EXEC-MRI: GRADIENT EXECUTION DEVIATIONS REVERSE MRI RECONSTRUCTION RANKINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

MRI reconstruction relies on knowing where signals were sampled in Fourier space. Hardware response can shift these locations, making the actual acquisition differ from the reconstruction model. We ask whether method rankings remain reliable under this mismatch. EXEC-MRI provides a controlled evaluation protocol for MRI with non-Cartesian sampling paths, such as spirals. It pairs simulated gradient execution errors with tests of trajectory knowledge, numerical settings, and training. On fastMRI Brain, $22/28$ method pairs change mean-score order from matched acquisition to moderate mismatch; reversals also occur on two external datasets. Numerical choices and model adaptation can themselves change the ordering, showing that robustness comparisons depend on the complete reconstruction pipeline rather than architecture alone. These retrospective simulations show why nominal performance alone does not establish relative performance under acquisition error. EXEC-MRI makes the conditions of such comparisons explicit and reproducible.

1 INTRODUCTION

MRI reconstruction turns measured signals into an image using a model of how the scanner produced them. Benchmarks have driven advances in learned reconstruction (Muckley et al., 2021; Sriram et al., 2020; Xin et al., 2023), but comparisons with an accurate acquisition model leave a question open: *if that model is slightly wrong, does the method that scored better still score better?* Beyond reducing image quality, model error may change the relative order of methods.

In MRI, each measurement samples a location in spatial-frequency space, called k-space. Magnetic-field gradients control the path through these locations, or *sampling trajectory*. Non-Cartesian acquisitions follow paths such as spirals rather than a regular grid. Hardware delays and other response effects can shift sampling locations away from the commanded path. We call this *gradient execution deviation* (GED). Reconstruction using the planned, or *nominal*, trajectory then interprets measurements at the wrong coordinates (Figure 1). We simulate this error with a gradient impulse response function (GIRF), an established linear model of hardware response (Vannesjo et al., 2013; 2016; Brodsky et al., 2009; Campbell-Washburn et al., 2016).

From calibration to method comparison. Trajectory error and its correction are well studied. We examine their implications for method comparisons through three questions: when do rankings change, what does supplying the actual trajectory recover, and how do numerical settings and training affect the ordering? A learned model may receive unfamiliar inputs even after coordinates are corrected, so trajectory knowledge and compatibility with training must be examined separately.

The EXEC-MRI protocol. We compare matched coordinates, mismatched coordinates, and reconstruction informed by the executed trajectory on the same images. The last comparison fixes measurements and learned weights. We then vary the numerical weights that account for uneven sampling and test training interventions. These paired comparisons locate the conditions under which rankings change. They evaluate specified implementations in retrospective simulation, without isolating each component's causal contribution or estimating clinical error prevalence.

Contributions. First, we provide a reproducible protocol for comparing methods under controlled acquisition-model mismatch, with paired scores and uncertainty over volumes. Second, we demonstrate ranking changes: $22/28$ method pairs reverse mean-score order by moderate GED on Brain;

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.

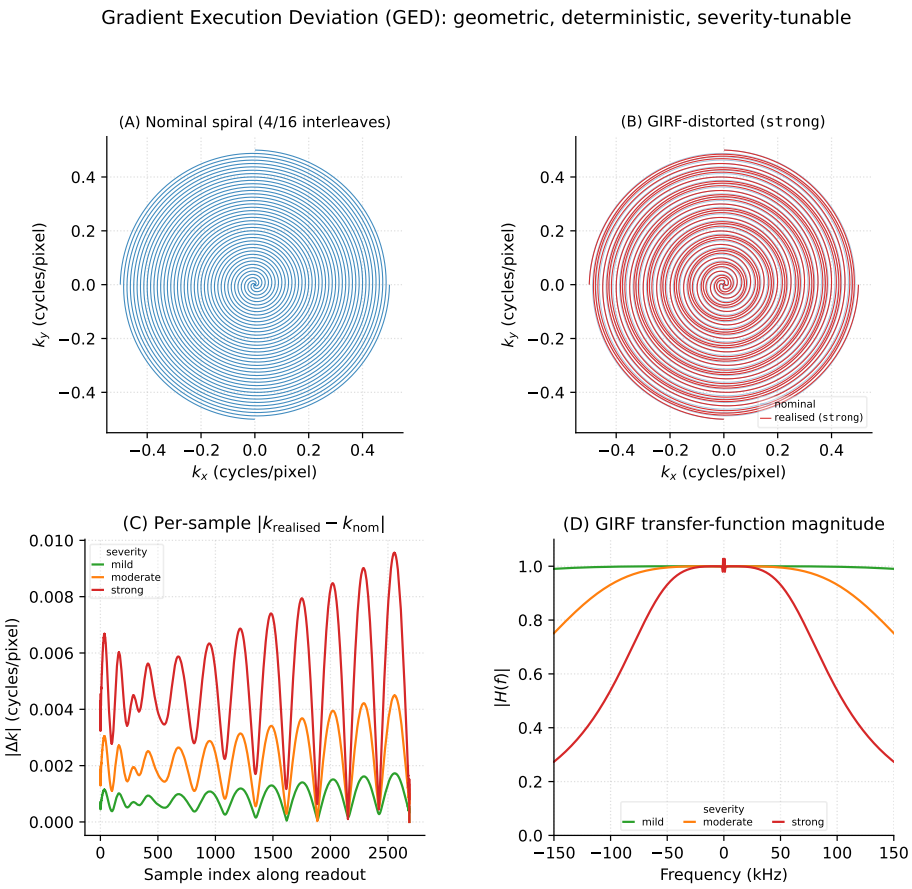

Figure 1: How gradient execution changes sampling locations. **(A)** Planned spiral path through spatial-frequency space (k-space) (4 of 16 interleaves). **(B)** Executed trajectory after passing the commanded gradient waveform through a `strong`-severity GIRF (red), overlaid on the nominal trajectory (blue). **(C)** Per-sample deviation $|k_{\text{realized}} - k_{\text{nominal}}|$ along the readout for three severity levels. **(D)** GIRF magnitude $|H(f)|$ used in panel (B); Appendix F.1 compares the presets with two scanner-measured GIRFs.

the highlighted SwinMR–E2E-VarNet reversal recurs at strong mismatch on two external datasets. Third, we show that the ordering is conditional on the complete reconstruction pipeline rather than on architecture alone. Supplying the executed trajectory does not reliably restore fixed learned models. Replacing the density weights removes the mild reversal at fixed checkpoints, brief nominal adaptation restores a smaller one (Figure 3), and training and selecting models for mismatch improves their robustness at a nominal cost. Together, these findings motivate comparing complete pipelines across acquisition conditions.

2 Related Work

MRI reconstruction benchmarks and robustness. fastMRI (Zbontar et al., 2018; Knoll et al., 2020) established a widely used setting for learned MRI reconstruction: comparisons on Cartesian (regular-grid) knee and brain k-space, with the 2020 challenge providing cross-team rankings (Muckley et al., 2021). Later datasets expanded anatomical, vendor, and field-strength coverage (Souza et al., 2018; Desai et al., 2021; Wang et al., 2023), and raw-data archives provide further acquisition data (Ong et al., 2018; Stanford Radiology Lab, 2014). Learned reconstruction has been stress-tested under adversarial perturbations (Antun et al., 2020), natural shifts in anatomy, scanner, or contrast (Darestani et al., 2021; 2022), clinically motivated perturbations (Cheng et al., 2020; Morshuis et al., 2022), and posterior-sampler robustness (Jalal et al., 2021). These studies establish the need to test robustness

beyond nominal data. We focus on whether relative method performance survives a structured error in sampling geometry, using a fixed waveform and GIRF to make that error reproducible.

Gradient hardware imperfections. Vannesjo et al. (2013) measured the gradient impulse response with a dynamic field camera, establishing the linear time-invariant (LTI) response model used here, and showed that incorporating it into the reconstruction operator restores image quality in spiral and EPI acquisitions (Vannesjo et al., 2016). Related work characterizes non-Cartesian trajectory errors (Brodsky et al., 2009), spiral eddy-current effects (Robison et al., 2019), calibration (Campbell-Washburn et al., 2016), concurrent field monitoring (Barmet et al., 2008), and thermal drift (Stich et al., 2020). This work establishes the physical source of trajectory error and the value of calibration. EXEC-MRI builds on that foundation to test a different outcome: the stability of reconstruction rankings, including how fixed learned pipelines respond when supplied with the executed trajectory.

Evaluation methodology. Our design draws on variance budgeting (Bouthillier et al., 2021), ranking instability under small distribution shifts (Recht et al., 2019), standardized robustness leaderboards (Croce et al., 2021), severity-indexed rank stability (Dünkel et al., 2025), documentation standards (Gebru et al., 2021; Mitchell et al., 2019; Paullada et al., 2021), and non-parametric statistics (Demšar, 2006; Efron & Tibshirani, 1986). EXEC-MRI applies these principles to MRI with controlled errors in sampling coordinates. Its contribution is an evaluation protocol and evidence about method comparisons: ranking changes, recovery when the executed trajectory is supplied, and dependence on numerical and training choices. The hardware-response model itself is established prior work.

3 The EXEC-MRI evaluation protocol

The central control is simple: separate the sampling coordinates used to generate measurements from those supplied to reconstruction. A *forward operator* maps a candidate image to predicted measurements; using the wrong coordinates makes this operator disagree with acquisition. We first define that disagreement, then the datasets, compared methods, and paired evaluation conditions.

3.1 Simulating sampling-coordinate errors

We model each in-plane gradient channel by a linear time-invariant filter (Vannesjo et al., 2013; 2016). With commanded gradient g_{cmd}, impulse response h, and $\bar{\gamma} = 42.576$ MHz/T (cycles per second per tesla),

$$g_{\mathrm{real}} = h * g_{\mathrm{cmd}}, \qquad \tilde{k}(t) = \bar{\gamma} \int_0^t g_{\mathrm{real}}(\tau)\, \mathrm{d}\tau.$$

Here $*$ denotes convolution. Integrating the filtered gradient gives the raw trajectory $\tilde{k}$; rescaling it to the nominal maximum radius gives the coordinates k_{real} used for simulation (Appendix B.3). For trajectory k, the forward operator is $E_k = \mathcal{F}_k \mathcal{C}$. The sensitivity maps $\mathcal{C}$ describe each receiver coil's spatial response; $\mathcal{F}_k$ evaluates the Fourier transform at coordinates k, implemented with a non-uniform fast Fourier transform (NUFFT). The simulator forms measurements $y = E_{k_{\mathrm{real}}} x$ from reference image x, without additional measurement noise. A reconstruction unaware of GED uses $E_{k_{\mathrm{nom}}}$ instead.

The primary trajectory comprises 16 spiral arms (interleaves), with approximately $2.38\times$ undersampling by nominal sample count. The response model combines low-pass filtering, delay, and resonance (Appendix B). Six presets, minimal through extreme, produce mean coordinate deviations of 0.08–5.87 % of the maximum sampling radius $k_{\max}$; nominal uses the identity response. The sweep tests path-shape changes after removing global radial scaling. Passing the same waveform through two publicly released measured GIRFs (Malik, 2015) gives 3.86 % for a 3 T system and 10.92 % for a 7 T system. These provide scale references, not measurements on our datasets or estimates of residual clinical error.

3.2 Primary and external datasets

The primary dataset is fastMRI Brain multi-coil (Zbontar et al., 2018; Knoll et al., 2020). Learned methods use 2,000 training slices and 800 validation slices, with no shared volume identifiers across

Table 1: Methods in the primary comparison. Learned architecture names denote our non-Cartesian implementations. DC indicates explicit measurement-consistency updates at inference.

Method	Type	Parameters	DC
CG-SENSE	classical	–	yes
L1-SENSE	classical	–	yes
U-Net	image-domain	7.76 M	no
SwinMR	image-domain	0.90 M	no
MoDL	unrolled	0.37 M	yes
E2E-VarNet	unrolled	3.65 M	yes
PromptMR	unrolled	4.27 M	yes
MoDL-Aug	augmented MoDL	0.37 M	yes

training, validation, and test. Evaluation uses 300 held-out slices from 52 volumes under all 13 conditions specified below. Images are cropped or padded to 320×320. Two external datasets use 200 slices from 20 volumes each without fine-tuning: fastMRI Knee and the 12-channel GE subset of CC359 (Knoll et al., 2020; Souza et al., 2018). Anatomy, contrast, vendor, and coil configuration can change together; these are combined dataset shifts. Appendix A documents reference-image construction and the exact subsets.

