WHEN MINIMIZING THE MEASUREMENT RESIDUAL SELECTS THE WRONG GUIDANCE STRENGTH

Anonymous authors
Paper under double-blind review

ABSTRACT

When a diffusion or flow prior solves an inverse problem, its trajectory is pulled toward the measurement at every step, and the strength of that pull is a free parameter. Reported scores tune it against reference images; without one, the natural choice is the strength that fits the measurement best. We measure when that choice fails. Under released CIFAR-10 and face diffusion priors, the measurement residual falls as the strength grows, all the way to the strongest setting on every CIFAR-10 blur and mask. Behind binary masks, that setting is also the most accurate. Where the operator measures some directions only weakly, near its noise floor, selecting on the residual fits noise in those directions and gives up 0.099 to 0.514 in correlation and 0.039 to 0.118 under a latent prior. The fraction of such directions, the weak band, is read from the operator and the noise level before any prior is fitted. It tracks the cost. On 10 accelerated-MRI and compressed-sensing forward operators added afterward, it separated the costly operators, which gave up 0.184 to 0.475, from the rest, which gave up at most 0.016. The discrepancy principle, which stops when the residual reaches the noise level, gives up at most 0.028 where the band is not empty. Projected SURE, which estimates the error in all but the weakest directions, gives up at most 0.001 across all 28 operators. Both require the noise level. Most of the loss belongs to the correction rule that replaces the current estimate; the gradient rule stops where its step begins to overshoot. In the real-data task of predicting the brain's response to stimulation with a conditional flow prior, the residual falls the same way, but neither criterion's cost can be distinguished from zero.

1 INTRODUCTION

Guided posterior sampling turns a generative prior into a solver for linear inverse problems. A trajectory runs from noise toward a sample; at every step, its current guess at the clean signal is pulled toward the observation, and the result is treated as a draw from the posterior. Published reconstruction scores use a pull strength tuned against the reference images that the method is meant to recover.

Many problems have no reference, only the measurement. When the measurement can be split, the strength follows from predicting a held-out part, as ZADS does for accelerated MRI (Alçalar et al., 2025). Holding out entries replaces A with a row subset whose spectrum differs from that of the full operator; we do not study that case. Without a split, the natural choice is the strength whose solution best fits the observation. Classical alternatives instead estimate the risk from the observation itself.

The trade-off behind this choice is classical. For regularized least squares, whose solution depends on the regularization parameter in closed form, fitting the measurement too closely amplifies noise in the directions the operator measures weakly, and criteria that stop earlier are well studied (Engl et al., 1996; Hansen, 1992). A guided sampler has no such closed form: its output averages trajectories through a trained denoiser, and the strength acts at every step. Whether the classical failure carries over, how large it is, which operators it reaches, and which classical criteria still work are open empirical questions, and tuning against references hides the answers. We answer them for released pixel and latent diffusion priors, operators from six families, and a real problem.

The residual can favor a strong setting after accuracy has begun to fall. The spectrum of the forward operator, available before any prior is fitted, predicts which operators pay this cost, so a benchmark default can be checked against a new operator before it is used. The size of the cost depends on the

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.

correction rule. Two classical criteria that need no reference do better: the discrepancy principle (Morozov, 1966), which matches the residual to the noise level, and, at twice the sampling cost, an estimate of the reconstruction error in the determined and weak directions (Eldar, 2009). The same tuning problem arises when predicting the brain's response to stimulation at a new site, where a flow prior conditioned on the turbulence of resting activity fills in what the sites already stimulated leave open; we measure it there with each response withheld until scoring.

Our contributions are the following.

- **A selection failure, measured.** Under the released pixel priors, selecting the strength on the residual gives up 0.099 to 0.514 in correlation wherever the operator carries weakly measured directions, and nothing behind binary masks; under a latent prior, it gives up 0.039 to 0.118. Settings are chosen on half of the images and scored on the other, and the separation holds over 1000 random halvings (Sections 3 to 5).
- **A spectral indicator and a test of it.** The weak band tracks the cost at 0.890 across the 18 operators measured first. Moving measured directions into it at fixed rank carries the cost from 0.000 to 0.514, and the added error lands in it at 1.9 to 23.5 times its share of directions. On accelerated-MRI and compressed-sensing operators added afterward, those with a band give up 0.184 to 0.475 and those without at most 0.016 (Sections 4 and 5).
- **Criteria without a reference.** The discrepancy principle gives up at most 0.028 where the band is not empty and up to 0.015 where it is empty. SURE and cross-validation of the fit share the residual's blind spot on CIFAR-10; projected SURE gives up at most 0.001 on all 28 operators, and 0.074 on the original 18 operators when the noise level is off by up to 10 percent (Section 6).
- **The correction rule.** The gradient rule's residual stops falling where its step starts to overshoot; its cost is indistinguishable from zero on 2 of 5 blurs (Section 7).
- **The value of the prior.** Tuned against the truth, the released prior beats retuned shrinkage on 5 of 5 blurs, while a flow prior fitted here on up to 30 000 images loses in 35 of 35 settings (Section 8).
- **A real problem, reference withheld.** At a held-out stimulation site, the residual falls at every step, and neither criterion's cost differs detectably from zero (Section 9).

2 SETUP

Problem. We observe $y = Ax + \eta$ with $\eta \sim \mathcal{N}(0, \sigma^2 I)$ and recover x from a generative prior. Write $A = \sum_i \varsigma_i u_i v_i^\top$ over its nonzero singular values. A least-squares fit sets the coordinate along each v_i to its true value plus $u_i^\top \eta / \varsigma_i$, which has standard deviation $1/s_i$ with $s_i = \varsigma_i / \sigma$; directions outside the sum are left to the prior. The ratios s_i give the error that such a fit injects in each measured direction, expressed in the coordinates where the correction acts. We call directions with $s_i > 1$, where that error is below one unit, *determined*; those with $0.1 < s_i \leq 1$, where it is between one and ten, the *weak band*; and those with $s_i \leq 0.1$ *invisible*. The round thresholds were fixed before any sweep. The weak band of an operator is the fraction of its directions in that range. It is not a signal-to-noise ratio: it depends on the operator, the noise level, and the units of x, here the $[-1, 1]$ pixel scale, but not on the image, and no prior needs to be fitted to read it. It describes where a residual criterion is heading, not the per-step gain of Equation 1, which also moves with ρ and t.

Guidance. We use a variance-preserving diffusion prior and the deterministic sampling trajectory of Song et al. (2021). At each step, the denoiser forms $\hat{x}_0$, its estimate of the clean image, and the correction replaces it with the Gaussian posterior mean that treats $\hat{x}_0$ as a prior mean of variance $\rho\,(1 - \bar{\alpha}_t)$,

$$\hat{x}_0 \leftarrow \hat{x}_0 + \rho\,(1 - \bar{\alpha}_t)\, A^\top \left(\rho\,(1 - \bar{\alpha}_t)\, AA^\top + \sigma^2 I\right)^{-1} (y - A\hat{x}_0)\,, \tag{1}$$

before the sampler continues. The scalar ρ is the guidance strength. As $\sigma \to 0$, Equation 1 becomes $A^+ y + (I - A^+ A)\hat{x}_0$, the range-null-space step of Wang et al. (2023). It shares with Song et al. (2023) the Gaussian approximation of $\hat{x}_0$ and the preconditioner $(vAA^\top + \sigma^2 I)^{-1}$, with v the variance that the approximation assumes, but PiGDM applies its correction to x_t through the derivative of the

denoiser. We call Equation 1 the replacement rule and compare it with the gradient rule of Chung et al. (2023). After each step, that rule moves x_t down the gradient $\nabla_{x_t}\|y - A\hat{x}_0\|$ by a fixed multiple, its strength, instead of enforcing the constraint on $\hat{x}_0$. This is equivalent to the squared-norm step of Chung et al. (2023) because their step size divides by $\|y - A\hat{x}_0\|$.

