UEPS: ROBUST AND EFFICIENT MRI RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep unrolled models (DUM) have become the state-of-the-art for accelerated MRI reconstruction, yet their robustness under domain shift remains a critical barrier to clinical adoption. In this work, we identify coil sensitivity map (CSM) estimation as the primary bottleneck limiting generalization. To address this, we propose UEPS, a novel DUM architecture featuring three key designs: (i) an Unrolled Expanded (UE) design that eliminates CSM dependency by reconstructing each coil independently; (ii) progressive resolution, which leverages k-space-to-image mapping for efficient coarse-to-fine refinement; and (iii) sparse attention tailored to MRI's 1D undersampling nature. These physics-grounded designs enable gains in robustness and computational efficiency. We construct a large-scale zero-shot transfer benchmark comprising 10 out-of-distribution (OOD) test sets spanning diverse clinical shifts—anatomy, view, contrast, vendor, field strength, and coil configurations. Extensive experiments demonstrate that UEPS consistently and substantially outperforms existing DUM, end-to-end, diffusion, and untrained methods across all OOD tests, achieving state-of-the-art robustness with low-latency inference suitable for real-time deployment.

1 INTRODUCTION

Magnetic resonance imaging (MRI) provides excellent soft-tissue contrast without ionizing radiation, but suffers from long acquisition times. Undersampled MRI aims to reconstruct high-fidelity images from a fraction of measurements in k-space to accelerate MRI scans. State-of-the-art methods predominantly use deep unrolled models (DUM) which integrate a physical forward model of the imaging process into learnable neural networks and are typically trained on large-scale datasets comprising multi-coil k-space measurements (Hammernik et al., 2018; Sriram et al., 2020a).

Despite their success on benchmarks with abundant training data, DUM struggle with domain shift and transfer poorly to unseen acquisition protocols (Knoll et al., 2019; Johnson et al., 2021). To enhance robustness, recent studies have explored training DUM on combined datasets aggregated from multiple sources (Lin & Heckel, 2023). However, MRI's wide contrast variability makes it difficult to curate a comprehensive training dataset. Untrained methods offer an alternative by avoiding dependence on training data and optimizing independently for each test case (Yaman et al., 2021; Darestani & Heckel, 2021; Feng et al., 2023). These methods are less susceptible to spurious correlations tied to a specific training distribution, but the limited information available in each undersampled input can constrain reconstruction quality (Darestani et al., 2021).

To address this robustness issue, we first identify the primary generalization bottleneck stems from the dependency of DUM on coil sensitivity maps (CSM). In multi-coil MRI, now the clinical standard, multiple receiver coils simultaneously acquire data, each with a characteristic CSM describing its spatial reception profile. CSM are used in DUM to convert individual coil images into a single image or vice versa. Our analysis reveals that learned CSM estimation becomes less reliable under domain shift, and the resulting inaccuracies can propagate through the CSM-dependent reconstruction pathway, potentially exacerbating reconstruction errors. We propose a simple yet effective redesign of DUM, termed Unrolled Expanded model (UE), which reconstructs each individual coil image separately, thereby eliminating the need for explicit CSM estimation. Reconstructing coil images before their final combination has precedents such as GrappaNet (Sriram et al., 2020b); we are the first to systematically investigate independent coil-wise reconstruction for robustness under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

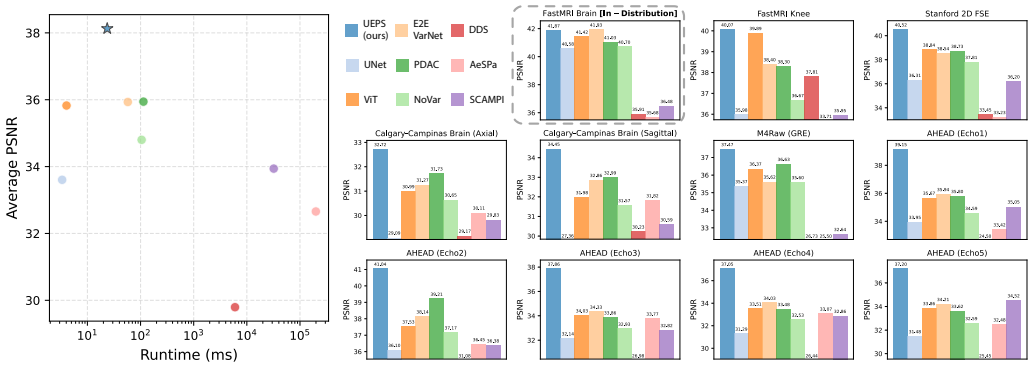

Figure 1: Zero-shot transfer benchmark results. Average PSNR: mean over 10 OOD sets. Runtime: measured on FastMRI Knee (15 coils). Compared methods include E2EVarNet (MICCAI 2020), PDAC (CVPR 2024), NoVar (CVPR 2025), DDS (ICLR 2024), AeSPa (CVPR 2025), and SCAMPI (MRM 2024), alongside UNet and ViT baselines.

domain shift. UE gives up explicit modeling of inter-coil correlations, yet achieves comparable in-distribution (ID) performance and improved out-of-distribution (OOD) performance. However, processing coils independently increases computation during both training and inference. To make this design more scalable, we introduce two features for efficiency as follows.

Unlike typical image-to-image tasks like image restoration (Liang et al., 2021), MRI reconstruction is a k-space-to-image mapping, suggesting that the standard down-sampling up-sampling architecture may be suboptimal here. We propose progressive resolution, whose core innovation lies in performing up-sampling by k-space expansion. Crucially, additional measurements are incorporated for the expanded k-space region, adding authentic high-frequency information, which cannot be achieved by image-space interpolation. In this up-sampling-only architecture, early stages run at lower resolution with substantially reduced computation. An additional benefit is the implicit curriculum learning: initial stages with denser k-space sampling pose an easier task, stabilizing training and providing improved initialization for subsequent high-resolution refinement. Sparse attention serves as another key feature for accelerating our transformer-based model. Recognizing that 1D MRI undersampling produces aliasing along a single axis, our sparse attention allocates more computation to this axis, achieving computational savings where full attention is unnecessary. Combining UE with progressive resolution and sparse attention yields UEPS.

In this work, we revisit the architecture of DUM to improve robustness under domain shifts in anatomy, view, contrast, vendor, field strength, and coil configuration while reducing reconstruction time. Our main contributions are as follows:

1. Motivated by our analysis of CSM dependence as a generalization bottleneck, we propose UE that improves OOD reconstruction while maintaining comparable ID performance, albeit at increased computational cost.
2. We further introduce progressive resolution and sparse attention, two design features grounded in the k-space-to-image mapping and 1D acceleration nature of MRI physics, which yield marked computational acceleration without compromising, and in fact improving, reconstruction quality.
3. To rigorously assess generalization, we establish a large-scale zero-shot transfer benchmark and show that UEPS outperforms all competing approaches by a significant margin, while maintaining acceptable runtime.

2 Related Work

2.1 Deep Unrolled Models

DUM represent a class of methods that unroll iterative optimization algorithms into network layers, combining the rigor of physics-based models with the representational power of deep learning

(Hammernik et al., 2018; Schlemper et al., 2017; Aggarwal et al., 2018). The evolution of DUM methods has seen a shift from using pre-computed CSM from external tools (Uecker et al., 2014; 2015) to jointly learning them with the reconstruction network (Sriram et al., 2020a; Jun et al., 2021), a paradigm that underpins all contemporary state-of-the-art (SOTA) DUM methods. Subsequent research has explored diverse architectural innovations to further improve performance, such as convolutional recurrent network (Yiasemis et al., 2022), hybrid convolutional-transformer network (Fabian et al., 2022), memory efficient transformer (Darestani et al., 2024), neural operator-based architecture (Jatyani et al., 2025), and leveraging adjacent slices input (Xin et al., 2024). Despite architectural advances, these works predominantly report performance on ID tests of large datasets like FastMRI (Zbontar et al., 2018); the real-world robustness of these models remains an open question.

