Metadata Matters: Fine-Grained Embeddings for Enhanced MRI Reconstruction

Joanna Kondylis
kondylis@mit.edu

Dewei Feng
dewei@mit.edu

Linbo Tang
linbotang@g.harvard.edu

Heng-Jui Chang
hengjui@mit.edu

Zoe Shleifer
zoe_shleifer@college.harvard.edu

Abstract

Magnetic Resonance Imaging (MRI) is a critical diagnostic modality but faces clinical limitations due to long acquisition times. Recent approaches leverage diffusion models conditioned on textual metadata to enhance Compressed Sensing (CS) MRI reconstruction. However, existing methods, such as ContextMRI, use general-purpose encoders like CLIP, which inadequately represent numerical MRI parameters (e.g., TR, TE, flip angle). This work proposes a fine-grained, structured embedding framework that explicitly encodes categorical and numerical metadata separately—categorical attributes via discrete embeddings and numerical attributes through sinusoidal encodings. Additionally, we introduce an out-of-distribution (OOD) training strategy by incorporating cardiac MRI data to force reliance on metadata embeddings. Experiments on the fastMRI brain dataset demonstrate significant improvements over ContextMRI, notably in scenarios with diverse or uncommon acquisition parameters. Our approach enhances reconstruction accuracy and provides explicit control over individual metadata factors, offering improved interpretability and robustness in MRI reconstruction.

1 Introduction

Magnetic resonance imaging (MRI) is indispensable in modern diagnostics, yet long acquisition times slow clinical workflows and increase motion artifacts. Compressed Sensing (CS) first addressed this bottleneck by exploiting transform sparsity to recover images from heavily undersampled k-space measurements [1]. More recently, Denoising Diffusion Probabilistic Models (DDPMs) [2] have become powerful learned priors for ill-posed inverse problems. Score-based generative reconstruction showed that a diffusion prior, combined with data-consistency projections, can accurately reconstruct multi-coil MR images across sampling patterns and provides calibrated uncertainty estimates [3]. Subsequent work, such as the Decomposed Diffusion Sampler (DDS), further accelerates inference without sacrificing quality [4].

An under-explored source of domain knowledge in MRI is the rich **metadata** shipped with every scan: imaging parameters (TR, TE, flip angle), sequence/contrast, anatomy, and often patient context (age, sex, pathology). These factors determine tissue contrast and constrain plausible anatomy, yet most diffusion reconstructions ignore them. ContextMRI, a text-conditioned diffusion model for MRI reconstruction, demonstrated that feeding metadata as a CLIP-encoded text prompt boosts reconstruction fidelity [5]. However, a free-form prompt encoded via a text encoder entangles heterogeneous information and might not represent continuous parameters such as TR or TE.

This work introduces a fine-grained, *type-aware* metadata conditioning scheme for diffusion-based MRI reconstruction. Categorical attributes (e.g. sequence type, anatomy) are mapped to learnable embeddings, whereas continuous fields (e.g. TR, TE, patient age) are encoded via sinusoidal or learned linear projections. These structured embeddings are fused—optionally alongside CLIP text vectors—into the diffusion U-Net. Our approach offers (i) higher reconstruction accuracy under strong undersampling, (ii) explicit control over individual metadata factors, and (iii) improved interpretability and robustness across varied protocols. To our knowledge, this is the first diffusion prior that leverages *structured numeric and categorical* MRI metadata.

2 Related Works

Diffusion models for medical image generation. Pinaya *et al.* used a latent diffusion model to synthesise 3-D brain MRIs conditioned on demographic covariates, releasing 100k realistic volumes [6]. SynthBrainGrow extends diffusion to longitudinal brain aging, generating two-year follow-up scans with realistic cortical thinning [7]. Med-DDPM concatenates tumour masks to enable pixel-level controllable 3-D brain-MRI synthesis and improves downstream segmentation [8].

Diffusion models for MRI reconstruction. ScoreMRI pioneered score-based posterior sampling for accelerated MRI [3]. DDS introduced Krylov subspace projections to cut sampling cost by an order of magnitude [4]. Other works incorporate physics constraints or plug-and-play solvers, but none exploit rich clinical metadata.

Metadata conditioning. Outside reconstruction, latent diffusion generators have accepted structured covariates (age, ventricular volume) [6]. In MRI reconstruction, ContextMRI is the first to include metadata, but encodes all fields with a CLIP text encoder, which is not designed for encoding MRI metadata [5]. Our method departs from prompt-only conditioning by introducing *field-wise, type-specific* embeddings, capturing subtle numeric differences (e.g. TR $=$ 5000 ms vs. 500 ms) and enabling fine-grained ablations. We thereby close the gap between unconditional diffusion priors and the information-rich reality of clinical MRI.

Summary and connection to our approach. The recent cascade of diffusion-based advances in MRI can be grouped into three threads: (i) *3-D generation* with latent or explicit diffusion backbones [6, 7, 8], (ii) *posterior reconstruction* via score-based sampling and its fast variants [3, 4], and (iii) the *first attempts at context-aware conditioning*, exemplified by ContextMRI [5]. Generative work demonstrates that diffusion models can ingest side information to steer macroscopic attributes (age, sex, tumor masks), yet those pipelines are detached from the physics-constrained reconstruction task. Conversely, score-based reconstructions show excellent performance but treat all scans as context–free, leaving protocol-specific ambiguities unresolved. ContextMRI closes this gap partially by feeding free-form text prompts, but its single CLIP embedding cannot disentangle or effectively encode continuous acquisition parameters.

Our contribution unifies the strengths of the above threads. We bring the fine-grained conditioning philosophy of the generative works *into the reconstruction setting* while preserving the data-consistency guarantees of score-based solvers. By designing *type-aware embeddings* for each metadata field, our model (i) captures subtle numeric effects (e.g. TE shifts), (ii) remains backwards-compatible with CLIP prompts, and (iii) yields an interpretable knob for analyzing how each clinical attribute shapes the posterior. This structured-conditioning design has, to our knowledge, not been explored in diffusion-based MRI reconstruction and directly addresses the open questions identified in prior literature.

3 Methods

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 [9]. Due to computational constraints, we train our models on a subset of 450 patient studies (7,185 scans) and evaluate performance on a separate test set of 150 patients (2,957 scans).

