding b=0). Lower values indicate lower residual noise. . 37

2.2 **Denoising performance on synthetic fastMRI data** across four SNR levels. Metrics reported as mean PSNR (dB), SSIM, and LPIPS over 100 test slices. Bold indicates best performance per SNR level. 39

3.1 **Quantitative results on the fastMRI brain dataset** (uniform 1D mask, 4× acceleration, CFG = 1.0). Bold indicates best performance per metric. 53

3.2 **Effect of CFG scale on reconstruction quality** for the OOD + Additive model. Bold indicates best performance per metric. 54

3.3 **Drop-one-field ablation.** Performance when all metadata fields are used except the listed one. Larger degradation indicates greater non-redundant contribution. 55

3.4 **Single-field ablation.** Performance when conditioning on only the listed field with all others dropped. Bold indicates best single-field performance. 56

3.5 **OOD noise level ablation.** Comparing high-noise and low-noise OOD training strategies for the OOD + Additive model. Differences fall within statistical error across all metrics. 59

3.6 **Ensemble reconstruction.** Pixel-wise averaging of five independent samples substantially improves all metrics at the cost of 5× inference time. 59

List of Figures

2.1 **Overview of the proposed noise-aware denoising architecture.** The network takes as input the noisy coil image and the measured noise prior Σ_c, which are encoded via separate dual-input embeddings before entering the encoder-decoder backbone. At each transformer block, FiLM-based conditioning modulates the intermediate feature maps using spatially varying gain γ and bias β parameters derived from Σ_c, suppressing high-variance regions before the attention calculation. The final skip connection is applied only to the image channel to produce the denoised coil output $\hat{S}_c$. 26

2.2 **Siemens Scanner — Diffusion-Weighted Images.** Representative DWIs at $b = 1500$ and $b = 3000\,\mathrm{s/mm^2}$ across all compared methods. Our method suppresses noise effectively at both b-values while preserving fine anatomical detail, with no structure visible in the residuals. 35

2.3 **GE Scanner — Diffusion-Weighted Images.** Representative DWIs at $b = 1000$ and $b = 2000\,\mathrm{s/mm^2}$, demonstrating zero-shot generalization to an unseen scanner platform with consistent noise suppression and preserved anatomical boundaries. 36

2.4 **Siemens Scanner — Fractional Anisotropy Maps.** FA maps derived from denoised DWI volumes. Our method produces sharper white matter boundaries and more stable diffusion contrast compared to all baselines, with no loss of fine structural detail. 37

2.5 **GE Scanner — Fractional Anisotropy Maps.** FA maps on the GE platform showing that improvements in downstream diffusion metrics transfer across scanner manufacturers without any retraining or fine-tuning. 38

3.1 **Proposed metadata conditioning framework.** The original ContextMRI framework (right branch) encodes all metadata as a single free-form text prompt via a frozen CLIP encoder. Our approach (left and center branches) introduces dedicated categorical condition embeddings for discrete attributes (anatomy, contrast, sequence, pathology) and numerical condition embeddings for continuous parameters (TR, TE, TI, flip angle, slice number), which are fused with the CLIP text embedding before being passed to the diffusion model. 48
3.2 **DDPM noise schedule.** The red dashed line at $\frac{3}{4}T_{\max}$ marks the minimum timestep at which OOD cardiac slices are noised during training. At timesteps above this threshold, $\bar{\alpha}_t$ is sufficiently small that anatomical content is effectively destroyed, leaving only the metadata embedding as a useful conditioning signal. 51
3.3 **Visualization of the OOD noising process.** Top row: original brain (in-distribution) and cardiac (OOD) slices. Bottom row: corresponding noised images. Brain features remain partially detectable after noising, while cardiac features are always completely obscured, ensuring the model must rely on metadata embeddings when processing OOD cardiac samples. 52
3.4 **No embeddings vs. OOD + Additive.** Reconstruction of a representative brain slice comparing an unconditioned model (no metadata embeddings) against our proposed OOD + Additive model. The conditioned reconstruction shows sharper gray-white matter boundaries and improved fidelity in low-SNR regions near the ventricles. 57
3.5 **ContextMRI vs. OOD + Additive.** Reconstructions of two representative brain slices. Each panel shows, from left to right, the undersampled measurement, the ContextMRI reconstruction, the OOD + Additive reconstruction, and the ground truth. The bottom row of each example shows zoomed-in regions, highlighting crisper ventricular edges and finer cortical detail in the OOD + Additive outputs. 58

A.1 **Distribution of categorical metadata fields in the fastMRI brain dataset.** Fields include system model, sequence type, and protocol name. The dominance of a small number of categories reflects the limited protocol diversity in the dataset. 70

A.2 **Distribution of continuous metadata fields in the fastMRI brain dataset.** Fields include TR, TE, TI, flip angle, and geometric parameters. Despite being recorded as continuous values, most parameters cluster at only a few discrete operating points, limiting the diversity of numerical conditioning seen during training. 71
A.3 **Distribution of categorical metadata fields in the CMRxRecon cardiac dataset.** Sequence types consist of long-axis and short-axis CINE acquisitions, and contrast type is uniformly balanced steady-state free precession — both distinct from the brain dataset protocol categories. 72
A.4 **Distribution of continuous metadata fields in the CMRxRecon cardiac dataset.** TR spans 3.6–50 ms, TE spans 1.1–1.6 ms, and flip angles range from 12 to 60 degrees, all lying outside the fastMRI brain distribution and providing the distributional distance necessary to force metadata reliance during OOD training. 73
A.5 **ContextMRI cross-attention weight analysis.** Token-level attention weights extracted from the mid-block cross-attention layer of the diffusion UNet, conditioned on a representative metadata prompt. The CLS token and anatomy token dominate the attention weights, while tokens corresponding to TR, TE, TI, and flip angle receive near-zero weight regardless of their numerical values. This confirms that the CLIP text encoder fails to surface acquisition parameters in a form the network can exploit, motivating the type-aware embedding approach of Chapter 3. 74

Acknowledgments

I am deeply grateful to Professor Yogesh Rathi, my thesis advisor, whose patience and careful guidance made this work possible. Working with Yogesh has been one of the most formative experiences of my graduate studies. He welcomed my curiosity even when my questions were half-formed, tolerated my initial unfamiliarity with the MRI domain, and never made me feel that the learning curve was anything other than a natural part of research. What I admire most about him is the way he engages with problems — with rigor, creativity, and genuine enthusiasm — and I have tried to carry that spirit into my own work. I leave this program a significantly better researcher because of him.

I owe a great deal to Professor Todd Zickler, who introduced me to the computer vision field and whose influence on how I think about visual problems runs through every chapter of this thesis. I will always remember visiting his office — he never failed to welcome me warmly, no matter how busy he was. His encouragement came at moments when I needed it most, and his kindness made me feel that I belonged in this field. The intellectual generosity he showed me is something I hope to one day extend to my own students.

I thank Professor Lipeng Ning for his thoughtful advice on the denoising project, and Qiang Liu for her practical expertise in MRI acquisition and for patiently introducing me to a modality I knew nothing about when I started. I am also grateful to Daniel Weinstock for serving on my thesis committee and for his steady guidance throughout my time in the master's program.

Finally, I thank my parents. They may not follow the technical details of this work, but they never needed to — they simply believed that what I was doing mattered, that it was correct and useful for the world, and that was enough. Their unconditional support and quiet confidence in me has been the most sustaining force throughout my graduate studies.

To everyone who has ever waited too long for a scan.

Introduction

Inverse problems are central to computer vision: given a degraded or incomplete observation, recover the underlying signal. In image denoising, the forward model is trivial but the noise process is complex; in accelerated imaging, the forward model is a severe undersampling operator. What unites them is ill-posedness — the observation alone does not uniquely determine the solution, and a prior over natural images is indispensable.

The computer vision and machine learning communities have made remarkable progress on both problems through learned priors. Convolutional denoisers (Zhang *et al.*, 2017), score-based generative models (Song and Ermon, 2019), and diffusion-based reconstruction networks (Chung and Ye, 2022) have pushed the state of the art on standard benchmarks by learning richer and richer implicit priors from data. Yet this progress has been accompanied by a quiet assumption that is rarely examined: the prior should come entirely from the training data, and the observation should be treated as a context-free input. The physical process that produced the observation — the scanner, the acquisition sequence, the noise characteristics — is treated as either unknown or irrelevant.

This assumption is not merely a modelling simplification. It is a deliberate discarding of information. Every image acquired by a physical sensor is accompanied by structured metadata describing the acquisition: noise statistics, hardware configuration, and protocol parameters that directly constrain the space of plausible reconstructions. In the computer vision literature, this information has largely been ignored in favour of scaling model capacity and dataset size.

The central thesis. This work investigates *informed inversion*: the design principle that explicitly conditioning deep networks on structured, acquisition-derived physical priors — rather than relying solely on learned implicit priors — yields consistent improvements in accuracy and cross-domain generalization. The argument is not that data-driven priors are insufficient, but that they are incomplete: physical acquisition context provides complementary information that is structured, reliable, and largely free, yet routinely discarded. When that structure is respected in the conditioning design, performance improves and generalization becomes possible across acquisition settings that were never seen during training.

We ground this argument in two concrete contributions, both situated within the broader ML and computer vision literature on inverse problems, and both evaluated on Magnetic Resonance Imaging as a demanding testbed. MRI is an ideal domain for this investigation: it is physically well-characterised, clinically high-stakes, and offers two qualitatively distinct forms of acquisition context — a measurable spatial noise field and rich structured metadata — that together cover the range of conditioning design challenges we wish to study.

Contribution I: Non-blind denoising with spatially varying noise priors. Chapter 2 revisits image denoising as a non-blind inverse problem. While the blind denoising literature has focused on estimating noise statistics from the image itself (Guo *et al.*, 2019), and recent self-supervised methods (Fadnavis *et al.*, 2020) avoid clean references entirely, neither framework exploits the case where the noise process is directly measurable. We show that conditioning on a spatially resolved noise map — obtained from a brief calibration scan — enables a single model to adapt to the noise statistics of any scanner without retraining. The model generalises across platforms without fine-tuning, achieving up to 23% lower residual noise than the strongest baseline. From a CV perspective, the central contribution is not the application but the conditioning design: we demonstrate that a dense spatial prior must be propagated throughout the feature hierarchy, not merely concatenated at the input.

Contribution II: Type-aware metadata conditioning for diffusion reconstruction. Chapter 3 addresses image reconstruction as a compressed sensing inverse problem and asks a question relevant to the broader conditional generative modelling literature: how should heterogeneous structured side information be encoded as a conditioning signal? Recent work in text-conditioned diffusion models (Rombach *et al.*, 2022) has shown that conditioning signal quality matters, but the encoding of non-textual, mixed-type structured data remains underexplored. We show that encoding acquisition metadata via a generic text encoder fails to represent continuous numerical parameters — confirmed through attention weight analysis — and that type-aware embeddings designed to match the structure of each field yield a 20% SSIM improvement and substantially stronger cross-domain generalization. The out-of-distribution training strategy we introduce to force active use of conditioning signals is likewise applicable beyond MRI.

Relationship between contributions. Both contributions address the same fundamental question from different angles: what is the right inductive bias for encoding physical acquisition context into a deep network? In Chapter 2, the prior is a dense spatial signal requiring hierarchical feature-level modulation. In Chapter 3, the prior is a heterogeneous mix of categorical and continuous scalars requiring type-aware embedding before attention-based integration. The conditioning mechanisms differ because the structure of the information differs — a design principle we articulate and defend in Chapter 4.

Organisation. Chapter 1 reviews the inverse problem framework, relevant deep learning architectures, conditioning mechanisms, and the MRI acquisition context motivating both problems. Chapters 2 and 3 present the two contributions. Chapter 4 synthesises the findings, situates them within the broader CV and ML literature, and outlines directions for future work.

Chapter 1

Background

This chapter reviews the technical foundations shared by both contributions of this thesis. We begin with the general inverse problem framework that unifies image denoising and reconstruction, survey the deep learning landscape for image restoration, cover diffusion models as learned priors for inverse problems, review conditioning mechanisms in deep networks, and conclude with the MRI acquisition physics that motivate the specific problems studied.

1.1 Imaging Inverse Problems

A broad class of image degradation and reconstruction problems can be written under a unified forward model:

$$\mathbf{y} = A\mathbf{x} + \mathbf{n}, \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the unknown clean image, $A : \mathbb{R}^d \to \mathbb{R}^m$ is a forward operator encoding the physics of the acquisition, $\mathbf{n}$ is additive noise, and $\mathbf{y} \in \mathbb{R}^m$ is the observation. The goal is to recover $\mathbf{x}$ from $\mathbf{y}$. When $m < d$, the system is underdetermined and the problem is ill-posed in the sense of Hadamard: the solution is not unique, and small perturbations in $\mathbf{y}$ can lead to large changes in $\hat{\mathbf{x}}$. Regularization — or equivalently, a prior over $\mathbf{x}$ — is therefore essential.

1.1.1 Denoising

When $A = I$ and $\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I)$, Eq. 1.1 reduces to additive white Gaussian noise (AWGN) removal. The maximum a posteriori (MAP) estimate is:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \ \|\mathbf{y} - \mathbf{x}\|^2 + \lambda \mathcal{R}(\mathbf{x}), \tag{1.2}$$

where $\mathcal{R}(\mathbf{x})$ encodes prior beliefs about natural images. Classical choices include total variation (Rudin *et al.*, 1992), which promotes piecewise smoothness, and sparsity in wavelet or DCT domains (Elad and Aharon, 2006). A key distinction in the denoising literature is between *blind* denoising — where σ is unknown and must be estimated from $\mathbf{y}$ — and *non-blind* denoising — where σ is provided. This thesis is concerned with the non-blind case, and specifically with the more general setting where the noise is spatially heterogeneous, i.e., $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$ with Σ a spatially varying diagonal covariance. Chapter 2 addresses how such a prior should be encoded into a deep network.

1.1.2 Compressed sensing reconstruction

When A is a partial measurement operator — for example, retaining only a subset of Fourier coefficients — the problem is severely underdetermined. Compressed sensing (CS) theory (Candès *et al.*, 2006; Donoho, 2006) establishes that exact recovery is possible when $\mathbf{x}$ is sparse in some basis Ψ, provided A satisfies a restricted isometry property (RIP). Practical recovery is achieved by solving:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \ \|A\mathbf{x} - \mathbf{y}\|^2 + \lambda \left\|\Psi\mathbf{x}\right\|_1. \tag{1.3}$$

In MRI, this framework was introduced by Lustig et al. (Lustig *et al.*, 2007), who showed that brain images, being approximately sparse in wavelet and finite-difference domains, could be faithfully reconstructed from as few as 25% of k-space measurements. However, the choice of prior $\mathcal{R}(\mathbf{x})$ critically determines reconstruction quality. Hand-crafted sparsity priors such as wavelet ℓ_1 penalties are outperformed by data-driven priors — particularly diffusion models — which implicitly capture the full statistics of natural images. Diffusion

models as priors for inverse problems are discussed in Section 1.3.

1.2 Deep Learning for Image Restoration

The past decade has seen deep neural networks become the dominant approach to image restoration. We review the major threads relevant to this thesis.

1.2.1 Convolutional denoisers

DnCNN (Zhang *et al.*, 2017) demonstrated that a plain residual convolutional network, trained with a noise level as a fixed hyperparameter, could outperform classical methods including BM3D (Dabov *et al.*, 2007) across a wide range of noise levels. FFDNet (Zhang *et al.*, 2018a) extended this to handle spatially varying noise by concatenating a noise level map to the network input — a design conceptually related to our approach, but limited to input-level injection without propagation through the feature hierarchy.

1.2.2 Blind denoising

CBDNet (Guo *et al.*, 2019) addresses real-world noise by estimating a pixel-wise noise map from the image itself, then feeding it to a denoising subnetwork. Noise2Noise (Lehtinen *et al.*, 2018) showed that a denoiser can be trained on pairs of independently corrupted images without access to clean targets, exploiting the zero-mean property of the noise. Self-supervised variants including Patch2Self (Fadnavis *et al.*, 2020) and Self2Self (Quan *et al.*, 2020) further remove the need for paired data by exploiting redundancy within a single acquisition.

1.2.3 Transformer-based restoration

Attention mechanisms have substantially improved image restoration by enabling long-range dependency modelling. SwinIR (Liang *et al.*, 2021) applied the Swin Transformer (Liu *et al.*, 2021) to restoration tasks including super-resolution and denoising. Restormer (Za-

mir *et al.*, 2022) introduced depth-wise separable attention computed across channels rather than spatial locations, enabling efficient processing of high-resolution images. Its encoder-decoder architecture uses PixelUnshuffle for downsampling and PixelShuffle for upsampling, preserving spatial detail at each scale. NAFNet (Chen *et al.*, 2022) proposed a simplified non-linear activation-free baseline that achieves competitive performance with lower computational cost.

1.2.4 Non-blind denoising with explicit priors

The majority of learning-based denoisers operate in the blind setting, either assuming a fixed noise level during training or estimating it from the corrupted input at test time (Guo *et al.*, 2019). The non-blind setting — where noise statistics are known — has received comparatively less attention in the deep learning era, despite being practically well-motivated in physical imaging systems where the noise process is directly characterisable from the hardware. Classical non-blind methods such as Wiener filtering (Wiener, 1949) exploit known noise spectra to design optimal linear filters, but cannot capture the complex signal-dependent structure of modern imaging noise. The gap between theoretically available physical information and what current deep denoisers actually exploit represents a clear opportunity, which Chapter 2 addresses in the context of multi-coil MRI.

1.2.5 Domain generalization in restoration

A recurring challenge in image restoration is generalization across acquisition settings — a problem particularly acute in medical imaging, where scanners from different manufacturers, field strengths, and coil configurations produce images with markedly different noise characteristics. Models trained on one scanner routinely degrade when deployed on another, necessitating costly re-training or fine-tuning on target-domain data (Karani *et al.*, 2018). Existing approaches to this problem broadly fall into two categories: data augmentation strategies that expose the model to diverse noise distributions during training (Foi *et al.*, 2008), and domain adaptation methods that align feature distributions across scanners (Guan

and Liu, 2022). Both require either anticipating the target distribution at training time or access to target-domain data at test time.

A more principled alternative is to condition the model on an explicit characterisation of the noise process, making the network's behaviour a function of the acquisition rather than a fixed mapping learned from a specific distribution. If the conditioning signal accurately describes the noise at test time, the model can adapt without any retraining — a property we term *scanner-adaptivity*. Chapter 2 realizes this through a measured, spatially resolved noise prior estimated from the scanner hardware itself, enabling a single model trained entirely on synthetic data to generalize zero-shot across multiple scanner platforms with no target-domain data required.

