PRIORSHARE: BENCHMARKING AND UNDERSTANDING HALLUCINATION IN DIFFUSION-BASED SCIENTIFIC IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

In scientific imaging, reference images are often unavailable, so a small measurement residual is the main check on a reconstruction from a reusable diffusion prior. The residual is insensitive to errors in source components that the measurement leaves unresolved, which we study as hallucinations. We introduce the *prior share*, the fraction of source energy in the directions a given operator and noise level leave unobservable. It is computed from the operator and representative source data without a test reference, and unlike operator-only summaries it accounts for how source energy is distributed across unresolved directions. We build the **PriorShare** benchmark around three new settings: weak-lensing mass mapping, galaxy PSF deconvolution, and reconstruction from real wide-field fluorescence observations. With multi-coil MRI and black-hole imaging added, it evaluates nine diffusion-prior solvers. In a controlled study that varies only the operator, agreement between residual and PSNR rankings decreases as the share grows. At a share of 72%, one solver reaches a relative residual of 0.6% while 99.95% of its squared reconstruction error lies in the discarded directions, mostly in the unmeasured mean level. In the scientific settings, four of nine solvers on BioSR and five on Galaxy incur larger squared error in the unresolved components than zero filling. On noiseless galaxy deconvolution, an untuned inverse without a learned prior outperforms all nine solvers in PSNR and SSIM. PriorShare thus reports, beside data fit and reference fidelity, how much source energy a residual cannot check.

1 INTRODUCTION

A trained diffusion model can be reused as a prior across acquisition settings (Song et al., 2021; Karras et al., 2022; Chung et al., 2023; Mardani et al., 2024). In scientific imaging, reference images are often unavailable. The convergence field of a survey patch is never observed directly, and no reference image of a black hole's horizon exists, so a small measurement residual becomes the main check on a reconstruction. Yet a measurement constrains only some directions of the source, and the residual is insensitive to errors in the remaining directions, which Bhadra et al. (2021) call null-space hallucinations (Gottschling et al., 2025). Existing benchmarks compare solvers against references on fixed operators (Kawar et al., 2022; Chung et al., 2023; Zheng et al., 2025a), and operator-level quantities such as conditioning, null-space size, or forward-model bounds (Iagaru et al., 2026) do not account for how source energy is distributed across operator directions. None of these quantifies how much source energy a measurement leaves unresolved.

We introduce the **prior share**, the fraction of source energy in directions that a measurement leaves unobserved or only weakly observed at a specified noise level. Prior share is computed from the operator, its noise level, and second-order statistics of representative sources, without a test reference, so it is available before any reconstruction. The number of unresolved directions can be small even when they carry most of the source energy. A Kaiser–Squires diagnostic operator that discards 0.78% of the frequency grid leaves 72% of the source energy unresolved, whereas a Moffat blur that pushes 80% of the grid below the 8-bit quantization floor leaves 3.4%.

The **PriorShare** benchmark, shown in Fig. 1, builds three scientific settings: weak-lensing mass mapping, galaxy PSF deconvolution on Galaxy10 DECaLS (Leung & Bovy, 2024; Walmsley et al.,

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.

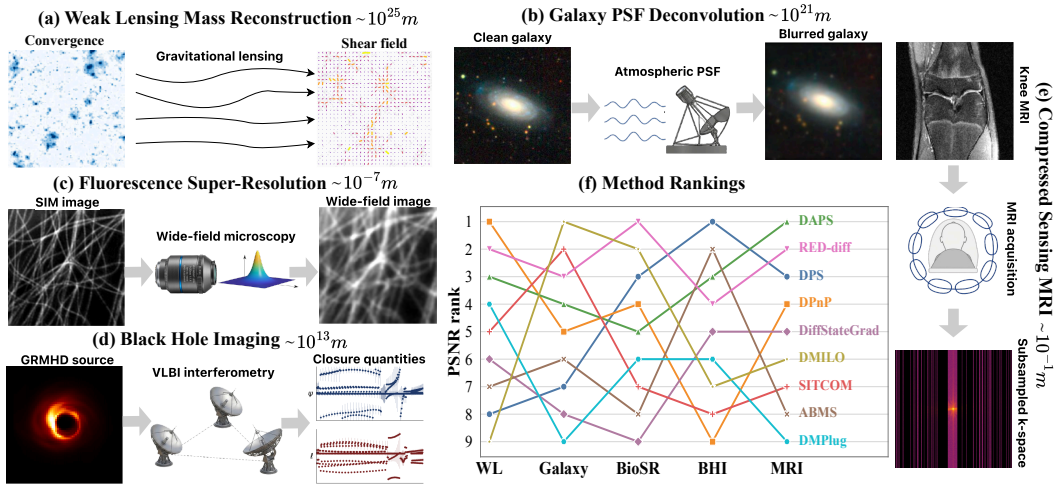


Figure 1: **The five scientific settings.** (a)–(e) Sources, acquisition, and measurements. (f) Solver rankings by PSNR in each setting.

2022; Dey et al., 2019), and reconstruction from real wide-field BioSR observations (Qiao & Li, 2024; Qiao et al., 2021). It joins them with multi-coil MRI and black-hole imaging from InverseBench to evaluate nine diffusion-prior solvers against domain baselines. A controlled study holds the Galaxy sources, prior, and solver configurations fixed and changes only the operator. There, residual and PSNR rankings agree less as the share grows. At a share of 72%, RED-diff reaches a relative residual of 0.6% while 99.95% of its squared error lies in the discarded directions, and the unmeasured mean level alone accounts for 84% of the total error. Weak lensing, Galaxy, and BioSR have shares of only 2–8%, yet on Galaxy and BioSR error decomposition still separates solvers that recover unresolved components from those that incur more squared error there than a zero fill. On noiseless Galaxy deconvolution, an untuned inverse without a learned prior outperforms all nine solvers in PSNR and SSIM.

Our contributions are:

1. **Conceptual.** The prior share, the fraction of source energy that a measurement leaves to the prior, computed from the operator, its noise level, and second-order source statistics before any reconstruction and without a test reference.
2. **Benchmark.** Three new scientific settings, weak-lensing mass mapping, galaxy PSF deconvolution, and wide-field BioSR reconstruction, joined with multi-coil MRI and black-hole imaging, and a controlled operator suite on fixed Galaxy sources, prior, and solvers.
3. **Empirical.** An evaluation of nine diffusion solvers and domain baselines that measures the error in unobservable directions against a zero fill. At the largest share, a pseudo-inverse filled with the training mean outperforms seven of the nine solvers.

Related work is reviewed in Appendix A.

2 What a measurement leaves to the prior

We distinguish *data fit*, whether $\mathcal{A}(\hat{x})$ reproduces the observation y, from *recovery*, whether $\hat{x}$ reproduces the source x. Measurement consistency constrains reconstruction error on the observed directions, through the smallest singular value there, and leaves the null space unconstrained.

Unobservable directions. Write a linear acquisition model as $y = \mathcal{A}x + \xi$ with noise of scale σ and singular system $\mathcal{A} = \sum_i s_i u_i v_i^\top$, and let $\mathcal{S}_i = \mathbb{E}_{x\sim p}\,|\langle v_i, x\rangle|^2$ be the source energy along v_i. We define

$$\mathcal{U}(\mathcal{A}) = \operatorname{span}\{v_i : \; s_i = 0 \text{ or } s_i^2\,\mathcal{S}_i < \sigma^2\}, \tag{1}$$

a subspace that contains the null space, where the data do not change, and the directions along which a source-scale change moves the data by less than σ. It plays the role of the null space in the

decomposition of Bhadra et al. (2021), and we refer to the reconstruction error projected onto $\mathcal{U}$ as hallucination.

Proposition 1 (Agreeing on the data constrains only the observable part)**.** *If two reconstructions both fit the data to within ε, so $\|\mathcal{A}\hat{x}_1 - y\| \le \varepsilon$ and $\|\mathcal{A}\hat{x}_2 - y\| \le \varepsilon$, then their difference obeys $\|P_{\mathcal{U}^\perp}(\hat{x}_1 - \hat{x}_2)\| \le 2\varepsilon/s_{\min}$, where $s_{\min}$ is the smallest singular value outside $\mathcal{U}$. Inside $\mathcal{U}$ the difference is unconstrained on the exact null space, and each remaining component is bounded only by $2\varepsilon/s_i$, which exceeds the source scale along that direction whenever $\varepsilon \ge \sigma/2$.*

Proposition 2 (Residuals bound only the observable error)**.** *For a reconstruction $\hat{x}$ with error $e = \hat{x} - x$, the residual satisfies $\|\mathcal{A}\hat{x} - y\| \ge s_{\min}\|P_{\mathcal{U}^\perp} e\| - \|\xi\|$. It is unchanged by error on the exact null space, and a source-scale error along any single direction of $\mathcal{U}$ changes it by less than the noise level.*

Both statements follow from the singular decomposition (Appendix B). For weakly observed directions, the sub-noise statement applies to an individual source-scale perturbation, not to an arbitrary error supported on all of $\mathcal{U}$.

Prior share. We call the fraction of source energy in $\mathcal{U}$ the *prior share* of an operator,

$$\rho(\mathcal{A}) = \frac{\mathbb{E}_{x\sim p}\|P_{\mathcal{U}(\mathcal{A})}\, x\|^2}{\mathbb{E}_{x\sim p}\|x\|^2}. \tag{2}$$

Prior share characterizes an operator–source pair, whereas hallucination error characterizes an individual reconstruction. For operators diagonal in the Fourier domain, Eq. 2 is the training power spectrum summed over $\mathcal{U}$ and normalized by its total, and it needs no test reference or solver run. Estimates from disjoint halves of the training data differ by less than 0.2 percentage points, and the share depends more on the assumed noise level (Appendix B.3). For zero-mean sources whose covariance is diagonal in the operator's singular basis, $\mathcal{U}$ is the complement of the likelihood-informed subspace of Bayesian inverse problems at the same threshold (Cui et al., 2014; Spantini et al., 2015). For nonzero-mean sources, $\mathcal{S}_i$ also includes a contribution from the source mean (Section 5.3).

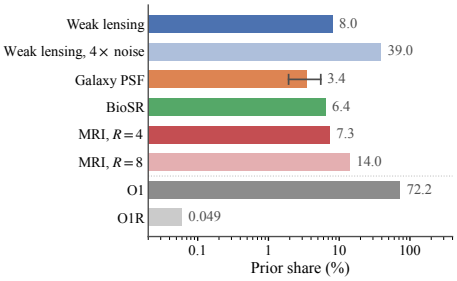


Figure 2: **Prior share of the benchmark operators.** The Galaxy whisker spans the three seeing levels, and MRI uses the single-coil sampling pattern.

Beyond operator-only summaries. Conditioning and null-space size describe the operator alone, whereas ρ weights its directions by source energy, as Fig. 2 shows for the benchmark operators. The diagnostic operator O1 of Section 5, built from the Kaiser–Squires kernel, discards 0.78% of the frequency grid and has condition number 126.5 elsewhere, yet $\rho = 72\%$, of which 67 percentage points lie in the mean level. The Galaxy blur O2 pushes 80% of the grid below the 8-bit quantization floor, yet $\rho = 3.4\%$, and the same weak-lensing operator has $\rho = 8.0\%$ or 39% at noise standard deviation 0.01 or 0.04. Conditioning still matters, since a residual bounds the observable error only through $s_{\min}$ in Proposition 2. Black-hole imaging and the magnitude-only Fourier operator O3 are nonlinear, and their measurements are invariant to translation of the source. We do not assign them a prior-share value and describe these invariances instead.

3 Constructing the PriorShare benchmark

The three new settings leave different parts of the source unresolved, as summarized in Table 1. Weak lensing loses the mean convergence and the frequencies where shear noise dominates, Galaxy deconvolution loses the frequencies where a controlled blur pushes the signal below the 8-bit quantization floor, and BioSR uses real microscope observations with a calibrated acquisition model. Each setting supplies sources, a measurement model, a trained prior, and domain evaluators.

Weak-lensing mass mapping. We connect simulated convergence fields $\kappa \in \mathbb{R}^{H\times W}$ to complex shear observations γ through the Kaiser–Squires map (Kaiser & Squires, 1993), which is diagonal in

Table 1: **Forward operators of the five settings.** Prior share at noise standard deviation 0.01 for weak lensing, FWHM $1.18''$ with an 8-bit quantization floor for Galaxy, and $R = 4$ for MRI, where it describes the single-coil sampling pattern.

Task	Forward operator	Operator chain	Class	Dataset	Res.	Prior share
Weak Lensing	Fourier-diagonal kernel	$\mathcal{F}^{-1}\hat{D}\,\mathcal{F}\,x$	Single linear	κ maps	128^2	8.0%
Galaxy PSF	Shift-invariant Moffat blur	$h * x$	Single linear	DECaLS	256^2	3.4%
Fluorescence	Calibrated affine PSF	$g(G * x) + b$	Calibrated affine	BioSR	128^2	6.4%
Black-Hole Imaging	Nonlinear closure phase	$\mathrm{cl}(\mathcal{V}\,x)$	Nonlinear	GRMHD sim.	64^2	invariance
Multi-Coil MRI	Coil-mod. $\times$ sparse Fourier	$M\,\mathcal{F}\,S\,x$	Composite linear	fastMRI	320^2	7.3%

the Fourier domain with unit modulus at every nonzero frequency and zero at the mean (Appendix C.3). Both shear components are observed with additive Gaussian noise of standard deviation 0.01. We train a prior on 128×128 convergence patches and hold out a simulation run for testing. Image errors assess recovery of the field, and power-spectrum and peak-distribution statistics assess its matter distribution.

Galaxy PSF deconvolution. This setting blurs 256×256 Galaxy10 DECaLS images (Leung & Bovy, 2024) with a normalized Moffat kernel applied to each color channel, at FWHM 0.8, 1.18, and 1.8 arcsec (Appendix C.3). The original survey image is the reconstruction target, and we add no synthetic noise. PSNR and SSIM assess image recovery, and adaptive-moment ellipticity error assesses shape preservation. The same sources and prior are also used in the controlled study of Section 5.1.

Real wide-field fluorescence reconstruction. BioSR pairs wide-field (WF) observations with higher-resolution structured illumination microscopy (SIM) references (Qiao & Li, 2024; Qiao et al., 2021). We use the Microtubules subset at 128×128, with observations acquired by the microscope rather than synthesized from the reference. The acquisition model maps the prior's normalized output to raw intensity and applies a reflection-padded Gaussian PSF with a gain and a background, $\mathcal{A}(x) = g\big(G_{\sigma,k} * x_{\mathrm{raw}}\big) + b$. The intensity map is fitted on 64 non-test pairs and selected on 16 disjoint pairs, and the PSF, gain, and background are fitted on training pairs, so test SIM images enter only the evaluator (Appendix C.3 and Fig. 6).

Multi-coil MRI and black-hole imaging extend the comparison to subsampled complex-valued measurements and nonlinear interferometric observables, and we reuse both settings and their priors from InverseBench (Zheng et al., 2025a). MRI uses $y_j = M\mathcal{F}(S_j x)$, with coil sensitivities S_j and a random Cartesian sampling mask M at acceleration $R = 4$ on fastMRI knee data (Zbontar et al., 2018). Coil sensitivities encode part of the unacquired k-space, so the unobservable set of this multi-coil operator is not the set of unacquired lines, and the share we report for MRI, that of the single-coil sampling model $M\mathcal{F}$, describes its sampling pattern. Black-hole imaging reconstructs from closure phases and log closure amplitudes, which are invariant to image translation. Each setting has its own prior checkpoint, shared by all solvers.

4 Evaluation protocol

We evaluate nine diffusion solvers, the sampling-based DPS (Chung et al., 2023), DAPS (Zhang et al., 2025), DiffStateGrad (Zirvi et al., 2025), SITCOM (Alkhouri et al., 2025), ABMS (Ren et al., 2026), and DPnP (Xu & Chi, 2024), and the optimization-based RED-diff (Mardani et al., 2024), DMPlug (Wang et al., 2024), and DMILO (Zheng et al., 2025b). All use the setting's EDM denoiser $D_\theta(x;\sigma)$ (Karras et al., 2022) with fixed weights, so they share the prior and differ in initialization, schedule, gradient path, and inner updates (Table 9 and Appendix D). The evaluation therefore compares complete reconstruction procedures.

For each method–task pair, we choose the configuration and iteration budget by PSNR on one to sixteen selection images, using translation-aligned PSNR for BHI. For MRI and BHI these came from the test split, so both are evaluated only on test images that no selection step used. Configurations are then frozen (Appendix C.4), and all methods on a task share the test set, with the

Table 2: **Operator transfer with a fixed prior.** Ten Galaxy images with the O2-selected configurations. Res. is the relative residual and Halluc. the fraction of error in the frequencies O1 discards, both in percent. Shares use the 8-bit quantization floor. O1R, in grey, is a noiseless control and is not ranked. SITCOM runs without the final mean correction used in its selection, on every operator.

	O2: Moffat PSF $\rho = 3.4\%$		O1: KS-derived $\rho = 72\%$			O1R: control $\rho = 0.05\%$		O3: Magnitude-Fourier invariance	
Method	**PSNR↑**	**Res. %**	**PSNR↑**	**Res. %**	**Halluc. %**	**PSNR↑**	**Res. %**	**PSNR↑**	**Res. %**
Sampling-based									
DPS	$25.43_{\pm 1.56}$	1.1	$29.61_{\pm 1.76}$	13	9.3	$30.05_{\pm 1.79}$	6.2	$13.05_{\pm 0.90}$	75
DAPS	$27.26_{\pm 1.61}$	0.35	$26.67_{\pm 3.69}$	1.5	84.7	$42.38_{\pm 0.97}$	0.58	$5.81_{\pm 0.06}$	200
DiffStateGrad	$25.18_{\pm 1.47}$	1.9	$27.28_{\pm 1.68}$	25	6.8	$27.59_{\pm 1.71}$	12	$12.60_{\pm 0.87}$	75
SITCOM	**30.53**$_{\pm 1.96}$	0.09	**37.29**$_{\pm 11.53}$	0.11	97.5	$62.84_{\pm 5.33}$	0.05	$14.60_{\pm 1.39}$	0.05
ABMS	$25.47_{\pm 1.57}$	3.0	$26.55_{\pm 1.58}$	30	5.7	$26.73_{\pm 1.59}$	14	**19.34**$_{\pm 1.67}$	15
DPnP	$26.26_{\pm 1.42}$	5.8	$14.12_{\pm 1.06}$	95	65.1	$17.34_{\pm 1.62}$	55	$1.73_{\pm 0.20}$	363
Optimization-based									
RED-diff	$29.31_{\pm 1.76}$	0.26	$11.46_{\pm 0.68}$	0.56	99.95	$47.85_{\pm 1.64}$	0.13	$16.28_{\pm 1.91}$	8.0
DMPlug	$21.96_{\pm 0.69}$	4.1	$25.60_{\pm 1.15}$	27	32.1	$26.75_{\pm 1.43}$	13	$18.64_{\pm 2.22}$	8.1
DMILO	$29.60_{\pm 1.82}$	1.1	$28.84_{\pm 1.76}$	6.0	85.8	$39.00_{\pm 1.38}$	2.8	$17.40_{\pm 1.55}$	2.6

few incomplete cells marked in the tables. Domain baselines are Kaiser–Squires inversion for weak lensing, Richardson–Lucy (Richardson, 1972; Lucy, 1974) for Galaxy and BioSR together with a conjugate-gradient inverse for Galaxy, closure-based regularized maximum likelihood (Chael et al., 2018) for black-hole imaging, and SENSE (Pruessmann et al., 1999), ℓ_1-wavelet (Lustig et al., 2007), and total-variation (Rudin et al., 1992) reconstruction for MRI. Table subscripts report between-image standard deviations, and Appendix E gives the search spaces and selected configurations.

