SPIRA: A SPATIALLY-VARIANT REINFORCEMENT-LEARNING AGENT FOR UNSUPERVISED IMAGING INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Imaging inverse problems arise ubiquitously in many applications, from medical imaging to consumer photography. In many scenarios, the lack of ground truth images prevents the development of data-driven supervised learning reconstruction approaches. Thus, unsupervised learning methods have gathered significant attention in recent years by developing different reconstruction loss functions, image priors, or neural architectural designs aimed at approximating a supervised loss function. However, the best unsupervised scheme varies not only across inverse problems (changes in the sensing matrix, imaging modality, noise distribution, etc) but across spatial regions. Thus, we propose a SPatIally-variant Reinforcement-learning Agent (SPIRA). We build a library of unsupervised reconstruction experts, each fitted only to the measurements, and train a policy that assigns one expert to each spatial region at each step of the reconstruction. The policy is trained entirely on computer-generated data, where the reference image is available at no cost, from a per-region reward that combines the improvement of the composite with the regret against the best available expert for that region; an auxiliary head predicts each expert's regional error and shapes the action logits. At inference time, the policy provides an instance-specific reconstruction pipeline that adapts to different image content and inverse problem physics. The proposed approach was validated in different imaging modalities, including image restoration (inpainting, demosaicing and deblurring), compressive sensing, sparse-view CT, and single-coil MRI, improving over any single unsupervised expert applied throughout while requiring one-third to one-half of the expert iterations of greedy expert selection.

1 INTRODUCTION

Imaging inverse problems aim to restore an unknown image from corrupted, undersampled and/or projected observations (Bertero et al., 2021; Ongie et al., 2020). These problems appear naturally in a wide range of areas, including medical imaging, which is fundamental for correct medical diagnostics (Lustig et al., 2008; Willemink et al., 2013), computational photography to enhance the quality of the images delivered by commercial cameras (Heide et al., 2014; Delbracio et al., 2021), and compressed sensing in which the scene is projected into low-dimensional observations (Candes & Wakin, 2008; Duarte et al., 2008), among others.

Since imaging inverse problems are ill-posed, traditional works have proposed using hand-crafted regularization such as sparsity through ℓ_1 or ℓ_0 norms (Tibshirani, 1996; Foucart & Rauhut, 2013), total variation (TV) (Rudin et al., 1992), or low-rank priors (Fazel et al., 2008; Peng et al., 2014). More recently, learning-based methods have employed deep learning models to map end-to-end from measurements to the reconstructed image. From this idea, many approaches have been developed, from black-box models (Zhang et al., 2017; McCann et al., 2017) to optimization-inspired unrolled networks (Gregor & LeCun, 2010; Monga et al., 2021; Xiang et al., 2021). Lately, plug-and-play approaches combine end-to-end modules such as denoiser networks (Venkatakrishnan et al., 2013; Zhang et al., 2021) or restoration networks that have been employed as implicit priors (Romano et al., 2017; Terris et al., 2024a). Under the same idea of plugging in learned models, generative priors such as generative adversarial networks (Bora et al., 2017), diffusion models (Chung

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.

et al., 2023; Zhu et al., 2023) and flow-matching models (Lipman et al., 2023; Martin et al., 2025) harness the learned data distribution to reconstruct realistic images. Nevertheless, all of these models require training on a huge corpus of clean images that, in several scenarios such as emerging medical modalities or scientific imaging, is highly limited or simply unavailable.

To address this significant problem, *unsupervised* reconstruction methods learn directly from corrupted measurements, replacing the supervised loss by a surrogate that involves only the observations and the forward operator. Three strategies dominate this literature. *Measurement splitting* holds out part of the observations and uses it as a proxy target, as in Noise2Noise (Lehtinen et al., 2018), its blind-spot variants (Batson & Royer, 2019; Krull et al., 2019), and SSDU for physics-guided reconstruction (Yaman et al., 2020). *Risk estimation* builds an unbiased estimate of the supervised error from the noise statistics through Stein's unbiased risk estimator and its generalized form (Eldar, 2009; Metzler et al., 2018; Monroy et al., 2025). *Invariance* exploits group symmetries of the signal set to recover the components that lie in the null space of the sensing operator (Chen et al., 2021; 2022; Terris et al., 2024b; Diaz-Delgado et al., 2026), with identifiability conditions given by sensing theorems for unsupervised learning (Tachella et al., 2023a). Orthogonal to all of them, instance-specific priors dispense with a training set altogether by re-parameterizing a single image with an untrained network, as in the deep image prior (DIP) (Ulyanov et al., 2018; Baguer et al., 2020), the deep decoder (Heckel & Hand, 2019), or sinusoidal implicit neural representations (SIRENs) (Sitzmann et al., 2020).