3.3 Methods and paired evaluation conditions

Table 1 compares eight methods. Classical solvers iteratively fit an image to the measurements. Image-domain networks map an initial reconstruction to an output image. Unrolled networks alternate learned image updates with *data consistency* (DC), which uses the forward operator to check agreement with measurements. MoDL-Aug adds GED-aware training; augmented E2E-VarNet and PromptMR are separate interventions. Learned baselines are compact non-Cartesian implementations, with adaptations and budgets in Appendix C. For example, our VarNet updates images with NUFFT residuals and supplied sensitivity maps, unlike the original k-space cascade and learned sensitivity estimator.

External rankings use seven common methods, excluding L1-SENSE because calibrated Knee measurements were incomplete. A score-based candidate is excluded after implementation audit (Appendix F.7). No estimated replacement scores enter any reported comparison.

The 13 conditions are one nominal condition and six mismatch/oracle pairs. *Nominal* uses the planned trajectory for both simulation and reconstruction. *Mismatch* generates measurements on the executed trajectory but reconstructs with the planned trajectory. The paired *oracle* condition uses the same measurements and supplies the executed trajectory to reconstruction, keeping learned weights fixed. It also recomputes the density-compensation function (DCF) and dependent inputs or initialization. These numerical weights account for uneven sampling and affect initial images and weighted consistency updates. Thus oracle tests the benefit of trajectory knowledge for a complete pipeline, not a coordinate-only intervention. The primary comparison uses frozen historical density weights; Section 4.4 tests standard interpolation-based weights with and without model adaptation.

For method M and severity s, we report two PSNR differences:

$$G_{M,s} = \mathrm{PSNR}_{M,\mathrm{oracle}(s)} - \mathrm{PSNR}_{M,\mathrm{mismatch}(s)}, \quad L_{M,s} = \mathrm{PSNR}_{M,\mathrm{nominal}} - \mathrm{PSNR}_{M,\mathrm{oracle}(s)}. \tag{1}$$

G measures improvement after trajectory substitution; L measures the remaining gap to nominal performance. Their sum is the total nominal-to-mismatch loss, but neither must be positive. A changed operator or input can reduce a fixed model's score. We therefore report absolute scores and signed differences, without treating oracle as an upper bound or the two differences as causal shares of the error.

3.4 Metrics and statistical protocol

Our main fidelity metric is peak signal-to-noise ratio (PSNR), measured in decibels (dB); higher values indicate lower pixel error relative to the reference image. We also report structural similarity (SSIM) (Wang et al., 2004), normalized root-mean-square error (NRMSE), perceptual distance

(LPIPS) (Zhang et al., 2018), and high-frequency error norm (HFEN). Higher SSIM and lower values of the latter three indicate better agreement. All metrics use per-slice least-squares amplitude alignment to the reference, so they assess images after correcting a global intensity scale.

Scores average the evaluated slices. To account for dependence among slices from one volume, paired 95% intervals use 20,000 bootstrap samples of whole volumes, retaining all constituent slices. Differences pair identical volume/slice keys; intervals are the 2.5th and 97.5th percentiles. They capture test-volume variation conditional on fixed checkpoints, not training or hardware variability. Primary and external comparisons use seed-0 checkpoints; separately labeled training interventions use their own checkpoints. No interval pools training seeds.

Mean-score ranks and their changes are descriptive. An interval excluding zero supports the highlighted directional contrast; rank-change counts are not counts of significant discoveries. Kendall τ measures rank agreement on a common method set. Appendices C and D document selection, normalization, and measurement provenance.

4 Results

We first ask whether method order changes, then what trajectory knowledge recovers. External datasets and other checks test how broadly the pattern appears; numerical and training interventions identify conditions that change it. Unless specified otherwise, results use 300 Brain slices from 52 held-out volumes, fixed seed-0 checkpoints, and paired volume-bootstrap intervals.

4.1 A nominal advantage can reverse under mild mismatch

SwinMR and E2E-VarNet illustrate the ranking question (Figure 2). At nominal, E2E-VarNet leads by 11.42 dB. At mild GED, SwinMR leads by 2.32 dB ($95\,\%$ CI $[2.01, 2.63]$), with absolute PSNRs of 27.69 and 25.37 dB (Table 2). Both methods lose fidelity, but at different rates: the nominal advantage no longer predicts their relative order. At strong, SwinMR retains a 0.67 dB lead ($[0.61, 0.74]$) and higher SSIM (0.403 vs. 0.225). Both image-domain models overtake both E2E-VarNet and PromptMR from mild onward; all corresponding paired PSNR intervals exclude zero (released `primary_pairwise.csv`).

Across eight methods, $22/28$ pairs ($79\,\%$) change mean-score order from nominal to moderate. This includes small gaps and does not mean 22 independently significant effects. We therefore accompany ranks with absolute fidelity and paired differences. Mild and moderate mean coordinate deviations are 0.17 and $0.46\,\%$ of $k_{\max}$; the labels describe this simulated waveform, not clinical severity. Appendix E shows nominal error maps and strong-mismatch images. Neither a higher rank nor a positive difference establishes acceptable reconstruction quality.

4.2 Trajectory knowledge does not guarantee recovery

Does the reversal occur because the measurements no longer support accurate reconstruction? The oracle comparison provides a partial answer by supplying the executed trajectory and dependent numerical components, without changing measurements or learned weights (Figure 5). At strong, CG-SENSE improves from 16.18 to 41.51 dB, recovering 25.33 dB while remaining 3.12 dB below its nominal score. The same measurements thus support substantially better recovery than its mismatched result.

Fixed learned models benefit less consistently. At strong, substitution improves U-Net by 1.34 dB, SwinMR by 0.64 dB, and MoDL by 6.77 dB, but slightly worsens E2E-VarNet and PromptMR (by 0.46 and 0.12 dB). Supplying the executed trajectory is therefore insufficient to restore these fixed learned pipelines to nominal performance. Changed inputs or operator scaling may contribute, but the intervention changes several components together. It cannot identify the refiner, learned step sizes, initialization, or density compensation as the cause.

4.3 Checks beyond the primary Brain comparison

Transfer to external datasets. The primary test holds dataset fixed. Without fine-tuning, strong-mismatch reversals also appear on Knee and CC359 (200 slices from 20 volumes each). SwinMR

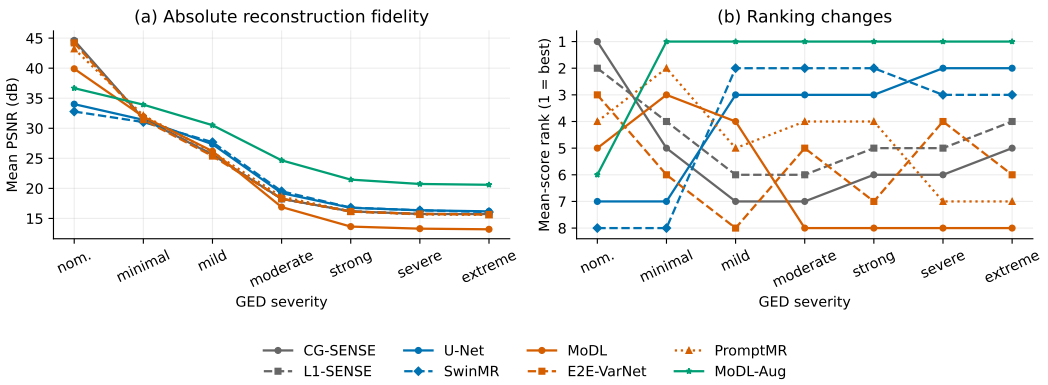

Figure 2: Eight-method Brain audit: absolute PSNR and mean-score rank across nominal and mismatch severities (300 slices, 52 volumes). The SwinMR–E2E-VarNet reversal begins between minimal and mild. Rank changes alone do not establish practical or statistical significance.

Table 2: Absolute PSNR (dB) and SSIM on fastMRI Brain (300 slices from 52 volumes). Mild and strong use the mismatched operator. Volume-bootstrap intervals and all metrics are released with the artifact.

Method	PSNR nom.	PSNR mild	PSNR strong	SSIM nom.	SSIM strong
CG-SENSE	44.63	25.60	16.18	0.990	0.289
L1-SENSE	44.41	25.63	16.18	0.989	0.288
U-Net	34.01	27.35	16.77	0.945	0.387
SwinMR	32.78	27.69	16.80	0.941	0.403
MoDL	39.90	26.19	13.64	0.977	0.418
E2E-VarNet	44.20	25.37	16.13	0.988	0.225
PromptMR	43.21	26.04	16.18	0.985	0.232
MoDL-Aug	36.67	30.52	21.46	0.962	0.641

exceeds E2E-VarNet by 0.94 dB on Knee ($95\,\%$ CI $[0.72, 1.18]$) and 1.04 dB on CC359 ($[0.99, 1.08]$). At strong, rankings of the seven common methods agree with Brain at Kendall $\tau = 0.810$ and 0.905, respectively (Table 3). L1-SENSE is excluded from this common set because Knee measurements were incomplete. The mean reversal starts at mild on both datasets, but the CC359 mild interval crosses zero ($+0.23$ dB, $[-0.08, 0.52]$). The evidence thus supports recurrence at strong mismatch, without establishing a shared mild threshold.

Metrics, trajectory, and training budget. The strong contrast persists under foreground-masked PSNR, SSIM, LPIPS, and a nonlinear gradient-saturation check (Table 4; Appendix F). HFEN, which emphasizes fine-detail errors, is an exception: its mean favors E2E-VarNet slightly and the paired interval includes zero. A second sampling path also gives a reversal: a radial-trained U-Net trails a radial-trained E2E-VarNet by 7.7 dB at nominal but leads by 0.67 dB at strong ($[0.46, 0.93]$). Severity labels do not imply equal coordinate deviations across spiral and radial paths. Nonlinear saturation retains the SwinMR reversal at moderate and strong but changes that pair's order again at severe.

A separate $5{,}002$-slice training configuration reproduces the mismatch reversal. Its regenerated preprocessing and sensitivity maps prevent attributing differences from the primary setting to training-set size alone. These checks extend the observed pattern across several settings; they do not establish invariance across metrics, perturbations, or training choices.

4.4 Numerical weights and adaptation change the comparison

Density compensation affects the images and updates seen by a learned model. We therefore test whether the reversal depends on these numerical weights, using one preselected slice from each of 52 Brain volumes. Figure 3 compares the historical protocol, a weight replacement at fixed checkpoints, and adaptation to the new weights.

Table 3: External strong-severity results (200 slices from 20 volumes each). Kendall τ compares the same seven methods with Brain. The PSNR difference is SwinMR minus E2E-VarNet; intervals resample volumes.

Dataset	τ vs. Brain	Difference (dB)	95% CI (dB)
fastMRI Knee	$+0.810$	$+0.94$	$[0.72, 1.18]$
CC359	$+0.905$	$+1.04$	$[0.99, 1.08]$

Table 4: Image-domain model versus E2E-VarNet at strong severity. The last column is a volume-bootstrap 95% interval for image-domain minus VarNet. PSNR units are dB; the other metrics are unitless. The 5k setting also changes preprocessing.

Variation	Image model	Image	VarNet	95% CI of difference
Main PSNR (dB)	SwinMR	16.80	16.13	$[+0.610, +0.741]$
SSIM	SwinMR	0.403	0.225	$[+0.167, +0.190]$
LPIPS (lower better)	SwinMR	0.421	0.576	$[-0.167, -0.142]$
HFEN (lower better)	SwinMR	1.267	1.261	$[-0.009, +0.024]$
5k setting PSNR (dB)	SwinMR	16.57	15.87	$[+0.599, +0.805]$
Foreground PSNR (dB)	SwinMR	13.58	13.23	$[+0.279, +0.430]$
Saturation PSNR (dB)	SwinMR	16.22	15.71	$[+0.451, +0.586]$
Radial PSNR (dB)	U-Net	24.85	24.18	$[+0.457, +0.930]$

Replacing historical weights with standard interpolation-based density compensation, matched in mean scale, changes the mild SwinMR–VarNet contrast from $+2.57$ to -1.80 dB. The strong contrast shrinks from $+0.67$ to $+0.06$ dB (95% CI $[0.009, 0.116]$). This is a negative result for the stability of the mild reversal. Nominal SwinMR also loses 5.74 dB, indicating a substantial training–inference input shift.