PSNR and SSIM measure reference fidelity, with translation-aligned and blurred PSNR for BHI. Weak-lensing statistics, Galaxy ellipticity, and Fourier ring correlation (FRC) assess statistical structure, shape, and frequency content, BHI closure statistics assess measurement consistency, and MRI full k-space error compares acquired and unacquired coefficients with the reference (Appendices C.4 and C.5).

5 Empirical Study

The controlled study of Sections 5.1–5.3 changes only the operator and keeps the Galaxy sources, the checkpoint, and each solver's configuration fixed. It asks two questions: whether a solver ranking survives a change of operator, and whether a residual still ranks solvers once most of the source energy is unobservable.

5.1 Operator transfer with a fixed prior

We reconstruct the same ten Galaxy images with a fixed checkpoint under four operators. Every solver keeps the configuration selected on the Moffat blur O2, whose share is 3.4%, so the other operators test transfer to a new, known operator without the reference-based re-tuning that a practitioner could not perform. The diagnostic O1 keeps only the real part of the Kaiser–Squires response, a real Fourier multiplier $(\ell_1^2 - \ell_2^2)/|\boldsymbol{\ell}|^2$ that discards the mean level and the diagonals $|\ell_1| = |\ell_2|$, 510 of 65536 frequencies carrying 72% of the source energy. O3 keeps every Fourier amplitude and discards phase. O1R augments O1 with measurements of the discarded coefficients and is a control for the error decomposition of Section 5.3. Data fit is measured by the relative residual $\|\mathcal{A}(\hat{x}) - y\|_2/\|y\|_2$.

A ranking obtained on one operator need not hold on the next, even with the prior and configurations fixed. From O2 to O3 every solver loses PSNR and 18 of the 36 pairwise orderings reverse, as shown in Table 2, with a central 95% interval of 17 to 20 when the ten images are resampled. From O2 to O1 the solvers move in opposite directions, with SITCOM and DPS gaining 6.8 and 4.2 dB while RED-diff loses 17.9 dB. These results describe the transfer of fixed configurations, without operator-specific re-tuning.

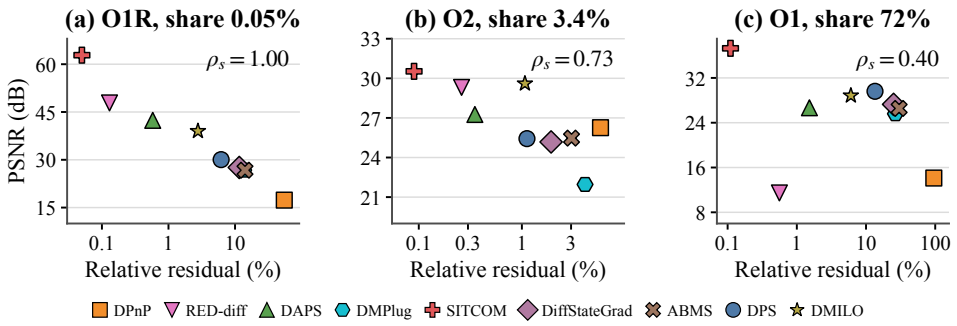


Figure 3: **Residual ranking across prior shares.** Relative residual against PSNR for the nine solvers under three operators. ρ_s is the Spearman correlation between negated residual and PSNR.

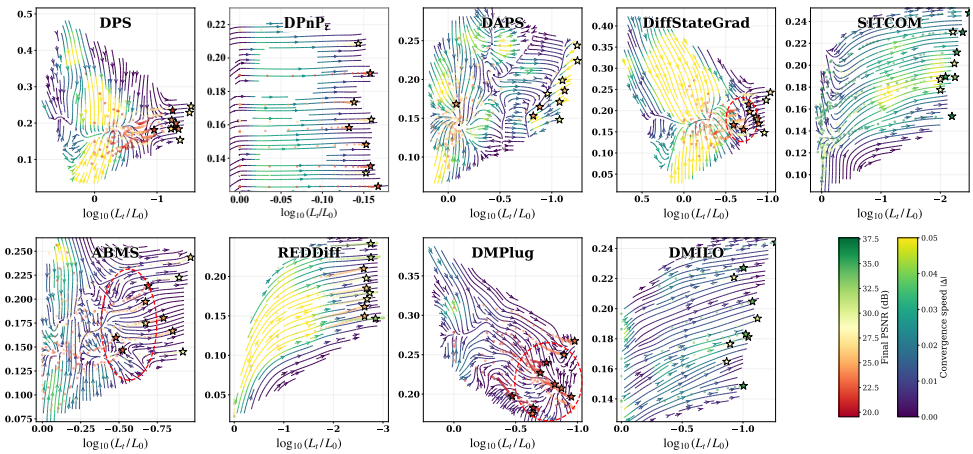


Figure 4: **Reconstruction dynamics on Galaxy PSF deconvolution.** Projected velocity fields of the nine methods over ten test images, with normalized log measurement loss on the horizontal axis and image intensity spread on the vertical axis. Endpoint color shows final PSNR.

5.2 Residual and PSNR rankings across operators

By Proposition 2, a residual certifies only the observable part of a reconstruction, so its ability to rank solvers should depend in part on the prior share. O1R, O2, and O1 differ in share while the sources, checkpoint, and configurations are fixed, and across them ranking agreement decreases as the share increases. The Spearman correlation between negated relative residual and PSNR across the nine solvers is 1.00 on O1R, where the share is 0.05%, 0.73 on O2 at 3.4%, and 0.40 on O1 at 72%, as shown in Fig. 3. This order holds in all $10{,}000$ paired bootstrap resamples of the ten images, and the central 95% intervals, $[0.98, 1.00]$, $[0.65, 0.77]$, and $[0.22, 0.47]$, do not overlap (Appendix F.6). O1R differs from O1 only by measuring the 510 discarded frequencies, 0.78% of the grid but 72% of the source energy, so the two form a paired comparison. The disagreement at 72% comes mainly from solvers that fit the data closely, and RED-diff has the second-lowest residual on O1 and the lowest PSNR.

5.3 Error in the discarded components

RED-diff shows how a close fit can hide a large error. On O1 its relative residual is 0.56%, yet 99.95% of its error lies in the discarded frequencies and 84% in the mean-level mode alone. Its mean output intensity stays near the mid-grey level that its zero initialization gives in the prior's $[-1, 1]$ coordinates, about 0.24 above the mean reference intensity, and O1 does not measure this mean level, so the offset is consistent with sensitivity to initialization in an unobserved direction. The fraction of reconstruction energy in the discarded set is also an incomplete diagnostic. RED-diff puts 93% of its energy in the discarded frequencies, above the 67 to 78% range of the training sources, but

Table 3: **Weak lensing and fluorescence microscopy.** Subscripts give between-image SD. †FRC grid limit. Censored FRC values enter the mean, so FRC is not ranked.

	Weak Lensing ($\sigma_{\text{noise}} = 0.01$)				Fluorescence Microscopy (BioSR)		
Method	**PSNR↑**	**SSIM↑**	**$P(\ell)$ bias↓**	**Peak χ^2 ↓**	**PSNR↑**	**SSIM↑**	**FRC res. (nm)↓**
Classical baselines							
Kaiser–Squires	$27.592_{\pm 4.269}$	$0.456_{\pm 0.226}$	$1.384_{\pm 2.051}$	$0.018_{\pm 0.020}$	–		
Richardson–Lucy			–		$25.658_{\pm 1.811}$	$0.757_{\pm 0.047}$	$127.19^{\dagger}$
Sampling-based							
DPS	$24.890_{\pm 5.450}$	$0.365_{\pm 0.244}$	$3.879_{\pm 6.669}$	$0.008_{\pm 0.008}$	$24.921_{\pm 1.828}$	$0.752_{\pm 0.057}$	$134.741_{\pm 23.635}$
DPnP	**$30.051_{\pm 2.939}$**	$0.472_{\pm 0.252}$	$0.398_{\pm 0.169}$	$0.015_{\pm 0.017}$	$24.832_{\pm 1.817}$	$0.736_{\pm 0.071}$	$134.903_{\pm 26.362}$
DAPS	$28.765_{\pm 3.394}$	$0.506_{\pm 0.198}$	$0.555_{\pm 0.749}$	$0.012_{\pm 0.013}$	$24.496_{\pm 1.795}$	$0.714_{\pm 0.068}$	$149.778_{\pm 44.095}$
DiffStateGrad	$27.266_{\pm 4.234}$	$0.407_{\pm 0.249}$	$1.384_{\pm 2.051}$	$0.020_{\pm 0.022}$	$21.599_{\pm 1.542}$	$0.633_{\pm 0.076}$	$131.644_{\pm 19.604}$
SITCOM	$28.109_{\pm 3.377}$	$0.467_{\pm 0.212}$	$0.590_{\pm 0.391}$	$0.022_{\pm 0.025}$	$23.635_{\pm 1.297}$	$0.674_{\pm 0.050}$	$266.672_{\pm 34.419}$
ABMS	$26.981_{\pm 3.540}$	$0.193_{\pm 0.351}$	$0.749_{\pm 0.188}$	$0.039_{\pm 0.030}$	$23.469_{\pm 2.127}$	$0.648_{\pm 0.077}$	$185.699_{\pm 65.579}$
Optimization-based							
RED-diff	$30.023_{\pm 2.836}$	**$0.551_{\pm 0.172}$**	**$0.388_{\pm 0.113}$**	**$0.007_{\pm 0.006}$**	**$26.559_{\pm 1.853}$**	**$0.807_{\pm 0.053}$**	$127.19^{\dagger}$
DMPlug	$28.325_{\pm 4.077}$	$0.468_{\pm 0.259}$	$0.690_{\pm 1.269}$	$0.016_{\pm 0.022}$	$23.869_{\pm 1.541}$	$0.666_{\pm 0.088}$	$145.286_{\pm 35.145}$
DMILO	$24.593_{\pm 3.009}$	$0.210_{\pm 0.207}$	$1.087_{\pm 1.591}$	$0.032_{\pm 0.018}$	$25.933_{\pm 1.917}$	$0.795_{\pm 0.061}$	$131.305_{\pm 20.172}$

DMILO and SITCOM, whose O1 errors lie 86% and 97.5% in those frequencies, stay inside that range (Appendix F.5).

Most of this share is the mean level, which the training sources predict well. After the sources are centered by the training-mean image, the energy in the discarded set falls from 72% to 3.8% of the original, uncentered source energy. A training-mean baseline, the pseudo-inverse of O1 with the discarded frequencies filled by the training mean, reaches 29.1 dB, above seven of the nine solvers. Replacing each solver's component in those frequencies by the training mean lowers PSNR for six solvers but raises it for DAPS, DPnP, and RED-diff, lifting RED-diff from 11.5 to 29.0 dB. Falling below the training-mean baseline therefore does not by itself imply that the mean would improve a solver's discarded components.

The O1R control tests whether measuring the discarded frequencies produces the gains that the error decomposition predicts. The four solvers with an O1 residual below 7% gain 10 to 36 dB, while for the five with residuals of 13% or more the decomposition predicts gains of 0.3 to 4.6 dB against measured gains of 0.2 to 3.2 dB.

5.4 Reconstruction in the low-share scientific settings

Unlike the controlled study, the scientific settings change sources, priors, and measurements together, so their rankings, shown in Fig. 1f, combine all three. Weak lensing, Galaxy, and BioSR leave only 2 to 8% of the source energy unobservable under the nominal noise levels and at most 14% under a soft threshold (Table 1 and Appendix B.3), and the sampling pattern of MRI leaves 7.3% on unacquired lines. Where the share is small, the measurement constrains most of the source energy, so solvers differ in how closely they fit the data and in what they place in the small unobservable set. On Galaxy deconvolution, Fig. 4 follows the reconstructions during inference, recording the image standard deviation $\sigma(\hat{x}_t)$ and the log measurement loss $\log_{10}(L_t/L_0)$ on ten test images that span the difficulty range. The three solvers with the highest PSNR, SITCOM, DMILO, and RED-diff, are the three whose paths run most directly toward lower measurement loss on these images. DPS and DMPlug follow more curved paths and attain lower final PSNR. Even so, a small share can hold most of the error. For the three best Galaxy solvers, 79 to 97% of the error lies in the unobservable set, whose training-estimated share is 3.4%.

Classical baselines remain competitive in these settings. Across all five, the best diffusion solver leads the best classical baseline by at most 2.5 dB, as shown in Tables 3, 4, and 5, and on Galaxy deconvolution it trails. In the other four settings, its lead ranges from 0.9 dB on BioSR to 2.5 dB on weak lensing. On noiseless Galaxy deconvolution, an untuned conjugate-gradient inverse with no prior outperforms every diffusion solver in PSNR and SSIM at all three seeing levels, while some diffusion solvers keep lower ellipticity error. On the ten controlled-study images, an evaluation

Table 4: **Galaxy PSF deconvolution.** Easy, Typical, and Hard seeing are FWHM $0.8''$, $1.18''$, and $1.8''$. Shape entries are mean/median $|\Delta e| \times 10^3$. Unless noted, results use 50 test images, except SITCOM at Hard (48) and DMPlug at Typical (41) and Hard (10). Marks rank diffusion solvers only.

Setting	Easy			Typical			Hard		
Methods	**PSNR↑**	**SSIM↑**	**$\lvert\Delta e\rvert\times10^3$ ↓**	**PSNR↑**	**SSIM↑**	**$\lvert\Delta e\rvert\times10^3$ ↓**	**PSNR↑**	**SSIM↑**	**$\lvert\Delta e\rvert\times10^3$ ↓**
Classical baselines									
Richardson–Lucy	$29.798_{\pm1.951}$	$0.769_{\pm0.062}$	0.037 / 0.021	$27.470_{\pm1.729}$	$0.576_{\pm0.073}$	0.066 / 0.055	$26.367_{\pm1.658}$	$0.429_{\pm0.082}$	0.127 / 0.089
Unregularized CG	$33.760_{\pm2.628}$	$0.915_{\pm0.035}$	0.081 / 0.048	$29.585_{\pm1.945}$	$0.765_{\pm0.063}$	0.039 / 0.038	$27.152_{\pm1.706}$	$0.552_{\pm0.073}$	0.056 / 0.042
Sampling-based									
DPS	$25.644_{\pm1.607}$	$0.463_{\pm0.069}$	0.168 / 0.165	$24.369_{\pm1.664}$	$0.356_{\pm0.070}$	0.150 / 0.119	$23.173_{\pm1.619}$	$0.267_{\pm0.068}$	1.260 / 0.796
DPnP	$25.540_{\pm1.515}$	$0.341_{\pm0.087}$	9.350 / 6.133	$25.341_{\pm1.424}$	$0.334_{\pm0.085}$	9.786 / 6.626	$25.006_{\pm1.305}$	$0.325_{\pm0.084}$	10.267 / 6.223
DAPS	$29.550_{\pm1.624}$	$0.783_{\pm0.058}$	0.077 / 0.075	$26.693_{\pm1.369}$	$0.562_{\pm0.054}$	0.069 / 0.064	$25.098_{\pm1.317}$	$0.383_{\pm0.058}$	0.083 / 0.070
DiffStateGrad	$25.194_{\pm1.525}$	$0.375_{\pm0.069}$	0.267 / 0.256	$24.329_{\pm1.655}$	$0.316_{\pm0.071}$	0.236 / 0.193	$23.021_{\pm1.532}$	$0.244_{\pm0.063}$	0.915 / 0.790
SITCOM	**$32.688_{\pm2.522}$**	**$0.888_{\pm0.045}$**	**0.010 / 0.010**	$28.340_{\pm1.726}$	**$0.697_{\pm0.063}$**	**0.021 / 0.017**	$26.346_{\pm1.734}$	**$0.485_{\pm0.073}$**	**0.036 / 0.034**
ABMS	$25.209_{\pm1.523}$	$0.337_{\pm0.073}$	0.951 / 0.937	$24.651_{\pm1.575}$	$0.307_{\pm0.075}$	0.431 / 0.390	$23.682_{\pm1.534}$	$0.261_{\pm0.071}$	0.284 / 0.222
Optimization-based									
RED-diff	$31.549_{\pm2.283}$	$0.846_{\pm0.056}$	0.021 / 0.019	$28.174_{\pm1.750}$	$0.646_{\pm0.070}$	0.029 / 0.026	$26.613_{\pm1.694}$	$0.462_{\pm0.080}$	0.040 / 0.035
DMPlug	$22.191_{\pm0.789}$	$0.204_{\pm0.032}$	4.572 / 3.234	$21.297_{\pm0.667}$	$0.171_{\pm0.022}$	2.112 / 1.736	$20.805_{\pm0.739}$	$0.154_{\pm0.023}$	1.199 / 1.109
DMILO	$31.874_{\pm2.196}$	$0.857_{\pm0.051}$	2.142 / 2.003	**$28.490_{\pm1.785}$**	$0.669_{\pm0.073}$	2.198 / 1.914	**$26.747_{\pm1.682}$**	$0.482_{\pm0.078}$	2.596 / 2.159

Table 5: **Black-hole imaging and multi-coil MRI.** Test images never used for selection. BHI closure columns give the median reduced χ^2, whose reference value is one.

(a) BHI at $r = 100\%$.