This abundance indeed turns reconstruction into a design problem: *which* unsupervised scheme should be used, and *where*? The answer changes across inverse problems, as the sensing matrix, the imaging modality, the measurement rate and the noise distribution change, but it also changes *within* a single image. Every prior encodes a different inductive bias: TV favors piecewise-constant regions, wavelet sparsity isolates edges and impulses, DIP reproduces natural textures, and SIRENs capture smooth gradients and periodic content. Natural scenes mix all of these, so a single global choice is necessarily a compromise, and the prior that recovers a flat background is rarely the one that recovers a textured foreground. Classical parameter-selection rules, such as the discrepancy principle (Morozov, 1966) or Tikhonov-based criteria (Golub et al., 1999), address a related question but only tune a scalar regularization weight, globally, once the prior has already been fixed. In the supervised setting, all-in-one restoration networks (Cui et al., 2025; Tian et al., 2025) adapt a single model to several degradations, and reinforcement learning (RL) has been used to adapt restoration pipelines by selecting a toolchain of restoration networks (Yu et al., 2018), by acting per pixel (Furuta et al., 2019), or by tuning plug-and-play hyperparameters (Wei et al., 2020). All of these agents nonetheless act on supervised restoration networks and are trained against clean references from the target domain; SPIRA instead composes unsupervised optimization procedures and learns its allocation from procedurally generated images alone.

Thus, we propose a **SP**at**I**ally-variant **R**einforcement-learning **A**gent (SPIRA) for unsupervised imaging inverse problems. We construct an expert library with different unsupervised reconstruction methods and partition the reconstruction domain into spatial regions. At every decision step, a policy observes the current reconstruction together with measurement-domain evidence of each expert, and assigns one expert to each region; a fixed budget of optimization iterations then advances the selected experts, and their outputs are recomposed through smooth partition-of-unity windows. Experts are persistent within an episode, so re-selecting one continues its optimization instead of restarting it, and only the experts actually selected are advanced, so the cost of a rollout is set by the number of distinct choices the policy makes. We first train the policy on a computer-generated synthetic dataset to maximize an expected return for improving reconstruction fidelity without using ground-truth images: the reward is built from held-out measurement residuals back-projected through the adjoint operator, so the ground truth never enters the reward, the expert losses, or the policy observations. Training uses only procedurally generated collages (dead leaves (Lee et al., 2001), octave noise, gratings, Voronoi mosaics and sparse impulses), in the spirit of pre-training with synthesized images (Baradad et al., 2021; Nakamura et al., 2023), and the policy is then deployed without adaptation. At inference time, the policy therefore provides an instance-specific reconstruction pipeline that adapts to different image content and inverse-problem physics. Seen from a different angle, SPIRA is a mixture of experts (Jacobs et al., 1991; Shazeer et al., 2017) in which the gating function is a spatial RL policy, and the experts are optimization procedures rather than learned layers.