Prior and baseline. The prior is the released CIFAR-10 (Krizhevsky, 2009) denoising diffusion model (Ho et al., 2020). Section 5 repeats the experiment at 256×256 with a released CelebA-HQ prior (Karras et al., 2018) and with a latent prior. The baseline is Tikhonov shrinkage toward the training mean, with its strength retuned for every operator. Reconstructions are scored by correlation with the true image over all three color channels; the stimulation experiment uses the same accuracy measure. The residuals are $\|y - A\hat{x}\|/\|y\|$. Section 5 repeats the residual criterion's cost in peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and a perceptual distance (LPIPS), each computed as Appendix A.5 describes.

Operators. Three families span the spectrum. *Blur* applies a circular Gaussian blur, diagonal in frequency, with a graded spectrum: five widths leave 303 to 2781 of 3072 directions at or below the noise floor. Pixels lie in $[-1, 1]$, and the observation noise is $\sigma = 0.05$, except 0.02 and 0.1 for the narrowest and widest blurs. *Mask* keeps a random fraction of pixels at full weight and discards the rest, so every direction is either far above the noise floor or absent, and the weak band is empty by construction. *Graded mask* starts from the binary mask that keeps half the pixels and weights down 30, 60, or 90 percent of the kept ones until they land in the weak band. This holds the rank at one half while the band moves from that mask's 0.000 to 0.450; the starting mask is counted once, in the mask family. The graded-mask family separates the weak band from the fraction measured. Section 5 adds super-resolution, accelerated MRI, and compressed sensing.

3 Selecting on the residual

Across all 12 operators of the three families, the residual falls monotonically as the strength increases, and a residual criterion selects the strongest setting. Accuracy against the true image is highest there only behind the binary masks.

The images are split in half: the most accurate and lowest-residual settings are selected on one half, and the accuracy difference between them is scored on the other 32 images. We resample the scoring images for a 95% interval while holding both settings fixed, so the interval is conditional on their selection. At blur width 1.2, the best reconstruction sits at $\rho = 0.003$ and the best fit at $\rho = 1$, scoring 0.973 against 0.620: the residual criterion gives up 0.353 [0.303, 0.404] in correlation (Figure 1).

4 Operators that pay

Behind all 4 binary masks, the most accurate and lowest-residual settings coincide: accuracy is still rising where the residual reaches 0.006, close to an exact fit, and varies by only 0.001 over the last three settings of the sweep. The 8 operators that do carry a weak band give up 0.250 to 0.514, every interval clear of zero; the two groups separate. Across all 12 operators, the weak band tracks the cost at 0.891 (Table 1). The separation does not depend on the original split. Across 1000 random halvings, each of the 4 binary masks gives up exactly nothing, whereas all 11 original weak-band operators and all 5 added later (Section 5) give up a positive cost every time.

Across families, the operators differ in several ways at once; the graded family holds the rank at one half. As its band grows from 0.000 to 0.450, the cost rises from 0.000 to 0.514 [0.487, 0.541] at a band of 0.300 and then holds, the last two steps overlapping. Rank does not explain the cost. Each step also moves directions out of the determined set, so the control holds the rank fixed without isolating the band. Measuring signal against noise instead, by reading the band from the singular values of $A\Sigma^{1/2}/\sigma$ with Σ the covariance of the training images, breaks the relation (Appendix A.4).

The algebra suggests why. As ρ grows, Equation 1 approaches the least-squares solution on everything the operator measures, which puts noise of standard deviation $1/s_i$ into each direction; our grids stop at a finite ρ, short of that limit. In the weak band, that noise ranges from one to ten units, so the prior should retain control of the estimate there.

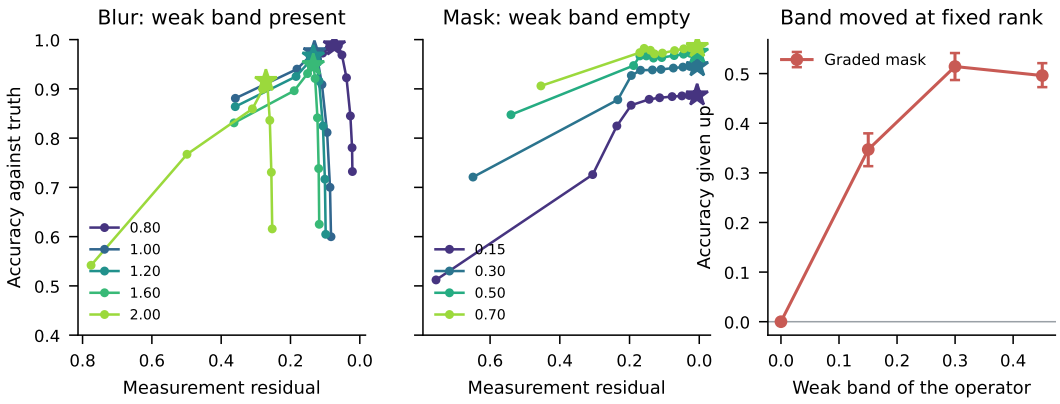

Figure 1: Guidance strength with the released CIFAR-10 diffusion prior. Each curve sweeps the strength; stars mark the most accurate setting over all images (a descriptive reading). The residual falls monotonically along every curve, and a residual criterion lands at its right end. Left: under a blur, weakly measured directions exist, and accuracy turns over. Center: behind a binary mask, accuracy is highest at the strongest setting, and the criterion is safe. Right: weighting down a growing share of the pixels kept by the half mask moves measured directions into the weak band, and the cost follows.

Table 1: Every operator, sorted by weak band within family. Best ρ and floor ρ denote the settings with the highest accuracy and lowest residual, respectively, on the selecting half of the images. The floor cost is the difference in accuracy between them on the scoring half, with its 95% interval; the target cost is the same difference for the discrepancy principle of Section 6. The stimulation row uses the flow prior of Section 9. It scores each participant at the settings selected on the others and lists the settings selected most often; its band is read in that prior's coordinates and omitted.