Method	Aligned PSNR↑	Blurred PSNR↑	χ^2_{cp}	χ^2_{logca}
Classical baseline				
EHT RML	$23.43_{\pm5.96}$	$29.44_{\pm8.39}$	2.61	2.14
Sampling-based				
DPS	**$25.71_{\pm4.96}$**	$32.49_{\pm7.13}$	1.40	1.27
DPnP	$19.33_{\pm2.35}$	$22.13_{\pm2.49}$	7.99	10.84
DAPS	$25.05_{\pm4.06}$	$30.83_{\pm5.64}$	1.40	1.27
DiffStateGrad	$22.97_{\pm4.05}$	$28.26_{\pm5.35}$	50.61	5.27
SITCOM	$21.61_{\pm3.76}$	$26.44_{\pm5.54}$	1.19	1.23
ABMS	$25.54_{\pm4.59}$	**$32.52_{\pm7.10}$**	1.11	1.17
Optimization-based				
RED-diff	$24.49_{\pm4.88}$	$29.52_{\pm6.83}$	1.58	1.55
DMPlug	$22.83_{\pm3.92}$	$27.73_{\pm4.95}$	2.07	1.32
DMILO	$22.08_{\pm2.69}$	$25.82_{\pm3.55}$	1.29	1.34

(b) MRI at $R = 4$.

Method	PSNR↑	SSIM↑	Full k Error↓
Classical baselines			
SENSE	$24.41_{\pm2.15}$	$0.377_{\pm0.059}$	$77.98_{\pm27.66}$
ℓ_1-wavelet	$26.40_{\pm1.55}$	$0.569_{\pm0.053}$	$71.73_{\pm23.69}$
TV	$26.36_{\pm1.39}$	$0.457_{\pm0.078}$	$70.84_{\pm22.68}$
Sampling-based			
DPS	$25.71_{\pm1.56}$	$0.534_{\pm0.123}$	$73.25_{\pm24.14}$
DPnP	$25.34_{\pm1.67}$	$0.588_{\pm0.071}$	$71.50_{\pm22.98}$
DAPS	**$28.08_{\pm1.81}$**	**$0.675_{\pm0.094}$**	$70.30_{\pm23.66}$
DiffStateGrad	$23.87_{\pm4.53}$	$0.557_{\pm0.107}$	$71.63_{\pm23.68}$
SITCOM	$20.42_{\pm4.37}$	$0.510_{\pm0.116}$	$80.13_{\pm21.79}$
ABMS	$17.99_{\pm2.96}$	$0.175_{\pm0.060}$	$94.81_{\pm21.69}$
Optimization-based			
RED-diff	$27.88_{\pm1.62}$	$0.642_{\pm0.098}$	**$70.15_{\pm23.22}$**
DMPlug	$17.12_{\pm4.77}$	$0.299_{\pm0.083}$	$91.98_{\pm16.42}$
DMILO	$23.21_{\pm2.11}$	$0.445_{\pm0.074}$	$73.50_{\pm23.03}$

reference that is exact outside $\mathcal{U}$ and zero inside it reaches 30.15 dB, above eight of the nine solvers, which motivates evaluating errors in observed and unresolved components separately.

Task-specific criteria still separate solvers within this regime. On Galaxy deconvolution, SITCOM and DMILO are within 0.15 dB in PSNR at Typical seeing, yet SITCOM's mean ellipticity error is 0.021×10^{-3} against 2.198×10^{-3}, an ordering that holds at all three seeing levels, so a shape-sensitive study would choose differently from a PSNR-based one. In weak lensing, RED-diff shares the PSNR lead with DPnP and also has the best SSIM, power-spectrum bias, and peak-distribution statistic. On real BioSR observations, a paired ablation of the Richardson–Lucy baseline shows that replacing its fitted acquisition gain with unit gain reduces PSNR by 6.04 dB (Appendix F.3), more than the 0.9 dB by which RED-diff leads it.

In MRI at $R = 4$, DAPS has the highest PSNR, 1.7 dB above the ℓ_1-wavelet baseline, and RED-diff and DAPS have the lowest full-reference k-space error, followed by total variation. In BHI, DPS, ABMS, and DAPS have the highest aligned PSNR, within 0.7 dB of one another and thus within the 1 dB standard error of a comparison that independent noise draws leave unpaired. DPS's mean closure-phase χ^2 is 22.4 against a median of 1.4, because a few of its reconstructions have large closure residuals, while DAPS has closure statistics near one.

Measurement residuals alone do not determine the error in unresolved components, so the benchmark measures that error directly where ground truth is available. For Galaxy, BioSR, and MRI we compare each solver's error in $\mathcal{U}$ with that of a zero-fill reference, which matches the source outside $\mathcal{U}$ and is zero inside it, so that its error there equals the source energy in $\mathcal{U}$. A ratio below one means that

Table 6: **Error in the unobservable set relative to a zero fill.** Mean per-image ratios, with values above one in bold. Galaxy uses the ten controlled-study images. DSG is DiffStateGrad. †Unacquired lines of the single-coil sampling model.

Setting	DPS	DPnP	DAPS	DSG	SITCOM	ABMS	RED-diff	DMPlug	DMILO
Galaxy PSF	**1.83**	0.99	**1.51**	**1.87**	0.89	**1.52**	0.95	**4.53**	0.93
BioSR	0.81	0.75	0.83	**2.32**	**1.10**	**1.08**	0.46	**1.20**	0.49
MRI†	**1.04**	0.84	0.69	0.94	**1.61**	**3.64**	0.62	**1.15**	**1.70**

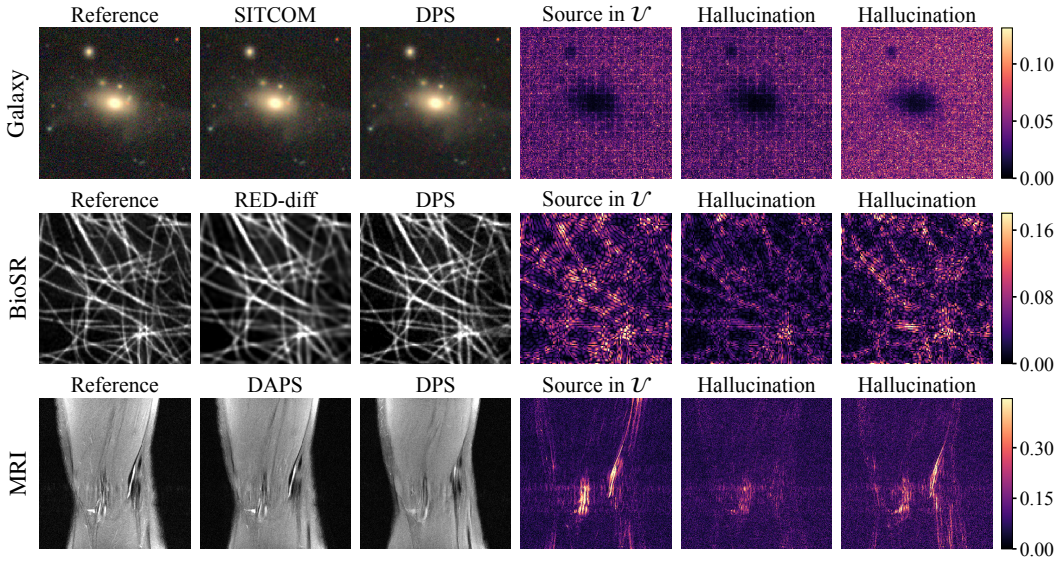


Figure 5: **Error in the unobservable set.** Reference, reconstructions of the highest-PSNR solver on the images of Table 6 and of DPS, the reference's component in $\mathcal{U}$, and both solvers' errors in $\mathcal{U}$, with one color scale per row.

the solver recovers part of the unresolved component. A ratio above one indicates greater error in $\mathcal{U}$ than the zero fill, which does not by itself rank the complete reconstruction. Squared error penalizes both missing reference structure and unsupported reconstructed structure, so the comparison favors a conservative fill. On BioSR, RED-diff and DMILO reduce this error to about half the zero-fill level, and four of the nine solvers exceed the zero fill on BioSR and five on Galaxy, as shown in Table 6 and Fig. 5. Only RED-diff and DPnP stay below the zero fill in all three settings, and SITCOM, the best solver on the Galaxy images, exceeds it on BioSR and MRI. In the Galaxy example, DPS places in $\mathcal{U}$ 80% as much energy as the reference has there, in structure uncorrelated with the source, which nearly doubles the zero-fill error, whereas SITCOM places 22%. For MRI, $\mathcal{U}$ is the set of unacquired lines, which coil encoding partly constrains, so that row measures error on those lines.

6 Conclusion

PriorShare complements measurement consistency and reference fidelity with a source-weighted measure of unresolved energy and with the error that solvers place in unresolved components. In the controlled study, residual and PSNR rankings agreed less as the share grew, and at the largest share a pseudo-inverse filled with the training mean outperformed seven of the nine solvers. Our computations assume operators diagonal in a known basis and second-order source statistics, and because O2 also differs from O1 in conditioning, the three operators show an association, with O1 and O1R as the only matched pair. Studies of diffusion priors in scientific imaging should therefore report prior share where applicable, the error in $\mathcal{U}$ relative to a zero fill where references exist, and simple non-diffusion baselines. A small residual is best read together with the share of the source that it cannot check.

Ethics statement

All datasets used here are publicly released, and the fastMRI data are de-identified by their providers. Diffusion-prior reconstructions can contain plausible structure that is absent from the source, and the benchmark is intended to help detect such errors. None of the reconstructions is intended for clinical or scientific decisions without independent validation.

Reproducibility and code availability

Task construction, prior training, observation models, and evaluator definitions are documented in Appendix C, and solver procedures and configurations appear in Appendices D and E. The benchmark separates solvers, operators, datasets, and evaluators. Selected configurations for all task–method pairs are listed in Appendix E. The code, which will be released with the paper, includes runnable configurations for the BioSR study and the Galaxy operator-transfer study, automated tests, and an end-to-end Galaxy reconstruction workflow, and exposes modular interfaces for adding solvers, forward operators, and evaluators.

AI use statement

Generative AI tools assisted with feedback on experimental design and methodology, implementation and debugging of experiment and analysis code, result analysis, manuscript drafting and editing, figure and table preparation, and LaTeX formatting. The authors verified AI-assisted analyses and all reported numbers against saved experiment records and tests and take responsibility for the final text, code, results, and claims.

References

Ismail R. Alkhouri, Shijun Liang, Cheng-Han Huang, et al. SITCOM: Step-wise triple-consistent diffusion sampling for inverse problems. In *International Conference on Machine Learning (ICML)*, 2025.

Harshith Bachimanchi and Giovanni Volpe. Diffusion models for super-resolution microscopy: A tutorial. *Journal of Physics: Photonics*, 7(1):013001, 2025. doi: 10.1088/2515-7647/ada101.

Sayantan Bhadra, Varun A. Kelkar, Frank J. Brooks, and Mark A. Anastasio. On hallucinations in tomographic image reconstruction. *IEEE Transactions on Medical Imaging*, 40(11):3249–3260, 2021. doi: 10.1109/TMI.2021.3077857.

Andrew A. Chael, Michael D. Johnson, Katherine L. Bouman, Lindy L. Blackburn, Kazunori Akiyama, and Ramesh Narayan. Interferometric imaging directly with closure phases and closure amplitudes. *The Astrophysical Journal*, 857(1):23, 2018.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, et al. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Tiangang Cui, James Martin, Youssef M. Marzouk, Antti Solonen, and Alessio Spantini. Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Problems*, 30(11):114015, 2014.

Alexander Denker, Johannes Hertrich, and Sebastian Neumayer. A stability benchmark of generative regularizers for inverse problems, 2026. URL `https://arxiv.org/abs/2605.10076`.

Arjun Dey, David J. Schlegel, Dustin Lang, et al. Overview of the DESI legacy imaging surveys. *The Astronomical Journal*, 157(5):168, 2019. doi: 10.3847/1538-3881/ab089d.

Nina M. Gottschling, Vegard Antun, Anders C. Hansen, and Ben Adcock. The troublesome kernel: On hallucinations, no free lunches, and the accuracy-stability tradeoff in inverse problems. *SIAM Review*, 67(1):73–104, 2025. doi: 10.1137/23M1568739.

David Iagaru, Nina M. Gottschling, Anders C. Hansen, and Josselin Garnier. On hallucinations in inverse problems: Fundamental limits and provable assessment methods. arXiv:2605.13146, 2026.

Pengfei Jin, Yiqi Tian, Kailong Fan, Bingjie Qi, and Quanzheng Li. Hallucination-aware diffusion sampling for inverse problems via robust prior updates. arXiv:2606.02331, 2026.

Nick Kaiser and Gordon Squires. Mapping the dark matter with weak gravitational lensing. *The Astrophysical Journal*, 404:441–450, 1993.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Henry W. Leung and Jo Bovy. Galaxy10 DECaLS, 2024. Dataset, Zenodo, doi:10.5281/zenodo.10845026.

Leon B. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical Journal*, 79:745–754, 1974.

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.

Morteza Mardani, Jiaming Song, Jan Kautz, et al. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Klaas P. Pruessmann, Markus Weiger, Markus B. Scheidegger, and Peter Boesiger. SENSE: Sensitivity encoding for fast MRI. *Magnetic Resonance in Medicine*, 42(5):952–962, 1999.

Chang Qiao and Di Li. BioSR: A biological image dataset for super-resolution microscopy, 2024. Dataset, version 9, figshare, doi:10.6084/m9.figshare.13264793.v9.

Chang Qiao, Di Li, Yuting Guo, et al. Evaluation and development of deep neural networks for image super-resolution in optical microscopy. *Nature Methods*, 18:194–202, 2021. doi: 10.1038/s41592-020-01048-5.

Minsi Ren, Wenhao Deng, Ruiqi Feng, et al. One step further with Monte-Carlo sampler to guide diffusion better. In *International Conference on Learning Representations (ICLR)*, 2026.

William Hadley Richardson. Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America*, 62(1):55–59, 1972.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, et al. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Alessio Spantini, Antti Solonen, Tiangang Cui, James Martin, Luis Tenorio, and Youssef Marzouk. Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM Journal on Scientific Computing*, 37(6):A2451–A2487, 2015.

Mike Walmsley, Chris Lintott, Tobias Géron, et al. Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2022. doi: 10.1093/mnras/stab2093.

Hengkang Wang, Xu Zhang, Taihui Li, et al. DMPlug: A plug-in method for solving inverse problems with diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jure Zbontar, Florian Knoll, Anuroop Sriram, et al. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018. arXiv:1811.08839.

Bingliang Zhang, Wenda Chu, Julius Berner, et al. Improving diffusion inverse problem solving with decoupled noise annealing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, et al. InverseBench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yang Zheng, Wen Li, and Zhaoqiang Liu. Integrating intermediate layer optimization and projected gradient descent for solving inverse problems with diffusion models. In *International Conference on Machine Learning (ICML)*, 2025b.

Rayhan Zirvi, Bahareh Tolooshams, and Anima Anandkumar. Diffusion state-guided projected gradient for inverse problems. In *International Conference on Learning Representations (ICLR)*, 2025.

Appendix contents

A Related work **13**

B Observable and unobservable directions **14**

B.1 Proofs of the propositions 14

B.2 Computing the prior share 14

B.3 Sensitivity of the prior-share estimate 15

C Implementation details and scientific task construction **15**

C.1 Per-task data and preprocessing 15

C.2 Diffusion prior training 16

C.3 Forward operators and observation models 16

C.4 Evaluation settings 18

C.5 Scientific metric definitions 19

D Solver attributes and pseudocode **21**

D.1 Solver attributes 21

D.2 Pseudocode 21

E Hyperparameter configurations and search landscapes **26**

E.1 Selection protocol 26

E.2 Search space and selected configurations 26

E.3 Per-task search landscapes 28

F Supplementary results and analyses **31**

F.1 Multi-coil MRI: configurations 31

F.2 Domain-standard baseline details 31

F.3 BioSR affine calibration and downstream RL reconstruction 31

F.4 Hallucination in the scientific settings 31

F.5 Energy in the discarded directions 33

F.6 Resampling the controlled study 33

G Scope and interpretation of scientific metrics **35**

H Qualitative reconstruction galleries **36**

A Related work

Diffusion solvers for inverse problems are usually evaluated on natural-image degradations with specified operators (Kawar et al., 2022; Chung et al., 2023; Mardani et al., 2024; Wang et al., 2024). InverseBench (Zheng et al., 2025a) established a physical-science evaluation with conventional baselines, and Denker et al. (2026) study shifts in the data distribution and errors in the forward or noise model. Hallucination has been tied to the kernel of the forward operator. Bhadra et al. (2021) split a reconstruction's error into null-space and measurement components, Gottschling et al. (2025)

relate hallucinations to the kernel through an accuracy–stability trade-off, Iagaru et al. (2026) bound their magnitude from the forward model alone, and Jin et al. (2026) modify the prior update of diffusion samplers to limit them. Likelihood-informed subspaces (Cui et al., 2014; Spantini et al., 2015) weigh operator directions by prior variance to reduce the dimension of Bayesian inference. The share uses the uncentered counterpart of this weighting, which also counts the mean, to report how much source energy a measurement leaves to the prior and to read solver errors against it.

B Observable and unobservable directions

This appendix proves the two propositions of Section 2 and states how the prior share of Eq. 2 is computed for each task.