1.3 Diffusion Models as Learned Priors

Diffusion models (Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020) have emerged as the dominant class of generative models for images, offering high sample quality, stable training, and principled probabilistic interpretation. Their application as priors for inverse problems in imaging is the foundation of Chapter 3.

1.3.1 Denoising Diffusion Probabilistic Models

Diffusion models are a class of latent variable generative models that learn a data distribution by reversing a gradual noising process. Ho et al. (Ho *et al.*, 2020) introduced DDPMs as a tractable instantiation of the non-equilibrium thermodynamics framework of Sohl-Dickstein et al. (Sohl-Dickstein *et al.*, 2015), establishing the modern training and sampling recipe.

Forward process. A DDPM defines a Markov chain that progressively corrupts data $\mathbf{x}_0 \sim p_0(\mathbf{x})$ with Gaussian noise over T timesteps. The conditional distribution at each step is:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t;\, \sqrt{\alpha_t}\,\mathbf{x}_{t-1},\, (1-\alpha_t)\,I)\,, \tag{1.4}$$

where $\{\alpha_t\}_{t=1}^{T}$ is a monotonically decreasing noise schedule. A key property of this Markov chain is that the marginal at any timestep t can be expressed in closed form:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_t;\, \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0,\, (1-\bar{\alpha}_t)\, I\right), \quad \bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s, \tag{1.5}$$

which allows direct sampling of $\mathbf{x}_t$ from $\mathbf{x}_0$ without simulating the intermediate steps. As $t \to T$, $\bar{\alpha}_t \to 0$ and $\mathbf{x}_T \approx \mathcal{N}(0, I)$, so the data is eventually destroyed into pure Gaussian noise regardless of its initial content.

Reverse process. The generative model learns to reverse this process. The true reverse conditional $q(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ is intractable, but for small noise steps it is approximately Gaussian (Sohl-Dickstein *et al.*, 2015). A neural network $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ is therefore trained to approximate it. Rather than directly predicting the mean and variance of the reverse Gaussian, Ho et al. (Ho *et al.*, 2020) reparameterise the network to predict the noise ϵ that was added to $\mathbf{x}_0$ to produce $\mathbf{x}_t$. The training objective, derived from the evidence lower bound (ELBO), simplifies to:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,\, \mathbf{x}_0 \sim p_0,\, \epsilon \sim \mathcal{N}(0,I)} \left[\left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t) \right\|^2 \right], \tag{1.6}$$

where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$ is computed via Eq. 1.5. This objective is equivalent to denoising score matching (Song and Ermon, 2019), where the network implicitly estimates the Stein score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ of the marginal distribution at each noise level, since:

$$\epsilon_\theta(\mathbf{x}_t, t) \approx -\sqrt{1-\bar{\alpha}_t}\, \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t). \tag{1.7}$$

This connection to score matching is fundamental to the use of diffusion models as priors for inverse problems, as discussed in Section 1.3.4.

Noise schedule. The noise schedule $\{\beta_t = 1 - \alpha_t\}$ controls the rate at which signal is destroyed. Ho et al. (Ho *et al.*, 2020) used a linear schedule from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$ over $T = 1000$ steps. Subsequent work (Nichol and Dhariwal, 2021) proposed cosine schedules that maintain more signal at intermediate timesteps, improving sample quality for complex

distributions. The choice of schedule affects both training stability and the difficulty of the denoising task at each timestep.

1.3.2 DDIM Sampling

The original DDPM sampling procedure follows the Markov chain in reverse, requiring $T = 1000$ sequential network evaluations. At roughly 10–50 seconds per step on modern hardware, this makes generation prohibitively slow for interactive or high-throughput applications.

Denoising Diffusion Implicit Models (DDIM) (Song *et al.*, 2020) address this by observing that the training objective Eq. 1.6 does not uniquely determine the reverse process — there is a family of non-Markovian processes that share the same marginals $q(\mathbf{x}_t \mid \mathbf{x}_0)$ and therefore the same training objective, but admit much faster sampling. DDIM defines such a process whose reverse update is:

$$\hat{\mathbf{x}}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(\mathbf{x}_t, t) \right), \tag{1.8}$$

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{\mathbf{x}}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\, \epsilon_\theta(\mathbf{x}_t, t) + \sigma_t\, \epsilon_t, \tag{1.9}$$

where $\hat{\mathbf{x}}_{0|t}$ is a denoised estimate of $\mathbf{x}_0$ given $\mathbf{x}_t$, and σ_t controls stochasticity. Setting $\sigma_t = 0$ for all t gives a fully deterministic trajectory: the same $\mathbf{x}_T$ always maps to the same $\mathbf{x}_0$. This determinism has two important consequences. First, the trajectory can be subsampled: rather than iterating over all T steps, one can select a subsequence of $S \ll T$ timesteps and still obtain high-quality samples. In practice, $S = 50$–100 steps suffice, yielding a $10\times$–$20\times$ speedup over DDPM with negligible quality loss. Second, deterministic sampling enables consistent latent-space interpolation and inversion, which is exploited in some reconstruction methods. Chapter 3 uses 100 DDIM steps for all reconstruction experiments, matching the ContextMRI methodology.

1.3.3 Conditional Diffusion Models

Unconditional diffusion models sample from the marginal $p(\mathbf{x})$. For controlled generation or physics-constrained reconstruction, we require sampling from the conditional $p(\mathbf{x} \mid \mathbf{c})$ for some conditioning signal $\mathbf{c}$. Several strategies have been proposed.

Classifier guidance. Dhariwal and Nichol (Dhariwal and Nichol, 2021) proposed training a noisy classifier $p_\phi(y \mid \mathbf{x}_t)$ and using its gradient to steer the reverse process toward a target class y:

$$\tilde{\epsilon}_\theta(\mathbf{x}_t, t, y) = \epsilon_\theta(\mathbf{x}_t, t) - \sqrt{1-\bar{\alpha}_t}\, w\, \nabla_{\mathbf{x}_t} \log p_\phi(y \mid \mathbf{x}_t), \tag{1.10}$$

where w is a guidance scale. While effective, this requires training a separate classifier on noisy images at every noise level, which is expensive.

Classifier-free guidance. Ho and Salimans (Ho and Salimans, 2022) proposed a more elegant approach that trains a single model jointly for conditional and unconditional generation. During training, the conditioning signal $\mathbf{c}$ is randomly replaced by a null token $\varnothing$ with probability p_{drop}, so the model learns both $\epsilon_\theta(\mathbf{x}_t, \mathbf{c})$ and $\epsilon_\theta(\mathbf{x}_t, \varnothing)$ with shared weights. At inference, the two predictions are combined:

$$\hat{\epsilon}_{\text{cfg}} = \epsilon_\theta(\mathbf{x}_t, \varnothing) + w\left[\epsilon_\theta(\mathbf{x}_t, \mathbf{c}) - \epsilon_\theta(\mathbf{x}_t, \varnothing)\right], \tag{1.11}$$

where $w \geq 1$ is the guidance scale. Geometrically, the correction term $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}) - \epsilon_\theta(\mathbf{x}_t, \varnothing)$ points in the direction that makes $\mathbf{x}_t$ more consistent with $\mathbf{c}$, and w controls how strongly this direction is followed. Larger w increases adherence to the condition at the cost of sample diversity and sometimes perceptual quality. In practice $w \in [1, 3]$ offers a good trade-off. CFG has become the standard conditioning mechanism in large-scale text-to-image models (Rombach *et al.*, 2022) and is adopted in ContextMRI (Chung *et al.*, 2025) and Chapter 3.

A critical but often overlooked aspect of CFG is that its effectiveness depends entirely on how $\mathbf{c}$ is encoded. If the encoding is lossy — collapsing semantically distinct fields

into an undifferentiated embedding — the correction term carries little useful information regardless of the guidance scale. This is precisely the failure mode of CLIP-based metadata encoding in ContextMRI, which we diagnose and address in Chapter 3.

1.3.4 Diffusion Models for Inverse Problems

The connection between diffusion models and inverse problems follows from the score-based perspective. Song et al. (Song *et al.*, 2021) showed that diffusion models can be interpreted as discretisations of stochastic differential equations, and that the reverse process requires only the score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$, which is approximated by ϵ_θ. For a noisy linear inverse problem $\mathbf{y} = A\mathbf{x} + \mathbf{n}$, the posterior score decomposes via Bayes' rule as:

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t \mid \mathbf{y}) = \underbrace{\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)}_{\text{prior score}} + \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t)}_{\text{likelihood score}}. \tag{1.12}$$

The prior score is provided by the trained diffusion model; the likelihood score encodes data consistency via the forward operator A.

ScoreMRI (Chung and Ye, 2022) pioneered this framework for accelerated MRI reconstruction, alternating between score-function denoising steps and projection onto the measurement constraint set, generalising across undersampling patterns without retraining. The Decomposed Diffusion Sampler (DDS) (Chung *et al.*, 2023) further accelerated inference using Krylov subspace projections, reducing computational cost by an order of magnitude without sacrificing reconstruction quality.

Both works treat the diffusion prior as entirely unconditional — sampling from $p(\mathbf{x})$ rather than $p(\mathbf{x} \mid \mathbf{c})$ — leaving acquisition metadata entirely unexploited. ContextMRI (Chung *et al.*, 2025) was the first to introduce metadata conditioning into this framework, but as we demonstrate through cross-attention weight analysis in Chapter 3, its CLIP-based text encoding fails to represent the numerical structure of acquisition parameters. Our contribution directly addresses this limitation through type-aware embeddings designed to respect the structure of each metadata field.

1.4 Conditioning Mechanisms in Deep Networks

The choice of conditioning mechanism determines how well a network can exploit side information. We review the two mechanisms central to this thesis.

Concatenation. The simplest approach is to concatenate the conditioning signal to the network input as additional channels. FFDNet (Zhang *et al.*, 2018a) uses this strategy for noise level conditioning. The limitation is that the conditioning signal is only explicitly present at the first layer; deeper layers must propagate it through the learned feature representations, and there is no guarantee the signal survives the depth of a modern encoder-decoder.

Feature-wise Linear Modulation (FiLM). FiLM (Perez *et al.*, 2018) was introduced for visual question answering and conditions intermediate feature maps via learned affine parameters derived from the conditioning input $\mathbf{z}$:

$$\mathbf{Y} = \mathbf{X} \odot (1 + \gamma(\mathbf{z})) + \beta(\mathbf{z}), \tag{1.13}$$

where $\gamma, \beta \in \mathbb{R}^C$ are gain and bias vectors matching the feature dimension, and $\odot$ denotes element-wise multiplication. FiLM is computationally lightweight, does not increase the sequence length of attention layers, and can be inserted at any depth in the network. Crucially, it allows the conditioning signal to be re-injected at every layer, preventing information loss in deep hierarchies. This property makes it well-suited for propagating a spatially varying noise prior through a hierarchical encoder-decoder, as explored in Chapter 2.

Cross-attention conditioning. In transformer-based architectures, cross-attention naturally accommodates sequence-form conditioning. Given a query Q derived from the image features and key-value pairs (K, V) derived from the conditioning embedding, the attended

output is:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \tag{1.14}$$

where d_k is the key dimension. Cross-attention enables the network to selectively attend to different parts of the conditioning embedding, which is desirable when the embedding contains multiple distinct fields. However, the quality of this attention depends on the embedding construction: if multiple semantically distinct fields are collapsed into a single sequence by a text encoder (as in ContextMRI), the attention mechanism cannot distinguish between them. Chapter 3 shows that type-aware embeddings, which preserve the identity of each field, enable more effective cross-attention conditioning.

Adaptive layer normalisation. DiT (Peebles and Xie, 2023) proposed adaLN-Zero, which modulates layer normalisation parameters from a conditioning signal, enabling efficient and expressive conditioning in diffusion transformer architectures. This is closely related to FiLM and has become standard in large-scale text-to-image models.

Low-Rank Adaptation (LoRA). When adapting a large pre-trained model to a new conditioning signal, full fine-tuning is computationally expensive. LoRA (Hu *et al.*, 2021) introduces trainable low-rank decompositions $\Delta W = AB$ with $A \in \mathbb{R}^{d \times r}$, $B \in \mathbb{R}^{r \times d}$, $r \ll d$, added to the frozen weight matrices. This reduces the number of trainable parameters dramatically while preserving the pre-trained model's capabilities. Chapter 3 evaluates LoRA as a parameter-efficient alternative to full fine-tuning for integrating type-aware metadata embeddings.

1.5 MRI Acquisition and Physical Conditioning Signals

Magnetic Resonance Imaging (MRI) is a non-ionising modality that images soft tissue by exploiting the quantum mechanical property of nuclear spin. We introduce the acquisition concepts necessary to understand the two types of physical conditioning signal studied in this thesis.

1.5.1 k-Space and the Fourier Relationship

MRI measurements are acquired in *k-space* — the spatial frequency domain of the image. The relationship between the image $\mathbf{x}(r)$ and the k-space measurement $\hat{\mathbf{x}}(k)$ is given by the Fourier transform:

$$\hat{\mathbf{x}}(k) = \int \mathbf{x}(r)\, e^{-2\pi i k \cdot r}\, \mathrm{d}r. \tag{1.15}$$

A fully sampled acquisition measures $\hat{\mathbf{x}}(k)$ on a Cartesian grid and recovers $\mathbf{x}(r)$ by inverse Fourier transform. Accelerated MRI reduces scan time by acquiring only a fraction of k-space — typically retaining low frequencies (which contain most signal energy) plus a random subset of high frequencies. The resulting reconstruction from undersampled measurements is an inverse problem of the form Eq. 1.1, with A being the partial Fourier operator.

1.5.2 Parallel Imaging and Receiver Coil Arrays

Modern MRI scanners use arrays of C receiver coils, each sensitive to a different spatial region. Let $\mathbf{S}_c \in \mathbb{C}^{H \times W}$ denote the sensitivity map of coil c. The signal measured by coil c is:

$$y_c(r) = \mathbf{S}_c(r) \cdot \mathbf{x}(r) + n_c(r), \tag{1.16}$$

where $n_c(r) \sim \mathcal{CN}(0, \sigma_c^2(r))$ is spatially varying complex Gaussian noise whose variance depends on the coil geometry and hardware. The final image is typically obtained by root-sum-of-squares (RSS) combination:

$$\hat{\mathbf{x}}(r) = \sqrt{\sum_{c=1}^{C} |\hat{\mathbf{x}}_c(r)|^2}, \tag{1.17}$$

where $\hat{\mathbf{x}}_c$ is the reconstructed image from coil c. Alternatively, coil-sensitivity maps estimated via ESPIRiT (Uecker *et al.*, 2014) can be used to compute the minimum-variance unbiased estimator (MVUE), which is used in Chapter 3 for data preprocessing.

1.5.3 Noise in MRI: The Noise Scan

The noise in each receiver coil is determined by the coil's geometry, number of elements, spatial positioning relative to the subject, and the scanner's hardware. This noise is *spatially varying* and *scanner-specific*: the same tissue imaged on different scanners will exhibit markedly different coil-wise noise patterns.

A *noise scan* is a brief acquisition using the same readout sequence as the diagnostic scan but with all radio-frequency (RF) excitation and refocusing pulses suppressed. Because no spin magnetisation is excited, the measured signal consists purely of thermal noise from the receiver coils. Let $N \in \mathbb{C}^{H \times W \times C \times T}$ denote the complex multi-coil noise data across C coils and T repetitions. Assuming temporally independent complex Gaussian noise per coil, the voxel-wise noise standard deviation is estimated as:

$$\Sigma_c(h, w) = \sqrt{\frac{1}{2T} \sum_{t=1}^{T} \left(|\Re(N_{h,w,c,t})|^2 + |\Im(N_{h,w,c,t})|^2 \right)}, \tag{1.18}$$

where the factor of $1/2$ accounts for the per-component variance of the complex Gaussian. Stacking all coils yields the full prior $\Sigma \in \mathbb{R}^{H \times W \times C}$, a spatially resolved map of the scanner's noise characteristics. This is the physical conditioning signal used in Chapter 2.

1.5.4 Diffusion MRI

Diffusion MRI (dMRI) is a specialised modality that probes tissue microstructure by measuring the diffusion of water molecules. By applying magnetic field gradients in different directions, the signal attenuation at each voxel encodes the directional diffusion profile, from which the diffusion tensor and downstream quantities such as fractional anisotropy (FA) and mean diffusivity (MD) are derived.

dMRI acquisitions at high b-values (the diffusion weighting parameter) suffer from particularly severe signal-to-noise ratio (SNR) degradation, because the diffusion encoding suppresses the MR signal by a factor of $e^{-b \cdot D}$, where D is the apparent diffusion coefficient. At $b = 3000\,\mathrm{s/mm^2}$, the remaining signal is a small fraction of the $b = 0$ baseline, making denoising essential for downstream tractography and microstructural modelling. This is the

application context of Chapter 2, which addresses coil-level denoising of multi-shell dMRI.

1.5.5 Acquisition Metadata

Every MRI scan is accompanied by rich protocol metadata that determines the tissue contrast and constrains the plausible appearance of the image. The key parameters relevant to Chapter 3 are:

- **Repetition time (TR):** The time between successive RF excitation pulses. TR controls T1 contrast: long TR reduces T1 weighting.
- **Echo time (TE):** The time between excitation and signal readout. TE controls T2 contrast: long TE increases T2 weighting and signal decay.
- **Inversion time (TI):** Used in inversion-recovery sequences to suppress specific tissues (e.g., FLAIR suppresses cerebrospinal fluid).
- **Flip angle:** The tip angle of the RF pulse. Flip angle interacts with TR to determine the Ernst angle and steady-state magnetisation.
- **Sequence type:** The pulse sequence design (e.g., TurboSpinEcho, FLASH), which determines the k-space sampling trajectory and contrast mechanism.
- **Anatomy and contrast:** Categorical labels indicating the imaged body region and contrast agent status.

These parameters are recorded continuously in the DICOM header of every clinical scan. Yet as we demonstrate in Chapter 3, the fastMRI brain dataset (Zbontar *et al.*, 2018) exhibits a narrow distribution over most of these parameters — TR, TE, and flip angle each cluster at only a few discrete values (see Appendix **??**). This distributional narrowness creates a challenge: a model that ignores metadata and relies on image content alone can perform well on in-distribution data. Our out-of-distribution training strategy addresses this by forcing the model to actively exploit metadata conditioning.

1.6 Evaluation Metrics

We use the following metrics throughout this thesis. Different metrics capture complementary aspects of reconstruction quality and are reported together to give a complete picture.

Peak Signal-to-Noise Ratio (PSNR). PSNR measures pixel-level fidelity in decibels:

$$\mathrm{PSNR}(\mathbf{x}, \hat{\mathbf{x}}) = 20 \log_{10} \left(\frac{\mathrm{MAX}}{\mathrm{RMSE}(\mathbf{x}, \hat{\mathbf{x}})} \right), \tag{1.19}$$

where MAX is the maximum possible pixel value and $\mathrm{RMSE} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2 / \sqrt{d}$. PSNR is monotonically related to MSE and does not capture structural or perceptual quality. Higher is better.