Family	Operator	Weak band	Best ρ	Floor ρ	Floor cost	Target cost
Blur	2	0.086	0.01	1	0.294 [0.249, 0.339]	same setting chosen
	0.8	0.098	0.001	1	0.250 [0.200, 0.300]	0.021 [0.014, 0.029]
	1.6	0.152	0.01	1	0.315 [0.269, 0.361]	same setting chosen
	1.2	0.246	0.003	1	0.353 [0.303, 0.404]	0.009 [0.004, 0.013]
	1	0.354	0.003	1	0.364 [0.312, 0.417]	0.016 [0.010, 0.022]
Mask	0.15	0.000	30	30	same setting chosen	0.005 [0.004, 0.007]
	0.3	0.000	30	30	same setting chosen	0.007 [0.005, 0.010]
	0.5	0.000	30	30	same setting chosen	0.011 [0.007, 0.015]
	0.7	0.000	30	30	same setting chosen	0.014 [0.009, 0.018]
Graded mask	0.3	0.150	0.03	30	0.347 [0.313, 0.380]	same setting chosen
	0.6	0.300	0.03	30	0.514 [0.487, 0.541]	same setting chosen
	0.9	0.450	0.03	30	0.496 [0.473, 0.521]	same setting chosen
Face SR $\times 2$	box	0.000	0.003	30	0.003 [0.003, 0.004]	0.001 [0.000, 0.001]
	1	0.000	0.01	30	0.033 [0.029, 0.036]	same setting chosen
	1.2	0.015	0.01	3	0.099 [0.089, 0.109]	same setting chosen
	1.5	0.083	0.01	3	0.226 [0.206, 0.245]	same setting chosen
Face SR $\times 4$	box	0.000	0.01	30	0.005 [0.004, 0.006]	same setting chosen
	3	0.031	0.03	10	0.234 [0.214, 0.253]	same setting chosen
Stimulation	–	–	30	300 000	−0.019 [−0.049, 0.046]	−0.010 [−0.055, 0.076]

The account also predicts where the added error lands. We decompose the squared error against the truth into the operator's bands for the scored reconstructions at both settings. Across all 8 CIFAR-10 operators with a weak band, 0.492 to 0.836 of the squared error added by the residual criterion falls in the band, 1.9 to 8.8 times the share of directions it occupies. Under the graded mask, the determined directions improve slightly, as expected when they are fitted more closely; under the narrowest blur, of width 0.8, they take 0.507. The two thresholds matter little: across 30 pairs of them, the band tracks the cost on these operators at 0.718 to 0.907 (Appendix A.3).

5 TRANSFER TO OTHER PRIORS AND OPERATORS

Because the weak band is a property of the operator, we test whether its relation to the cost survives simultaneous changes in the prior, resolution, and dataset. We keep the rule and protocol fixed, replace the prior with the released CelebA-HQ diffusion model, and use FFHQ images (Karras et al., 2019) at 256×256. The operator is a prefilter followed by subsampling, the forward model used for super-resolution in this literature. No scored image belongs to the prior's training corpus, although all are faces at eight times the linear resolution of CIFAR-10.

Subsampling also provides the fixed-rank control. Sampling at stride k leaves a rank of at most $1/k^2$ of the dimension, and exactly that under the prefilters used here: box rows do not overlap, and a Gaussian transfer function has no zeros. The rows are shifts of one another, so $AA^\top$ is circulant on the low-resolution grid, and Equation 1 stays exact. A box prefilter averages each block before subsampling and sets all measured singular values equal, so the weak band is empty; widening the prefilter to a Gaussian pushes the measured high frequencies toward the noise floor at the same rank, as the graded masks of Section 4 do.

Across the 6 operators, each scored on 32 images, the 3 that carry a weak band give up 0.099 to 0.234 in correlation, with every interval clear of zero. The 3 without one give up at most 0.033; the two groups separate. The weak band tracks the cost at 0.842 here and at 0.890 across all 18 operators of both domains (Figure 2).

Because the band counts directions between two fixed thresholds, it fails on one operator. Among those with an empty band, that operator gives up 0.033 [0.029, 0.036], about 6.5 times the worst of the 6 others, 0.005. Its weakest visible direction sits at 1.7 times the noise, against 5.0 or more for the rest, just above the band's upper threshold of one. The floor margin, the smallest s_i above the invisible cutoff, recovers the ordering, and we report it alongside the band from here.

Correlation is scale-invariant and ignores the errors of contrast and brightness that the other metrics count. The failure persists under them. On the weak-band operators, a residual criterion gives up 7.7 to 16.4 dB of PSNR and 0.287 to 0.729 of SSIM (Wang et al., 2004). On the operators with an empty band, it gives up at most 2.2 dB and 0.149, except on the one singled out above, which loses 8.1 dB and 0.460, within the range of the weak-band operators. The metric changes which spectral summary predicts the cost. Across both domains, the weak band tracks the cost at 0.890 when accuracy is measured by correlation, 0.499 in PSNR, and 0.417 in SSIM; the floor margin, read as $-\log_{10}$ of itself so that it grows as the weakest visible direction nears the cutoff, tracks it at 0.863, 0.936, and 0.912. Correlation computed after the same clipping moves no cost by more than 0.0001: the clipping does not explain the difference. Appendix A.5 gives a possible explanation and the costs in LPIPS (Zhang et al., 2018).

On the 3 face operators that carry a weak band, 0.320 to 0.728 of the squared error added by the residual criterion falls inside it, 8.3 to 23.5 times its share of the directions. That range reaches well above the 1.9 to 8.8 of Section 4, because subsampling leaves the band a much smaller share. Where the band is narrowest, at 0.015 of the directions, the determined directions take 0.527 of the added error, more than the band does.

The residual falls all the way to the strongest setting in only 3 of the 6 face operators: under the three prefilters that create a weak band, it reaches its lowest point inside the grid. The last step of the trajectory may carry too little variance at this resolution to drive the weakest directions to an exact fit. The residual criterion then selects that interior setting and still pays the cost reported above.

A latent prior. We also swap the pixel prior for the released CelebA-HQ latent diffusion model (Rombach et al., 2022), applying Equation 1 to the decoded estimate and encoding the result again at every step (Appendix A.6). The same operators pay: the 3 with a weak band give up 0.039 to 0.118, every interval clear of zero, and the 3 without one at most 0.017. The two groups separate, and the band tracks the cost at 0.976. The costs are smaller than under the pixel prior, 0.099 to 0.234.

Accelerated MRI and compressed sensing. Two families were added after the relation above was measured, each band computed before any reconstruction (Appendix A.7). Accelerated MRI keeps lines of the Fourier transform of each face image. One coil that keeps the central lines and a random share of the rest (Zbontar et al., 2018) leaves the band empty. With eight coils of known sensitivity

