ACTIVE DATA ACQUISITION WITH SIDE INFORMATION VIA DISCRETE DIFFUSION PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Acquiring data is costly: higher measurement fidelity costs power and storage and risks collecting irrelevant content, while aggressive cost reduction can discard information that later analysis needs. We address this trade-off with an information-theoretic framework that acquires data relevant to a broad set of tasks rather than to one model. A mask policy, conditioned on side information, chooses which pixels to measure so as to maximize the mutual information between a discrete image and its partial observation under a budget; since the image entropy does not depend on the mask, this is equivalent to minimizing the conditional entropy. A frozen discrete denoising diffusion model (D3PM) supplies the posterior, and we use it in two ways: as an entropy surrogate for training a one-shot mask generator, and as the criterion for sequential greedy acquisition. The one-shot generator outperforms random masks only with care, including an unbiased gradient estimator for binary masks. With sequential acquisition, on MNIST the prior makes $8\times$ fewer errors than random at a 10% budget, and on CIFAR-10 it gains 0.9–3.4 dB. On fastMRI, our proposed technique using a static mask outperforms the well-known methods such as variable density and LOUPE.

Keywords: Mutual information; active sensing; discrete diffusion; side information; accelerated MRI.

1 INTRODUCTION

Measuring everything is expensive, and measuring too little discards information later tasks need. In budget-constrained pixel sensing, such as medical imaging, remote sensing and bandwidth-limited transmission, only a fraction of locations can be measured, and the question is which ones carry the most information about the whole picture. In real-time systems such as autonomous driving, sensor bandwidth, power and latency limit how much can be measured and processed for each frame (Gehrig & Scaramuzza, 2024), so choosing what to measure also reduces the computation downstream; adaptive LiDAR, for instance, samples only part of the scene (Bergman et al., 2020; Pittaluga et al., 2020).

We take an information-theoretic view. Rather than tailoring acquisition to one downstream model, we score measurements by their mutual information with the target, a model-agnostic quantity (Cover & Thomas, 2006). A sensing action X is chosen from side information Z, i.e., a class label, a coarse summary of the scene and a budget, to produce a measurement Y with maximal information about the target C (Figure 1). Here the sensing action is a binary mask over pixels, the setting of our experiments. Learned masks have been trained jointly with a reconstructor (Bahadir et al., 2020; Huijben et al., 2020; Aggarwal & Jacob, 2020), as sequential policies that choose the next measurement from the current reconstruction (Zhang et al., 2019; Pineda et al., 2020; Bakker et al., 2022), and, in Bayesian experimental design, against a bound or neural estimate of the mutual information between measurements and the quantity of interest (Kleinegesse & Gutmann, 2020; Foster et al., 2021); Vuong et al. (2025) learn masks for continuous images this way, pairing a U-Net mask generator (Ronneberger et al., 2015) with MINE (Belghazi et al., 2018). Here the target is discrete, and a frozen absorbing-state D3PM (Austin et al., 2021) supplies a surrogate for $H(C \mid Y)$.

Training a mask generator lowers the value of this surrogate objective, but a lower objective value does not by itself mean better masks. We therefore judge every mask by a criterion it was not trained

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.

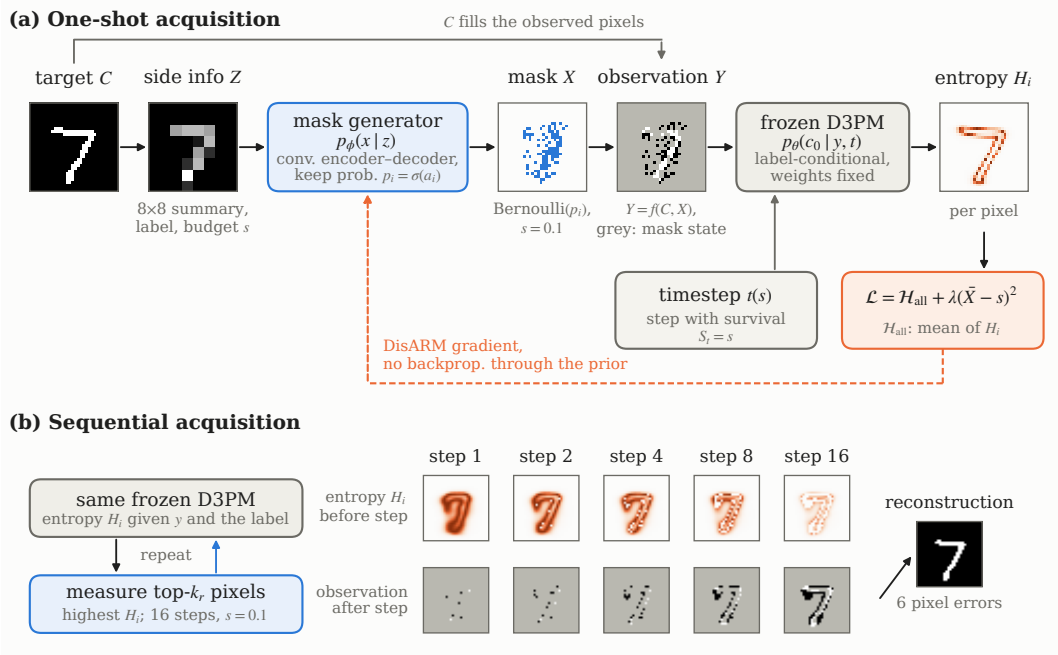

Figure 1: The two uses of the prior, on one MNIST test digit at $s = 0.1$. (a) Training the one-shot mask generator. From the side information Z, the generator gives each pixel a keep probability p_i; a sampled mask X yields the observation Y, in which unobserved pixels take the mask state (grey). The frozen D3PM, queried at the timestep $t(s)$ whose survival matches the budget, gives per-pixel entropies H_i, and the loss is their mean plus the budget penalty. The generator is trained with DisARM from the loss of antithetic mask pairs; no gradient passes through the prior. (b) Sequential acquisition with the same prior: starting from no observations, with only the label, each of 16 steps measures the unobserved pixels of highest entropy and updates the posterior. Insets are outputs of the trained generator and prior.

on: how well the image is reconstructed from the pixels the mask observed; on accelerated MRI, this is the reconstruction error of a network that played no part in choosing the mask.

Those checks separate two ways of optimizing the same objective. In **one-shot** acquisition, a mask generator is trained against the surrogate and chooses the whole mask before anything is measured. In **sequential** acquisition, the posterior is evaluated after each round of measurements and the next pixels are chosen where it is most uncertain, similar to Nolan et al. (2025). For noiseless pixel observations that uncertainty is exactly the information gained, $I(C; Y_i \mid y) = H(C_i \mid y)$, so the sequential method is greedy maximization of $I(Y; C)$ itself.

1.1 Contributions

- We cast budgeted acquisition with side information as maximizing $I(Y; C)$ under a budget constraint for discrete images. The objective is task-agnostic: it values a measurement by what it reveals about the whole target, not by one reconstructor's or classifier's loss, so one mask serves many downstream tasks. The policy conditions on side information Z (class label, coarse summary, budget), and a frozen D3PM, queried at a timestep calibrated to the budget, gives a tractable surrogate for the conditional entropy $H(C \mid Y)$. We measure what each part of Z contributes (Table 1).
- We compare two ways of using the prior. One-shot optimization of the entropy surrogate needs care: averaging over observed pixels is degenerate, since it rewards observing predictable background, and the straight-through Gumbel-Softmax gradient points away from the whole-image objective. Trained with an unbiased estimator (DisARM), the one-shot mask makes fewer errors than random masks on MNIST at every budget, but on CIFAR-10 its gain is under 1 dB and appears only with the target's own summary. Sequential greedy

information gain works better: in 16 steps it makes $8\times$ fewer errors than random on MNIST at $s = 0.1$, and on CIFAR-10 it improves on random by 0.9–3.4 dB PSNR.

- We also test on the fastMRI dataset: the one-shot surrogate disagrees with reconstruction quality, instance conditioning does not contribute meaningful improvement, mask ranking depends on the reconstructor, and a static mask optimized against a reconstructor has lower reconstruction error than variable density and LOUPE (Bahadir et al., 2020).

2 Related Work

Active learning and sensing. Active learning selects samples to label for a particular model, for instance by their mutual information with the unknown labels (Guo & Greiner, 2007); Shayovitz & Feder (2021) minimize conditional information in a universal framework. Radar detection and classification likewise optimize for a fixed task (Kay, 1998; Han et al., 2021). Scoring a batch by the sum of its items' marginal entropies is a known failure mode, because it ignores redundancy between the selected items (Kirsch et al., 2019), and our per-pixel surrogate inherits that weakness. We instead score spatial acquisition masks by mutual information under explicit budgets.

Prior acquisition work. Acquisition has been cast as Bayesian experimental design, choosing measurements to maximize their expected information about the quantity of interest (Kleinegesse & Gutmann, 2020; Foster et al., 2021), and, for imaging, as a sequential policy trained on reconstruction error (Zhang et al., 2019; Pineda et al., 2020). Vuong et al. (2025) pose acquisition with side information as a family of problems (P1–P4) and estimate $I(Y; C)$ with MINE on CelebA (Liu et al., 2015); we follow the budgeted form P4, a policy $p(x \mid z)$ under a sparsity constraint, and move to discrete states with a D3PM prior. Closest to our setting, Nolan et al. (2025) drive active subsampling with a pre-trained diffusion model under a maximum-entropy criterion, selecting the locations whose measurements are least predictable, the direction our observed-pixel form inverts, and the direction our sequential method follows. Our method differs in using a discrete absorbing prior, in reading each pixel's information gain in closed form from one forward pass per round rather than from posterior samples, and in showing that a separate mask state matters for absorbing priors (Appendix E).