Structural Similarity Index (SSIM). SSIM (Wang *et al.*, 2004) measures perceptual similarity by comparing luminance μ, contrast σ, and structure σ_{xy} between patches:

$$\mathrm{SSIM}(\mathbf{x}, \hat{\mathbf{x}}) = \frac{(2\mu_x \mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)}, \tag{1.20}$$

where c_1, c_2 are stability constants. SSIM values lie in $[-1, 1]$; for natural images the range is effectively $[0, 1]$. It is more sensitive to structural distortions such as blurring and ringing than PSNR. Higher is better.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS (Zhang *et al.*, 2018b) measures perceptual distance using deep feature activations from a pre-trained VGG or AlexNet network:

$$\mathrm{LPIPS}(\mathbf{x}, \hat{\mathbf{x}}) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\phi_l(\mathbf{x})_{hw} - \phi_l(\hat{\mathbf{x}})_{hw}) \|_2^2, \tag{1.21}$$

where ϕ_l denotes the feature map at layer l and w_l are learned perceptual weights. LPIPS correlates more strongly with human perceptual judgements than PSNR or SSIM, particularly for textured or high-frequency content. Lower is better.

Residual noise level (σ_{MAD}). For the denoising evaluation in Chapter 2, we follow Rajan *et al.* (2010) and estimate residual noise using the local-difference median absolute deviation:

$$\sigma_{\text{MAD}} = \frac{\text{median}(|\mathbf{r} - \text{median}(\mathbf{r})|)}{0.6745}, \tag{1.22}$$

where $\mathbf{r}$ is the local difference image. This estimate is robust to outliers and insensitive to the signal structure, making it preferable to variance-based estimates in the presence of anatomical edges. Lower indicates less residual noise.

Chapter 2

Non-Blind Denoising with Explicit Noise Priors

2.1 Introduction and Motivation

Image denoising is one of the oldest and most studied problems in computer vision, yet a practically important regime remains underexplored: the case where the noise process is not merely unknown, but is *directly measurable* from the imaging hardware. This chapter addresses that regime, and argues that when physical noise information is available, it should be treated as a first-class input to the denoising network — not estimated, not ignored, but explicitly conditioned upon.

2.1.1 The non-uniform nature of MRI noise

In natural image denoising, noise is commonly modeled as additive white Gaussian noise (AWGN) with a spatially uniform standard deviation σ. This assumption, while convenient, is rarely satisfied in practice. In Magnetic Resonance Imaging, the noise is fundamentally different in character. Each receiver coil in a modern parallel imaging array contributes its own noise, whose spatial distribution is determined by the coil's geometry, its distance from the subject, the number of array elements, and the scanner's hardware properties. The result

is a noise field that is *spatially heterogeneous* — varying significantly across the image plane — and *scanner-specific* — markedly different between scanner platforms, field strengths, and coil configurations. A voxel near a peripheral coil element may experience several times more noise than a voxel near the isocentre. This spatial structure is not an artifact; it is a physical consequence of the electromagnetic geometry of the receiver array. Crucially, it is a structure that carries information — information that existing deep learning denoisers systematically discard.

2.1.2 The noise scan: a freely available physical prior

A key insight motivating this work is that the scanner's coil-wise noise characteristics can be measured directly through a brief *noise scan* — an acquisition that uses the same readout sequence as the diagnostic scan but suppresses all radio-frequency (RF) excitation and refocusing pulses. Because no spin magnetization is excited, the scanner acquires pure thermal noise from the receiver coils, with no tissue signal. This acquisition takes only a few seconds and is already part of many clinical and research protocols. From the resulting multi-coil noise data, one can estimate a spatially resolved noise standard deviation map $\Sigma_c(h, w)$ for each coil c, capturing the precise spatial distribution of noise across the image plane (see Section 2.3.1 for the formal derivation).

This noise prior is remarkable in several respects. It is *physically grounded* — derived directly from the scanner hardware rather than inferred from image statistics. It is *spatially resolved* — providing voxel-wise noise estimates at the same resolution as the image. It is *coil-specific* — capturing the distinct noise profile of each receiver element in the array. And critically, it is *freely available* — requiring no additional diagnostic imaging, no clean reference images, and no assumptions about image content. Given all of this, it is surprising that existing deep learning denoisers do not exploit it.

2.1.3 From prior to generalization

Beyond improving denoising quality on a single scanner, the noise prior enables something more fundamental: *scanner-adaptive generalization*. A conventional deep denoiser learns to remove noise by implicitly adapting to the noise distribution seen during training. When deployed on a different scanner — with a different coil geometry, a different number of elements, a different noise floor — its implicit assumptions are violated and performance degrades. This is a well-documented failure mode in medical image analysis more broadly, and is one reason why models trained at one institution often fail to generalize to data from another (Karani *et al.*, 2018).

Our approach breaks this dependency. By conditioning on the measured Σ_c at inference time, the network does not need to know in advance which scanner produced the image — it is told. A single model, trained once on synthetic data, can adapt its denoising behavior to any scanner for which a noise scan is available. This is the central advantage of explicit prior conditioning over implicit learning: it shifts the burden of scanner-specificity from the training distribution to the conditioning signal, where it can be resolved at inference time with no additional training.

2.1.4 The coil-level processing advantage

A second contribution of this chapter concerns *where* denoising is performed. Conventional approaches denoise the combined magnitude image produced by root-sum-of-squares (RSS) combination of the individual coil signals. We argue this is suboptimal for two reasons. First, the RSS combination is a nonlinear operation that mixes the noise from all coils; denoising after combination cannot disentangle coil-specific noise patterns that are still individually structured before combination. Second, and more fundamentally, there is a provable SNR advantage to denoising before combination: if each coil's residual noise variance is reduced by a factor $\eta_c < 1$, the post-RSS noise power is $\sum_c \eta_c \sigma_c^2$ rather than $\sum_c \sigma_c^2$, yielding a multiplicative SNR gain of $1/\sqrt{\overline{\eta}}$ that is unavailable when denoising the combined image. Performing denoising at the individual coil level, using the coil-specific noise prior

Σ_c, exploits this advantage while naturally incorporating the physically meaningful coil structure.

2.1.5 Training without MRI data

A further distinguishing feature of our approach is that the denoising network is trained entirely on synthetic data — natural images corrupted with physics-based simulations of coil noise — without requiring any real MRI acquisitions. Coil sensitivity maps are simulated via the Biot-Savart model (Esin, 2017), and spatially varying noise is synthesised by smoothing and scaling the simulated coil images. This synthetic training strategy is enabled precisely because the noise prior is explicit: the network does not need to learn scanner-specific noise characteristics from real data, because they are provided at inference time through Σ_c. The result is a model that generalises zero-shot to in-vivo acquisitions from multiple scanner platforms without any domain-specific training data.

2.2 Related Work

2.2.1 Statistical dMRI Denoising

The dominant classical approaches exploit redundancy along the diffusion encoding dimension. MPPCA (Veraart *et al.*, 2016), Hybrid-PCA (Neto Henriques *et al.*, 2022), Kernel PCA (Ramos-Llordén *et al.*, 2021), and NORDIC (Moeller *et al.*, 2021) identify and remove noise-dominated components under assumptions of low-rank signal structure and locally stationary Gaussian noise. While data-efficient and requiring no training data, these methods derive their statistical assumptions from idealized acquisition models. When the true noise violates these assumptions — as it does under unusual coil geometry or strong spatial correlation — performance degrades with no mechanism for correction. Crucially, none incorporate any information about the scanner's actual noise characteristics.

2.2.2 Learning-Based Denoisers

Chapter 1 reviewed the broader landscape of deep learning for image restoration, including transformer-based architectures (Restormer (Zamir *et al.*, 2022), NAFNet (Chen *et al.*, 2022)) and self-supervised methods (Patch2Self (Fadnavis *et al.*, 2020), Self2Self (Quan *et al.*, 2020)) relevant to dMRI. A common thread across all of these methods is that the noise process is treated as an implicit property to be inferred from the corrupted image, either through exposure to a particular noise distribution at training time or through auxiliary estimation at test time. None condition on explicit, hardware-derived noise statistics, and none operate at the individual receiver-coil level.

2.2.3 Blind and Non-Blind Noise Estimation

As discussed in Chapter 1, FFDNet (Zhang *et al.*, 2018a) demonstrated that accepting a user-provided noise level map as input improves both performance and flexibility over fixed-noise-level denoisers. Our work extends this principle in three directions: the noise map is derived from physical hardware measurement rather than user specification, it is propagated throughout the network hierarchy via feature modulation rather than injected only at the input, and it is applied at the individual coil level rather than to the combined image.

2.2.4 Scanner Adaptability and Domain Generalization

As noted in Chapter 1, generalizing across scanner platforms is a persistent challenge in medical image analysis (Karani *et al.*, 2018). Existing approaches — domain adaptation (Guan and Liu, 2022), physics-based noise augmentation (Foi *et al.*, 2008), and test-time adaptation — all require either target-domain data or knowledge of the target noise distribution at training time. Our approach sidesteps these requirements entirely: by conditioning on the measured Σ_c at inference time, the model adapts to any scanner without retraining, target-domain data, or distribution matching.

2.2.5 Coil-Level Processing in MRI

Parallel imaging methods including SENSE (Pruessmann *et al.*, 1999), GRAPPA (Griswold *et al.*, 2002), and ESPIRiT (Uecker *et al.*, 2014) exploit the spatial encoding provided by multi-coil arrays for accelerated acquisition and optimal coil combination. In the denoising literature, however, processing at the individual coil level prior to RSS combination has received virtually no attention. Existing learned denoisers operate almost universally on the combined magnitude image, discarding the coil structure that carries spatially resolved noise information. To the best of our knowledge, our work is the first to perform deep learning denoising at the coil level for dMRI, with a provable SNR advantage over post-combination denoising derived in Section 2.3.3.

2.3 Method

2.3.1 Scanner-Specific Noise Prior Estimation

Let $N \in \mathbb{C}^{H\times W\times C\times T}$ denote the complex multi-coil pure-noise data across C coils and T temporal frames acquired during the noise scan. Assuming independent complex Gaussian noise per coil, the voxel-wise noise standard deviation is estimated as:

$$\Sigma_c(h,w) = \sqrt{\frac{1}{2T}\sum_{t=1}^{T}\left(|\Re(N_{h,w,c,t})|^2 + |\Im(N_{h,w,c,t})|^2\right)}, \tag{2.1}$$

where the factor of $1/2$ accounts for the per-component variance of the complex Gaussian, and $\Sigma_c \in \mathbb{R}^{H\times W}$ is a spatially varying noise-σ map for coil c. Stacking all coils yields the full prior $\Sigma \in \mathbb{R}^{H\times W\times C}$, which captures non-uniform noise patterns arising from coil geometry, sensitivity fall-off, and hardware performance. This map is computed once per scan session and reused across all diffusion directions and b-values, adding negligible overhead to the acquisition protocol.

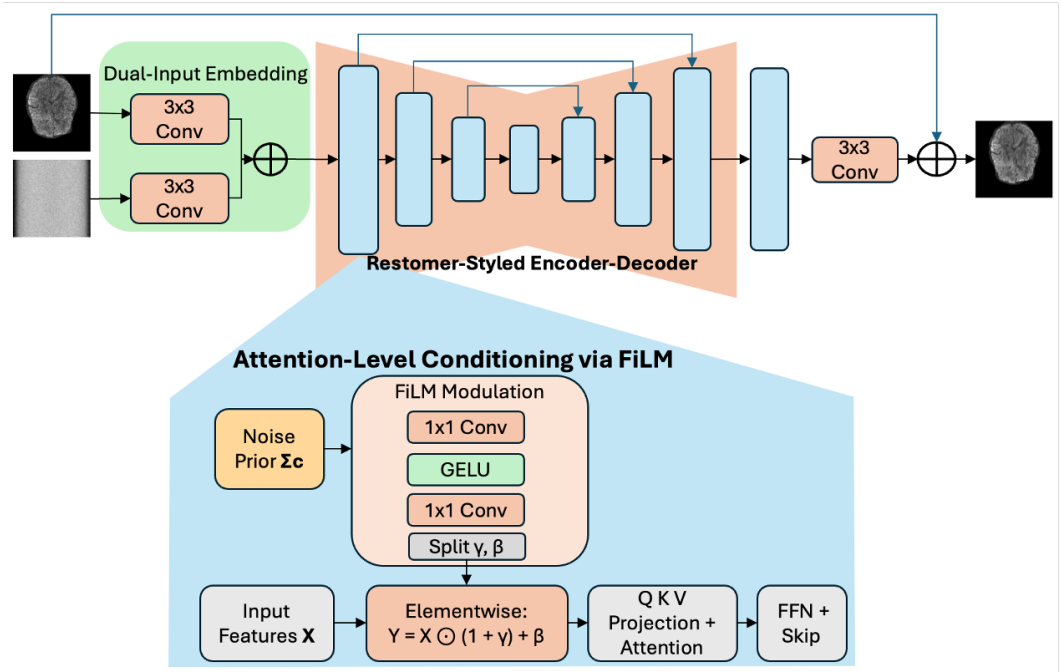


Figure 2.1: ***Overview of the proposed noise-aware denoising architecture.*** *The network takes as input the noisy coil image and the measured noise prior Σ_c, which are encoded via separate dual-input embeddings before entering the encoder-decoder backbone. At each transformer block, FiLM-based conditioning modulates the intermediate feature maps using spatially varying gain γ and bias β parameters derived from Σ_c, suppressing high-variance regions before the attention calculation. The final skip connection is applied only to the image channel to produce the denoised coil output $\hat{S}_c$.*

2.3.2 Noise-Aware Transformer Architecture

Our denoising network is designed around a modular conditioning philosophy: the dual-input embedding and FiLM conditioning blocks are self-contained components that can be inserted into any hierarchical encoder-decoder denoising framework without modifying the backbone architecture. This modularity is a deliberate design choice — the conditioning mechanisms are agnostic to the specific backbone, and their benefits transfer across architectures. The overall architecture is illustrated in Figure 2.1. We describe each component in turn, beginning with the motivation for spatial feature modulation.

Motivation: Spatially Varying Noise and Attention Degradation

Before describing our architecture, we motivate the need for explicit spatial noise conditioning within the attention mechanism. Standard self-attention computes pairwise interactions between feature vectors via scaled dot-products:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \tag{2.2}$$

The attention weights are determined by the dot-products $QK^\top$, which are sensitive to the magnitude of the query and key vectors. In natural image denoising with spatially uniform AWGN, feature magnitudes are roughly homogeneous across spatial locations, and the attention mechanism behaves as intended.

In dMRI, however, coil sensitivity creates a spatially varying noise field: regions near the periphery of the coil array experience substantially higher noise variance than regions near the isocentre. When a noisy image is passed through the network, these high-variance regions produce feature vectors with disproportionately large magnitudes. During the dot-product calculation, these vectors *dominate* the attention weights — effectively shouting over the quieter, structurally informative regions. The attention mechanism consequently attends to noise-corrupted locations rather than anatomically relevant ones, propagating artifacts into the global context representation and throughout the decoder. We refer to this failure mode as the *shouting effect*: high-variance noisy regions hijack the attention mechanism precisely because they are the most corrupted.

The root cause is that standard attention has no mechanism to account for the reliability of different spatial locations. Our FiLM-based conditioning addresses this directly by normalizing feature magnitudes according to the measured noise prior before the attention calculation.

Dual-Input Embedding

For each coil c, the network input is $\mathbf{X}_c = \text{concat}(|S_c|\,, \Sigma_c)$, where $|S_c| \in \mathbb{R}^{H\times W}$ is the coil magnitude image and $\Sigma_c \in \mathbb{R}^{H\times W}$ is the corresponding noise prior map. At the first stage,

two separate 3×3 convolutions encode the image and noise channels independently into half the embedding dimension $d/2$ each, and the resulting feature tensors are concatenated to form the initial embedding of dimension d:

$$\mathbf{F}_0 = \text{concat}(\text{Conv}_{3\times3}(|S_c|),\ \text{Conv}_{3\times3}(\Sigma_c)) \in \mathbb{R}^{d\times H\times W}. \tag{2.3}$$

Encoding the two channels through separate convolutions rather than concatenating them as raw inputs preserves a clean separation between structural content and noise statistics from the very first layer, preventing the network from conflating image features with noise features in early representations. The subsequent encoder-decoder follows the standard Restormer hierarchy, with downsampling and upsampling via PixelUnshuffle and PixelShuffle operations.

FiLM-Based Spatial Feature Modulation

The dual-input embedding injects noise information at the input level, but deeper transformer layers can progressively lose this conditioning as features become more abstract. To maintain noise awareness throughout the hierarchy and directly address the shouting effect, we extend each transformer block with a lightweight FiLM (Perez *et al.*, 2018) module that performs spatially varying affine modulation of the feature maps before the attention calculation.

Specifically, the noise prior Σ_c is passed through a two-layer 1×1 convolutional network with a GELU nonlinearity to produce spatially resolved gain γ and bias β maps that match the spatial dimensions and channel count of the attention input:

$$[\gamma,\ \beta] = \text{Conv}_{1\times1}(\text{GELU}(\text{Conv}_{1\times1}(\Sigma_c))), \quad \gamma, \beta \in \mathbb{R}^{C\times H\times W}. \tag{2.4}$$

These parameters are used to modulate the query-key-value (QKV) input features $\mathbf{X}$ at every spatial location before the attention projection:

$$\mathbf{Y} = \mathbf{X} \odot (1+\gamma) + \beta. \tag{2.5}$$

The effect of this modulation is to locally rescale and shift feature magnitudes according to the measured noise level at each spatial location. In high-noise regions, γ learned to attenuate feature magnitudes, effectively *turning down* the contribution of unreliable locations before they participate in the dot-product calculation. In low-noise regions, the modulation is close to identity, preserving the feature magnitudes of structurally reliable content. This directly suppresses the shouting effect: high-variance noisy regions are normalized before the attention weights are computed, preventing them from dominating the global context.

Critically, the FiLM parameters are *spatially varying* — γ and β are full spatial maps rather than scalar or channel-wise parameters. This allows the modulation to adapt to the precise noise profile of each coil independently, rather than applying a uniform correction across the image. The module is inserted at every transformer block in the encoder-decoder, ensuring that noise awareness is maintained throughout all levels of the feature hierarchy without introducing significant computational overhead.