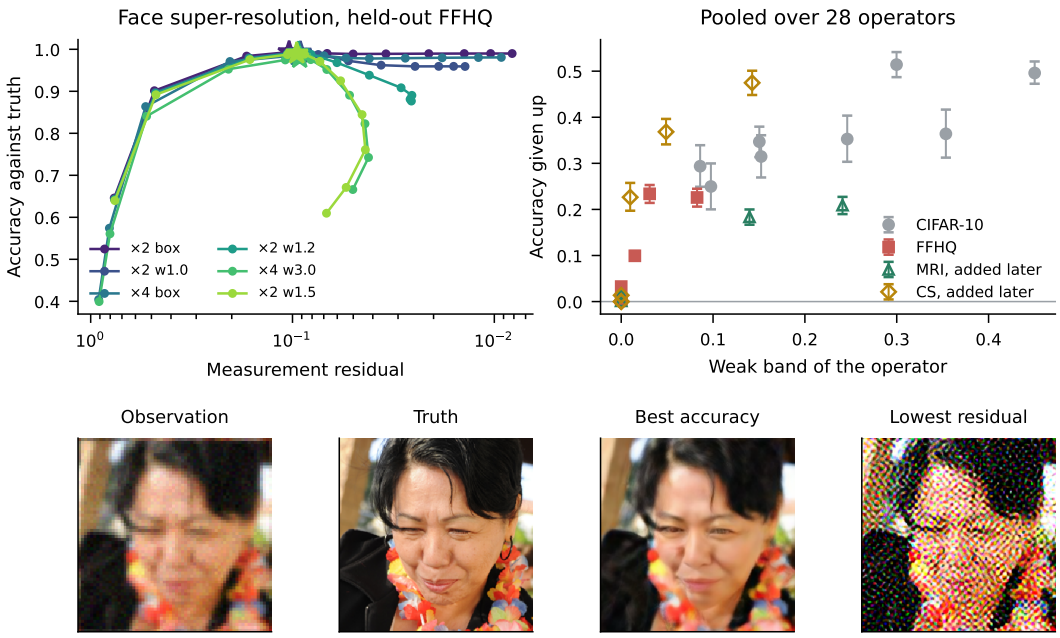

Figure 2: Super-resolution on FFHQ faces under a CelebA-HQ prior. Left: each curve sweeps the guidance strength for one operator, stars mark the most accurate setting over all images, and color runs with the weak band. Right: what each operator gives up against its weak band, for the face and CIFAR-10 operators on the same axis, with the accelerated-MRI and compressed-sensing operators added afterward as open markers. Below: an observation, the truth, and the reconstructions produced by the two settings for the operator that gives up the most.

and every second, fourth, or eighth line kept, the band grows from nothing to 0.241, because the coils cannot fully separate the pixels that the missing lines fold together. Parallel imaging measures this noise amplification by its g-factor (Pruessmann et al., 1999). Compressed sensing multiplies each channel of a CIFAR-10 image by a Gaussian matrix (Donoho, 2006) with a quarter, a half, or three quarters as many rows as pixels. Its band is empty at the noise level used so far. Raising the noise on the largest matrix moves directions into it while the matrix remains fixed. The 5 operators with a band give up 0.184 to 0.475 and the 5 without one at most 0.016. Across all 28 image operators, the band tracks the cost at 0.805. Within the two new families, the correlation is 0.964 for MRI and 0.873 for compressed sensing, compared with 0.578 when they are pooled. The band identifies costly operators across families, but ranks the size of the cost only within a family. The discrepancy principle gives up at most 0.028 on them, and projected SURE at most 0.0002.

6 The discrepancy principle

A residual has both a target and a floor. An estimate that explains the observation only to the noise level leaves a relative residual near the nominal $\sigma\sqrt{m}/\|y\|$, since $\mathbb{E}\|\eta\|^2 = m\sigma^2$; the operator and the observation supply this target without any reference image. Choosing ρ to hit it is the discrepancy principle. It needs one thing the residual criterion does not: the noise level, which in practice comes from a calibration or from a region without signal.

On the 11 operators that carry a weak band, where the residual criterion gives up at least 0.099, the discrepancy principle gives up 0.000 to 0.021 and is indistinguishable from the best available setting on 8 of them. On the 7 operators that carry none, the ordering reverses, less cleanly. The residual criterion selects the best setting on 4 of them and gives up as much as 0.033 on the rest. The discrepancy principle gives up as much as 0.014 and is indistinguishable from the best setting on 2. On every one of the 18 operators, at least one criterion pays (Table 1, last two columns).

We also test how accurate that level must be. Because Equation 1 depends on ρ and σ only through ρ/σ^2, a correction run with σ misjudged by a factor c returns at strength ρ the reconstruction that the

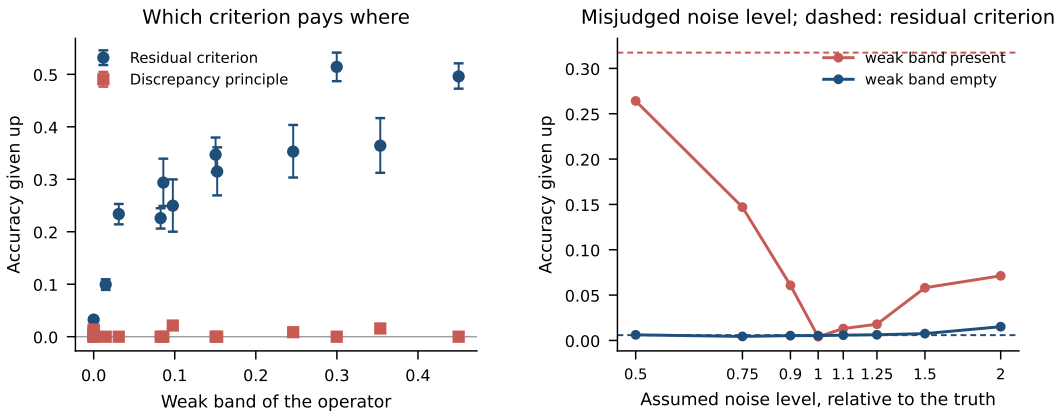

Figure 3: The two criteria that need no reference image. Left: what each gives up, against the weak band of the operator. Right: the discrepancy principle when the noise level is misjudged, averaged within each group, against the residual criterion on the same group (dashed).

true σ returns at ρ/c^2. We rescale the target alone and select again among the sweep's reconstructions. This is equivalent to running the misjudged correction on a grid multiplied by c^2. Holding the grid fixed would shift the candidates, a case not tested here. For factors between 0.5 and 2, a misjudged noise level costs more than the residual criterion on 0 of the 11 weak-band operators (Figure 3). Within 10 percent, the principle still beats the residual criterion on 10 of them, and at a factor of two on 5. The worst cost inside that narrow range, 0.294, belongs to the one operator whose residual barely reaches its target. Underestimation is more dangerous: it gives up as much as 0.514, compared with 0.322 for overestimation, because a target set too low turns the principle into the residual criterion.

A switch fitted before the operator is seen. A switch reads one number off the forward operator and stops at the noise level on one side of a threshold, minimizing the residual on the other. Threshold and side are fitted with one of the 3 operator families held out and then applied to it; super-resolution counts as one family and both mask families as another. The fit uses other operators' costs on the same scoring images, so it holds out operator families but not images. Thresholding the floor margin, the switch minimizes the residual on 0 of the 18 held-out operators: only the binary masks favor the residual criterion, and they are either held out or set the threshold above every other margin. The switch is then the discrepancy principle throughout and saves 0.192 [0.100, 0.306] over the residual criterion. Thresholding the weak band does worse: fitted on blurs and masks, it sends 5 held-out face operators to a residual floor that costs each of them, gives up 0.021 [0.000, 0.062] against the discrepancy principle, and beats it on 0 of them. These intervals use the 27 ordered draws obtained by resampling the 3 families with replacement, with every cost and threshold fixed. They describe these families only. The quantity that predicts the size of the loss does not locate the operators on which the other criterion avoids it.