Discrete and continuous diffusion. D3PMs (Austin et al., 2021) extend denoising diffusion (Ho et al., 2020) to categorical data; the absorbing-state kernel matches masked prediction. We use an absorbing D3PM with a cosine schedule (Nichol & Dhariwal, 2021), built on the implementation of Ryu (2024), and a DiT backbone (Peebles & Xie, 2023) on CIFAR-10. Masked diffusion of this kind now scales to large language models, trained on a bound on the likelihood (Nie et al., 2025); we use the prior differently, as a frozen scorer, and our surrogate is a mean of per-pixel entropies of its c_0-predictor. The usual relaxation for training over binary masks, Gumbel-Softmax (Jang et al., 2017; Maddison et al., 2017), gives a gradient that points the wrong way here, so we train with DisARM (Dong et al., 2020).

Learned sampling for MRI. Variable-density random sampling with a fully sampled low-frequency block is the standard Cartesian heuristic (Lustig et al., 2007). Learned patterns follow two lines of work: combinatorial selection, building masks greedily against reconstruction error (Gözcü et al., 2018), and relaxation, training a probabilistic mask jointly with a reconstruction network, as in LOUPE (Bahadir et al., 2020). Our MRI experiments re-derive both to test whether an information-theoretic, instance-adaptive policy improves on them; it does not, and we run LOUPE itself as a baseline. Bakker et al. (2022) already found adaptive acquisition policies roughly matching non-adaptive ones on multi-coil fastMRI, so our conditioning result replicates theirs in the single-coil Cartesian setting rather than establishing something new. Alkan et al. (2025) maximize a variational mutual-information bound for 3D k-space design with the reconstructor trained jointly, and improve on variable density, the regime our own positive result points to.

3 Problem Formulation

We write C for the target discrete image, X for the sensing action (a binary mask) chosen by a policy $p(x \mid z)$, Y for the measurement, a partial observation of C, and Z for the side information. Images

are discretized to N bins per channel, so $C \in \{0, \dots, N-1\}^{C_{\text{ch}} \times H \times W}$ over C_{ch} channels. A mask $X \in \{0,1\}^{1 \times H \times W}$ selects observed pixels, and $Y = f(C, X)$ replaces every unobserved pixel with the D3PM's absorbing token. The token has a state of its own: 0 means unobserved and nothing else, and the data bins are $1, \dots, N$, giving 3 states on MNIST (0 = mask, 1 = black, 2 = white) and 9 on CIFAR-10. Were the token to share a value with the data, an observed black pixel and an unobserved pixel would be the same input to the prior; Appendix E measures what the separate state adds. Side information Z is an 8×8 grayscale summary of C, the class label (MNIST), and a budget s drawn per sample from $\mathcal{U}(s_{\min}, s_{\max})$.

Problem.

$$\max_{p(x|z)} \; I(Y;C) \quad \text{s.t.} \quad \mathbb{E}\Big[\tfrac{1}{HW}\textstyle\sum_i X_i\Big] = s. \tag{1}$$

Since $I(Y;C) = H(C) - H(C \mid Y)$ and $H(C)$ does not depend on X, this is equivalent to minimizing $H(C \mid Y)$. Optimal solutions of this lie at extreme points of convex feasible sets, and enumerating them is intractable in high dimensions (Vuong et al., 2025); we instead parameterize $p_\phi(x \mid z)$ and minimize a differentiable surrogate.

4 Method

4.1 A D3PM surrogate for $H(C \mid Y)$

A frozen absorbing-state D3PM (Austin et al., 2021) with denoiser $p_\theta(c_0 \mid c_t, t, c_{\text{lbl}})$ is applied to the observation Y at a timestep $t(s)$ matched to the budget. Its per-pixel entropy is

$$H_i = -\sum_{k=0}^{N-1} p_i(k) \log p_i(k), \qquad p_i = \text{softmax}(\hat{z}_i). \tag{2}$$

Subadditivity makes $\sum_i H_i$ an upper bound on the joint conditional entropy under the model. Instead of the sum, we average H_i in one of two ways:

Observed-pixel aggregation. Averaging H_i over the observed pixels only, our original design, measures how predictable the observed pixels are rather than $H(C \mid Y)$, so it rewards observing whatever is easiest to predict: on MNIST the background, and with a weak budget penalty a nearly empty mask. We therefore average over every pixel.

Whole-image aggregation. Averaging over every pixel,

$$\mathcal{H}_{\text{all}} = \frac{1}{HWC_{\text{ch}}} \sum_i H_i, \tag{3}$$

so a mask can lower the objective's value only through what its observations add to Y. On MNIST, observing a background pixel leaves Y unchanged and earns nothing.

The D3PM is queried at a timestep $t(s)$ matched to the budget. Each step of its forward process absorbs more pixels into the mask token, so a mask that keeps a fraction s of the pixels looks like a forward-corrupted image at the step where a fraction s of pixels still survive. We estimate this survival curve, $S_t = \mathbb{E}_C\big[\frac{1}{HWC_{\text{ch}}} \sum_i \mathbf{1}[c_t^{(i)} = c_0^{(i)}]\big]$, the expected fraction of pixels not yet absorbed at step t, by Monte Carlo on training images, and set $t(s)$ to the step with $S_t = s$, so the prior sees the kind of input it was trained on at that step. S_t falls monotonically from 1 to 0, so each budget has one timestep; we estimate it on 20 training batches at every tenth step and interpolate (on MNIST, $s = 0.1$ gives $t = 935$ of $T = 1000$). The timestep follows the target budget, not the density a mask actually achieves: a mask far below budget presents many more mask tokens than the prior expects at $t(s)$, so it is queried off its training distribution. The surrogate is therefore only meaningful when the budget holds.

4.2 Mask generator and objective

The mask generator concatenates the 8×8 summary in Z with an MLP embedding of s (and a class embedding on MNIST), each broadcast over the 8×8 grid, and decodes per-pixel keep/drop logits

with a convolutional encoder–decoder (Figure 1). Each pixel is kept independently with probability $p_i = \sigma(a_i)$, a_i the difference of its two logits, and the generator is trained with DisARM (Dong et al., 2020), an unbiased gradient estimator for binary masks; Section 5 explains why we do not use straight-through Gumbel-Softmax. The loss is

$$\mathcal{L} = \mathcal{H}_{\text{all}} + \lambda\, \mathbb{E}\big[(\bar{X} - s)^2\big], \tag{4}$$

where $\bar{X}$ is the fraction of pixels a sampled mask keeps and s is drawn per sample from the budget range (Table 5). The first term is the whole-image surrogate computed on that mask's observation at $t(s)$; the second, weighted by λ, holds the expected density at the budget. The D3PM is frozen: its weights never change, it only scores the observations of sampled masks, and no gradient passes through it, since DisARM estimates the gradient with respect to the keep logits from the loss values of antithetic pairs of masks.

4.3 Sequential greedy information gain

Measuring pixel i noiselessly reveals C_i, so its information gain given the current observation y is

$$I(C; Y_i \mid y) = H(Y_i \mid y) - H(Y_i \mid C, y) = H(C_i \mid y), \tag{5}$$

the prior's entropy at i. Greedy acquisition therefore needs no surrogate: at each acquisition step it measures the k_r unobserved pixels of highest $H(C_i \mid y)$, then recomputes the posterior. Taking all k pixels in one step is suboptimal (Kirsch et al., 2019); splitting the budget over R steps lets each step see what the last one revealed. The first step starts either from a small random probe or from no observations at all, in which case the label in the side information is the only thing placing the first pixels. For this method to work the prior must distinguish "observed black" from "unobserved", which the separate mask state of Section 3 provides.

5 One-shot masks on MNIST and CIFAR-10

5.1 Setup

We compare five methods here and in Section 6: **random** (budget-exact uniform masks), **variable density** (density decaying away from the image center), **LOUPE** (Bahadir et al., 2020) (a learned probabilistic mask, ported to the pixel domain and trained per budget with its own U-Net), **one-shot (ours)** (the generator trained on $\mathcal{H}_{\text{all}}$ with DisARM) and **sequential (ours)** (greedy information gain over 16 steps), all reconstructed by the same prior unless stated otherwise. Both datasets use the encoding of Section 3. MNIST is padded to 32×32 and binarized (N=3 states); CIFAR-10 is $32 \times 32 \times 3$ with 8 intensity bins and the mask token (N=9). The frozen D3PMs use T=1000 steps and a cosine schedule: a small convolutional c_0-predictor on MNIST and a DiT (d=1024) on CIFAR-10. Mask generators are trained with $\mathcal{H}_{\text{all}}$ for 200 epochs. All numbers are on the hold-out test split (2560 images), over 5 seeds. Hidden pixels are reconstructed at the D3PM's argmax; we report errors per image and foreground recovery (digit pixels recovered) on MNIST, and PSNR with hidden pixels at the posterior mean on CIFAR-10. Learned masks are drawn on budget: the $\text{round}(sHW)$ pixels with the largest keep log-odds plus logistic noise. Appendix A gives the hyperparameters and the full evaluation protocol.