2.2 Multi-Resolution

Multi-resolution architectures are commonly used in low-level vision tasks such as image restoration and medical image reconstruction, leveraging coarse resolutions for computational efficiency and global context while preserving high-resolution pathways for spatial precision. In image restoration, early seminal works (Wang et al., 2022; Zamir et al., 2022) employ a transformer-based U-Shaped structure with downsampling and upsampling, to capture hierarchical multi-scale representations and maintain computational efficiency, which becomes gold standard for subsequent approaches (Zhou et al., 2025; He et al., 2025). In DUM-based MRI reconstruction, each cascade contains a denoiser, typically with a multi-resolution architecture, ranging from UNet in early works (Sriram et al., 2020a) to transformer-based (Darestani et al., 2024; Wang et al., 2025) or hybrid convolution-transformer designs (Fabian et al., 2022; Wang et al., 2024) in recent approaches. Note this multi-resolution design is within each cascade. To the best of our knowledge, no prior work has explored multi-resolution across cascades. Learned upsampling operations are used in these multi-scale models, such as transposed convolutions (Dumoulin & Visin, 2016) or pixel shuffle (Shi et al., 2016), which are well-suited for image-to-image tasks by learning to synthesize details. However, they cannot utilize the available high-frequency information present in k-space for MRI reconstruction.

2.3 Sparse Attention

Sparse attention has become essential for scaling transformers to high-resolution images, where standard self-attention's quadratic complexity becomes prohibitive. Various strategies have been proposed to restrict attention to relevant subsets to reduce computational cost while largely preserving representational power. Early breakthroughs focused on fixed sparsity patterns like shifted non-overlapping local windows (Liu et al., 2021), sliding window (Hassani et al., 2023), and factorized attention along axes (Wang et al., 2020; Dong et al., 2022). Beyond static patterns, dynamic sparse attention predicts the most relevant tokens for each query based on the input content. For instance, DynamicViT progressively prunes redundant visual tokens using a lightweight prediction module (Rao et al., 2021), and SpargeAttn predicts the sparse mask by compressing each block of queries or keys to a single token (Zhang et al., 2025). Although sparse attention has been extensively studied for vision and language tasks, it remains unexplored in MRI reconstruction.

3 Proposed Method

UEPS, our proposed architecture, is a DUM redesigned for robustness without sacrificing efficiency. It consists of three main features, the Unrolled Expanded model (UE), progressive resolution, and sparse attention, with the entire model shown in fig. 2. UE primarily improves robustness, but sacrifices speed. Progressive resolution and sparse attention, each motivated by MRI-specific principles, both reduce computation and further improve reconstruction quality.

3.1 Problem Formulation

Modern MRI uses multiple receiver coils, each of which measures Fourier components of the imaged volume multiplied by a complex-valued position-dependent CSM (Zbontar et al., 2018). The measured k-space data from the ith coil, $\mathbf{k}_i$, are given by

$$\mathbf{k}_i = \mathcal{F}(S_i\mathbf{m}) + \epsilon_i, \quad i = 1, 2, \ldots, N \tag{1}$$

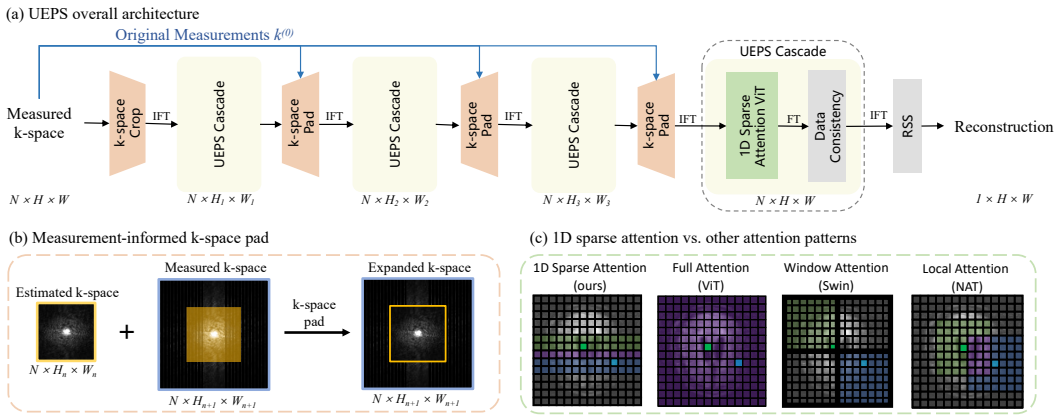

Figure 2: Overview of the proposed UEPS architecture, illustrating progressive resolution across cascades (e.g., 64 × 64, 128 × 128, 256 × 256 and 320 × 320) (a); schematic of measurement-informed k-space padding (b); and comparison of attention patterns, with different colours indicating the attention regions of different query tokens (c).

where $\mathbf{m}$ is the imaging object, S_i is the CSM of the ith coil, $\mathcal{F}$ is the Fourier transform operator, ϵ_i is measurement noise for the ith coil, and N is the number of coils. For fully sampled k-space data, the ith coil image $\hat{\mathbf{m}}_i$ can be reconstructed by an inverse Fourier transform:

$$\hat{\mathbf{m}}_i = \mathcal{F}^{-1}(\mathbf{k}_i) \tag{2}$$

Then, the individual coil images are combined with pixel-level root-sum-of-squares (RSS) to provide the final image estimate $\hat{\mathbf{m}}$ (Zbontar et al., 2018).

$$\hat{\mathbf{m}} = \mathrm{RSS}(\hat{\mathbf{m}}_1, \ldots, \hat{\mathbf{m}}_N) = \sqrt{\sum_{i=1}^{N} |\hat{\mathbf{m}}_i|^2} \tag{3}$$

In undersampled MRI, only part of the k-space is acquired to accelerate scan

$$\tilde{\mathbf{k}}_i = \mathcal{M}\mathcal{F}(S_i\mathbf{m}) + \epsilon_i, \quad i = 1, 2, \ldots, N \tag{4}$$

where $\mathcal{M}$ is the binary sampling mask and $\tilde{\mathbf{k}}_i$ denotes the undersampled k-space data. DUM solves this ill-posed inverse problem of reconstructing undersampled MRI by integrating the physical forward model eq. (1) into a cascaded neural network inspired by iterative optimization algorithms. Let $\mathbf{k}^{(0)} \in \mathbb{C}^{N\times H\times W}$ denote the Cartesian k-space data from all coils with unsampled points filled with zeros. H/W is the number of readout samples/phase encoding in k-space, equivalent to height/width in image-space for Cartesian MRI. Let $\mathbf{x}^{(t)} \in \mathbb{C}^{N\times H\times W}$ denote the intermediate multi-coil image estimate at the t-th cascade of the unrolled model, and $\mathbf{k}^{(t)} \in \mathbb{C}^{N\times H\times W}$ the corresponding intermediate k-space estimate. Each refinement stage can be formulated as below with image-space intermediate quantities

$$\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)} + \mathcal{E}(\mathcal{D}_\theta^{(t)}(\mathcal{R}(\mathbf{x}^{(t)}))) \tag{5}$$

$$\mathbf{k}^{(t)} = \mathcal{F}(\tilde{\mathbf{x}}^{(t)}) \tag{6}$$

$$\mathbf{k}^{(t+1)} = \mathcal{DC}_\eta(\mathbf{k}^{(t)}, \mathbf{k}^{(0)}) \tag{7}$$

$$\mathbf{x}^{(t+1)} = \mathcal{F}^{-1}(\mathbf{k}^{(t+1)}) \tag{8}$$

where $\mathcal{R}$ and $\mathcal{E}$ are reduction and expansion operators to convert multi-coil images into a single image or vice versa by leveraging CSM with details given by Sriram et al. (2020a). $\mathcal{D}_\theta^{(t)}$ is a learnable image denoiser, such as UNet (Sriram et al., 2020a), Restormer (Darestani et al., 2024), or ViT (Dosovitskiy et al., 2020) used in this work. $\mathcal{DC}_\eta$ is the data consistency module with learnable weight η. While originally inspired by the gradient descent formulation of compressed sensing

MRI, this module can accommodate various implementations, provided that data fidelity with acquired measurements is maintained. Notably, with the specific instantiation below with scalar η, eqs. (5) to (8) are mathematically equivalent to the formulation of Sriram et al. (2020a).

$$\mathbf{k}^{(t+1)} = \mathcal{DC}_\eta(\mathbf{k}^{(t)}, \mathbf{k}^{(0)}) = \mathbf{k}^{(t)} - \eta\mathcal{M}(\mathbf{k}^{(t)} - \mathbf{k}^{(0)}) \tag{9}$$

$\mathbf{x}^{(0)}$, derived by eq. (2), serves as an initial estimate with aliasing artifacts. The iterative updates concluded with coil combination step as in eq. (3), form the complete MRI reconstruction pipeline.