Selective Residual Skip

While noise awareness is maintained throughout the encoder-decoder via FiLM-modulated features, including the noise prior channel in the final residual connection would risk reintroducing noise-related components directly into the image output, since Σ_c represents variance rather than signal content. We therefore apply the final skip connection only to the image channel. Letting $\Delta_c = F_\theta(\mathbf{X}_c)$ denote the residual correction predicted by the network, the denoised coil output is:

$$\hat{S}_c = |S_c| + \Delta_c. \tag{2.6}$$

In this formulation, Δ_c is the network's learned estimate of the noise component, while the noise prior Σ_c influences the prediction entirely through the FiLM modulation rather than through direct addition to the output. A single set of weights is shared across all coils; each coil is denoised independently using identical parameters but its own noise prior Σ_c, preserving coil-specific noise conditioning while keeping the parameter count constant

regardless of the number of coils.

2.3.3 Coil Combination and SNR Analysis

After denoising all C coils individually, the final diffusion-weighted magnitude image is reconstructed by root-sum-of-squares (RSS) combination:

$$\hat{M} = \sqrt{\sum_{c=1}^{C} \left|\hat{S}_c\right|^2}. \tag{2.7}$$

We now derive the SNR advantage of performing denoising before RSS combination. Let each coil measurement be $y_c = s_c + n_c$, where s_c is the true coil signal and $n_c \sim \mathcal{CN}(0, \sigma_c^2)$ is independent complex Gaussian noise. Before any denoising, the total noise power contributing to the RSS image is $\sum_c \sigma_c^2$. If coil-level denoising reduces each coil's residual noise variance by a factor $\eta_c \in (0, 1)$, the effective post-RSS noise power becomes $\sum_c \eta_c \sigma_c^2$. The resulting SNR gain relative to the undenoised case is:

$$\text{Gain}_{\text{SNR}} \propto \sqrt{\frac{\sum_c \sigma_c^2}{\sum_c \eta_c \sigma_c^2}} = \frac{1}{\sqrt{\bar{\eta}}}, \quad \bar{\eta} = \frac{\sum_c \eta_c \sigma_c^2}{\sum_c \sigma_c^2} < 1. \tag{2.8}$$

Because the RSS operation combines squared magnitudes before taking the square root, even small per-coil reductions in noise variance contribute additively to the total noise energy, yielding a multiplicative improvement in the final SNR. This gain is fundamentally unavailable when denoising is applied to the combined image, since RSS is a nonlinear operation that irreversibly mixes coil-specific noise structures. Furthermore, the coil-level noise prior Σ_c provides the precise variance information needed to achieve near-optimal per-coil noise reduction, maximizing $1/\sqrt{\bar{\eta}}$.

2.3.4 Synthetic Training Data

A distinguishing feature of our approach is that no real dMRI data is required for training. We generate paired training data from 20,000 natural images through a physics-based simulation pipeline. For each image I_{GT}, coil sensitivity maps $C \in \mathbb{C}^{H \times W \times C}$ are simulated using the Biot-Savart electromagnetic model (Esin, 2017), which computes the expected

sensitivity profile of a surface coil array at 3T. A target coil c is randomly selected, and the clean coil image is synthesized as:

$$S_c = |I_{\text{GT}} \odot C_c| \,. \tag{2.9}$$

To emulate the spatially varying, scanner-specific noise characteristic of real acquisitions, a noise standard deviation map $\Sigma_c(x, y)$ is constructed by repeatedly applying Gaussian smoothing to S_c and optionally inverting it, producing a smooth spatial envelope that mimics the sensitivity fall-off of a real coil. The local noise level is drawn uniformly from $[\sigma_{\min}, \sigma_{\max}]$ and combined with a small global noise floor σ_{global}:

$$\Sigma_c(x, y) = \sqrt{(\sigma_{\text{local}} \cdot f_{\text{smooth}}(S_c))^2 + \sigma_{\text{global}}^2}. \tag{2.10}$$

The noisy observation is then drawn as:

$$\tilde{S}_c = S_c + \mathcal{N}\big(0,\, \Sigma_c(x, y)^2\big) \,. \tag{2.11}$$

Each training sample consists of the two-channel input $[\tilde{S}_c,\, \Sigma_c]$ and the clean target S_c. This simulation strategy produces diverse, spatially heterogeneous noise realizations while maintaining physical consistency with real coil sensitivity profiles. The explicit noise prior Σ_c is included as both the second input channel and as the conditioning signal for FiLM modulation, enabling the network to learn the mapping from measured noise statistics to optimal denoising behavior without ever seeing real MRI data during training.

2.4 Experiments

2.4.1 Training Setup

We train the network using standard supervised learning with paired noisy–clean coil images generated entirely from synthetic data, as described in Section 2.3.4. Training data are derived from 20,000 natural images synthetically converted into coil-domain representations, yielding a large and diverse set of spatially heterogeneous noise realizations

that cover a wide range of coil geometries and noise levels.

The network is optimized with an ℓ_2 reconstruction loss between the denoised coil output $\hat{S}_c$ and the clean coil reference S_c:

$$\mathcal{L} = \left\| \hat{S}_c - S_c \right\|_2^2 . \tag{2.12}$$

Training is performed at the *coil level* rather than on combined images, preserving coil-specific noise statistics and enabling the network to learn the mapping from the measured noise prior Σ_c to optimal per-coil denoising behavior. Inputs are 2D patches of size 128×128, randomly cropped from the synthesized coil images and augmented with random horizontal and vertical flips and 90-degree rotations to improve spatial generalization.

Optimization uses the Adam optimizer (**?**) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an initial learning rate of 2×10^{-4} with cosine decay scheduling. Gradient clipping is applied to prevent instability during early training when the FiLM modulation parameters are being initialized. No learning rate warmup is used. The network is trained for 200,000 iterations on a single NVIDIA A100 GPU, with a batch size of 8 patch pairs per iteration. The FiLM modulation parameters γ and β are initialized to zero, so that at the start of training the modulation is close to identity and the network begins from the unmodulated baseline, allowing stable convergence before the conditioning mechanism becomes active.

At inference time, the measured noise prior Σ_c from the noise scan is provided directly to the network, requiring no additional estimation or preprocessing beyond the noise standard deviation computation in Eq. 2.1. Each coil is denoised independently using the shared network weights and its own Σ_c, after which the denoised coils are combined via RSS to produce the final diffusion-weighted image.

2.4.2 In-Vivo Multi-Scanner Evaluation

To demonstrate the scanner-adaptive capability of our framework under realistic clinical conditions, we evaluate on in-vivo diffusion MRI data acquired from a healthy volunteer on two 3T scanner platforms with different manufacturers, coil configurations, and acquisition

protocols. This study was conducted with the approval of the Institutional Review Board of Mass General Brigham, and the subject provided written informed consent prior to scanning.

Acquisition Protocol

The **Siemens MAGNETOM Prisma** acquisition used a 32-channel head coil with 1.5 mm isotropic resolution, b-values of 1500 and 3000 s/mm^2, and 60 diffusion directions per shell. The **GE SIGNA MAGNUS** acquisition used a 32-channel head coil with 1.25 mm isotropic resolution, b-values of 1000 and 2000 s/mm^2, and 30 diffusion directions per shell. Both acquisitions used vendor-agnostic diffusion sequences (**??**) to minimize cross-scanner protocol differences. Each dataset included a corresponding noise scan acquired immediately prior to the diagnostic sequence, from which the coil-wise noise prior Σ_c was estimated using Eq. 2.1 and provided to the denoising network at inference time without any fine-tuning or domain adaptation.

High b-value data — particularly $b = 3000\,\mathrm{s/mm^2}$ on the Siemens scanner — exhibits substantially lower SNR due to exponential signal attenuation $e^{-b \cdot D}$, and displays strong spatially varying noise amplification from coil sensitivity fall-off at the image periphery. These conditions provide a rigorous test of the proposed framework's ability to suppress spatially structured noise without introducing over-smoothing or anatomical distortion.

Qualitative Results

Figures 2.2 and 2.3 show representative diffusion-weighted images at multiple b-values for the Siemens and GE scanners respectively, across all compared methods. For both scanners, our method effectively suppresses noise while preserving fine anatomical details including cortical boundaries, subcortical structures, and white matter tracts. The residual images — the difference between the noisy input and the denoised output — contain no visible anatomical structure on either scanner, confirming that the denoising is accurate and exclusive: only noise is removed with no signal leakage. Statistical methods including MP-

PCA and Hybrid-PCA leave residual structured noise particularly at high b-values, where their low-rank assumptions are less well-satisfied. Learning-based methods without explicit noise conditioning (Restormer, NAFNet) produce visible over-smoothing in peripheral regions, consistent with the shouting effect described in Section 2.3.2.

Figures 2.4 and 2.5 show the corresponding fractional anisotropy (FA) maps derived from the denoised DWI volumes. FA maps provide a particularly informative evaluation of downstream diffusion metric quality, as errors in denoising — whether residual noise or over-smoothing — propagate directly into the estimated diffusion tensor and its scalar indices. Our denoised FA maps on both scanners retain sharp white matter boundaries and realistic diffusion contrast, with fiber orientation estimates consistent with known neuroanatomy. In contrast, statistical baselines produce noisier FA estimates in peripheral white matter regions where coil sensitivity is lower, and learning-based baselines without noise conditioning show over-smoothed FA maps that blur fine white matter structures.

Quantitative Comparison

Table 2.1 reports residual noise levels for all methods using the σ_{MAD} estimator (Rajan *et al.*, 2010), computed on DWI volumes within each diffusion shell (excluding $b = 0$). Our method achieves the lowest residual noise across all four scanner–b-value combinations. The improvement is most pronounced at high b-values: at $b = 3000\,\text{s/mm}^2$ on the Siemens scanner, our method reduces residual noise by 23% relative to the next best-performing baseline (NAFNet), where the benefit of explicit spatial noise conditioning is most critical. Restormer, despite its strong performance on natural image benchmarks, ranks last among learning-based methods across all conditions — a direct consequence of the shouting effect that our FiLM conditioning is specifically designed to suppress. Performance is consistent across both scanner platforms, confirming that the measured noise prior Σ_c provides a scanner-agnostic conditioning signal that transfers across manufacturer boundaries without any retraining.

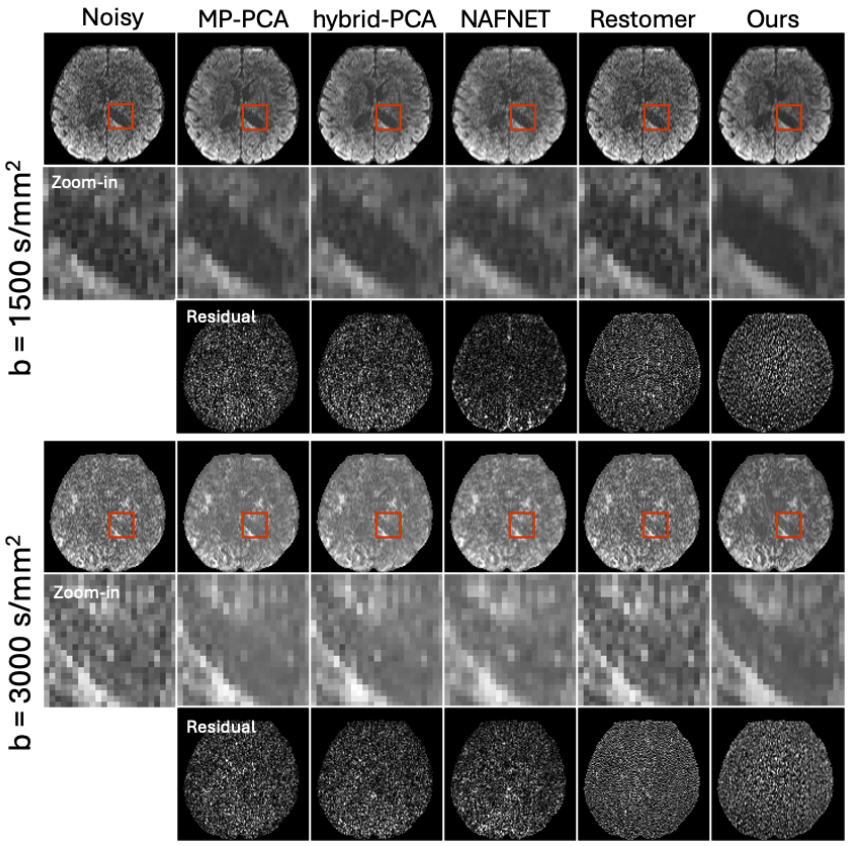


Figure 2.2: ***Siemens Scanner — Diffusion-Weighted Images.*** *Representative DWIs at $b = 1500$ and $b = 3000\,s/mm^2$ across all compared methods. Our method suppresses noise effectively at both b-values while preserving fine anatomical detail, with no structure visible in the residuals.*

2.4.3 Synthetic Noise Experiments

To evaluate the proposed method under controlled conditions where ground-truth clean images are available, we conduct experiments on a synthetic noise benchmark derived from the fastMRI brain dataset (Zbontar *et al.*, 2018). We randomly sample 100 slices from the fully sampled fastMRI volumes and corrupt them with spatially varying Gaussian noise at four input SNR levels: 5, 10, 15, and 20. Noise maps are generated following the coil simulation procedure described in Section 2.3.4, producing spatially heterogeneous noise patterns that mimic real coil sensitivity profiles. We compare against MP-PCA (Veraart *et al.*, 2016), Hybrid-PCA (Neto Henriques *et al.*, 2022), and NAFNet (Chen *et al.*, 2022), along with the noisy input as a lower bound. All baselines operate in the standard blind denoising

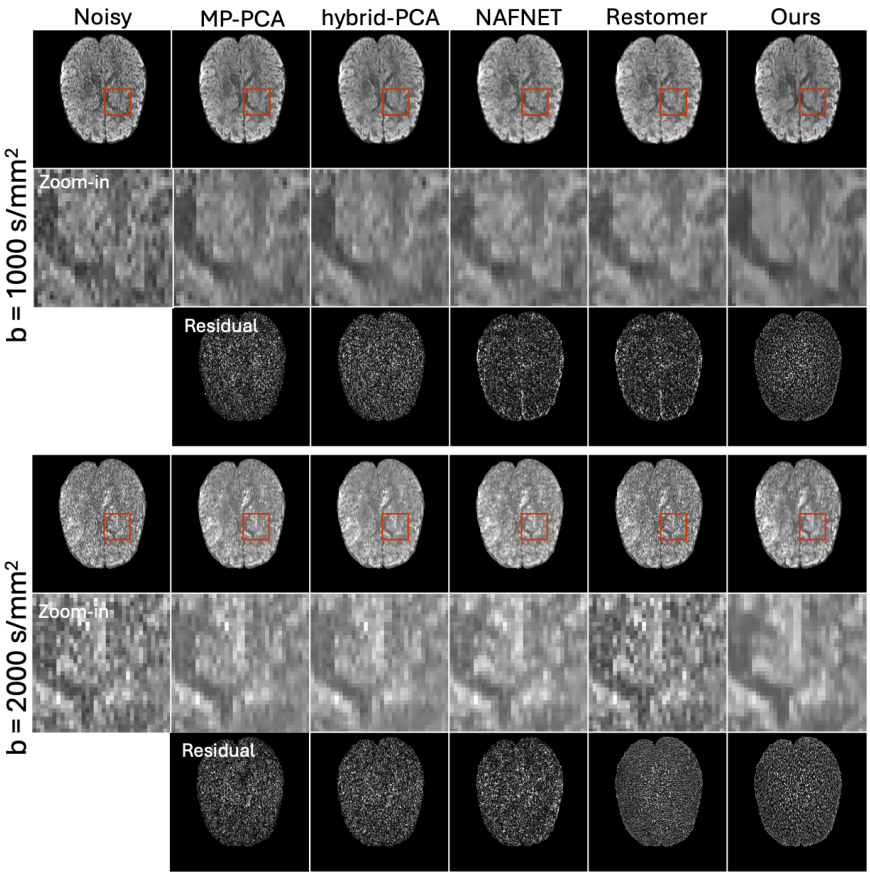


Figure 2.3: ***GE Scanner — Diffusion-Weighted Images.*** *Representative DWIs at $b = 1000$ and $b = 2000\,s/mm^2$, demonstrating zero-shot generalization to an unseen scanner platform with consistent noise suppression and preserved anatomical boundaries.*

setting without access to the noise prior, while our method receives the ground-truth noise map Σ_c at inference time.

Results

Table 2.2 reports PSNR, SSIM, and LPIPS across all four SNR levels. Our method achieves the highest PSNR at every SNR level by a substantial margin. The advantage is most pronounced at low SNR: at SNR $= 5$, our method achieves 21.47 dB compared to 11.90 dB for the next best-performing baseline (Hybrid-PCA), a gap of nearly 10 dB. This extreme low-SNR regime is where the explicit noise prior is most valuable — when the image is severely corrupted, knowing precisely where and how much noise is present allows the

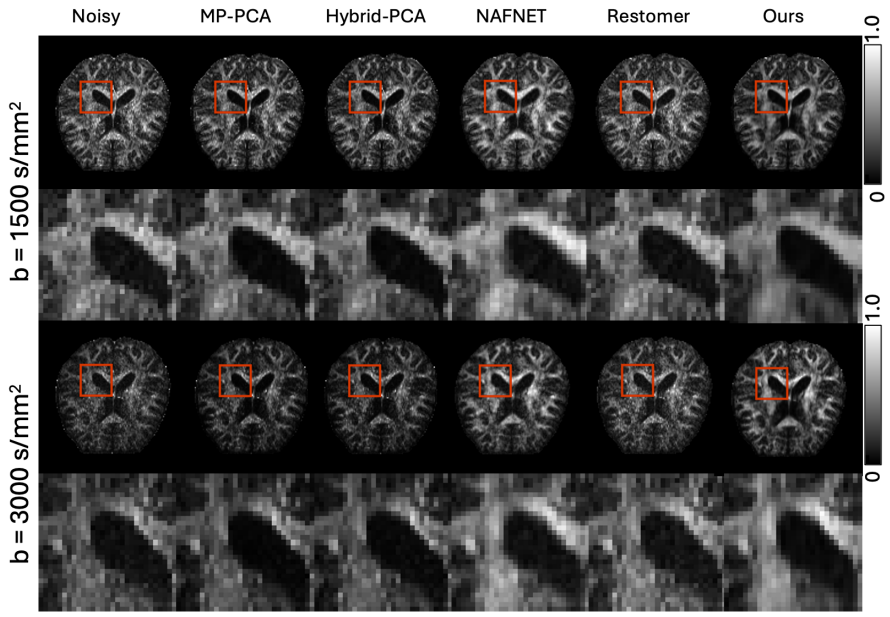


Figure 2.4: ***Siemens Scanner — Fractional Anisotropy Maps.*** *FA maps derived from denoised DWI volumes. Our method produces sharper white matter boundaries and more stable diffusion contrast compared to all baselines, with no loss of fine structural detail.*

Method	**Siemens** b=1500	**Siemens** b=3000	**GE** b=1000	**GE** b=2000
Noisy	0.0486	0.0361	0.0543	0.0485
MP–PCA	0.0325	0.0170	0.0349	0.0324
Hybrid–PCA	0.0293	0.0182	0.0316	0.0293
Restormer	0.0454	0.0327	0.0541	0.0454
NAFNet	0.0260	0.0179	0.0296	0.0260
Ours	**0.0253**	**0.0138**	**0.0260**	**0.0253**

Table 2.1: ***Estimated residual noise levels*** *(σ_{MAD}) averaged across DWIs within each diffusion shell (excluding b=0). Lower values indicate lower residual noise.*

network to make far more informed denoising decisions than any blind approach can.