Gradient estimator. The straight-through Gumbel-Softmax gradient does not train these generators. Against the exact gradient of the expected loss, estimated by flipping one pixel of a sampled mask at a time, it has cosine -0.73 to -0.97 on the generator's weights, so it points uphill, and no checkpoint trained with it reached a lower objective than a random mask. DisARM (Dong et al., 2020) has cosine 0.94 to 0.99 at the same cost, and under it the objective falls from the first epoch (Appendix A).

Results. On MNIST the generator lowers $\mathcal{H}_{\text{all}}$ below random at every budget, and the gain carries over to criteria it was not trained on: at $s = 0.1$ it recovers 90% of digit pixels against 78% and makes 23.0 errors per image against 46.7, and 41% of its observed pixels are digit pixels, against 9% for a random mask (Figure 2; every budget and metric in Appendix B).

The generator sees an 8×8 summary of the target image (Table 1). Replacing it with another test image's summary raises errors at $s = 0.1$ from 23.0 to 26.8, and removing it gives 26.4, still

Table 1: What the one-shot generator gets from its side information: the target's own 8×8 summary, another test image's summary, or none (an all-zero summary). Budget-exact sampled masks, hold-out test split, 5 seeds; seed spread at most 0.10 errors per image and 0.02 dB. On MNIST the generator outperforms random even without the true summary; on CIFAR-10 its gain needs it.

	MNIST (errors/img)		**CIFAR-10** (PSNR, dB)		
Summary	$s = 0.1$	0.3	$s = 0.1$	0.3	0.5
target's own	23.0	4.74	20.45	24.80	28.01
another image's	26.8	5.22	20.12	24.25	27.40
none	26.4	5.31	19.89	23.78	26.42
random (no generator)	46.7	24.0	20.14	24.33	27.47

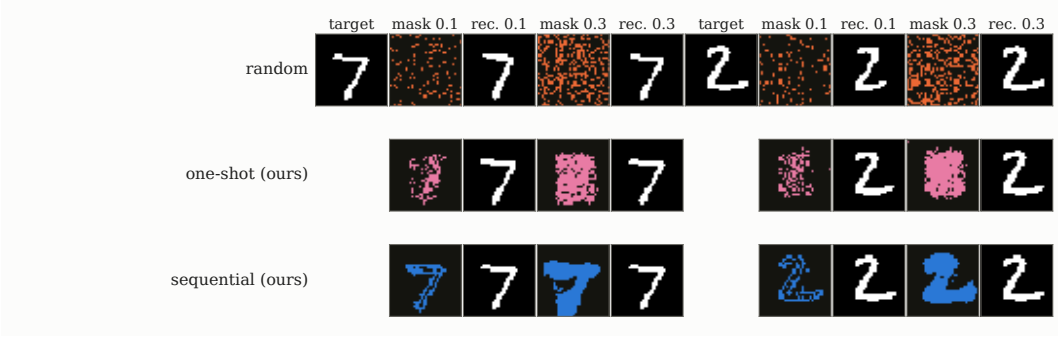

Figure 2: MNIST test images at $s = 0.1$ and 0.3, one row per method; masks are tinted by method, and each reconstruction keeps the observed pixels and fills the rest from the prior. The one-shot generator, trained with DisARM and sampled on budget, covers the region where the digit lies, densely by $s = 0.3$; sequential acquisition (label-greedy, 16 steps; Section 6) places its pixels on and around the stroke, because the prior's entropy concentrates there, as Figure 1(b) shows step by step. On all 2560 test images the one-shot generator makes 22.9 errors per image at $s = 0.1$ against 46.6 for random (4.7 against 24.0 at $s = 0.3$), and sequential acquisition makes 5.6 and 0.01.

far below random (46.7). With deterministic top-k selection, the mode of the learned masks, the summary matters more (27.6, 39.7 and 42.1 against 46.5), and by $s = 0.3$ it hardly matters (3.66 to 4.54 errors against 24.0). Sequential acquisition, which uses only the label, still makes about $4\times$ fewer errors at $s = 0.1$ (5.6; Section 6).

On CIFAR-10 the gain is small, 0.3 to 0.7 dB PSNR, against a seed spread of at most 0.02 dB. The whole gain comes from the target-derived summary (Table 1): with another image's summary the generator is level with random, and with none it is worse. At $s = 0.1$ its objective ends slightly above random's while its PSNR keeps improving, and a plain Bernoulli draw overshoots the budget (density 0.187) and appears to gain 2.6 dB, most of it from the overshoot (Appendix B).

6 Sequential acquisition

We apply Section 4.3 to the MNIST prior on the same 2560 test images. Strategies share images, probes and random masks, so comparisons are paired, and the probe takes 20% of the budget. Because recovery saturates once whole digits are reconstructed, we report errors per image and the fraction of images reconstructed exactly (Table 2).

A single greedy step after a random probe already removes 44% of the errors a random mask leaves at $s = 0.1$, and 92% at $s = 0.3$. The best method starts from no observations and takes 16 steps: at $s = 0.1$ it leaves 5.6 errors per image against 46.7 for random, and at $s = 0.2$ it reconstructs 92% of test digits exactly, against none for random. Starting from the label is better than starting from a random probe at every budget: the class-conditional prior already knows where a digit of that class is uncertain. On CIFAR-10 the opposite holds: label-greedy with 16 steps is level with a random mask at $s = 0.1$ and 0.3 and about 2 dB behind a random probe followed by variance-greedy steps, since a CIFAR-10 class label says little about where an image's informative pixels are. Against the

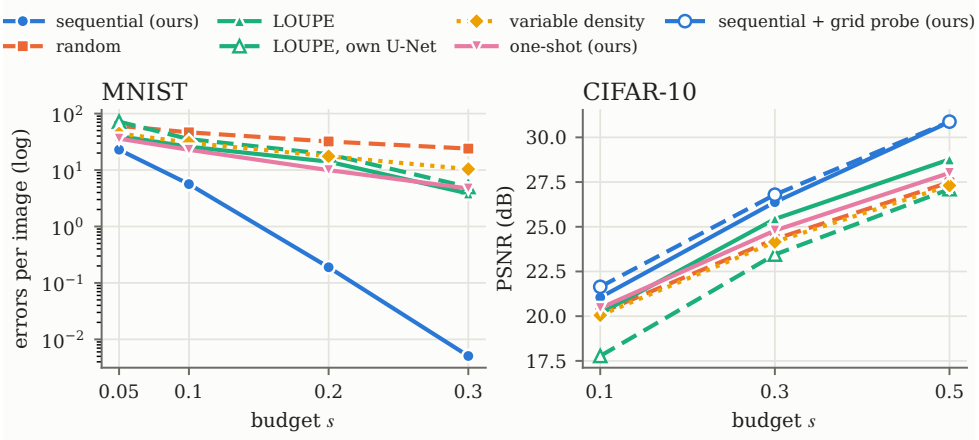

Figure 3: Sequential acquisition against the baselines on the test split; every mask is reconstructed by the same prior except "LOUPE, own U-Net". MNIST: label-greedy with 16 steps, errors per image (log scale). CIFAR-10: variance-greedy with 16 steps, alone and after a jittered-grid probe, PSNR. The grid probe is used on CIFAR-10 only: on MNIST, starting from the label is better than starting from a probe (Table 2). LOUPE is shown at the better of its top-k and budget-exact sampled masks (Appendix C). The one-shot mask sees the target's own 8×8 summary.

Table 2: MNIST errors per test image (lower is better; fraction of images reconstructed exactly in parentheses).

Strategy	$s = 0.05$	$s = 0.1$	$s = 0.2$	$s = 0.3$
random	60.5 (0%)	46.7 (0%)	32.1 (0%)	24.0 (0%)
probe + greedy, 1 step	40.0 (0%)	26.1 (1%)	9.22 (14%)	1.97 (50%)
probe + greedy, 16 steps	26.9 (4%)	8.45 (17%)	0.42 (83%)	0.02 (99%)
label-greedy, 4 steps	26.8 (4%)	10.0 (14%)	0.64 (81%)	0.03 (99%)
label-greedy, 16 steps	**22.9** (6%)	**5.63** (26%)	**0.19** (92%)	**0.01** (100%)

baselines (Figure 3), sequential acquisition outperforms LOUPE, variable density and the one-shot mask at every budget on both datasets; on CIFAR-10 it reaches 21.6, 26.8 and 30.9 dB at $s = 0.1$, 0.3 and 0.5, 1.2 to 2.9 dB above the one-shot mask. Appendix C gives the per-step curves and example masks.

7 FastMRI

7.1 Setup

We use the fastMRI single-coil knee data (Zbontar et al., 2018), one mid-slice per volume: 973 training and 199 hold-out validation slices at 300×300 pixels. A mask selects rows of k-space around a 32-row auto-calibration (ACS) block that is always sampled, and the budget is enforced exactly by top-k. Side information Z is a central scout reconstruction, and the one-shot policy is an instance-adaptive row selector trained on zero-filled reconstruction error plus the coarse row-profile entropy surrogate of a frozen D3PM. Two U-Nets are trained on the training split under heuristic masks only: **U-Net A** (1.9M parameters), which masks are optimized against, and **U-Net B** (4.3M parameters, different seed and mask draws), a hold-out judge used only for scoring. We report per-slice NMSE, $\|\hat{c} - c\|^2/\|c\|^2$, under U-Net B; stochastic baselines are averaged over 5 seeds, and a margin inside two standard deviations is not counted as an improvement.