Following the ContextMRI pipeline, each raw multi-coil fastMRI volume undergoes an inverse Fourier transform and is combined with coil-sensitivity maps estimated via ESPIRiT [10], resulting

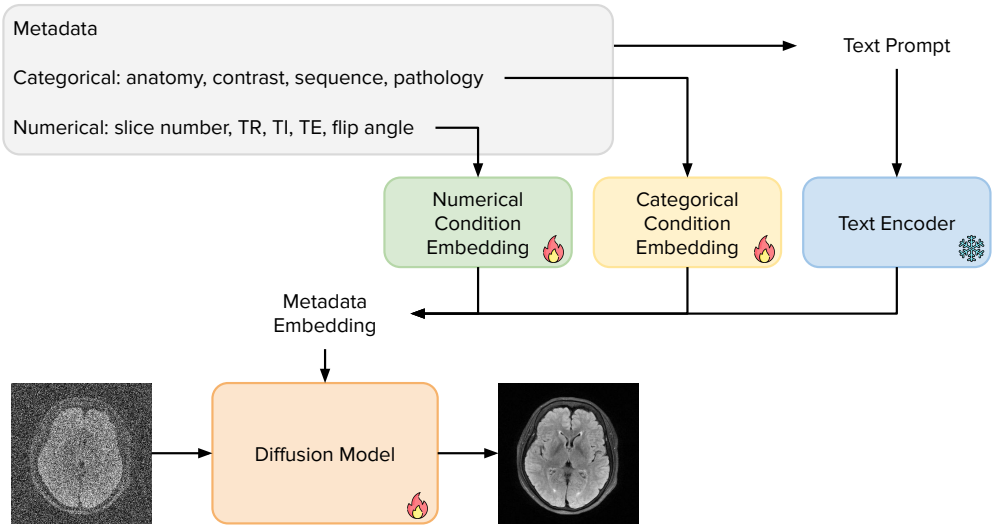


Figure 1: Proposed metadata conditioning MRI diffusion model framework. The original ContextMRI framework [5] only uses the text encoder. This project proposes to encode metadata with separate numerical and categorical embeddings to represent conditions explicitly.

in minimum-variance unbiased (MVUE) complex-valued images for each slice. These complex images are then center-cropped to a resolution of 320×320 pixels to remove zero-padding from k-space. Subsequently, real and imaginary components are stacked into a two-channel tensor. Finally, 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.

3.2 Theoretical Overview

Diffusion models utilize a continuous sequence of densities $p_t(\boldsymbol{x})$ indexed by diffusion time t [2]. These models begin with samples from the data distribution $\boldsymbol{x}_0 \sim p_0(\boldsymbol{x})$ and progressively add Gaussian noise according to the conditional distribution $p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\sqrt{\bar{\alpha}_t})\, I_d\right)$ until the data approximate a standard normal distribution $\boldsymbol{x}_T \sim \mathcal{N}(0, I_d)$, where $\boldsymbol{x} \in \mathbb{R}^d$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. During training, a neural network $s_\theta(\boldsymbol{x}_t)$ estimates the score function $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ through denoising score matching [11], which, for Gaussian perturbations, is equivalent to predicting the added noise. Typically, this is achieved by predicting noise with a network $\epsilon_\theta(\boldsymbol{x}_t)$, minimizing the following loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, \boldsymbol{x}_0 \sim p_0, \epsilon \sim \mathcal{N}(0,I)} \left[\|\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t)\|^2 \right]. \tag{1}$$

For inference, Denoising Diffusion Implicit Models (DDIM) [12] generate samples deterministically by iteratively applying the following updates from $t = T, T-1, \ldots, 0$, starting with random Gaussian noise:

$$\hat{\boldsymbol{x}}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} \left(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(\boldsymbol{x}_t, c) \right), \tag{2}$$

$$\boldsymbol{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{\boldsymbol{x}}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\epsilon_\theta(\boldsymbol{x}_t, c) + \sigma_t \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, I_d), \tag{3}$$

where σ_t is set to zero for deterministic sampling.

To incorporate condition c, such as metadata, diffusion models can be trained to sample from conditional distributions $p(\boldsymbol{x}|c)$. Text-conditioned models commonly use paired data $(\boldsymbol{x}, c)$ along with random conditioning dropout, allowing both conditional $\epsilon_\theta(\boldsymbol{x}_t, c)$ and unconditional $\epsilon_\theta(\boldsymbol{x}_t)$ sampling, i.e., $\epsilon_{\text{cfg}} = \epsilon_\theta(\boldsymbol{x}_t, c) + w\left[\epsilon_\theta(\boldsymbol{x}_t, c) - \epsilon_\theta(\boldsymbol{x}_t)\right]$, where w is a weight that can be tuned during inference. During inference, the predicted noise ϵ_{cfg} is substituted into Eqs. 2 and 3 in place of ϵ_θ.

3.3 Background: ContextMRI

We build upon the ContextMRI framework [5], a diffusion model trained on complex-valued MR images and conditioned on metadata to enhance MR image generation quality. ContextMRI transforms metadata into textual prompts, encoding these prompts into continuous representations using the pre-trained, frozen CLIP text encoder [13]. For example, if an MR image is scanned with contrast `AX R1 POST_FBB` and TR of 3150, the text prompt would be "`AX R1 POST_FBB, TR: 3150`."

Despite integrating metadata into the generation process, improvements reported in [5] are minimal, with the performance gap between conditioned and unconditioned reconstructions smaller than the variability of evaluation metrics. We hypothesize two reasons for this limitation: (1) the CLIP encoder was not trained on such specialized text prompts, and (2) the encoder inadequately captures numerical metadata values. Consequently, we propose an enhanced metadata conditioning approach employing fine-grained embeddings, detailed in the subsequent section.

To emphasize the complexity and motivate the need for improved metadata encoding strategies, we present visualizations of continuous and categorical metadata distributions in Figures 6 and 7, respectively, in Appendix A.

3.4 Fine-grained Metadata Condition Embedding

Categorical and Numerical Condition Embedding. We introduce a fine-grained metadata encoding approach that maps MRI metadata into continuous embeddings using attribute-specific embedding modules. For categorical attributes (e.g., contrast type, pathology), embedding tables transform each category into continuous vectors. For numerical attributes (e.g., TR, TE), values are normalized to the range $[0, 1]$ and encoded with sinusoidal positional embeddings, analogous to timestep encoding in diffusion models.

Additive and Concatenated Condition Embedding. We propose two methods to integrate these fine-grained embeddings with text embeddings. The additive method sums all fine-grained embeddings and combines them with the original CLIP text embeddings. Conversely, the concatenation method appends fine-grained embeddings sequentially to text embeddings, separated by a special token `[SEP]`. Although concatenation requires greater memory and computational resources, it allows easier utilization of condition embeddings via cross-attention mechanisms in the diffusion model's U-Net architecture.

Classifier-free Guidance. Following ContextMRI, we employ classifier-free guidance (CFG) [14] to enhance generation quality. Each fine-grained embedding is independently dropped during training, simulating real-world scenarios with incomplete metadata. The overall framework is depicted in Figure1.

3.5 LoRA

Adapting large-scale diffusion models to new metadata conditioning typically requires extensive computational resources. To mitigate this issue, we use Low-Rank Adaptation (LoRA) [15], introducing trainable low-rank matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, such that the weight matrix W becomes $W + \alpha AB$, where α scales the contribution of the update and $r \ll d$. Applied to U-Net cross-attention layers, LoRA efficiently integrates fine-grained metadata embeddings with minimal resource requirements and maintains robustness in metadata-scarce conditions. Although initial LoRA results slightly trail full fine-tuning (Table 1), performance may improve with optimal rank selection.