B.1 Proofs of the propositions

Throughout, $\mathcal{A} = \sum_i s_i u_i v_i^\top$ is the singular decomposition of a linear acquisition model, $\mathcal{U}$ is the subspace of Eq. 1, and $s_{\min} = \min\{s_i : v_i \notin \mathcal{U}\}$. Because $\mathcal{U}$ is spanned by right singular vectors, $\mathcal{A}$ maps $\mathcal{U}$ and $\mathcal{U}^\perp$ to orthogonal subspaces, so for any w

$$\|\mathcal{A}w\|^2 = \|\mathcal{A}P_{\mathcal{U}}w\|^2 + \|\mathcal{A}P_{\mathcal{U}^\perp}w\|^2 \;\geq\; s_{\min}^2 \|P_{\mathcal{U}^\perp}w\|^2. \tag{3}$$

Proof of Proposition 1. Let $d = \hat{x}_1 - \hat{x}_2$. The triangle inequality gives $\|\mathcal{A}d\| \leq \|\mathcal{A}\hat{x}_1 - y\| + \|y - \mathcal{A}\hat{x}_2\| \leq 2\varepsilon$, and Eq. 3 gives $\|\mathcal{A}d\| \geq s_{\min}\|P_{\mathcal{U}^\perp}d\|$. Combining the two yields $\|P_{\mathcal{U}^\perp}d\| \leq 2\varepsilon/s_{\min}$. For $v_i \in \mathcal{U}$ with $s_i > 0$ the same identity gives $|\langle v_i, d\rangle| \leq \|\mathcal{A}d\|/s_i \leq 2\varepsilon/s_i$, and since $s_i^2\mathcal{S}_i < \sigma^2$ this exceeds $\sqrt{\mathcal{S}_i}$ whenever $\varepsilon \geq \sigma/2$. On the null space nothing constrains d: replacing $\hat{x}_2$ by $\hat{x}_2 + u$ with $\mathcal{A}u = 0$ leaves both residuals unchanged. □

Proof of Proposition 2. With $e = \hat{x} - x$ and $y = \mathcal{A}x + \xi$, the residual is $\mathcal{A}\hat{x} - y = \mathcal{A}e - \xi$. Hence $\|\mathcal{A}\hat{x} - y\| \geq \|\mathcal{A}e\| - \|\xi\| \geq s_{\min}\|P_{\mathcal{U}^\perp}e\| - \|\xi\|$ by Eq. 3. For the second claim, perturbing $\hat{x}$ by u changes the residual by $\mathcal{A}u$. If $\mathcal{A}u = 0$ the residual is unchanged. If $u = c\,v_i$ with $v_i \in \mathcal{U}$ and $c^2 \leq \mathcal{S}_i$, then $\|\mathcal{A}u\|^2 = s_i^2c^2 \leq s_i^2\mathcal{S}_i < \sigma^2$ by Eq. 1. The bound holds per direction; a perturbation spread over many directions of $\mathcal{U}$ can change the residual by more than σ in total. □

B.2 Computing the prior share

Weak lensing, Galaxy, BioSR, and the controlled operators O1, O1R, and O2 are diagonal in the Fourier basis, exactly for the FFT-based operators and under the periodic approximation for the padded convolutions of Galaxy and BioSR, which ignores boundary handling. For these, Eq. 2 becomes a sum of the training power spectrum over the unobservable index set. For MRI we compute the share of the single-coil sampling model $M\mathcal{F}$, which ignores coil-sensitivity encoding and therefore describes the sampling pattern rather than the multi-coil operator. Black-hole imaging and O3 are nonlinear and have no share. We estimate $\mathcal{S}$ as the mean periodogram of training sources, never test sources, in an orthonormal transform so that per-mode noise equals the per-pixel variance and energy shares are exact under Parseval. Weak lensing, Galaxy, and BioSR use their full training splits of $10{,}240$, $17{,}631$, and $15{,}000$ images. Subsets are avoided because the weak-lensing training file is ordered by simulation, and its first maps are systematically lower in amplitude, which at a fixed noise level would roughly double the share. MRI uses the validation split with 6 volumes, as the released data ship no separate training LMDB, and validation is disjoint from test.

The noiseless operators need a convention, since $s_i^2\mathcal{S}_i < \sigma^2$ is vacuous at $\sigma = 0$. We define the unobservable set as $\{s_i = 0\} \cup \{s_i^2\mathcal{S}_i < \sigma^2\}$, which reduces to the exact null set when $\sigma = 0$, and we additionally report the share under an 8-bit quantization floor $\sigma = 1/(255\sqrt{12})$ representing the precision at which the stored images themselves are meaningful. The floor applies to the Galaxy sources, which are stored as 8-bit images, and Tables 1 and 2 report the share under it. Under the exact null set the share is 0 for Moffat blur at every seeing level and for O1R, and 72.2% for O1, so the contrasts in Section 2 hold under either convention. MRI is computed at $\sigma = 0$, where the unobservable set of the single-coil model is the unacquired k-space.

For the Kaiser–Squires cross-operator, taking the real part after the inverse transform is not an orthogonal projection. Since the source is real and the kernel satisfies $K(-\boldsymbol{\ell}) = K(\boldsymbol{\ell})$ away from the

Nyquist row and column, the operator stays diagonal with multiplier $M(\boldsymbol{\ell}) = \big(K(\boldsymbol{\ell}) + \overline{K(-\boldsymbol{\ell})}\big)/2$, which equals $(\ell_1^2 - \ell_2^2)/|\boldsymbol{\ell}|^2$ off those lines and retains an imaginary part on them. Its exact zero set is the DC bin together with the two diagonals $|\ell_1| = |\ell_2|$, which is 510 of the 65536 frequencies at 256×256.

B.3 Sensitivity of the prior-share estimate

Table 7 reports how each share moves when the assumed noise level is halved or doubled, under a soft threshold that replaces the hard cut $|H|^2\mathcal{S} < \sigma^2$ by the per-frequency linear minimum-mean-square error $\mathcal{S}\sigma^2/(|H|^2\mathcal{S} + \sigma^2)$, and on the test images themselves. Recomputing the spectrum from two disjoint halves of the training data changes no share by more than 0.13 percentage points. Exact nulls are insensitive to all of these choices, so O1 stays at 72%. Noise-limited shares are not, and weak lensing ranges from 4.4% to 21.8% as its noise level is halved or doubled. MRI, computed without noise, rises to at most 7.5% at $R = 4$ and 14.2% at $R = 8$ with the thermal noise estimated from the corners of fully sampled k-space. The contrasts drawn in Section 2 hold under every variant.

Table 7: **Precision of the prior share.** Shares in percent under the nominal noise level, half and twice that level, the soft threshold, and for the test images of each task.

Setting	Nominal	$\sigma/2$	2σ	Soft	Test images
Weak lensing, $\sigma = 0.01$	7.98	4.40	21.78	13.97	9.53
Weak lensing, $\sigma = 0.04$	38.99	21.78	59.84	41.24	41.35
Galaxy PSF, $0.8''$	1.93	1.45	2.46	1.99	–
Galaxy PSF, $1.18''$	3.44	2.89	4.16	3.55	3.23
Galaxy PSF, $1.8''$	5.49	5.03	5.93	5.46	–
BioSR	6.36	4.53	8.95	7.02	7.29
MRI, $R = 4$, single-coil	7.33	–	–	7.33	4.46
MRI, $R = 8$, single-coil	14.01	–	–	14.01	8.35
O1	72.25	–	–	72.31	–
O1R	0.049	–	–	0.115	–

C Implementation details and scientific task construction

C.1 Per-task data and preprocessing

Weak lensing. Convergence fields are taken from the Weak Lensing Mass Mapping project, comprising five independent N-body simulation runs. Four runs form the training set and the held-out fifth run forms the test set. Each run yields 40 redshift slices cropped into 64 non-overlapping 128×128 single-channel patches. Global min–max normalization maps physical κ to $[-1, 1]$ via $\tilde{x} = 2(x - x_{\min})/(x_{\max} - x_{\min}) - 1$, with $(x_{\min}, x_{\max}) \approx (-0.1200, 1.2870)$. Training augmentation uses horizontal and vertical flips at probability 0.5.

Galaxy PSF deconvolution. Galaxy10 DECaLS (Leung & Bovy, 2024; Walmsley et al., 2022; Dey et al., 2019) provides 17,736 three-band $g/r/z$ images at 256×256 from the DESI Legacy Imaging Surveys. Morphological labels are not used. Images from the official release are normalized as $\tilde{x} = x/127.5 - 1$. The split with random seed 42 gives 17,631 training, 5 validation, and 100 test images. The main Galaxy evaluation uses 50 test images; the fixed-source cross-operator study uses 10.

Fluorescence microscopy. BioSR (Qiao & Li, 2024; Qiao et al., 2021) provides paired wide-field and higher-resolution SIM images across cell structures and signal levels. The prior is trained on the Microtubules SIM targets, without conditioning on WF observations. Each processed 128×128 SIM image is stored as a 32-bit TIFF and normalized per image to $[-1, 1]$, using output zero for constant images. Reconstruction uses the fixed intensity decoder below, not the test SIM minimum and maximum. All methods are evaluated on the same 50 held-out test pairs under the frozen acquisition model.

Black-hole imaging and multi-coil MRI. The priors are reused from InverseBench (Zheng et al., 2025a). BHI uses GRMHD source simulations; MRI uses fastMRI knee data (Zbontar et al., 2018). Architecture and source split sizes are recorded in Table 8. The MRI full-k-space evaluation and displayed baseline comparison use the same test sources and masks as the corresponding diffusion results.

C.2 Diffusion prior training

All tasks use unconditional diffusion priors following the EDM framework (Karras et al., 2022), with the prior interface adopted from InverseBench (Zheng et al., 2025a). The denoiser takes the form

$$D_\theta(x,\sigma) = c_{\text{skip}}(\sigma)x + c_{\text{out}}(\sigma)F_\theta(c_{\text{in}}(\sigma)x, c_{\text{noise}}(\sigma)), \tag{4}$$

with preconditioning coefficients

$$c_{\text{skip}} = \frac{\sigma_{\text{data}}^2}{\sigma^2+\sigma_{\text{data}}^2}, \quad c_{\text{out}} = \frac{\sigma\sigma_{\text{data}}}{\sqrt{\sigma^2+\sigma_{\text{data}}^2}}, \quad c_{\text{in}} = \frac{1}{\sqrt{\sigma^2+\sigma_{\text{data}}^2}}, \quad c_{\text{noise}} = \frac{\log\sigma}{4}.$$

Training minimizes

$$\mathcal{L} = w(\sigma)\|D_\theta(x+n;\sigma)-x\|_2^2, \qquad w(\sigma) = \frac{\sigma^2+\sigma_{\text{data}}^2}{(\sigma\sigma_{\text{data}})^2}, \quad n\sim\mathcal{N}(0,\sigma^2 I).$$

The noise scale is sampled as $\log\sigma \sim \mathcal{N}(P_{\text{mean}}, P_{\text{std}}^2)$ with $P_{\text{mean}} = -1.2$, $P_{\text{std}} = 1.2$, and $\sigma_{\text{data}} = 0.5$ unless noted otherwise. Shared training defaults are Adam, learning rate 10^{-4}, 100,000 steps with 500-step warmup, and EMA half-life corresponding to 204,800 images. The default standalone EDM sampler uses 40 steps, $\sigma_{\min} = 0.002$, $\sigma_{\max} = 80$, and $\rho = 7.0$; the inverse solvers use the method-specific schedules in Appendix E. Table 8 records task-specific specifications. Its test-split sizes are dataset sizes, not the number of reconstructions contributing to every result row.

Table 8: **Task-specific prior specifications.** BHI and MRI checkpoints are reused from InverseBench without modification. Training/test sizes describe the source splits; evaluated subsets are specified separately.

	Weak lensing	**Galaxy PSF**	**BioSR**	**BHI**	**MRI**
Resolution / channels	128^2 / 1	256^2 / 3	128^2 / 1	64^2 / 1	320^2 / 2
Training / test images	10,240 / 2,560	17,631 / 100	15,000 / 100	50,000 / 100	25,012 / 94
U-Net base channels	64	128	128	128	128
Channel multipliers	$[1,2,4]$	$[1,1,2,2,2]$	$[1,1,2,2]$	$[1,2,2]$	$[1,1,1,2,2]$
Attn. blocks (enc/dec)	0	5	4	3	5
Res. blocks per resolution	enc $[1,1]$, mid 2, dec $[1,1]$	2	2	1	1
Self-attention	none	16^2	16^2	$\{16\}$	$\{16\}$
Parameters	4.72 M	46.4 M	34.1 M	20.0 M	26.0 M
Batch size	64	24	24	256	64
Learning rate	2×10^{-4}	10^{-4}	10^{-4}	10^{-4}	10^{-4}
Gradient clipping	1.0	none	none	none	none
Training steps	100,000	100,000	100,000	50,000	100,000
σ_{data}	0.0361	0.5	0.5	0.5	0.5
GPU	V100 32 GB	V100 32 GB	V100 32 GB	V100 32 GB	V100 32 GB

C.3 Forward operators and observation models

Weak lensing. The prior output $x \in [-1,1]^{H\times W}$ is decoded to physical convergence as

$$\kappa = x_{\min} + \tfrac{1}{2}(x+1)(x_{\max}-x_{\min}), \qquad (x_{\min}, x_{\max}) \approx (-0.1200, 1.2870).$$

The Kaiser–Squires kernel is diagonal in the Fourier domain,

$$\widehat{\gamma}(\boldsymbol{\ell}) = P(\boldsymbol{\ell})\widehat{\kappa}(\boldsymbol{\ell}), \qquad P(\boldsymbol{\ell}) = \frac{\ell_1^2-\ell_2^2+2i\ell_1\ell_2}{\ell_1^2+\ell_2^2}, \quad P(\mathbf{0}) = 0. \tag{5}$$

where $\boldsymbol{\ell}$ indexes the two-dimensional Fourier grid, and the two-component observation model is

$$\mathcal{A}(\kappa) = \left[\Re\mathcal{F}^{-1}(P \odot \mathcal{F}\kappa),\ \Im\mathcal{F}^{-1}(P \odot \mathcal{F}\kappa)\right], \qquad y = \mathcal{A}(\kappa) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_{\text{noise}}^2 I).$$

For nonzero frequencies $|P| = 1$, while $P(\mathbf{0}) = 0$ leaves the mean unobserved. The analytic inverse has Fourier representation $\widehat{\kappa}_{\text{KS}} = P^*\widehat{\gamma}$. DiffStateGrad uses its normalized output as initialization; other solvers retain their method-specific noise, prior, latent, or zero initializations. Reconstruction uses $\sigma_{\text{noise}} = 0.01$, and $\sigma_{\text{noise}} = 0.04$ enters only the prior-share comparison of Section 2. Fig. 8a shows a convergence map with its clean and noisy shear. The data repository also contains observations at 0.05 and 0.1336. Patches have 5.0° field of view at a fixed source-redshift bin. The diagnostic O1 in Section 5.1 is a separate Galaxy wrapper; its label should not be taken as asserting equivalence to this two-component acquisition.

Galaxy PSF deconvolution. The normalized Moffat kernel

$$h(i,j) = \frac{1}{Z}\left(1 + \frac{i^2 + j^2}{\alpha^2}\right)^{-\beta}, \qquad \alpha = \frac{\text{FWHM}}{s\,2\sqrt{2^{1/\beta} - 1}}. \tag{6}$$

is convolved independently with the three channels. The integer indices (i, j) are offsets from the kernel center; Z gives unit sum. We use $\beta = 2.5$, pixel scale $s = 0.262$ arcsec/pixel, and FWHM in $\{0.8, 1.18, 1.8\}$ arcsec, which Fig. 7b shows for one source. The measurements contain no additional synthetic noise. Data-consistency gradients use the transposed-convolution adjoint. Because the reference is an observed survey image, this task measures recovery from the additional specified degradation rather than inversion of the complete survey acquisition process.

Fluorescence microscopy. The observation is the paired WF image at the highest signal level. For normalized prior output, the frozen intensity decoder is

$$u = (x + 1)/2, \qquad x_{\text{raw}} = a + s_{\text{int}}u, \qquad (a, \log s_{\text{int}}) = (0.0011476, -0.0541517),$$

with $s_{\text{int}} = 0.9473$. A reflection-padded Gaussian PSF, global gain, and background define

$$\mathcal{A}(x) = g\big(G_{\sigma,k} *_{\text{ref}} x_{\text{raw}}\big) + b.$$

The intensity map and optical parameters are fitted separately. Constant, quantile-based, and ridge intensity policies are fitted on 64 official validation pairs; a disjoint set of 16 pairs selects the constant policy by normalized raw-intensity decoding MSE. The policy is frozen before test evaluation. Optical fitting uses a coarse grid over $k \in \{11, 15, 19\}$ and $\sigma \in \{1.2, 1.4, \ldots, 2.8\}$ on 1,024 training pairs with closed-form least squares for g, b, a finer σ search on 4,096 pairs, and final gain/background fitting on 8,192 pairs. The selected model is

$$\mathcal{A}(x) = 1.5054\big(G_{1.9,19} *_{\text{ref}} x_{\text{raw}}\big) + 0.0321.$$

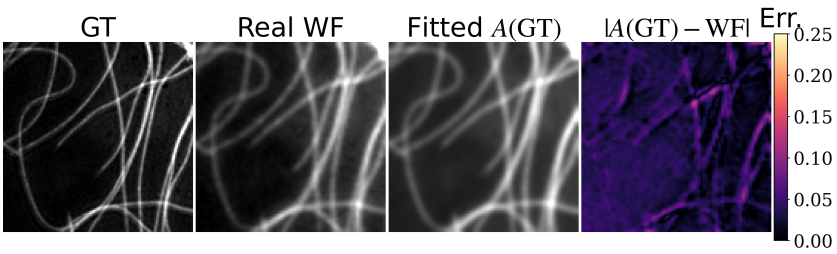
The Gaussian L2 data term uses $\sigma_y^2 = 10^{-2}$ as its scale parameter. The fitted model has validation MSE 3.01×10^{-3} and correlation 0.9476 with real WF observations. Test SIM values are used only by PSNR, SSIM, and FRC evaluation. Figure 6 shows calibration examples.

Black-hole imaging. Following the closure construction in InverseBench (Zheng et al., 2025a), station-pair visibilities are $V_{ij} = g_i g_j^* \widetilde{x}(u_{ij}, v_{ij}) + n_{ij}$, where g_i are station gains and n_{ij} is thermal noise. The observables are closure phases and log closure amplitudes,

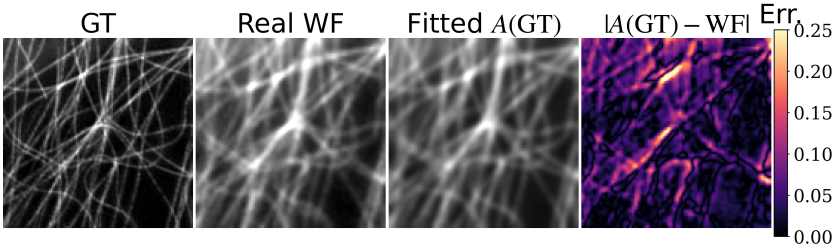
$$\psi_{ijk} = \arg V_{ij} + \arg V_{jk} + \arg V_{ki}, \qquad \ell_{ijkl} = \log \frac{|V_{ij}|\,|V_{kl}|}{|V_{ik}|\,|V_{jl}|}.$$

These nonlinear quantities remove station gain dependence and are invariant to source translation. We therefore report translation-aligned PSNR. We use the full observation time, $r = 100\%$. The dirty image in Fig. 7a visualizes the sampled visibilities only; it is not used in the closure-domain data term.

