Improving MRI Models using Metadata

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 powerful diagnostic tool, but its long acquisition times limit clinical throughput. Compressed sensing (CS) aims to accelerate MRI by reconstructing full-resolution images from undersampled data. Recent work such as ContextMRI leverages diffusion models conditioned on textual metadata to improve CS reconstruction quality. However, these approaches rely on CLIP text encoders that may not effectively represent numerical scan attributes like TR, TE, and flip angle. In this work, we propose an improved metadata conditioning framework that encodes each metadata attribute separately, using discrete embeddings for categorical variables and sinusoidal encodings for normalized numerical values. We also explore a hybrid approach that combines our structured conditioning with CLIP embeddings for compatibility with pre-trained models. Preliminary experiments on the fastMRI brain dataset demonstrate that our method can match or outperform the CLIP-only baseline, especially for slices with uncommon or varied acquisition parameters. These findings suggest that more precise handling of metadata can enhance diffusion-based MRI reconstruction and pave the way for more adaptable and interpretable CS models.

1 Introduction

Magnetic Resonance Imaging (MRI) plays a critical role in medical diagnostics, yet its long acquisition times remain a key bottleneck in clinical workflows. Compressed Sensing (CS) has emerged as a promising solution by enabling reconstruction of high-fidelity images from sparsely sampled k-space data. While early CS approaches relied on handcrafted priors or traditional regularization schemes, recent advances in generative modeling—particularly denoising diffusion probabilistic models (DDPMs)—have significantly improved reconstruction fidelity by learning rich priors from data.

Diffusion-based inverse solvers such as DDS and Score-MRI offer powerful frameworks for solving ill-posed imaging problems through stochastic posterior sampling. These methods have shown strong results across a variety of imaging modalities and undersampling schemes, highlighting the versatility of learned generative priors.

One notable advancement is ContextMRI, which incorporates clinically relevant metadata—such as imaging parameters, anatomical region, and pathology—into a text-conditioned diffusion model for MRI reconstruction. By leveraging a frozen CLIP text encoder, ContextMRI converts structured metadata into embeddings that guide the generative process. However, this approach presents key

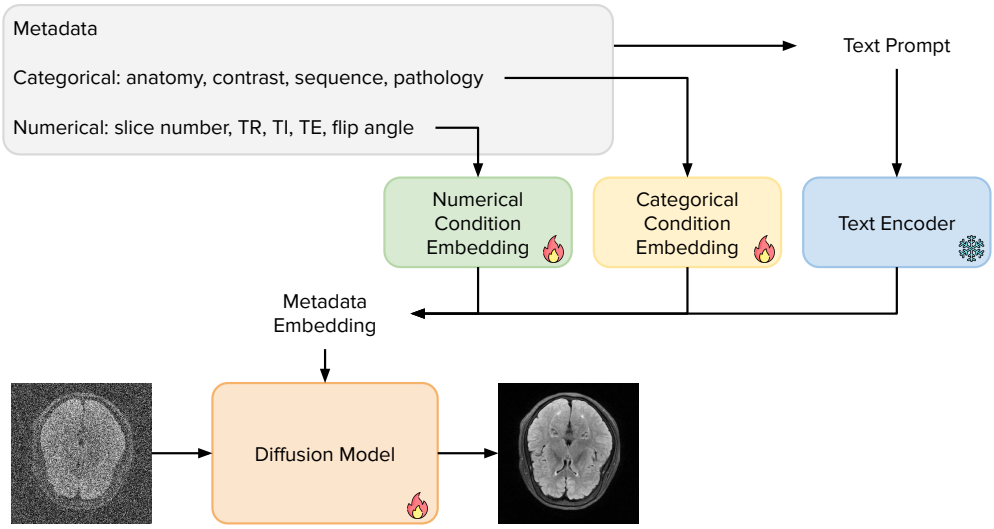

Figure 1: Proposed metadata conditioning MRI diffusion model framework. The original ContextMRI framework [1] only uses the text encoder. This project proposes to encode metadata with separate numerical and categorical embeddings to represent conditions explicitly.

limitations: the CLIP encoder was not trained on medical text or numerical values, and may fail to faithfully represent continuous scan parameters such as TR, TE, or flip angle, which are critical for accurate image synthesis.

In this work, we propose a more granular and structured approach to metadata conditioning in MRI reconstruction. Instead of representing all metadata as free-form text, we separately encode each attribute based on its type: categorical values are mapped via embedding lookup tables, while continuous values are normalized and embedded using sinusoidal encodings. We also introduce a hybrid conditioning scheme that combines these structured embeddings with the original CLIP embeddings, enabling compatibility with pre-trained ContextMRI weights and reducing the need to retrain from scratch.