To test adaptation to that shift, both models receive six nominal-training epochs under standard weights, selected only by nominal validation PSNR. SwinMR then trails VarNet by 2.38 dB at nominal but leads by 0.40 dB at mild ($[0.337, 0.465]$) and 0.37 dB at strong ($[0.328, 0.404]$). The adapted differences are positive on $51/52$ selected slices at mild and $52/52$ at strong; their medians are 0.37 and 0.34 dB. Only $1/52$ favors SwinMR at mild with standard weights and fixed models. Thus the smaller adapted reversal is spread across the selected volumes, but its magnitude and even its presence depend on numerical and training choices. This bounded warm-start check adds optimization steps; it is neither a from-scratch comparison nor an isolation of the density algorithm's effect (Appendix F.6). The dependence is itself part of the evaluation problem: robustness conclusions attached only to architecture names can change when numerically consequential pipeline components change.

4.5 Training for mismatch changes the accuracy trade-off

The previous adaptation uses nominal acquisition only. We next expose MoDL, E2E-VarNet, and PromptMR to GED during training, using the same trajectory-sampling protocol for each. These augmented models are also selected by strong-mismatch validation PSNR, whereas their plain counterparts use nominal validation. The comparison therefore tests training exposure and checkpoint selection together (Appendix C.5).

All three augmented versions outperform their plain counterparts and SwinMR at each tested mismatch severity, at a 2.5–3.5 dB nominal cost (Table 5). Paired intervals are released for these comparisons. Their strong scores of 21.46, 24.62, and 24.34 dB remain below CG-SENSE supplied with the executed trajectory (41.51 dB). This compares different available information: the augmented models still use the nominal trajectory at inference. It is an empirical reference gap, not an upper bound on blind reconstruction. Training for mismatch and supplying trajectory knowledge answer distinct questions; both conditions belong in the interpretation of a method comparison.

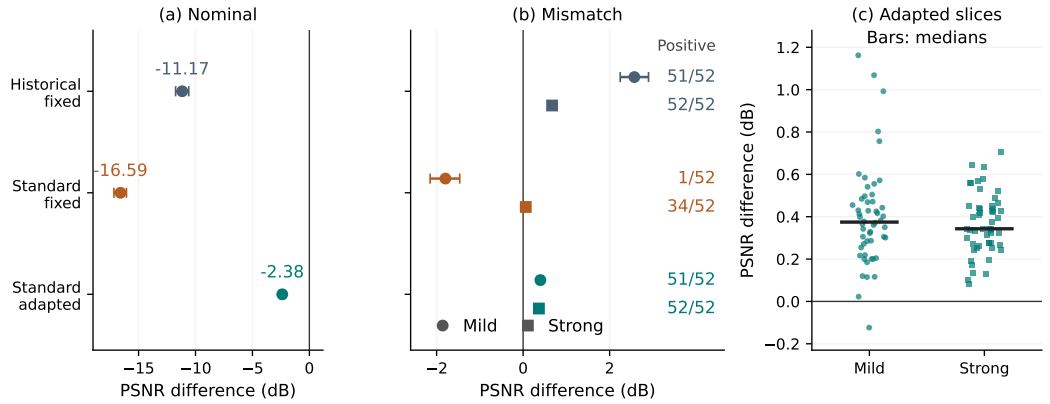

Figure 3: Numerical and training conditions change the comparison. All panels use the same one preselected slice from each of 52 Brain volumes. (a,b) SwinMR minus E2E-VarNet PSNR; positive favors SwinMR. Intervals resample volumes (95%). Counts in (b) are positive slice contrasts. (c) Every adapted-model contrast, with median bars. Fixed models use original checkpoints; adapted models receive six nominal-training epochs under standard density weights.

Table 5: Plain / GIRF-augmented PSNR (dB), fastMRI Brain, 300 slices. Bold marks the higher score in each pair. Augmented checkpoints use strong-mismatch validation selection; test mismatch columns retain nominal reconstruction operators.

Architecture	Nominal	Mild	Moderate	Strong
MoDL	**39.90** / 36.67	26.19 / **30.52**	16.89 / **24.66**	13.64 / **21.46**
E2E-VarNet	**44.20** / 40.75	25.37 / **34.78**	18.29 / **28.99**	16.13 / **24.62**
PromptMR	**43.21** / 40.70	26.04 / **34.90**	18.62 / **29.12**	16.18 / **24.34**

5 Discussion

What a nominal ranking establishes. A method's nominal advantage can disappear or reverse when actual sampling coordinates differ from those supplied to reconstruction, even on the same test dataset. Nominal fidelity alone therefore does not establish relative performance under acquisition-model error. This evidence comes from retrospective simulation with specified implementations; it does not measure how often reversals occur on deployed scanners or what they mean clinically.

Why trajectory knowledge and training must be distinguished. The large CG-SENSE recovery confirms that the strong-mismatch measurements retain useful information, consistent with the established value of calibration (Vannesjo et al., 2016; Campbell-Washburn et al., 2016; Barmet et al., 2008). The smaller or negative gains for several learned models show that supplying the trajectory need not restore a fixed pipeline. Training and selecting for mismatch improves the tested unrolled models while the executed trajectory remains unavailable at inference, at a cost in nominal fidelity.

The complete pipeline determines the comparison. Standard density weights remove the mild reversal at fixed checkpoints; nominal adaptation restores a smaller one. The primary gap is therefore a property of the evaluated pipeline rather than of the architecture alone: numerical weights, initialization, scaling, training coverage, and checkpoint selection all help define the evaluated method, and a ranking reported under an architecture name is incomplete until they are specified. EXEC-MRI accordingly pairs ranking comparisons with tests of trajectory knowledge and of compatibility between training and inference.

Implications for evaluation. Report absolute scores and paired differences across nominal and mismatch conditions, together with a trajectory-informed reference that specifies every changed component. Separate descriptive rank changes from uncertainty in individual comparisons, and resample volumes for multi-slice data. Releasing numerical arrays and training conventions makes

these conditions inspectable. The same model-selection question arises in other inverse problems whose acquisition models are imperfect.

6 LIMITATIONS

Retrospective simulation and hardware coverage. All reconstruction results use measurements simulated from reference images derived from Cartesian data. The main sweep uses a parametric LTI GIRF with one deterministic response per severity; a separate check adds nonlinear saturation but is not a complete nonlinear hardware model. Two measured GIRFs provide scale references, not prospective acquisitions or a distribution of residual scanner errors. Simulation and reconstruction also share discretization and sensitivity maps, so changing the trajectory removes their exact coordinate agreement but not every favorable simulation assumption.

Implementations, training, and normalization. We study compact non-Cartesian adaptations of named architectures, trained on 2,000 slices; conclusions concern these implementations, not every model in their families. Main results use one training seed, and volume-bootstrap intervals describe test-set variation conditional on those checkpoints. Some methods use ground-truth-derived amplitude scaling at inference; this controlled convention is not deployable without an input-derived estimator (Appendix D).

Numerical and adaptation dependence. The historical density-weight iteration is numerically sensitive. We release its exact arrays; changing CPU reduction order when regenerating weights can materially alter nominal scores (Appendix B). The primary findings therefore depend on these numerical conventions as well as on the trained models. The six-epoch nominal adaptation is a bounded warm-start check; it does not establish convergence from scratch or invariance across training seeds.

Trajectory, dataset, and endpoint coverage. The main readout is spiral, with a four-method radial check. External evaluation covers two datasets and seven common methods, without separating every anatomy, contrast, vendor, and coil effect. Image-quality metrics are not clinical endpoints; reader studies or downstream tasks would be needed to assess diagnostic consequences. Additional scope details appear in Appendix G.

REPRODUCIBILITY STATEMENT

The anonymized supplement provides simulation and reconstruction code, training drivers for all reported models, external-data adapters, split manifests, and the scripts for the checks in Appendix F. It also releases measured per-slice scores, a source manifest with hashes, and frozen trajectory and density-weight arrays. Running `scripts/regenerate_submission.py` reproduces the quantitative tables and figures from those measurements without retraining. Raw scanner data and trained weight files are not included; image-level reruns require separate data access and checkpoints. The GIRF conventions, training budgets, and result provenance are detailed in Appendices B, C, and D. Primary training used one RTX 3080; the separate larger-training configuration is documented in Appendix F.2.

ETHICS STATEMENT

All experiments use publicly available, de-identified MRI datasets (fastMRI brain and knee, Calgary-Campinas CC359) under their original data-use agreements; no new human-subject data were collected. The work evaluates the robustness of existing reconstruction methods and does not propose a clinical tool.

AI USE STATEMENT

In this work, we used generative AI assistants for literature search, research planning and execution, manuscript drafting, and language editing. Research assistance included feedback on experimental design, software implementation, analysis of results, and review of technical arguments. Verification

included reference checking, software testing, and comparison of the text against experimental outputs. AI-generated suggestions were not treated as experimental measurements: all reported quantitative comparisons use the measured result files listed in the released manifest. The authors are responsible for the final paper and all reported results.

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.

Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

Christoph Barmet, Nicola De Zanche, and Klaas P Pruessmann. Spatiotemporal magnetic field monitoring for MR. *Magnetic Resonance in Medicine*, 60(1):187–197, 2008.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Naz Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Christopher Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, 2021.

Ethan K Brodsky, Alexey A Samsonov, and Walter F Block. Characterizing and correcting gradient errors in non-Cartesian imaging: Are gradient errors linear time-invariant (LTI)? *Magnetic Resonance in Medicine*, 62(6):1466–1476, 2009.

Adrienne E Campbell-Washburn, Hui Xue, Robert J Lederman, Anthony Z Faranesh, and Michael S Hansen. Real-time distortion correction of spiral and echo planar images using the gradient system impulse response function. *Magnetic Resonance in Medicine*, 75(6):2278–2285, 2016.

Kaiyang Cheng, Francesco Calivá, Rutwik Shah, Misung Han, Sharmila Majumdar, and Valentina Pedoia. Addressing the false negative problem of deep learning MRI reconstruction models by adversarial attacks and robust training. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121, pp. 121–135. PMLR, 2020. URL https://proceedings.mlr.press/v121/cheng20a.html.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021. Leaderboard concept and external evaluation invitation.

Mohammad Zalbagi Darestani, Akshay S Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *International Conference on Machine Learning (ICML)*, 2021.

Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *International Conference on Machine Learning (ICML)*, 2022.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Arjun D Desai, Andrew M Schmidt, Elka B Rubin, Christopher M Sandino, Marianne S Black, Valentina Mazzoli, Kathryn J Stevens, Robert Boutin, Christopher Ré, Garry E Gold, et al. SKM-TEA: A dataset for accelerated MRI reconstruction with dense image labels for quantitative clinical evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Olaf Dünkel, Artur Jesslen, Jiahao Xie, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. CNS-Bench: Benchmarking image classifier robustness under continuous nuisance shifts. In *International Conference on Computer Vision (ICCV)*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Dunkel_CNS-Bench_Benchmarking_Image_Classifier_Robustness_Under_Continuous_Nuisance_Shifts_ICCV_2025_paper.pdf.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

Jiahao Huang, Yingying Fang, Yinzhe Wu, Huanjun Wu, Zhifan Gao, Yang Li, Javier Del Ser, Jun Xia, and Guang Yang. Swin transformer for fast MRI. *Neurocomputing*, 493:281–304, 2022.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon I Tamir. Robust compressed sensing MRI with deep generative priors. In *Advances in Neural Information Processing Systems*, 2021.

Florian Knoll, Jure Zbontar, Anuroop Sriram, Matthew J Muckley, Mary Bruno, Aaron Defazio, Marc Parente, Krzysztof J Geras, Joe Katsnelson, Hersh Chandarana, et al. fastMRI: A publicly available raw k-space and DICOM dataset of knee images for accelerated MR image reconstruction using machine learning. *Radiology: Artificial Intelligence*, 2(1):e190007, 2020.

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

Shaihan J. Malik. reVERSE-GIRF: VERSE guided RF pulse design with modification for imperfect gradient performance, including measured 3T and 7T GIRF data. Software release, https://github.com/mriphysics/reVERSE-GIRF, 2015.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019.

Jan Nikolas Morshuis, Sergios Gatidis, Matthias Hein, and Christian F Baumgartner. Adversarial robustness of MR image reconstruction under realistic perturbations. *Lecture Notes in Computer Science (MICCAI Workshop)*, 2022.