3.2 Unrolled Expanded Model

UE reformulates reconstruction in terms of individual coil images, removing the CSM-based reduction and expansion operations in eq. (5).

$$\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)} + \mathcal{D}_\theta^{(t)}(\mathbf{x}^{(t)}) \tag{10}$$

$$\mathbf{k}^{(t)} = \mathcal{F}(\tilde{\mathbf{x}}^{(t)}) \tag{11}$$

$$\mathbf{k}^{(t+1)} = \mathcal{DC}_\eta(\mathbf{k}^{(t)}, \mathbf{k}^{(0)}) \tag{12}$$

$$\mathbf{x}^{(t+1)} = \mathcal{F}^{-1}(\mathbf{k}^{(t+1)}) \tag{13}$$

where $\mathcal{D}_\theta^{(t)}$ takes $\mathbf{x}^{(t)} \in \mathbb{C}^{N\times H\times W}$ as input, rather than $\mathcal{R}(\mathbf{x}^{(t)}) \in \mathbb{C}^{H\times W}$ in section 3.1. We treat multi coils as a batch dimension, keeping $\mathcal{D}_\theta^{(t)}$ unchanged for fair comparison with standard DUM. This design trades explicit modeling of inter-coil correlations for independence from estimated CSMs. This additional batch dimension increases computation, which is alleviated by the features introduced below.

3.3 Progressive Resolution

Progressive resolution is designed for efficient reconstruction of high-resolution images by delaying high-resolution processing to later stages, minimizing computation in early cascades. Our core innovation is performing up-sampling via k-space expansion, which populates the extended frequency region with actual high-frequency measurements that image-space methods cannot provide. This novel up-sampling operation integrates naturally into the unrolled pipeline, leading to the following formulation for each stage

$$\tilde{\mathbf{x}}^{(t)} = \mathbf{x}^{(t)} + \mathcal{D}_\theta^{(t)}(\mathbf{x}^{(t)}) \tag{14}$$

$$\mathbf{k}^{(t)} = \mathcal{F}(\tilde{\mathbf{x}}^{(t)}) \tag{15}$$

$$\tilde{\mathbf{k}}^{(t)} = \mathcal{DC}_\eta(\mathbf{k}^{(t)}, \mathbf{k}^{(0)}) \tag{16}$$

$$\mathbf{k}^{(t+1)} = \mathcal{P}(\tilde{\mathbf{k}}^{(t)}, \mathbf{k}^{(0)}) \tag{17}$$

$$\mathbf{x}^{(t+1)} = \mathcal{F}^{-1}(\mathbf{k}^{(t+1)}) \tag{18}$$

where $\mathbf{x}^{(t)}, \tilde{\mathbf{x}}^{(t)}, \mathbf{k}^{(t)}, \tilde{\mathbf{k}}^{(t)} \in \mathbb{C}^{N\times h^{(t)}\times w^{(t)}}$, $h^{(t)}$ and $w^{(t)}$ define the progressive resolution schedule for the t-th stage. $\mathcal{P}$ is the k-space padding operator, filling the expanded region with acquired measurements where available, and zero elsewhere. Unlike UE and standard DUM, the initial estimate $\mathbf{x}^{(0)}$ is derived by eq. (2) after cropping $\mathbf{k}^{(0)}$ to the central region of size $h^{(0)} \times w^{(0)}$.

Almost all existing DUM approaches fall within a two-loop paradigm, an outer cascaded refinement loop with constant resolution, and an inner multi-resolution loop with a U-shaped backbone. Our progressive resolution design is, to the best of our knowledge, the first design to unify cascaded refinement and resolution scheduling into a single, up-sampling-only loop, which substantially improves efficiency by operating at lower resolutions for the bulk of the computation. Typical MRI sampling masks are denser in the center, leading to a higher sampling density of the initial low-resolution cascades with smaller k-space support, translating to a lower effective acceleration factor, making the early reconstruction considerably easier. Thus, our design inherently strengthens the progressive refinement nature of DUM.

3.4 Sparse Attention

Despite substantial acceleration from progressive resolution, the high-resolution cascades are still computationally heavy for high-resolution images, when employing a standard ViT as $\mathcal{D}_\theta$, whose full self-attention incurs quadratic complexity. Sparse attention is introduced by designing a sparsity pattern that determines which spatial locations attend to each other, computing attention only where it is meaningful and omitting it elsewhere to overcome the quadratic complexity.

We propose a novel sparsity pattern tailored for MRI, exploiting the fact that MRI undersampling, and thus aliasing, is anisotropic, a critical distinction from natural image tasks that assume isotropic spatial correlations. Acceleration in MRI is accomplished by undersampling along the phase-encode axis, while the readout direction is always fully sampled. From Fourier principles, multiplying k-space by a separable 1D mask corresponds to convolving the image with a 1D kernel, producing aliasing exclusively along that dimension. Following the convention of Zbontar et al. (2018), we consider a 1D sampling mask along the k_x axis, with the resulting aliasing along the x axis.

Since the primary goal of $\mathcal{D}_\theta$ is to remove x-axis aliasing, meaningful attention is expected to occur predominantly along the x-axis. As shown in fig. 3, attention maps from UE with a standard ViT confirm this assumption, with scores concentrated along the x-axis. Rather than learning the sparsity pattern from data with a standard ViT, we design it explicitly to maximize efficiency. Concretely, we define that each token attends to all other tokens within the same row and its n adjacent rows on each side, as shown in fig. 2. Additionally, we interleave sparse attention layers with full attention layers to preserve global context, following SOTA sparse attention designs in large language models (Team et al., 2024; Lu et al., 2025). By faithfully leveraging MRI physical principles, namely k-space acquisition and 1D undersampling, UEPS achieves simultaneous gains in efficiency and performance, circumventing the typical trade-off between the two.

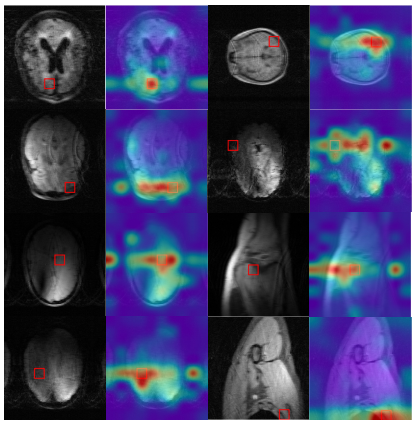

Figure 3: Attention maps from UE with a full-attention ViT backbone.

4 Experiments and Results

4.1 Experimental Setups

We construct a zero-shot transfer benchmark with ten OOD test sets from five public resources containing fully sampled multi-coil Cartesian k-space data (table 6). FastMRI Brain (Zbontar et al., 2018), the largest dataset among these resources, provides a data-rich training source for studying cross-dataset transfer. Its held-out split serves as the ID test set, while the remaining datasets evaluate OOD transfer across anatomy, view, sequence, contrast, vendor, field strength, and coil configuration without target-domain fine-tuning. All inputs are standardized to 320×320 through image-space cropping or padding, following Fabian et al. (2022); Wang et al. (2024).

The benchmark includes DUM (Sriram et al., 2020a; Wang et al., 2024; Jatyani et al., 2025), end-to-end networks (Ronneberger et al., 2015; Lin & Heckel, 2022), diffusion-based reconstruction (Chung et al., 2023), and untrained methods (Siedler et al., 2024; Joo et al., 2025). DUM and end-to-end baselines are retrained to match our settings; DDS uses released pretrained weights, and untrained methods are directly applied. Benchmark baselines retain their original architectural configurations, including cascade counts where applicable, while UEPS uses four cascades. Controlled ablations use matched configurations and training protocols, as detailed in section B.

Unless otherwise specified, we retrospectively undersample k-space using equispaced masks with $4\times$ acceleration and an 8% fully sampled center (Zbontar et al., 2018). We fix the acceleration factor to focus on cross-dataset transfer; substantial OOD degradation is already observed at $4\times$. We evaluate PSNR and SSIM against fully sampled RSS references, using the reference volume maximum for normalization. Runtime is measured on a single RTX 5090 GPU unless noted. DDS

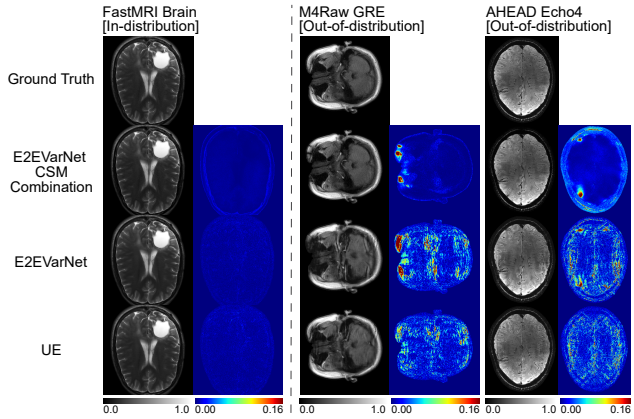

Figure 4: CSM combination and reconstruction on ID and OOD examples. CSM combination uses E2EVarNet-estimated CSMs to combine fully sampled coil images; E2EVarNet and UE reconstructions use undersampled inputs. All error maps are computed relative to the RSS reference.

