MARMI: MARGINALIZING ANATOMICAL UNCERTAINTY FOR MRI MOTION INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Patient motion in multi-shot magnetic resonance imaging (MRI) introduces k-space inconsistencies that depend jointly on the motion trajectory and the underlying anatomy, making the two difficult to disentangle from the acquired measurements. A motion-free auto-calibration signal (ACS) anchors the reference pose but only partially constrains the corresponding anatomy, leaving motion estimation intrinsically ambiguous. We introduce **MARMI (Marginalized Anatomical References for Motion Inference)**, a framework that models ACS-conditioned anatomical uncertainty and marginalizes it during amortized shot-wise motion inference on a compact central k-space representation. MARMI consists of two complementary components. Specifically, **Conditional Anatomical Residual Sampling (CARS)** generates plausible motion-free references by modeling the anatomical content not determined by the ACS, providing a direct physical interpretation of the unobserved frequencies. **Reference-Marginalized Motion Inference (RAMI)** then evaluates these references through a motion-aware MRI forward model and aggregates their physical evidence for motion estimation. On fastMRI FLAIR, marginalizing multiple ACS-conditioned references reduces normalized motion error by 10–13% relative to a single sampled reference. MARMI also outperforms JSMoCo and MoPED in motion-corrected reconstruction across the tested ACS and acceleration settings, and extends to 0.3 T M4Raw and an adapted in-house 0.55 T cohort. Together, these results show that modeling the partially observed reference as a latent distribution provides a principled strategy for disentangling motion from anatomy.

1 INTRODUCTION

Patient motion is a major source of image degradation in magnetic resonance imaging (MRI), particularly in acquisitions with extended scan times(Andre et al., 2015; Zaitsev et al., 2015). Because k-space is acquired sequentially, motion causes measurements collected at different poses to become inconsistent with a single stationary encoding model, producing characteristic reconstruction artifacts (Zaitsev et al., 2015; Spieker et al., 2024). In segmented multi-shot Cartesian sequences such as fast spin-echo (FSE), each echo train acquires a subset of phase-encoded k-space lines, and the complete k-space is assembled across successive shots (Hossbach et al., 2023; Polak et al., 2023). Under a shot-wise rigid-motion approximation, the pose is treated as approximately constant within each shot, while pose changes across shots define the motion trajectory. Retrospective correction therefore requires recovering these motion states from the corrupted measurements. This is intrinsically difficult because the k-space signature of a given motion trajectory depends on the underlying anatomy, coupling the inference of acquisition parameters with an unknown latent image.

Existing retrospective MRI motion-correction methods can be broadly distinguished by how they handle this anatomy-motion coupling. One class treats the motion-free image and motion trajectory as coupled unknowns, optimizing them jointly or alternately under a motion-aware forward model (Batchelor et al., 2005; Cordero-Grande et al., 2016). Recent variants of this approach leverage learned reconstruction models, test-time optimization, or subject-specific neural representations to regularize this highly ill-posed joint inverse problem (Haskell et al., 2019; Singh et al., 2024; Klug et al., 2024; Levac et al., 2024; Chen et al., 2025; Ortiz-Gonzalez et al., 2026; Wu et al., 2025). A second class reduces the anatomical degrees of freedom by introducing auxiliary motion-free information. Methods like SAMER, MoPED, and QUEEN rely on low-resolution scouts, additional

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.

motion-free references, or navigator measurements to anchor the motion estimation (Polak et al., 2022; Hossbach et al., 2023; Brackenier et al., 2024; Rizzuti et al., 2024). While these approaches successfully constrain the anatomy, they generally rely on external observed anatomical references that may not be available in routine clinical protocols. Generative approaches to inverse problems provide a complementary perspective by using learned image distributions to constrain solutions under incomplete or corrupted measurements (Chung et al., 2023; Aali et al., 2025; Park et al., 2026). In MRI motion correction, generative priors have likewise been incorporated into joint image-motion estimation, where they provide anatomical regularization while the motion trajectory is inferred from corrupted measurements (Levac et al., 2024; Chen et al., 2025; Ortiz-Gonzalez et al., 2026). Together with reference-assisted methods, these developments highlight the value of constraining the underlying anatomy for reliable motion estimation.

These observations motivate us to seek such anatomical constraints directly from the acquired data, without requiring an additional motion-free reference acquisition. In the acquisition considered here, we assume that the central auto-calibration signal (ACS) is acquired in a motion-consistent state and use its pose as the reference for relative shot-wise motion estimation. However, because the ACS covers only a subset of k-space, it constrains but does not uniquely determine the corresponding motion-free reference. Motion inference must therefore account for anatomical ambiguity that remains unresolved by the ACS. To account for this uncertainty, we introduce **MARMI (Marginalized Anatomical References for Motion Inference)**. Rather than committing to a single estimate of the partially observed reference anatomy, MARMI represents the reference image r as a latent variable governed by an ACS-conditioned distribution $p(r \mid a)$, where a denotes the observed ACS. MARMI then marginalizes over this anatomical uncertainty when inferring the shot-wise motion trajectory. To make this joint reasoning computationally tractable, MARMI operates on a compact central k-space inference window. This window retains essential low-frequency anatomical structure and motion-dependent shot measurements while delegating full-resolution image recovery to downstream reconstruction.

Specifically, MARMI realizes this probabilistic formulation through two synergistic components. **Conditional Anatomical Residual Sampling (CARS)** learns to model the anatomical content unresolved by the ACS, generating a diverse set of plausible motion-free references consistent with the same observation. We theoretically show that under any fixed linear complex-valued reconstruction map, this generated residual corresponds exactly to the reconstructed contribution of frequencies unobserved by the ACS, lending a rigorous physical interpretation to the latent variable. Subsequently, **Reference-Marginalized Motion Inference (RAMI)** evaluates these sampled references through a motion-aware MRI forward model. A finite-reference variational formulation yields a reference-index evidence lower bound (ELBO) that connects the learned reference weighting to physical evidence, enabling amortized shot-wise motion estimation. The resulting trajectory can seamlessly initialize existing physics-based correction or motion-aware reconstruction pipelines.

Our contributions are threefold:

- **Probabilistic Latent-Reference Formulation:** We reframe MRI motion estimation as a coupled inverse problem where the reference anatomy is only partially observed. By marginalizing over an ACS-conditioned reference distribution, we systematically propagate object uncertainty into the estimation of acquisition parameters.

- **Physically Grounded Generative Marginalization:** We develop CARS and RAMI to instantiate this framework. This includes a theoretical physical interpretation of the conditionally generated anatomical residual and a variational objective (ELBO) that links reference marginalization to motion-dependent evidence.

- **Demonstrated Gains from Uncertainty Propagation:** On the fastMRI FLAIR dataset, marginalizing multiple anatomical references reduces motion-estimation error by 10%–13% compared to single-reference inference. When used to initialize downstream pipelines, MARMI yields substantial improvements, including a 9.65 dB PSNR gain over zero-motion initialization in controlled settings. Experiments on low-field M4Raw further validate the framework's generalizability.

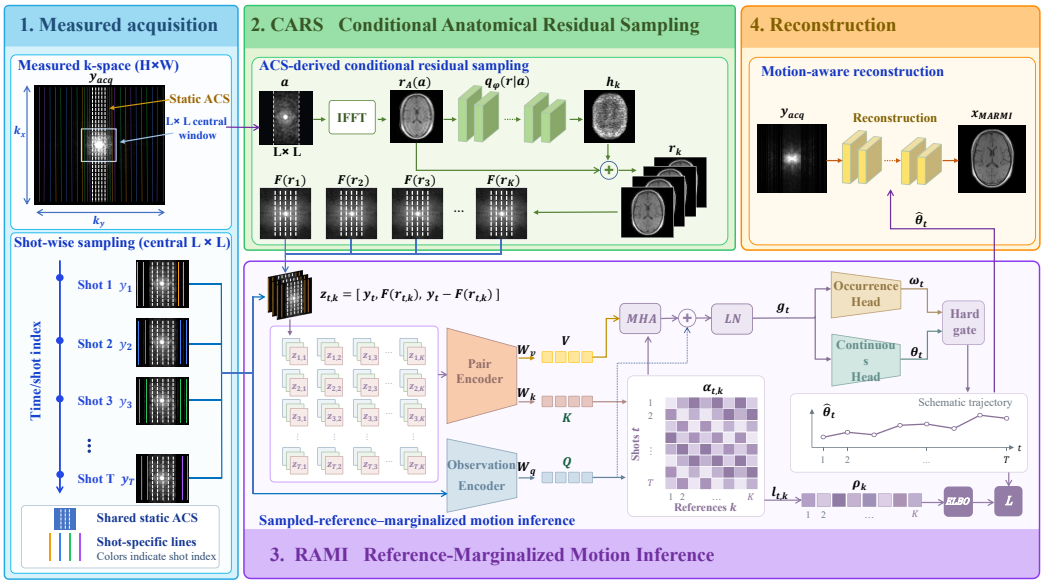

Figure 1: Given the measured multi-shot k-space and static ACS, CARS samples plausible motion-free anatomical references, while RAMI marginalizes their motion evidence to infer the shot-wise motion trajectory. The estimated motion is subsequently used for motion-aware reconstruction..

2 MARMI

MARMI infers shot-wise motion by marginalizing over plausible motion-free references consistent with the observed ACS. Figure 1 provides an overview. We first define the multi-shot acquisition model within the central k-space inference window (Section 2.1) and derive the corresponding reference-marginalized motion posterior (Section 2.2). CARS then learns an ACS-conditioned distribution over plausible reference images consistent with the observed ACS (Section 2.3), while RAMI performs amortized motion inference over a finite sampled reference set (Section 2.4). Finally, a reference-index evidence lower bound connects the learned reference weighting to physical evidence and yields the training objective (Section 2.5). Additional derivations and the scope of the forward model are provided in Appendix A.1.