Matthew J Muckley, Bruno Riemenschneider, Alireza Radmanesh, Sunwoo Kim, Geunu Jeong, Jingyu Ko, Yohan Jun, Hyungseob Shin, Dosik Hwang, Mahmoud Mostapha, et al. Results of the 2020 fastMRI challenge for machine learning MR image reconstruction. *IEEE Transactions on Medical Imaging*, 40(9):2306–2317, 2021.

Frank Ong, Sara Amin, Shreyas Vasanawala, and Michael Lustig. mridata.org: An open archive for sharing MRI raw data. *Proceedings of the ISMRM*, pp. 3425, 2018.

Amandalynne Paullada, Inioluwa Deborah Raji, Emily M Bender, Emily Denton, and Alex Hanna. Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021.

Klaas P Pruessmann, Markus Weiger, Peter Börnert, and Peter Boesiger. Advances in sensitivity encoding with arbitrary k-space trajectories. *Magnetic Resonance in Medicine*, 46(4):638–651, 2001.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, 2019.

Ryan K Robison, Zhiqiang Li, Dinghui Wang, Melvyn B Ooi, and James G Pipe. Correction of B0 eddy current effects in spiral MRI. *Magnetic Resonance in Medicine*, 81(4):2501–2513, 2019. doi: 10.1002/mrm.27583.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

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 *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 64–73, 2020.

Stanford Radiology Lab. Stanford fully-sampled 3D FSE knees dataset. http://mridata.org/list?project=Stanford%20Fullysampled%203D%20FSE%20Knees, 2014.

Manuel Stich, Christiane Pfaff, Tobias Wech, Anne Slawig, Gudrun Ruyters, Andrew Dewdney, Ralf Ringler, and Herbert Köstler. The temperature dependence of gradient system response characteristics. *Magnetic Resonance in Medicine*, 83(4):1519–1527, 2020.

S Johanna Vannesjo, Maximilian Haeberlin, Lars Kasper, Matteo Pavan, Bertram J Wilm, Christoph Barmet, and Klaas P Pruessmann. Gradient system characterization by impulse response measurements with a dynamic field camera. *Magnetic Resonance in Medicine*, 69(2):583–593, 2013.

S Johanna Vannesjo, Nadine N Graedel, Lars Kasper, Simon Gross, Julia Busch, Maximilian Haeberlin, Christoph Barmet, and Klaas P Pruessmann. Image reconstruction using a gradient impulse response model for trajectory prediction. *Magnetic Resonance in Medicine*, 76(1):45–58, 2016.

Chengyan Wang, Jun Lyu, Shuo Wang, Chen Qin, Kunyuan Guo, Xinyu Zhang, Xiaotong Yu, Yan Li, Fanwen Wang, Jianhua Jin, et al. CMRxRecon: An open cardiac MRI dataset for the competition of accelerated image reconstruction. *arXiv preprint arXiv:2309.10836*, 2023.

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.

Bingyu Xin, Meng Ye, Leon Axel, and Dimitris N Metaxas. Fill the k-space and refine the image: Prompting for dynamic and multi-contrast mri reconstruction. *arXiv preprint arXiv:2309.13839*, 2023.

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.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A Datasheets for the three datasets

This appendix explains how raw datasets become reference images, sensitivity maps, and evaluation subsets, following Gebru et al. (2021). Dataset-level descriptions should not be read as balanced sampling across scanners, contrasts, or clinical populations. All raw images remain subject to the upstream access agreements and are not redistributed in the submission package.

A.1 fastMRI Brain: primary training and test set

Source and retained population. We use multi-coil fastMRI Brain data (Zbontar et al., 2018; Knoll et al., 2020). The retained data pool contains 10,442 slices from 1,790 volume identifiers, with recorded coil counts ranging from 2 to 28. These are counts of the local retained pool, not the full upstream dataset. The corpus contains multiple brain contrasts; no contrast-balanced sampling claim is made. The local metadata does not establish a field-strength breakdown for the evaluated subset.

Volume-level partition. The released `manifests/splits_seed0.json` specifies the exact ordered split: 5,002 training slices from 860 volumes, 800 validation slices from 136 volumes, and 802 test slices from 136 volumes. Primary training uses the first 2,000 training entries (344 volumes), and evaluation the first 300 test entries (52 volumes). Volume identifiers do not overlap across these three splits. This check does not establish patient-level independence when multiple volume identifiers might refer to one patient. The full manifest and its SHA-256 digest are provided; subset order matters for reproducing the caps.

Preprocessing. The preprocessing driver removes the first and last five slices of each volume, then retains slices whose reference-image maximum is at least 5% of the maximum among the candidate slices *within that volume*. The reference is the stored root-sum-of-squares (RSS) image, center-cropped to 320×320; incompatible or empty references are skipped. Sensitivity maps are estimated using the `direct` option: coil images from the central 24×24 calibration region are normalized by their RSS. This is not ESPIRiT eigenvector calibration. The same retained maps are used for simulation and reconstruction. Non-Cartesian measurements are synthesized from these references and maps. Initial network inputs are formed by mapping the measurements back to coil images with the adjoint operator and combining their magnitudes by RSS. No additional measurement noise is added.

Limitations. The slice cap and volume ordering constrain the evaluated population. Magnitude RSS references discard the original object's complex phase. Shared sensitivity maps and retrospective simulation remove sources of error present in prospective scanning; these results do not establish clinical reconstruction quality or diagnostic accuracy.

A.2 fastMRI Knee: external anatomy and acquisition shift

We evaluate on multi-coil fastMRI Knee (Zbontar et al., 2018; Knoll et al., 2020) without retraining or fine-tuning. The reported result files contain 200 slices, ten from each of 20 volume identifiers, and five acquisition conditions. The external preprocessing driver drops the end slices and applies the within-volume signal rule, then selects evenly spaced retained slice indices, with a cap of ten per volume. Ground truth comes from stored RSS images; direct sensitivity estimation and the 320×320 simulation matrix follow the Brain pipeline. The adapter accepts both proton-density and fat-suppressed knee acquisitions. The exact evaluated `volume_id`/`slice_id` keys are released with the measured results.

Anatomy, contrast, coil configuration, and acquisition can change together. This evaluation tests transfer to this retained knee subset; it does not isolate the effect of anatomy or establish performance on all knee scans.

A.3 Calgary–Campinas: external brain acquisition shift

The Calgary–Campinas evaluation uses the 12-channel multi-coil brain data (Souza et al., 2018), with 200 evaluated slices from 20 volume identifiers and no retraining. The source is the T1-weighted GE cohort. The same ten-slices-per-volume budget and five conditions are used as for Knee. Vendor, contrast, coil configuration, and native resolution can all differ from the primary corpus, so this is not a vendor-only intervention.

Adapter and reference construction. The released `external_eval/adapters.py` converts the interleaved real/imaginary coil dimension into complex multi-coil arrays. The upstream release already contains hybrid (x, k_y, k_z) data: the scanner has applied an inverse Fourier transform along the readout direction. Our adapter selects spatial x slices and applies a 2D inverse transform over the

remaining encoded directions. The source includes partially acquired, zero-filled slice-encoding data; we do not add a separate partial-Fourier reconstruction step. The upstream format is documented at https://www.ccdataset.com/mr-reconstruction-challenge. Images are center-padded/cropped to 320×320 in the image domain, with corresponding Fourier conversion for the coil data. Unlike fastMRI, this source does not supply the same stored RSS reference; the adapter constructs RSS from coil images. This difference in reference construction is another source of distribution shift.

Access and scope. Raw fastMRI and Calgary–Campinas data must be obtained through their respective upstream distribution channels. The submission provides adapters and measured result keys, not raw scans. The retained external populations each cover 20 volumes, and their bootstrap intervals should be interpreted with that sample size and the fixed checkpoint choice.

B GIRF MODEL AND NUMERICAL CONVENTIONS

We specify both the hardware-response model and the numerical steps that turn it into sampling coordinates. These details define the tested perturbation: filtering is circular, trajectory radius is normalized, and the historical density weights are frozen for reproducibility.

B.1 TRANSFER FUNCTION

The same scalar response is applied separately to the two in-plane gradient channels. This is a controlled simplification; it does not represent channel coupling or scanner-specific axis asymmetry. The frequency response $H(f)$ describes how each temporal frequency is changed by low-pass filtering, delay, and resonance:

$$H(f) = H_{\mathrm{lpf}}(f) H_{\mathrm{delay}}(f) H_{\mathrm{res}}(f), \tag{2}$$

$$H_{\mathrm{lpf}}(f) = [1 + \sqrt{2} j (f/f_c) + (jf/f_c)^2]^{-1}, \tag{3}$$

$$H_{\mathrm{delay}}(f) = \exp(-j 2\pi f \tau_g), \tag{4}$$

$$H_{\mathrm{res}}(f) = 1 + \frac{A_{\mathrm{res}}}{1 - (f/f_{\mathrm{res}})^2 + 2j\zeta(f/f_{\mathrm{res}})}. \tag{5}$$

Here j is the imaginary unit, f_c the cutoff frequency, τ_g the delay, A_{res} the resonance amplitude, f_{res} its frequency, and ζ its damping factor. The implementation adds 10^{-12} to the resonance denominator for numerical stability. We fix $f_{\mathrm{res}} = 800\,\mathrm{Hz}$ and $\zeta = 0.10$. These functional forms are motivated by gradient-response modeling (Vannesjo et al., 2013; 2016); the complete parameter set is an experimental construction, not a fit to the scanners that acquired the reference datasets.

B.2 SEVERITY PRESETS

The main evaluator calls `from_severity` with the explicit presets in Table 6. The nominal condition bypasses the filter and uses the original trajectory. The labels order these presets; they are not clinical maintenance or calibration categories. For reproducibility, the table gives the actual filter parameters rather than the separate legacy `from_scale` helper, whose argument is scaled relative to the severe preset.

Table 6: Parameters used by the evaluator. Deviations are mean coordinate differences after radial-extent normalization, relative to $k_{\max}$, on the benchmark spiral; they are not RMS errors.

Preset	f_c (kHz)	τ_g (μs)	A_{res}	Mean deviation (%)
minimal	800	0.1	0.001	0.08
mild	400	0.2	0.002	0.17
moderate	160	0.5	0.005	0.46
strong	80	1.0	0.010	1.01
severe	40	2.0	0.020	2.28
extreme	16	5.0	0.050	5.87

B.3 INTEGRATION, NORMALIZATION, AND RESAMPLING

The gradient waveform is sampled every $10\,\mu$s, and the analog-to-digital converter (ADC) records a measurement every $4\,\mu$s. For each interleave, the code samples H at the FFT frequencies of the gradient array, multiplies each channel's FFT by H, and takes the real part of its inverse FFT. No temporal zero-padding is applied; this finite-array implementation has circular boundary conditions. The integrated trajectory is $\tilde{k}_n = \bar{\gamma}\Delta t \sum_{j \le n} g_{\mathrm{real},j}$, with $\bar{\gamma} = 42.576\,\mathrm{MHz/T}$. The code then normalizes its maximum radial extent to that of the nominal trajectory:

$$k_{\mathrm{real},n} = \tilde{k}_n \, \frac{\max_j |k_{\mathrm{nom},j}|}{\max_j |\tilde{k}_j|}.$$

Consequently this study isolates coordinate-shape changes after removing a global radial scaling; it does not test arbitrary gradient gain errors. There is no additional subtraction enforcing zero initial displacement. Real and imaginary coordinates are linearly interpolated to the ADC grid and converted to NUFFT radians in the order (k_y, k_x). The nominal trajectory has maximum radius 0.5 before multiplying by 2π.

The reported deviation is $\langle |k_{\mathrm{real}} - k_{\mathrm{nom}}| \rangle / \max |k_{\mathrm{nom}}|$. The code also exposes the corresponding maximum; no Gaussian model or fixed conversion between mean, maximum, and RMS deviations is assumed. The measured-GIRF overlay uses the same radial-extent convention (Appendix F.1).

B.4 SCOPE OF THE SIMULATOR

The simulator forms multi-coil data from an image and supplied sensitivity maps with the same NUFFT discretization used in inversion. No extra measurement noise, motion, off-resonance phase evolution, thermal variation, cross-axis response, or concomitant fields are added. A nonlinear slew/amplitude limiter is a separate check, not part of the main LTI sweep. Circular filtering, radial normalization, and exact sensitivity information limit physiological and hardware realism. Adding omitted effects need not monotonically worsen every method or preserve any ranking.