Table 1: CSM combination comparison.

Dataset	Metric	BART	E2EVarNet	PDAC	NoVar
FastMRI Brain [ID]	PSNR↑	**44.95**	44.28	43.84	41.98
	SSIM↑	**99.00**	98.97	98.85	98.85
Average [OOD]	PSNR↑	**41.71**	38.16	36.82	37.09
	SSIM↑	**96.50**	94.44	92.24	94.31

Table 2: Reconstruction comparison.

Dataset	Metric	E2EVarNet	UE
FastMRI Brain [ID]	PSNR↑	**41.93**	41.90
	SSIM↑	**96.27**	96.16
Average [OOD]	PSNR↑	35.33	**37.54**
	SSIM↑	90.70	**92.82**

follows its original evaluation protocol, using Gaussian sampling masks and an MVUE reference (Chung et al., 2023).

4.2 Evaluation of CSM Estimation

As ground-truth CSM are unavailable, we assess CSM quality by comparing the CSM-combined multi-coil image (CSM combination) with the RSS combination, which an ideal CSM should recover. BART (Uecker et al., 2015) is a well-established untrained baseline for CSM estimation, providing a consistent reference for assessing the intrinsic difficulty of CSM estimation across datasets. As shown in table 1, E2EVarNet performs comparably to BART on ID data, but their PSNR gap widens from 0.67 dB to 3.55 dB on average across OOD test sets. PDAC and NoVar also fall substantially behind BART on OOD data, highlighting the limited cross-dataset generalization of learned CSM estimation. Figure 4 shows visual examples of how CSM errors relate to reconstruction errors. In the OOD examples, localized CSM estimation errors propagate through the reconstruction network, resulting in substantial reconstruction failure in the same regions. UE, without explicit CSM estimation, avoids CSM-induced regional artifacts. This observation is supported by the quantitative comparison in table 2: UE maintains comparable ID performance while improving average OOD PSNR by 2.21 dB. Together, these results identify dependence on learned CSM estimation as a bottleneck to OOD generalization.

4.3 Zero-shot Transfer Benchmark Results

Figure 1, together with the qualitative comparison in fig. 5, provides clear evidence that UEPS markedly improves robustness, consistently surpassing existing SOTA methods across all OOD test sets and achieving a large performance margin in aggregate, with complete results in table 8. The strong ID performance of other DUM methods and end-to-end methods fails to generalize to OOD settings, underscoring the need for zero-shot transfer evaluation for real-world deployment. Untrained methods trade off ID performance as expected; however, this sacrifice brings no consistent OOD advantage over supervised techniques. Remarkably, UEPS delivers robust performance with low latency, enabled by two efficiency-driven features that make real-time deployment feasible.

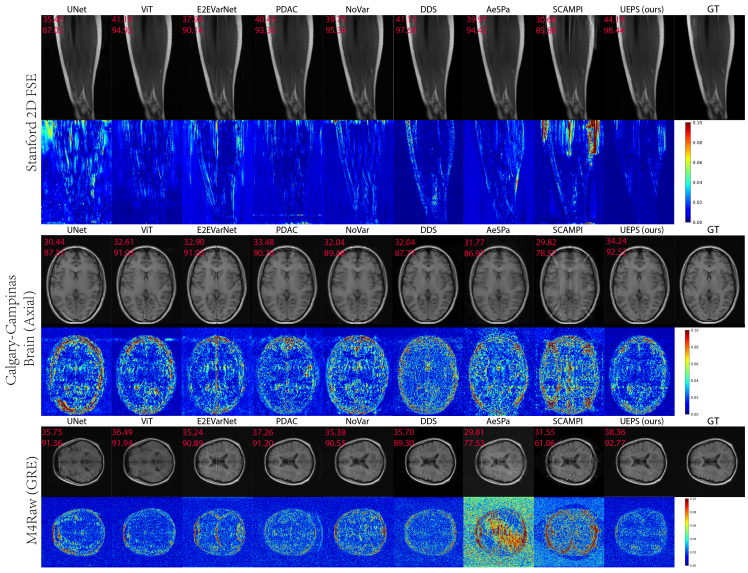

Figure 5: Reconstruction examples from the zero-shot transfer benchmark, each with PSNR/SSIM shown in the top-left corner.

4.4 Ablation Studies

We evaluate progressive resolution and sparse attention using the same ViT-based UE framework, adding each independently, and combining both as UEPS. All four configurations use the same training protocol and comparable parameter budgets. As shown in table 3, both components improve average reconstruction quality and reduce latency relative to UE. Incorporating progressive resolution with sparse attention substantially reduces inference latency with comparable reconstruction quality. Figure 6 illustrates their complementary benefits: progressive resolution reduces computation across cascades, while sparse attention limits attention cost as the token count increases.

4.5 Multi-Resolution Comparison

We evaluate our progressive resolution design against competitive multi-resolution architectures from both the general vision domain and prior MRI reconstruction studies. We select Restormer (Zamir et al., 2022), a cornerstone in image restoration with a U-shaped transformer, along with a recent improved variant (Zhou et al., 2025). Given the high memory footprint of these two models, we use a reduced configuration with approximately 30M parameters and a single UE cascade, the maximum feasible within our GPU memory. Two prior multi-resolution MRI methods are included: GrappaNet (Sriram et al., 2020b), a UNet-based coil-wise approach similar to our UE, and PDAC (Wang et al., 2024), a DUM approach with inner-cascade multi-resolution. GrappaNet and our model are likewise configured with 30M parameters for fair comparison, while PDAC uses its original configuration with roughly 130M parameters.

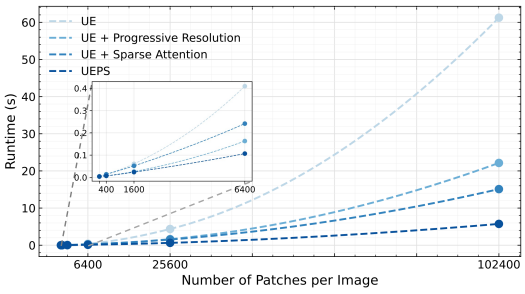

Figure 6: Runtime comparison.

Table 4 shows our model achieves top robustness, followed by Restormer and ASTv2, while PDAC lags due to its fragile CSM estimation. The strong representational capacity of Restormer and ASTv2 comes at the cost of substantial computational burden and GPU memory consumption. In contrast, UEPS merges cascaded refinement with progressive resolution scheduling in a single, up-

Table 3: Ablation of progressive resolution and sparse attention. The final row corresponds to UEPS. Best values are bold and second best underlined.

UE	Progressive resolution	Sparse attention	M4Raw GRE			FastMRI Knee			AHEAD Echo3			Average		
			PSNR	SSIM	Runtime	PSNR	SSIM	Runtime	PSNR	SSIM	Runtime	PSNR	SSIM	Runtime
✓	—	—	37.80	91.93	16.51	39.91	92.08	60.78	37.34	93.11	132.56	37.90	93.08	97.28
✓	✓	—	37.62	92.05	7.31	39.97	92.31	26.29	37.57	93.41	57.93	37.98	93.25	42.32
✓	—	✓	**37.81**	**92.11**	13.71	**40.07**	92.37	51.77	37.75	**93.66**	111.01	**38.19**	**93.47**	81.28
✓	✓	✓	37.47	91.94	**6.50**	**40.07**	**92.38**	**23.61**	**37.86**	**93.66**	**51.85**	38.13	93.39	**38.02**

Table 4: Multi-resolution comparison.

Dataset	Metric	Model				
		GrappaNet	Restormer	ASTv2	PDAC	UEPS(Ours)
M4Raw GRE	PSNR↑	36.40	36.50	36.93	36.63	**37.32**
	SSIM↑	91.01	91.30	91.59	90.31	**91.76**
	Runtime↓	4.96	107.34	87.21	98.99	**2.32**
FastMRI Knee	PSNR↑	38.88	39.83	**39.90**	38.30	39.87
	SSIM↑	91.54	92.38	**92.40**	90.38	92.08
	Runtime↓	15.68	194.67	185.55	113.22	**7.43**
AHEAD Echo3	PSNR↑	35.34	35.82	35.38	33.86	**37.24**
	SSIM↑	91.41	91.94	91.45	87.85	**93.01**
	Runtime↓	39.44	755.35	607.07	122.62	**16.14**
Average	PSNR↑	36.03	36.71	36.72	35.94	**37.57**
	SSIM↑	91.59	92.24	92.10	90.28	**92.76**
	Runtime↓	28.66	543.00	439.88	114.94	**11.85**

Table 5: Sparse-attention comparison.