Our preliminary experiments on the fastMRI brain dataset demonstrate that the proposed conditioning method provides more stable and accurate reconstructions under limited-resource settings. These results suggest that fine-grained metadata representations can improve model interpretability and robustness—especially in clinical scenarios involving varied imaging protocols.

2 Methods

2.1 Background

We build upon the ContextMRI framework [1], which uses a diffusion model trained on complex-valued MR images and conditioned on metadata for better MR image generation. The metadata for each MR scan is first transformed into a text prompt by concatenating all information into a string. 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`." Then, the text prompt is passed through a pre-trained and frozen text encoder like CLIP text encoder [2] to extract continuous representations as the condition embedding for the diffusion model.

Although ContextMRI injects metadata into the diffusion generation process, the improvements shown in [1] are insignificant since the performance difference between unconditioned and metadata-conditioned reconstruction is less than the standard deviation of multiple metrics. We suspect the cause of this phenomenon might be 1) the text encoder was not trained with this type of text prompt, and 2) the text encoder cannot accurately encode the numerical values. Hence, in the next section, we present an improved metadata conditioning approach via fine-grained condition embeddings to address these issues.

To illustrate the diverse nature of the metadata and motivate the need for specialized encoding methods, we provide visualizations of metadata distributions for both continuous (Figure 2) and categorical attributes (Figure 3) in the appendix.

2.2 Fine-grained Metadata Condition Embedding

We propose a fine-grained metadata conditioning method to encode MRI metadata into continuous representations effectively. For each metadata attribute, we use a separate embedding module according to the attribute's data type. For categorical attributes like contrast type and pathology, we use an embedding table similar to the text embedding layer in deep learning-based language models to transform each possible category into continuous vectors. For continuous numerical attributes like TR and TE, we normalize the values to between zero and one and apply a sinusoidal positional encoding to map the normalized values into continuous embeddings, analogous to how diffusion models encode the timestep.

Following ContextMRI, we use classifier-free guidance (CFG) [3] to improve generation quality. We proposed two condition embedding dropout approaches for CFG training. First, we drop all conditions simultaneously with a probability p_{CFG}. The second method drops each fine-grained condition embedding independently, simulating real-world scenarios where not all metadata attributes are always present. The overall framework is shown in Figure 1.

3 Preliminary Experiments

We use the official ContextMRI codebase and checkpoint to evaluate baseline reconstruction quality on a subset of the fastMRI brain dataset.[1] Due to limited computing resources, we scaled down the experiments regarding dataset size, batch size, and trial count.

3.1 Text Prompt Embedding Probing

We first examine whether the metadata information can be recovered from the metadata encoder in ContextMRI by probing the CLIP text embeddings with a light-weight prediction head. We randomly sample metadata attributes and combine them into text prompts. The text prompts are then passed through the CLIP text encoder to extract a sequence of embeddings. We use separate prediction heads to perform regression and classification tasks to recover numerical and categorical metadata. Each prediction head aggregates the text embedding sequence into a single vector with a multi-head attention layer and a learnable query vector. Then, the vector is passed through a linear layer to predict either a single value or multiple logits, depending on the task. For numerical attributes, the learning target value x is first normalized to a range of $[0, 1]$ by

$$\hat{x} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}, \tag{1}$$

where x_{max} and x_{min} are respectively the maximum and minimum possible values of the attribute.

3.2 ContextMRI and Small Scale Fine-tuning

First, we install and run the original ContextMRI codebase and model to set a baseline for CLIP text encoder-based metadata conditioning. Next, we run a small-scale fine-tuning using our proposed conditioning strategies. We use Peak Signal-to-noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) as evaluation metrics. Note that we made some mistakes with the proposed fine-grained condition embeddings, making the current results suboptimal.

4 Preliminary Findings

4.1 Text Prompt Embedding Probing

As shown in Table 1, numerical attributes can be recovered from the CLIP text embeddings with a low RRMSE, implying that numerical values can be embedded with a text-based encoder but with

[1] `https://github.com/DoHunLee1/ContextMRI`

Numerical (RRMSE ↓)				Categorical (ACC ↑)			
TR	TE	TI	Flip Angle	Anatomy	Contrast	Sequence	Pathology
6.4%	4.5%	5.2%	3.2%	100%	100%	100%	100%

Table 1: Text prompt embedding probing experiments. Numerical and categorical attributes are evaluated with relative root-mean-squared error (RRMSE) and accuracy (ACC), respectively.

some precision loss. Meanwhile, the categorical attributes in the metadata can be easily recovered since these attributes are originally represented in text form. Although the metadata can mostly be effectively encoded with the CLIP text encoder, it remains unclear whether the diffusion model is utilizing this information efficiently. Hence, this motivates us to propose a more explicit way to encode metadata.