2.1 Acquisition Model Construction

We consider multi-shot Cartesian MRI with shot-wise rigid motion(Batchelor et al., 2005; Cordero-Grande et al., 2016), where the object pose is approximately constant within each shot but may vary across shots. Rather than performing motion inference on the full reconstruction grid, MARMI operates on a compact central k-space window Ω_L with side length L. Let $r \in \mathbb{C}^{L\times L}$ denote the corresponding band-limited reference image. For shot $t \in \{1,\dots,T\}$ and receiver coil $c \in \{1,\dots,C\}$, the measurements within this window are modeled as

$$y_{t,c} = \mathcal{A}^{\mathrm{BL}}_{t,c}(r,\theta_t) + \epsilon_{t,c}, \tag{1}$$

Here $\mathcal{A}^{\mathrm{BL}}_{t,c}$ is the band-limited reference-based forward operator defined in Eq.18: it synthesizes a coil view from r, embeds it on the original grid, applies motion to the coil image in original-grid coordinates, and retains the acquired central samples. We write $\theta_t = (\varphi_t, t_{x,t}, t_{y,t})$, with rotation in degrees and translations in original-grid pixels, and $\Theta = \{\theta_t\}_{t=1}^{T}$. Equation equation 1 specifies the compact-reference working model; it is not an algebraic rewrite of the full-grid simulator.

We assume that the central auto-calibration signal (ACS) is acquired in a motion-consistent reference state and define this pose as $\theta_A = 0$. Let $\Omega_A \subset \Omega_L$ denote the ACS support. For coil c, the ACS measurement is

$$a_c = M_A \mathcal{F}_L\,[S_c \odot r] + \epsilon_{a,c}, \tag{2}$$

where M_A selects the ACS frequencies. We collect the dynamic-shot and ACS observations as

$$y = \{y_t\}_{t=1}^{T}, \qquad y_t = \{y_{t,c}\}_{c=1}^{C}, \qquad a = \{a_c\}_{c=1}^{C}.$$

The ACS therefore defines the reference pose for Θ, but does not fully specify the associated reference anatomy. Since Ω_A occupies only a subset of Ω_L, multiple band-limited references may be consistent with the same ACS observation. The next subsection formulates motion inference under this unresolved reference uncertainty.

2.2 Reference-Marginalized Motion Posterior

To account for this ambiguity, we model the unresolved reference anatomy with an ACS-conditioned distribution

$$p(r \mid a). \tag{3}$$

Our goal is to infer the posterior over the shot-wise motion trajectory, $p(\Theta \mid y, a)$. By Bayes' rule,

$$p(\Theta \mid y, a) \propto p(y \mid \Theta, a)\, p(\Theta \mid a), \tag{4}$$

where evaluating the likelihood $p(y \mid \Theta, a)$ requires accounting for the unknown reference r.

Under the acquisition model in Section 2.1, we assume fixed-covariance circular complex Gaussian measurement noise, independent ACS and shot errors, conditional independence of the dynamic shots given r and Θ, and conditional independence between the reference and motion trajectory before observing the dynamic measurements,

$$p(r, \Theta \mid a) = p(r \mid a)p(\Theta \mid a).$$

Marginalizing the latent reference then gives

$$p(\Theta \mid y, a) \propto p(\Theta \mid a) \int p(r \mid a) \prod_{t=1}^{T} p(y_t \mid r, \theta_t)\, dr. \tag{5}$$

The same latent reference r is shared across all shots, so their likelihoods are combined before marginalizing over the anatomical uncertainty. Equation 5 therefore integrates motion evidence over the ACS-conditioned reference distribution. The next section introduces CARS to obtain a tractable, sampleable approximation to $p(r \mid a)$.

2.3 CARS: ACS-Conditioned Anatomical References

CARS represents the part of the reference that the ACS leaves unresolved. For a noiseless central k-space realization $k^\star = k_A^\star + k_U^\star$, let $k_A^\star$ be its ACS-supported component and $k_U^\star$ the complement within the inference window. If $\mathcal{C}$ is a fixed complex-linear map from this k-space representation to the reference image, $r^\star = \mathcal{C}(k^\star)$ and $r_A^\star = \mathcal{C}(k_A^\star)$ imply

$$h^\star := r^\star - r_A^\star = \mathcal{C}(k_U^\star). \tag{6}$$

This identity concerns a noiseless realization. For measured ACS data a, the baseline is instead $r_A(a)$; CARS models $h = r - r_A(a)$ with a conditional diffusion distribution $q_\psi(h \mid a)$ (Ho et al., 2020), thereby inducing $q_\psi(r \mid a)$ through $r = r_A(a) + h$. Learned by conditional residual denoising, it approximates the unknown $p(r \mid a)$ without identifying the noisy baseline with $r_A^\star$. Replacing that posterior in Eq.5 defines

$$\widetilde{p}_\psi(y \mid \Theta, a) = \int q_\psi(r \mid a) \prod_{t=1}^{T} p(y_t \mid r, \theta_t)\, dr. \tag{7}$$

2.4 RAMI: Motion Inference with a Shared Reference Set

RAMI draws $\mathcal{R}_K = \{r^{(k)}\}_{k=1}^{K}$ from $q_\psi(r \mid a)$. Let $\pi_k > 0$ be normalized reference weights, with $\pi_k = 1/K$ for equally weighted samples. A single index Z selects the reference shared by the entire case, giving

$$p_K(y \mid \Theta, \mathcal{R}_K) = \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} p(y_t \mid r^{(k)}, \theta_t). \tag{8}$$

The ordering of the product and sum preserves the common anatomy. At fixed Θ, Bayes' rule on this realized set yields its likelihood-derived responsibility

$$\gamma_k(\Theta) := p(Z = k \mid y, \Theta, \mathcal{R}_K) = \frac{\pi_k \prod_t p(y_t \mid r^{(k)}, \theta_t)}{\sum_j \pi_j \prod_t p(y_t \mid r^{(j)}, \theta_t)}. \tag{9}$$

This is exact within the finite mixture for a proposed trajectory. Because the trajectory is unknown, RAMI predicts from the observations and the reference set. For shot t, let o_t be its measured k-space view with static ACS context, and $b_{t,k}$ the corresponding k-space view synthesized from $r^{(k)}$ on the same support. A shared pair encoder uses the observation, reference, and their difference to form $u_{t,k} = E_{\text{pair}}(o_t, b_{t,k})$; an observation encoder forms the query $q_t = E_{\text{obs}}(o_t)$. With $U_t = \{u_{t,k}\}_{k=1}^K$ as keys and values, set attention (Lee et al., 2019)produces

$$\begin{aligned} \ell_{t,k} &= \frac{1}{J} \sum_{j=1}^{J} \frac{(W_Q^{(j)} q_t)^\top (W_K^{(j)} u_{t,k})}{\sqrt{d_j}}, \\ g_t &= \text{MHA}(q_t, U_t, U_t). \end{aligned} \tag{10}$$

Here J is the number of heads, d_j their projected key dimension, and $W_Q^{(j)}, W_K^{(j)}$ learned projections. Each head uses its own softmax over references to aggregate values into the shot representation g_t. The head-averaged, unnormalized scores $\ell_{t,k}$ instead define an amortized, case-level reference distribution:

$$\rho_{\omega,k} = \text{softmax}_k \left(\log \pi_k + \sum_{t=1}^{T} \ell_{t,k} \right). \tag{11}$$

The learned scores are not constrained to equal physical log-likelihoods. The resulting ρ_ω is an amortized variational weight used by the reference-index objective at the predicted trajectory; it need not equal $\gamma(\Theta)$ for an arbitrary proposed trajectory. Within-shot attention weights are not reference posterior responsibilities.

The same g_t feeds separate occurrence and continuous-motion heads. The former gives $p_t = \sigma(f_{\text{occ}}(g_t))$, the latter gives a mean μ_t and diagonal covariance Σ_t in the physical coordinates of θ_t. The resulting shot-wise marginal is

$$q_\phi(\theta_t \mid y, a, \mathcal{R}_K) = (1 - p_t)\delta_0 + p_t \mathcal{N}(\mu_t, \Sigma_t), \qquad p_t = q_\phi(z_t = 1 \mid y, a, \mathcal{R}_K). \tag{12}$$

Here $z_t \in \{0, 1\}$ indicates whether shot t moves and δ_0 is a point mass at zero motion. The parameters ω and ϕ distinguish outputs of an inference network with shared encoders and attention. These shot-wise marginals do not specify every dependency in a joint posterior over Θ.

2.5 Reference-Index ELBO and Training Objective

The finite mixture connects the amortized weights to the physical evidence. For fixed Θ and $\mathcal{R}_K$, Jensen's inequality gives the reference-index bound

$$\mathcal{E}_{\text{ref}}(\Theta, \rho_\omega) := \sum_{k=1}^{K} \rho_{\omega,k} \log p(y \mid \Theta, r^{(k)}) - D_{\text{KL}}(\rho_\omega \| \pi) \le \log p_K(y \mid \Theta, \mathcal{R}_K), \tag{13}$$

Expanding the divergence to the physical responsibilities gives the exact gap

$$\log p_K(y \mid \Theta, \mathcal{R}_K) - \mathcal{E}_{\text{ref}}(\Theta, \rho_\omega) = D_{\text{KL}}(\rho_\omega \| \gamma(\Theta)). \tag{14}$$

Thus the bound is tight at $\rho_\omega = \gamma(\Theta)$ when the variational family can represent those responsibilities. This claim concerns a fixed motion hypothesis and realized reference set, rather than an ELBO over the continuous reference law or motion uncertainty.

For $\log p(y \mid \Theta, r^{(k)}) = -E_k(\Theta)/\tau + C_0$, with $\tau > 0$ and a constant C_0 shared across references and trajectories, the negative bound in energy units is

$$\mathcal{L}_{\text{ELBO}}(\Theta, \rho_\omega) = \sum_{k=1}^{K} \rho_{\omega,k} E_k(\Theta) + \tau D_{\text{KL}}(\rho_\omega \| \pi). \tag{15}$$

The first term weights reference-dependent explanations of the dynamic shots under Θ; the second penalizes departure from the reference prior. For simulated trajectories with movement labels $z_t^\star$ and motion labels $\theta_t^\star$, the implemented RAMI training objective is

$$\mathcal{L}_{\mathrm{RAMI}} = \lambda_{\mathrm{occ}}\mathcal{L}_{\mathrm{BCE}}(p, z^\star) + \lambda_{\mathrm{mot}}\mathcal{L}_{\mathrm{NLL}}(\mu, \Sigma; \Theta^\star \mid z^\star = 1) + \lambda_{\mathrm{ref}}\mathcal{L}_{\mathrm{ELBO}}(\mu, \rho_\omega). \tag{16}$$

The binary cross-entropy uses all shots, whereas the diagonal-Gaussian negative log-likelihood is averaged over moving shots only. The reference-index term is evaluated at the predicted motion means $\mu = (\mu_t)_t$ before occurrence thresholding.