Two observations are particularly noteworthy. First, NAFNet performs comparably to the noisy input at SNR $= 5$, confirming the shouting effect described in Section 2.3.2: at low SNR, the attention mechanism is dominated by high-variance regions, producing artifacts rather than clean reconstructions. Our FiLM conditioning suppresses this failure mode by normalizing feature magnitudes before the attention calculation. Second, statistical methods (MP-PCA, Hybrid-PCA) perform competitively at higher SNR levels (15 and 20), where their low-rank assumptions are better satisfied. However, they degrade sharply at low SNR,

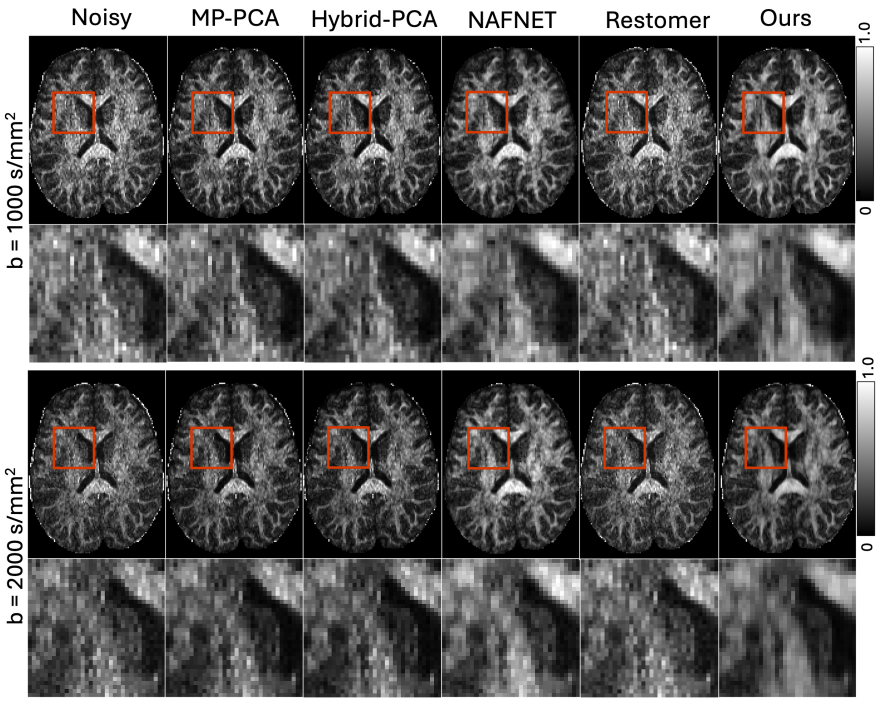


Figure 2.5: ***GE Scanner — Fractional Anisotropy Maps.*** *FA maps on the GE platform showing that improvements in downstream diffusion metrics transfer across scanner manufacturers without any retraining or fine-tuning.*

where noise dominates the principal component spectrum and the low-rank assumption breaks down.

In terms of SSIM and LPIPS, our method achieves the best or near-best performance across all SNR levels. At higher SNR (15 and 20), NAFNet and Hybrid-PCA achieve marginally better LPIPS, suggesting that blind denoisers can produce perceptually smoother outputs when noise is mild. However, this comes at the cost of substantially lower PSNR, indicating over-smoothing of fine structural detail. The consistency of our improvements across all four SNR levels demonstrates that explicit noise conditioning provides robust benefits regardless of degradation severity.

Method	PSNR↑	SSIM↑	LPIPS↓
SNR = 5			
Noisy Input	6.674	0.0345	1.1808
MP-PCA	10.345	0.9188	0.1079
Hybrid-PCA	11.901	0.9226	0.1022
NAFNet	6.803	0.9373	0.0742
Ours	**21.473**	**0.9624**	**0.0386**
SNR = 10			
Noisy Input	15.557	0.0705	0.9177
MP-PCA	19.205	0.9541	0.0543
Hybrid-PCA	19.560	0.9563	0.0489
NAFNet	15.765	0.9699	0.0343
Ours	**24.612**	**0.9720**	**0.0356**
SNR = 15			
Noisy Input	20.615	0.0908	0.7651
MP-PCA	23.483	0.9710	0.0285
Hybrid-PCA	23.528	0.9721	0.0260
NAFNet	20.522	0.9816	**0.0201**
Ours	**26.600**	**0.9787**	0.0276
SNR = 20			
Noisy Input	24.015	0.1035	0.6348
MP-PCA	26.217	0.9803	0.0161
Hybrid-PCA	26.180	0.9809	**0.0151**
NAFNet	23.263	0.9865	0.0160
Ours	**28.098**	**0.9837**	0.0217

Table 2.2: ***Denoising performance on synthetic fastMRI data*** *across four SNR levels. Metrics reported as mean PSNR (dB), SSIM, and LPIPS over 100 test slices. Bold indicates best performance per SNR level.*

2.5 Discussion

2.5.1 Summary of Contributions

This chapter presented a scanner-adaptive coil-level denoising framework for diffusion MRI that conditions on an explicitly measured, spatially resolved noise prior. The results across both synthetic and in-vivo experiments consistently demonstrate that explicit noise conditioning — when designed to propagate throughout the feature hierarchy rather than being limited to the input layer — substantially outperforms both classical statistical methods and learning-based blind denoisers, particularly under severe noise conditions.

2.5.2 Advantages of Explicit Noise Conditioning

The central advantage of our approach is that it decouples the denoising network from any particular noise distribution. Conventional deep denoisers implicitly encode noise characteristics in their weights during training, which means their behavior is fixed at deployment time. When the test distribution differs from the training distribution — as it invariably does across scanner manufacturers, field strengths, and coil configurations — performance degrades with no mechanism for correction. By contrast, our method receives the scanner's actual noise statistics at inference time through the measured prior Σ_c. The network does not need to have seen this noise profile during training; it simply conditions on it. This shift from implicit to explicit noise representation is what enables zero-shot generalization: a single model trained on synthetic data from a simulated coil array generalizes to both a Siemens Prisma and a GE MAGNUS without any fine-tuning, target-domain data, or distribution matching.

The spatially varying nature of the conditioning is equally important. A scalar noise level per image — as used in FFDNet (Zhang *et al.*, 2018a) — would not capture the coil-induced spatial structure that varies significantly across the image plane. By conditioning on a full $H \times W$ noise map per coil, the network can locally adapt its denoising strength, suppressing noise aggressively in high-variance peripheral regions while preserving fine structural detail in low-noise central regions. The FiLM modulation mechanism is what makes this local adaptation possible throughout the entire feature hierarchy, not just at the input.

2.5.3 Training Without MRI Data

A practically significant advantage of the proposed framework is that it requires no real MRI data for training. The physics-based synthetic data generation pipeline — simulating coil sensitivity maps via the Biot-Savart model and constructing spatially varying noise envelopes from natural images — produces a diverse and physically consistent training distribution that generalizes to real in-vivo acquisitions. This is made possible precisely because the noise prior is explicit: since the network is told the noise statistics at inference time, it does

not need to internalize scanner-specific noise patterns from real data during training. This is a meaningful practical advantage. Acquiring large datasets of paired noisy and clean dMRI data is prohibitively expensive — clean ground truth does not exist in practice, and high-SNR references require impractically long scan times. Our synthetic training strategy bypasses this bottleneck entirely, making the framework immediately deployable on any scanner for which a noise scan is available, without the need for site-specific data collection or model retraining.

2.5.4 Limitations

Despite these advantages, several limitations warrant acknowledgment. First, the framework requires a noise scan to be acquired alongside the diagnostic sequence. While this scan takes only a few seconds and is already standard in many research protocols, it represents an additional step that may not always be available in retrospective studies or time-constrained clinical settings. Second, the current implementation denoises each coil independently, ignoring potential cross-coil correlations that could provide additional signal for noise estimation. Incorporating cross-coil information into the conditioning signal may further improve performance. Third, while our synthetic training strategy generalizes well in practice, the Biot-Savart coil simulation is an approximation of the true electromagnetic environment inside the bore, and may not capture all hardware-specific noise sources such as digitization noise or gradient interference.

2.5.5 Future Directions

Several directions present natural extensions of this work. The most immediate is to investigate learned noise prior estimation — training an auxiliary network to predict Σ_c from the dMRI data itself, eliminating the requirement for a separate noise scan. Self-supervised approaches such as Noise2Noise (Lehtinen *et al.*, 2018) or Patch2Self (Fadnavis *et al.*, 2020) could provide the training signal for such a network without requiring clean references.

A second direction is to extend the coil-level processing framework to three-dimensional volumes, exploiting redundancy across diffusion directions as well as spatial dimensions. The current 2D patch-based approach discards through-plane correlations that carry additional signal in multi-shell acquisitions.

Finally, the conditioning design developed in this chapter — modular FiLM blocks that propagate a spatial prior throughout a hierarchical encoder-decoder — is architecture-agnostic and not specific to MRI. Future work could investigate its application to other physical imaging modalities where spatially varying noise is measurable from hardware, such as low-light photography, electron microscopy, or ultrasound, where analogous calibration acquisitions could provide explicit noise priors at inference time.

Chapter 3

Structured Conditioning for Diffusion-Based Reconstruction

3.1 Introduction and Motivation

Accelerated MRI reconstruction is an instance of a compressed sensing inverse problem: the forward operator A retains only a fraction of k-space measurements, and the reconstruction must recover the full image from a severely underdetermined system. As established in Chapter 1, diffusion models have emerged as powerful learned priors for this setting, with score-based posterior sampling (Chung and Ye, 2022) and its accelerated variants (Chung *et al.*, 2023) demonstrating state-of-the-art reconstruction fidelity across a range of undersampling patterns. These methods, however, treat the diffusion prior as entirely unconditional — sampling from $p(\mathbf{x})$ rather than $p(\mathbf{x} \mid \mathbf{c})$ — and leave a rich source of domain knowledge entirely unexploited.

Every MRI scan is accompanied by structured acquisition metadata: imaging parameters including repetition time (TR), echo time (TE), inversion time (TI), and flip angle; sequence and contrast type; anatomical region; and often patient-level context. These parameters are not incidental — they directly determine tissue contrast and constrain the space of plausible image appearances. A T1-weighted scan with short TR and short TE looks fundamentally

different from a T2-weighted FLAIR scan, even of the same anatomy. A diffusion model that is unaware of these parameters must maintain a prior broad enough to cover all possible contrasts simultaneously, reducing its specificity as a prior for any given acquisition. Conditioning on metadata narrows this prior to the plausible appearances for the specific acquisition at hand, directly reducing reconstruction ambiguity.

ContextMRI (Chung *et al.*, 2025) was the first diffusion model for MRI reconstruction to incorporate acquisition metadata, encoding it as free-form text via a frozen CLIP text encoder (Radford *et al.*, 2021). While this approach improved over the unconditioned baseline, the reported gains were modest — smaller than the variability of the evaluation metrics. We identify two reasons for this failure. First, the CLIP encoder was not trained on specialized medical text prompts and does not understand the domain-specific semantics of acquisition parameters. Second, and more fundamentally, the encoder inadequately captures numerical metadata values: representing TR as a string of digits in a text prompt gives the model no meaningful representation of its physical magnitude or its relationship to neighboring values. A prompt with TR = 500 and one with TR = 5000 differ by only one character, yet correspond to fundamentally different tissue contrasts. The narrow metadata distribution of the fastMRI dataset (Appendix A.1) further masks this failure: when most scans share the same TR and TE values, the model can achieve good reconstruction without attending to these parameters at all. We provide direct mechanistic evidence for this failure through cross-attention weight analysis in Appendix A.3, which shows that numerical metadata tokens receive near-zero attention weight regardless of their values, with conditioning effectively collapsed into the CLS summary token.

The failure of CLIP-based encoding is fundamentally a representation problem: heterogeneous structured data — a mix of categorical labels and continuous numerical parameters — is being forced through an encoder designed for unstructured natural language. The fix is to respect the type structure of the metadata. Categorical attributes such as anatomy, sequence type, and contrast are discrete and non-ordinal; they should be encoded via learnable discrete embedding tables. Continuous attributes such as TR, TE, and flip angle are

numerical — their magnitude matters, and nearby values should map to nearby embeddings — and should be encoded via sinusoidal positional encodings that preserve ordinal structure and enable interpolation to unseen values. A subtler challenge is that even with better embeddings, the model may learn to ignore the conditioning signal if it can achieve good reconstruction from image content alone, since the fastMRI metadata distribution is narrow. We address this through an out-of-distribution (OOD) training strategy using cardiac MRI data that forces active exploitation of the conditioning signal.

Contributions. This chapter makes the following contributions to the problem of conditioning design for diffusion-based inverse problem solvers:

- We propose type-aware embeddings — discrete tables for categorical fields and sinusoidal encodings for continuous ones — fused into the diffusion model via cross-attention with per-field classifier-free guidance dropout.
- We introduce an OOD training strategy that forces active exploitation of conditioning signals, improving both in-distribution reconstruction and cross-domain generalization.
- We provide comprehensive ablation studies that isolate the contribution of each metadata field, offering interpretable insights into which acquisition parameters most constrain the reconstruction prior.

3.2 Related Work

3.2.1 Diffusion Models for Image Reconstruction

Chapter 1 covered the score-based framework for inverse problems and the key reconstruction methods — ScoreMRI (Chung and Ye, 2022) and DDS (Chung *et al.*, 2023) — in detail. Relevant to our experimental setup, Shen et al. (Shen *et al.*, 2024) showed that averaging multiple independent stochastic reconstructions from the same undersampled input reduces

variance and improves overall quality, exploiting the probabilistic nature of diffusion models. We adopt this ensemble strategy in Section 3.4.7 and evaluate its interaction with our metadata conditioning approach.

3.2.2 Conditioning in Generative Models

The use of structured side information to condition generative models has been explored in the medical imaging literature, primarily in the context of image synthesis rather than reconstruction. Pinaya et al. (Pinaya *et al.*, 2022) conditioned a latent diffusion model on demographic covariates (age, sex, ventricular volume) for brain MRI synthesis, demonstrating that structured covariates steer macroscopic image attributes. SynthBrainGrow (**?**) extended this to longitudinal aging simulation, and Med-DDPM (**?**) incorporated tumor masks for controllable 3D brain synthesis. These works share the insight that structured conditioning improves generative specificity, but condition on macroscopic semantic variables rather than precise numerical acquisition parameters where the encoding precision matters. Our work addresses this gap by focusing specifically on the encoding design for heterogeneous mixed-type metadata.

In the natural image domain, text-conditioned latent diffusion models (Rombach *et al.*, 2022) have demonstrated the power of conditioning for high-quality image generation, but rely on CLIP encoders designed for natural language — a choice that transfers poorly to structured numerical data as we demonstrate in Section 3.1.

3.3 Method

3.3.1 Data Curation

Our primary dataset is derived from the fastMRI brain dataset, comprising 6,970 fully sampled brain MRI scans released by NYU Langone Hospital (Zbontar *et al.*, 2018). Due to computational constraints, we train on a subset of 450 patient studies (7,185 scans) and evaluate on a separate held-out test set of 150 patients (2,957 scans).

Following the ContextMRI pipeline (Chung *et al.*, 2025), each raw multi-coil fastMRI volume undergoes an inverse Fourier transform and is combined with coil-sensitivity maps estimated via ESPIRiT (Uecker *et al.*, 2014), resulting in minimum-variance unbiased (MVUE) complex-valued images for each slice. These complex images are center-cropped to 320×320 pixels to remove zero-padding from k-space. Real and imaginary components are stacked into a two-channel tensor, and intensities are normalized by the 99th percentile magnitude within each slice, yielding values in an approximate range of $[-1.5, 1.5]$. This normalization stabilizes training and facilitates direct pixel-space diffusion modeling.

3.3.2 Background: ContextMRI and its Limitations

We build upon ContextMRI (Chung *et al.*, 2025), the first diffusion model for MRI reconstruction conditioned on acquisition metadata. ContextMRI transforms metadata into textual prompts and encodes them using a pre-trained, frozen CLIP text encoder (Radford *et al.*, 2021). For example, an image acquired with contrast `AX R1 POST_FBB` and TR of 3150 is represented as the prompt "`AX R1 POST_FBB, TR: 3150`".

Despite integrating metadata into the generation process, the improvements reported in ContextMRI are minimal — the performance gap between conditioned and unconditioned reconstructions is smaller than the variability of the evaluation metrics. We identify two reasons for this failure. First, the CLIP encoder was not trained on specialized medical text prompts and does not understand the domain-specific semantics of acquisition parameters. Second, and more fundamentally, the encoder inadequately captures numerical metadata values: representing TR as a string of digits in a text prompt gives the model no meaningful representation of its physical magnitude or its relationship to neighboring values. A prompt with TR $= 500$ and one with TR $= 5000$ differ by only one character, yet correspond to fundamentally different tissue contrasts.