7.2 Results

The entropy surrogate disagrees with reconstruction quality. At $s = 0.25$ the learned policy attains the lowest entropy of any method and still loses on NMSE to every heuristic tested. Table 3 explains why: masks that spread out gain most from a reconstructor, while center-concentrated

masks gain least, because a reconstructor with a prior already predicts the low-frequency center: spread-out masks (ACS + equispaced or random) gain about 30%, while the energy oracle and the learned policies gain only 14–16%.

Table 3: Val NMSE at $s = 0.25$. Gain is the relative NMSE reduction from zero-filled to U-Net B reconstruction, $1 - \mathrm{NMSE_B}/\mathrm{NMSE_{zf}}$. Center-concentrated masks gain least from a reconstructor that carries a prior.

Mask	Zero-filled	U-Net B	Gain
ACS + equispaced	0.03820	0.02623	31.3%
ACS + random	0.03819	0.02696	29.4%
static search (ours)	0.03283	0.02389	27.2%
ACS + variable density	0.03382	0.02474	26.8%
learned policy, static (ours)	0.03431	0.02887	15.9%
learned policy, adaptive (ours)	0.03337	0.02840	14.9%
energy oracle	0.03147	0.02715	13.7%

Table 4: Hold-out val NMSE under the judge U-Net B. ACS + VD: mean ± sd over 5 seeds. Sequential: rows ranked by the coarse row-profile prior's entropy, 16 steps.

s	Static search (ours)	ACS + VD	LOUPE	Sequential
0.25	**0.02389**	0.02471 ± 0.00006	0.02759	0.02589
0.40	**0.01860**	0.01889 ± 0.00005	0.02213	0.01992
0.50	**0.01523**	0.01546 ± 0.00007	0.01995	0.01559
0.75	**0.00751**	0.00764 ± 0.00001	0.00966	0.00807

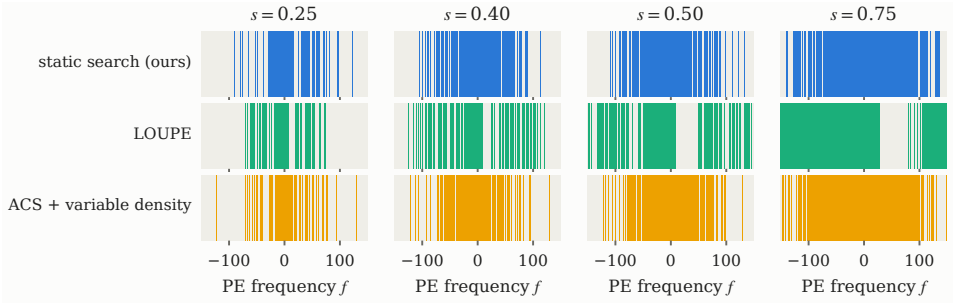

Figure 4: Phase-encode rows sampled by the static search mask, LOUPE and ACS + variable density ($f = 0$ is DC); Figure 15 in Appendix D shows every static mask.

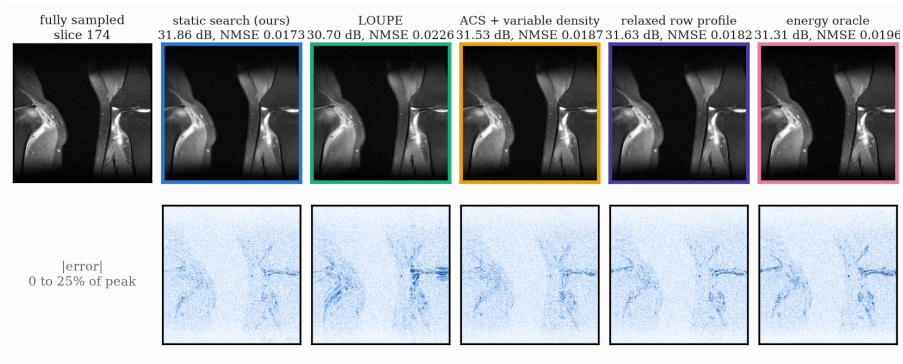

Figure 5: A validation slice of median difficulty for ACS + variable density, at $s = 0.25$, reconstructed by the hold-out judge. Top: reconstructions. Bottom: absolute error |reconstruction − fully sampled|, from 0 (white) to a quarter of the slice's 99.5th-percentile intensity (dark blue). PSNR and NMSE are per slice.

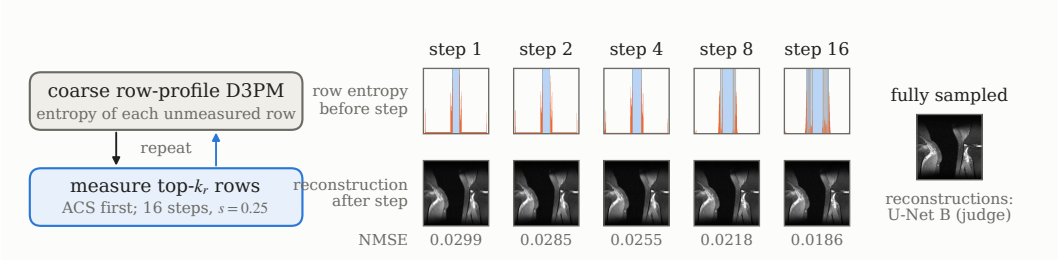

Figure 6: Sequential acquisition on the slice of Figure 5 at $s = 0.25$: from the ACS block, each of 16 steps adds the unmeasured k-space rows of highest prior entropy. Top: per-row entropy before the step (DC in the middle; blue: measured rows). Bottom: U-Net B reconstruction after the step, with its NMSE.

Optimizing the discrete mask against the reconstructor works better. A greedy row-swap search on the hard mask, in which the gradient proposes candidate swaps and exact evaluation on the training split accepts them, is best at every budget (Table 4), also on PSNR and under zero-filled reconstruction; a little over half of its gain comes from starting at the best fixed variable-density draw, the rest from the swaps. LOUPE, restricted to whole phase-encode rows, is worse than every heuristic (Figure 4 shows the rows each mask samples), and a relaxed row profile trained through a straight-through top-k mask improves on variable density only at $s = 0.50$. Sequential acquisition with the only prior that trained well here, a coarse row-profile D3PM, loses to variable density at $s = 0.25$ and 0.40 and never reaches the search mask's NMSE: after the ACS block the coarse profile carries little slice-specific information. Figure 5 shows a validation slice of median difficulty at $s = 0.25$: the search mask has the lowest NMSE (0.0173, against 0.0187 for variable density), and LOUPE's errors are largest along tissue edges. Figure 6 follows sequential acquisition on the same slice: its 16 steps lower NMSE from 0.0299 to 0.0186, level with variable density and behind the search mask. Appendix D gives the remaining figures, every sequential variant and the LOUPE details.

8 Discussion and Conclusion

We cast budgeted acquisition with side information as maximizing $I(Y; C)$ and used a frozen absorbing D3PM in two ways: as a surrogate for $H(C \mid Y)$ that a one-shot mask generator is trained against, and as the posterior for sequential greedy acquisition. The two behave differently. For noiseless pixels, the criterion of sequential acquisition is the information gain itself: in 16 steps it leaves $8\times$ fewer errors than random on MNIST at $s = 0.1$ and gains 0.9–3.4 dB on CIFAR-10. The one-shot generator gains less, and only with care: it needs an unbiased gradient estimator, whole-image averaging and budget-exact sampling, and on CIFAR-10 its whole gain comes from the target's own summary. The prior's encoding matters too: a separate mask state gives 10–15% fewer errors than a prior in which the mask shares its value with black (Appendix E).

Limitations. The surrogate averages per-pixel entropies, an upper bound that ignores dependence between pixels, so greedy steps can pick redundant pixels (Kirsch et al., 2019); the categorical entropy also ignores the order of intensity bins, one reason the objective and PSNR can part ways on CIFAR-10. The information-gain identity assumes noiseless measurements. Our images are small (32×32), and sequential acquisition costs one pass of the prior per step, a one-shot mask one pass of the generator. On fastMRI the coarse row-profile prior carries little slice-specific information, so sequential acquisition trails a static mask optimized against a reconstructor.

Future work. Stronger k-space priors, batch-aware criteria, noisy measurements, and an amortized sequential policy with the speed of a one-shot mask, which would suit real-time sensing such as adaptive LiDAR in autonomous driving (Bergman et al., 2020).

AI USE STATEMENT

We used Claude Code (Anthropic) to assist with quick model architecture experimentation, code debugging, and monitoring training jobs; Claude was also used to check grammatical errors, detect ambiguous or confusing phrasing in some paragraphs, and help with exploring alternative designs; Gemini and Claude were also used to select pleasant color palettes for plotting, and help with formatting LaTeX figures correctly to avoid overflow. LLM-based Google Scholar Labs was used to identify relevant references. The research question, problem formulation, core implementation, and main writing are the authors'; no LLM was used during the ideation and proof-of-concept phase. The authors reviewed all code, results and text and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

All datasets are public: MNIST, CIFAR-10 and the fastMRI single-coil knee data (Zbontar et al., 2018). Sections 4 and 5 and Table 5 give the priors, the mask-generator architecture and its training settings; Sections 4.3 and 6 the sequential acquisition procedure; Section 7 the reconstructors, masks and splits used on fastMRI; and Appendix E the comparison of priors. Every image result is on the hold-out test split with fixed evaluation seeds, and we report the spread over seeds. We will release the code, trained checkpoints and evaluation scripts after the review period.