4.2 ContextMRI and Small Scale Fine-tuning

Our preliminary experiments indicate promising improvements in reconstruction quality using our fine-grained metadata conditioning approach compared to the original pre-trained ContextMRI. As shown in Table 2, at a guidance scale of 1.0, our model achieves a PSNR of 30.32 ± 6.25, outperforming ContextMRI's baseline PSNR of 28.39 ± 6.23.

These improvements suggest that structured, attribute-specific embeddings provide a more precise representation of numerical scan parameters, leading to better-informed image generation by the diffusion model. Interestingly, our results for ContextMRI differ slightly from those reported in the original publication, likely due to testing only on a smaller subset (test_batch_0 with 200 patients) rather than the full evaluation dataset used in their paper (due to limited computational resources). These preliminary findings highlight the potential of fine-grained metadata embeddings to enhance the fidelity and robustness of diffusion-based MRI reconstruction.

Method	**Guidance**	PSNR ↑	SSIM $\cdot 10^2$ ↑	LPIPS $\cdot 10^2$ ↓
Our Model	0.0 (Uncond)	$--$	$--$	$--$
	1.0	30.32 ± 6.25	78.64 ± 22.45	26.16 ± 12.27
	2.0	$--$	$--$	$--$
	3.0	$--$	$--$	$--$
ContextMRI	0.0 (Uncond)	$--$	$--$	$--$
	1.0	28.39 ± 6.23	65.84 ± 22.05	33.31 ± 11.87
	2.0	$--$	$--$	$--$
	3.0	$--$	$--$	$--$

Table 2: Quantitative preliminary results on the FastMRI brain dataset (Uniform 1D mask, Acceleration $\times 4$). We compare our model against ContextMRI across different classifier-free guidance scales.

5 Planned Experiments

5.1 Better Text Encoder

For the current project, we are using the simple CLIP text encoder to encode some of the metadata. However, there are evidences showing that CLIP text encoder might not be the best to capture the meaning of the words. Instead, we can use specialized text encoders like T5 to perform better encoding.

5.2 Better Fine-tuning Approach (LoRA)

For our current project, we are using the simplest fine-tuning method to incorporate our new method. However, because of the limitation in the number of data, fine-tuning on this small subset of data might lead to overfitting and poor generalization. To address this, we plan to explore a more efficient and scalable fine-tuning appraoch using Low-Rank Adaptation (LoRA). LoRA enables adaption of large pretrained models by injecting trainable rank-decomposition matrices into each layer's weights, significantly reducing the number of parameters that need to be updated during training. This not only makes fine-tuning more memory- and compute-efficient, but also helps preserve the generalization capability of the base model. By integrating LoRA with metadata, we aim to guide the model to learn more context-aware diffusion representations, improving performance across different MRI sources.

5.3 Out-of-distribution Data

We plan to explore using out-of-distribution data to supplement the FastMRI brain scans. The current dataset contains very few possible values for many of the metadata attributes (see Appendix A for metadata distributions). This may make it difficult for the model to make use of the continuous valued metadata features. By incorporating more MRIs, we hope to get better coverage. We plan to make use of MRIs of anatomy (such as knees) at high noise levels as one data source.

5.4 Scaling Up Training Data

To ensure stable and effective training, we plan to scale up the training dataset and the batch size. Choosing a suitable batch size is important for training stability. Given that the original configuration uses a batch size of 64, scaling up may require more powerful computing resources.

5.5 Impact of each attribute

We will conduct ablation studies comparing three configurations: CLIP-only conditioning (baseline), our hybrid conditioning approach, and structured conditioning without CLIP.

Next, we aim to assess the relative importance of individual metadata attributes by systematically removing or isolating them and measuring the impact on reconstruction quality. This will help us understand which metadata signals are most valuable for guiding image synthesis.

Additionally, we will perform qualitative evaluations by comparing reconstructions with and without specific metadata cues. If time and compute allow, we will also explore using more advanced diffusion architectures to further improve image fidelity.

5.6 Cross-attention visualization

In our planned cross-attention visualization experiment, we will analyze whether the ContextMRI model's UNet architecture effectively utilizes the text conditioning information during inference. Specifically, we will probe 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. By visualizing these maps at various timesteps in the reverse diffusion process, we aim to understand how much specific anatomical or acquisition-related tokens in the conditioning vector influence reconstruction outcomes.

References

[1] 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.

[2] 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.