We confirm this failure quantitatively through cross-attention weight analysis, reported in full in Appendix A.3. Probing the mid-block cross-attention layer of the diffusion UNet during inference reveals that the CLS token and anatomy token dominate the attention

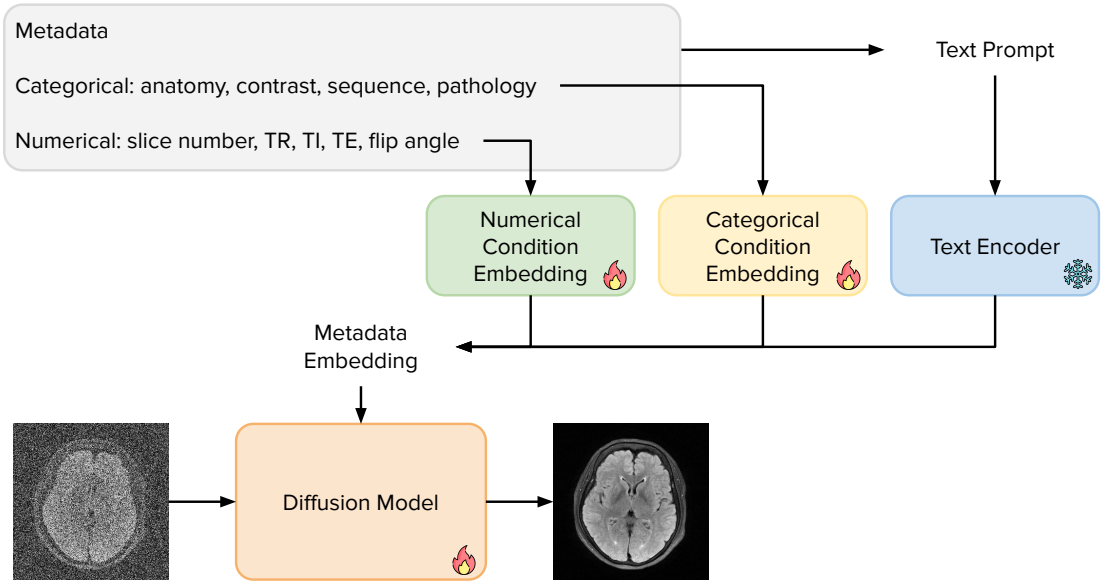


Figure 3.1: ***Proposed metadata conditioning framework.*** *The original ContextMRI framework (right branch) encodes all metadata as a single free-form text prompt via a frozen CLIP encoder. Our approach (left and center branches) introduces dedicated categorical condition embeddings for discrete attributes (anatomy, contrast, sequence, pathology) and numerical condition embeddings for continuous parameters (TR, TE, TI, flip angle, slice number), which are fused with the CLIP text embedding before being passed to the diffusion model.*

weights, while tokens corresponding to TR, TE, TI, and flip angle receive near-zero weight regardless of their numerical values. The CLIP encoder effectively compresses all numerical metadata into the CLS summary token, discarding the individual parameter values entirely. Figure 3.1 illustrates our proposed framework alongside the original ContextMRI design, with dedicated numerical and categorical condition embedding modules that directly address this failure mode. The embedding design is described in detail in Section 3.3.3.

3.3.3 Type-Aware Metadata Embeddings

The core insight motivating our approach is that metadata fields differ in their type, and the encoding must respect that structure. We distinguish two categories of metadata fields.

Categorical and Numerical Condition Embeddings

Categorical attributes — anatomy, contrast type, sequence type, and pathology — take values from a finite discrete set with no ordinal structure. Each categorical attribute a_k with N_k possible values is mapped by a learnable embedding table $E_k \in \mathbb{R}^{N_k \times d}$, producing a d-dimensional vector. This allows each category to occupy an independently learned position in embedding space without imposing any spurious ordering.

Continuous attributes — TR, TE, TI, flip angle, and slice number — are numerical parameters whose magnitude carries physical meaning. Each continuous attribute v_l is first normalized to $[0, 1]$ and then encoded with sinusoidal positional embeddings, analogous to timestep encoding in diffusion models:

$$\phi(v)_i = \begin{cases} \sin\left(v/10000^{2i/d}\right) & i \text{ even} \\ \cos\left(v/10000^{2(i-1)/d}\right) & i \text{ odd} \end{cases} \tag{3.1}$$

This encoding preserves ordinal structure — TR $= 5000$ encodes as more distinct from TR $= 500$ than from TR $= 4900$ — and enables smooth interpolation to acquisition parameter values not seen during training. Distributions of both categorical and continuous metadata fields in the fastMRI brain dataset are provided in Appendix **??**, illustrating the narrow clustering that motivates the OOD training strategy in Section 3.3.4.

Fusion Strategies

We propose two methods for integrating the type-aware embeddings with the existing CLIP text embeddings from ContextMRI, maintaining backward compatibility with the pre-trained checkpoint.

Additive fusion sums all fine-grained embeddings element-wise and adds the result to the CLIP text embedding, preserving the sequence length expected by the cross-attention layers without any architectural modification.

Concatenated fusion appends the fine-grained embeddings sequentially to the text embedding sequence, separated by a special `[SEP]` token. This requires greater memory

and increases the cross-attention sequence length, but allows the network to attend to individual metadata fields independently rather than receiving their sum.

Classifier-Free Guidance with Per-Field Dropout

Following ContextMRI, we employ classifier-free guidance (CFG) (Ho and Salimans, 2022) to control conditioning strength at inference time. Crucially, each fine-grained embedding is independently dropped during training with probability p_{drop}, rather than dropping all metadata simultaneously. This per-field dropout serves two purposes: it simulates real-world scenarios with incomplete metadata, and it enables per-field ablation studies at inference time by selectively dropping individual fields. The CFG inference rule follows Eq. 1.11 from Chapter 1, with the guidance scale w tuned on the validation set.

3.3.4 Out-of-Distribution Training Strategy

Even with type-aware embeddings, the model may learn to ignore the conditioning signal if it can achieve adequate reconstruction from image content alone. This is a genuine risk for the fastMRI dataset, where TR, TE, and flip angle cluster at only a few discrete values (Appendix A.1): the model can perform well by implicitly learning the dominant contrast mode without attending to metadata. We address this with an out-of-distribution (OOD) training strategy that forces the model to actively exploit the conditioning signal.

The key idea is that when image content is destroyed by heavy noise, metadata becomes the only available source of information about the reconstruction prior. We therefore introduce training samples whose anatomical content is completely masked, leaving only the metadata embedding as a useful signal. If the model learns to condition effectively on metadata in these extreme cases, the learned mappings transfer to the in-distribution reconstruction task where image content is only partially corrupted.

For OOD samples, we use the CMRxRecon cardiac cine dataset (Wang *et al.*, 2024), which provides raw k-space images of the heart from 300 healthy volunteers scanned at 3T with both short-axis (SAX) and long-axis (LAX) views. The cardiac metadata distribution lies

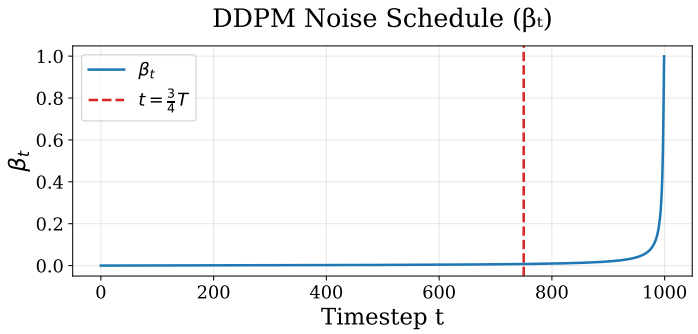


Figure 3.2: ***DDPM noise schedule.*** *The red dashed line at $\frac{3}{4}T_{\max}$ marks the minimum timestep at which OOD cardiac slices are noised during training. At timesteps above this threshold, $\bar{\alpha}_t$ is sufficiently small that anatomical content is effectively destroyed, leaving only the metadata embedding as a useful conditioning signal.*

well outside the fastMRI brain distribution: TR spans 3.6–50 ms, TE spans 1.1–1.6 ms, and flip angles range from 12 to 60 degrees (Appendix A.2). This distributional distance ensures that the model cannot rely on familiar image statistics when processing OOD samples and must instead attend to the metadata embeddings.

During training, cardiac slices are sampled at a 1:5 ratio to brain slices and noised to timesteps between $\frac{3}{4}T_{\max}$ and $T_{\max}$, as illustrated by the noise schedule in Figure 3.2. At these high timesteps, the DDPM forward process (Eq. 1.5) reduces $\bar{\alpha}_t$ to near zero, driving the noisy image $\mathbf{x}_t$ toward pure Gaussian noise and effectively masking all anatomical content. Figure 3.3 shows representative brain and cardiac slices before and after this noising process: brain features may remain partially detectable at moderate timesteps, while cardiac features are always completely obscured, ensuring full reliance on metadata. The model trained with this strategy using additive fusion is referred to as the **OOD + Additive** model throughout our experiments.

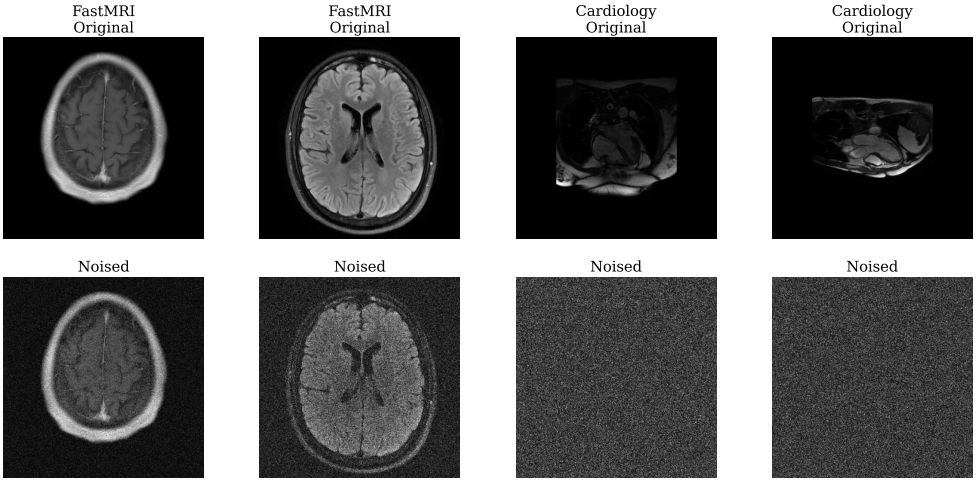


Figure 3.3: ***Visualization of the OOD noising process.*** *Top row: original brain (in-distribution) and cardiac (OOD) slices. Bottom row: corresponding noised images. Brain features remain partially detectable after noising, while cardiac features are always completely obscured, ensuring the model must rely on metadata embeddings when processing OOD cardiac samples.*

3.4 Experiments

3.4.1 Experimental Setup

All models are fine-tuned from the pre-trained ContextMRI checkpoint rather than trained from scratch, allowing us to leverage the robust reconstruction capabilities of the baseline while isolating the effect of our proposed conditioning improvements. Training uses a batch size of 4, a constant learning rate of 5×10^{-5}, and runs for 10,000 steps on two NVIDIA Volta V100 GPUs. All variants share identical hyperparameters to ensure fair comparison.

Our experimental evaluation focuses on the $4\times$ acceleration task, which retains only 25% of k-space measurements through uniform 1D undersampling, simulating a fourfold reduction in acquisition time. All methods use 100 DDIM inference steps, matching the ContextMRI methodology. Evaluation metrics — PSNR, SSIM (Wang *et al.*, 2004), and LPIPS (Zhang *et al.*, 2018b) — are computed on a per-slice basis and aggregated across the test set. All results are reported at CFG $= 1.0$ unless otherwise indicated.

Method	PSNR↑	SSIM·10^2 ↑	LPIPS·10^2 ↓
ContextMRI	28.39 ± 6.23	65.84 ± 22.05	33.31 ± 11.87
Additive Embedding	30.32 ± 6.25	78.64 ± 22.45	26.16 ± 12.27
Concatenated Embedding	30.23 ± 6.82	79.00 ± 23.98	$\mathbf{23.51 \pm 13.78}$
OOD + Additive	$\mathbf{30.40 \pm 6.62}$	$\mathbf{79.04 \pm 21.97}$	26.20 ± 11.98

Table 3.1: ***Quantitative results on the fastMRI brain dataset*** *(uniform 1D mask, 4× acceleration, CFG = 1.0). Bold indicates best performance per metric.*

3.4.2 Main Results

Table 3.1 reports quantitative results for all methods on the fastMRI brain test set. Both the Additive and Concatenated Embedding variants substantially outperform the ContextMRI baseline across all three metrics, confirming that type-aware embeddings provide a significantly more effective conditioning signal than CLIP-based text encoding. The Additive method achieves marginally higher PSNR, suggesting slightly better pixel-level fidelity, while the Concatenated method achieves stronger LPIPS, indicating superior perceptual quality. The two strategies are otherwise comparable, consistent with the interpretation that the primary gain comes from type-aware encoding rather than the specific fusion mechanism.

The OOD + Additive model achieves the best overall performance, with the highest PSNR and SSIM among all tested methods. The improvement over the plain Additive model confirms that the OOD training strategy provides an additional benefit beyond the embedding design alone, by compelling the network to actively exploit conditioning signals rather than relying passively on image content. All subsequent experiments use the OOD + Additive model.

3.4.3 Effect of Conditioning Strength

Table 3.2 reports reconstruction performance of the OOD + Additive model as the CFG scale w is varied from 0 (unconditioned) to 3. Even mild guidance at $w = 1.0$ yields a substantial improvement over the unconditioned baseline — a gain of over 2 PSNR points and more

CFG	PSNR↑	SSIM·10^2 ↑	LPIPS·10^2 ↓
0.0 (Uncond)	28.36 ± 6.60	67.62 ± 20.83	33.15 ± 11.33
1.0	30.40 ± 6.62	79.04 ± 21.97	$\mathbf{26.20 \pm 11.98}$
2.0	30.37 ± 6.53	$\mathbf{79.19 \pm 21.34}$	26.24 ± 11.49
3.0	$\mathbf{30.50 \pm 6.42}$	77.97 ± 23.08	26.26 ± 11.44

Table 3.2: ***Effect of CFG scale on reconstruction quality*** *for the OOD + Additive model. Bold indicates best performance per metric.*

than 11 SSIM points — confirming that the type-aware embeddings carry meaningful conditioning information that the model actively exploits. Performance remains stable as w increases from 1 to 3, with PSNR showing a marginal further improvement at $w = 3.0$ while SSIM peaks at $w = 2.0$. This stability indicates that our conditioning signal is well-calibrated: stronger guidance does not destabilize reconstruction quality, and moderate values of $w \in [1, 2]$ offer a robust operating point.

3.4.4 Per-Field Metadata Ablations

We conduct two complementary ablation studies to isolate the contribution of each metadata field to reconstruction quality.

Dropping One Field at a Time

Table 3.3 reports performance when one metadata field is dropped at inference while all others are retained. Dropping semantic fields — Slice Number, Anatomy, and TI — produces the largest performance degradation, suggesting these fields provide the most non-redundant information to the reconstruction prior. Anatomy and sequence type likely inform the model about global structural priors that are difficult to infer from the undersampled measurement alone. Continuous acquisition parameters such as TR, TE, and Flip Angle produce smaller but consistent degradations when dropped, suggesting they contribute complementary refinements to the reconstruction that become apparent only in combination with the semantic fields.

Dropped Field	PSNR↑	SSIM$\cdot 10^2$ ↑	LPIPS$\cdot 10^2$ ↓
Slice Number	29.52 ± 6.90	73.41 ± 26.21	28.43 ± 13.34
Anatomy	29.87 ± 6.82	74.53 ± 25.82	27.65 ± 12.92
Contrast	30.12 ± 6.46	74.88 ± 25.70	27.41 ± 12.75
Sequence	30.03 ± 6.58	74.85 ± 25.69	27.47 ± 12.71
TE	29.76 ± 6.77	74.19 ± 25.98	27.87 ± 13.13
TI	29.54 ± 6.97	73.31 ± 26.88	28.24 ± 13.66
TR	29.97 ± 6.70	74.48 ± 26.04	27.62 ± 12.95
Flip Angle	30.08 ± 6.66	74.98 ± 25.99	27.51 ± 13.09

Table 3.3: ***Drop-one-field ablation.*** *Performance when all metadata fields are used except the listed one. Larger degradation indicates greater non-redundant contribution.*

Conditioning on a Single Field

Table 3.4 reports performance when only one metadata field is retained at inference, with all others dropped. This complements the drop-one analysis by probing the standalone informativeness of each field. Semantic fields — Anatomy, Sequence, Slice Number, and Flip Angle — support strong reconstructions even in isolation, achieving SSIM above 75 and LPIPS close to the fully conditioned model. This suggests these fields encode rich structural priors about scan content that are largely sufficient on their own.

Continuous acquisition parameters TR, TE, and TI, by contrast, perform substantially worse in isolation, with SSIM declining by over 6 points and LPIPS increasing sharply. These fields appear to provide their greatest benefit in combination with other context — particularly Contrast, which may help the model identify the dominant contrast mode before the numerical parameters refine it further. Interestingly, Flip Angle achieves the strongest single-field performance, likely because it implicitly captures contrast and sequence information, making it highly predictive in isolation despite being relatively redundant when the full set of metadata is available.

Single Field	**PSNR**↑	**SSIM**$\cdot 10^2$ ↑	**LPIPS**$\cdot 10^2$ ↓
Slice Number	29.96 ± 6.21	75.45 ± 24.07	26.57 ± 11.68
Anatomy	29.90 ± 6.05	75.42 ± 23.43	26.52 ± 11.03
Contrast	29.71 ± 6.67	74.97 ± 24.57	26.79 ± 11.81
Sequence	29.93 ± 6.26	75.19 ± 24.08	26.53 ± 11.40
TE	28.90 ± 6.24	69.58 ± 19.74	32.27 ± 10.78
TI	28.56 ± 6.44	67.88 ± 21.00	33.10 ± 11.35
TR	28.68 ± 6.44	68.25 ± 20.89	32.90 ± 11.25
Flip Angle	$\mathbf{29.99 \pm 6.25}$	$\mathbf{75.59 \pm 23.93}$	$\mathbf{26.46 \pm 11.64}$

Table 3.4: ***Single-field ablation.*** *Performance when conditioning on only the listed field with all others dropped. Bold indicates best single-field performance.*

3.4.5 Qualitative Results

Figure 3.4 compares reconstructions from an unconditioned model — with no metadata embeddings — and our OOD + Additive model on a representative brain slice. For a broader comparison across multiple examples, Figure 3.5 provides two additional slices comparing ContextMRI directly against our OOD + Additive model. The metadata-conditioned reconstruction yields noticeably higher anatomical fidelity, particularly in regions prone to low SNR such as the ventricular system and gray-white matter boundaries. The improvement is most pronounced in areas where the undersampled measurement provides the least information, demonstrating that type-aware metadata conditioning effectively compensates for the ambiguity introduced by k-space subsampling.