REFERENCES

Hemant Kumar Aggarwal and Mathews Jacob. J-MoDL: Joint model-based deep learning for optimized sampling and reconstruction. *IEEE Journal of Selected Topics in Signal Processing*, 14: 1151–1162, 2020.

Cagan Alkan, Morteza Mardani, Congyu Liao, Zhitao Li, Shreyas S. Vasanawala, and John M. Pauly. AutoSamp: Autoencoding k-space sampling via variational information maximization for 3d MRI. *IEEE Transactions on Medical Imaging*, 44(1):270–283, 2025.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Cagla D. Bahadir, Alan Q. Wang, Adrian V. Dalca, and Mert R. Sabuncu. Deep-learning-based optimization of the under-sampling pattern in MRI. *IEEE Transactions on Computational Imaging*, 6:1139–1152, 2020.

Tim Bakker, Matthew Muckley, Adriana Romero-Soriano, Michal Drozdzal, and Luis Pineda. On learning adaptive acquisition policies for undersampled multi-coil MRI reconstruction. In *Medical Imaging with Deep Learning (MIDL)*, 2022.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 531–540, 2018.

Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Deep adaptive LiDAR: End-to-end optimization of sampling and depth completion at low sampling rates. In *IEEE International Conference on Computational Photography (ICCP)*, pp. 1–11, 2020. doi: 10.1109/ICCP48838.2020.9105252.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2 edition, 2006.

Zhe Dong, Andriy Mnih, and George Tucker. DisARM: An antithetic gradient estimator for binary latent variables. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *International Conference on Machine Learning (ICML)*, pp. 3384–3395, 2021.

Daniel Gehrig and Davide Scaramuzza. Low-latency automotive vision with event cameras. *Nature*, 629(8014):1034–1040, 2024. doi: 10.1038/s41586-024-07409-w.

Baran Gözcü, Rabeeh Karimi Mahabadi, Yen-Huan Li, Efe Ilıcak, Tolga Çukur, Jonathan Scarlett, and Volkan Cevher. Learning-based compressive MRI. *IEEE Transactions on Medical Imaging*, 37(6):1394–1406, 2018.

Yuhong Guo and Russell Greiner. Optimistic active learning using mutual information. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 823–829, 2007.

Sudan Han, Linjie Yan, Yuxuan Zhang, Pia Addabbo, Chengpeng Hao, and Danilo Orlando. Adaptive radar detection and classification algorithms for multiple coherent signals. *IEEE Transactions on Signal Processing*, 69:560–572, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Iris A. M. Huijben, Bastiaan S. Veeling, and Ruud J. G. van Sloun. Deep probabilistic subsampling for task-adaptive compressed sensing. In *International Conference on Learning Representations (ICLR)*, 2020.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, 2017.

Steven M. Kay. *Fundamentals of Statistical Signal Processing, Volume II: Detection Theory*. Prentice Hall, 1998.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Steven Kleinegesse and Michael U. Gutmann. Bayesian experimental design for implicit models by mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, pp. 5316–5326, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

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.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, pp. 8162–8171, 2021.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

Oisin Nolan, Tristan S. W. Stevens, Wessel L. van Nierop, and Ruud J. G. van Sloun. Active diffusion subsampling. *Transactions on Machine Learning Research*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4172–4182, 2023.

Luis Pineda, Sumana Basu, Adriana Romero, Roberto Calandra, and Michal Drozdzal. Active MR k-space sampling with reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 23–33, 2020.

Francesco Pittaluga, Zaid Tasneem, Justin Folden, Brevin Tilmon, Ayan Chakrabarti, and Sanjeev J. Koppal. Towards a MEMS-based adaptive LIDAR. In *International Conference on 3D Vision (3DV)*, pp. 1216–1226, 2020. doi: 10.1109/3DV50981.2020.00132.

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

Simo Ryu. Minimal implementation of D3PM in PyTorch. `https://github.com/cloneofsimo/d3pm`, 2024.

Shachar Shayovitz and Meir Feder. Universal active learning via conditional mutual information minimization. *IEEE Journal on Selected Areas in Information Theory*, 2(2):720–734, 2021.

An Vuong, Anthony Q. Nguyen, and Thinh Nguyen. Active data acquisition: An information theoretic approach. In *International Conference on Computing, Networking and Communications (ICNC)*, pp. 207–211, 2025. doi: 10.1109/ICNC64010.2025.10993621.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, et al. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018.

Zizhao Zhang, Adriana Romero, Matthew J. Muckley, Pascal Vincent, Lin Yang, and Michal Drozdzal. Reducing uncertainty in undersampled MRI reconstruction with active acquisition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

APPENDIX

A EXPERIMENTAL DETAILS

Priors. Each D3PM is initialized from a D3PM trained on the same data without the mask state and then trained under the encoding of Section 3 for 60 epochs on MNIST and 30 on CIFAR-10. Table 5 lists the mask-generator settings.

Table 5: Mask-generator training; DisARM uses one antithetic pair per image.

Setting	**MNIST**	**CIFAR-10**
Batch size	512	128
Learning rate	3×10^{-4}	2×10^{-5}
Sparsity weight λ	10	10
Budget range	$[0.05, 0.95]$	$[0.1, 0.9]$
Gradient estimator	DisARM	DisARM
Epochs	200	200

Gradient estimator. For independent per-pixel masks the exact gradient of the expected loss is $p_i(1-p_i)\,\mathbb{E}\big[\mathcal{L}(X_i{=}1) - \mathcal{L}(X_i{=}0)\big]$, which we estimate by flipping one pixel of a sampled mask at a time. On the MNIST generator, untrained and after training, the straight-through gradient has cosine -0.73 to -0.97 with it on the generator's weights. Trained with it, the generator's whole-image entropy rose at learning rate 3×10^{-4} and did not fall at 10^{-4} or 3×10^{-5}, and no checkpoint reached a lower objective than a random mask. DisARM needs two forward passes of the frozen prior per antithetic pair and no backward pass through it, and has cosine 0.94 to 0.99 with the exact gradient.

Evaluation. All image numbers are on the hold-out test split (20 batches of 128), at fixed budgets $s \in \{0.1, 0.3, 0.5, 0.7\}$, against budget-exact uniformly random masks. Besides the surrogate objective itself, we report several criteria the mask was not optimized for. We reconstruct each image with its observed pixels at their measured values and its hidden pixels at the D3PM's argmax, and score it by full-image accuracy and by foreground recovery, the fraction of nonzero (digit) pixels recovered correctly; since $\sim$90% of MNIST pixels are background, an all-black guess already

reaches $\sim$0.90 full-image accuracy, so foreground recovery is the informative one. On CIFAR-10 we add PSNR, with hidden pixels at the D3PM's posterior mean on a $[0, 1]$ scale; the reference is the 8-bin quantized target the D3PM models, not the original image, so quantization error is excluded from every PSNR we report. The informative fraction is the share of observed pixels that are not black (on MNIST, digit pixels). We also report accuracy on the hidden pixels, overall and on the hidden foreground; these are biased against a mask that observes more of the digit, so we read them alongside the whole-reconstruction scores. Learned masks are drawn on budget: the $\mathrm{round}(sHW)$ pixels with the largest keep log-odds plus logistic noise, which is a Bernoulli draw's own noise with its threshold moved to meet the budget exactly. A plain draw overshoots it, by up to 87% on CIFAR-10.

B Additional one-shot results

Tables 6 and 7 give every metric for the one-shot generators. On MNIST the objective and the criterion improve together over training (Figure 7): at $s = 0.7$, digit recovery stays above random from epoch 10 (0.992) to epoch 200 (1.000, against 0.968). The advantage narrows as the budget grows (Figure 8); only accuracy on the digit pixels the mask leaves hidden falls below random, from $s = 0.3$, a measure biased against a mask that observes more of the digit. On CIFAR-10 the generator lowers the surrogate below random only from $s = 0.3$; at $s = 0.1$ its objective ends slightly above random's while its PSNR keeps improving (Figure 7). Figure 9 shows example masks on CIFAR-10.

Table 6: MNIST mask generator trained with $\mathcal{H}_{\mathrm{all}}$, learned / random, hold-out test split, 5 seeds, budget-exact sampled masks (bold = better). Hidden-fg acc. scores only the digit pixels left hidden; it is biased against good masks, since a mask that observes most of the digit leaves only the pixels the prior finds hardest.

s	$\mathcal{H}_{\mathrm{all}}$	Full acc.	Fg recovery	Informative	Hidden-fg acc.	Errors/img
0.1	**0.0564** / 0.0685	**0.978** / 0.954	**0.903** / 0.777	**0.41** / 0.09	**0.829** / 0.753	**23.0** / 46.7
0.3	**0.0495** / 0.0527	**0.995** / 0.977	**0.978** / 0.894	**0.28** / 0.09	0.813 / **0.849**	**4.74** / 24.0
0.5	**0.0476** / 0.0488	**0.999** / 0.986	**0.996** / 0.939	**0.19** / 0.09	0.734 / **0.879**	**0.68** / 14.4
0.7	**0.0467** / 0.0471	**1.000** / 0.992	**1.000** / 0.968	**0.13** / 0.09	0.578 / **0.893**	**0.06** / 7.88

Table 7: CIFAR-10 mask generator trained with $\mathcal{H}_{\mathrm{all}}$, learned / random, hold-out test split, 5 seeds, budget-exact sampled masks (bold = better). Non-black rec. is the fraction of non-black pixels reconstructed correctly.

s	$\mathcal{H}_{\mathrm{all}}$	Full acc.	Non-black rec.	PSNR (dB)
0.1	0.790 / **0.786**	**0.675** / 0.670	**0.676** / 0.670	**20.45** / 20.14
0.3	**0.420** / 0.435	**0.824** / 0.816	**0.824** / 0.815	**24.80** / 24.33
0.5	**0.236** / 0.250	**0.901** / 0.894	**0.901** / 0.893	**28.01** / 27.47
0.7	**0.116** / 0.126	**0.951** / 0.946	**0.951** / 0.946	**31.54** / 30.87

C Additional sequential results

Figure 10 plots Table 2. Going from 1 to 16 steps reduces errors $3\times$ at $s = 0.1$ and $22\times$ at $s = 0.2$. On CIFAR-10, label-greedy with 16 steps reaches 20.0 and 24.4 dB at $s = 0.1$ and 0.3, against 20.2 and 24.3 for random and 21.1 and 26.4 for a random probe followed by variance-greedy steps; with 4 steps it falls below 15 dB. In Figure 3, LOUPE is shown at the better of two test protocols at each budget: the top-k of its learned map, as for MRI, or a budget-exact sample from it, the distribution it trained on (top-k degenerates to the image border on CIFAR-10 at $s = 0.1$, where the map never polarized). Figures 11 and 12 show example masks of every method.