3 Experiments

3.1 Experimental Setup

Datasets We evaluate MARMI on two public multi-coil brain MRI datasets and one in-house low-field cohort. We use 200 subjects with axial FLAIR acquisitions from fastMRI (Zbontar et al., 2018) and 158 subjects with T1-weighted acquisitions from the 0.3 T M4Raw dataset (Lyu et al., 2023). Each public dataset is split by subject into training, validation, and test sets at an $8{:}1{:}1$ ratio, with identical splits used for CARS and RAMI. We further acquire T1-weighted data from 22 subjects on a 0.55 T Siemens MAGNETOM Free.Star system, divided into 18/2/2 subjects for training/validation/test. For this target-domain experiment, both CARS and RAMI are initialized from their M4Raw-trained weights and fine-tuned on the 0.55 T training set before evaluation on the held-out test subjects. All splits are subject-disjoint.

Motion simulation and evaluation For both RAMI training and motion evaluation, we synthesize eight-shot Cartesian acquisitions with the static ACS defining the zero-motion reference pose. Unless otherwise specified, MARMI uses a central 64×64 k-space window with $K = 8$ ACS-conditioned references. We consider 24- and 32-line ACS configurations, corresponding to five and four dynamic lines per shot, respectively.

The simulation protocol contains 13 motion conditions: one no-motion condition and 12 combinations of 2, 4, 6, or 8 moving shots with maximum rotation/translation magnitudes of $1^\circ/1$ pixel, $3^\circ/3$ pixels, or $5^\circ/5$ pixels. RAMI is trained on these conditions, and the same protocol is applied to held-out subjects with independently simulated trajectories for evaluation. We additionally include four $8^\circ/8$-pixel stress-test conditions for evaluation only. Motion estimation is quantified using normalized motion error (NME), with rotation and translation errors normalized by 8° and 8 pixels, respectively. Ground-truth trajectories are used for simulation, training supervision, and evaluation, but are not available to MARMI at inference.

Reconstruction protocol For reconstruction evaluation, four of the eight shots are assigned motion under a $3^\circ/3$-pixel rotation/translation bound. We evaluate both $2\times$ and $3\times$ accelerated acquisitions. The trajectory estimated by MARMI initializes three rounds of alternating motion and image refinement, with 20 motion updates and 50 image updates per round (learning rates 0.01 and 0.03, respectively). After refinement, the motion trajectory is fixed and the image is reconstructed from the measured k-space using 20 conjugate-gradient iterations with ridge 10^{-5}.

3.2 Network and Training Details

Reference generation. We train separate CARS models for fastMRI and M4Raw under the 24- and 32-line ACS configurations. Each CARS model is an ϵ-predicting residual DDPM with a U-Net using 64 base channels, channel multipliers $(1, 2, 4)$, and one residual block per level. We use a 1,000-step linear diffusion schedule. In addition to the denoising objective, training includes residual, missing-frequency multi-coil, and measured-ACS consistency losses with weights 0.15, 0.10, and 0.05, respectively. CARS is trained with batch size 16 and an initial learning rate of 10^{-4}, and reference samples are generated using 1,000 reverse diffusion steps.

Motion inference. To accommodate datasets with different coil receiver-channel counts, RAMI selects four coils according to their measured ACS energy. For each shot-reference pair, the acquired shot view, reference-predicted view, and their difference are represented by their real and imaginary

components, yielding six input channels. The resulting pair features are aggregated by the reference-attention module and passed to the occurrence and conditional-motion heads defined in Section 2. At inference, shots with $p_t < 0.5$ are assigned zero motion.

RAMI is trained with Eq. 16, using unit weights for the occurrence and moving-shot motion losses, $\lambda_{\text{ref}} = 0.1$, $\tau = 0.1$, and a uniform reference prior. Training runs for 200 epochs with batch size 4, two-step gradient accumulation, learning rate 10^{-4}, and weight decay 10^{-4}. The reference-evidence discrepancy E_k is computed as the mean squared error between the reference-predicted and acquired dynamic-shot k-space samples within the inference window. Detailed encoder and prediction-head architectures are provided in Appendix A.2.

3.3 Reconstruction Results

fastMRI We compare MARMI with JSMoCo (Chen et al., 2025), a generative joint image–motion method, and MoPED (Hossbach et al., 2023), which uses an additional low-resolution reference, on fastMRI FLAIR under ACS24/32 and $2\times$ /$3\times$ acceleration (Table 1). Across all four settings, MARMI+CG20 achieves higher PSNR and SSIM than both baselines, while MARMI+MoDL gives the best overall performance. Figure 2 provides qualitative comparisons through matched reconstructions, residual maps, and ROI enlargements.

To separate the contribution of motion estimation from that of the reconstruction backend, MARMI+CG20 and MARMI+MoDL use identical MARMI-estimated trajectories and the same motion-refinement procedure. Replacing only the final CG20 reconstruction with MoDL increases PSNR by 0.39-0.49 dB at $2\times$ and 2.09-2.11 dB at $3\times$, with consistent SSIM improvements. Importantly, MARMI+CG20 already outperforms both reconstruction baselines in every setting, indicating that the gains are not solely due to the learned MoDL backend. The accuracy of the MARMI-estimated trajectories is examined directly using NME. Additional quantitative results and protocol details are provided in Appendix A.4.

Table 1: Reconstruction comparison on the matched fastMRI FLAIR test dataset with 50% moved shots (at most 3°/3 pixels). Each cell gives PSNR (dB) / SSIM; dashes await matched test results. MoPED-GT uses a reference derived from each test target and iterative residual refinement followed by CG20.

ACS	R	MoPED-GT	JSMoCo	MARMI+CG20	MARMI+MoDL
32	2	37.31 / 0.9693	37.61 / 0.9150	38.03 / 0.9707	**38.52 / 0.9754**
	3	31.96 / 0.9388	32.57 / 0.8971	32.94 / 0.9426	**35.03 / 0.9567**
24	2	37.33 / 0.9701	37.41 / 0.9129	39.21 / 0.9788	**39.60 / 0.9826**
	3	32.21 / 0.9400	32.66 / 0.8946	33.47 / 0.9492	**35.58 / 0.9619**

Low-field M4Raw and in-house 0.55 T datasets We further evaluate MARMI on low-field T1-weighted MRI at $2\times$ acceleration (Table 2). On the 0.3 T M4Raw dataset, MARMI achieves 36.77 dB/0.9562 PSNR/SSIM with ACS24 and 35.65 dB/0.9425 with ACS32, with corresponding NME values of 0.0154 and 0.0238. These results show that MARMI maintains both accurate motion estimation and high reconstruction quality in a low-field acquisition setting distinct from the fastMRI FLAIR experiments.

We next evaluate transfer to the in-house 0.55 T cohort. Starting from the M4Raw-trained weights, both CARS and RAMI are fine-tuned on the target-domain training set before evaluation on held-out subjects. MARMI achieves 39.81 dB/0.9713 PSNR/SSIM with ACS32 and 36.75 dB/0.9481 with ACS24, with NME values of 0.0185 and 0.0319, respectively. These results support the applicability of MARMI after adaptation to a different field strength and acquisition system. Interestingly, ACS24 performs better on M4Raw, whereas ACS32 performs better on the 0.55 T cohort, suggesting that the preferred ACS configuration may depend on the acquisition setting.

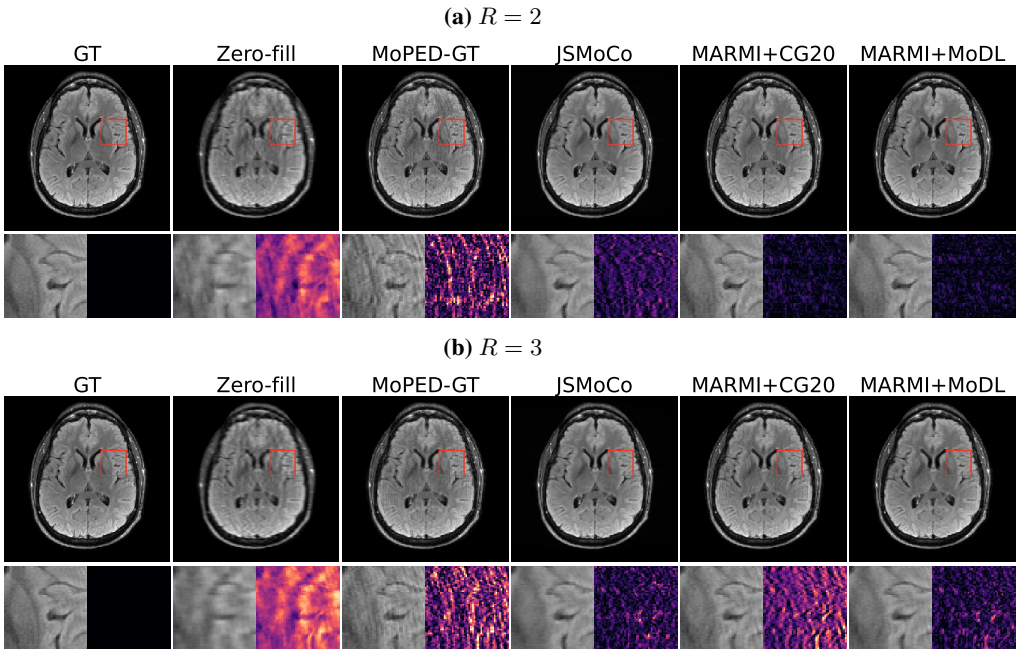

Figure 2: FastMRI FLAIR reconstruction comparison with ACS24 at $R = 2$ (a) and $R = 3$ (b). Columns show the ground truth, zero-filled reconstruction, MoPED-GT, JSMoCo, MARMI+CG20, and MARMI+MoDL. The lower panels show $3\times$ magnified regions of interest and corresponding absolute-error maps. All reconstructions use the same grayscale window within each acceleration setting.