3.6 Out-of-Distribution Metadata Conditioning

To encourage the diffusion model to place greater weight on metadata features during reconstruction, we tested incorporating out-of-distribution (OOD) samples, with a significantly different metadata distribution, during training. When pixel information is heavily corrupted by noise, metadata embeddings become the primary source of reliable signal. Interleaving such OOD cases—drawn from a domain whose metadata values lie well outside those in the fastMRI brain dataset—compels the network to learn effective mappings from metadata to reconstruction priors.

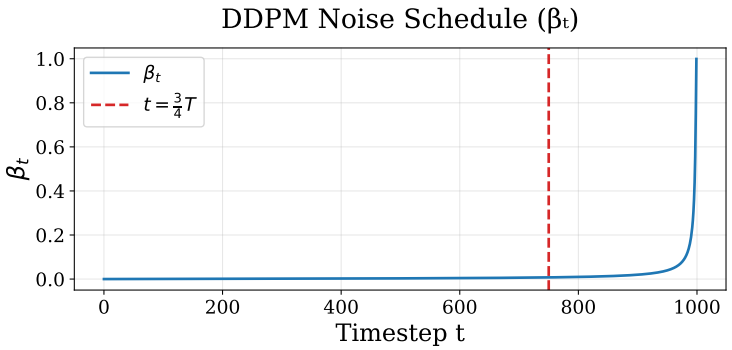


Figure 2: Noising schedule with the red line at $\frac{3}{4}T_{\text{max}}$, indicating the start of OOD noising.

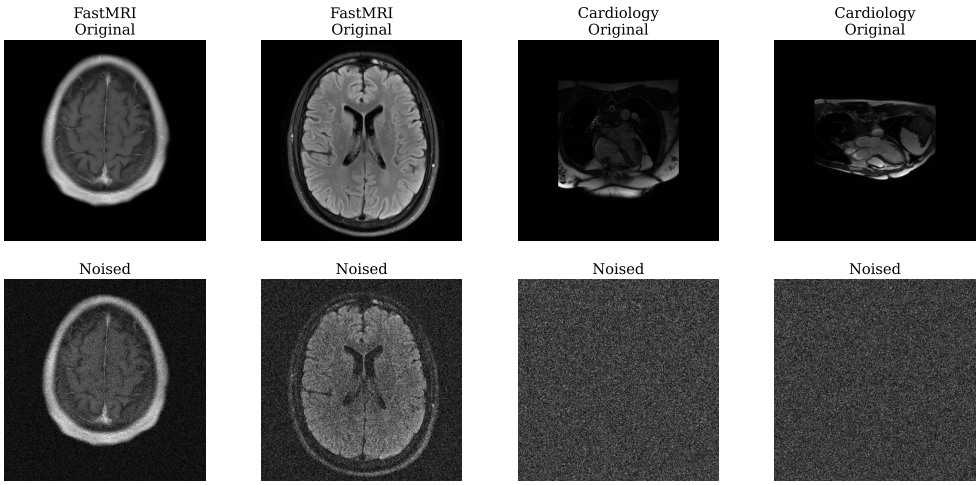


Figure 3: Visualization of original images (top row) and noised images (bottom row). Brain features can remain detectable, while cardiology (OOD) features are always obscured.

Our in-distribution dataset is comprised of the fastMRI brain repository, consisting of approximately 7,000 two-dimensional slices acquired under a limited set of repetition times (TR), echo times (TE), flip angles, and field-of-view settings. Although metadata fields such as TR, TE, and flip angle are recorded continuously, most assume only a few discrete values (see Appendix A).

For OOD augmentation we selected the CMRxRecon cardiac cine dataset [16], 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. Imaging parameters span TR$=$ 3.6ms–50ms, TE=1.1ms–1.6ms, and flip angles of $12 - 60$ degrees, offering coverage beyond the fastMRI distribution. For pre-processing the OOD data, we applied the same coil-combining and z-score normalization as for brain data. During training, cardiology slices were sampled at a $1 : 5$ ratio to brain, and noised at time steps between $\frac{3}{4}T_{\text{max}}$ and T_{max}, effectively masking anatomical content and forcing reliance on metadata embeddings. We provide a visualization of the OOD metadata distribution in Appendix B (Figures 8 and 9) and of the noising process (Figure 2 and Figure 3).

The OOD version of the model was trained in conjunction with the additive embedding implementation and is thus coined the OOD + Additive model.

4 Experiments

4.1 Baselines and Setting

For all training variants, we maintained consistent hyperparameters to ensure fair comparison: training batch size of 4 samples, gadient accumulation steps set to 1, learning rate of $5e-5$ with a constant scheduler, and a training duration of $10,000$ total steps. All models were trained using two NVIDIA Volta V100 GPUs.

Importantly, all of our models build upon the pre-trained ContextMRI checkpoint rather than training from scratch. This approach allows us to leverage the robust reconstruction capabilities already developed in the baseline model while introducing our targeted improvements through fine-tuning. This methodological choice significantly reduces computational requirements while enabling fair comparison of our enhancements against the original architecture.

Our experimental evaluation focuses on the $4\times$ acceleration task, which simulates a quadruple reduction in MRI acquisition time by retaining only 25% of k-space measurements through uniform undersampling. All of our proposed methods are benchmarked against the pre-trained ContextMRI model.

During inference, we employ 100 diffusion steps for all reconstruction tasks, which balances computational efficiency with reconstruction quality and matches the ContextMRI methodology.

For evaluation, we compute quality metrics on a per-slice basis, calculating Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [17], and Learned Perceptual Image Patch Similarity (LPIPS) [18] individually for each reconstructed slice before aggregating statistics across the test set. Our metrics are reported for CFG $= 1$ unless otherwise indicated.

4.2 Main Results: Fine-Grained Embeddings Outperform ContextMRI

Table 1 provides a comprehensive comparison of our metadata embedding approaches against the baseline ContextMRI model for fastMRI brain reconstruction. Our analysis reveals significant performance improvements across our proposed methods.

When comparing the Additive and Concatenated Embedding approaches, we observe similar performance in terms of PSNR and SSIM metrics. The Additive method achieves slightly higher PSNR, suggesting marginally better signal fidelity, while the Concatenated approach shows stronger LPIPS scores, indicating superior perceptual quality. We ultimately selected the Additive Embedding method for training jointly with the out-of-distribution approach.

Building upon the Additive Embedding foundation, our OOD + Additive approach demonstrates the strongest overall performance among all tested methods. This enhanced performance can be attributed to the strategic introduction of out-of-distribution cardiac MRI samples during training. By incorporating heavily noised OOD samples, we effectively force the model to rely more heavily on metadata embeddings on top of image content. This training strategy enhances the model's ability to extract meaningful information from acquisition parameters and apply it during the reconstruction process, even when working with in-distribution brain MRI data.