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

A Contributions Sections

- Joanna: Created data pre-processing scripts (to generate ESPIRIT maps and slice files) following the methodology described in ContextMRI. Plotted metadata. Prepared data on server and ran train/inference scripts on server.

B Distribution of Metadata

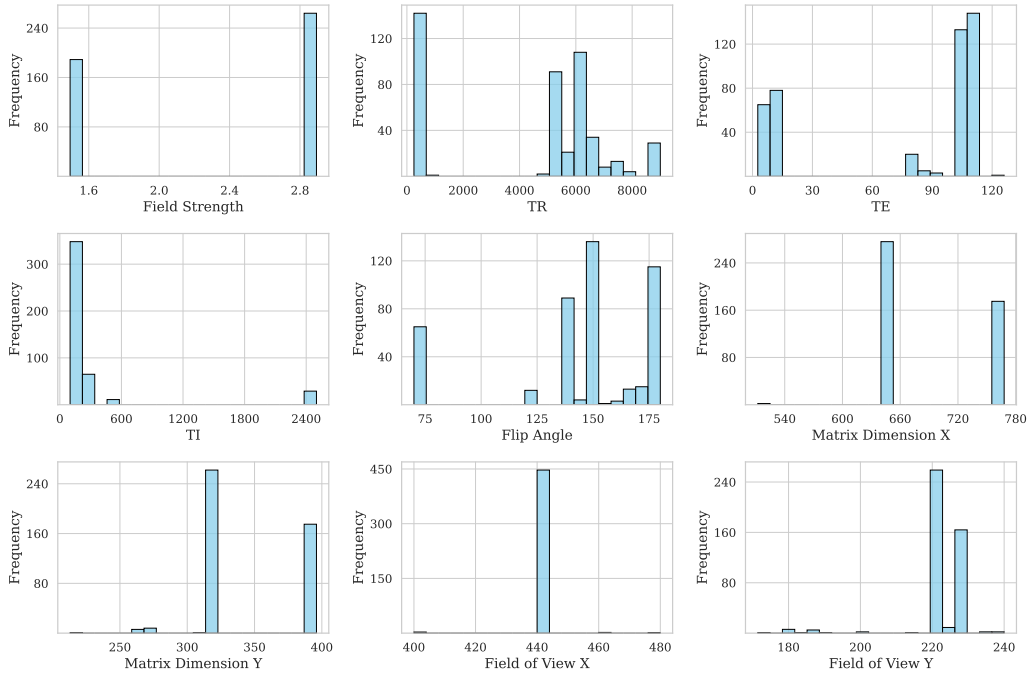

Figure 2: Distribution of the continuous-valued metadata.

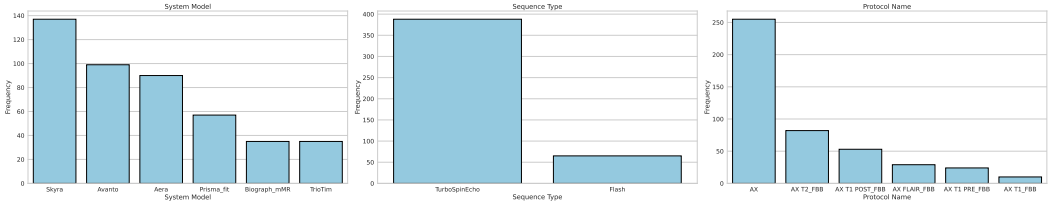

Figure 3: Distribution of the categorical metadata.

C Qualitative visualization of our conditioning model

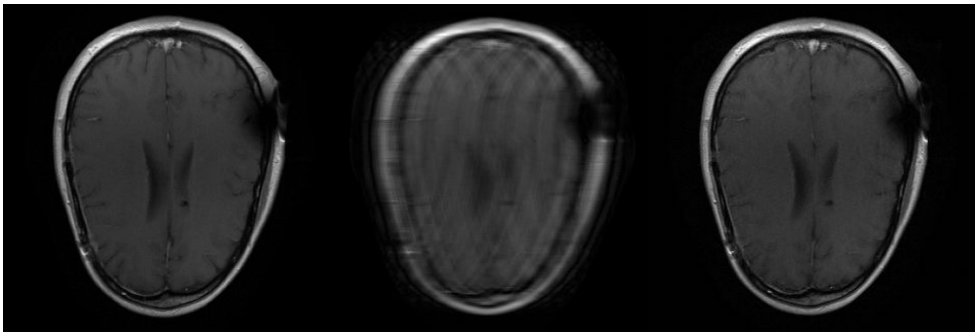

Figure 4: A visualization of one of our conditioned model's successful outputs.