Table 2: Reconstruction results across datasets and field strengths at $2\times$ acceleration with 50% moving shots and motion bounded by $3^\circ/3$ pixels on each dataset's native grid. Entries report slice-averaged PSNR, SSIM, and post-refinement normalized motion error (NME).

Test cohort	ACS	PSNR (dB)↑	SSIM↑	NME↓
M4Raw T1	24	36.77	0.9562	0.0154
	32	35.65	0.9425	0.0238
In-house T1	24	36.75	0.9481	0.0319
	32	39.81	0.9713	0.0185

3.4 Ablation Studies

We examine two design choices in MARMI motion inference: the number of ACS-conditioned reference samples K and the ACS sampling configuration. We first study the effect of K on fastMRI, and then compare ACS24 and ACS32 on both fastMRI and M4Raw using $K = 8$.

Number of references On fastMRI with ACS32, we vary $K \in \{1, 2, 4, 8, 16\}$ while keeping the reference prefixes, motion conditions, and training schedule fixed. Two independently trained RAMI models are evaluated for each K. For this ablation, $\lambda_{\text{ref}} = 0$, so the experiment isolates sensitivity to the number of sampled references without reference-evidence supervision. The evaluation includes the 12 moving conditions of the base motion grid and four additional stress-test conditions, for which the maximum motion magnitude is fixed at $8^\circ/8$ pixels and the number of moving shots is set to 2, 4, 6, or 8. Across these conditions, increasing K from 1 to 8 reduces mean NME from 0.0701 to 0.0610 on the base motion grid, corresponding to a 13.0% reduction, and from 0.0918 to 0.0826 on the $8^\circ/8$-pixel stress-test conditions, corresponding to a 10.0% reduction (Figure 3). Increasing the reference set further to $K = 16$ yields NME values of 0.0611 and 0.0826, respectively, showing little additional improvement. We therefore use $K = 8$ in the remaining experiments.

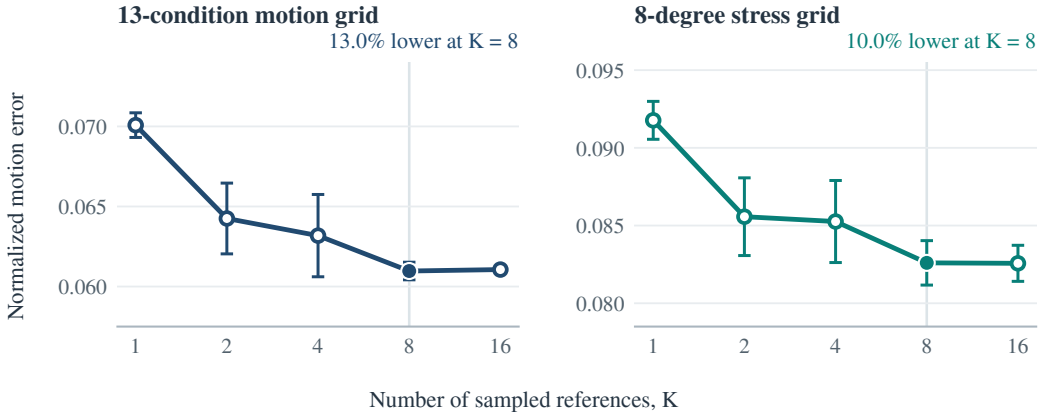

Figure 3: Reference-count ablation on fastMRI FLAIR with ACS32. NME is macro-averaged over the 13 moving conditions of the base motion grid (left) and four additional stress-test conditions with a fixed maximum motion of $8^\circ/8$ pixels and 2, 4, 6, or 8 moving shots (right). Error bars denote the sample SD across two independently trained models. Filled points indicate $K = 8$. Lower is better.

Sensitivity to ACS configuration With $K = 8$, we compare ACS24 and ACS32 under matched motion protocols on fastMRI and M4Raw (Table 3). On both datasets, ACS24 yields lower NME than ACS32 on the base motion grid, with NME decreasing from 0.0610 to 0.0441 on fastMRI and from 0.0806 to 0.0597 on M4Raw. The same trend is observed on the stress grid, particularly on M4Raw, where NME decreases from 0.1204 to 0.0778.

These results indicate that motion-estimation performance is sensitive to the ACS configuration under the evaluated protocols. However, they do not imply a universally preferred ACS setting, as the relative performance of ACS24 and ACS32 can vary across acquisition settings. In addition, ACS24 provides five dynamic lines per shot compared with four for ACS32, and the two configurations use separately trained CARS and RAMI models. The comparison therefore reflects the complete ACS24 and ACS32 configurations rather than the isolated effect of ACS width.

Table 3: Sensitivity of motion estimation to the ACS configuration. NME is macro-averaged over the 13 moving conditions of the base motion grid and the four additional $8^\circ/8$-pixel stress-test conditions. Lower is better.

Dataset	ACS lines	13-condition NME ↓	Stress-grid NME ↓
fastMRI FLAIR	24	0.0441	0.0805
	32	0.0610	0.0826
M4Raw T1	24	0.0597	0.0778
	32	0.0806	0.1204

4 CONCLUSION

In this paper, we introduced MARMI, a framework for MRI motion inference that explicitly accounts for uncertainty in the reference anatomy given only the observed ACS. Rather than collapsing this partial anatomical information into a single deterministic reference, CARS generates multiple plausible ACS-conditioned references, while RAMI marginalizes their motion evidence to estimate the shot-wise motion trajectory. Experiments on fastMRI, M4Raw, and in-house 0.55 T data show that the resulting motion estimates support effective motion-corrected reconstruction across different acquisition settings. The reference-count study further shows that incorporating multiple plausible references improves motion estimation over relying on a single sampled reference. More broadly, MARMI reframes reference-based motion inference as inference under anatomical uncertainty, providing a principled way to combine generative anatomical priors with acquisition physics when the reference anatomy is only partially observed.

DECLARATIONS

AI USE STATEMENT

In this work, we used generative AI tools for research ideation and execution, including code modification. We did not use them for other required-disclosure tasks; the remaining tasks are not applicable. Additionally, we used them to find related work and to draft and polish parts of the manuscript. The authors reviewed all AI-assisted work, checked the code, citations, claims, and results, and take responsibility for the final content, including AI-assisted text and artifacts.

ETHICS STATEMENT

This study uses the publicly available fastMRI and M4Raw datasets and an in-house 0.55 T T1-weighted MRI cohort. The in-house data were used with participant consent and were de-identified before analysis. No patient identifiers are included in this paper. The simulated-motion experiments evaluate an image reconstruction method and do not establish its suitability for clinical diagnosis.

REPRODUCIBILITY STATEMENT

Section 2 specifies the acquisition model, the CARS and RAMI methods, and the training objective. Section 3 describes the datasets, subject-disjoint evaluation, training settings, and reconstruction comparisons. Appendix A.1 provides the theoretical derivation; Appendices A.2, A.3, and A.4 document the network implementation, evaluation metrics, and baseline protocols. Upon acceptance, we will release all code developed for this study as open source, including scripts for data preparation, motion simulation, training, inference, reconstruction, and evaluation. The fastMRI and M4Raw datasets can be obtained through their respective providers; access to the in-house 0.55 T data is governed by participant consent and privacy requirements.

REFERENCES

Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alex Dimakis, and Jonathan Tamir. Ambient diffusion posterior sampling: Solving inverse problems with diffusion models trained on corrupted data. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2dd7f33ffbb59b4ff987be5442a13016-Abstract-Conference.html.

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, 2019. doi: 10.1109/TMI.2018.2865356.

Jalal B. Andre, Brian W. Bresnahan, Mahmud Mossa-Basha, Michael N. Hoff, C. Patrick Smith, Yoshimi Anzai, and Wendy A. Cohen. Toward quantifying the prevalence, severity, and cost associated with patient motion during clinical MR examinations. *Journal of the American College of Radiology*, 12(7):689–695, 2015. doi: 10.1016/j.jacr.2015.03.007.

Philip G. Batchelor, David Atkinson, Pablo Irarrazaval, Derek L. G. Hill, Joseph V. Hajnal, and David J. Larkman. Matrix description of general motion correction applied to multishot images. *Magnetic Resonance in Medicine*, 54(5):1273–1280, 2005. doi: 10.1002/mrm.20656.

Yannick Brackenier, Nan Wang, Congyu Liao, Xiaozhi Cao, Sophie Schauman, Mahmut Yurt, Lucilio Cordero-Grande, Shaihan J. Malik, Adam Kerr, Joseph V. Hajnal, and Kawin Setsompop. Rapid and accurate navigators for motion and B0 tracking using QUEEN: Quantitatively enhanced parameter estimation from navigators. *Magnetic Resonance in Medicine*, 91(5):2028–2043, 2024. doi: 10.1002/mrm.29976.

Lixuan Chen, Xuanyu Tian, Jiangjie Wu, Ruimin Feng, Guoyan Lao, Yuyao Zhang, Hongen Liao, and Hongjiang Wei. Joint coil sensitivity and motion correction in parallel MRI with a self-calibrating score-based diffusion model. *Medical Image Analysis*, 102:103502, 2025. doi: 10.1016/j.media.2025.103502.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Lucilio Cordero-Grande, Rui Pedro A. G. Teixeira, Emer J. Hughes, Jana Hutter, Anthony N. Price, and Joseph V. Hajnal. Sensitivity encoding for aligned multishot magnetic resonance reconstruction. *IEEE Transactions on Computational Imaging*, 2(3):266–280, 2016. doi: 10.1109/TCI.2016.2557069.

Melissa W. Haskell, Stephen F. Cauley, Berkin Bilgic, Julian Hossbach, Daniel N. Splitthoff, Josef Pfeuffer, Kawin Setsompop, and Lawrence L. Wald. Network accelerated motion estimation and reduction (NAMER): Convolutional neural network guided retrospective motion correction using a separable motion model. *Magnetic Resonance in Medicine*, 82(4):1452–1461, 2019. doi: 10.1002/mrm.27771.

Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. doi: 10.6028/jres.049.044.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Julian Hossbach, Daniel Nicolas Splitthoff, Stephen Cauley, Bryan Clifford, Daniel Polak, Wei-Ching Lo, Heiko Meyer, and Andreas Maier. Deep learning-based motion quantification from k-space for fast model-based magnetic resonance imaging motion correction. *Medical Physics*, 50(4):2148–2161, 2023. doi: 10.1002/mp.16119.