The significant improvements over the ContextMRI baseline highlight the importance of how metadata is represented and integrated into diffusion models. These results validate our hypothesis that structured, attribute-specific embeddings provide more precise representations of numerical scan parameters than the original approaches. The consistent performance gains across different embedding strategies demonstrate that fine-grained metadata conditioning substantially enhances the fidelity and robustness of diffusion-based MRI reconstruction.

Unless otherwise indicated, all further experiments in this section are carried out with the OOD + Additive model, as it achieves the best overall performance.

4.3 Varying Conditioning Strength

To evaluate the effect of conditioning strength on reconstruction performance, we conducted experiments varying the classifier-free guidance (CFG) scale used during sampling. Specifically, we tested CFG values of 1.0, 2.0, and 3.0, comparing them against an unconditioned baseline. As shown

Method	PSNR ↑	SSIM $\cdot 10^2$ ↑	LPIPS $\cdot 10^2$ ↓
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	$\mathbf{23.51 \pm 13.78}$
LoRA	28.36 ± 6.01	63.83 ± 22.25	33.83 ± 11.50
OOD + Additive	$\mathbf{30.40 \pm 6.62}$	$\mathbf{79.04 \pm 21.97}$	26.20 ± 11.98

Table 1: Quantitative results on the FastMRI brain dataset (Uniform 1D mask, Acceleration $\times 4$ and CFG $= 1.0$). We compare our models against ContextMRI in a variety of metrics.

CFG	PSNR ↑	SSIM $\cdot 10^2$ ↑	LPIPS $\cdot 10^2$ ↓
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 2: For the OOD + Additive model quantitative results when varying the conditioning strength (CFG).

in Table 2, increasing the CFG value had a modest but consistent effect on performance: PSNR and SSIM remained stable across values, with a slight improvement at CFG = 3.0, while LPIPS remained low, indicating perceptual fidelity was preserved. Notably, even mild guidance (CFG = 1.0) outperformed the unconditioned model, affirming the utility of metadata-aware conditioning. These results suggest that moderate CFG values offer a robust trade-off between faithful adherence to metadata and reconstruction quality.

4.4 Impact of MRI Imaging Parameters

4.4.1 Dropping One Metadata Field at a Time

To understand the contribution of each MRI acquisition parameter to the reconstruction quality, we conduct two complementary ablation studies. First, in Table 3, we drop one metadata field at a time during inference while keeping all other fields intact. As a result, different metadata fields have varying levels of influence on reconstruction quality. Dropping semantic fields such as Contrast, Sequence, and Anatomy tends to result in a slightly more significant degradation across PSNR, SSIM, and LPIPS, indicating that these attributes provide essential contextual cues that guide the model's predictions. These fields likely inform the model about global scan characteristics or anatomy-specific priors that are difficult to infer from the image alone.

In contrast, removing acquisition parameters such as TE, TI, and Slice Number leads to relatively smaller performance declines. While these values do carry scan-specific information, they are often continuous and may vary in scale and distribution, making them harder for the model to consistently exploit during inference. Their contribution appears to be more subtle, potentially refining rather than shaping the core reconstruction.

4.4.2 Conditioning on a Single Metadata Field

To further isolate the contribution of individual metadata fields, Table 4 reports reconstruction performance when conditioning on a single metadata attribute at a time while removing all others. This complements our drop-one-field analysis by probing how much signal each field contributes on its own. Semantic metadata—particularly Anatomy, Sequence, and Slice Number—again emerge as the most informative. These features support strong reconstructions even in isolation, achieving SSIM scores above 75 and LPIPS values near those of the fully conditioned model, indicating that they encapsulate rich structural priors about the scan content. In contrast, continuous acquisition

Dropped Metadata Feature	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: Ablation study on dropping individual metadata features. Each row shows the model's performance when all metadata is used *except* for the listed feature.

Metadata Feature	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	29.99 ± 6.25	75.59 ± 23.93	26.46 ± 11.64

Table 4: Ablation study on individual metadata features. We examine reconstruction performance when using single acquisition parameters in isolation. In other words, we only condition on the one metadata feature and all other conditioning parameters are dropped.

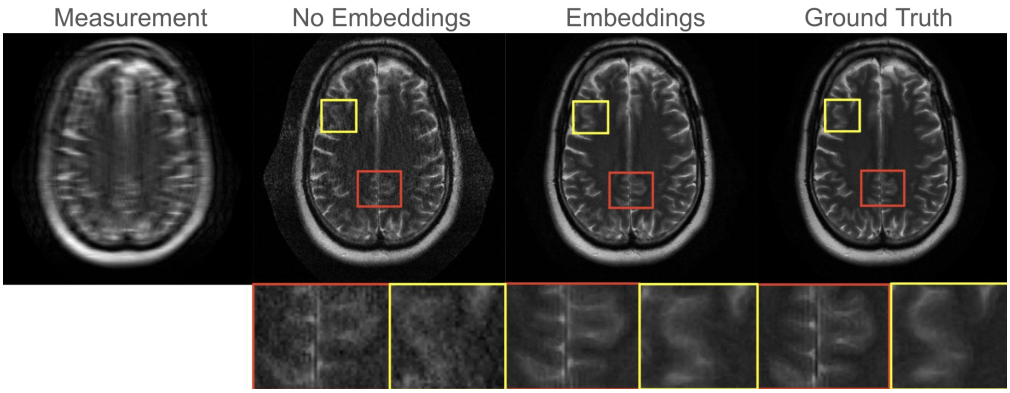
parameters such as TE, TI, and TR perform substantially worse when used alone, leading to declines of over 6 SSIM points and substantial increases in LPIPS. These fields appear to be most useful when combined with other context. In particular, it may be that these numerical parameters are most useful when Contrast information is accessible, so that the model can identify which of a few modes the sample falls into. Interestingly, Flip Angle achieves the best performance in this setting, likely because it implicitly captures contrast and sequence variations on its own, making it highly predictive in isolation, even though it is redundant when richer metadata is available.

4.5 Comparing Visualizations

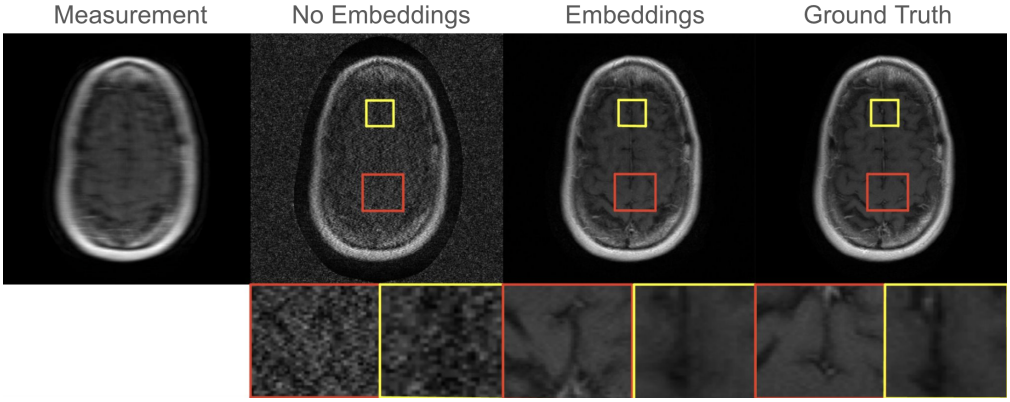
To qualitatively assess the effect of fine-grained metadata embeddings, we visualize reconstructed brain slices from models trained with and without metadata conditioning. Figure 4 presents two representative examples comparing outputs from an unconditioned model (no embeddings) to our proposed OOD + Additive model using structured embeddings.