D fastMRI details

Reconstructors. Both U-Nets are trained under six heuristic mask families, random, equispaced and variable density, each with and without the ACS block, with $s \sim \mathcal{U}(0.10, 0.75)$.

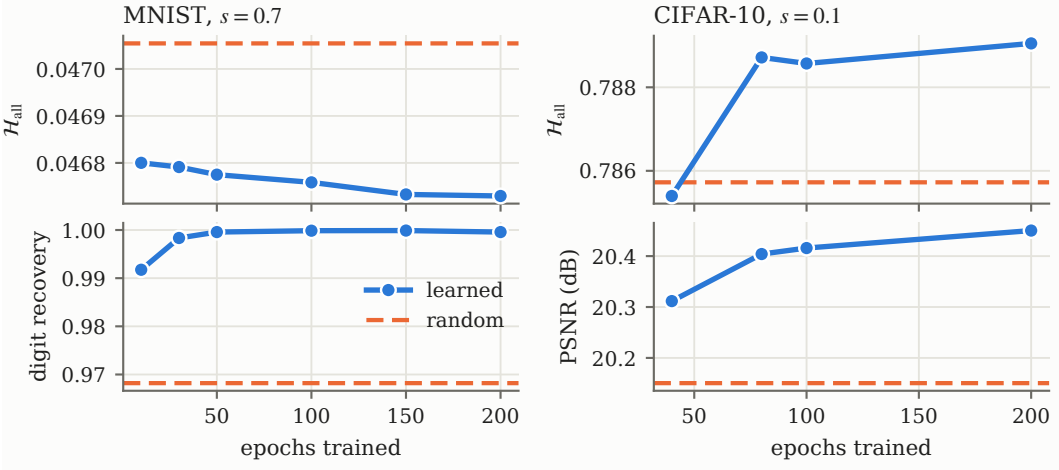

Figure 7: Top: the value of the whole-image surrogate objective the mask is trained on, with DisARM. Bottom: a criterion the mask is not optimized for, on the same checkpoints, with budget-exact sampled masks. On MNIST at $s = 0.7$ both improve together, and digit recovery stays above random throughout (0.992 at epoch 10 and 1.000 at 200, against 0.968). On CIFAR-10 at $s = 0.1$ they part ways: the objective rises above random's after epoch 40 while PSNR keeps improving and stays above random (20.31 dB at epoch 40, 20.45 at 200, against 20.15).

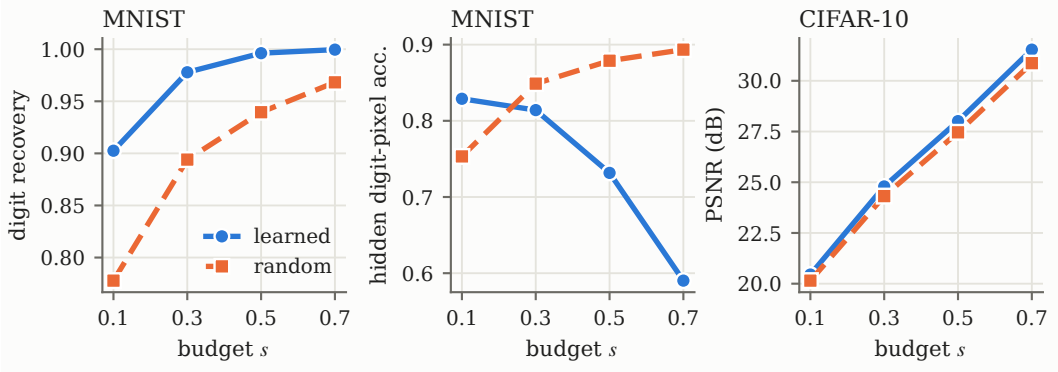

Figure 8: The MNIST advantage in digit recovery (left) holds at every budget and narrows as the budget grows (0.903 vs. 0.777 at $s = 0.1$, 1.000 vs. 0.968 at $s = 0.7$); accuracy on the digit pixels the mask leaves hidden (centre) is worse than random for every $s \geq 0.3$, a measure biased against a mask that observes more of the digit. Mask trained with DisARM, final checkpoint, budget-exact sampled masks. On CIFAR-10 (right) the mask is 0.3 to 0.7 dB above random at every budget.

Static search. Every run of the row-swap search stopped at a local optimum after 0–11 accepted swaps; a little over half the gain (57% over all budgets) comes from starting at the best fixed variable-density draw rather than redrawing per slice, and the rest from the swaps. The static mask is the best of those tested on NMSE and PSNR at every budget (Figures 13, 14 and 5), and it also has lower NMSE than variable density under plain zero-filled reconstruction at every budget. SSIM agrees with NMSE against variable density at every budget, but center-concentrated masks score up to 1.0 SSIM points ($\times 100$) higher at $s \geq 0.40$. Figure 15 adds the relaxed row profile and the energy oracle to Figure 4; both concentrate their rows at low frequencies. Figure 16 shows a harder validation slice, of 75th-percentile difficulty: the search mask again has the lowest NMSE (0.0339, against 0.0352 for variable density).

Relaxed row profile. A relaxed row profile, 300 logits trained against U-Net A through a straight-through top-k mask, leaves its energy initialization only with a small step (2×10^{-4}; at 5×10^{-3} no epoch improves on it at $s \geq 0.40$). Retuned, it matches variable density at $s = 0.40$ and 0.75,

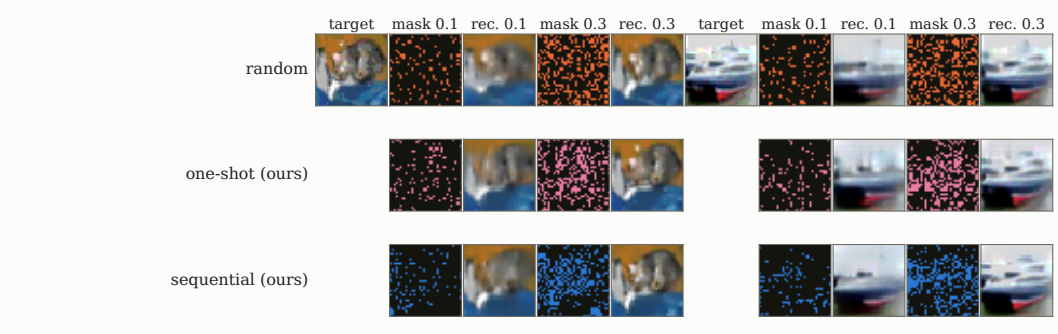

Figure 9: CIFAR-10 test images at $s = 0.1$ and 0.3, one row per method; masks are tinted by method, and each reconstruction keeps the observed pixels and fills the rest with the prior's posterior mean. The one-shot generator, trained with DisARM and sampled on budget, scatters its pixels over the whole image much as a random mask does; sequential acquisition (variance-greedy, 16 steps; Section 6) spreads them more evenly and more densely where the image has structure.

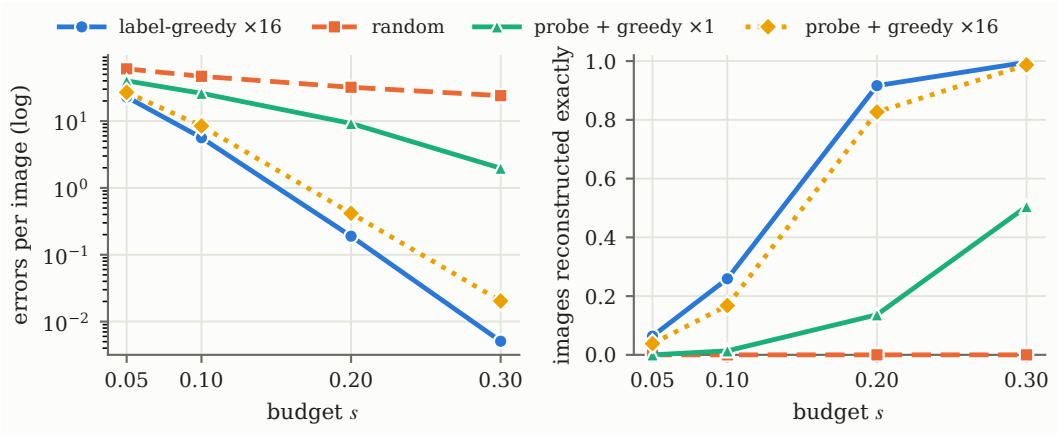

Figure 10: Sequential acquisition on MNIST (exact values in Table 2). Left: errors per image, log scale. Right: fraction of test images reconstructed exactly.

improves on it only at $s = 0.50$ (0.01527 against 0.01546 ± 0.00007), loses at $s = 0.25$, and trails the search mask at every budget.