Tobit Klug, Kun Wang, Stefan Ruschke, and Reinhard Heckel. MotionTTT: 2D test-time-training motion estimation for 3D motion corrected MRI. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1285. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/47a6e9e2c3019f13ad94a0f259fe4970-Abstract-Conference.html.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Brett Levac, Sidharth Kumar, Ajil Jalal, and Jonathan I. Tamir. Accelerated motion correction with deep generative diffusion models. *Magnetic Resonance in Medicine*, 92(2):853–868, 2024. doi: 10.1002/mrm.30082.

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*, 10:264, 2023. doi: 10.1038/s41597-023-02181-4.

Antonio Ortiz-Gonzalez, Erich Kobler, Lukas Schletter, and Alexander Effland. MotionDPS: Motion-compensated 3-d brain MRI reconstruction. *IEEE Transactions on Medical Imaging*, 45(8):4591–4604, 2026. doi: 10.1109/TMI.2026.3705847.

Chicago Y. Park, Shirin Shoushtari, Hongyu An, and Ulugbek Kamilov. Measurement score-based diffusion model. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/9bcc5e97c6343b52dc4ff7b034cc640c-Abstract-Conference.html.

Daniel Polak, Daniel Nicolas Splitthoff, Bryan Clifford, Wei-Ching Lo, Susie Y. Huang, John Conklin, Lawrence L. Wald, Kawin Setsompop, and Stephen Cauley. Scout accelerated motion estimation and reduction (SAMER). *Magnetic Resonance in Medicine*, 87(1):163–178, 2022. doi: 10.1002/mrm.28971.

Daniel Polak, Julian Hossbach, Daniel Nicolas Splitthoff, Bryan Clifford, Wei-Ching Lo, Azadeh Tabari, Min Lang, Susie Y. Huang, John Conklin, Lawrence L. Wald, and Stephen Cauley. Motion guidance lines for robust data consistency-based retrospective motion correction in 2D and 3D MRI. *Magnetic Resonance in Medicine*, 89(5):1777–1790, 2023. doi: 10.1002/mrm.29534.

Gabrio Rizzuti, Tim Schakel, Niek R. F. Huttinga, Jan Willem Dankbaar, Tristan van Leeuwen, and Alessandro Sbrizzi. Towards retrospective motion correction and reconstruction for clinical 3D brain MRI protocols with a reference contrast. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 37:807–823, 2024. doi: 10.1007/s10334-024-01161-y.

Nalini M. Singh, Neel Dey, Malte Hoffmann, Bruce Fischl, Elfar Adalsteinsson, Robert Frost, Adrian V. Dalca, and Polina Golland. Data consistent deep rigid MRI motion correction. In *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pp. 368–381. PMLR, 2024. URL https://proceedings.mlr.press/v227/singh24a.html.

Veronika Spieker, Hannah Eichhorn, Kerstin Hammernik, Daniel Rueckert, Christine Preibisch, Dimitrios C. Karampinos, and Julia A. Schnabel. Deep learning for retrospective motion correction in MRI: A comprehensive review. *IEEE Transactions on Medical Imaging*, 43(2):846–859, 2024. doi: 10.1109/TMI.2023.3323215.

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. doi: 10.1109/TIP.2003.819861.

Qing Wu, Chenhe Du, Xuanyu Tian, Jingyi Yu, Yuyao Zhang, and Hongjiang Wei. Moner: Motion correction in undersampled radial MRI with unsupervised neural representation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OdnqG1fYpo.

Maxim Zaitsev, Julian Maclaren, and Michael Herbst. Motion artifacts in MRI: A complex problem with many partial solutions. *Journal of Magnetic Resonance Imaging*, 42(4):887–901, 2015. doi: 10.1002/jmri.24850.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI. arXiv preprint arXiv:1811.08839, 2018. URL https://arxiv.org/abs/1811.08839.

A Appendix

A.1 Complete Theoretical Derivation and Proof

This appendix develops the probabilistic argument underlying Section 2. It distinguishes the full-grid acquisition used to generate observations, the inference-window model used to express the latent-reference posterior, and the band-limited operator used to evaluate reference-dependent evidence. The posterior and ELBO identities below are exact for their stated likelihood and realized reference set; the learned reference distribution and the band-limited forward operator introduce separate approximations.

Full-grid observations and the inference-window likelihood. Let $x \in \mathbb{C}^{H\times W}$ be a motion-free object on the original image grid, $S_c^{H,W}$ its fixed full-grid sensitivity map for coil c, $\mathcal{F}_{H,W}$ the centered Fourier transform, and $\mathcal{T}_{\theta_t}^{H,W}$ the rigid warp at that grid size. The motion vector $\theta_t = (\varphi_t, t_{x,t}, t_{y,t})$ is relative to the static ACS pose; its translations are expressed in original-grid pixels.

Let $\mathcal{C}_L$ crop Fourier data to a central $L \times L$ window, and let $M_A^{H,W}$ and $M_t^{H,W}$ select the disjoint ACS and dynamic-shot samples on the full grid. The simulator's two observation paths are

$$\begin{aligned} a_c &= \mathcal{C}_L M_A^{H,W} \mathcal{F}_{H,W}(S_c^{H,W} \odot x) + \epsilon_{a,c}, \\ y_{t,c} &= \mathcal{C}_L M_t^{H,W} \mathcal{F}_{H,W}\left[\mathcal{T}_{\theta_t}^{H,W}(S_c^{H,W} \odot x)\right] + \epsilon_{t,c}. \end{aligned} \tag{17}$$

Thus motion is applied before the Fourier transform and central crop. Equation 1 in the main text is an effective model written in terms of a central-window object r; it is not an algebraic rewrite of Eq.17. In particular, cropping and rotation need not commute, and warping $S_c^{H,W} \odot x$ is not generally equivalent to multiplying the warped object by an unwarped sensitivity map.

For clarity about the reference-based computation, let $\mathcal{P}_L$ embed an $L \times L$ k-space array into the center of the $H \times W$ grid, and let $\widehat{k}_c^L(r)$ denote the coil k-space view synthesized from the compact reference (for example, $\mathcal{F}_L(S_{c,L} \odot r)$, where $S_{c,L}$ corresponds to the effective S_c in Eq.1). The band-limited prediction used for a candidate reference is

$$\mathcal{A}_{t,c}^{\mathrm{BL}}(r, \theta_t) = M_{t,L} \mathcal{C}_L \mathcal{F}_{H,W}\left[\mathcal{T}_{\theta_t}^{H,W}\left(\mathcal{F}_{H,W}^{-1} \mathcal{P}_L \widehat{k}_c^L(r)\right)\right], \tag{18}$$

where $M_{t,L}$ restricts the cropped result to the acquired samples of shot t. This operator preserves the original-grid motion coordinates, but discards the outer k-space content before warping. It also uses a reference-synthesized coil view in place of the true full-resolution coil image. These are modeling approximations, not consequences of the Fourier crop. The formulas below hold with either a stipulated inference-window likelihood or a likelihood built from $\mathcal{A}^{\mathrm{BL}}$; their probability identities do not imply that the two forward models are equal.

Write $a = \{a_c\}_c$ and $y_t = \{y_{t,c}\}_c$ for the distinct observed samples, with $y = \{y_t\}_{t=1}^T$. Let $\mathcal{A}_t(r, \theta_t)$ stack the predicted dynamic samples across coils under the chosen reference-based forward model. If the errors on those acquired coordinates are independent across shots and independent of ACS errors conditional on the latent variables, and each dynamic error is proper complex Gaussian with a fixed positive-definite covariance Γ_t, then

$$p(y_t \mid r, \theta_t) = \frac{\exp\left[-(y_t - \mathcal{A}_t(r, \theta_t))^H \Gamma_t^{-1} (y_t - \mathcal{A}_t(r, \theta_t))\right]}{\det(\pi \Gamma_t)}. \tag{19}$$

An analogous density under the static ACS forward relation defines $p(a \mid r)$. The covariance and the set of acquired coordinates are fixed with respect to r and Θ when comparing reference hypotheses. ACS data condition the reference distribution and are not counted again as independent dynamic-shot measurements in the product below. Gaussianity gives a quadratic discrepancy, while the reference-marginalization argument requires only a well-defined conditional likelihood and the stated conditional independences.

A latent reference and the motion posterior. The ACS-conditioned anatomy satisfies $p(r \mid a) \propto p(a \mid r)p(r)$. Assume that, before observing the dynamic shots, $p(r, \Theta \mid a) = p(r \mid a)p(\Theta \mid a)$ and that the dynamic-shot errors are conditionally independent given their common r and Θ. Then

$$p(y, r, \Theta \mid a) = p(\Theta \mid a)\, p(r \mid a) \prod_{t=1}^{T} p(y_t \mid r, \theta_t). \tag{20}$$

Integrating the joint density first over r and then normalizing over Θ yields

$$p(\Theta \mid y, a) = \frac{p(\Theta \mid a) \int p(r \mid a) \prod_t p(y_t \mid r, \theta_t)\, dr}{\int p(\Theta' \mid a) \left[\int p(r \mid a) \prod_t p(y_t \mid r, \theta_t')\, dr\right] d\Theta'}. \tag{21}$$

The same r appears in every factor inside the integral. Marginalizing a separate reference for each shot would instead define a different latent-variable model.

CARS and the learned reference distribution. The residual representation can be derived without identifying a noisy ACS image with a noiseless anatomical component. For one noiseless central multi-coil k-space realization $k^\star$, let P_A project onto the ACS-supported coordinates and set

$k_A^\star = P_A k^\star$ and $k_U^\star = (I - P_A)k^\star$. If $\mathcal{C}$ is the same fixed complex-linear k-space-to-object map in both reconstructions, linearity gives

$$\begin{aligned} r^\star &= \mathcal{C}(k_A^\star + k_U^\star), \qquad & r_A^\star &= \mathcal{C}(k_A^\star), \\ h^\star &:= r^\star - r_A^\star = \mathcal{C}(k_U^\star). \end{aligned} \tag{22}$$

This identity explains the ACS-unresolved frequency contribution in the noiseless setting. For a measured, potentially noisy ACS observation a, CARS instead uses its derived baseline $r_A(a)$ and defines the random residual $h = r - r_A(a)$. Conditional residual sampling $h \sim q_\psi(h \mid a)$ induces, by this translation, a distribution $q_\psi(r \mid a)$ on $r = r_A(a) + h$. Neither Eq.22 nor conditioning on a proves $q_\psi(r \mid a) = p(r \mid a)$.