In both cases, the metadata-conditioned reconstructions yield higher anatomical fidelity, particularly in regions prone to low signal-to-noise ratio (SNR). For instance, subtle boundaries within the ventricular system and surrounding gray-white matter structures are noticeably sharper in the conditioned outputs. This suggests that the diffusion model effectively leverages acquisition parameters (such as TE and flip angle) to refine its reconstructions.

The improvements are especially visible in areas with poor initial input quality, demonstrating that fine-grained metadata conditioning can compensate for information loss due to undersampling. These visual results support our quantitative findings and emphasize the utility of structured conditioning in improving clinical interpretability.



(a) Example 1: No Embeddings vs. With Embeddings



(b) Example 2: No Embeddings vs. With Embeddings

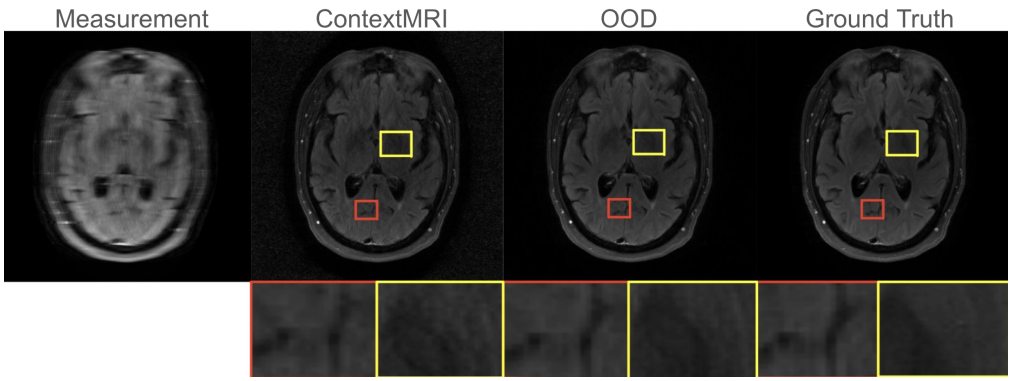
Figure 4: Comparison between unconditioned (no embeddings) and fine-grained metadata embedding models. Embeddings improve fidelity in subtle brain structures, especially in low-SNR regions.

Furthermore, we also visualize reconstructed brain slices from models trained with ContextMRI and our OOD-aware model. Figure 5 presents two representative examples comparing outputs from ContextMRI to our OOD-aware model. We observe a similar trend where the OOD-aware model produces reconstructions with improved structural detail and anatomical sharpness, particularly in regions with subtle or low-contrast features. The zoomed-in views in the bottom row highlight these differences more clearly—edges around ventricles and fine cortical structures are more crisply defined in the OOD-aware reconstructions. This qualitative improvement reinforces the hypothesis that introducing heavily noised, out-of-distribution samples during training encourages the model to rely more heavily on metadata embeddings, resulting in reconstructions that are both more accurate and more robust to variation in acquisition parameters.

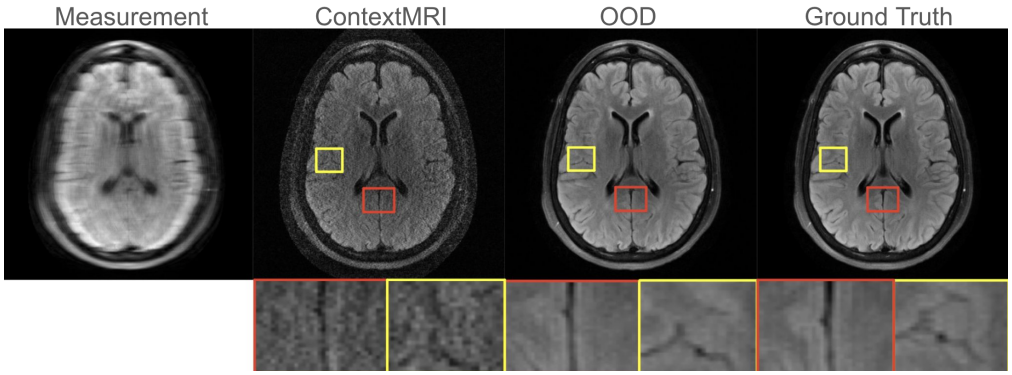
4.6 OOD with Low Noise

While our primary OOD approach focused on heavily corrupted samples to emphasize metadata dependence, we hypothesized that an alternative strategy might enhance reconstruction of fine anatomical details. In this experiment, we implemented a complementary conditioning paradigm using OOD cardiac data with minimal noise corruption, restricting the noise schedule to 0 to $\frac{1}{4}T_{\text{max}}$.

This approach is motivated by the observation that diffusion models learn different features at different noise levels—with the early, low-noise regime particularly crucial for capturing subtle tissue boundaries and fine structural details. By concentrating OOD samples in this low-noise region, we aimed to enhance the model's sensitivity to minute anatomical features that significantly impact diagnostic quality but may be underrepresented in the brain-only training distribution.



(a) Example 1: ContextMRI vs. OOD



(b) Example 2: ContextMRI vs. OOD

Figure 5: Comparison of reconstructions between ContextMRI and our OOD-aware model on two different brain slices. The bottom row of each image shows zoomed-in regions with enhanced recovery of structure and detail in OOD + Additive reconstructions.

Method	PSNR ↑	SSIM $\cdot 10^2$ ↑	LPIPS $\cdot 10^2$ ↓
High Noise for OOD	$\mathbf{30.40 \pm 6.62}$	79.04 ± 21.97	26.20 ± 11.98
Low Noise for OOD	30.33 ± 6.64	$\mathbf{79.29 \pm 22.23}$	$\mathbf{25.77 \pm 12.53}$

Table 5: For the OOD + Additive model comparing training with OOD data at low noise versus high noise.

The results are summarized in Table 5 and demonstrate remarkably similar reconstruction quality between high-noise and low-noise OOD conditioning strategies. Across all three metrics the differences fall well within the margins of statistical error. The high-noise approach achieved marginally better PSNR, while the low-noise approach showed slight improvements in structural similarity and perceptual quality.