B.5 FROZEN NUMERICAL PROTOCOL

The historical density weights were produced by ten iterations of Fourier forward/adjoint reweighting in single precision. This is not the interpolation-only Pipe–Menon density-compensation algorithm. The calculation is sensitive to floating-point reduction order: on the illustrative slice, recomputing with four rather than ten CPU threads changed the U-Net nominal score from 35.82 to $27.73\,\mathrm{dB}$ without changing the checkpoint. The released protocol therefore freezes both sampled coordinates and the actual density weights as hashed arrays. These arrays reproduce the historical inputs without requiring the same CPU thread setting. This establishes a reproducible numerical stress test; it does not demonstrate that every ordering persists under a different, better-conditioned density-compensation scheme. Such a change also changes training inputs; Appendix F.6 reports both a fixed-checkpoint intervention and a bounded adaptation control.

B.6 VISUAL SEVERITY LADDER

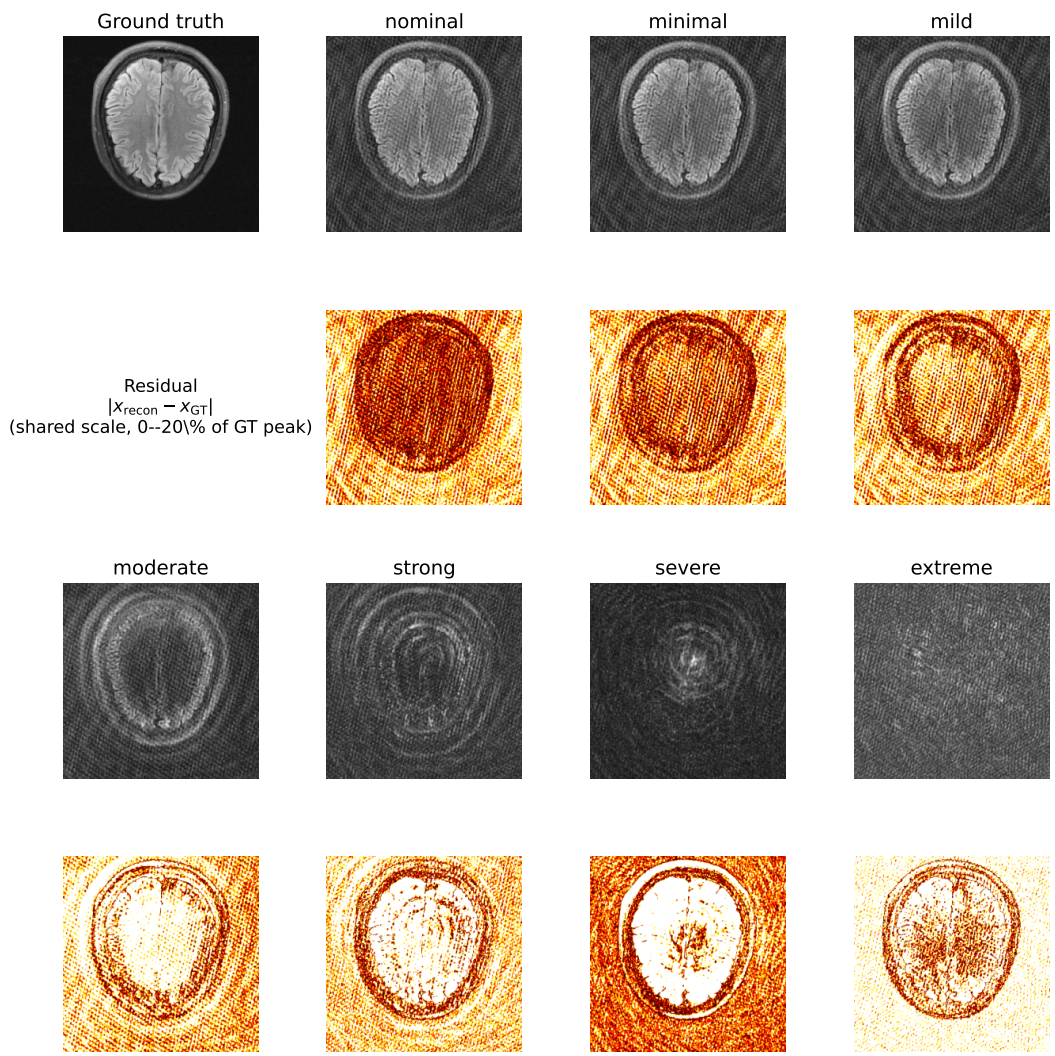

Figure 4: Nominal-operator adjoint reconstructions across the severity presets on one Brain slice. Images and residuals use shared windows within their respective groups. This visualization illustrates the simulated input distortion and does not establish a method-level failure mechanism or clinical consequence.

C Implementations and training configuration

Architecture names identify the models on which our implementations are based. We document their adaptations, optimization, and checkpoint selection separately, because each can affect the reported comparison.

C.1 Data and optimization

The primary models use the first 2,000 entries (344 volumes) of the seed-0 training manifest and all 800 validation entries. The full training manifest contains 5,002 slices and the test manifest 802; the primary audit uses its first 300 entries from 52 volumes. Volume identifiers are disjoint across training, validation, and test. The exact manifest is distributed without raw images.

All primary learned models use Adam. U-Net and SwinMR optimize $0.8L_1 + 0.2(1 - \mathrm{SSIM})$ on magnitude images; MoDL, E2E-VarNet, PromptMR, and their augmented variants use magnitude L_1 loss. Learning rates are 10^{-3} (U-Net), 2×10^{-4} (SwinMR, E2E-VarNet, PromptMR), and 10^{-4}

(MoDL). StepLR halves the rate every 40 epochs for U-Net, 30 for SwinMR/MoDL, and 15 for E2E-VarNet/PromptMR. Image-domain validation runs each epoch; unrolled validation runs every two epochs, with the first three also evaluated. Early-stopping patience is 15 validation checks; manual interruption and epoch caps also limit the historical runs. Checkpoints maximize nominal validation PSNR unless explicitly identified as augmented selection below.

C.2 Architecture adaptations

Classical methods implement CG-SENSE and wavelet-regularized SENSE (Pruessmann et al., 2001; Lustig et al., 2007; Beck & Teboulle, 2009). Learned models are non-Cartesian implementations based on U-Net, SwinMR, MoDL, E2E-VarNet, and PromptMR (Ronneberger et al., 2015; Huang et al., 2022; Aggarwal et al., 2018; Sriram et al., 2020; Xin et al., 2023). They are retrained in this simulator rather than evaluated as original published checkpoints. Their architecture configurations are stored in the retained checkpoint metadata; the release includes that metadata and weight-file hashes.

Table 7: Architecture settings of the evaluated implementations. The classical solver settings are given in Appendix D.5.

Method	Configuration
U-Net	Encoder widths 32/64/128/256; single-channel magnitude input from a density-weighted adjoint RSS image.
SwinMR	Embedding width 60; four residual Swin blocks of depth six; six attention heads; window size eight.
MoDL	Four unrolls; three CG iterations per data-consistency solve; five denoiser blocks, width 64; learned regularization weight.
E2E-VarNet	Six image-domain NUFFT residual-update cascades; normalized U-Net refiner, width 18, three pooling levels; learned DC step initialized to 0.1.
PromptMR	Six adapted cascades; width 18, three pooling levels; prompt dimension 64 and six prompts; prompt-conditioned refiner.
MoDL-Aug	Same architecture as MoDL; perturbed acquisition during training and nominal DC operator.

Our adapted VarNet receives supplied sensitivity maps and evolves an image-domain iterate, whereas the original architecture evolves multi-coil k-space and estimates sensitivities. The normalized refiner predicts a residual and restores its scale without adding back the input mean. Our PromptMR adaptation uses cascade-indexed prompt embeddings with feature-wise linear modulation (FiLM) on a single slice, rather than the original input-adaptive visual prompts and adjacent-input design. These choices and the compact cascade budgets can affect robustness; statements in this paper are restricted to the tested implementations. A CG iteration entails both forward and adjoint operations, so matching cascade counts is not a matched-compute design.

C.3 Checkpoint record

The seed-0 checkpoint files record the following selected epochs and validation scores: U-Net, epoch 50 (32.89 dB); SwinMR, epoch 71 (31.85 dB); MoDL, epoch 98 (40.07 dB); E2E-VarNet, epoch 46 (44.57 dB); PromptMR, epoch 58 (43.54 dB). These are validation scores, not the test scores reported in the main tables. The main training runs used an RTX 3080 (10 GB); the separate 5k configuration used an A100 and regenerated preprocessing. We do not use nominal validation differences to infer uncertainty in mismatch rankings.

C.4 Training-seed coverage

Every primary and external population comparison uses the seed-0 checkpoint. An additional U-Net training run is retained, but does not enter the reported mean or interval. The bootstrap resamples test volumes and cannot substitute for independently repeated training. No claim is made that additional seeds would leave every rank unchanged.

C.5 GIRF-AWARE AUGMENTATION AND CHECKPOINT SELECTION

For MoDL-Aug, E2E-VarNet-Aug, and PromptMR-Aug, 70% of training iterations simulate the acquisition on a distorted trajectory selected uniformly from the six presets; the other 30% use nominal acquisition. The data-consistency layers always use the nominal operator. These models use the same 2,000/800 training/validation subsets as their plain counterparts, with architecture-specific optimization settings specified in the training drivers.

Validation evaluates both nominal and mismatch strong. The primary augmented checkpoint, `best_model.pt`, maximizes strong validation PSNR. A second `best_model_nom.pt` maximizes nominal validation PSNR. Early stopping follows the strong-validation stopper. Both checkpoint result files are released; the primary text uses the strong-selected version. This selection choice is part of the intervention, so it is not described as identical nominal-only checkpoint selection across all methods.

C.6 BUDGET AND REPRODUCIBILITY BOUNDARIES

The evaluated configurations use a 2,000-slice training cap and one population-evaluated training seed. MoDL uses four unrolls with three inner CG steps. Primary classical evaluation uses 30 CG-SENSE and 100 L1-SENSE iterations. These are fixed compute choices; we do not claim an exhaustive convergence or architecture-capacity study. The released training drivers allow different budgets, and the checkpoint metadata records the evaluated architecture. Raw data and weight files are not included in the submission ZIP. Training anew can therefore reproduce the protocol but is not guaranteed to yield bit-identical predictions; the released measured CSVs reproduce the quantitative tables and plots without retraining.

D MEASURED RESULTS, STATISTICS, AND PROVENANCE

The tables below connect each reported comparison to its measured sample set. We first specify coverage and uncertainty, then give full scores, implementation conventions, and reproduction checks.

D.1 COVERAGE AND SOURCE POLICY

The primary audit reports eight ranked methods on 300 slices from 52 Brain volumes, each under 13 conditions. External rankings report the same seven methods on Brain, Knee, and CC359 when computing Kendall τ; the two external sets each contain 200 slices from 20 volumes and five conditions. A second MoDL-Aug checkpoint is provided as a selection-control result, not a ninth ranked method.

Only directly computed per-slice measurements from accepted implementations are included. A score-based candidate is excluded after a predictor-sign error was found in its local sampler; its measured outputs do not constitute a valid Score-MRI baseline (Appendix F.7). L1-SENSE is excluded from both external ranking cohorts because calibrated Knee measurements were incomplete. No estimated replacement, cross-dataset offset, or another solver's per-slice values are used to fill missing cells. The released `results/manifest.json` records the included files, their SHA-256 hashes, sample counts, and excluded comparisons. Any future completed evaluations should be reported as additional measurements with their actual sample set.

D.2 UNCERTAINTY AND PAIRING

All reported inferential intervals use 20,000 bootstrap draws of whole volumes (seed 0), preserving the constituent slices on each draw. The score is the mean over slices, so volumes contribute in proportion to the number of evaluated slices they contain. Differences pair `volume_id` and `slice_id` within each condition; incomplete or duplicate keys cause the analysis to fail. Intervals are percentile 95% intervals conditional on fixed checkpoints and a fixed trajectory preset. They do not include variation from training seeds, scanner hardware, or the choice of perturbation family.

Ranks are determined by mean scores. The $22/28$ count is descriptive: a pair is counted when its mean difference changes sign between nominal and moderate. It is not a count of simultaneous

Table 8: Mean PSNR (dB) across nominal and mismatch severities on Brain (300 slices). Full volume-bootstrap intervals for five metrics are in the released generated tables.