Replacing the unknown ACS posterior in Eq.21 by this learned distribution defines

$$\widetilde{p}_\psi(y \mid \Theta, a) = \int q_\psi(r \mid a) \prod_{t=1}^{T} p(y_t \mid r, \theta_t)\, dr. \tag{23}$$

The tilde marks a likelihood under the learned reference law. It should not be read as an exact evaluation of the likelihood under the unknown $p(r \mid a)$.

A finite set with one shared reference index. Draw $\mathcal{R}_K = \{r^{(k)}\}_{k=1}^K$ from $q_\psi(r \mid a)$. Within this realized set, let $Z \sim \mathrm{Cat}(\pi_1, \ldots, \pi_K)$ be a single reference index shared by the case, where $\pi_k > 0$ and $\sum_k \pi_k = 1$. The resulting finite-reference likelihood is

$$p_K(y \mid \Theta, \mathcal{R}_K) = \sum_{k=1}^{K} \pi_k \prod_{t=1}^{T} p(y_t \mid r^{(k)}, \theta_t). \tag{24}$$

The product must precede the sum: $\prod_t \sum_k \pi_k p(y_t \mid r^{(k)}, \theta_t)$ assigns an independent latent index to every shot and no longer represents one shared anatomy. The finite-set motion posterior is $p_K(\Theta \mid y, a, \mathcal{R}_K) \propto p(\Theta \mid a) p_K(y \mid \Theta, \mathcal{R}_K)$.

Suppose that, conditional on a, the references $r^{(1)}, \ldots, r^{(K)}$ are independent and identically distributed samples from $q_\psi(r \mid a)$. At fixed Θ, define $L(r) = \prod_{t=1}^{T} p(y_t \mid r, \theta_t)$. For deterministic, nonnegative weights satisfying $\sum_{k=1}^{K} \pi_k = 1$, and provided that $L(r)$ has a finite second moment,

$$\begin{aligned} \mathbb{E}_{\mathcal{R}_K}[p_K(y \mid \Theta, \mathcal{R}_K)] &= \mathbb{E}_{q_\psi}[L(r)] = \widetilde{p}_\psi(y \mid \Theta, a), \\ \mathrm{Var}_{\mathcal{R}_K}[p_K(y \mid \Theta, \mathcal{R}_K)] &= \mathrm{Var}_{q_\psi}[L(r)] \sum_{k=1}^{K} \pi_k^2. \end{aligned} \tag{25}$$

Thus deterministic normalized weighting preserves unbiasedness; uniform weights $\pi_k = 1/K$ additionally give the variance $\mathrm{Var}_{q_\psi}[L(r)]/K$. If the weights depend on the sampled references, unbiasedness requires separate justification. When p_K is positive and the expectations exist, Jensen's inequality also gives $\mathbb{E}_{\mathcal{R}_K}[\log p_K] \le \log \widetilde{p}_\psi(y \mid \Theta, a)$. Hence unbiasedness of p_K does not extend to its logarithm. None of these statements guarantees that increasing K improves the learned motion estimator.

For a fixed proposed trajectory, Bayes' rule on the realized finite set gives the physical responsibility

$$\gamma_k(\Theta) := p(Z = k \mid y, \Theta, \mathcal{R}_K) = \frac{\pi_k \prod_t p(y_t \mid r^{(k)}, \theta_t)}{\sum_j \pi_j \prod_t p(y_t \mid r^{(j)}, \theta_t)}. \tag{26}$$

It is exact for the finite mixture and chosen likelihood at that Θ, not for the original continuous-reference posterior.

Amortized reference and shot-motion distributions. The likelihood-derived $\gamma(\Theta)$ depends on a proposed trajectory. RAMI instead amortizes inference from observations and the sampled set. For each shot, an observation-only query q_t attends to pair tokens $u_{t,k}$ that encode the measured shot, the kth reference view, and their discrepancy. With $U_t = \{u_{t,k}\}_k$, the multi-head computation can

be written

$$\ell_{t,k}^{(j)} = \frac{(W_Q^{(j)} q_t)^\top (W_K^{(j)} u_{t,k})}{\sqrt{d_j}}, \quad \alpha_{t,k}^{(j)} = \mathrm{softmax}_k(\ell_{t,k}^{(j)}),$$
$$g_t = \mathrm{MHA}(q_t, U_t, U_t), \qquad \ell_{t,k} = \frac{1}{J}\sum_{j=1}^{J} \ell_{t,k}^{(j)}. \tag{27}$$

Here j indexes attention heads and d_j is their projected key dimension. The per-head softmax weights aggregate value tokens into g_t. The head-averaged *raw* logits, rather than averaged normalized weights, define the case-level categorical distribution

$$\rho_{\omega,k} = \mathrm{softmax}_k\left(\log \pi_k + \sum_{t=1}^{T} \ell_{t,k}\right). \tag{28}$$

Thus $\alpha_{t,k}^{(j)}$ describes how shot t reads a reference token, whereas $\rho_{\omega,k}$ supplies one distribution over the shared Z. The learned logits are not stipulated to equal shot log-likelihoods. In particular, ρ_ω does not condition explicitly on a candidate Θ and need not equal $\gamma(\Theta)$.

Separate heads applied to g_t produce a motion-occurrence probability p_t and the mean and diagonal covariance of a moved-shot rigid-motion distribution. The corresponding shot-wise marginal is

$$q_\phi(\theta_t \mid y, a, \mathcal{R}_K) = (1 - p_t)\delta_0 + p_t \mathcal{N}(\mu_t, \Sigma_t), \qquad p_t = q_\phi(z_t = 1 \mid y, a, \mathcal{R}_K). \tag{29}$$

Here $z_t = 0$ denotes exact zero motion, and δ_0 is its point mass. If the continuous head outputs normalized mean $\bar{\mu}_t$ and diagonal covariance $\bar{\Sigma}_t$, a diagonal scale map D gives $\mu_t = D\bar{\mu}_t$ and $\Sigma_t = D\bar{\Sigma}_t D^\top$ in the physical coordinates of θ_t. These are shot-wise marginals; they do not by themselves define all dependencies in a joint $q_\phi(\Theta \mid y, a, \mathcal{R}_K)$. The reference-index ELBO below conditions on a fixed trajectory and does not integrate this motion distribution.

Reference-index ELBO and its exact finite-set gap. For fixed Θ and $\mathcal{R}_K$, abbreviate $p_k = p(y \mid \Theta, r^{(k)}) = \prod_t p(y_t \mid r^{(k)}, \theta_t)$. Let ρ be any categorical distribution whose support lies within that of the strictly positive π and p_k. Inserting ρ_k/ρ_k into Eq.24 and applying Jensen's inequality yields

$$\begin{aligned} \log p_K(y \mid \Theta, \mathcal{R}_K) &= \log \sum_{k=1}^{K} \rho_k \frac{\pi_k p_k}{\rho_k} \\ &\geq \sum_{k=1}^{K} \rho_k \log \frac{\pi_k p_k}{\rho_k} \\ &= \sum_{k=1}^{K} \rho_k \log p_k - D_{\mathrm{KL}}(\rho \| \pi) =: \mathcal{E}_{\mathrm{ref}}(\Theta, \rho). \end{aligned} \tag{30}$$

For zero components of ρ, the same formula uses the standard $0 \log 0 = 0$ convention and can be obtained by continuity. Substituting Eq.26 into the KL divergence gives the exact identity

$$\begin{aligned} D_{\mathrm{KL}}(\rho \| \gamma(\Theta)) &= \sum_k \rho_k \log \frac{\rho_k p_K(y \mid \Theta, \mathcal{R}_K)}{\pi_k p_k} \\ &= \log p_K(y \mid \Theta, \mathcal{R}_K) - \mathcal{E}_{\mathrm{ref}}(\Theta, \rho). \end{aligned} \tag{31}$$

The bound is tight at $\rho = \gamma(\Theta)$ if ρ may vary freely on the simplex. Setting $\rho = \rho_\omega(y, a, \mathcal{R}_K)$ restricts that family, so equality is not guaranteed for every Θ. The identity concerns $\log p_K$ at a realized reference set and fixed motion hypothesis. It is not a lower bound on $\log \widetilde{p}_\psi$, and it does not integrate over uncertainty in Θ.

Energy form and the objective actually evaluated. Suppose the chosen conditional likelihood has a common-scale form $\log p_k = -E_k(\Theta)/\tau + C_0$, with $\tau > 0$ and C_0 independent of k and Θ. Multiplying the negative ELBO by τ and dropping the common additive constant gives

$$\mathcal{L}_{\mathrm{ELBO}}(\Theta, \rho) = \sum_{k=1}^{K} \rho_k E_k(\Theta) + \tau D_{\mathrm{KL}}(\rho \| \pi). \tag{32}$$

The corresponding finite-set negative log marginal, its optimizer over unrestricted ρ, and the energy-unit gap are

$$\begin{aligned} \mathcal{L}_{\mathrm{marg}}(\Theta) &= -\tau \log \sum_k \pi_k e^{-E_k(\Theta)/\tau}, \\ \gamma_k(\Theta) &= \frac{\pi_k e^{-E_k(\Theta)/\tau}}{\sum_j \pi_j e^{-E_j(\Theta)/\tau}}, \\ \mathcal{L}_{\mathrm{ELBO}}(\Theta, \rho) - \mathcal{L}_{\mathrm{marg}}(\Theta) &= \tau D_{\mathrm{KL}}(\rho \| \gamma(\Theta)). \end{aligned} \tag{33}$$

For $K = 1$, both categorical distributions equal one and the KL term vanishes. If E_k is a normalized or otherwise surrogate discrepancy, these formulas still define an algebraic finite-mixture bound for positive scores $e^{-E_k/\tau}$. Calling γ a *physical Bayesian posterior* then additionally requires a properly normalized likelihood whose normalizing factor does not change across reference and motion hypotheses. A surrogate alone does not establish that calibration.