Multi-coil MRI. The MRI model is $y_j = M\mathcal{F}(S_j x)$ for $j = 1, \ldots, C$, with estimated sensitivity maps S_j and a Cartesian sampling mask M at acceleration $R = 4$. Every solver uses the k-space likelihood $\sum_j \|M\mathcal{F}(S_j\hat{x}) - y_j\|^2$ through its own update. The full-reference metric evaluates $\|\mathcal{F}(S\hat{x}) - k_{\text{full}}\|_2$ after reconstruction, so the fully sampled test reference does not enter inference. Fig. 8b shows the sampling masks at $R = 4$ and at $R = 8$, which enters only the prior-share comparison.

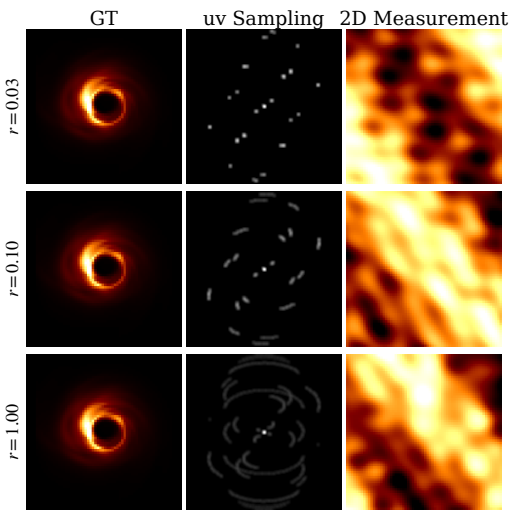


(a) Validation sample 1.

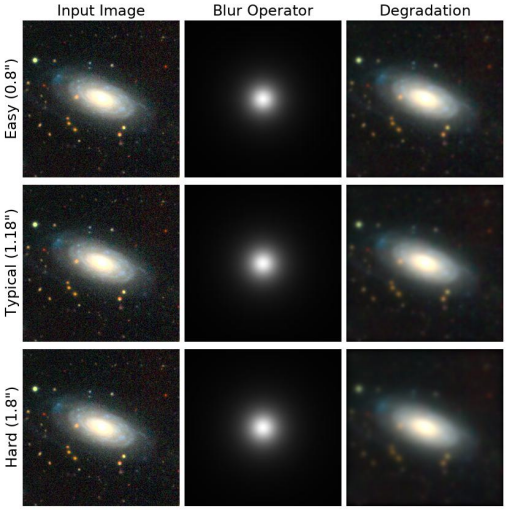


(b) Validation sample 2.

Figure 6: **BioSR forward-operator calibration.** Left to right: SIM reference, real wide-field observation, calibrated forward prediction $A(\mathrm{GT})$, and absolute error map.



(a) Black-hole imaging.

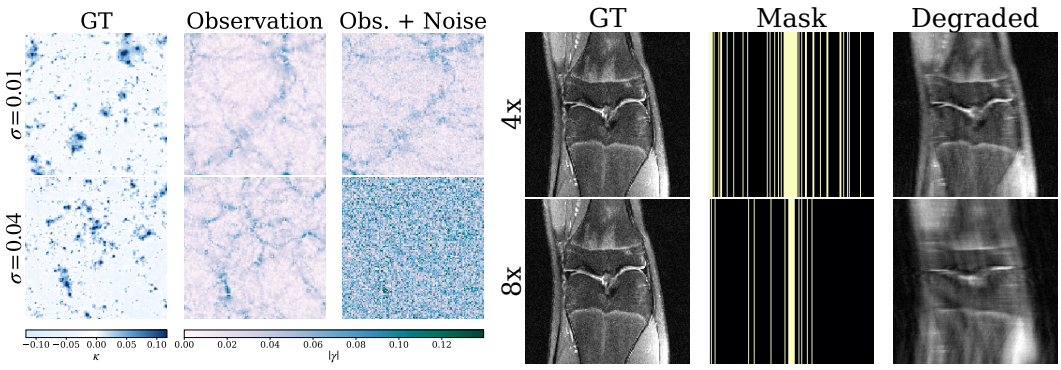


(b) Galaxy PSF degradation.

Figure 7: **Measurement construction for BHI and Galaxy.** The BHI dirty image is an illustration, not the observation used in reconstruction. Galaxy examples show the three controlled seeing levels.

C.4 Evaluation settings

Weak-lensing evaluation uses a fixed held-out test subset at noise 0.01, shared by all nine methods and the unsmoothed Kaiser–Squires baseline. Its fifty maps come from the 1941 of 2560 test patches whose target standard deviation is at least 0.005. An initial random draw of fifty contained ten near-flat patches, which we noticed after a DPS run returned strongly negative PSNR on several of them; they were replaced by a seeded random draw from the eligible pool. The rule depends only on the target maps, but the check that prompted it was a solver result. Galaxy evaluation uses three seeing levels, 0.8, 1.18, and 1.8 arcsec. PSNR and SSIM describe full-image recovery, while ellipticity error is conditional on successful adaptive-moment measurement as defined below. These shape



(a) Weak lensing.

(b) Multi-coil MRI.

Figure 8: **Measurement construction for weak lensing and MRI.** Convergence-to-shear mapping with noise, and Cartesian subsampling at two MRI accelerations.

measurements are benchmark image statistics, not a direct evaluation of downstream cosmological parameter inference.

For BioSR, the paired highest-signal WF observations are reconstructed with the frozen intensity and optical calibration above. No test-reference intensity range is supplied to the reconstructor. FRC is computed between reconstruction and SIM reference by radially averaging the normalized cross-spectrum, smoothing with a one-ring Gaussian kernel, and taking its first crossing q_* below $1/7$. We report $r_{\mathrm{FRC}} = (2r_{\max}/q_*)p$, with $r_{\max} = 64$ and processed pixel size $p = 62.6$ nm. The source SIM pixel size is 31.3 nm (Qiao & Li, 2024) and the processed images are downsampled by two (Bachimanchi & Volpe, 2025). If there is no crossing, $q_* = 63$ gives 127.19 nm, marked by a dagger. Equal censored readouts do not establish equal physical resolution.

BHI is evaluated at $r = 100\%$ using aligned PSNR, Gaussian-blurred PSNR at $15\,\mu$as FWHM, and the reduced χ^2 of closure phases and log closure amplitudes. MRI is evaluated with a random Cartesian mask at $R = 4$. We report PSNR, SSIM, and full k-space error against the fully sampled test reference. The latter includes unacquired coefficients and is distinct from observational self-consistency.

C.5 Scientific metric definitions

Weak-lensing statistics. For image x, define radial power in integer-frequency ring q as

$$P_x(q) = \frac{1}{|\mathcal{R}_q|} \sum_{\boldsymbol{k} \in \mathcal{R}_q} |\widehat{x}(\boldsymbol{k})|^2, \qquad \mathcal{R}_q = \{\boldsymbol{k} : \lfloor \|\boldsymbol{k}\|_2 \rfloor = q\}.$$

The DC ring is excluded. Valid bins satisfy $P_{\mathrm{GT}}(q) > 10^{-4} \max_{q'} P_{\mathrm{GT}}(q')$. The reported power bias is

$$\operatorname*{median}_{q \in \mathcal{V}} \frac{|P_{\mathrm{rec}}(q) - P_{\mathrm{GT}}(q)|}{|P_{\mathrm{GT}}(q)| + 10^{-12}}.$$

Peaks are pixels no smaller than their eight neighbors. We form 20 linear bins between the 1st and 99th percentiles of reference peak amplitudes, clip both sets of peaks to that interval, and normalize their histograms to probability vectors p, g. The peak statistic is

$$\chi^2_{\mathrm{peak}} = \frac{1}{20} \sum_{b=1}^{20} \frac{(p_b - g_b)^2}{p_b + g_b + 10^{-12}}.$$

Because the histograms are normalized, this statistic measures their relative distribution, not the total number of peaks.

Galaxy ellipticity. Images are clipped to $[0, 1]$ and converted to luminance $I = 0.299R + 0.587G + 0.114B$. GalSim HSM adaptive moments at pixel scale 0.262 arcsec/pixel give distortion components

(e_1, e_2). The error is

$$|\Delta e| = \sqrt{(e_{1,\mathrm{rec}} - e_{1,\mathrm{GT}})^2 + (e_{2,\mathrm{rec}} - e_{2,\mathrm{GT}})^2}.$$

A sample is valid only if both reference and reconstruction measurements succeed without error. The table reports mean and median over those valid measurements; these conditional summaries are not a measure of shape-measurement success rate.

D SOLVER ATTRIBUTES AND PSEUDOCODE

D.1 SOLVER ATTRIBUTES

Table 9: Technical attributes of the nine evaluated methods, grouped into sampling-based and optimization-based paradigms. The rightmost columns flag subspace projection and inner optimization.

Method	Operating space	Inner mechanism	Likelihood integration	Subspace	Inner loop
Sampling-based					
DPS (Chung et al., 2023)	Image	None	Tweedie gradient	✗	✗
DAPS (Zhang et al., 2025)	Image	Langevin per σ	Direct likelihood gradient	✗	✓
ABMS (Ren et al., 2026)	Image	Finite stochastic DDIM directions	Monte Carlo noise-space gradient	✗	✗
DiffStateGrad (Zirvi et al., 2025)	Image	Travel + pixel/latent AdamW	Projected likelihood + ReSample	✓	✓
SITCOM (Alkhouri et al., 2025)	Image	Per-level latent Adam	Denoise–fit–re-noise	✗	✓
DPnP (Xu & Chi, 2024)	Image	Exact-OU likelihood Langevin	Alternating PCS + conditional DDS	✗	✓
Optimization-based					
RED-diff (Mardani et al., 2024)	Image	None	Score regularization	✗	✗
DMPlug (Wang et al., 2024)	Initial latent z_0	Full-trajectory backprop	Reconstruction loss	✗	✗
DMILO (Zheng et al., 2025b)	Per-step latent z_t	Per-step inner Adam	Reconstruction loss	✗	✓

Table 9 records the implementation attributes of the nine solvers. The pseudocode below specifies the EDM interface and method-specific updates; the shared prior leaves each solver its own schedule, gradient path, and computational cost, which is what the comparison varies.

D.2 PSEUDOCODE

Shared helpers. The EDM scheduler emits noise levels $\{\sigma_i\}$, input scales $\{s_i\}$, output scales $\{c_i\}$, and step factors $\{f_i\}$. For the source implementation, the shared score conversion and step helpers are

$$\begin{aligned}
\mathrm{score}_i(x_i, \hat{x}_0) &= (\hat{x}_0 - x_i/s_i)/(\sigma_i^2 s_i), \\
\mathrm{EDMStep}(x_i, \hat{x}_0; i, \mathrm{mode}) &= \begin{cases} c_i x_i + f_i\,\mathrm{score}_i + \sqrt{f_i}\,\varepsilon, & \mathrm{mode} = \mathrm{SDE}, \\ c_i x_i + (f_i/2)\,\mathrm{score}_i, & \mathrm{mode} = \mathrm{ODE}, \end{cases} \qquad \text{(D.1)} \\
\mathrm{EulerMeanStep}(D_\theta, x_i, i) &= \big(\mathrm{EDMStep}(x_i, \hat{x}_0; i, \mathrm{ODE}), \hat{x}_0, \mathrm{info}\big), \\
&\qquad \hat{x}_0 = D_\theta(x_i/s_i; \sigma_i), \qquad \text{(D.2)} \\
\mathrm{Sampler}(D_\theta, z; \text{sub-sched}) &= \text{deterministic iteration of (D.1) from } x_0 = z, \qquad \text{(D.3)} \\
\mathrm{Denoise}(D_\theta, x, \sigma, s) &= D_\theta(x/s; \sigma). \qquad \text{(D.4)}
\end{aligned}$$

Here $\varepsilon \sim \mathcal{N}(0, I)$. The returned info contains next-step noise and scaling values. Adam-step denotes an Adam update with default $\beta = (0.9, 0.999)$ unless specified otherwise; Aloss denotes the task-specific data term. The finite stochastic DDIM update for ABMS below is specified separately from the generic helper.

Algorithm 1 DPS (Chung et al., 2023): per-step Tweedie likelihood gradient.

Require: y, A, D_θ, scheduler $\{\sigma_i, s_i, c_i, f_i\}_{i=0}^{N-1}$, guidance ζ, mode
Ensure: Reconstruction $\hat{x}$

```
x_0 ← σ_max ξ, ξ ~ N(0, I)
for i = 0, ..., N − 1 do
    x̂_0 ← Denoise(D_θ, x_i, σ_i, s_i)
    g_i ← ∇_{x_i} ‖y − A(x̂_0(x_i))‖_2          ▷ non-squared L2 through the denoiser
    x_{i+1} ← EDMStep(x_i, x̂_0; i, mode) − ζ g_i
end for
return x_N
```

Algorithm 2 DAPS (Zhang et al., 2025): annealed Langevin chain.

Require: y, A, D_θ, scheduler $\{\sigma_i\}_{i=0}^{N}$ with $\sigma_N = 0$, diffusion sub-config, Langevin parameters $(M_{\text{lgvd}}, \tau, \eta_{\max}, r_{\min})$

$x_0 \leftarrow \sigma_{\max}\xi$
for $i = 0, \ldots, N-1$ **do**
  sub-sched $\leftarrow$ Scheduler(diff-cfg, $\sigma_{\max} = \sigma_i$)
  $\hat{x}_0 \leftarrow \text{Sampler}(D_\theta, x_i; \text{sub-sched})$
  $x \leftarrow \hat{x}_0$; $\eta_l \leftarrow \eta_{\max}(1 + (i/N)(r_{\min} - 1))$
  for $j = 1, \ldots, M_{\text{lgvd}}$ **do**
    $g_{\text{meas}} \leftarrow \nabla_x \|y - A(x)\|_2^2/(2\tau^2)$
    $g_{\text{prior}} \leftarrow (x - \hat{x}_0)/\sigma_i^2$
    $x \leftarrow x - \eta_l(g_{\text{meas}} + g_{\text{prior}}) + \sqrt{2\eta_l}\varepsilon$
  end for
  $x_{i+1} \leftarrow x + \sigma_{i+1}\varepsilon$
end for
return x_N

Algorithm 3 ABMS (Ren et al., 2026): finite stochastic DDIM with noise-space guidance.

Require: $y, A, D_\theta, \{\sigma_i, s_i\}_{i=0}^{N}$, DDIM noise η, guidance γ, direction count D_{dir}, interval K

$x_0 \leftarrow s_0\sigma_0\xi$
for $i = 0, \ldots, N-1$ **do**
  $\hat{x}_0 \leftarrow \text{Denoise}(D_\theta, x_i, \sigma_i, s_i)$; $\hat{\varepsilon}_i \leftarrow (x_i/s_i - \hat{x}_0)/\sigma_i$
  $\sigma_{\text{up}} \leftarrow \eta\sigma_{i+1}\sqrt{\max\{0, 1 - (\sigma_{i+1}/\sigma_i)^2\}}$
  $\sigma_{\text{down}} \leftarrow \sqrt{\max\{0, \sigma_{i+1}^2 - \sigma_{\text{up}}^2\}}$
  $\bar{x}_{i+1} \leftarrow s_{i+1}(\hat{x}_0 + \sigma_{\text{down}}\hat{\varepsilon}_i)$; $\delta_i \leftarrow s_{i+1}\sigma_{\text{up}}$
  if $\sigma_{i+1} = 0$ **then**
    return $\hat{x}_0$
  end if
  if $i > 0$ and $i \bmod K = 0$ **then**
    for $d = 1, \ldots, D_{\text{dir}}$ **do**
      $e_d \sim \mathcal{N}(0, I)$
      $\hat{x}^{(d)} \leftarrow \text{clamp}(\text{Denoise}(D_\theta, \bar{x}_{i+1} + \delta_i e_d, \sigma_{i+1}, s_{i+1}), [-1, 1])$
      $\Delta_d \leftarrow -\gamma\nabla_{e_d}\sqrt{\text{Aloss}(\hat{x}^{(d)}, y)}/\|\nabla_{e_d}\sqrt{\text{Aloss}(\hat{x}^{(d)}, y)}\|_2$
    end for
    $x_{i+1} \leftarrow \bar{x}_{i+1} + \delta_i(e_{\text{ref}} + D_{\text{dir}}^{-1}\sum_d \Delta_d)$, $e_{\text{ref}} \sim \mathcal{N}(0, I)$
  else
    $x_{i+1} \leftarrow \bar{x}_{i+1} + \delta_i e$, $e \sim \mathcal{N}(0, I)$
  end if
end for
return $\text{Denoise}(D_\theta, x_N, \sigma_N, s_N)$

Algorithm 4 DiffStateGrad (Zirvi et al., 2025): projected optimization and resampling.

Require: y, A, D_θ, scheduler, initializer Init, guidance ζ, travel interval K, travel length J, projection period P, cutoff c, resampling scale κ

$x \leftarrow \mathrm{Init}(y)$ ▷ normalized KS for WL; otherwise prior noise
for $t = 0, \ldots, N-1$ **do**
  $x_{\mathrm{prev}} \leftarrow x$; $\hat{x}_0 \leftarrow \mathrm{Denoise}(D_\theta, x_{\mathrm{prev}}, \sigma_t, s_t)$
  $\tilde{x} \leftarrow \mathrm{EDMStep}(x_{\mathrm{prev}}, \hat{x}_0; t, \mathrm{mode})$
  $\alpha_t \leftarrow (1+\sigma_t^2)^{-1}$; $\alpha_{t+1} \leftarrow (1+\sigma_{t+1}^2)^{-1}$; $c_t \leftarrow s_t s_{t+1}/\sqrt{\alpha_t \alpha_{t+1}}$
  $x \leftarrow \tilde{x} - (\zeta/2)\alpha_t c_t \nabla_{x_{\mathrm{prev}}} \|y - A(\hat{x}_0)\|_2$
  if t is a travel step **then**
    Save $x_{\mathrm{save}} \leftarrow x$; advance J prior steps to $(z, \hat{z}_0)$; compute rank-r SVD basis (U_r, V_r) from z
    Optimize $z^* \leftarrow \arg\min_z \|y - A(z)\|_2^2$ using stage-dependent pixel/latent AdamW
    Every Pth gradient is projected as $g \leftarrow U_r U_r^\top g V_r V_r^\top$
    $\omega \leftarrow \kappa \frac{1-\alpha_{t+1}}{1-\alpha_t} \max\{1 - \alpha_t/\alpha_{t+1}, 0\}$; $b \leftarrow 1 - \alpha_{t+1}$
    $x \leftarrow \frac{\omega s_{t+1} z^* + b x_{\mathrm{save}}}{\omega + b} + \frac{s_{t+1}}{\sqrt{\alpha_{t+1}}} \sqrt{\frac{\omega b}{\omega + b}} \varepsilon$
  end if
end for
Optionally apply projected latent refinement at the terminal state; return the result.