Method	Nom.	Minimal	Mild	Moderate	Strong	Severe	Extreme
CG-SENSE	44.63	31.57	25.60	18.25	16.18	15.74	15.72
L1-SENSE	44.41	31.63	25.63	18.25	16.18	15.74	15.73
U-Net	34.01	31.40	27.35	19.22	16.77	16.36	16.17
SwinMR	32.78	31.03	27.69	19.52	16.80	16.35	16.07
MoDL	39.90	31.93	26.19	16.89	13.64	13.30	13.19
E2E-VarNet	44.20	31.43	25.37	18.29	16.13	15.75	15.66
PromptMR	43.21	32.14	26.04	18.62	16.18	15.67	15.59
MoDL-Aug	36.67	33.93	30.52	24.66	21.46	20.73	20.60

Table 9: Trajectory substitution at strong severity (PSNR, dB). Gain and remaining loss are signed diagnostic contrasts, not causal components.

Method	Nominal	Mismatch	Oracle	Gain	Remaining
CG-SENSE	44.63	16.18	41.51	+25.33	+3.12
L1-SENSE	44.41	16.18	41.22	+25.04	+3.19
U-Net	34.01	16.77	18.11	+1.34	+15.91
SwinMR	32.78	16.80	17.44	+0.64	+15.35
MoDL	39.90	13.64	20.42	+6.77	+19.49
E2E-VarNet	44.20	16.13	15.67	-0.46	+28.53
PromptMR	43.21	16.18	16.06	-0.12	+27.15
MoDL-Aug	36.67	21.46	22.41	+0.95	+14.27

hypothesis rejections. We give volume-level paired intervals for the highlighted contrasts. The optional positive-bootstrap fraction in the released CSV is the fraction of resampled means above zero, not a p-value or a posterior probability, and is not used as a 0.75 significance rule.

D.3 Absolute results and trajectory substitution

The full sweep separates two outcomes: how much image fidelity is lost under mismatch, and how much of that loss a fixed pipeline recovers when supplied with the executed trajectory.

The trajectory-substitution gain G and remaining loss L are defined in Eq. 1. At mild, for example, SwinMR's substitution gain is negative (-1.39 dB); at strong E2E-VarNet's is also negative (-0.46 dB). The optional ratio $G/(G+L)$ can therefore lie outside $[0, 1]$ and becomes unstable when total loss is near zero. We release signed contrasts across all severities and use absolute scores for interpretation. A correct reconstruction operator is not an optimality guarantee for a fixed trained estimator.

D.4 Multi-metric rank patterns

PSNR and NRMSE both depend on pixel error, so agreement between them is not independent evidence. LPIPS is a natural-image perceptual metric applied here as an additional proxy, not a validated diagnostic endpoint. HFEN is based on a high-frequency error filter and gives a different mean order for SwinMR versus E2E-VarNet at strong; the volume-bootstrap interval includes zero. All five metrics are retained to expose rather than conceal this dependence.

D.5 Baseline implementation and amplitude conventions

CG-SENSE uses 30 conjugate-gradient iterations of the density-weighted regularized normal system $(A^{\mathsf{H}}DA + \lambda I)x = A^{\mathsf{H}}Dy$, with $\lambda = 10^{-6}$. Here A is the encoding operator, A^{H} its adjoint, D the diagonal density-weight matrix, and I the identity. These are specified solver settings in this simulator, not evidence that every baseline has reached its best achievable performance. L1-SENSE uses a Daubechies-4 wavelet penalty, 100 FISTA iterations, and weight $10^{-4}\,\mathrm{gt}_{\max}$ for the primary result, where $\mathrm{gt}_{\max}$ is the reference-image maximum. The relative amplitude scaling avoids applying a unit-scale penalty to raw small-amplitude fastMRI images; it also uses information that would not ordinarily be available at deployment.

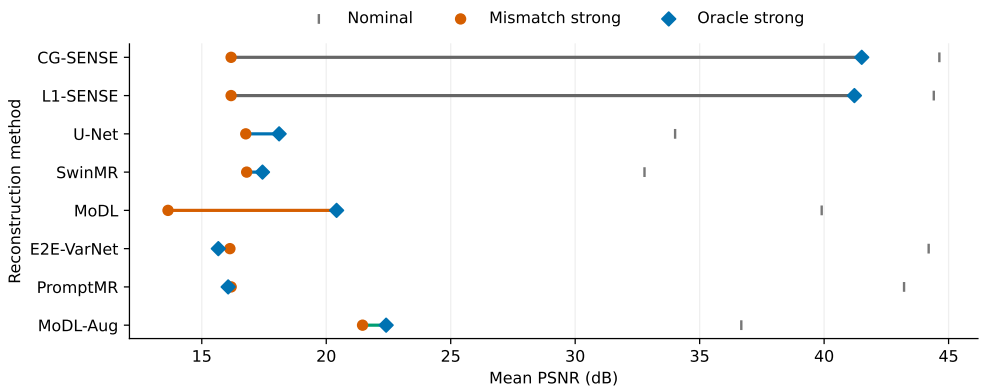

Figure 5: Absolute PSNR under nominal, mismatch strong, and oracle strong conditions. A line connects each method's mismatch and oracle scores. All comparisons use the same 300 slices; oracle changes the reconstruction operator, density compensation, and dependent initialization, but does not retrain the model.

Table 10: SwinMR versus E2E-VarNet at strong severity. Changes count mean-order reversals among eight methods from nominal to moderate; they are not counts of significant pairwise tests. Intervals are volume-bootstrap differences (SwinMR minus VarNet).

Metric	Better	SwinMR	VarNet	Changes	95% CI
PSNR	higher	16.804	16.130	22/28	$[+0.610, +0.741]$
SSIM	higher	0.403	0.225	23/28	$[+0.167, +0.190]$
NRMSE	lower	0.644	0.696	22/28	$[-0.057, -0.047]$
LPIPS	lower	0.421	0.576	22/28	$[-0.167, -0.142]$
HFEN	lower	1.267	1.261	21/28	$[-0.009, +0.024]$

MoDL and MoDL-Aug normalize the measurements by the reference image's maximum to reproduce their training amplitude convention and rescale the output afterward. U-Net and SwinMR normalize by the maximum of the adjoint input; the adapted unrolled refiners normalize internally. All methods' output magnitudes are additionally least-squares aligned to the reference before scoring. This latter scoring convention cannot remove differences caused by nonlinear processing under different input scales. VarNet training normalizes targets by their reference-image maximum (GT-max), whereas inference uses raw-amplitude measurements; fixed normalization epsilons make scale equivariance approximate. Ground-truth-based normalization is an explicit controlled simulation assumption; an input-estimated scale and no-alignment sensitivity analysis would be needed for deployment claims.

D.6 Reproducing the quantitative artifact

Run `scripts/regenerate_submission.py -results results -paper generated` from the released package. It verifies source hashes, rejects missing pairs, and produces the main quantitative figures, LaTeX tables, per-cell intervals, rank CSVs, augmentation comparisons, and a JSON summary. The `brain_ci_*`, `fastmri_knee_ci_*`, and `cc359_ci_*` CSVs include both slice and volume counts. Reconstructing images requires separately obtained raw data and retrained or locally retained checkpoints, which are not included in the submission ZIP.

Image-level reproduction check. A fresh check covers 96 reconstructions: three fixed test indices (0, 12, 299), four conditions (nominal, mild mismatch, strong mismatch, strong oracle), and eight methods. With frozen density weights, all 72 learned-method cells agree with historical PSNR within 0.0027 dB. Classical solvers differ by up to 0.258 dB, failing the all-cell 0.01 dB agreement criterion. L1-SENSE's Lipschitz power iteration uses a random initial vector whose historical seed was not recorded; finite-precision iterative solves also contribute variation. The package includes the check and its failed criterion; historical measured CSVs remain unchanged. This is a limited reproduction check, not a population rerun or an extra statistical sample.

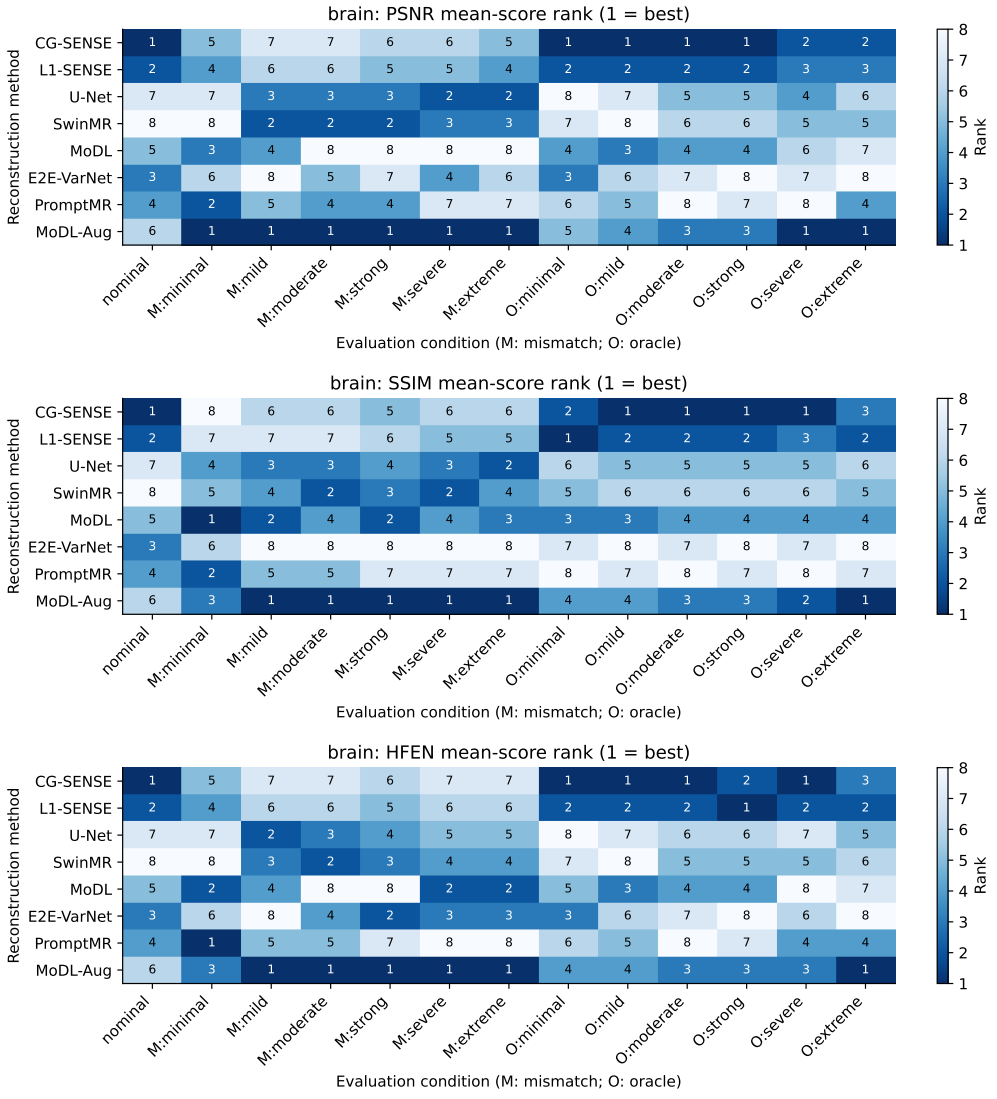

Figure 6: Brain mean-score ranks under PSNR, SSIM, and HFEN. Eight methods, 13 conditions, 300 slices per cell. Remaining metrics and full mean/interval tables are generated from the same released data.