In the training objective of Section 2.4, $\mathcal{L}_{\mathrm{ELBO}}$ is evaluated at the predicted trajectory $\widehat{\Theta}$ and combined with supervised motion terms. This plug-in evaluation is distinct from an ELBO that would also take an expectation over $q_\phi(\Theta \mid y, a, \mathcal{R}_K)$. Accordingly, the exact gap in Eq.31 characterizes reference-index inference conditional on $\widehat{\Theta}$; it is not a proof that the amortized shot-motion marginals equal the true motion posterior.

A.2 CARS and RAMI Network Implementation Details

This appendix records architectural and optimization details omitted from the main experimental description. The main text emphasizes what each model receives and predicts; the configurations below specify how those representations are implemented.

CARS. We train separate CARS models for fastMRI and M4Raw at each ACS width. CARS is an epsilon-predicting residual DDPM with a U-Net whose base width is 64 channels, whose level multipliers are $(1, 2, 4)$, and which uses one residual block per level. The linear noise schedule has 1,000 steps. The objective combines the denoising loss with auxiliary residual, missing-frequency multi-coil k-space, and measured-ACS losses, weighted by 0.15, 0.10, and 0.05, respectively. Sampling uses 1,000 reverse steps. Training uses batch size 16 and initial learning rate 10^{-4}.

RAMI input encoders and set attention. For each case, four coils are selected by the mean squared magnitude of their measured ACS k-space. Each observation-reference pair is jointly scaled by the larger magnitude peak of the two views. The pair encoder concatenates the complex shot observation, the corresponding reference-synthesized view, and their difference; real and imaginary components give six input channels. A shared per-coil CNN uses 3×3 convolutions with channel widths $32, 32, 64, 64, 128$, GroupNorm and SiLU activations, and two stride-two reductions. Adaptive average pooling to 4×4 is followed by a linear projection, SiLU, and LayerNorm to form a 256-dimensional token. The selected-coil features are combined by a mask-weighted mean.

An observation-only query encoder takes the complex observation as two real/imaginary channels and uses the same convolutional widths, downsampling pattern, pooling grid, and 256-dimensional projection. Its selected-coil features are likewise mask-averaged. Four-head multi-head attention uses this query to aggregate the set of reference-conditioned pair tokens for each shot; a residual connection and LayerNorm produce the shared shot representation.

RAMI output heads. The occurrence head is a $256 \to 64 \to 1$ MLP with GELU and predicts the movement logit. The conditional-motion head is a $256 \to 128 \to 6$ MLP with GELU and LayerNorm. Three outputs parameterize the normalized motion mean and are passed through $\tanh$ before scaling to rotation and translation units; the other three parameterize the diagonal log variance, clamped to $[-8, 3]$. The resulting p_t, μ_t, and Σ_t define the shot-wise zero-motion and Gaussian mixture in Eq.12.

A.3 Evaluation Metrics

This appendix defines the motion and image metrics used in Section 3. Image metrics are computed for each reconstructed slice, with one vote per slice. For slice i and shot t, let $\theta_{i,t} = (\alpha_{i,t}, t_{x,i,t}, t_{y,i,t})$ be the simulated rigid motion and $\widehat{\theta}_{i,t}$ its estimate. Rotation is measured in degrees and translations in pixels on each cohort's native grid. For a set Q of slice–shot pairs, we use

$$\begin{aligned} \mathrm{MAE}_q(Q) &= \frac{1}{|Q|} \sum_{(i,t)\in Q} |\widehat{q}_{i,t} - q_{i,t}|, \quad q \in \{\alpha, t_x, t_y\}, \\ \mathrm{nMAE}(Q) &= \frac{1}{3}\left(\frac{\mathrm{MAE}_\alpha(Q)}{8^\circ} + \frac{\mathrm{MAE}_{t_x}(Q)}{8\,\mathrm{px}} + \frac{\mathrm{MAE}_{t_y}(Q)}{8\,\mathrm{px}} \right), \\ \mathrm{NME}(Q) &= \frac{1}{|Q|} \sum_{(i,t)\in Q} \left\| \mathrm{diag}(8^\circ, 8\,\mathrm{px}, 8\,\mathrm{px})^{-1} (\widehat{\theta}_{i,t} - \theta_{i,t}) \right\|_2 . \end{aligned} \tag{34}$$

Table 2 uses the dimensionless nMAE after per-case motion refinement; Table 4 gives its component MAEs in degrees and pixels. Both average all eight shots per test slice before taking slice-equal cohort means. The reference-count and ACS-budget experiments use NME on their stated grids. Their moving-condition macro mean weights each of the 12 or 16 moving conditions equally; the all-static condition is excluded, while static shots within a moving condition remain included. The factors of eight are fixed evaluation scales, not the motion bounds of each condition. Unlike nMAE, NME takes the Euclidean norm of the scaled error for each shot before averaging.

Table 4: Post-refinement component motion MAEs for the completed $2\times$ results in Table 2. Entries are slice-equal means over all eight shots; arrows denote zero-shot $\rightarrow$ both-stage fine-tuning.

Test cohort	ACS	Rotation (°)	t_x (px)	t_y (px)
fastMRI FLAIR	24	0.092	0.095	0.054
	32	0.137	0.111	0.097
M4Raw T1	24	0.103	0.102	0.163
	32	0.186	0.148	0.237
In-house T1	32	$0.429 \rightarrow 0.187$	$0.125 \rightarrow 0.105$	$0.330 \rightarrow 0.153$
	24	$2.568 \rightarrow 0.270$	$1.218 \rightarrow 0.155$	$2.164 \rightarrow 0.342$

Full-image magnitude normalization. For each reconstructed slice, let $u_j = |x_j^{\mathrm{target}}|$ and $v_j = |x_j^{\mathrm{recon}}|$ at pixel $j \in \{1, \ldots, N\}$. Prediction and target are independently normalized on the full image:

$$\widetilde{u}_j = \frac{u_j - \min_k u_k}{\max_k u_k - \min_k u_k}, \qquad \widetilde{v}_j = \frac{v_j - \min_k v_k}{\max_k v_k - \min_k v_k}. \tag{35}$$

If an image has zero intensity range, its normalized image is defined as zero, matching the metric implementation's finite-value safeguard. Independent normalization removes absolute intensity-scale differences and does not assess their calibration.

PSNR and related image errors. Set $d_j = \widetilde{v}_j - \widetilde{u}_j$ and $\mathrm{MSE} = N^{-1} \sum_j d_j^2$. With unit data range, the per-slice metrics are

$$\begin{aligned} \mathrm{PSNR} &= 10 \log_{10}(1/\mathrm{MSE}), \qquad \mathrm{NMSE} = \frac{\sum_j d_j^2}{\sum_j \widetilde{u}_j^2}, \\ \mathrm{NRMSE} &= \sqrt{\mathrm{NMSE}}, \qquad \mathrm{MAE}_{\mathrm{img}} = N^{-1} \sum_j |d_j|, \\ \mathrm{RMSE} &= \sqrt{\mathrm{MSE}}. \end{aligned} \tag{36}$$

The implementation clamps a vanishing MSE or target-energy denominator to a small positive value before division. Image MAE is distinct from rotation MAE in Eq.34. PSNR and NRMSE occur in the main reconstruction results; NMSE, image MAE, and RMSE identify supplementary per-slice calculations.

Structural similarity. SSIM (Wang et al., 2004) is computed on $\widetilde{u}$ and $\widetilde{v}$. At each local window, let μ_u, μ_v be Gaussian-weighted means, σ_u^2, σ_v^2 the corresponding population variances, and σ_{uv} their population covariance. The local score is

$$\mathrm{SSIM}_{\mathrm{local}} = \frac{(2\mu_u\mu_v + C_1)(2\sigma_{uv} + C_2)}{(\mu_u^2 + \mu_v^2 + C_1)(\sigma_u^2 + \sigma_v^2 + C_2)}, \qquad C_1 = (0.01)^2, \quad C_2 = (0.03)^2. \tag{37}$$

Gaussian weights use an 11×11 window with $\sigma = 1.5$, population covariance, and data range one. Per-slice SSIM is the mean local score over valid window centers on the full image, without an anatomical mask. The pending in-house entries in Table 2 will use this protocol.

A.4 Reconstruction Comparison Protocols

Common cohort and scoring. Table 1 uses 575 fastMRI FLAIR test slices for each ACS24/32 and $2\times/3\times$ setting. Reconstructed and target magnitudes are independently normalized per slice and scored over the full image, as defined in Appendix A.3. PSNR and SSIM are averaged with equal weight per slice. A baseline enters the 575-slice table only after its test slices, acquired measurements, ACS width, acceleration, and scoring protocol are matched.

MARMI reconstruction backends. CARS supplies references for RAMI motion estimation. Starting from the RAMI trajectory, we perform three rounds of per-case refinement against the acquired measurements, with 20 motion-parameter updates and 50 auxiliary-image updates per round (learning rates 0.01 and 0.03, respectively). This refinement updates the estimated trajectory and the auxiliary image while keeping all network weights fixed. MARMI+CG20 and MARMI+MoDL then use the same refined trajectory for each case. The former fixes it and applies 20 conjugate-gradient iterations (Hestenes & Stiefel, 1952) with ridge 10^{-5}; the latter fixes it within a motion-aware forward operator and reconstructs with a separately trained MoDL network.

Our MoDL is a PyTorch reimplementation of the published architecture (Aggarwal et al., 2019), not a run of the authors' released code or pretrained weights. We retain the released code's five-layer, 64-channel residual CNN and shared-weight unrolling, and use five unrolls with ten complex-CG steps per data-consistency update. We adapt the data-consistency operator to the estimated shot-wise motion at test time and constrain its learned regularization parameter to be positive with a softplus parameterization. Ground truth is used for supervised MoDL training and offline scoring, not as a test-time input.

MoPED-GT adaptation. MoPED (Hossbach et al., 2023) infers shot-wise motion from corrupted low-resolution k-space and a clean reference. Because a runnable official implementation and checkpoint were unavailable, we trained a FLAIR-specific PyTorch adaptation with paper-style DenseBlocks and eight shot-wise motion heads. Its four-coil, 64×64 input pairs the corrupted shot data with one clean reference. For this experiment, that reference was constructed offline from the corresponding test target, making the arm GT-assisted (MoPED-GT). The ACS32/2$\times$ entry in Table 1 uses a checkpoint trained on 500 source-training slices with eight motion realizations per slice for 100 epochs. At test time, the predicted motion is updated through alternating image data-consistency and motion-residual steps, followed by fixed-motion CG20 reconstruction. This residual procedure is our physics-based adaptation of the published iteration, whose complete implementation was not available. The inference code reads the precomputed reference and measured data; test-target information enters through reference construction and later metric scoring.