Sequential acquisition. Starting from the ACS block, each of R steps adds the unobserved rows where the coarse row-profile prior's predictive entropy is highest. Table 8 gives every variant. Under the judge it loses to variable density at $s = 0.25$ and 0.40 (4.8% and 5.5% higher NMSE with $R = 16$), is level at $s = 0.50$, and never reaches the static search mask's NMSE. At $s = 0.75$ a single step is best (1.8% lower NMSE than variable density, level with the search mask), but that mask is the same for every slice: after the ACS block the coarse profile carries no slice-specific information, so the first step reads a static mask off the prior. Under zero-filled reconstruction it has lower NMSE than variable density at every budget, the pattern of Table 3. Ranking rows by the prior's expected energy instead of its entropy is worse at every budget and number of steps, so the entropy carries more than energy, but a prior over row energy does not tell the reconstructor what it is missing.

LOUPE. We ported LOUPE (Bahadir et al., 2020) to PyTorch (mask layers matching the original to 3×10^{-7}), restricted it to whole PE rows, and trained it until its row set stopped changing. Its mask is 15.5–31.0% worse than ours under the judge, and worse than every heuristic at every budget, although it trains its reconstructor jointly. Its forward model assumes Hermitian k-space: at $s \leq 0.50$ its masks hold only 7–8 mirrored $\pm f$ row pairs (ours: 17–57) and at $s = 0.75$ sample one half of k-space.

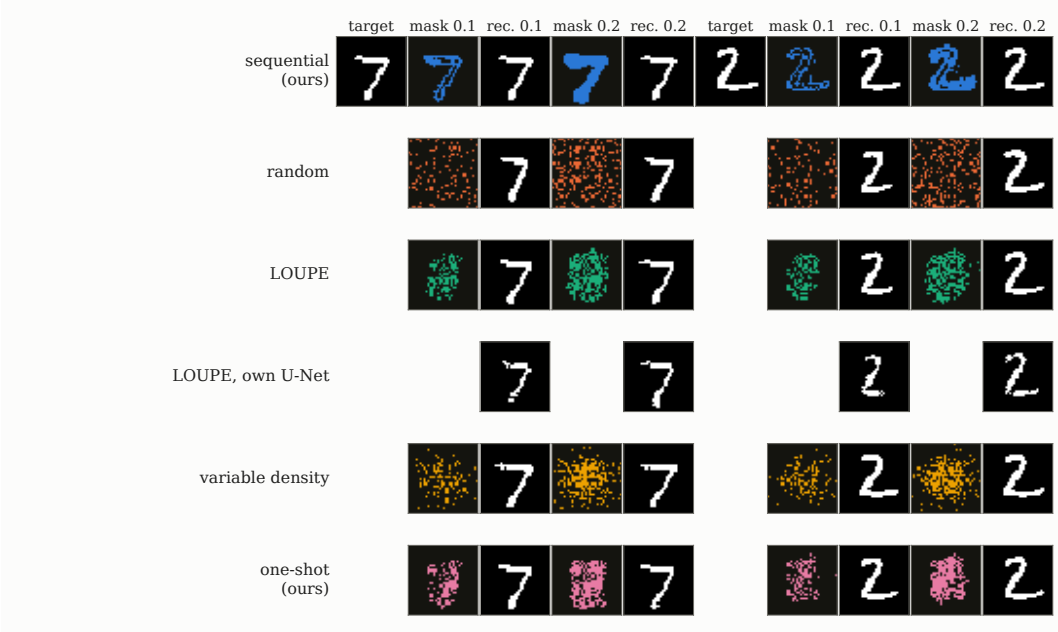

Figure 11: MNIST test images at $s = 0.1$ and 0.2: each method's mask, tinted in its colour, and the reconstruction under the prior (LOUPE's own U-Net in its own row). LOUPE's masks are sampled from its learned map.

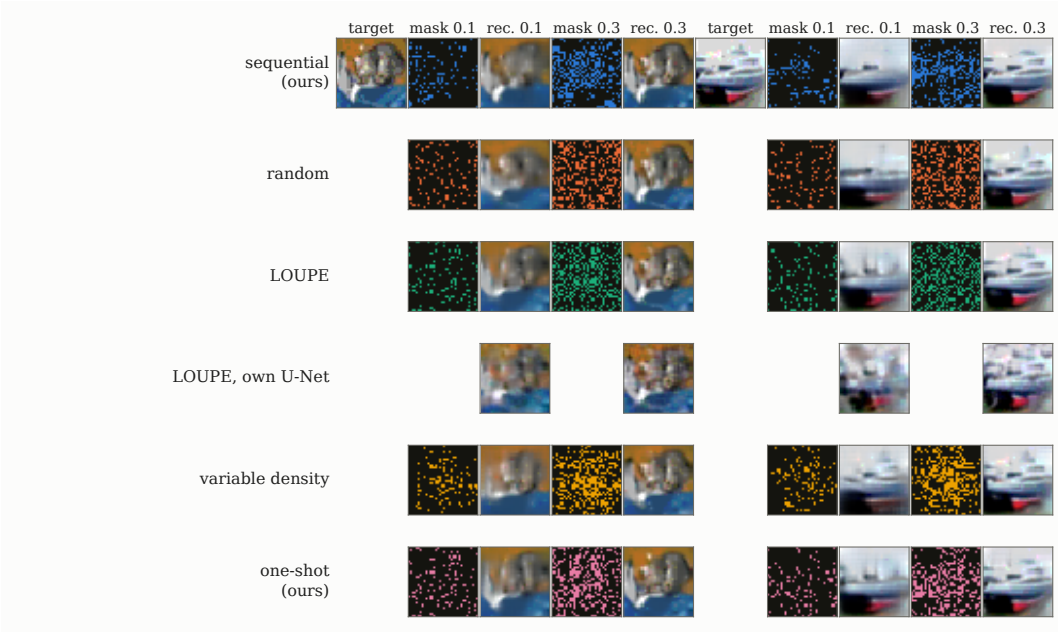

Figure 12: CIFAR-10 test images at $s = 0.1$ and 0.3, as in Figure 11; the sequential-acquisition row is variance-greedy with 16 steps. The one-shot mask, trained with DisARM and sampled on budget, scatters its pixels over the image much as a random mask does.

Table 8: Hold-out val NMSE under the judge U-Net B for sequential row acquisition with the coarse row-profile prior: starting from the ACS block, rows are added in 1, 4 or 16 steps, ranked by the prior's predictive entropy or, as a control, by its expected energy. ACS + variable density is the mean over 5 seeds. Bold: the lowest NMSE per budget, and any entry within two standard deviations of it, measured by the seed spread of ACS + variable density.

Mask	$s = 0.25$	$s = 0.40$	$s = 0.50$	$s = 0.75$
ACS + variable density	0.02471	0.01889	0.01546	0.00764
static search (ours)	**0.02389**	**0.01860**	**0.01523**	**0.00751**
sequential acquisition, entropy, 1 step	0.03031	0.02571	0.02061	**0.00751**
sequential acquisition, entropy, 4 steps	0.02670	0.02254	0.01916	0.00941
sequential acquisition, entropy, 16 steps	0.02589	0.01992	0.01559	0.00807
control, expected energy, 1 step	0.03359	0.02928	0.02381	0.01145
control, expected energy, 4 steps	0.02982	0.02438	0.01940	0.01209
control, expected energy, 16 steps	0.02699	0.02122	0.01790	0.00966

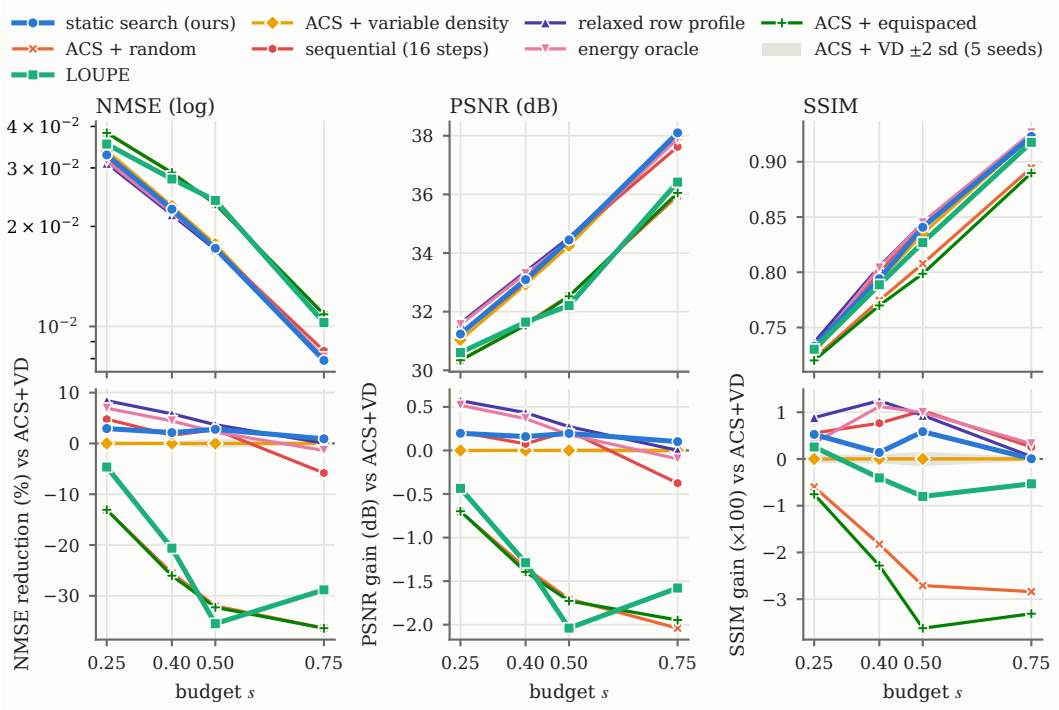

Figure 13: The same masks under zero-filled reconstruction, with no network. The sequential-acquisition line has lower NMSE than variable density here at $s \leq 0.50$.