Figure 3.5 compares ContextMRI and our OOD + Additive model across two representative brain slices. In both examples, the OOD-aware model produces reconstructions with improved structural detail and anatomical sharpness. The zoomed-in regions in the bottom row of each example highlight these differences most clearly: edges around ventricles and fine cortical structures are more crisply defined in our reconstructions, while the ContextMRI outputs show residual blurring in these high-frequency regions. These qualitative observations are consistent with the quantitative improvements reported in Table 3.1, and reinforce the interpretation that active metadata conditioning — forced by the OOD training

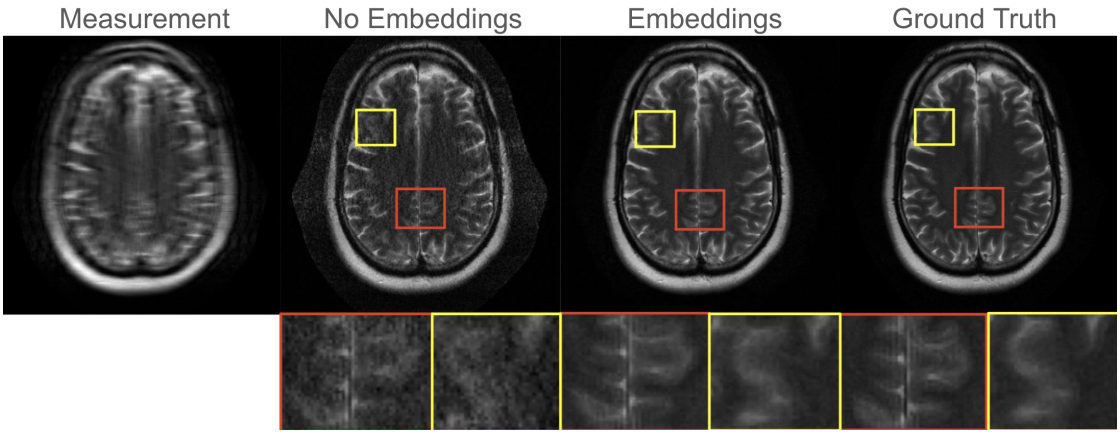


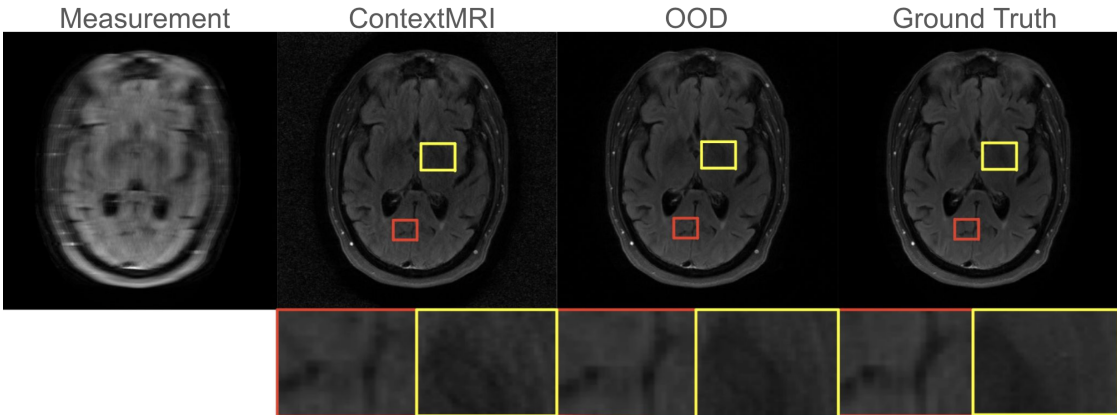
Figure 3.4: ***No embeddings vs. OOD + Additive.*** *Reconstruction of a representative brain slice comparing an unconditioned model (no metadata embeddings) against our proposed OOD + Additive model. The conditioned reconstruction shows sharper gray-white matter boundaries and improved fidelity in low-SNR regions near the ventricles.*

strategy — produces a more informed and spatially precise reconstruction prior.

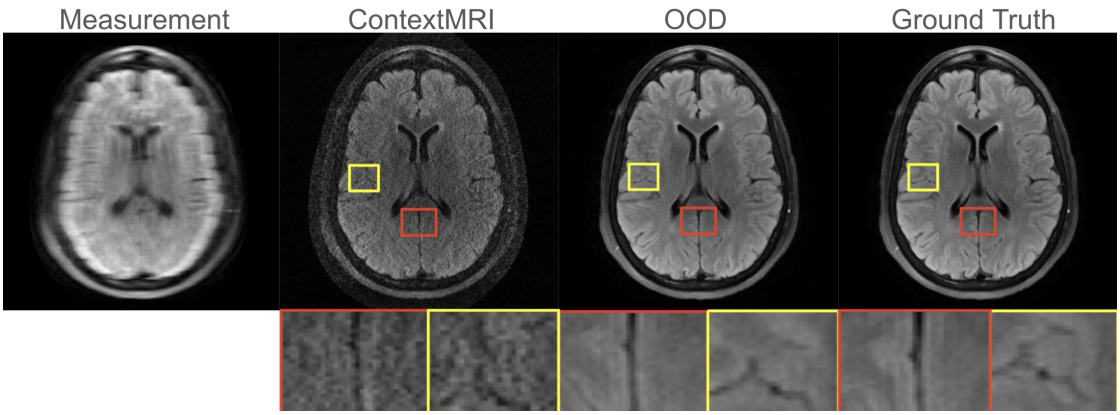
3.4.6 OOD Noise Level Ablation

Our primary OOD strategy noises cardiac samples at timesteps between $\frac{3}{4}T_{\max}$ and $T_{\max}$, destroying anatomical content to force metadata reliance. We also investigate a complementary strategy that restricts OOD noising to timesteps between 0 and $\frac{1}{4}T_{\max}$, motivated by the observation that diffusion models learn different image features at different noise levels — with low-noise timesteps particularly important for capturing fine structural details and tissue boundaries.

Table 3.5 shows that both strategies perform remarkably similarly across all three metrics, with differences well within the margins of statistical error. The high-noise approach achieves marginally better PSNR, while the low-noise approach shows slight improvements in SSIM and LPIPS. This robustness suggests that the benefit of OOD training is not sensitive to the specific noise level regime, and that practitioners can choose either strategy without meaningful performance trade-offs.



(a) *Example 1: ContextMRI vs. OOD + Additive.*



(b) *Example 2: ContextMRI vs. OOD + Additive.*

Figure 3.5: ***ContextMRI vs. OOD + Additive.*** *Reconstructions of two representative brain slices. Each panel shows, from left to right, the undersampled measurement, the ContextMRI reconstruction, the OOD + Additive reconstruction, and the ground truth. The bottom row of each example shows zoomed-in regions, highlighting crisper ventricular edges and finer cortical detail in the OOD + Additive outputs.*

3.4.7 Ensemble Reconstruction

Diffusion models are inherently stochastic, producing different reconstructions from the same undersampled input depending on the random seed. This leads to substantial per-sample variance in quality metrics (± 6 PSNR across slices). Following Shen et al. (Shen *et al.*, 2024), we address this by generating five independent reconstructions per acquisition and averaging them pixel-wise, reducing the impact of individual poor samples.

Table 3.6 shows that ensemble averaging yields large improvements over single-sample

OOD Noise Level	PSNR↑	SSIM·10^2 ↑	LPIPS·10^2 ↓
High Noise ($\frac{3}{4}T_{\max}$ to $T_{\max}$)	$\mathbf{30.40 \pm 6.62}$	79.04 ± 21.97	26.20 ± 11.98
Low Noise (0 to $\frac{1}{4}T_{\max}$)	30.33 ± 6.64	$\mathbf{79.29 \pm 22.23}$	$\mathbf{25.77 \pm 12.53}$

Table 3.5: ***OOD noise level ablation.*** *Comparing high-noise and low-noise OOD training strategies for the OOD + Additive model. Differences fall within statistical error across all metrics.*

Sampling Strategy	PSNR↑	SSIM·10^2 ↑	LPIPS·10^2 ↓
Single Sample	30.40 ± 6.62	79.04 ± 21.97	26.20 ± 11.98
5-Sample Ensemble	$\mathbf{34.60 \pm 7.11}$	$\mathbf{85.53 \pm 20.12}$	$\mathbf{17.32 \pm 11.51}$

Table 3.6: ***Ensemble reconstruction.*** *Pixel-wise averaging of five independent samples substantially improves all metrics at the cost of* $5\times$ *inference time.*

reconstruction: PSNR increases from 30.40 to 34.60 dB, SSIM from 79.04 to 85.53, and LPIPS nearly halves from 26.20 to 17.32. These gains come at a linear increase in inference time with the number of samples, and the pixel-wise averaging may introduce mild smoothing of fine structural details. Nevertheless, the overall quality improvement substantially outweighs these costs, and the ensemble strategy is recommended when inference time is not a hard constraint.

3.5 Discussion

3.5.1 Summary

This chapter presented a type-aware metadata conditioning framework for diffusion-based MRI reconstruction, demonstrating that respecting the structure of the conditioning signal — discrete embeddings for categorical fields, sinusoidal encodings for continuous ones — substantially outperforms generic text-based encoding. The OOD training strategy further improves performance by compelling the model to actively exploit conditioning signals rather than bypassing them. Together these contributions establish that conditioning design matters as much as model capacity for diffusion-based inverse problem solvers.

3.5.2 Advantages of Type-Aware Conditioning

The central advantage of our approach is that it treats the encoding of each metadata field as a design problem in its own right, rather than delegating all fields to a single general-purpose encoder. By matching the encoding to the type of each field, the resulting embedding preserves the information that is physically meaningful — ordinal structure for continuous parameters, independent identity for categorical ones — and makes that information available to the cross-attention mechanism in a form it can exploit. The consistent improvement across all three evaluation metrics, and across both in-distribution and out-of-distribution test sets, confirms that this is a principled and robust design choice rather than an artifact of a specific experimental configuration.

The OOD training strategy addresses a subtler but equally important problem: ensuring the model actually uses the conditioning signal it is given. The narrow metadata distribution of the fastMRI dataset creates an incentive for the model to ignore metadata and rely on image content alone. By introducing cardiac data whose anatomical content is destroyed at training time, we remove this shortcut and force the model to learn active mappings from metadata to reconstruction prior. The resulting model generalizes better not only within the fastMRI domain but also to unseen cardiac data, demonstrating that active conditioning is a transferable capability.

3.5.3 Limitations

Several limitations constrain the scope of the conclusions that can be drawn from these experiments. First, computational resources significantly constrained the experimental scale. Training was limited to two NVIDIA V100 GPUs and 10,000 steps, compared to the eight NVIDIA H100 GPUs and substantially larger effective batch sizes used in the original ContextMRI training. This gap in compute likely accounts for some of the remaining performance difference relative to a fully re-trained baseline, and the true benefit of type-aware conditioning may be larger at greater scale.

Second, both training and evaluation were conducted on subsets of the available data

due to storage and computational constraints. The generalizability of the results across the full diversity of patient populations, acquisition protocols, and pathologies present in complete fastMRI dataset remains to be established.

Third, while sinusoidal encodings provide a principled and parameter-free approach to numerical embedding, they are a fixed representation that does not adapt to the distribution of values seen during training. Learned numerical encodings — whether linear projections or non-linear mappings tailored to the specific range and distribution of each clinical parameter — may provide further improvements, particularly for parameters with highly non-uniform distributions such as TR and TI.

3.5.4 Future Directions

The most immediate extension is to scale these experiments to match the computational resources of the original ContextMRI training, which would provide a cleaner assessment of the benefit of type-aware conditioning at full scale. A second direction is to investigate learned numerical encoding strategies beyond sinusoidal embeddings, including adaptive encodings that account for the empirical distribution of each parameter in the training data.

Beyond reconstruction, the conditioning framework developed here applies naturally to other imaging inverse problems where structured acquisition metadata is available. Dynamic MRI and functional MRI (fMRI) introduce temporal metadata — frame index, repetition number, task condition — that could be encoded using the same type-aware approach. Multi-modal imaging scenarios, where metadata bridges different acquisition modalities, represent a further opportunity.

Finally, investigating metadata conditioning under clinically realistic conditions — where metadata fields may be missing, corrupted, or inconsistently recorded — is critical for real-world deployment. The per-field dropout strategy introduced during training provides a natural starting point for handling missing fields at inference, but robustness to systematic metadata errors or distribution shift in the parameter values themselves remains an open problem.

Chapter 4

Discussion

4.1 A Taxonomy of Acquisition-Context Conditioning

The two contributions in this thesis represent different points in a design space defined by two axes: the *structure* of the conditioning signal, and the *depth* at which it must be injected.

Chapter 2 conditions on a *dense spatial signal* — a noise map with the same spatial resolution as the image. Such signals must be modulated throughout the feature hierarchy; injecting them only at the input is insufficient because deeper layers lose spatial specificity through downsampling. FiLM-style modulation at every transformer block is the natural solution, providing locally adaptive feature normalization at every scale.

Chapter 3 conditions on a *heterogeneous structured signal* — a mix of categorical and continuous scalars. Such signals should not be concatenated or homogenized into a single embedding space, as CLIP does; they must be encoded in a type-aware manner that respects the semantics of each field. Cross-attention then provides a natural mechanism for the network to selectively attend to relevant metadata.

These two design principles are complementary and could be combined in a single model that simultaneously conditions on a measured noise prior and rich acquisition metadata — an avenue we discuss in Section 4.3.

4.2 Explicit Priors and Cross-Domain Generalization

A recurring finding across both chapters is that explicit conditioning signals improve cross-domain generalization. In Chapter 2, the model trained exclusively on synthetic natural images generalizes to in-vivo dMRI on both Siemens and GE platforms without any fine-tuning. The explicit noise prior is what makes this transfer possible: the model does not need to learn scanner-specific noise patterns from real data because they are provided directly at inference time through Σ_c.

In Chapter 3, the OOD + Additive model outperforms the ContextMRI baseline by a substantial margin on both in-distribution brain reconstruction and out-of-distribution cardiac data. The metadata conditioning provides a domain-invariant handle: when the anatomical image features are unfamiliar, the acquisition parameters still carry useful information about expected image structure that guides the reconstruction prior.

This suggests a broader design principle: *explicit conditioning reduces the implicit distributional burden on learned representations*, enabling better transfer to unseen acquisition settings without retraining or domain adaptation. When the conditioning signal accurately describes the degradation or acquisition context at test time, the network can adapt its behavior accordingly — a capability that is simply unavailable to models that encode this information implicitly in their weights.

4.3 Future Directions

4.3.1 Combining Noise and Metadata Conditioning

The most natural extension of this work is a reconstruction model that simultaneously conditions on a measured noise prior and structured metadata. In compressed sensing MRI, both signals are available and complementary: the noise prior characterizes the measurement uncertainty on the observation side, while the metadata characterizes the image prior on the signal side. Integrating both into a unified conditioning framework — using FiLM for the spatial noise prior and type-aware embeddings for metadata — would

address both sources of acquisition-specific information within a single model.

4.3.2 Type-Aware Embeddings Beyond MRI

The sinusoidal numerical encoding and type-aware embedding design are entirely general and do not depend on MRI-specific knowledge. Camera RAW image restoration is an immediate candidate: ISO, exposure time, and aperture are continuous acquisition parameters that shape the noise distribution in ways a CLIP encoder cannot represent precisely. Similarly, electron microscopy and ultrasound imaging both involve rich acquisition metadata whose structure would benefit from type-aware encoding. The OOD training strategy is equally transferable: any domain with a narrow in-distribution metadata distribution could benefit from OOD augmentation that forces active conditioning.

4.3.3 Learned Noise Prior Estimation

The noise scan used in Chapter 2 adds a brief but non-negligible overhead to the acquisition protocol. A natural extension is to learn to estimate Σ_c from the dMRI data itself, potentially through a self-supervised auxiliary network trained alongside the denoiser. Noise2Noise-style (Lehtinen *et al.*, 2018) approaches could provide the training signal without requiring a separate calibration acquisition, making the framework applicable to retrospective datasets where noise scans were not collected.

4.3.4 Extension to 3D and Temporal Imaging

Both methods currently operate on 2D slices. Extending coil-level denoising to 3D volumes would allow exploitation of through-plane correlations in diffusion encoding, potentially improving SNR further. For the metadata conditioning framework, dynamic MRI and functional MRI (fMRI) introduce temporal metadata — frame index, cardiac phase, task condition — that could be encoded using the same type-aware approach, extending the framework to the full richness of clinical imaging protocols.

4.4 Conclusion

This thesis has presented two contributions to the design of conditioning in deep networks for imaging inverse problems, both instantiated on MRI as a demanding and physics-rich testbed. The core finding is consistent across both contributions: when acquisition-side information is available and explicitly represented in a structure-respecting manner, it substantially improves both accuracy and generalization relative to methods that ignore it or encode it generically.

The design principles — hierarchical injection of dense spatial priors via feature modulation, and type-aware encoding of heterogeneous structured metadata via dedicated embedding modules — are grounded in the structure of the conditioning signal rather than in MRI-specific knowledge. A dense spatial prior must be propagated throughout the feature hierarchy at the appropriate spatial resolution; a heterogeneous structured signal must be encoded in a manner that preserves the semantics of each field. When these conditions are met, a single model can adapt to unseen acquisition settings at inference time without retraining, bridging the gap between the rich physical context that accompanies every medical image and the information that deep networks currently exploit.

The broader implication is that the design of conditioning mechanisms deserves as much attention as the design of network architectures and training objectives. As imaging systems become more instrumented and acquisition metadata more routinely available, the principles developed here — explicit, structure-respecting, physically grounded conditioning — offer a path toward deep imaging models that are more accurate, more generalizable, and more transparent in their use of the physical world they are trying to reconstruct.

References

Candès, E. J., Romberg, J. and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, **52** (2), 489–509.

Chen, L., Chu, X., Zhang, X. and Sun, J. (2022). Simple baselines for image restoration. In *European Conference on Computer Vision*, pp. 17–33.

Chung, H., Lee, D., Wu, Z., Kim, B.-H., Bouman, K. L. and Ye, J. C. (2025). ContextMRI: Enhancing compressed sensing MRI through metadata conditioning. *arXiv preprint arXiv:2501.04284*.

—, Lee, S. and Ye, J. C. (2023). Decomposed diffusion sampler for accelerating large-scale inverse problems. *arXiv preprint arXiv:2303.05754*.

— and Ye, J. C. (2022). Score-based diffusion models for accelerated MRI. *Medical Image Analysis*, **80**, 102479.

Dabov, K., Foi, A., Katkovnik, V. and Egiazarian, K. (2007). Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, **16** (8), 2080–2095.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, **52** (4), 1289–1306.

Elad, M. and Aharon, M. (2006). Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, **15** (12), 3736–3745.

Esin, Y. E. (2017). *Coil sensitivity map calculation using Biot-Savart law at 3 Tesla*. Ph.D. thesis, Middle East Technical University.

Fadnavis, S., Batson, J. and Garyfallidis, E. (2020). Patch2self: Denoising diffusion MRI with self-supervised learning. *Advances in Neural Information Processing Systems*, **33**, 16293–16303.

Foi, A., Trimeche, M., Katkovnik, V. and Egiazarian, K. (2008). Practical Poissonian-Gaussian noise modeling and fitting for single-image raw-data. *IEEE Transactions on Image Processing*, **17** (10), 1737–1754.

Griswold, M. A., Jakob, P. M., Heidemann, R. M., Nittka, M., Jellus, V., Wang, J., Kiefer, B. and Haase, A. (2002). Generalized autocalibrating partially parallel acquisitions (GRAPPA). *Magnetic Resonance in Medicine*, **47** (6), 1202–1210.