Risk estimates without a reference. Stein's unbiased risk estimate (SURE) (Stein, 1981) and generalized cross-validation (Golub et al., 1979) also need no reference image. All three criteria below require the divergence of the reconstruction map. We estimate it by a finite difference, using one extra reconstruction per image and setting at $y + \epsilon b$, where b is standard normal and $\epsilon = 10^{-3}$, with the same sampler noise (Ramani et al., 2008). This estimator doubles the sampling cost and is unbiased only as $\epsilon \to 0$. We minimize three quantities, each averaged over the selecting half: SURE of the error in the measurement, $\|A(\hat{x} - x)\|^2$; generalized cross-validation, with that estimate in place of the trace of the influence matrix and no noise level; and SURE of the error projected onto the directions with $s_i > 0.1$ (Eldar, 2009), comprising the determined directions and the weak band, up to a constant independent of the setting. The first two weigh each direction by its squared singular value, as the residual does, and share its blind spot: on the CIFAR-10 operators with a weak band, both give up as much as , the worst the residual criterion itself takes there. The projected estimate weighs the weak band as heavily as the determined directions. Across all 18 operators of both domains, its costs lie between −0.002 and 0.001; every interval contains zero. It beats the

discrepancy principle on 9 of them while losing on 1. On the 6 face operators, the blind spot does not show: SURE matches the tuned setting 6 times and the projected estimate 6 times; cross-validation matches it 5 times and gives up 0.001 on the other. Like the discrepancy principle, projected SURE requires an accurate noise level. A wrong level changes its variance term and, because s_i is divided by that level, changes which directions pass the cutoff. Given levels off by the same factors, its worst cost is 0.074 within 10 percent of the truth and 0.514 when the level is halved, against 0.294 and 0.514 for the discrepancy principle. It is more forgiving near the true level, not at half that level.

In practice. The band and the floor margin need the operator's singular values and the noise level, and no reconstruction. Where the band is empty and the floor margin is 3.21 or wider, the residual criterion gave up at most 0.017 here and nothing at all behind the binary masks, and it needs no noise level; at the one narrower margin, 1.70, it gave up 0.033. Where the band is not empty, it is unsafe. With a trusted noise level, the discrepancy principle is the cheap choice. Projected SURE is safer when the pseudo-inverse above the cutoff can be applied and a second reconstruction per setting is affordable. Both fail once the noise level is badly underestimated; an uncertain level should be calibrated first. The band gives the size of the risk, not the choice between the two criteria: the switch above never beats the discrepancy principle on any of the 3 families held out.

7 THE CORRECTION RULE

The failure is concentrated in one rule. The gradient rule of Chung et al. (2023) is swept over its own grid, from 1 to 3000. Its strength parameter is on a different scale from ρ. On the scoring half, the gradient rule's residual on every blur bottoms out at a strength between 3 and 10, with a relative residual of 0.043 to 0.262. It then climbs as the step overshoots, reaching 14.1 to 15.5 times the observation's norm at the strongest step tried. The replacement rule descends further on 5 of 5 blurs, to 0.021 to 0.251, and loses most of its accuracy on that last stretch: read at the gradient rule's floor, it still scores 0.896 to 0.943, against 0.615 to 0.741 at its own. Where their residuals overlap, the two rules agree in accuracy to within 0.062. A residual criterion takes each rule to its own floor, and the gradient rule's cost is then indistinguishable from zero on 2 of 5 blurs, while the 3 that do pay give up between 0.028 and 0.074, a small fraction of what the replacement rule gives up on the same operators. The largest cost has an interval of [0.052, 0.096]. The gradient rule is protected by a floor that its fixed step does not pass, rather than by a better trade-off between accuracy and fit.

8 THE VALUE OF THE PRIOR

At a well-chosen strength, the prior helps. Under the replacement rule, where the failure is concentrated, the guided prior beats retuned shrinkage in 5 of 5 blurs by 0.003 to 0.025 in correlation. The margin grows as more directions fall below the noise floor: 0.025 where 2781 directions are undetermined, against 0.003 where only 303 are. Under the gradient rule of Chung et al. (2023), the prior wins on 3 of 5; the margin does not grow with the number of undetermined directions, and the point rests on the replacement comparison. The residual does not find that strength.

A flow prior fitted here does not help. It is trained with fixed width and depth on 7 training-set sizes from 30 to 30 000 images, with its strength tuned inside the grid (Appendix A.2). Scored on the same five blurs, it beats retuned shrinkage in 0 of 35 settings; its margin runs from −0.068 at the smallest size to between −0.011 and −0.005 at the largest. Performance saturates before the largest training set. Because the flow works on grayscale images in $[0, 1]$ with another architecture and objective, the comparison with the released prior is uncontrolled and covers this configuration alone.

9 BRAIN STIMULATION WITH THE REFERENCE WITHHELD

Focal brain stimulation is delivered at a site chosen before anyone knows what it does elsewhere. The sites already stimulated in a person constrain their response operator on part of its domain. A conditional flow prior fitted on other participants covers the rest and is conditioned on the person's resting activity, including the spread of local phase alignment across spatial scales. Every site scored here was stimulated, but its measured response is withheld from the method and from every selection,

and used only for scoring. Dividing each row by its noise standard deviation gives the noise model of Section 2 with $\sigma = 1$, except for correlations between rows, which remain (Appendix A.8).

Sweeping the replacement rule over 16 settings on real intracranial stimulation data (Thompson et al., 2020) lowers the residual at 15 of 15 steps, from 1.192 to 0.124. Pooled over participants, accuracy climbs to a plateau; past its peak at $\rho = 30$, it falls by 0.013 [−0.014, 0.038], an interval that includes zero, while the residual falls a further 4.94 times. By comparison, the weak-band image operators give up a median 0.359 past their peaks. This curve pools the participants it describes, and they disagree about their own peaks: 3 of 15 do best at the strongest setting.

Held out, each participant is scored at the settings selected on the other 14, from folds refitted without that participant. The strength selected as most accurate on the others, most often $\rho = 30$, scores 0.046. Minimizing the residual selects the strongest setting, $\rho = 300\,000$, for all participants and scores 0.065, a cost of −0.019 [−0.049, 0.046]. The discrepancy principle selects $\rho = 1$ most often and scores 0.056, a cost of −0.010 [−0.055, 0.076]; a negative cost means the criterion's setting scored higher on the held-out participant than the strength tuned on the others. Neither cost can be distinguished from zero, but the intervals are wide relative to accuracies below 0.1. The discrepancy principle stops early because the measurement is mostly noise: its nominal targets, 0.66 to 1.09 with median 0.94, are reached between $\rho = 0.3$ and 3, before accuracy reaches its plateau.

10 Related work

Guided posterior sampling. Rules that apply data consistency to the current estimate during sampling fall into two families. One replaces the estimate with a posterior mean or a projection (Wang et al., 2023; Kawar et al., 2022; Zhu et al., 2023); the other takes a gradient step through the denoiser (Song et al., 2023; Chung et al., 2023; 2022; Mardani et al., 2024). Both extend plug-and-play reconstruction, which substitutes a denoiser for an explicit prior (Venkatakrishnan et al., 2013; Romano et al., 2017; Kadkhodaie & Simoncelli, 2021), and both inherit its free parameter: how hard to pull the iterate onto the measurement (Daras et al., 2024; Ongie et al., 2020).

Latent priors. Most deployed priors work in a latent space (Rombach et al., 2022), which puts the operator and the trajectory in different spaces. Such priors have their own solvers: penalizing latents that are not fixed points of the encoder-decoder composition (Rout et al., 2023), or enforcing consistency by an inner optimization and then remapping onto the noisy manifold (Song et al., 2024).