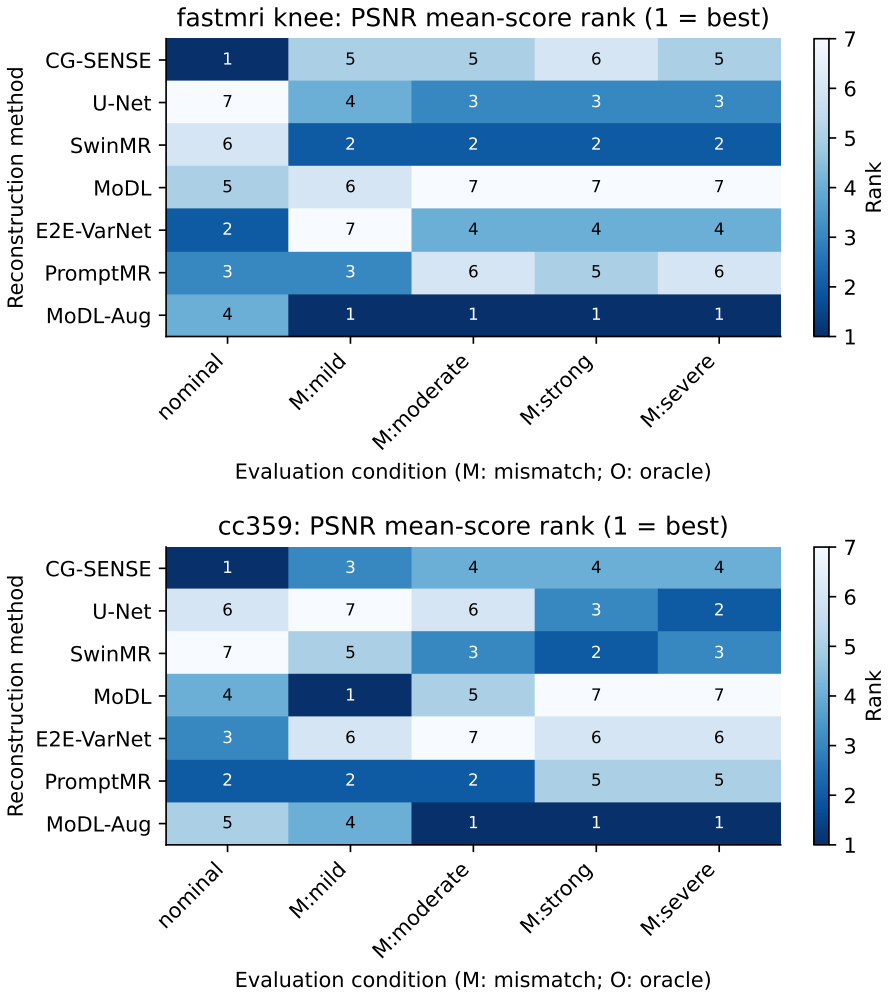

Figure 7: External PSNR rankings of the seven common measured methods. Each dataset has 200 slices from 20 volumes. Comparisons with Brain use this same subset; external L1-SENSE and Score-MRI are not imputed.

E VISUAL FAILURE GALLERY

These examples make the absolute fidelity changes visible. The fixed illustration slice does not establish how often an artifact occurs; population scores and paired differences are reported separately.

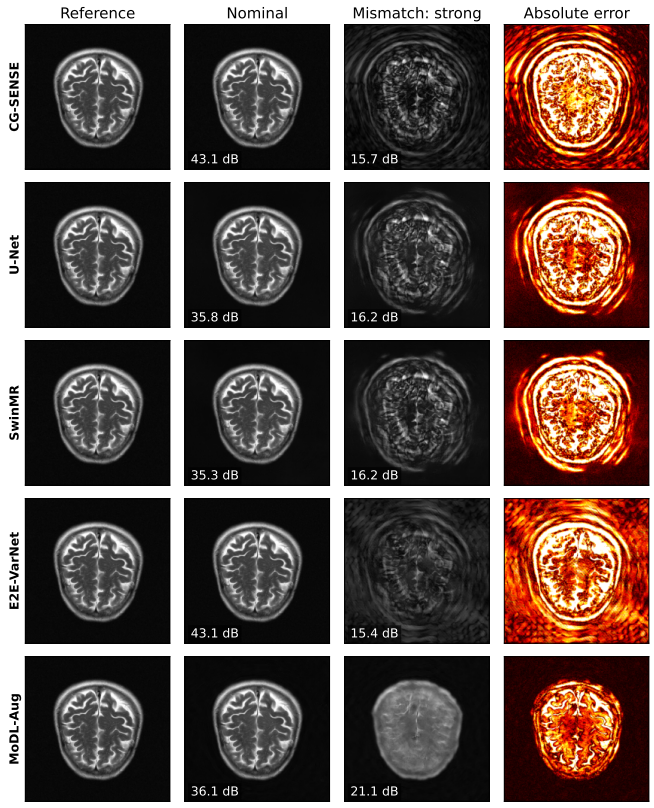

Figure 8: Compact qualitative excerpt on a fixed fastMRI Brain test slice (manifest index 12), selected for illustration (CG-SENSE, U-Net, SwinMR, E2E-VarNet, MoDL-Aug). Columns: ground truth, `nominal`, `mismatch_strong`, and the magnitude difference at `mismatch_strong`. Inset PSNR values are computed directly from the displayed, amplitude-aligned reconstructions.

Why nominal reconstructions that look alike differ by several dB. In Figure 8 the nominal reconstructions of the methods look similar at the displayed scale, yet their per-slice PSNR differs by about 8 dB. Figure 9 shows where that difference lies. With an error colour scale set for the nominal residuals, rather than the much larger strong-severity scale, the image-domain residuals trace the tissue structure, consistent with some loss of fine detail, while the CG-SENSE and E2E-VarNet residuals are smaller. The displayed residuals locate a fine-detail component of the gap. It is not specific to this slice: over the test set the same ordering holds on HFEN, which weights high spatial frequencies (0.017 for E2E-VarNet vs. 0.157 for SwinMR at `nominal`), and restricting PSNR to the anatomical foreground does not close it (41.6 vs. 29.6 dB). This nominal advantage does not preserve the ordering under GED (Table 2).

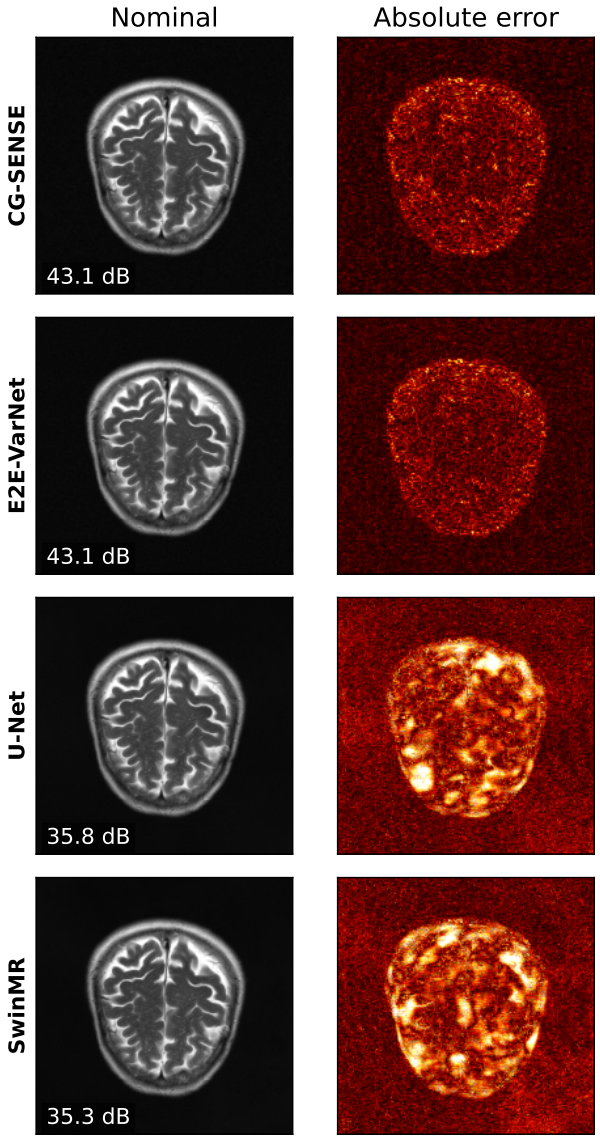

Figure 9: Nominal reconstructions and error maps on the slice of Figure 8. Rows: CG-SENSE and E2E-VarNet (data consistency at inference), U-Net and SwinMR (image domain). The error column uses one colour scale shared across rows (99th percentile of all nominal residuals). Inset PSNR values are computed directly from the displayed, amplitude-aligned reconstructions.

F ROBUSTNESS CHECKS

These checks ask whether the observed reversal depends on the chosen severity scale, training configuration, metric, sampling path, or numerical weights. Table 4 gives paired volume-bootstrap intervals for the highlighted strong-mismatch contrasts. Additional tables below give descriptive means; uncertainty is not inferred by treating slices from the same volume as independent.

F.1 SEVERITY REFERENCES FROM MEASURED GIRFS

We pass the benchmark commanded waveform through two publicly released measured GIRFs from reVERSE (Malik, 2015). Their induced mean coordinate deviations, under the radial-extent normalization used in the simulator, are 3.86% (3T London, 2014) and 10.92% (7T Zurich, 2015). The former lies between severe and extreme; the latter exceeds extreme. Thus one reference lies inside the sweep and the other above it. Within the band containing 99.9% of the commanded gradient energy (below 5.6 kHz), the measured filters also differ in spectral shape from our presets. Matching a scalar deviation magnitude does not imply equivalent reconstruction effects. We use these responses for a scale comparison, not as a population model of residual clinical trajectory error. No reconstruction result in this paper is a prospective measurement on either scanner.

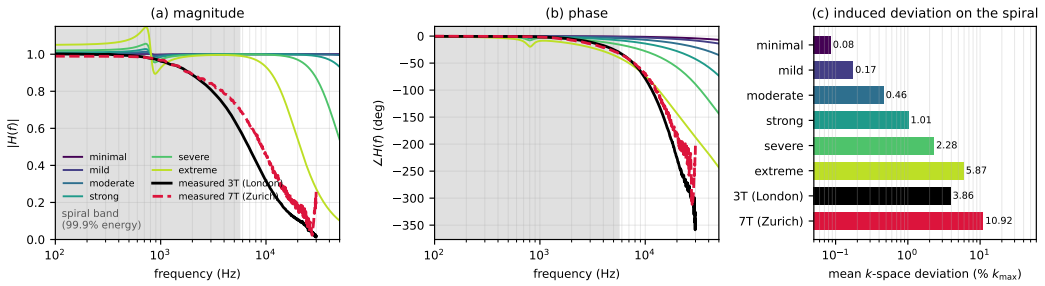

Figure 10: Parametric and scanner-measured GIRFs: response magnitude, phase, and induced mean coordinate deviation. The shaded band contains 99.9% of the commanded gradient energy. Measured response curves summarize the in-plane axes; the trajectory integration uses the axis-specific responses. Deviation values use the same radial-extent normalization as the primary simulator.

F.2 LARGER TRAINING CONFIGURATION

SwinMR, E2E-VarNet, and PromptMR were additionally trained on 5,002 slices, using a 40-epoch cap and patience 12 on an A100. Preprocessing and sensitivity maps were regenerated on that machine. Comparisons among the three 5k models are paired, but their difference from the 2k results is confounded with preprocessing and optimization budget. We therefore report replication of the ranking pattern under this configuration, not an isolated training-set-size effect. At strong, SwinMR exceeds E2E-VarNet by 0.70 dB, with a volume-bootstrap interval $[0.60, 0.80]$ dB. No greater-than-10k or matched-nominal-performance claim is made.

Table 11: Mean PSNR (dB) at 2,000 vs. 5,002 training slices ($n = 300$ per cell).

Condition	E2E-VarNet		SwinMR		PromptMR	
	2k	5k	2k	5k	2k	5k
nominal	44.20	41.60	32.78	32.42	43.21	43.42
mismatch minimal	31.43	32.51	31.03	30.84	32.14	31.28
mismatch mild	25.37	26.58	27.69	27.46	26.04	25.18
mismatch moderate	18.29	19.00	19.52	19.58	18.62	17.94
mismatch strong	16.13	15.87	16.80	16.57	16.18	15.88
mismatch severe	15.75	15.35	16.35	16.12	15.67	15.67
mismatch extreme	15.66	15.23	16.07	15.64	15.59	15.56

F.3 METRICS AND FOREGROUND MASKING

This check asks whether the ranking change is specific to PSNR or to background pixels. The five-metric comparison is in Table 10. SwinMR exceeds E2E-VarNet at strong on PSNR and SSIM and has lower NRMSE and LPIPS. HFEN gives a slightly lower mean for E2E-VarNet, but the volume-bootstrap difference interval includes zero. Mean-rank changes among the eight methods are counted separately from these paired contrasts.

The sensitivity maps are normalized across the field of view and are not foreground segmentations. We therefore construct an image mask from the ground truth using Otsu thresholding, hole filling, the largest connected component, and two-pixel dilation (mean area approximately 33%). Foreground PSNR at strong is 13.58 dB for SwinMR and 13.23 dB for E2E-VarNet; their difference interval is $[0.28, 0.43]$ dB. At nominal the foreground still favors E2E-VarNet (41.6 versus 29.6 dB). These results argue against background pixels alone explaining the reversal; they do not establish diagnostic value.