Dataset	Metric	Model				
		Swin	CSWin	NAT	Sparge	UEPS(Ours)
M4Raw GRE	PSNR↑	37.18	37.47	**37.59**	36.58	37.47
	SSIM↑	91.61	**91.95**	91.92	91.09	91.94
	Runtime↓	6.91	6.58	6.77	8.60	**6.50**
FastMRI Knee	PSNR↑	39.94	39.90	39.98	39.71	**40.07**
	SSIM↑	92.29	92.20	92.35	92.10	**92.38**
	Runtime↓	24.60	**23.54**	24.14	29.79	23.61
AHEAD Echo3	PSNR↑	37.05	37.60	37.25	36.98	**37.86**
	SSIM↑	92.94	93.43	93.00	92.69	**93.66**
	Runtime↓	53.93	**51.15**	53.25	64.72	51.85
Average	PSNR↑	37.62	38.01	37.80	37.36	**38.13**
	SSIM↑	92.92	93.28	93.07	92.56	**93.39**
	Runtime↓	39.72	**37.41**	38.96	47.57	38.02

sampling-only loop, thus concentrating most computation at low resolutions for superior efficiency. Note that UEPS performance in table 4 is worse than that in table 3 due to reduced model size.

4.6 Sparse Attention Comparison

We compare our design with static sparse attention methods, Swin (Liu et al., 2021), CSWin (Dong et al., 2022), and NAT (Hassani et al., 2023), and the dynamic method Sparge (Zhang et al., 2025). Representative patterns appear in fig. 2(c). All methods use UE with progressive resolution, varying only the attention mechanism. Note that the runtime of Sparge is measured on a 4090 GPU and converted to a 5090-equivalent estimate because of hardware incompatibility. As shown in table 5, our MRI-tailored sparse attention outperforms general vision counterparts by explicitly leveraging MRI's 1D aliasing nature.

5 Limitations

Our study focuses on architectural robustness, leaving diverse training data, which improve robustness in MRI (Lin & Heckel, 2023; Lin et al., 2025) and vision (Radford et al., 2021; Kirillov et al., 2023), for future integration. Generalization to unseen sampling masks and acceleration factors remains untested. Prior work indicates this could be alleviated through training with diverse sampling patterns (Lin & Heckel, 2023). We also focus on 2D reconstruction. Reported benefits of 3D inputs (Xin et al., 2024) motivate extending our sparse attention to volumetric reconstruction.

6 Conclusion

We significantly improve MRI reconstruction robustness through three MRI-grounded design features: unrolled expanded model, progressive resolution, and sparse attention. These innovations yield simultaneous gains in performance and speed. A large-scale zero-shot transfer benchmark across 10 diverse OOD datasets confirms that UEPS consistently and substantially outperforms existing methods, establishing a new standard for robust real-world MRI reconstruction.

AI USE STATEMENT

Generative AI was used to assist with language polishing and improve the clarity of the manuscript. All revisions were reviewed and verified by the authors.

REFERENCES

Hemant K Aggarwal, Merry P Mani, and Mathews Jacob. Modl: Model-based deep learning architecture for inverse problems. *IEEE transactions on medical imaging*, 38(2):394–405, 2018.

Anneke Alkemade, Martijn J Mulder, Josephine M Groot, Bethany R Isaacs, Nikita van Berendonk, Nicky Lute, Scott JS Isherwood, Pierre-Louis Bazin, and Birte U Forstmann. The amsterdam ultra-high field adult lifespan database (ahead): A freely available multimodal 7 tesla submillimeter magnetic resonance imaging database. *NeuroImage*, 221:117200, 2020.

Joseph Cheng. Stanford 2d fse dataset. mridata.org, 2018. URL `http://mridata.org/list?project=Stanford%202D%20FSE`.

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

Mohammad Zalbagi Darestani and Reinhard Heckel. Accelerated mri with un-trained neural networks. *IEEE Transactions on Computational Imaging*, 7:724–733, 2021.

Mohammad Zalbagi Darestani, Akshay S Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *International Conference on Machine Learning*, pp. 2433–2444. PMLR, 2021.

Mohammad Zalbagi Darestani, Vishwesh Nath, Wenqi Li, Yufan He, Holger R Roth, Ziyue Xu, Daguang Xu, Reinhard Heckel, and Can Zhao. Ir-frestormer: Iterative refinement with fourier-based restormer for accelerated mri reconstruction. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 7655–7664, 2024.

Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *arXiv preprint arXiv:2412.05496*, 2(3):4, 2024.

Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12124–12134, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016.

Zalan Fabian, Berk Tinaz, and Mahdi Soltanolkotabi. Humus-net: Hybrid unrolled multi-scale network architecture for accelerated mri reconstruction. *Advances in Neural Information Processing Systems*, 35:25306–25319, 2022.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024.

Ruimin Feng, Qing Wu, Jie Feng, Huajun She, Chunlei Liu, Yuyao Zhang, and Hongjiang Wei. Imjense: scan-specific implicit representation for joint coil sensitivity and image estimation in parallel mri. *IEEE Transactions on Medical Imaging*, 43(4):1539–1553, 2023.

Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6185–6194, 2023.

Xuyi He, Yuhui Quan, Ruotao Xu, and Hui Ji. A universal scale-adaptive deformable transformer for image restoration across diverse artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12731–12741, 2025.

Armeet Singh Jatyani, Jiayun Wang, Aditi Chandrashekar, Zihui Wu, Miguel Liu-Schiaffini, Bahareh Tolooshams, and Anima Anandkumar. A unified model for compressed sensing mri across undersampling patterns. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26004–26013, 2025.

Patricia M Johnson, Geunu Jeong, Kerstin Hammernik, Jo Schlemper, Chen Qin, Jinming Duan, Daniel Rueckert, Jingu Lee, Nicola Pezzotti, Elwin De Weerdt, et al. Evaluation of the robustness of learned mr image reconstruction to systematic deviations between training and test data for the models from the fastmri challenge. In *International Workshop on Machine Learning for Medical Image Reconstruction*, pp. 25–34. Springer, 2021.

Jinho Joo, Hyeseong Kim, Hyeyeon Won, Deukhee Lee, Taejoon Eo, and Dosik Hwang. Aespa: Attention-guided self-supervised parallel imaging for mri reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5217–5226, 2025.

Yohan Jun, Hyungseob Shin, Taejoon Eo, and Dosik Hwang. Joint deep model-based mr image and coil sensitivity reconstruction network (joint-icnet) for fast mri. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5270–5279, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Florian Knoll, Kerstin Hammernik, Erich Kobler, Thomas Pock, Michael P Recht, and Daniel K Sodickson. Assessment of the generalization of learned image reconstruction and the potential for transfer learning. *Magnetic resonance in medicine*, 81(1):116–128, 2019.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1833–1844, 2021.

Kang Lin and Reinhard Heckel. Vision transformers enable fast and robust accelerated mri. In *International Conference on medical imaging with deep learning*, pp. 774–795. PMLR, 2022.

Kang Lin and Reinhard Heckel. Robustness of deep learning for accelerated mri: benefits of diverse training data. *arXiv preprint arXiv:2312.10271*, 2023.

Kang Lin, Anselm Krainovic, Kun Wang, and Reinhard Heckel. Improving deep learning for accelerated mri with data filtering. *arXiv preprint arXiv:2508.13822*, 2025.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Enzhe Lu, Zhejun Jiang, Jingyuan Liu, Yulun Du, Tao Jiang, Chao Hong, Shaowei Liu, Weiran He, Enming Yuan, Yuzhi Wang, et al. Moba: Mixture of block attention for long-context llms. *arXiv preprint arXiv:2502.13189*, 2025.

Mengye Lyu, Lifeng Mei, Shoujin Huang, Sixing Liu, Yi Li, Kexin Yang, Yilong Liu, Yu Dong, Linzheng Dong, and Ed X Wu. M4raw: A multi-contrast, multi-repetition, multi-channel mri k-space dataset for low-field mri research. *Scientific Data*, 10(1):264, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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*, pp. 8748–8763. PmLR, 2021.

Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jo Schlemper, Jose Caballero, Joseph V Hajnal, Anthony N Price, and Daniel Rueckert. A deep cascade of convolutional neural networks for dynamic mr image reconstruction. *IEEE transactions on Medical Imaging*, 37(2):491–503, 2017.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Thomas M Siedler, Peter M Jakob, and Volker Herold. Enhancing quality and speed in database-free neural network reconstructions of undersampled mri with scampi. *Magnetic Resonance in Medicine*, 92(3):1232–1247, 2024.

Roberto Souza, Oeslle Lucena, Julia Garrafa, David Gobbi, Marina Saluzzi, Simone Appenzeller, Letícia Rittner, Richard Frayne, and Roberto Lotufo. An open, multi-vendor, multi-field-strength brain mr dataset and analysis of publicly available skull stripping methods agreement. *NeuroImage*, 170:482–494, 2018.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *International conference on medical image computing and computer-assisted intervention*, pp. 64–73. Springer, 2020a.

Anuroop Sriram, Jure Zbontar, Tullie Murrell, C Lawrence Zitnick, Aaron Defazio, and Daniel K Sodickson. Grappanet: Combining parallel imaging with deep learning for multi-coil mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14315–14322, 2020b.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

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.

Martin Uecker, Frank Ong, Jonathan I Tamir, Dara Bahri, Patrick Virtue, Joseph Y Cheng, Tao Zhang, and Michael Lustig. Berkeley advanced reconstruction toolbox. In *Proc. Intl. Soc. Mag. Reson. Med*, volume 23, pp. 9, 2015.

Chong Wang, Lanqing Guo, Yufei Wang, Hao Cheng, Yi Yu, and Bihan Wen. Progressive divide-and-conquer via subsampling decomposition for accelerated mri. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25128–25137, 2024.

Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European conference on computer vision*, pp. 108–126. Springer, 2020.

Puyang Wang, Pengfei Guo, Keyi Chai, Jinyuan Zhou, Daguang Xu, and Shanshan Jiang. Sdum: A scalable deep unrolled model for universal mri reconstruction. *arXiv preprint arXiv:2512.17137*, 2025.

Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17683–17693, 2022.

Bingyu Xin, Meng Ye, Leon Axel, and Dimitris N Metaxas. Rethinking deep unrolled model for accelerated mri reconstruction. In *European Conference on Computer Vision*, pp. 164–181. Springer, 2024.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, and Mehmet Akçakaya. Zero-shot self-supervised learning for mri reconstruction. *arXiv preprint arXiv:2102.07737*, 2021.

George Yiasemis, Jan-Jakob Sonke, Clarisa Sánchez, and Jonas Teuwen. Recurrent variational network: a deep learning inverse problem solver applied to the task of accelerated mri reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 732–741, 2022.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

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.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattention: Accurate and training-free sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025.

Shihao Zhou, Jinshan Pan, and Jufeng Yang. Learning an adaptive sparse transformer for efficient image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

A Dataset Summary

Table 6: Summary of MRI datasets. 3D datasets are treated as multi-slice 2D data in selected anatomical views: CC-359 is converted to axial and sagittal views as two separate test sets, and AHEAD to the axial view. The AHEAD sequence contains five echoes, each forming a separate test set. Sources: FastMRI (Zbontar et al., 2018), Stanford 2D (Cheng, 2018), CC-359 (Souza et al., 2018), M4Raw (Lyu et al., 2023), and AHEAD (Alkemade et al., 2020).

Split	Dataset Name	Part	View	Sequence	Vendor	B0	Coils	Subj.	Slices
Train	FastMRI Brain	brain	axial	2D FSE	Siemens	3T/1.5T	2–28	4469	70748
Test(ID)	FastMRI Brain	brain	axial	2D FSE	Siemens	3T/1.5T	2–24	558	8852
Test (OOD)	FastMRI Knee	knee	coronal	2D FSE	Siemens	3T/1.5T	15	150	5427
	Stanford 2D	mix	mix	2D FSE	GE	3T	3–32	89	2037
	CC-359(Axi)	brain	axial	3D MP-RAGE	GE	3T	12	67	13802
	CC-359(Sag)	brain	sagittal	3D MP-RAGE	GE	3T	12	67	9490
	M4Raw(GRE)	brain	axial	2D GRE	Xingaoyi	0.3T	4	183	6588
	AHEAD(Echo1)	brain	axial	3D MP2RAGEME	Philips	7T	32	20	3240
	AHEAD(Echo2)	brain	axial	3D MP2RAGEME	Philips	7T	32	20	3240
	AHEAD(Echo3)	brain	axial	3D MP2RAGEME	Philips	7T	32	20	3240
	AHEAD(Echo4)	brain	axial	3D MP2RAGEME	Philips	7T	32	20	3240
	AHEAD(Echo5)	brain	axial	3D MP2RAGEME	Philips	7T	32	20	3240

B Additional Experimental Setups

B.1 UEPS Architecture

The unrolled architecture of UEPS consists of 4 cascades with a progressive resolution schedule of 64×64, 128×128, 256×256, and finally 320×320. The full-attention ViT baseline is built on DiT (Peebles & Xie, 2023) and incorporates 2D rotary position embeddings (RoPE) (Su et al., 2024; Fang et al., 2024) and SwiGLU feedforward networks (Shazeer, 2020). The ViT backbone within each cascade is configured identically with a patch size of 8, 10 transformer layers, an embedding width of 512, 8 attention heads, and an MLP intermediate size of 1280, yielding approximately 120M parameters for the complete UEPS model. The sparse-attention ViT is implemented using FlexAttention (Dong et al., 2024). Its 1D sparse attention is applied in cascades where the number of patches per image exceeds 256; otherwise, standard scaled dot-product attention is used (*e.g.*, in the initial 64×64 low-resolution cascade). Within each 10-layer transformer, we interleave sparse- and full-attention layers: the first and last layers use full attention, while the remaining layers use sparse attention with $n = 1$ adjacent row on each side.

B.2 Training

Our model was trained end-to-end on the FastMRI Brain (Zbontar et al., 2018) training split with Mean Absolute Error (MAE) loss and the Adam optimizer for 30 epochs with a batch size of 8. The learning rate started with a linear warmup to the base value of 0.0003 over the first 1% of total iterations, and later decayed via a cosine annealing schedule down to 10% of the base value. To accelerate training, we utilized Automatic Mixed Precision (AMP) alongside PyTorch's model compilation (`torch.compile`). To ensure fair evaluation, all models compared in our ablation and comparative experiments share the same training configuration, unless otherwise specified.

B.3 Ablation Studies

Using the ViT-based UE model as the baseline, we ablate progressive resolution and sparse attention by enabling each component independently and then jointly in UEPS. All four configurations use four unrolled cascades, approximately 120M parameters, and the same training protocol.

For the computational efficiency analysis in fig. 6, we measure the inference runtime of all four variants at a fixed input resolution of 320×320. Patch sizes vary over 32, 16, 8, 4, 2, and 1,

corresponding to 100, 400, 1600, 6400, 25600, and 102400 patches per full-resolution image, respectively.

B.4 Multi-Resolution Comparison

Training configuration was adjusted for all models in this study to accommodate the substantial memory and computational costs associated with Restormer (Zamir et al., 2022) and ASTv2 (Zhou et al., 2025). Specifically, all models were trained for 20 epochs with 8-coil training data, randomly sampled from the original multi-coil data (which contains a variable number of coils). Restormer (Zamir et al., 2022) and ASTv2 (Zhou et al., 2025) have approximately 26M and 10M parameters, respectively, matching the configurations in their original publications.

B.5 Sparse Attention Comparison

To isolate the impact of the attention pattern itself, all models share the same parameter capacity of 120M. Furthermore, the sparsity patterns were configured to ensure that the effective receptive field (attention scope) for each patch is roughly comparable across models.

C Additional CSM Evaluation Results

Table 7 provides the complete dataset-wise results corresponding to table 1. Estimated CSMs are used to combine fully sampled coil images, and the combined images are evaluated against the RSS reference. PSNR is computed over foreground pixels; SSIM is obtained by averaging the full-image SSIM map within the same foreground mask. Slices with empty foreground masks are excluded.

Table 7: Complete CSM evaluation results on the ID and ten OOD test sets. Values are mean ± standard deviation across valid slices. Average is computed over the ten OOD test sets. PSNR is reported in dB and SSIM in %. Best values are bold and second best underlined.