Choosing a strength without a reference. Besides the discrepancy principle (Morozov, 1966), classical regularization offers corner rules that compare fit with solution norm (Hansen, 1992), cross-validation (Golub et al., 1979), and unbiased risk estimates (Stein, 1981; Eldar, 2009). Monte Carlo SURE estimates the divergence these need from a black-box reconstruction (Ramani et al., 2008), and under an operator that is not invertible, it covers only the error in the measured directions (Eldar, 2009). Hamiltonian Monte Carlo over the initial noise of a deterministic sampler replaces the per-step correction and needs no task-specific tuning, even at an unknown noise level (**?**).

11 Scope and future work

Every operator permits a closed form for Equation 1: each is diagonal or block diagonal in a known basis, has circulant $AA^\top$ under subsampling, or is small enough to factorize. Iterative solves, nonlinear operators, non-Gaussian likelihoods, and high-resolution natural scenes are untested. MRI uses real face images, known coil sensitivities, and Cartesian lines; clinical data, complex images, and other sampling patterns are untested. The only latent prior is an unconditional face model with a plain decode-and-encode correction. Stimulation covers one operator family and 15 participants, ignores row correlations, and tests neither the spectral account nor non-generative baselines. Scores use the average of the samples; posterior calibration is untested. Risk estimates use one probe per image; stimulation uses none. Future work should test decoupled annealing (Zhang et al., 2025), latent solvers, samplers without a per-step correction, and strengths that vary along the trajectory. Corner rules, a smooth weighting of directions in place of projected SURE's cutoff, and criteria for an unknown noise level remain to be compared. The band could be read from the Jacobian of a nonlinear operator, and from randomized singular-value estimates for an operator too large to factorize.

Ethics Statement

This work analyzes two public image corpora and one public data resource from concurrent intracranial electrical stimulation and fMRI. No new data were collected, and no human participant was contacted. The stimulation resource is de-identified and was collected under ethical approval obtained by its original investigators, as reported in its data descriptor (Thompson et al., 2020); we used only files from its public release. The face images are public and were used to evaluate a publicly released prior; no attempt was made to identify any individual, and no reconstruction of a real person is reported beyond the published corpus. The paper reports a limitation of a selection rule, not a new capability. The stimulation analysis measures how that rule behaves in 15 participants when a population model predicts one person's response; it is not clinical guidance.

Reproducibility Statement

Every number in this paper is produced by a single pipeline from public inputs; the refit tolerance quoted below is printed by a released script instead. Reruns reproduce the replacement rule's image reconstructions exactly, and refitting a stimulation prior, whose training is seeded, reproduces its curves to about 10^{-7}. The gradient rule differentiates through the denoiser, whose GPU kernels for that derivative are not deterministic: in a rerun, individual reconstructions at the strongest steps changed markedly, but no selected strength did. Section 2 states the priors, operators, noise level, and scoring rule; the operator families and their spectra are defined there in closed form. Every selection cost is scored on data other than that on which its settings are selected: one half of the images for the other half (Section 3); for the stimulation experiment, settings are selected on the other participants from folds refitted without the participant being scored (Appendix A.8). Curves, their stars, and the floors of Section 7 are descriptive. The appendix records the sampler settings, describes the stimulation data, operator, and prior, and checks the correction against the closed-form Gaussian posterior mean; the posterior covariance is not checked. Source code and the fixed analysis settings are included in the supplementary material, together with scripts that regenerate every figure, table, and number.

Use of Generative AI

Generative AI was used only to polish the writing; it was not used for the research, code, analysis, or results.

References

Yaşar Utku Alçalar, Junno Yun, and Mehmet Akçakaya. Automated tuning for diffusion inverse problem solvers without generative prior retraining. *arXiv preprint arXiv:2509.09880*, 2025.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G. Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Gustavo Deco and Morten L. Kringelbach. Turbulent-like dynamics in the human brain. *Cell Reports*, 33(10):108471, 2020.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Yonina C. Eldar. Generalized SURE for exponential families: Applications to regularization. *IEEE Transactions on Signal Processing*, 57(2):471–481, 2009.

Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, 1996.

Oscar Esteban, Christopher J. Markiewicz, Ross W. Blair, Craig A. Moodie, A. Ilkay Isik, Asier Erramuzpe, James D. Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fMRIprep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, 16:111–116, 2019.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Per Christian Hansen. Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34(4):561–580, 1992.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Zahra Kadkhodaie and Eero P. Simoncelli. Stochastic solutions for linear inverse problems using the prior implicit in a denoiser. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13242–13254, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. In *International Conference on Learning Representations*, 2024.

V. A. Morozov. On the solution of functional equations by the method of regularization. *Soviet Mathematics Doklady*, 7:414–417, 1966.

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory*, 1(1):39–56, 2020.

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.

Sathish Ramani, Thierry Blu, and Michael Unser. Monte-carlo SURE: A black-box optimization of regularization parameters for general denoising algorithms. *IEEE Transactions on Image Processing*, 17(9):1540–1554, 2008.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *International Conference on Learning Representations*, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023.

Charles M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981.

William Hedley Thompson, Reeti Nair, Hiroyuki Oya, Oscar Esteban, James M. Shine, Christopher I. Petkov, Russell A. Poldrack, Matthew Howard, and Ralph Adolphs. A data resource from concurrent intracranial stimulation and functional MRI of the human brain. *Scientific Data*, 7:258, 2020.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *IEEE Global Conference on Signal and Information Processing*, pp. 945–948, 2013.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. In *International Conference on Learning Representations*, 2023.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, et al. fastMRI: An open dataset and benchmarks for accelerated MRI. *arXiv preprint arXiv:1811.08839*, 2018.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20895–20905, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023.

A Appendix

A.1 The correction reproduces the closed-form posterior mean

When the prior is Gaussian with known covariance, the posterior mean has a closed form against which the correction can be checked. A single application of Equation 1 at prior variance v reproduces the Tikhonov solution at strength $1/v$ to a relative error below 10^{-4} in the deblurring case and below 10^{-5} in the masking case, both verified in the test suite. When a flow trained on Gaussian draws is guided with a random linear measurement, the mean of 64 samples stays within 0.068 of the

closed-form posterior mean in relative norm. The covariance of the samples is not checked. Every reconstruction using a generative prior here averages 8 to 64 samples. The results concern these averages, not posterior uncertainty. Their Monte Carlo error enters the intervals over images and is not reported separately.

A.2 Settings

Reconstructions with the released priors average 16 posterior samples over 100 sampling steps on CIFAR-10 and 8 samples over 50 steps at 256×256. The guidance grid is $\rho \in \{10^{-4}, 3 \times 10^{-4}, 10^{-3}, \dots, 1\}$ for the blur family and extends to 30 for the two mask families, where the residual floor lies further out; the gradient rule is swept over $\{1, 3, 10, 30, 100, 300, 1000, 3000\}$. Shrinkage strengths are searched over 21 logarithmically spaced values per operator. The fitted flow of Section 8 averages 32 samples over 100 Euler steps. Its guidance strength is tuned over 9 values from 0.01 to 100, and its shrinkage comparator is tuned over 29 values from 10^{-3} to 10^{4}. Both are tuned on the same 64 validation images. In every setting, the selected strength lies between 0.3 and 10, inside the grid. For the selection costs, 64 held-out CIFAR-10 test images are split evenly between selection and scoring. The intervals are not adjusted for the number of operators. Each comparison of Section 8 scores 100 test images disjoint from its validation set: 16 validation images for the released prior and the 64 above for the fitted flow. Those 64 include 48 of the released prior's test images, among them all 32 scoring images. The face experiment uses the same grid, split, and image count, with images drawn from the last FFHQ shard, at observation noise $\sigma = 0.05$ in $[-1, 1]$ units. Over the random halvings of Section 4, the central 95% range of each weak-band operator's cost is no wider than 0.063.