Algorithm 5 SITCOM (Alkhouri et al., 2025): per-level consistency optimization.

Require: y, A, D_θ, Karras schedule $(N, \sigma_{\min}, \sigma_{\max}, \rho)$, inner cap M, learning rate η_{in}, optional threshold L_*

Construct $\{\sigma_k\}_{k=0}^N$ with $\sigma_N = 0$; $x \leftarrow \sigma_{\max}\xi$
for $k = 0, \ldots, N-1$ **do**
  $z \leftarrow x$
  for $j = 1, \ldots, M$ **do**
    $\hat{x}_0 \leftarrow \mathrm{clamp}(D_\theta(z, \sigma_k), [-1, 1])$; $\ell \leftarrow \mathrm{Aloss}(\hat{x}_0, y)$
    $z \leftarrow \text{Adam-step}(z; \nabla_z \ell, \eta_{\mathrm{in}})$
    if $L_* \neq \varnothing$ and $\ell \leq L_*$ **then**
      break
    end if
  end for
  $x \leftarrow \hat{x}_0 + \sigma_{k+1}\varepsilon$
end for
return x

Algorithm 6 DPnP (Xu & Chi, 2024): likelihood steps and conditional denoising.

Require: y, A, D_θ, schedule $\{\eta_i\}_{i=0}^{K-1}$, PCS steps M_{PCS}, DDS steps M_{DDS}, scale ν, step scale c_h, regularizer λ

$x \leftarrow \eta_0 \xi$
for $i = 0, \dots, K-1$ **do**
  $a \leftarrow x$; $v \leftarrow x$; $\eta \leftarrow \eta_i$
  for $j = 1, \dots, M_{\mathrm{PCS}}$ **do**
    $\ell \leftarrow \mathrm{Aloss}(v, y)$; $g \leftarrow \nabla_v \ell$; $h \leftarrow c_h \min\{\nu^2, \eta^2\}/(\lambda + \sqrt{\ell})$
    $c \leftarrow 1 - \exp(-h/\eta^2)$
    $v \leftarrow v + c(a - v) - \eta^2 c\, g/(2\nu^2) + \eta\sqrt{1 - \exp(-2h/\eta^2)}\varepsilon$
  end for
  $x \leftarrow v$; $u \sim \mathcal{N}(0, I)$; $\Delta\theta \leftarrow \pi/(2M_{\mathrm{DDS}})$
  for $m = 0, \dots, M_{\mathrm{DDS}} - 1$ **do**
    $\theta \leftarrow m\Delta\theta$; $q \leftarrow x + \eta \sin\theta\, u$; $\sigma \leftarrow \eta \cos\theta$
    $\hat{x}_0 \leftarrow D_\theta(q; \sigma)$; $u \leftarrow u - \Delta\theta(q - \hat{x}_0)/\sigma$
  end for
  $x \leftarrow x + \eta u$
end for
return x

Algorithm 7 RED-diff (Mardani et al., 2024): image-space optimization.

Require: y, A, D_θ, scheduler $\{\sigma_i\}$, observation weight w, RED weight $\lambda(\sigma)$, learning rate η

$\mu \leftarrow 0$
for $i = 0, \dots, N-1$ **do**
  $\sigma \leftarrow \sigma_i$; $\varepsilon \sim \mathcal{N}(0, I)$; $x_t \leftarrow \mu + \sigma\varepsilon$
  $\hat{\varepsilon} \leftarrow (x_t - D_\theta(x_t, \sigma))/\sigma$ ▷ no denoiser gradient
  $\ell \leftarrow \|y - A(\mu)\|_2^2$
  $g \leftarrow w\nabla_\mu \ell + \lambda(\sigma)(\hat{\varepsilon} - \varepsilon)$
  $\mu \leftarrow \text{Adam-step}(\mu; g, \eta)$
end for
return μ

Algorithm 8 DMPlug (Wang et al., 2024): full-trajectory latent optimization.

Require: y, A, D_θ, scheduler, outer steps K, learning rate η

$z \leftarrow \sigma_{\max}\xi$; $\ell^* \leftarrow +\infty$; $\hat{x}^* \leftarrow \mathrm{None}$
for $k = 0, \dots, K-1$ **do**
  $\hat{x}_0 \leftarrow \mathrm{Sampler}(D_\theta, z; \text{full schedule})$
  $\ell \leftarrow \mathrm{Aloss}(\hat{x}_0, y)$
  $g_z \leftarrow (\partial\hat{x}_0/\partial z)^\top \nabla_{\hat{x}_0}\ell$
  $z \leftarrow \text{Adam-step}(z; g_z, \eta)$
  if $\ell < \ell^*$ **then**
    $\ell^* \leftarrow \ell$; $\hat{x}^* \leftarrow \hat{x}_0$
  end if
end for
return $\hat{x}^*$ ▷ measurement-loss selection, not test-reference selection

Algorithm 9 DMILO (Zheng et al., 2025b): layer-wise latent and offset optimization.

Require: y, A, D_θ, scheduler, outer steps K, inner steps M_{inner}, learning rates $(\text{lr}_z, \text{lr}_\nu)$, weights (λ_ν, λ_z), fidelity $\mathcal{L}_{\text{fid}}$

$x_{\text{layer}}[0] \leftarrow \sigma_{\max}\xi$
For $\ell = 0, \ldots, N-1$, set $x_{\text{layer}}[\ell+1]$ to the first output of $\text{EulerMeanStep}(D_\theta, x_{\text{layer}}[\ell], \ell)$
$\nu_{\text{layer}}[0..N-1] \leftarrow 0$
for $k = 0, \ldots, K-1$ **do**
  for $\ell = N-1, \ldots, 0$ **do**
    $\text{target} \leftarrow x_{\text{layer}}[\ell+1]$; $z \leftarrow x_{\text{layer}}[\ell]$; $\nu \leftarrow \nu_{\text{layer}}[\ell]$
    for $j = 1, \ldots, M_{\text{inner}}$ **do**
      $x_{\text{prop}}, _, \text{info} \leftarrow \text{EulerMeanStep}(D_\theta, z, \ell)$; $x_{\text{prop}} \leftarrow x_{\text{prop}} + \nu$
      if $\ell = N-1$ **then**
        $L_{\text{recon}} \leftarrow \mathcal{L}_{\text{fid}}(\text{Denoise}(D_\theta, x_{\text{prop}}, \text{info}.\sigma_{\text{next}}, \text{info}.s_{\text{next}}), y)$
      else
        $L_{\text{recon}} \leftarrow \text{MSE}(x_{\text{prop}}, \text{target})$
      end if
      $L \leftarrow L_{\text{recon}} + \lambda_\nu \|\nu\|_1 + \lambda_z \|z\|_2^2$
      $(z, \nu) \leftarrow \text{Adam-step}((z, \nu); \nabla L, \text{lr}_z, \text{lr}_\nu)$
    end for
    $x_{\text{layer}}[\ell] \leftarrow z$; $\nu_{\text{layer}}[\ell] \leftarrow \nu$
  end for
  Rebuild the forward chain: take the first EulerMeanStep output at each layer and add $\nu_{\text{layer}}[\ell]$
end for
return $\text{Denoise}(D_\theta, x_{\text{layer}}[N], \sigma_{\min}, s_{\min})$

E HYPERPARAMETER CONFIGURATIONS AND SEARCH LANDSCAPES

E.1 SELECTION PROTOCOL

Hyperparameters are selected by mean PSNR on small selection sets, which we report per task because they are smaller than a typical validation split. Weak lensing selects on one validation map (index 1512) and Galaxy on one validation image (ID 2816); neither belongs to the test set. BioSR selects DPnP and ABMS on sixteen validation pairs and the other solvers on one, and the four-iteration Richardson–Lucy baseline was chosen on one validation pair from eight iteration counts. The Galaxy Richardson–Lucy baseline uses a fixed 100 iterations and the unregularized CG baseline a fixed 80 conjugate-gradient iterations; neither was tuned. MRI selection read test images 1 to 22, 40, and 68, and the configurations used here were fixed on image 4, so we report images 23 to 32. BHI selection read test images 0, 1, 2, 6, 12, 16, 23, 29, and 37 across the nine solvers, so we report the remaining 41. DPnP's BHI configuration follows its reported search table because the search run could not be recovered, and every BHI solver observes an independent noise realization of the same array. Continuous parameters use log-scale searches and discrete parameters use finite candidate sets. BHI selection uses translation-aligned PSNR without additional blurring. The BioSR intensity policy is frozen before solver selection. The budget itself is a selection dimension rather than a requirement to run all solvers to their maximum number of steps. The selected configurations are frozen before test evaluation; runtime and effective-step information is reported separately.

E.2 SEARCH SPACE AND SELECTED CONFIGURATIONS

Table 10 lists the search space of each method and the configuration selected on each task.

Table 10: **Search spaces and selected configurations.** Continuous parameters use log-scale searches; selection uses PSNR on the per-task selection sets described above. Iteration counts denote effective steps.

Method / parameter	Search space	Weak lensing	Galaxy PSF	Fluorescence	BHI	MRI
DPS						
guidance scale	$[10^{-3}, 10^{3}]$	10.0	6.0	3.0×10^{-1}	0.00225	0.00225
data-fidelity domain	{k-space, image}	image	image	image	—†	k-space
reverse diffusion steps	$[20, 1000]$	20	1000	500	1000	1000
DAPS						
annealing steps	$[20, 200]$	200	170	62	200	200
diffusion steps	$[1, 10]$	8	3	2	4	5
Langevin step size	$[10^{-8}, 10^{-2}]$	3.0×10^{-6}	1.70×10^{-5}	1.49×10^{-6}	1.0×10^{-8}	1.0×10^{-4}
Langevin steps	$[1, 100]$	50	82	44	20	100
reg./noise temp.	$[10^{-4}, 10^{1}]$	1.0×10^{-3}	8.24×10^{-4}	4.03×10^{-2}	5.0×10^{-3}	1.0×10^{-2}
DiffStateGrad						
variance threshold	$[0.90, 1.00]$	0.9	0.93	0.979	0.98	0.999
projection frequency	$[0, 10]$	8	2	10	1	2
reverse diffusion steps	$[20, 1000]$	100	500	200	220	1000
SITCOM						
correction step size	$[10^{-5}, 5 \times 10^{-2}]$	1.0×10^{-2}	6.72×10^{-3}	4.75×10^{-3}	1.5×10^{-2}	2.5×10^{-3}
correction steps	$[5, 50]$	8	50	10	28	10
diffusion steps	$[10, 80]$	10	78	20	16	50
ABMS						
guidance scale	$[10^{-3}, 10^{1}]$	1.0×10^{-3}	1.89×10^{-2}	2.0×10^{-1}	5.0×10^{-2}	3.0×10^{-3}
MC paths	{1,2,4,9}	2	9	4	1	1
correction interval	$[1, 10]$	1	1	2	1	1
stochastic reverse update	{on,off}	off	off	on	off	on
reverse diffusion steps	$[100, 1000]$	1000	100	100	300	100
DPnP						
consistency step size	$[10^{-3}, 1]$	1.0	1.64×10^{-2}	4.21×10^{-1}	3.0×10^{-2}	0.35
consistency iters.	$[5, 500]$	500	5	64	16	50
regularization weight	$[10^{0}, 10^{2}]$	100.0	13.14	10.79	50.0	50.0
eff. noise level	$[2 \times 10^{-3}, 5 \times 10^{-1}]$	1.0×10^{-2}	1.18×10^{-2}	1.20×10^{-1}	0.10	3.0×10^{-3}
prior denoise steps	$[1, 8]$	6	5	4	1	4
outer prior iterations	$[2, 20]$	$6/16$	12	$4/6$	$4/12$	$2/16$
RED-diff						
learning rate	$[10^{-4}, 10^{-1}]$	1.0×10^{-3}	1.25×10^{-2}	1.78×10^{-2}	5.0×10^{-2}	2.0×10^{-2}
regularization / obs. weight	$[10^{-5}, 1]$	1.0	1.68×10^{-3}	9.56×10^{-1}	4.0×10^{-4}	2.5×10^{-1}
gradient weight	$[10^{-2}, 10^{2}]$	1.0×10^{-1}	8.89	5.22×10^{-1}	2.5×10^{-1}	1.0
reg. schedule	{const.,linear,sqrt}	constant	constant	constant	constant	constant
data-fidelity domain	{k-space,image}	image	image	image	—†	k-space
optimization steps	$[20, 1000]$	20	1000	250	1000	1000
DMPlug						
latent learning rate	$[10^{-4}, 2 \times 10^{-1}]$	3.0×10^{-2}	1.59×10^{-2}	2.29×10^{-2}	1.0×10^{-2}	3.0×10^{-2}
optimization steps	$[220, 1000]$	800	460	1000	300	220
diffusion / decode steps	$[5, 200]$	150	11	40	40	5
DMILO						
latent step size	$[10^{-4}, 1.6 \times 10^{-2}]$	5.0×10^{-3}	8.0×10^{-3}	8.65×10^{-3}	1.2×10^{-2}	3.0×10^{-3}
offset step size	$[10^{-5}, 10^{-3}]$	5.0×10^{-4}	1.0×10^{-3}	9.09×10^{-5}	1.0×10^{-3}	2.0×10^{-4}
outer opt. steps	$[3, 18]$	4	10	5	12	14
inner opt. steps	$[25, 200]$	100	100	200	50	40
interm. layer / diffusion steps	$[2, 15]$	3	3	3	3	3
sparsity reg.	$[10^{-2}, 1]$	2.0×10^{-2}	1.0×10^{-1}	1.0×10^{-1}	1.0×10^{-1}	1.0×10^{-1}
latent prior reg.	$[10^{-5}, 10^{-3}]$	1.0×10^{-5}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}

MRI settings correspond to $R = 4$. DPnP fractions denote starting/total prior sampling steps. BioSR uses conditional DDS (64 consistency steps, four denoising steps), four ABMS directions with stochastic DDIM ($\eta = 1$), DiffStateGrad guidance 0.5, and DMILO pixel-MSE fidelity. †BHI fidelity is in the closure domain. Auxiliary scheduler details remain in the experiment configurations.

E.3 PER-TASK SEARCH LANDSCAPES

Figures 9–12 show the recorded searches for weak lensing, Galaxy, BioSR, and black-hole imaging, and the MRI selection is described in Appendix F.1. Selection evaluates configuration and budget jointly: some candidate trajectories improve the data term without improving validation PSNR, so the selected budget need not be the largest tested.

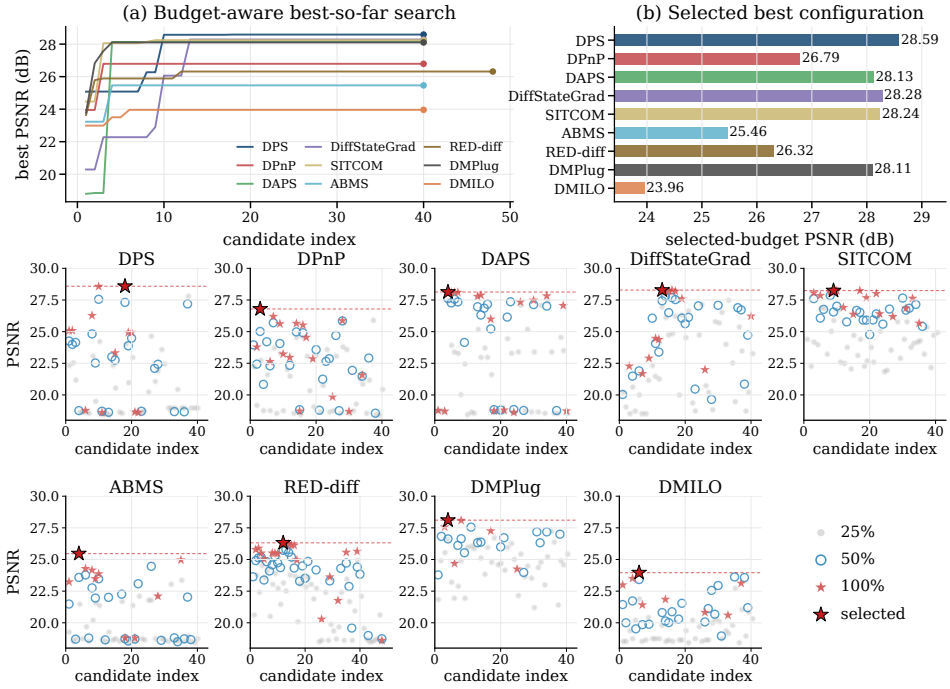


Figure 9: **Weak-lensing configuration search** at noise 0.01. The panels show search progress and evaluated candidates.

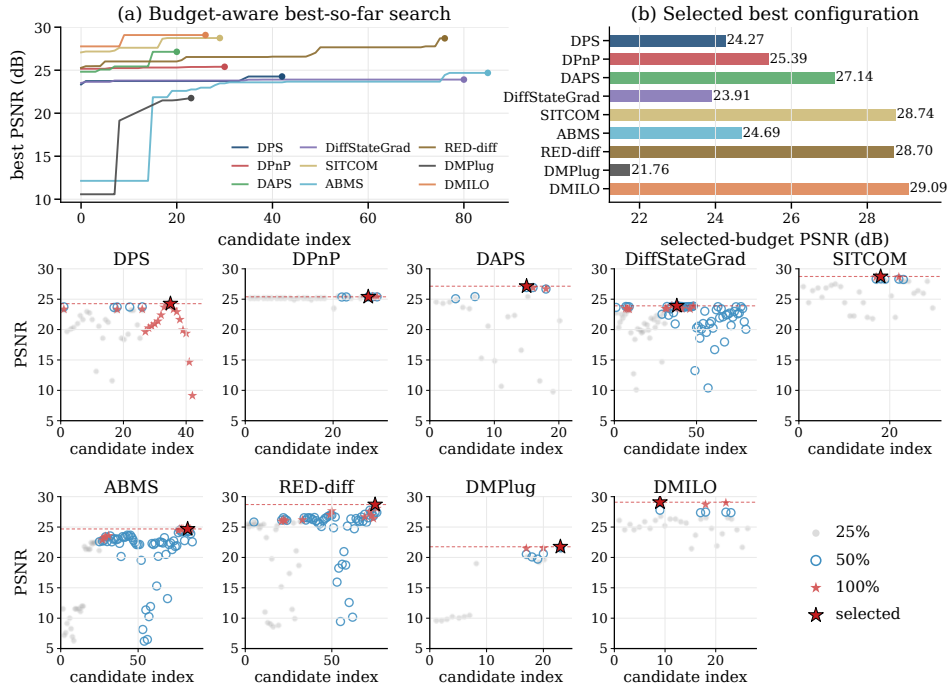


Figure 10: **Galaxy PSF configuration search.** The selected parameter values are recorded in Table 10.