These findings suggest that the model's capacity to leverage OOD metadata for brain image reconstruction remains robust regardless of the noise level applied to the cardiac training data. Moreover, we see that metadata conditioning provides consistent benefits across different noise schedules, offering flexibility in implementation without compromising reconstruction quality.

4.7 Ensemble Reconstruction for Enhanced Reliability

To address the inherent stochasticity in diffusion-based models for MRI reconstruction, we implement a multiple sampling strategy as proposed by Shen et al [19]. This approach involves generating several reconstructions from the same undersampled k-space data and subsequently averaging them

Method	PSNR ↑	SSIM $\cdot 10^2$ ↑	LPIPS $\cdot 10^2$ ↓
Single Sample	30.40 ± 6.62	79.04 ± 21.97	26.20 ± 11.98
5 Sample Reconstruction Average	34.60 ± 7.11	85.53 ± 20.12	17.32 ± 11.51

Table 6: For the OOD + Additive model comparing single sampling to multi-sampling metrics.

at the pixel level. The method leverages the probabilistic nature of diffusion models, which can produce diverse yet plausible reconstructions from identical input conditions. For each undersampled acquisition, we generate five independent reconstructions by varying the random seed, then compute their pixel-wise average as the final output, reconstructed image.

Our primary motivation was to decrease the variance in quality metrics across different imaging samples, which exhibited substantial fluctuations (± 6 PSNR). While the variance across different imaging samples remained relatively constant despite our interventions, we observed significant improvements in both the robustness and overall performance of our method. Our results are summarized in Table 6. This ensemble approach effectively mitigates the impact of individual poor reconstructions, providing more reliable and higher-quality outputs. While this technique enhances reconstruction fidelity, it introduces a computational trade-off, as multiple sampling increases reconstruction time linearly with the number of samples. Furthermore, the averaging process may introduce slight smoothing that affects fine detail preservation, though the overall improvement in reconstruction quality generally outweighs this limitation.

5 Limitations and Future Works

Our study proposes multiple advancements in metadata conditioning for diffusion-based MRI reconstruction models. However, the results cannot fully reflect the effectiveness of these methods for multiple reasons. First, computational resources significantly constrained our experimental design. Due to limited access to GPU infrastructure, specifically only a few NVIDIA V100 GPUs at any given time, we could not match the computational scale utilized by previous studies like ContextMRI, which employed eight NVIDIA H100 GPUs with substantially larger effective batch sizes. While gradient accumulation and LoRA fine-tuning were employed to mitigate these constraints, the resulting training stability and performance were suboptimal. Future efforts should focus on scaling computational resources to match existing benchmarks, enabling more robust and comprehensive evaluations.

Second, dataset limitations also affected our analysis. Due to the large size and storage requirements of MRI datasets, we trained and evaluated our models on subsets of available data. Consequently, the generalizability and robustness of our models across diverse patient populations and scan protocols might not be fully captured. Future work should incorporate larger-scale experiments, utilizing complete datasets and potentially diverse imaging modalities to rigorously evaluate generalization and performance across varied clinical scenarios.

Additionally, although our fine-grained embedding approach demonstrated some benefits, further exploration into embedding optimization is desired. Alternative numerical encoding strategies beyond sinusoidal embeddings could be investigated, including learned linear or non-linear mappings tailored explicitly to clinical parameters.

Lastly, future research should extend the proposed methods to additional imaging contexts, such as dynamic MRI, functional MRI (fMRI), or multi-modal imaging scenarios. Incorporating patient-specific metadata like demographic or pathological information into these settings could further enhance model interpretability and diagnostic performance. Moreover, investigating metadata conditioning under clinically realistic scenarios, including incomplete or noisy metadata inputs, would improve the practicality and applicability of diffusion-based MRI reconstruction in real-world clinical workflows.

6 Conclusion

This project introduces a novel, structured approach to embedding MRI metadata within diffusion-based reconstruction models, significantly enhancing reconstruction quality and interpretability. By explicitly separating numerical and categorical metadata and employing dedicated embedding methods, we offer fine-grained control over metadata integration, which addresses the limitations of previous generalized approaches such as ContextMRI. Our results demonstrate substantial improvements in reconstruction accuracy, particularly in challenging scenarios involving uncommon or diverse MRI acquisition parameters. Integrating out-of-distribution cardiac MRI data into our training regimen further reinforces the effectiveness of our approach, compelling the model to better utilize metadata embeddings for reliable reconstructions.

Future research directions include scaling experiments with larger computational resources and datasets, exploring alternative numerical embedding strategies, and extending this approach to dynamic, functional, or multi-modal imaging contexts. Investigating the robustness of our method under real-world clinical conditions with incomplete or noisy metadata will also be critical. Ultimately, our approach paves the way toward more robust, interpretable, and clinically viable MRI reconstruction models.

References

[1] Michael Lustig, David Donoho, and John M Pauly. Sparse mri: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[3] Hyungjin Chung and Jong Chul Ye. Score-based diffusion models for accelerated mri. *Medical image analysis*, 80:102479, 2022.

[4] Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. *arXiv preprint arXiv:2303.05754*, 2023.

[5] Hyungjin Chung, Dohun Lee, Zihui Wu, Byung-Hoon Kim, Katherine L Bouman, and Jong Chul Ye. Contextmri: Enhancing compressed sensing mri through metadata conditioning. *arXiv preprint arXiv:2501.04284*, 2025.

[6] Walter HL Pinaya, Petru-Daniel Tudosiu, Jessica Dafflon, Pedro F Da Costa, Virginia Fernandez, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. Brain imaging generation with latent diffusion models. In *MICCAI Workshop on Deep Generative Models*, pages 117–126. Springer, 2022.

[7] Anna Zapaishchykova, Benjamin H Kann, Divyanshu Tak, Zezhong Ye, Daphne A Haas-Kogan, and Hugo JWL Aerts. Synthbraingrow: Synthetic diffusion brain aging for longitudinal mri data generation in young people. In *MICCAI Workshop on Deep Generative Models*, pages 75–86. Springer, 2024.

[8] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[9] Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastmri: An open dataset and benchmarks for accelerated mri. *arXiv preprint arXiv:1811.08839*, 2018.

[10] Martin Uecker, Peng Lai, Mark J. Murphy, Patrick Virtue, Michael Elad, John M. Pauly, Shreyas S. Vasanawala, and Michael Lustig. Espirit—an eigenvalue approach to autocalibrating parallel mri: Where sense meets grappa. *Magnetic Resonance in Medicine*, 71(3):990–1001, 2014.

[11] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[12] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arxiv 2021. *arXiv preprint arXiv:2106.09685*, 2021.

[16] Chengyan Wang, Jun Lyu, Shuo Wang, Chen Qin, Kunyuan Guo, Xinyu Zhang, Xiaotong Yu, Yan Li, Fanwen Wang, et al. Cmrxrecon: A publicly available k-space dataset and benchmark to advance deep learning for cardiac mri. *Scientific Data*, 11:687, 2024.