A.3 Where the band's thresholds are drawn

The band counts directions whose s_i lies in $(0.1, 1]$. Recounted with the lower threshold anywhere from 0.03 to 0.3 and the upper from 0.5 to 3, 30 pairs in all, the band tracks the cost across the 18 image and face operators at 0.750 to 0.907, and across the CIFAR-10 operators alone at 0.718 to 0.907. The published pair ranks 8 of 30; the best, $(0.3, 1.5]$, is not used, since choosing it here would fit the thresholds to the costs they are meant to predict (Table 2).

Table 2: Correlation between the weak band and the cost of selecting on the residual across the 18 image and face operators, as the band's two thresholds move. The published pair is in bold.

Lower	Upper threshold					
threshold	0.5	0.7	1	1.5	2	3
0.03	0.750	0.814	0.869	0.859	0.846	0.820
0.05	0.789	0.845	0.883	0.876	0.862	0.834
0.1	0.835	0.875	**0.890**	0.889	0.877	0.848
0.2	0.879	0.896	0.885	0.901	0.893	0.863
0.3	0.897	0.902	0.880	0.907	0.899	0.861

A.4 The band is not a signal-to-noise ratio

The band counts directions according to the noise injected by a least-squares fit. The operator and noise level determine this quantity. An alternative is a measure of signal against noise, which would weigh each direction by how much of the image it carries. Reading the band from the singular values of $A\Sigma^{1/2}/\sigma$, with Σ the covariance of the training images, does so, and ranks directions differently from the singular values of A. That version breaks the relation in Section 4. Across the same 12 operators, it tracks the cost at –0.222, compared with 0.891 for the band itself. Along the graded series, it falls from 0.226 to 0.106 as the cost rises and reaches 0.336 behind masks that cost nothing.

A.5 Squared error and a perceptual distance

PSNR, SSIM, and LPIPS are computed after clipping the reconstruction to the pixel range, with SSIM on intensities rescaled to $[0, 1]$. A possible explanation, not tested here, follows from squared

error. A least-squares fit adds expected squared error $\sum_i s_i^{-2}$ over the measured directions, a sum that its smallest s_i can dominate, whereas the band counts directions without weighting them. SSIM compares local luminance, contrast, and structure, so this explanation does not extend to it. LPIPS measures different things at 32^2 and at 256^2, so we read it within each domain: the band tracks its cost at 0.792 on the 12 CIFAR-10 operators and at 0.789 on the 6 face operators, and its transformed floor margin at 0.740 and 0.824.

A.6 The latent prior

The latent prior is the released CelebA-HQ latent diffusion model (Rombach et al., 2022), whose trajectory runs in the $3 \times 64 \times 64$ latent of a vector-quantized autoencoder. At each step, the estimate of the clean latent is decoded and corrected by Equation 1 in pixel space with prior variance $\rho\,(1-\bar{\alpha}_t)$ on the latent schedule. It is then clipped to the pixel range and encoded again; the reconstruction is the decoded final latent. The weak band is read in pixel coordinates, where the correction acts; the prior itself is trained on the autoencoder's latents, so here the band describes the correction, not errors in the prior's own coordinates. The 64 images, their split, the observation noise, the operators, the guidance grid, and the 8 samples over 50 steps are those of the pixel face prior. Encoding and decoding alone return the images at a mean correlation of 0.995 and at least 0.971. These values indicate the autoencoder's fidelity; they do not bound what guidance can reach. The most accurate latent settings score a median 0.019 below the corresponding pixel-prior settings. The costliest operator without a band is the box at stride four, which gives up 0.017 [0.013, 0.022]; its floor margin is wide, so neither spectral summary predicts that cost, and the lossy round trip after the last correction is one possible cause, not tested here. As under the pixel prior, the residual is monotone under the 3 operators without a weak band and not under the 3 with one. Unlike the pixel-prior case, it reaches its minimum inside the grid only under the stride-four prefilter; under the two stride-two prefilters, it rises briefly mid-range and is lowest at the strongest setting, which the residual criterion then selects. On each of the 6 operators the residual stays above that operator's own nominal target: the lowest residuals run 0.096 to 0.131 and the targets 0.093 to 0.098, which the pixel prior undercuts at 0.009 to 0.044. The latent correction's looser fit may help explain its smaller costs, but this possibility is not tested. Solvers built for latent priors (Rout et al., 2023; Song et al., 2024) replace this correction and are not tested.

A.7 Accelerated MRI and compressed sensing

For MRI, each color channel of a face image is weighted by each coil's sensitivity and transformed by the unitary two-dimensional discrete Fourier transform, and whole lines along one axis are kept; every real and imaginary entry of the kept lines carries noise of standard deviation $\sigma = 0.05$. The single-coil operator has uniform sensitivity and keeps the central 20 of the 256 lines and 44 of the remaining lines drawn at random, one line in four. This follows the fastMRI sampling pattern (Zbontar et al., 2018) but keeps a fixed number of outer lines rather than meeting the acceleration in expectation. The multi-coil operators use eight coils spaced evenly on a circle at 1.5 times the image's half-width. Each coil's sensitivity falls in inverse proportion to the distance from the coil, and its phase turns with the angle around it. The sensitivities are scaled so that their root sum of squares is one at every pixel. The correction uses their exact values, and every second, fourth, or eighth line is kept. Because the images are real, the Gram matrix of the correction is the real part of A^*A. For one coil it is diagonal in frequency, since a real image ties the coefficients at k and $-k$. For several coils it couples only the pixels that one pattern of kept lines folds together: it is block diagonal, and Equation 1 stays exact. Neither type of observation can be fitted exactly. Independent noise breaks the conjugate symmetry that ties the coefficients at k and $-k$, equal in real part and opposite in imaginary part, while a multi-coil observation has more real entries than the image has pixels. For compressed sensing, the 1024 pixels of each channel are multiplied by one Gaussian matrix with entries of variance $1/m$, for $m/n = 0.25$, 0.5, and 0.75 at $\sigma = 0.05$, and the matrix of ratio 0.75 again at $\sigma = 0.2$, 0.3, and 0.5. The images, split, grid, samples, and steps are those of the face and CIFAR-10 sweeps, and the risk estimates take one probe per selecting image (Table 3).

A.8 The stimulation problem

The data are the intracranial electrical stimulation resource of Thompson et al. (2020) (OpenNeuro ds002799), in which patients with implanted electrodes had a bipolar contact pair stimulated during

Table 3: Accelerated MRI under the face prior and compressed sensing (CS) under the CIFAR-10 prior. The band and the floor margin were computed before any reconstruction. Costs are in correlation, with 95% intervals, for the residual criterion (floor), the discrepancy principle (target), and projected SURE, each selected on one half of the images and scored on the other.