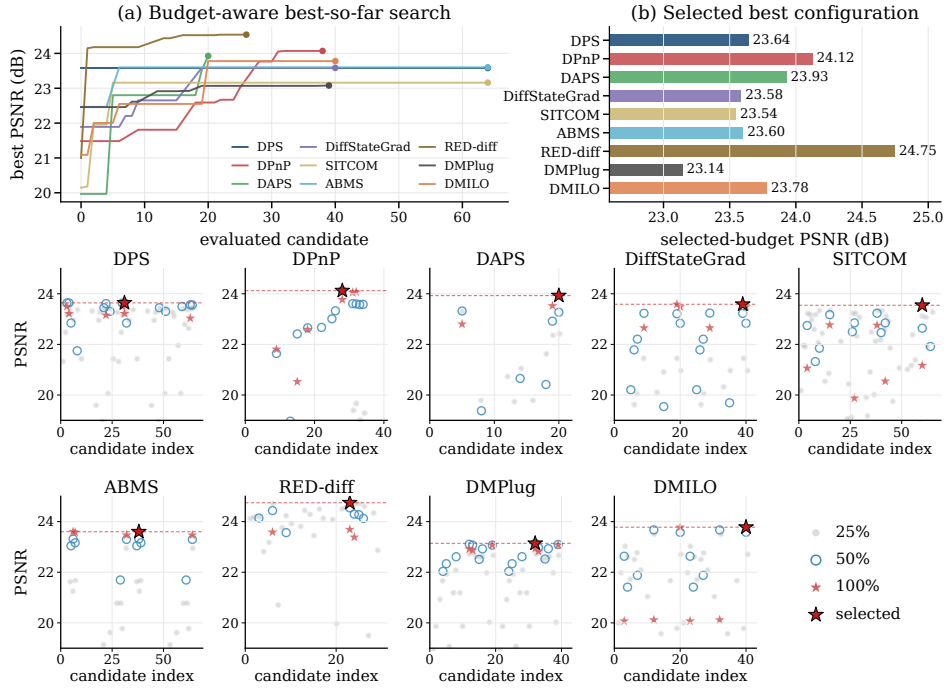


Figure 11: **BioSR configuration search.** Intensity calibration is fixed before solver selection; selected configurations are frozen before testing.

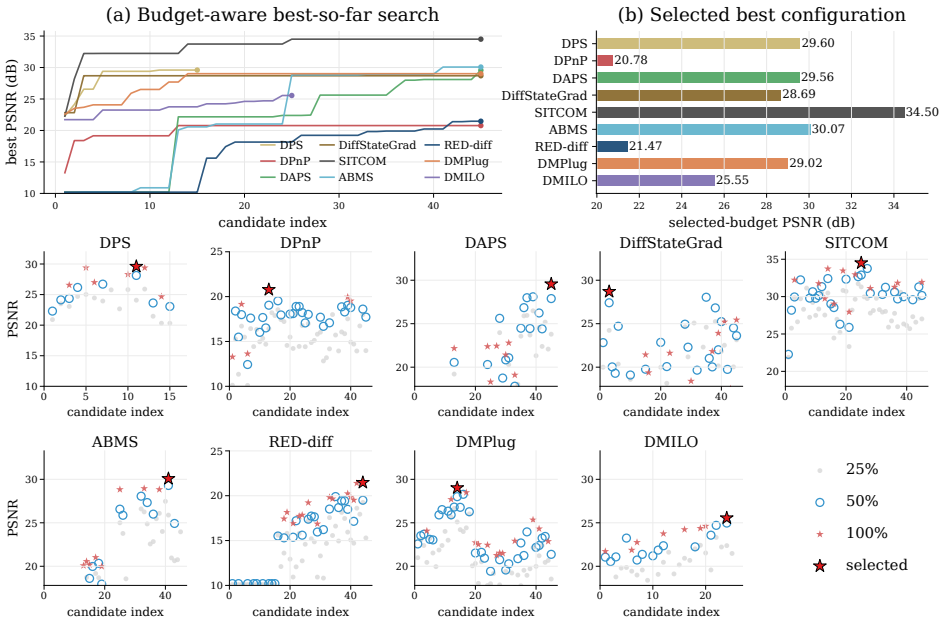


Figure 12: **BHI configuration search.** Solvers fit closure measurements, while aligned PSNR on the selection images selects configurations. These are different inference and selection criteria.

F SUPPLEMENTARY RESULTS AND ANALYSES

The main tables report complete means and between-image standard deviations, including all three Galaxy seeing levels. The subsections below add the MRI configurations, baseline details, the BioSR calibration ablation, the hallucination measurements, and two checks on the controlled study.

F.1 MULTI-COIL MRI: CONFIGURATIONS

Configurations are selected on image 4 of the test split, which no evaluation uses, and every solver is evaluated on images 23 to 32. SITCOM uses the MRI setting of its authors, 50 steps with 10 inner steps and no proximal term (Alkhouri et al., 2025), with the learning rate chosen on image 4; a 30-step inner loop selected earlier collapses on that image to 8.1 dB. The projection variant used in DiffStateGrad's earlier search is not available in the released code, so its guidance scale was reselected on image 4, and DMPlug's learning rate was chosen there among three values. ABMS uses an Euler–Maruyama transition, which runs 0.4 to 0.7 dB below the original implementation on two check images.

F.2 DOMAIN-STANDARD BASELINE DETAILS

The reported baselines are unsmoothed Kaiser–Squires on the main weak-lensing observations; four-iteration Richardson–Lucy with the BioSR observation-only calibration; EHT RML using closure terms, entropy weight 1, flux weight 100, and 80 iterations; and SigPy SENSE and ℓ_1-wavelet MRI reconstructions with regularization weight 10^{-2}. BioSR, BHI, and MRI choices were frozen from non-test validation runs. The MRI baseline and diffusion results use the same test sources and masks. The RML closure-phase and amplitude statistics have medians of 2.61 and 2.14 and means of 18.5 and 17.8, respectively.

F.3 BIOSR AFFINE CALIBRATION AND DOWNSTREAM RL RECONSTRUCTION

We evaluate the same 50 held-out WF/SIM pairs with four Richardson–Lucy iterations and the fixed reflection-padded Gaussian PSF ($\sigma = 1.9$, kernel size 19). The reference condition uses the affine intensity, gain, and background parameters fitted on non-test data. Each control replaces only the named component with its nominal value: intensity offset and scale $(0, 1)$, gain 1, or background 0. Figure 13 shows the paired PSNR difference between the reference condition and each control. Relative to unit gain, the fitted gain raises mean PSNR by 6.04 dB and improves 49 of the 50 images, and the fitted background adds 0.81 dB on average. Replacing the intensity decoder alone changes mean PSNR by 0.003 dB for RL. Removing both gain and background lowers mean PSNR by 7.80 dB. The calibrated and unit-gain settings obtain mean SSIM 0.757 and 0.732, respectively; zero background instead has slightly higher SSIM (0.769). FRC readouts remain at the grid limit and do not resolve these differences. This fixed-solver ablation isolates the affine parameters for the conventional baseline, not their effect on diffusion solvers.

F.4 HALLUCINATION IN THE SCIENTIFIC SETTINGS

For a linear setting with a defined share, a reconstruction's error splits into its part in the unobservable set $\mathcal{U}$, its hallucination, and the rest. Table 11 reports the hallucination as a share of each solver's error and relative to the source energy, $\|P_{\mathcal{U}}(\hat{x} - x)\|^2/\|x\|^2$, and compares the latter with a zero-fill reference that is exact outside $\mathcal{U}$ and zero inside it, averaging per-image ratios. The sets $\mathcal{U}$ are those of Table 1. For MRI this is the unacquired k-space of the single-coil sampling model, which coil encoding partly constrains, so the MRI block reports error on unacquired lines rather than hallucination in the strict sense. Figure 5 shows one example per setting. Weak lensing is omitted because the reconstructions behind its table came from the earlier implementation and were not retained, and black-hole imaging has no share. The Galaxy block uses the ten images of the controlled study under O2, and on the fifty Galaxy test images the prior-free conjugate-gradient and Richardson–Lucy inverses have ratios of 0.85 and 0.97.

Table 11: **Error in the unobservable set.** Its share of each solver's error, its size relative to the source energy, and its ratio to a zero fill that is exact outside the set and zero inside it. Ratios above one are bold. For MRI the set is the unacquired k-space of the single-coil sampling model, which ignores coil encoding, so that block measures error on unacquired lines rather than hallucination in the strict sense.

Method	PSNR	Share of error (%)	Relative energy (%)	Ratio to zero fill
Fluorescence (BioSR), 50 test images				
RED-diff	26.56	49.8	3.65	0.46
DMILO	25.93	45.2	3.95	0.49
DPS	24.92	57.9	6.36	0.81
DPnP	24.83	52.5	6.06	0.75
DAPS	24.50	54.6	6.52	0.83
DMPlug	23.87	63.2	10.03	**1.20**
SITCOM	23.63	60.5	9.03	**1.10**
ABMS	23.47	55.8	8.40	**1.08**
DiffStateGrad	21.60	75.3	19.59	**2.32**
Zero fill	26.33	100.0	8.03	1.00
Galaxy PSF, the ten controlled-study images				
SITCOM	30.53	96.5	1.86	0.89
DMILO	29.60	81.8	1.95	0.93
RED-diff	29.31	78.7	2.00	0.95
DAPS	27.26	77.5	3.04	**1.51**
DPnP	26.26	41.7	2.08	0.99
ABMS	25.47	51.8	3.07	**1.52**
DPS	25.43	61.9	3.72	**1.83**
DiffStateGrad	25.18	58.9	3.68	**1.87**
DMPlug	21.96	66.4	8.53	**4.53**
Zero fill	30.15	100.0	2.09	1.00
Multi-coil MRI ($R = 4$), test images 23–32, unacquired lines as $\mathcal{U}$				
DAPS	28.08	75.0	3.13	0.69
RED-diff	27.88	71.7	2.80	0.62
DPS	25.71	72.5	4.75	**1.04**
DPnP	25.34	69.1	3.86	0.84
DiffStateGrad	23.87	56.3	4.32	0.94
DMILO	23.21	72.3	7.86	**1.70**
SITCOM	20.42	47.5	7.41	**1.61**
ABMS	17.99	47.7	17.00	**3.64**
DMPlug	17.12	19.6	5.28	**1.15**
Zero fill	–	100.0	4.58	1.00

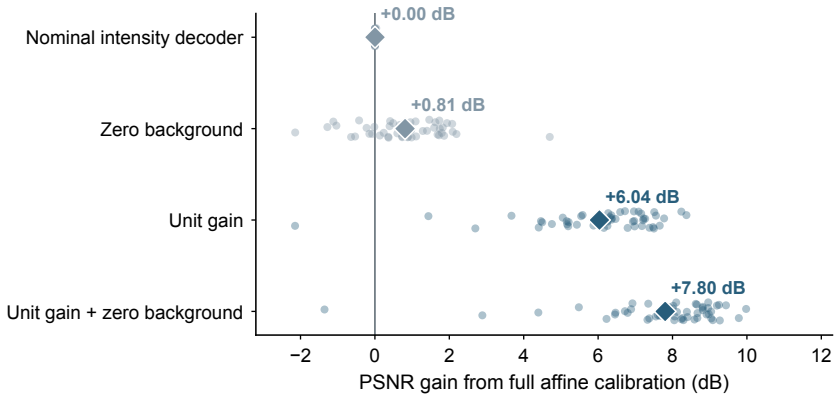


Figure 13: **Downstream effect of BioSR affine calibration.** Paired PSNR gain of the fully calibrated model over four component controls for the same 50 held-out images and four-iteration Richardson–Lucy solver. All conditions retain the same Gaussian PSF. Dots denote individual images and diamonds the mean paired difference; positive values favor the calibrated condition.

F.5 Energy in the discarded directions

A reconstruction's own energy in the unobservable set can be compared with that of the training sources without ground truth. For each O1 reconstruction we compute the fraction of its energy in the 510 discarded frequencies, in the image representation used for the prior share, and compare it with the same fraction over 2000 training images, whose central 90% range is 66.7% to 78.2% and whose energy-weighted mean reproduces the 72.2% share. RED-diff lies above that range at 93.0% on average, the signature of its brightness offset, and DAPS and DPnP lie below it at 65.1% and 36.0%. The other six solvers lie inside, including DMILO and SITCOM, whose O1 errors lie 85.8% and 97.5% in the discarded frequencies. The check detects a gross over- or under-filling of the open directions, not a wrong fill of the right size.

F.6 Resampling the controlled study

The controlled study evaluates every operator on the same ten Galaxy images, so its two summary statistics can be checked by resampling those images. SITCOM runs without a final mean correction that was used when its configuration was selected. In each of $10{,}000$ paired bootstrap draws we resample ten image indices with replacement, apply the same draw to all nine solvers and all four operators, and recompute each solver's mean PSNR and mean relative residual. Of the 36 solver pairs, 18 reverse their PSNR ordering from O2 to O3 on the evaluated images; the central 95% interval is 17 to 20, and no draw falls outside 16 to 21. The Spearman correlation between negated residual and PSNR has central 95% intervals $[0.98, 1.00]$ on O1R, $[0.65, 0.77]$ on O2, and $[0.22, 0.47]$ on O1, and it decreases from O1R to O2 to O1 in every draw (Fig. 14). Its level on O1 is the least certain of the three, but no draw places it above its level on O2. The resampling measures sensitivity to which images are evaluated; it does not retrain the prior or reselect configurations.

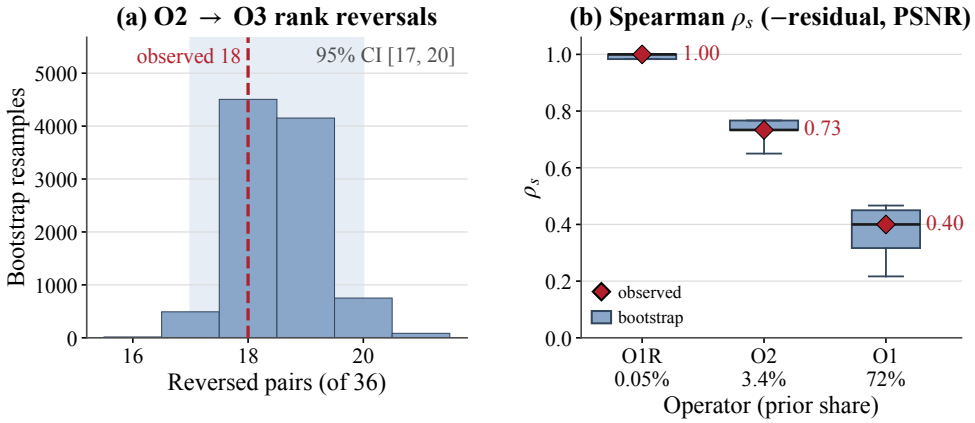


Figure 14: **Resampling the ten Galaxy images.** Each of 10,000 paired bootstrap draws is applied to all solvers and operators. (a) Solver pairs whose mean-PSNR ordering reverses from O2 to O3; the band marks the central 95% interval. (b) Spearman correlation between negated residual and PSNR, with operators ordered by prior share; whiskers mark the central 95% interval.

G SCOPE AND INTERPRETATION OF SCIENTIFIC METRICS

Weak lensing. Weak gravitational lensing relates the projected convergence field to shear. In the benchmark's flat-sky, Fourier-domain model, the full two-component Kaiser–Squires map has unit-modulus response at nonzero frequencies and an unobserved mean. The benchmark separates reference reconstruction error from agreement of selected statistical summaries. Its power and normalized peak-distribution metrics do not by themselves evaluate a complete cosmological inference pipeline or the total abundance of detected peaks.

Galaxy PSF deconvolution. The Galaxy task concerns a different astronomical measurement process: atmospheric and instrumental blur in survey images (Dey et al., 2019; Walmsley et al., 2022). The Galaxy construction applies an additional known PSF to a survey-image reference. It therefore provides a controlled reconstruction problem rather than a complete simulation of all observing effects. Adaptive moments measure a selected shape property of the reconstructed benchmark images; their agreement is distinct from full-image fidelity and downstream shear-calibration accuracy.

Fluorescence microscopy. For microscopy, BioSR provides real WF/SIM pairs for evaluating reconstruction (Qiao et al., 2021). Here the prior models normalized microtubule structure, while a non-test intensity decoder and fitted PSF link this representation to WF measurements. Reference FRC summarizes frequency-domain agreement. Its censored grid-limit readout cannot distinguish methods once both curves remain above threshold, nor does it establish identical optical resolution.

Black-hole imaging. Black-hole imaging instead uses nonlinear interferometric closure quantities from simulated source intensities and an observation geometry (Zheng et al., 2025a). Source translation is not determined by those observables, so the image metric permits alignment. Closure fit and aligned reference fidelity test different aspects of a reconstruction; neither alone characterizes uncertainty over possible source images.

Multi-coil MRI. The MRI setting combines coil sensitivities with Fourier subsampling on fastMRI data (Zbontar et al., 2018; Zheng et al., 2025a). Reference-image scores and full-k-space errors are reconstruction metrics, while acquired-coefficient residuals concern measurement consistency. No diagnostic or clinical-utility assessment is performed by the benchmark.

H QUALITATIVE RECONSTRUCTION GALLERIES

Figures 15–19 show held-out reconstructions from the nine diffusion solvers on weak lensing, Galaxy, BioSR, black-hole imaging, and MRI. The black-hole and MRI images are test images excluded from hyperparameter selection, selected at regular intervals from the ordered evaluation IDs.