JSMoCo. We adapt the authors' released PyTorch JSMoCo framework (Chen et al., 2025) to the same eight-shot FLAIR measurements, including the static ACS samples. The solver retains score-based joint image, motion, and coil-sensitivity estimation. We trained its NCSNv2 score prior on source-domain fastMRI FLAIR images. The corrected sensitivity estimate csm supplies the initial coil map. The direct-initialization arm uses the method's seeded motion initialization and no MARMI image or motion estimate; motion and coil-sensitivity estimation remain enabled. Recorded runs use a k-space phase-ramp motion operator, 611 outer steps, and 15 inner steps. JSMoCo receives no test target or simulated trajectory during reconstruction.

B Additional Qualitative Results

B.1 M4Raw T1 ACS24 scan at two acceleration factors

Figures 4 and 5 show the two middle slices of M4Raw scan at $R = 2$ and $R = 3$. The two horizontal rows show the acceleration factors; the columns show ground truth, zero-motion CG20, N3+CG20 reconstruction after motion inference and refinement, and absolute reconstruction error. Within each slice, the magnitude panels share the ground-truth 99.5th-percentile display window across both rates. Error is computed after independent per-slice min–max normalization of reconstruction and target, as in Appendix A.3; all error panels use the same $[0, 0.10]$ color scale.

For slice 03, reconstructed PSNR is 40.19 dB at $R = 2$ and 35.14 dB at $R = 3$, compared with 27.67 and 26.03 dB for zero-motion CG20. For slice 04, the corresponding reconstruction values are 39.31 and 36.09 dB, compared with 29.13 and 27.16 dB for zero-motion CG20. The reconstructions suppress much of the visible motion-related ripple; residual error is more apparent along the outer brain boundary and inferior structures at $R = 3$. These two slices are an illustrative selection from one scan, not cohort-level evidence.

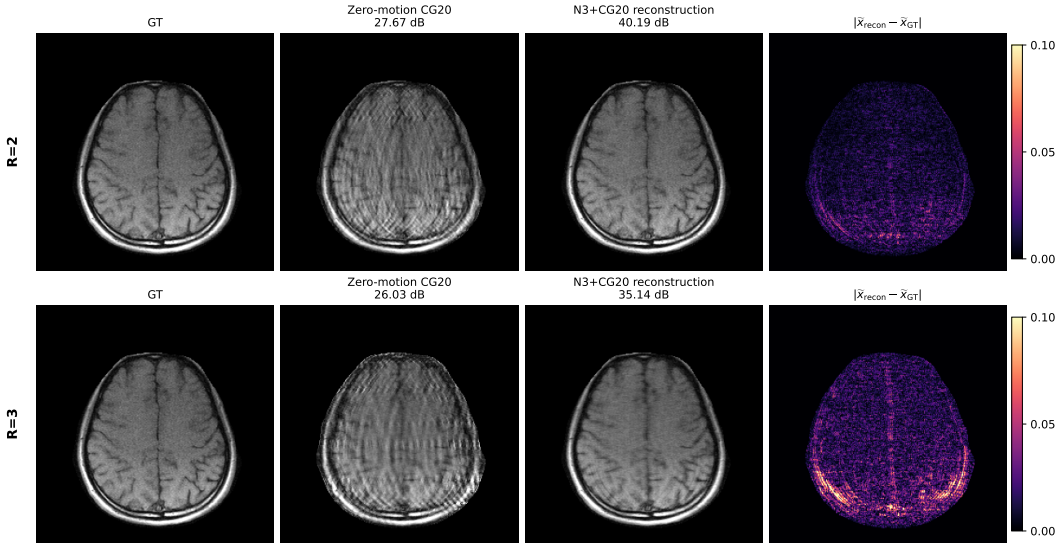

Figure 4: M4Raw T1 ACS24 scan, middle slice 03. Rows show $R = 2$ and $R = 3$; columns show ground truth, zero-motion CG20, N3+CG20 reconstruction, and normalized absolute reconstruction error.

B.2 In-house 0.55-T T1 ACS32 representative slices

Figures 6 and 7 show slices 12 and 13 of held-out subject of 0.55-T in-house dataset at $R = 2$ with ACS32. The four panels show the full-resolution target, zero-filled motion-corrupted observation, N3+CG20 reconstruction after fine-tuning both CARS and RAMI, and absolute reconstruction error. The zero-filled image is obtained by centered inverse Fourier transformation of the same simulated eight-shot measurement used for reconstruction, followed by sensitivity-weighted coil combination. It is a direct zero-fill baseline, whereas the M4Raw comparison above uses a zero-motion CG20 baseline. The target is the normalized full-resolution image formed from the stored multi-coil k-space with the physical sensitivity estimate. Within each slice, the magnitude panels share the target's 99.5th-percentile display window.

For slice 12, full-image independently min–max-normalized PSNR improves from 30.41 to 42.80 dB; for slice 13 it improves from 31.37 to 44.38 dB. The reconstructions reduce the visible blurring and streaking in the zero-filled images while retaining the ventricular shape and cortical folds. Remaining error is concentrated around high-contrast outer boundaries and appears as weak spatial texture within the brain, particularly on slice 12. The error is the absolute difference between independently min–max-normalized reconstruction and target, computed over the full image as in Appendix A.3; both figures use the same $[0, 0.10]$ error color scale.

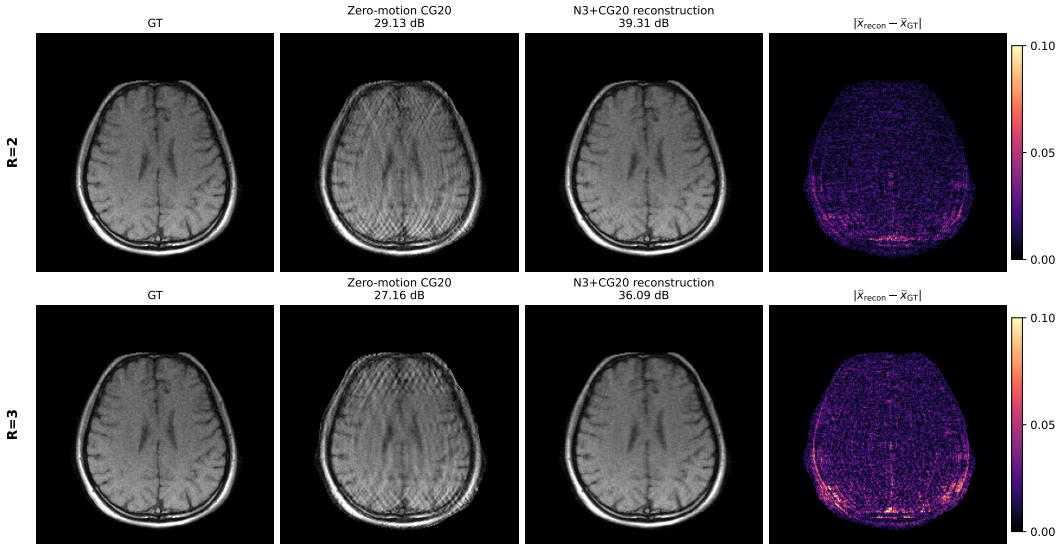

Figure 5: M4Raw T1 ACS24 scan, middle slice 04, with the same layout and error color scale as Figure 4.

These two slices are an illustrative selection from the nine retained slices of this subject. Across all slices, mean PSNR is 31.17 dB for zero-fill and 40.69 dB for N3+CG20. These single-subject values do not replace the two-subject mean, and the figures do not isolate the contributions of motion estimation, per-case refinement, and CG20.

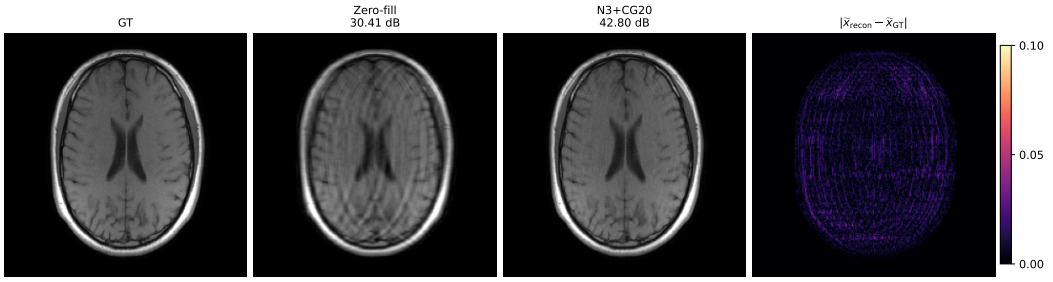

Figure 6: In-house 0.55-T T1 ACS32 subject, slice 12. Panels show target, zero-fill, N3+CG20 reconstruction, and normalized absolute reconstruction error. Panel PSNRs use full-image independent min–max normalization.

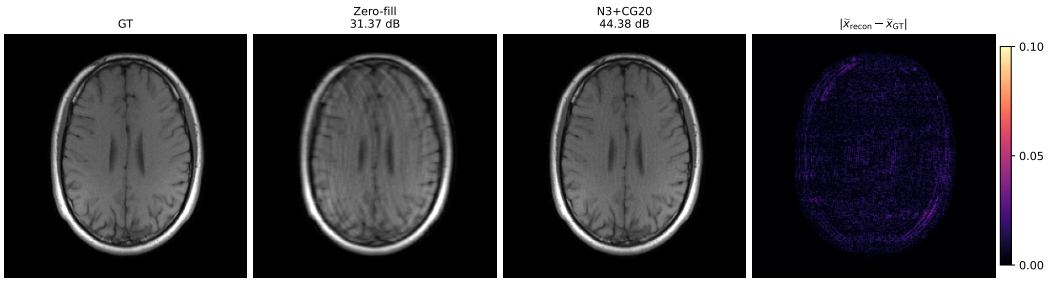

Figure 7: In-house 0.55-T T1 ACS32 subject
, slice 13, with the same layout and $[0, 0.10]$ error color scale as Figure 6.