Family	Operator	Weak band	Floor margin	Floor cost	Target cost	Projected SURE
MRI, one coil	$\times 4$	0.000	14.1	0.001 [0.001, 0.002]	same setting chosen	0.0002 [0.0000, 0.0003]
MRI, eight coils	$\times 2$	0.000	6.32	0.016 [0.014, 0.017]	same setting chosen	same setting chosen
	$\times 4$	0.140	0.490	0.184 [0.167, 0.200]	same setting chosen	same setting chosen
	$\times 8$	0.241	0.100	0.209 [0.190, 0.227]	0.014 [0.012, 0.017]	same setting chosen
CS, $m/n = 0.25$	$\sigma = 0.05$	0.000	20.3	same setting chosen	0.002 [0.001, 0.003]	same setting chosen
CS, $m/n = 0.5$	$\sigma = 0.05$	0.000	8.45	same setting chosen	0.009 [0.006, 0.012]	same setting chosen
CS, $m/n = 0.75$	$\sigma = 0.05$	0.000	3.21	0.014 [0.009, 0.019]	0.015 [0.009, 0.022]	same setting chosen
	$\sigma = 0.2$	0.010	0.803	0.226 [0.197, 0.258]	same setting chosen	same setting chosen
	$\sigma = 0.3$	0.049	0.535	0.368 [0.341, 0.396]	0.028 [0.021, 0.034]	same setting chosen
	$\sigma = 0.5$	0.143	0.321	0.475 [0.448, 0.501]	same setting chosen	same setting chosen

whole-brain BOLD acquisition. We work from its released fMRIPrep derivatives (Esteban et al., 2019), registered to a template and averaged within 400 cortical parcels (Schaefer et al., 2018), and from its released event, channel, and electrode tables, which give the stimulation onsets, the stimulated contacts, and their coordinates. Of the 173 stimulation runs from 26 participants, 41 could not be prepared: 25 came from three participants without an electrode table, 12 lacked an event or channel table, and 4 had stimulated contacts that could not be located. Another 31 exceeded the motion limits: mean framewise displacement above 0.5 mm or more than 40% of volumes above 0.5 mm. Ten repeated a site already stimulated in that participant. The remaining 91 runs come from 20 participants, with one to ten runs per participant, all at different sites for that participant. A participant's runs fill numbered slots in a fixed random order, stored with the code, so a slot names a different site in each participant. A site can be held out only in a participant with at least two other sites, which leaves 15 participants and 84 participant–site pairs to score.

Each run gives a parcel-wise response over post-stimulus lags. We score the 249 parcels for which at least half of the voxels are covered in at least 95% of the 91 runs, using the same support in every fold. The delivered field is modeled as the magnitude of a bipolar point-source field at the parcel centroids, softened by 5 mm and scaled to unit peak at each site. Response and field are contrasted against the mean of the participant's other runs, the lag axis is compressed onto a basis fitted on training contrasts, and the parcel axis onto the leading 64 eigenvectors of the normalized Laplacian of a group-average connectome. The participant's response operator is restricted to a banded matrix of half-width 3 in that basis, so each stimulated site contributes linear equations $y = Ax + \eta$ in the operator's coefficients x. The noise variance is estimated per spatial mode from eight design-matched null maps for each of the participant's runs, and each row of A and y is divided by its standard deviation (Section 9); correlations between rows, which cannot be estimated from so few null maps, are ignored; Equation 1 is the Gaussian posterior mean only under this diagonal approximation. The discrepancy target $\sqrt{m}/\|y\|$ in these units needs only the marginal variances, since the expected squared norm of the noise is their sum whatever the correlations, although correlations widen its spread. After the flow is fitted, the correction multiplies the noise variance by a factor chosen from $\{0.03, 0.1, 0.3, 1, 3, 10\}$. For up to eight training participants with at least three sites, taken in a fixed order, three sites are withheld in turn from each participant's measurement and predicted from 16 samples. The factor that predicts them best is kept. The flow and the temporal basis still include those sites, so this choice is not fully held out, though no scored participant enters it. Every fit chose 10. Because Equation 1 depends on ρ and the noise variance only through their ratio, a single factor shared by every fold only rescales the strength axis. The relative residual and its target are computed without that factor, although the reconstruction depends on it.

For every participant in the prior set with at least two training sites, operator coefficients are fitted from those sites by generalized ridge regression, and again with each site left out in turn while at least two remain. A conditional flow-matching model (Lipman et al., 2023) of width 256 and depth 4 is trained on all these estimates, 39 to 78 per fitted prior (median 60) from 9 to 14 participants (median 11), for 6000 AdamW steps with an exponential moving average of its weights. Estimates from one participant are not independent, and participants with more sites contribute more of them. The flow is conditioned on descriptors of each participant's resting activity. These descriptors are computed from the phase of band-passed parcel time series using the local Kuramoto order parameter

at eight spatial scales. They comprise the spread of that parameter across parcels and time at each scale (the amplitude turbulence of Deco & Kringelbach, 2020); the correlation of its time courses between adjacent scales (the information cascade); the metastability of the global order parameter; the scaling exponents of its structure functions of orders two to six; and the share of resting power at nine spatial scales. Participants without a resting scan carry a flag and the mean descriptors of the participants used to train that fold, so no held-out data enter the fold's inputs. Coefficients are scaled by the standard deviations of a zero-mean Gaussian population prior fitted in the same fold, and the flow is trained and guided in those coordinates, so the band of this operator depends on a fitted prior. Sampling integrates the velocity field over 100 Euler steps. At each step, it applies Equation 1 to the current estimate of the coefficients in whitened units, with prior variance $\rho\,(1-t)^2/(t^2+(1-t)^2)$ on the linear path. The prediction averages 64 samples. Participants are split into four cross-fitting groups, and the prior used for a participant is fitted on the other three groups. The temporal basis has three components. To measure accuracy, the prediction and measured contrast are projected onto that basis, and their canonical-response amplitude maps are read by projecting the lag axis onto a double-gamma response. Accuracy is the correlation between these maps over parcels more than 30 mm from the stimulated contacts. The residual is $\|y - A\hat{x}\|/\|y\|$ in whitened units. The strength runs over 16 settings from 10^{-2} to 3×10^5, and every setting uses the same fitted prior within a fold.

Each participant is scored at the settings that three criteria select on the other 14 participants, never at a setting chosen on that participant's own data. The criteria select the most accurate setting, the one with the smallest residual, and the one whose residual is closest to its nominal target. Each criterion is averaged over the other participants' site pairs, just as the image criteria are averaged over images. No quantity fitted for the other curves may include the scored participant, so every fold is built repeatedly: once with the held-out group removed, giving the curves that are scored, and once for each other scored participant, with that participant also removed from the fold's training data, including its temporal basis and population prior. This takes 399 fitted priors over 33 folds. The intervals resample participants and repeat every selection inside each of 2000 draws; the fitted priors are held fixed, so the intervals omit the variation that refitting would add. A single split into a selecting and a scoring half, the protocol used for images, is unreliable here. In a descriptive analysis of the scored curves, which are not refitted for each split, the residual criterion's cost across all 6435 ways to select on seven participants and score on eight ranges from −0.118 to 0.024, with median 0.