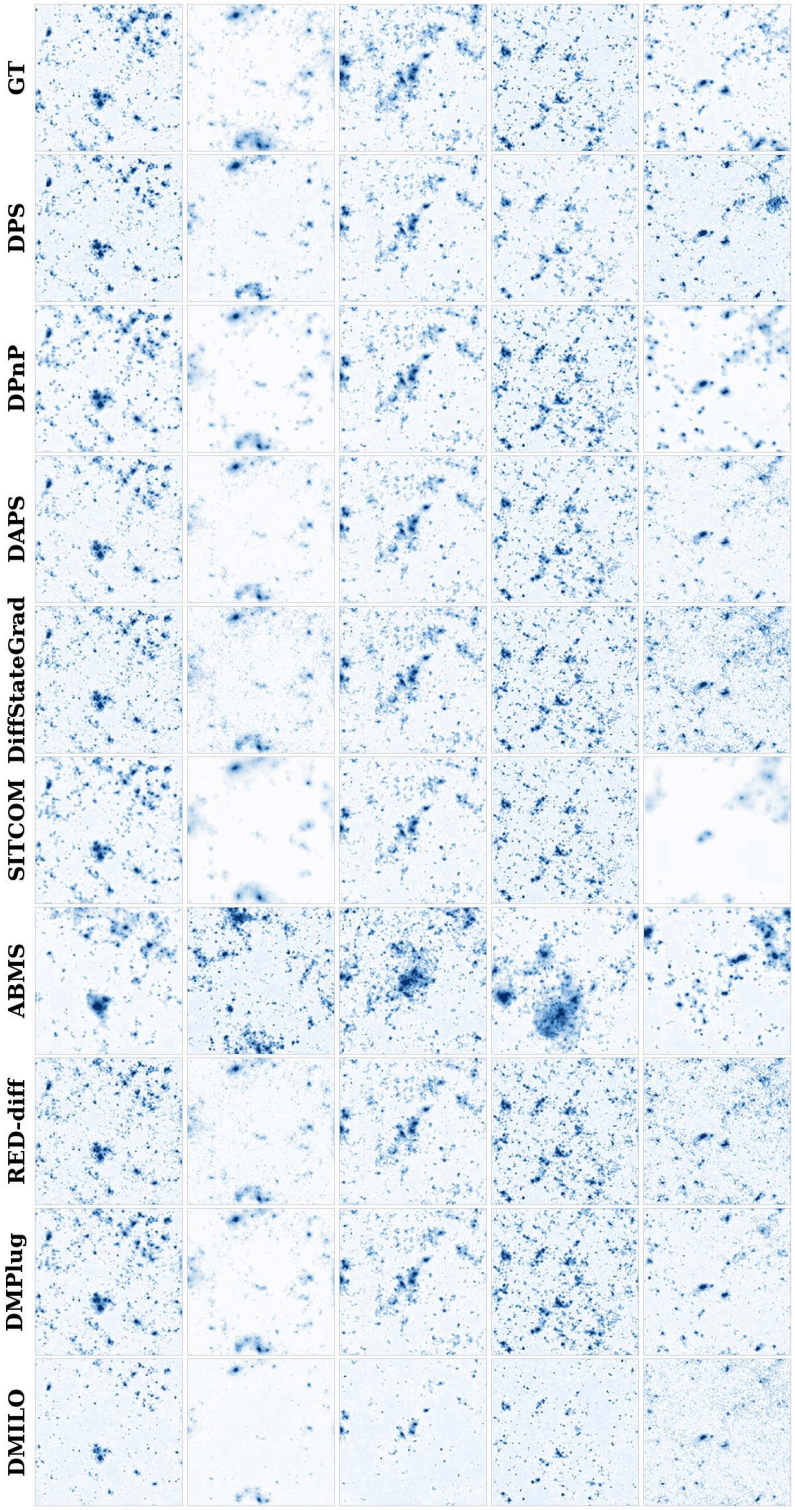


Figure 15: **Additional weak-lensing reconstructions.** Five test examples across the nine methods, with convergence references.

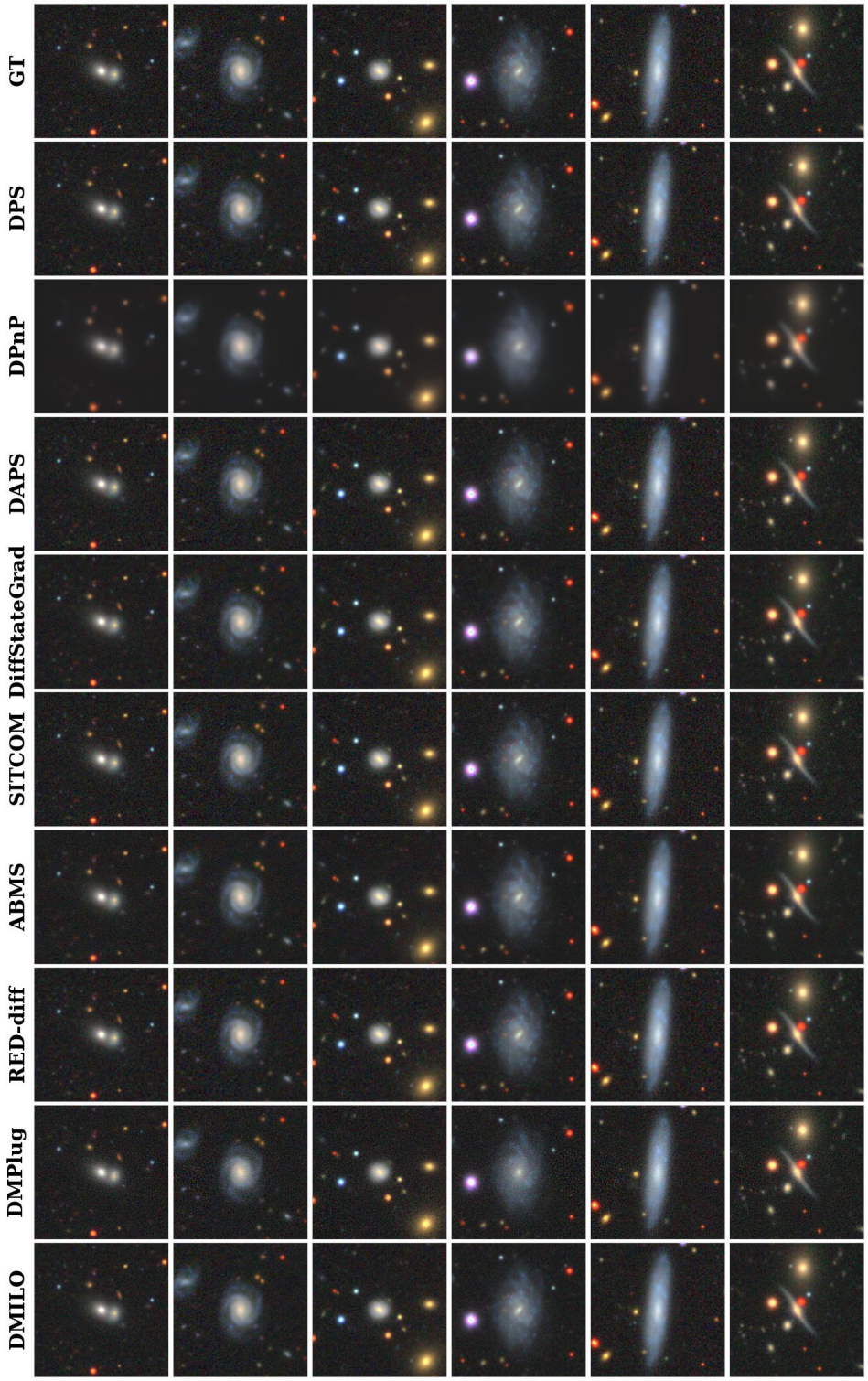


Figure 16: **Additional Galaxy reconstructions.** Six sources under the Typical setting, across the nine methods.

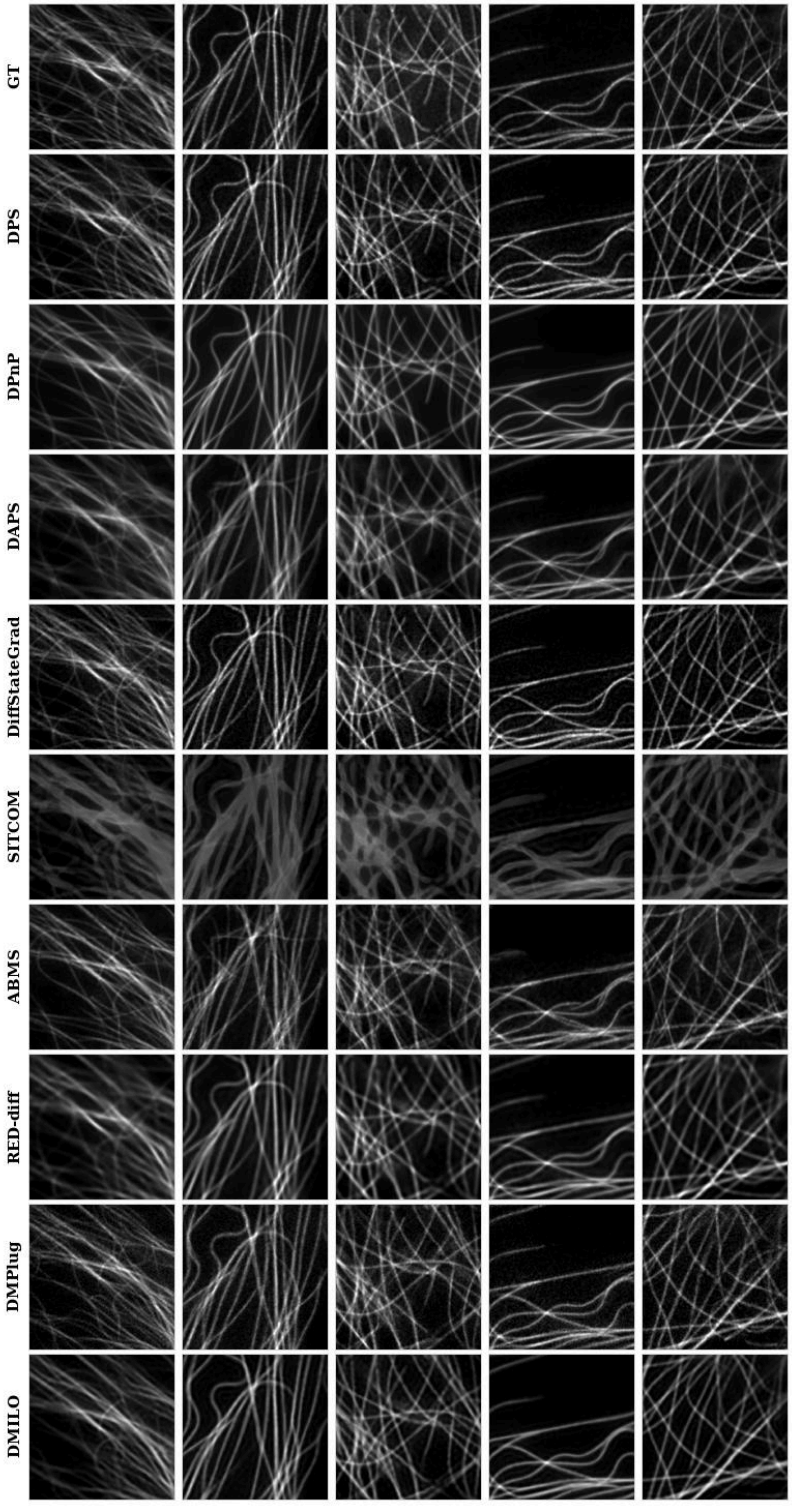


Figure 17: **BioSR qualitative gallery.** Five held-out examples across the nine methods.

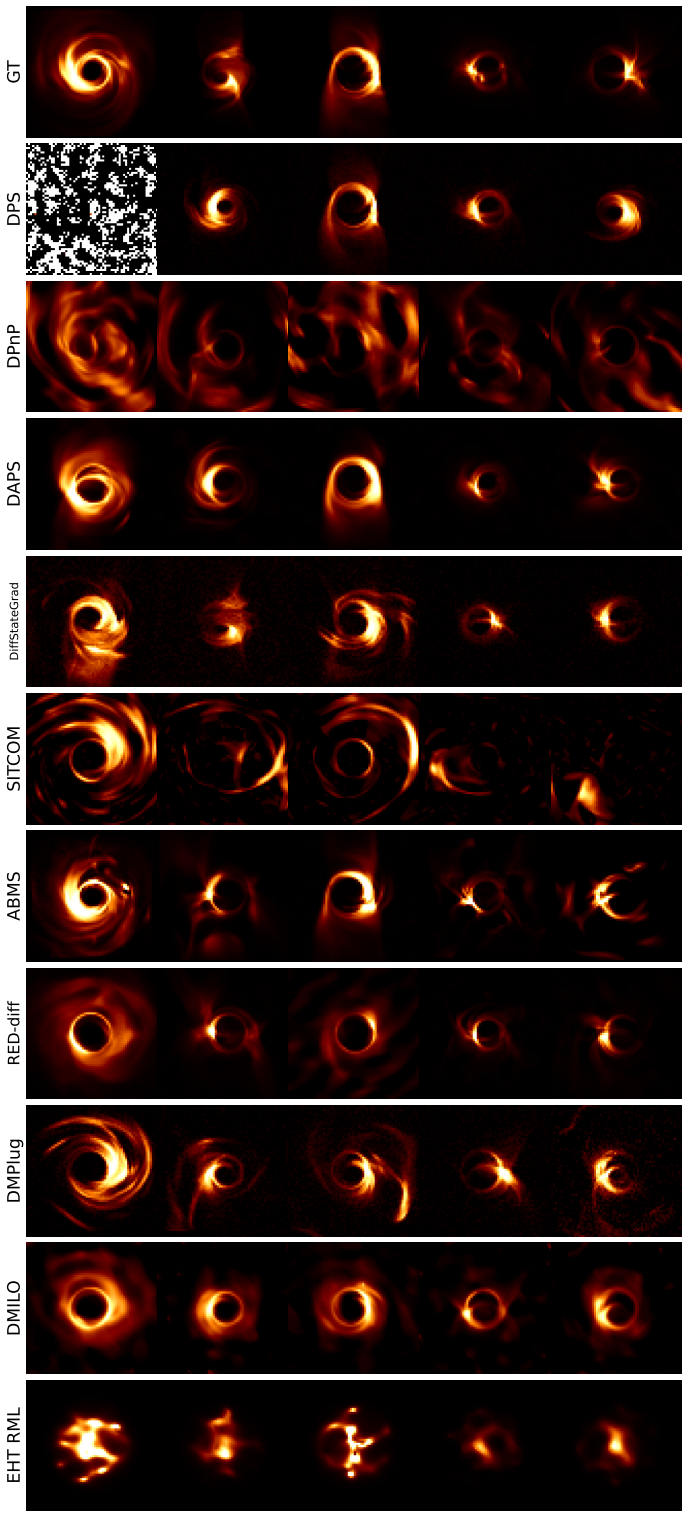


Figure 18: **Additional black-hole imaging reconstructions.** Test images 3, 15, 27, 39, and 49 at $r = 100\%$ for the nine methods and EHT RML. DPS diverges on image 3.

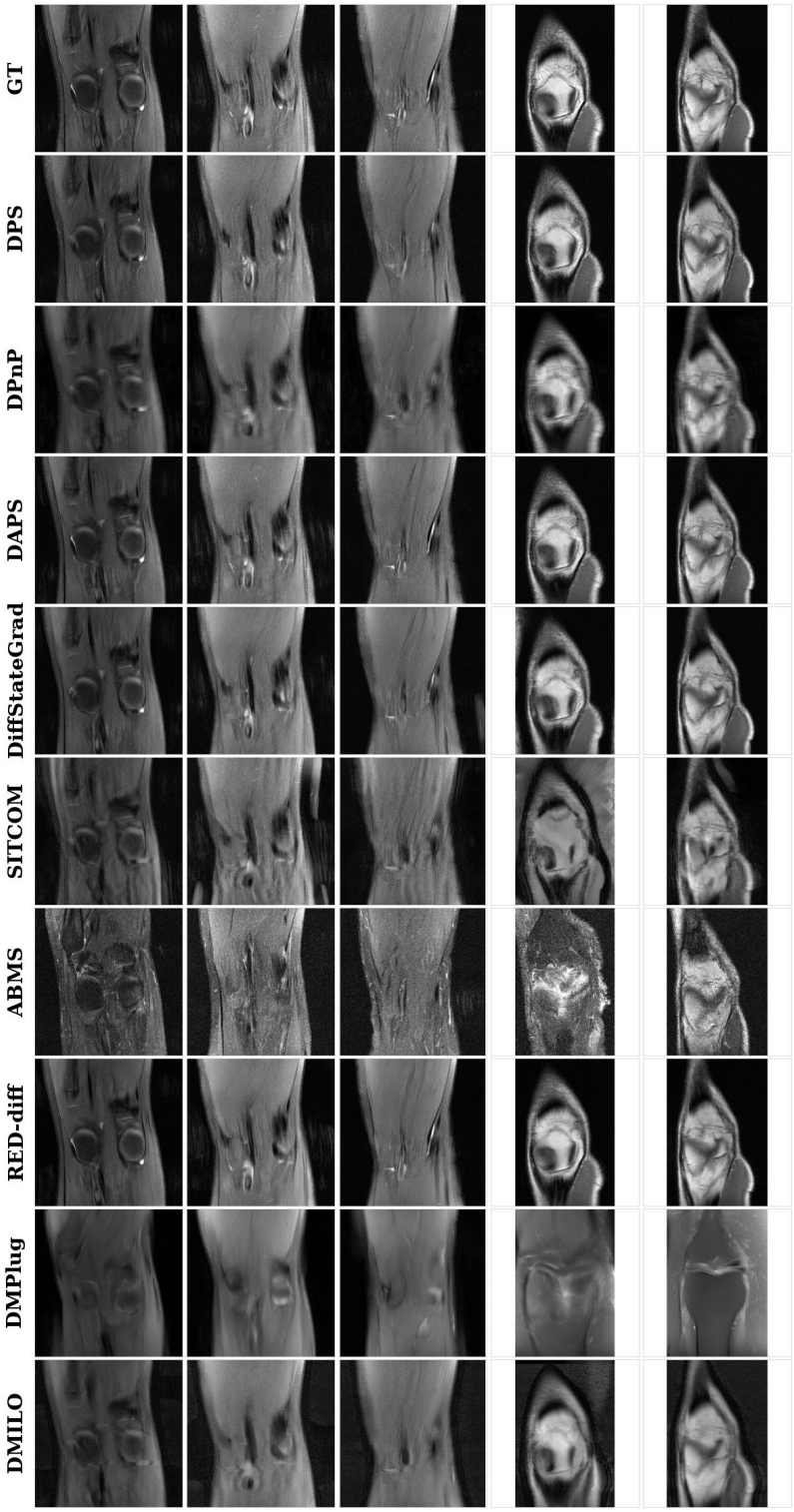


Figure 19: **Additional MRI reconstructions.** Random-mask $R = 4$ test images 23, 25, 27, 29, and 31 for the nine methods.