[17] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[18] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[19] Guoyao Shen, Mengyu Li, Chad W. Farris, Stephan Anderson, and Xin Zhang. Learning to reconstruct accelerated mri through k-space cold diffusion without noise. *Scientific Reports*, 14:21877, 2024.

A Distribution of fastMRI Brain Metadata

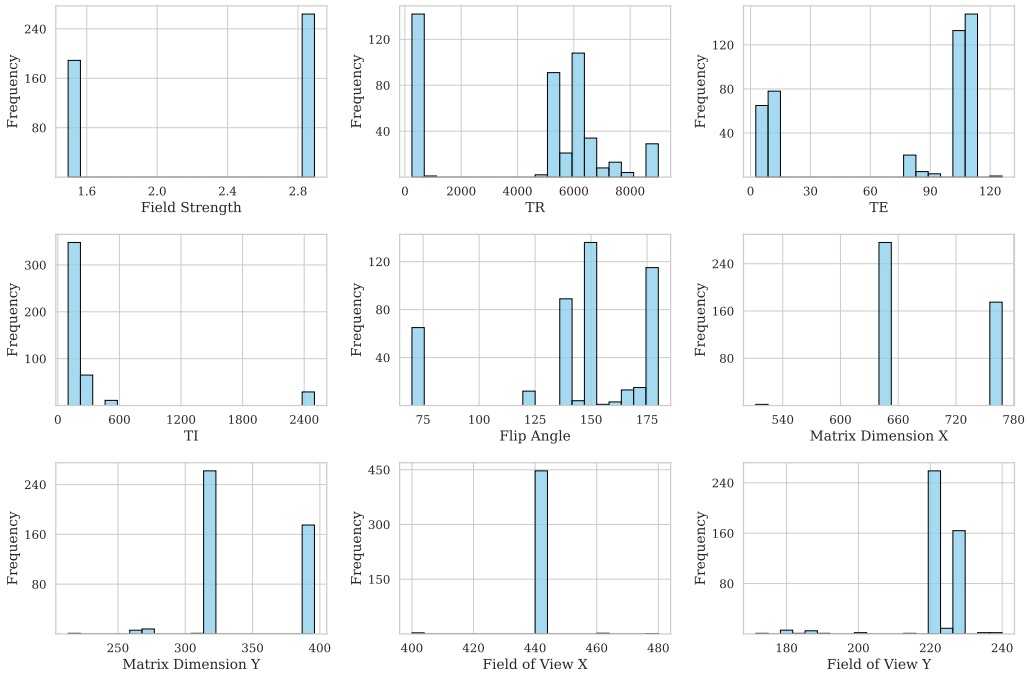


Figure 6: Distribution of the fastMRI continuous-valued metadata.

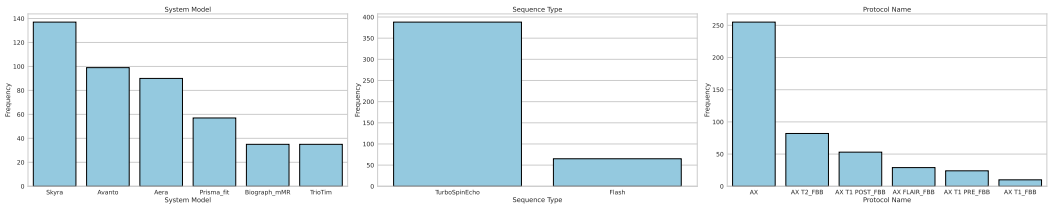


Figure 7: Distribution of the fastMRI categorical metadata.

B Distribution of OOD Cardiology Metadata

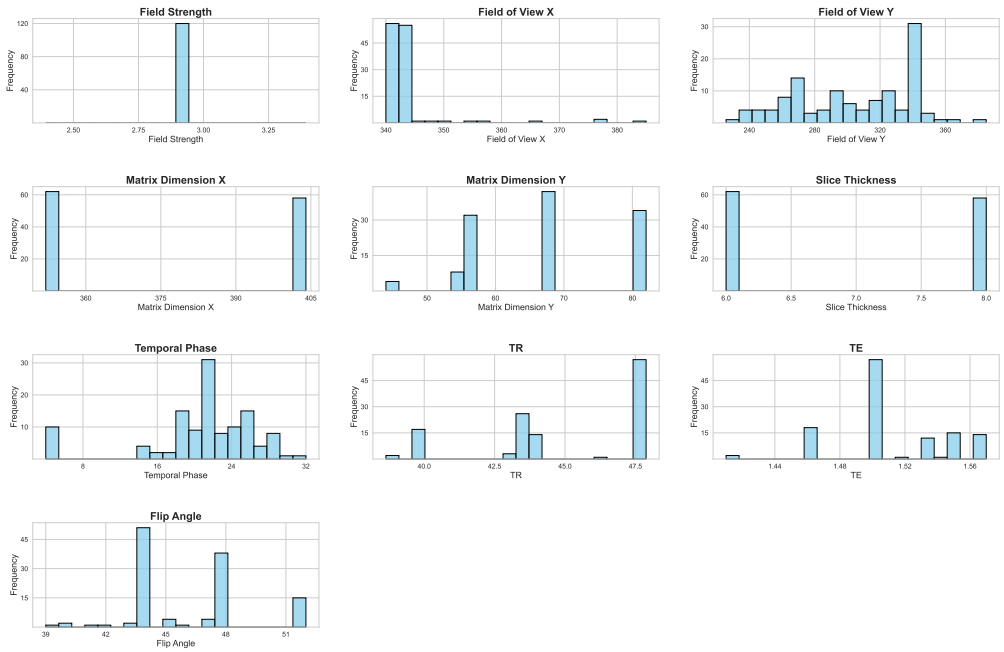


Figure 8: Distribution of the OOD continuous-valued metadata.

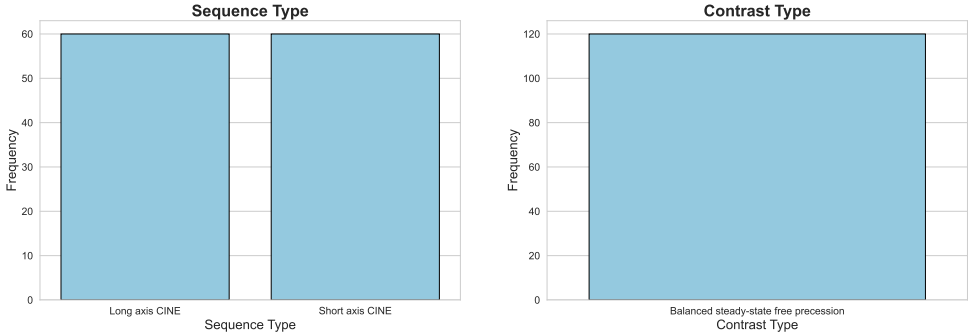


Figure 9: Distribution of the OOD categorical metadata.