Guan, H. and Liu, M. (2022). Domain adaptation for medical image analysis: A survey. *IEEE Transactions on Biomedical Engineering*, **69** (3), 1173–1185.

Guo, S., Chen, Z., Shi, W., Chen, Y. and Zhang, L. (2019). Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1712–1722.

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

— and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Karani, N., Chaitanya, K., Baumgartner, C. and Konukoglu, E. (2018). A lifelong learning approach to brain MR segmentation across scanners and protocols. In *Medical Image Computing and Computer Assisted Intervention*, pp. 476–484.

Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M. and Aila, T. (2018). Noise2noise: Learning image restoration without clean data. *arXiv preprint arXiv:1803.04189*.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L. and Timofte, R. (2021). SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 1833–1844.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.

Lustig, M., Donoho, D. and Pauly, J. M. (2007). Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, **58** (6), 1182–1195.

Moeller, S. *et al.* (2021). Noise reduction with distribution corrected (NORDIC) PCA in dMRI. *NeuroImage*, **226**, 117539.

Neto Henriques, R., Nørhøj Jespersen, S. and Shemesh, N. (2022). Hybrid PCA denoising. In *Proceedings of the 30th Scientific Meeting of the ISMRM*, p. 84.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171.

Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205.

Perez, E., Strub, F., de Vries, H., Dumoulin, V. and Courville, A. (2018). FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

Pinaya, W. H., Tudosiu, P.-D. *et al.* (2022). Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pp. 117–126.

Pruessmann, K. P., Weiger, M., Scheidegger, M. B. and Boesiger, P. (1999). SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, **42** (5), 952–962.

Quan, Y., Chen, M., Pang, T. and Ji, H. (2020). Self2self with dropout: Learning self-supervised denoising from single image. In *Proceedings of the IEEE/CVF CVPR*, pp. 1890–1898.

Radford, A., Kim, J. W., Hallacy, C. *et al.* (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763.

Rajan, J., Poot, D., Juntu, J. and Sijbers, J. (2010). Noise measurement from magnitude MRI using local estimates. *Physics in Medicine & Biology*, **55** (16), N441.

Ramos-Llordén, G., Vegas-Sánchez-Ferrero, G., Liao, C., Westin, C.-F., Setsompop, K. and Rathi, Y. (2021). SNR-enhanced diffusion MRI with structure-preserving low-rank denoising in reproducing kernel Hilbert spaces. *Magnetic Resonance in Medicine*, **86** (3), 1614–1632.

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

Rudin, L. I., Osher, S. and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, **60** (1-4), 259–268.

Shen, G., Li, M., Farris, C. W., Anderson, S. and Zhang, X. (2024). Learning to reconstruct accelerated MRI through k-space cold diffusion. *Scientific Reports*, **14**, 21877.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. and Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*.

Song, J., Meng, C. and Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, Y. and Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, **32**.

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

Uecker, M., Lai, P., Murphy, M. J. *et al.* (2014). ESPIRiT—an eigenvalue approach to autocalibrating parallel MRI. *Magnetic Resonance in Medicine*, **71** (3), 990–1001.

Veraart, J., Novikov, D. S., Christiaens, D., Ades-Aron, B., Sijbers, J. and Fieremans, E. (2016). Denoising of diffusion MRI using random matrix theory. *NeuroImage*, **142**, 394–406.

Wang, C., Lyu, J., Wang, S. *et al.* (2024). CMRxRecon: A publicly available k-space dataset for cardiac MRI. *Scientific Data*, **11**, 687.

Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13** (4), 600–612.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. MIT Press.

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S. K. and Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF CVPR*, pp. 5728–5739.

Zbontar, J., Knoll, F., Sriram, A. *et al.* (2018). fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*.

Zhang, K., Zuo, W., Chen, Y., Meng, D. and Zhang, L. (2017). Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, **26** (7), 3142–3155.

—, — and Zhang, L. (2018a). FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, **27** (9), 4608–4622.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E. and Wang, O. (2018b). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF CVPR*, pp. 586–595.

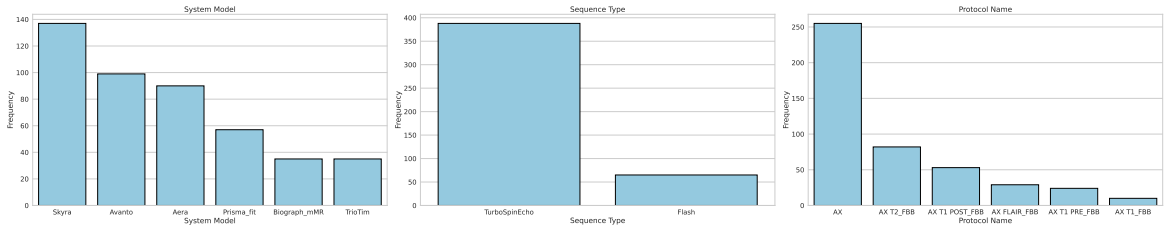


Figure A.1: ***Distribution of categorical metadata fields in the fastMRI brain dataset.*** *Fields include system model, sequence type, and protocol name. The dominance of a small number of categories reflects the limited protocol diversity in the dataset.*

Appendix A

Supplementary Results

A.1 fastMRI Brain Metadata Distributions

The following figures show the distributions of categorical and continuous metadata fields in the fastMRI brain dataset used for training and evaluation. The narrow clustering of continuous parameters such as TR, TE, and flip angle around a small number of discrete values motivates the out-of-distribution training strategy described in Section 3.3.4.

A.2 OOD Cardiology Metadata Distributions

The following figures show the metadata distributions of the CMRxRecon cardiac cine dataset used for out-of-distribution training. Compared to the fastMRI brain distributions

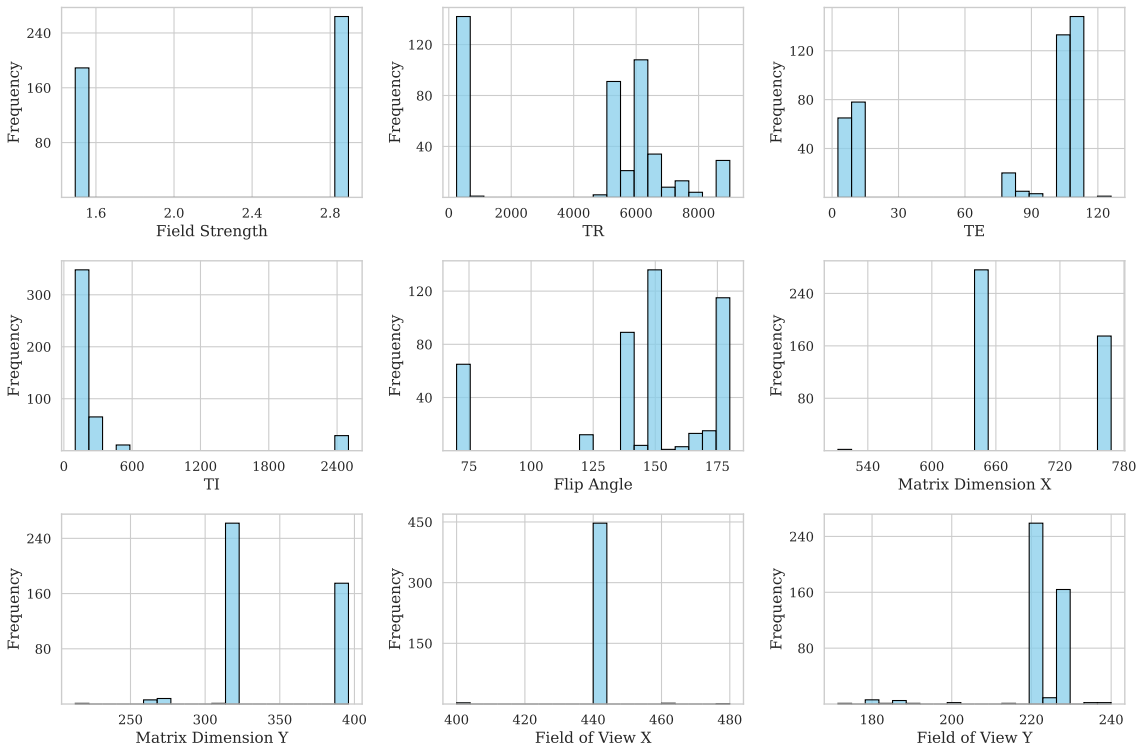


Figure A.2: ***Distribution of continuous metadata fields in the fastMRI brain dataset.*** *Fields include TR, TE, TI, flip angle, and geometric parameters. Despite being recorded as continuous values, most parameters cluster at only a few discrete operating points, limiting the diversity of numerical conditioning seen during training.*

in Section A.1, the cardiac dataset covers substantially different ranges of TR, TE, and flip angle, confirming that it provides genuine out-of-distribution coverage that forces active metadata conditioning during training.

A.3 ContextMRI Cross-Attention Weight Analysis

To diagnose why ContextMRI's CLIP-based metadata conditioning fails to improve reconstruction quality, we probe the cross-attention layers of the diffusion model's UNet during inference. Specifically, we extract attention maps from the `UNet2DConditionModel` mid-block attention processor, which indicate how much the model attends to each token in the input text embedding when generating features at a given spatial location.

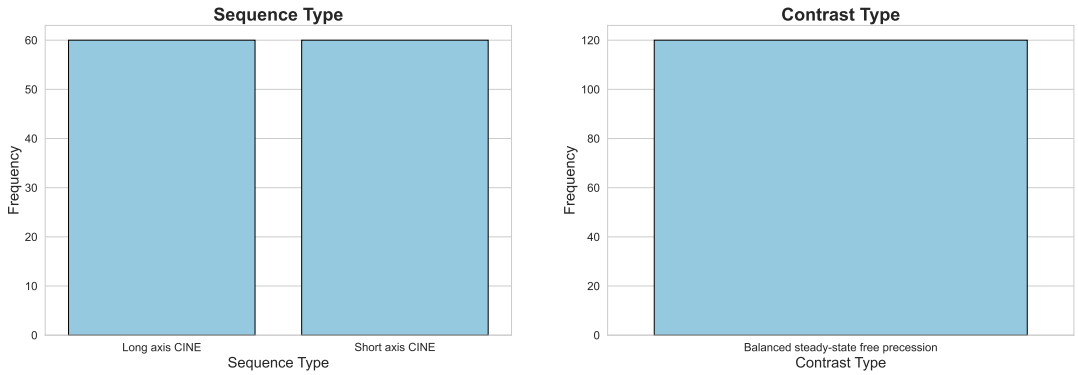


Figure A.3: ***Distribution of categorical metadata fields in the CMRxRecon cardiac dataset.*** *Sequence types consist of long-axis and short-axis CINE acquisitions, and contrast type is uniformly balanced steady-state free precession — both distinct from the brain dataset protocol categories.*

Figure A.5 shows the resulting token-level attention weights for a representative metadata prompt. The CLS token — the summary token that aggregates the entire sequence — and the anatomy token receive the dominant share of attention weight. Tokens corresponding to numerical acquisition parameters including TR, TE, TI, and flip angle receive near-zero weight, regardless of their values. This confirms that the cross-attention mechanism effectively ignores the numerical metadata tokens, collapsing all conditioning information into the CLS summary representation and discarding the individual parameter values. This analysis provides direct mechanistic evidence for the failure mode identified in Section 3.3.2 and motivates the type-aware embedding design proposed in Section 3.3.3.

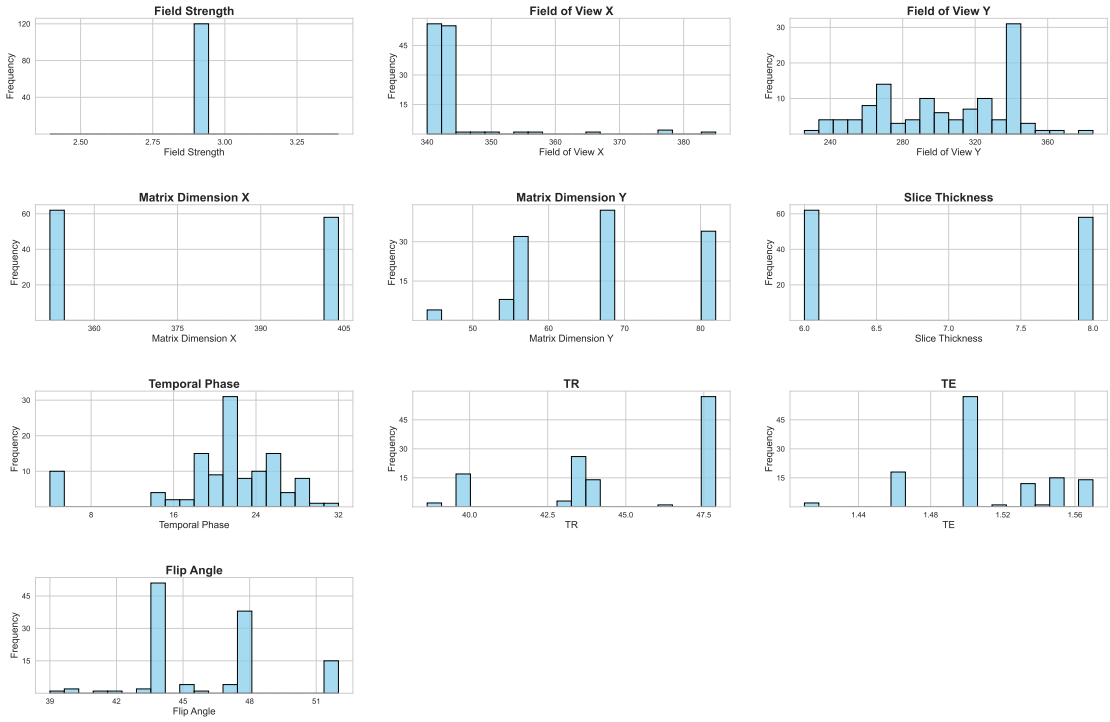
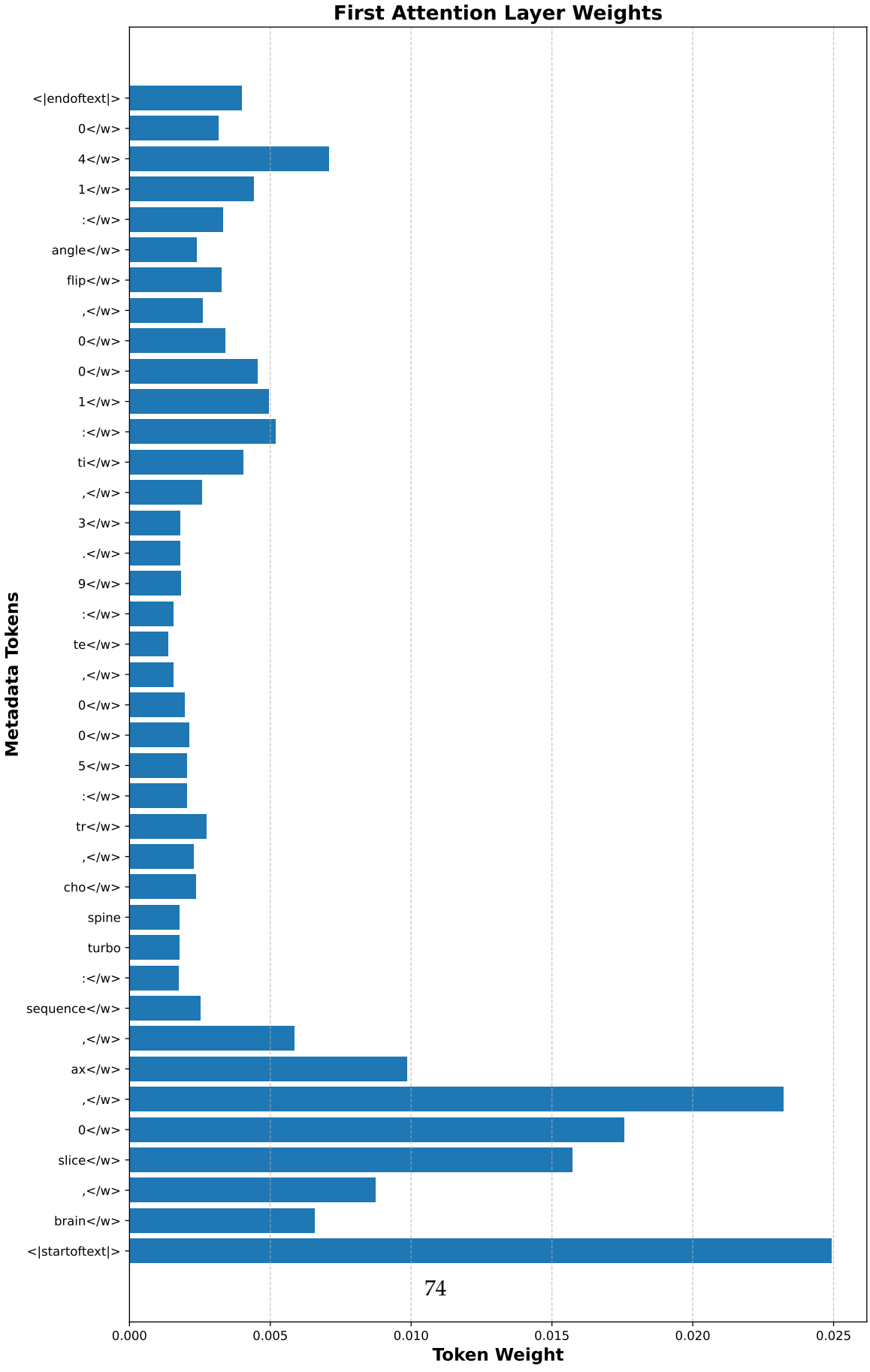


Figure A.4: ***Distribution of continuous metadata fields in the CMRxRecon cardiac dataset.*** *TR spans 3.6–50 ms, TE spans 1.1–1.6 ms, and flip angles range from 12 to 60 degrees, all lying outside the fastMRI brain distribution and providing the distributional distance necessary to force metadata reliance during OOD training.*


First Attention Layer Weights
Metadata Tokens
<|endoftext|>
0</w>
4</w>
1</w>
:</w>
angle</w>
flip</w>
,</w>
0</w>
0</w>
1</w>
:</w>
ti</w>
,</w>
3</w>
.</w>
9</w>
:</w>
te</w>
,</w>
0</w>
0</w>
5</w>
:</w>
tr</w>
,</w>
cho</w>
spine
turbo
:</w>
sequence</w>
,</w>
ax</w>
,</w>
0</w>
slice</w>
,</w>
brain</w>
<|startoftext|>
0.000
0.005
0.010
0.015
0.020
0.025
Token Weight