		CSM source			
Dataset	Metric	BART	E2EVarNet	PDAC	NoVar
FastMRI Brain [ID]	PSNR↑	**44.95**±6.52	<u>44.28</u>±4.59	43.84±4.67	41.98±4.33
	SSIM↑	**99.00**±2.03	<u>98.97</u>±2.30	98.85±2.58	98.85±2.48
CC-359 Axi	PSNR↑	**41.86**±3.10	<u>38.96</u>±3.30	37.22±3.72	38.83±2.94
	SSIM↑	**98.74**±0.96	98.34±1.02	97.76±1.62	<u>98.41</u>±0.92
CC-359 Sag	PSNR↑	**42.02**±3.24	<u>38.05</u>±3.02	37.63±3.22	36.89±2.52
	SSIM↑	**98.82**±0.72	<u>98.31</u>±0.79	98.02±1.14	98.21±0.75
M4Raw GRE	PSNR↑	**45.30**±1.61	<u>44.40</u>±1.68	44.38±1.77	44.03±1.70
	SSIM↑	**99.27**±0.25	<u>99.20</u>±0.28	99.18±0.32	99.19±0.28
AHEAD Echo1	PSNR↑	**40.34**±2.23	<u>37.09</u>±2.27	35.61±2.22	36.80±2.40
	SSIM↑	**95.11**±2.14	92.63±3.78	89.82±6.38	<u>92.89</u>±3.33
AHEAD Echo2	PSNR↑	**48.78**±4.03	<u>44.40</u>±3.83	43.91±3.85	41.81±3.16
	SSIM↑	**99.03**±1.04	<u>98.34</u>±2.09	97.75±3.10	98.06±2.14
AHEAD Echo3	PSNR↑	**42.90**±3.71	<u>36.19</u>±3.42	33.89±2.90	35.03±2.91
	SSIM↑	**96.85**±3.15	<u>93.33</u>±5.69	89.81±8.52	92.88±4.95
AHEAD Echo4	PSNR↑	**38.62**±4.04	<u>33.74</u>±3.15	32.07±3.17	32.97±2.81
	SSIM↑	**93.33**±5.43	<u>88.37</u>±7.89	83.14±11.68	87.83±6.44
AHEAD Echo5	PSNR↑	**35.86**±4.08	<u>32.35</u>±2.59	30.98±2.91	31.98±2.38
	SSIM↑	**89.32**±7.25	<u>84.09</u>±9.39	77.93±13.50	83.72±7.77
FastMRI Knee	PSNR↑	**38.76**±6.47	<u>36.73</u>±6.80	34.33±6.46	34.85±5.40
	SSIM↑	**96.55**±4.23	95.11±6.28	93.14±8.33	<u>95.25</u>±5.79
Stanford 2D	PSNR↑	**42.68**±6.25	<u>39.67</u>±6.60	38.22±6.17	37.71±5.28
	SSIM↑	**98.01**±2.60	<u>96.70</u>±4.72	95.81±6.41	96.61±4.39
Average [OOD]	PSNR↑	**41.71**	<u>38.16</u>	36.82	37.09
	SSIM↑	**96.50**	<u>94.44</u>	92.24	94.31

D Additional Zero-shot Transfer Benchmark Results

Complete results for each test set, including quantitative results (Table 8) and visual examples (Figure 7 and fig. 8).

Table 8: Complete results for the zero-shot transfer benchmark in fig. 1. Average is computed over the ten OOD test sets. Best values are in bold and second best underlined.

		Model								
Dataset	Metric	UNet	ViT	E2EVarNet	PDAC	NoVar	DDS	AeSPa	SCAMPI	UEPS$_{(Ours)}$
FastMRI Brain [ID]	PSNR(dB)↑	40.58±3.02	41.42±2.81	**41.93**±2.86	41.03±2.81	40.70±2.87	35.91±5.37	35.68±2.61	36.48±2.42	<u>41.87</u>±2.87
	SSIM(%)↑	95.61±3.49	95.95±3.49	**96.27**±3.34	95.69±3.63	95.68±3.62	85.52±21.49	89.05±5.16	76.48±10.54	<u>96.15</u>±3.49
CC-359 Axi	PSNR(dB)↑	29.09±3.45	30.99±3.52	31.27±3.46	<u>31.73</u>±3.48	30.65±3.50	29.17±5.42	30.11±2.47	29.83±2.52	**32.72**±3.25
	SSIM(%)↑	83.40±6.54	87.34±5.19	<u>87.86</u>±5.16	87.36±5.09	86.03±5.76	74.10±25.72	82.16±3.70	73.17±7.98	**90.35**±3.23
CC-359 Sag	PSNR(dB)↑	27.36±5.08	31.98±3.32	32.86±3.32	<u>32.99</u>±3.30	31.57±3.47	30.23±4.72	31.82±2.13	30.59±2.96	**34.45**±3.33
	SSIM(%)↑	77.01±13.56	88.81±4.32	<u>90.51</u>±4.05	89.69±4.09	87.65±5.24	79.90±20.66	83.43±3.14	69.62±5.48	**92.01**±3.03
M4Raw GRE	PSNR(dB)↑	35.37±2.34	36.37±2.08	35.62±2.01	<u>36.63</u>±1.93	35.60±2.24	26.73±9.07	25.50±4.48	32.64±1.50	**37.47**±1.77
	SSIM(%)↑	90.16±2.18	<u>91.01</u>±1.85	90.05±1.97	90.31±1.93	89.92±2.16	58.03±29.52	66.45±9.90	60.72±3.89	**91.94**±1.51
AHEAD Echo1	PSNR(dB)↑	33.95±2.21	35.67±2.24	<u>35.94</u>±2.26	35.80±2.23	34.59±2.20	24.50±3.56	33.42±1.80	35.05±1.53	**39.15**±2.08
	SSIM(%)↑	90.09±2.49	92.40±1.95	<u>93.18</u>±1.77	92.05±2.18	90.98±2.15	36.19±21.01	84.45±5.15	82.94±5.43	**95.60**±1.32
AHEAD Echo2	PSNR(dB)↑	36.10±2.08	37.53±2.03	38.14±2.00	<u>39.21</u>±2.10	37.17±1.93	31.08±5.31	36.45±2.22	36.38±1.68	**41.04**±2.04
	SSIM(%)↑	93.45±2.41	94.67±1.91	95.13±1.54	<u>95.40</u>±1.95	94.01±2.04	68.71±30.43	84.24±7.12	90.39±2.30	**96.75**±1.19
AHEAD Echo3	PSNR(dB)↑	32.14±2.83	34.03±2.72	<u>34.33</u>±2.54	33.86±2.68	32.93±2.61	26.98±4.25	33.77±2.86	32.82±2.04	**37.86**±2.54
	SSIM(%)↑	85.08±7.69	89.13±4.15	<u>89.63</u>±4.15	87.85±4.62	86.67±4.91	50.04±27.62	83.89±5.67	77.22±4.68	**93.66**±2.52
AHEAD Echo4	PSNR(dB)↑	31.29±3.74	33.51±3.51	<u>34.03</u>±3.28	33.48±3.37	32.53±3.35	26.44±3.84	33.07±2.37	32.86±1.57	**37.05**±3.13
	SSIM(%)↑	81.58±9.69	87.17±5.23	<u>88.23</u>±5.03	85.90±5.69	84.86±6.09	45.90±25.30	82.80±5.80	70.48±5.53	**92.15**±3.19
AHEAD Echo5	PSNR(dB)↑	31.48±4.36	33.86±3.93	34.21±3.55	33.62±3.75	32.59±3.68	25.45±3.30	32.48±3.17	<u>34.52</u>±2.23	**37.20**±3.47
	SSIM(%)↑	80.73±11.35	87.10±5.71	<u>88.11</u>±5.20	85.80±6.14	84.78±6.48	37.31±19.64	81.94±6.17	72.16±4.93	**91.85**±3.49
FastMRI Knee	PSNR(dB)↑	35.98±5.31	<u>39.89</u>±4.02	38.40±4.69	38.30±3.92	36.67±4.63	37.81±4.61	33.71±4.12	35.95±2.58	**40.07**±3.93
	SSIM(%)↑	86.74±9.46	**92.43**±5.87	91.33±6.21	90.38±6.59	89.25±6.99	89.19±12.75	79.99±11.71	76.91±14.43	<u>92.38</u>±6.08
Stanford 2D	PSNR(dB)↑	36.31±6.00	<u>38.84</u>±5.88	38.54±5.44	38.73±5.58	37.81±4.94	33.45±8.69	33.23±7.34	36.20±5.54	**40.52**±5.85
	SSIM(%)↑	90.34±8.70	<u>93.11</u>±6.64	92.96±6.98	92.69±7.09	91.87±6.30	78.04±26.40	76.55±14.50	75.95±13.17	**94.39**±6.45
Average [OOD]	PSNR(dB)↑	32.91	35.27	35.33	<u>35.43</u>	34.21	29.18	32.35	33.69	**37.75**
	SSIM(%)↑	85.86	90.31	<u>90.70</u>	89.74	88.60	61.74	80.59	74.95	**93.11**