F.4 NONLINEAR GRADIENT EXECUTION

To test one departure from the linear hardware model, we cap the GIRF-filtered gradient amplitude and its rate of change (slew rate) before integrating the trajectory. It is a first-order saturation model, not a full nonlinear hardware emulator. On the same 300 Brain slices, it increases mean coordinate deviation from 0.46 to 1.03% at moderate, from 1.01 to 4.40% at strong, and from 2.28 to 16.1% at severe. All percentages use the normalized trajectory convention; direct equivalence to a measured scanner is not assumed.

At strong, SwinMR remains ahead of E2E-VarNet by 0.51 dB, with a volume-bootstrap interval $[0.45, 0.59]$ dB. At severe, the SwinMR mean instead falls 0.12 dB below E2E-VarNet, and several methods have slightly higher PSNR than under LTI. The reversal is therefore not universal across saturation strengths, and increasing coordinate deviation need not lower every metric monotonically.

Table 12: Mean PSNR (dB) under the LTI GED model and with slew/amplitude saturation (`clip`), $n = 300$ per cell. The table reports descriptive means; the strong image-domain contrast has a volume-bootstrap interval in Table 4.

	moderate		strong		severe	
Method	LTI	clip	LTI	clip	LTI	clip
CG-SENSE	18.25	16.62	16.18	15.77	15.74	15.88
U-Net	19.22	17.16	16.77	16.30	16.36	16.03
SwinMR	19.52	17.13	16.80	16.22	16.35	15.80
MoDL	16.89	13.68	13.64	13.24	13.30	14.78
E2E-VarNet	18.29	16.34	16.13	15.71	15.75	15.92
PromptMR	18.62	16.58	16.18	15.67	15.67	15.78

F.5 RADIAL-NATIVE MODELS

To test a second sampling geometry, the radial acquisition uses 201 straight, full-diameter paths (spokes) (approximately $2.5\times$ undersampling) and a prephasing lobe. U-Net and E2E-VarNet are trained on the radial operator using the same 2,000/800 training/validation subset sizes, with epoch caps 80 and 40. CG-SENSE and L1-SENSE require no training. Evaluation covers 150 test slices from 26 volumes per condition.

At strong the radial mean deviation is 0.36% of $k_{\max}$, compared with 1.01% on the spiral. The severity names are thus not matched physical doses across readouts. The radial-native U-Net trails E2E-VarNet by 7.7 dB at nominal and leads by 0.67 dB at strong (95% volume-bootstrap CI $[0.46, 0.93]$ dB). This supports a second-trajectory example of reversal, rather than trajectory-universal thresholds or family-level conclusions.

Table 13: Radial GED sweep, mean PSNR (dB), $n = 150$ per cell; mismatch conditions except `nominal`.

Method	nominal	mild	moderate	strong	severe	extreme
CG-SENSE (training-free)	41.93	35.82	29.69	24.76	20.30	16.82
L1-SENSE (training-free)	43.25	36.00	29.66	24.74	20.29	16.81
U-Net (radial-native)	33.27	32.00	28.84	**24.85**	**20.65**	**17.53**
E2E-VarNet (radial-native)	40.99	35.01	28.95	24.18	20.21	16.50

Table 14: Density-weight sensitivity on one fixed slice from each of 52 test volumes. Standard weights are mean-matched to historical weights. Fixed models retain the original checkpoints; adapted models receive six additional nominal-training epochs, selected only by nominal validation PSNR. Differences are SwinMR minus E2E-VarNet; 95% intervals resample volumes.

Weights / models	Condition	SwinMR	VarNet	Difference	95% CI
Historical	nominal	33.01	44.18	-11.17	$[-11.75, -10.60]$
Historical	mild	27.91	25.34	$+2.57$	$[+2.24, +2.90]$
Historical	strong	16.69	16.02	$+0.67$	$[+0.60, +0.74]$
Standard (fixed models)	nominal	27.27	43.86	-16.59	$[-17.17, -16.06]$
Standard (fixed models)	mild	24.34	26.14	-1.80	$[-2.15, -1.47]$
Standard (fixed models)	strong	16.25	16.19	$+0.06$	$[+0.01, +0.12]$
Standard (adapted)	nominal	42.89	45.27	-2.38	$[-2.59, -2.18]$
Standard (adapted)	mild	26.54	26.14	$+0.40$	$[+0.34, +0.46]$
Standard (adapted)	strong	16.53	16.16	$+0.37$	$[+0.33, +0.40]$

F.6 Density-weight sensitivity and bounded adaptation

The historical density-weight iteration has numerical sensitivity (Appendix B). We separately replace its nominal weights with `torchkbnufft.calc_density_compensation_function`, which uses interpolation-based Pipe density compensation (ten iterations), and rescale the resulting vector to the historical vector's mean. This controls global weight scale while changing its distribution across samples. Both SwinMR and E2E-VarNet retain their original checkpoints; all acquisition coordinates, reconstruction coordinates, sensitivity maps, and amplitude-scoring conventions remain the same.

We select the middle entry within each volume's ordered entries in the primary 300-slice manifest, giving 52 slices from 52 volumes. Both weight choices are freshly evaluated on this same subset at nominal, mild, and strong conditions: 624 measured reconstruction cells in total. Table 14 includes the negative mild result.

Under historical weights, the subset reproduces the primary reversal. Under standard weights, SwinMR loses at mild and its strong advantage falls to 0.06 dB. Nominal SwinMR fidelity also falls substantially, whereas VarNet changes less. Thus changing weights creates an input shift for these fixed models; it cannot establish the ordering of models separately trained with standard weights. It does establish that the mild reversal does not survive this fixed-checkpoint numerical change. Threshold transfer across density-compensation implementations must not be inferred from the main volume-bootstrap intervals.

Predeclared warm-start control. To examine adaptation to the new input distribution, we additionally train both original seed-0 checkpoints for six epochs using standard weights, the same first 2,000 training and 800 validation slices, and no GIRF augmentation. The design is recorded before training. We use fresh Adam states at constant learning rate 5×10^{-5}, batch sizes four (SwinMR) and one (VarNet), and the original architecture-specific losses: 0.8 L1 plus $0.2(1 - \mathrm{SSIM})$ for SwinMR and magnitude L1 for VarNet. SwinMR retains the original flip augmentation; VarNet gradients are clipped at norm one. Selection maximizes nominal validation PSNR over epochs zero through six; no test scores enter selection. Validation uses normalized training targets without amplitude alignment; reported test PSNR uses the common least-squares alignment convention. Both selected checkpoints are epoch six. Nominal validation PSNR changes from 26.81 to 42.57 dB for SwinMR and from 43.49 to 45.76 dB for VarNet.

Only after selection, we evaluate the same 52-volume, three-condition subset, adding 312 measured cells. Adapted SwinMR/VarNet PSNR is $42.89/45.27$ dB at nominal, $26.54/26.14$ dB at mild, and $16.53/16.16$ dB at strong. The paired mild difference is $+0.399$ dB (95% CI $[0.337, 0.465]$); the strong difference is $+0.366$ dB ($[0.328, 0.404]$). The reversal therefore appears again under this adapted standard-weight setting, with smaller gaps than under historical weights. This is a bounded warm-start control, not evidence of convergence from scratch, matched optimization, or invariance across seeds. Because it also adds optimization steps, it does not isolate the density algorithm from additional training. The fixed-model negative result remains part of the comparison. The artifact includes the predeclared design, epoch histories, checkpoint hashes, and measured cells.

Descriptive consistency across volumes. For the primary 300-slice comparison, we first average paired slice PSNR differences within each of the 52 volumes. All 52 volume means favor VarNet at nominal and SwinMR at mild and strong mismatch. The density-weight controls use one preselected slice per volume, so their per-volume differences are single-slice observations, not whole-volume reconstruction assessments. Under standard weights and fixed models, SwinMR leads on $1/52$ selected slices at mild and $34/52$ at strong; after adaptation the counts are $51/52$ and $52/52$. The adapted medians are 0.375 and 0.343 dB. Removing any one volume leaves the adapted mean difference between 0.384 and 0.409 dB at mild, and between 0.359 and 0.371 dB at strong. These descriptive checks show that no single selected volume determines the sign of the adapted mean contrast; they add neither independent experiments nor evidence about training variability or clinical success. Released tables contain all paired values, sign counts, medians, and leave-one-volume-out ranges.

Bounded audit of the density-weight comparisons. We checked frozen acquisition/reconstruction coordinate identities, density-weight hashes and mean matching, volume-disjoint splits, method/condition keys, original/adapted checkpoint hashes, and adapted selection against all seven nominal-validation entries. A fresh rerun of 36 cells (two fixed subset entries, two methods, three conditions, three density/training protocols) differs from stored PSNR by at most 0.00024 dB. On one selected slice, comparing raw versus GT-max-scaled inference changes aligned PSNR by at most 0.0574 dB for VarNet and 0.000022 dB for SwinMR across those protocols and conditions. This checks a limited amplitude convention, not population invariance. Three random complex inner-product checks of the supplied-map NUFFT forward/adjoint pair have relative errors below 3×10^{-6}. The released audit records these checks separately from population measurements. They do not validate every model or replace the earlier failed classical-solver reproduction criterion; original checkpoint selection histories are insufficient for a full independent replay.

F.7 Excluded score-based implementation

An implementation audit identified a reverse-drift sign error in the local score-based candidate. Its network predicts noise $\hat{\epsilon}$ and defines the score as $s_\theta = -\hat{\epsilon}/\sigma$. For a descending variance-exploding (VE) noise schedule, the reverse predictor's score increment is $(\sigma_i^2 - \sigma_{i+1}^2)s_\theta$. The local sampler instead used $(\sigma_{i+1}^2 - \sigma_i^2)s_\theta$, reversing this drift. Consequently, neither its historical measurements nor the resumed 200-step measurements support claims about Score-MRI or generative-method robustness. The run was stopped and its outputs excluded; no partially completed population is presented as $n = 300$.

The artifact provides the corrected experimental code and an analytic regression check. With Gaussian score $s(x, \sigma) = -x/(1 + \sigma^2)$, initial value $x = 2$, schedule $1 \to 0.5 \to 0$, and noise, corrector, and data consistency disabled, the two predictor updates should yield $x = 1$. The legacy implementation yields 3.3; the correction yields 1 to numerical precision. This verifies the sign correction only, not the full MRI sampler, its training, or population performance. No corrected-sampler population comparison is claimed. The eight-method primary cohort and the density-weight adaptation control do not use this sampler.

G Extended limitations and follow-up tests

The main results establish conditional comparisons. The following untested factors delimit what can be inferred and identify experiments needed for stronger claims.

Training interventions. GIRF-aware training is tested on three unrolled implementations, each with one training seed and one perturbation family. Augmentation and strong-mismatch checkpoint selection change together. Image-domain augmentation, alternate severity mixtures, input-scale estimators, trajectory-estimation modules, and matched-compute or matched-nominal-fidelity comparisons remain untested. Similar scores from two architectures do not establish that capacity is irrelevant or that an information-theoretic limit has been reached.

Trajectory and hardware coverage. The radial experiment covers two classical solvers and two radial-native networks on 150 slices from 26 volumes. Its strong label produces a smaller coordinate deviation than the spiral strong label, so reversal thresholds should not be transferred quantitatively between readouts. EPI and prospectively acquired non-Cartesian data with independently measured trajectories remain outside this study. A single deterministic GIRF per preset does not characterize robustness to a distribution of hardware responses.

Uncertainty beyond test volumes. Resampling whole volumes addresses dependence among sampled slices within a volume. It does not establish independence between different volumes from the same person, nor does it account for model initialization, training subset selection, checkpoint selection, metric choice, or simulator design. Split integrity is verified at the volume-identifier level; stronger patient-level guarantees require linkage metadata.

Diagnostic interpretation. Changing the reconstruction trajectory also changes density compensation and initialization. A learned data-consistency step may respond to the resulting operator norm and conditioning. To isolate a mechanism, future tests should independently vary initialization, data-consistency coordinates, density weights, learned step sizes, and the refiner. The current oracle intervention is useful for comparing complete fixed pipelines but does not identify those modules' causal contributions.

External datasets and endpoints. Knee and CC359 add combined anatomy, contrast, coil, and scanner shifts; we cannot attribute differences to vendor alone. Other field strengths, acquisition contrasts, datasets, and clinical tasks require separate validation. PSNR, SSIM, NRMSE, LPIPS, and HFEN quantify image differences, not lesion detectability or clinical acceptability.