E What the mask state adds

To measure what the separate mask state of Section 3 adds, we fine-tune two MNIST priors from the same D3PM with the same recipe (60 epochs, learning rate 2×10^{-4}, batch 512): one with the mask state, and a control that keeps 0 as both black and unobserved. The pair is repeated for 5 fine-tuning seeds. Both priors drive the same acquisition, a random probe of 20% of the budget followed by one greedy step on the prior's entropy, on the same 2560 test images with the same probes (5 evaluation seeds). With the mask state the prior makes 10–15% fewer errors on average (Table 9), in every fine-tuning seed at $s \leq 0.3$ and in 4 of 5 at $s = 0.4$. The spread comes from the fine-tunes, not the evaluation, whose standard deviation over evaluation seeds is below 0.15 errors; it grows with the budget as the error counts shrink.

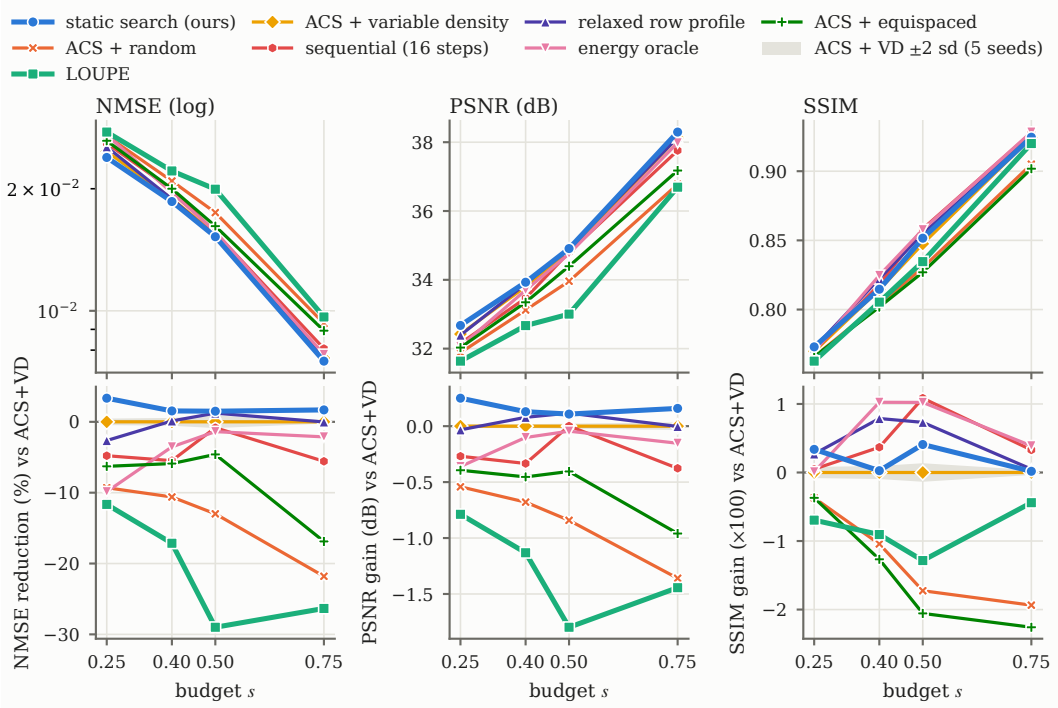

Figure 14: Row masks on the 199 hold-out validation slices, reconstructed by the hold-out judge U-Net B. Top: NMSE (log scale), PSNR and SSIM against the budget. Bottom: gain over ACS + variable density. The static search mask is best on NMSE and PSNR at every budget and LOUPE is last; the center-concentrated relaxed row profile and energy oracle, with sequential acquisition, lead on SSIM at $s = 0.40$–0.75. The sequential-acquisition line (16 steps; Table 8) trails variable density on NMSE at every budget except $s = 0.50$, where it is level.

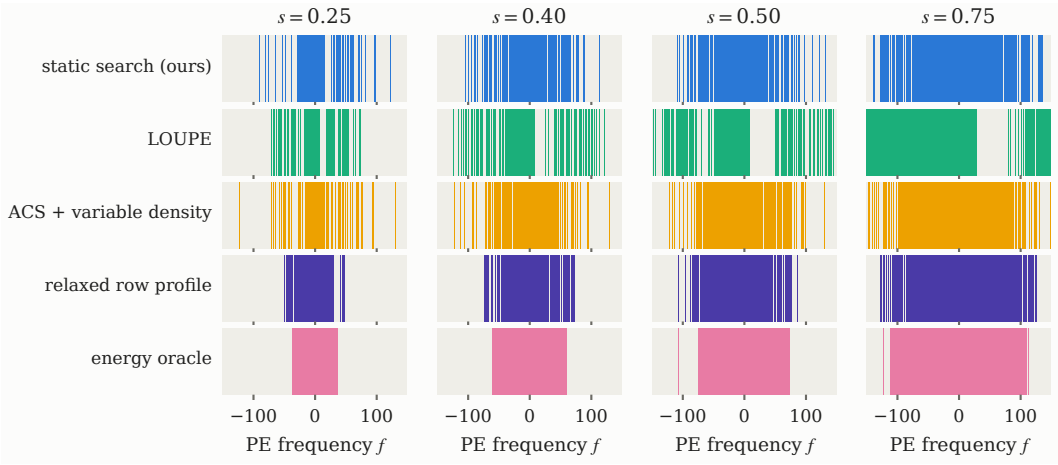

Figure 15: Phase-encode rows sampled by every static mask, including the relaxed row profile and the energy oracle ($f = 0$ is DC).

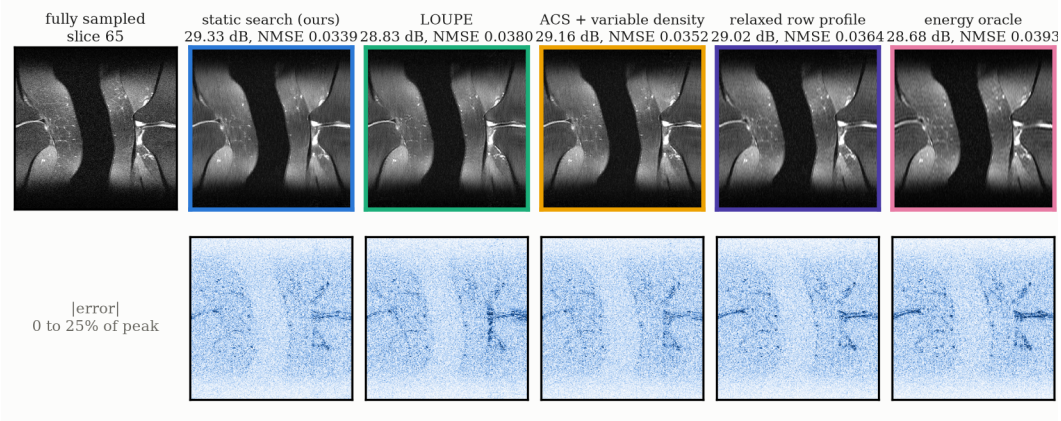

Figure 16: A validation slice of 75th-percentile difficulty for ACS + variable density, at $s = 0.25$, reconstructed by the hold-out judge. Top: reconstructions. Bottom: absolute error $|\text{reconstruction} - \text{fully sampled}|$, from 0 (white) to a quarter of the slice's 99.5th-percentile intensity (dark blue). PSNR and NMSE are per slice.

Table 9: MNIST errors per test image with and without the separate mask state, probe + one greedy step, mean $\pm$ standard deviation over 5 fine-tuning seeds. The last column counts the seeds in which the mask state has fewer errors.

s	**Without**	**With mask state**	**Fewer errors**	**Seeds**
0.10	28.72 ± 0.36	$\mathbf{25.73} \pm 0.44$	$10.4 \pm 0.9\%$	5/5
0.20	10.08 ± 0.22	$\mathbf{8.83} \pm 0.33$	$12.4 \pm 2.8\%$	5/5
0.25	4.98 ± 0.14	$\mathbf{4.28} \pm 0.25$	$14.2 \pm 4.5\%$	5/5
0.30	2.07 ± 0.08	$\mathbf{1.76} \pm 0.15$	$15.0 \pm 7.4\%$	5/5
0.40	0.22 ± 0.01	$\mathbf{0.19} \pm 0.02$	$14.4 \pm 10.8\%$	4/5