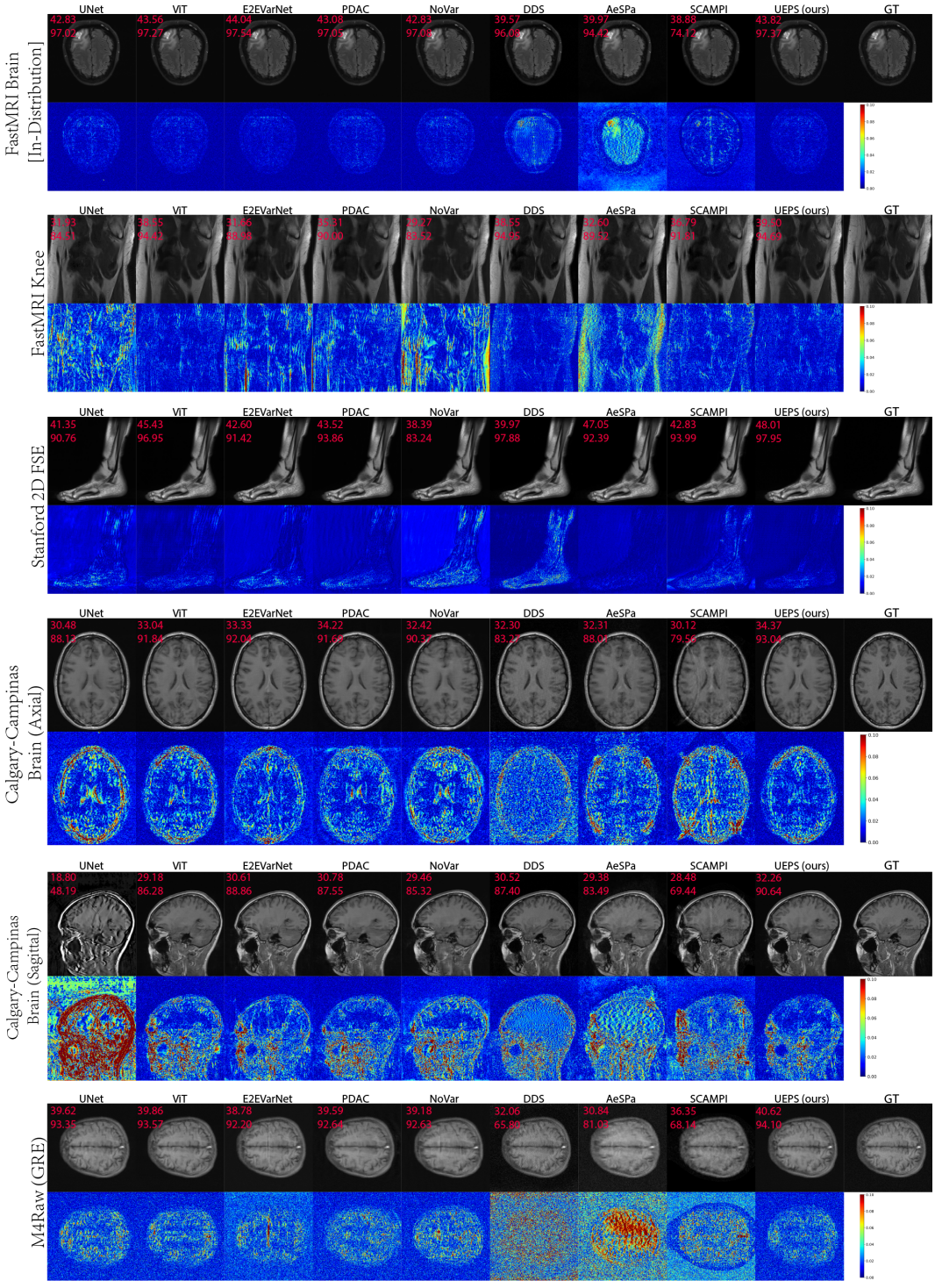

Figure 7: Additional reconstruction examples, each with PSNR/SSIM shown in the top-left corner (Part 1).

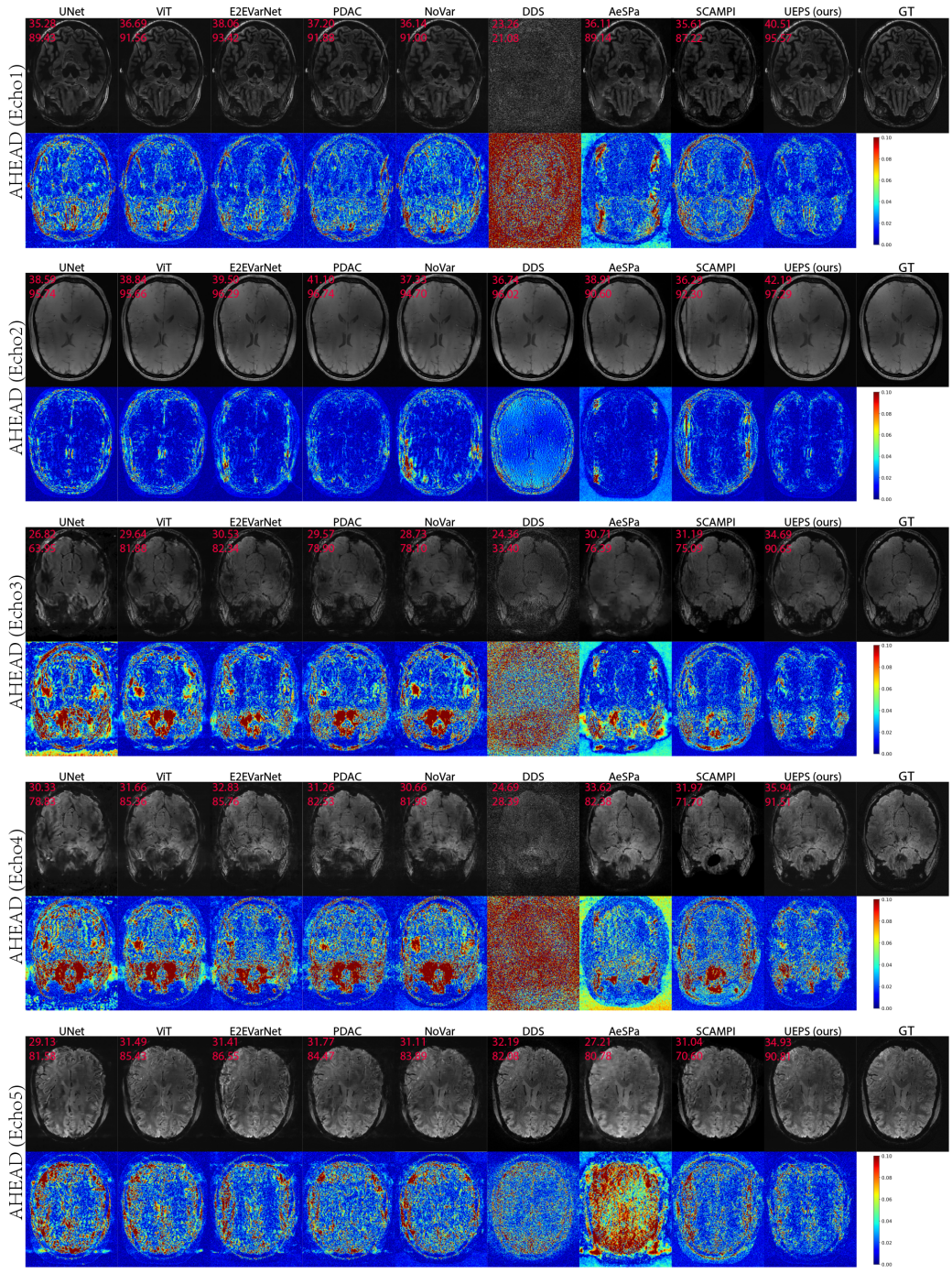

Figure 8: Additional reconstruction examples, each with PSNR/SSIM shown in the top-left corner (Part 2).