C Visualizing ContextMRI's Attention Layer Weights

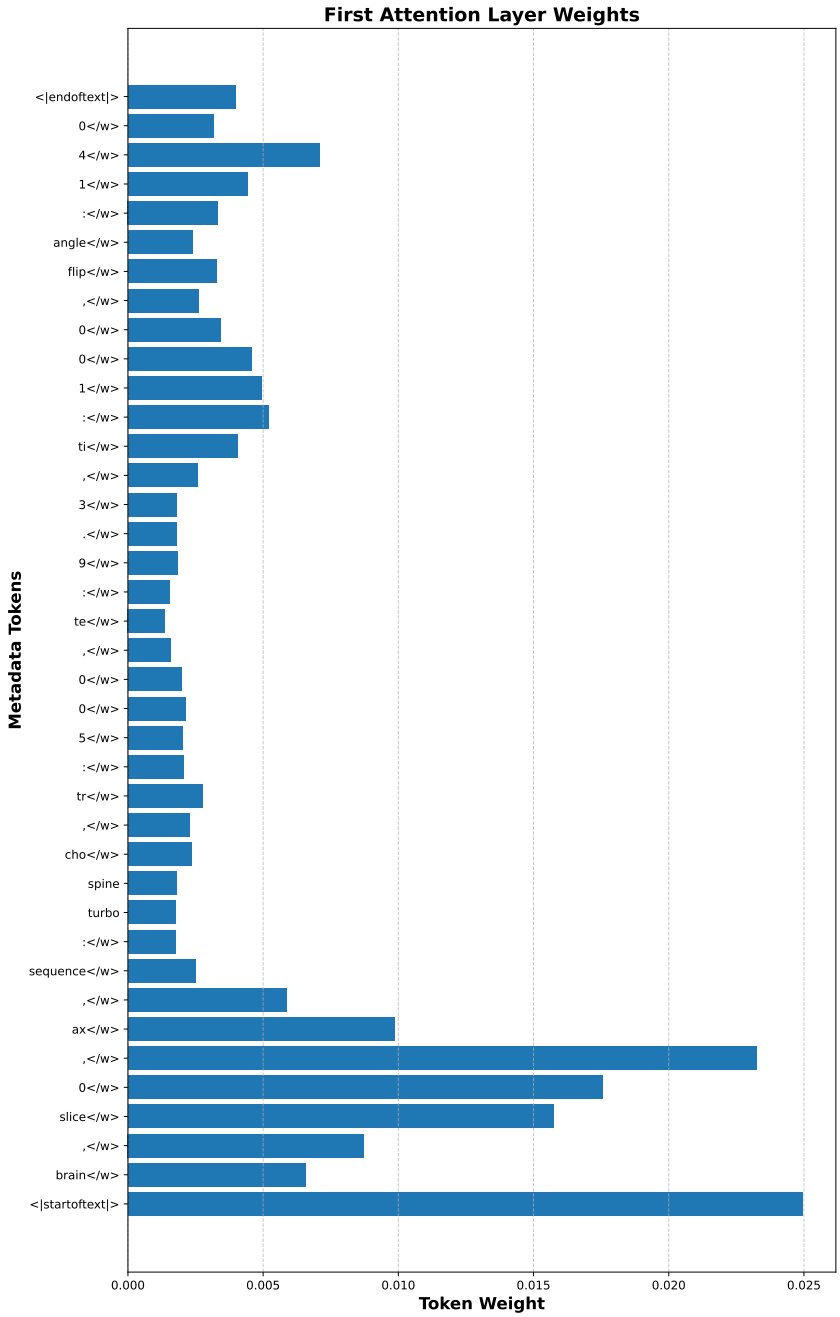


Figure 10: A visualization of ContextMRI's first attention layer. We can see that the CLS (start token) and the slice-level token have the most weight.

Here we analyze whether the ContextMRI model's UNet architecture effectively utilizes the text conditioning information during inference. We probed the cross-attention layers of the diffusers.UNet2DConditionModel to extract attention maps that indicate how much the model attends to different dimensions of the input text embeddings. From our plot we see that the slice token and CLS token (summary token of the entire embedding) have significant weight, while the other metrics are largely ignored.

D Performance of Models on OOD Cardiology Data

To assess the effectiveness of our metadata conditioning approach for cross-domain generalization, we compared two models, our OOD + Additive approach and a finetuned ContextMRI version, on the CMRxRecon cardiac dataset. First, we fine-tuned the baseline ContextMRI model using the same 1:5 sampling ratio (OOD cardiology data to fastMRI data) employed in our OOD + Additive approach. When finetuning ContextMRI, we followed the exact same protocol as we did for training our OOD + Additive model.

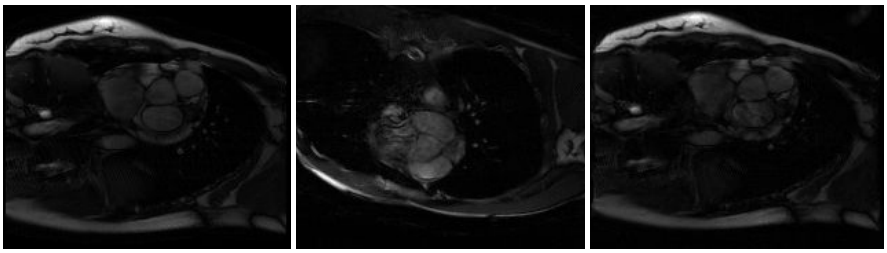
We tested our models on 400 cardiology images, half were CINE-SAX and the remaining CINE-LAX. The quantitative results in Table 7 reveal a clear trend across all evaluation metrics. The OOD + Additive model outperforms the fine-tuned ContextMRI model in terms of all metrics. This performance advantage is notable because it demonstrates that our metadata conditioning strategy enables better generalization than traditional fine-tuning approaches.

Moreover, the fine-tuned ContextMRI model exhibits higher variability in its performance metrics, particularly for SSIM, suggesting less consistent reconstruction quality across different cardiac images. In contrast, the OOD + Additive model produces more reliable results with lower standard deviations across all metrics.

In Figure 11 we show an example image comparing reconstruction quality of the finetuned ContextMRI to the OOD + Additive model.

Method	PSNR ↑	SSIM $\cdot 10^2$ ↑	LPIPS $\cdot 10^2$ ↓
OOD + Additive	35.30 ± 4.39	89.72 ± 7.91	21.54 ± 4.66
Finetuned ContextMRI	30.20 ± 4.68	64.43 ± 18.18	36.08 ± 7.15

Table 7: Comparing quantitative metrics for the OOD + Additive model vs a Finetuned ContextMRI model on the OOD cardiology data.



Label Image Finetuned ContextMRI OOD + Additive

Figure 11: A visualization of a Label (true) image against a reconstruction created by the Finetuned ContextMRI model and the OOD + Additive model for the cardiology OOD dataset.