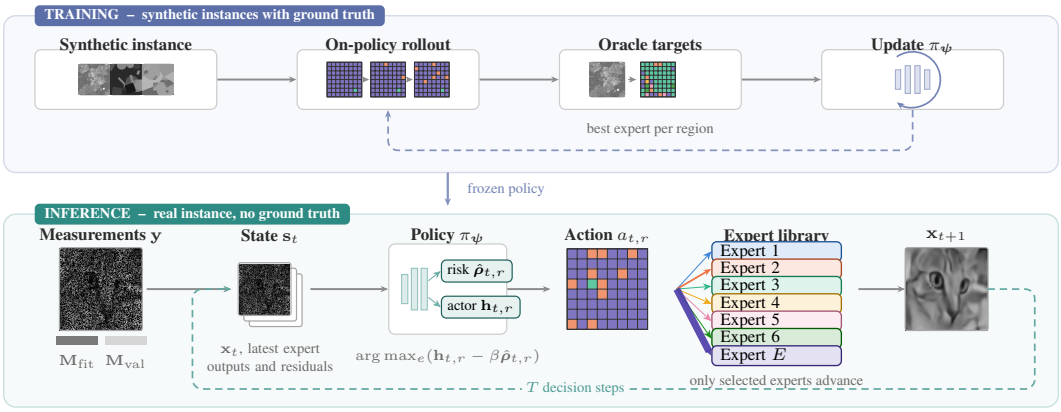

Figure 1: **SPIRA learns *which* unsupervised expert to use and *where*. Top:** the policy is trained on procedurally generated instances, each drawing its own signal, forward operator, noise level and measurement rate. Because the signal is synthetic, its ground truth is available during training and is used in two places only: to label the best expert of every region, which a *risk head* learns to predict, and to score each decision by its regret against that winner. Experts are advanced on-policy, exactly as at inference. **Bottom:** the frozen policy is deployed on a real instance whose ground truth is never observed. At every decision step it reads the current reconstruction together with the latest output and back-projected held-out residual of each expert, predicts the regional risk of every expert and selects one expert per region; only the selected experts are optimized by K gradient descent iterations; the others keep their last state, and the selected outputs are recomposed through smooth partition-of-unity windows. The grids and the arrow widths are the allocation the policy actually executes on a natural inpainting instance: the last decision mixes several experts, one of which covers most of the regions, and only the selected experts are advanced.

2 Method

2.1 Problem Formulation

We recover an image $\mathbf{x}^\star \in \mathbb{R}^N$ from measurements $\mathbf{y} = \mathbf{H}\mathbf{x}^\star + \boldsymbol{\omega} \in \mathbb{R}^M$, where $\mathbf{H} \in \mathbb{R}^{M\times N}$ is a known linear operator and $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_M)$. We are given E unsupervised reconstruction methods, the *experts* $e \in \{1,\dots,E\}$, and a partition of the image into R regions $r \in \{1,\dots,R\}$ with pixel sets B_r. Let $\varepsilon_r(\mathbf{x})$ denote the mean squared error of an estimate $\mathbf{x}$ over region r. Because every prior fits some image structures better than others, the expert with the lowest ε_r — the *regional oracle* — changes with the operator, the noise level and, within one image, from region to region. It cannot be identified at test time, since ε_r depends on the unknown $\mathbf{x}^\star$. SPIRA learns, from synthetic images where $\mathbf{x}^\star$ is available, a policy that predicts how the regional error of every expert will evolve and selects one expert per region accordingly; at test time the policy observes only $(\mathbf{y}, \mathbf{H}, \sigma)$, and the reconstruction is produced entirely by the unsupervised experts. The complete formulation, with all dimensions and definitions, is given in Appendix A.

2.2 Unsupervised Experts and Measurement Evidence

Expert e represents its estimate as $\mathbf{x}^{(e)} = f_e(\boldsymbol{\theta}^{(e)})$ with parameters $\boldsymbol{\theta}^{(e)} \in \mathbb{R}^{D_e}$. The measurements are split once into a fitting and a held-out part, selected by diagonal binary masks $\mathbf{M}_{\text{fit}}, \mathbf{M}_{\text{val}} \in \{0,1\}^{M\times M}$, and every expert minimizes

$$\mathcal{J}_e\big(\boldsymbol{\theta}^{(e)}\big) = \frac{\big\|\mathbf{M}_{\text{fit}}\big(\mathbf{H}f_e(\boldsymbol{\theta}^{(e)}) - \mathbf{y}\big)\big\|_2^2}{\operatorname{tr}(\mathbf{M}_{\text{fit}})} + \lambda_e\, h_e\big(f_e(\boldsymbol{\theta}^{(e)})\big), \tag{1}$$

where the prior is explicit through a regularizer h_e weighted by $\lambda_e \geq 0$ (e.g. total variation), implicit through the architecture of f_e with $h_e \equiv 0$ (deep image priors, the deep decoder, a sinusoidal network), or both. The held-out measurements provide the only test-time evidence of quality: back-projecting the held-out residual of an estimate gives a residual-energy map $\widetilde{\boldsymbol{\rho}}(\mathbf{x}) \in \mathbb{R}^{N_p}_{\geq 0}$ over the N_p

pixels, and the regional score of expert e before decision t is

$$s_{t,e,r} = \frac{1}{|B_r|} \sum_{i \in B_r} \widetilde{\rho}_i\big(\mathbf{x}_t^{(e)}\big). \tag{2}$$

This score is informative but biased: it is blind to errors in the null space of $\mathbf{H}$ and it compares experts after unequal amounts of optimization.

2.3 Sequential Spatial Selection

Reconstruction proceeds over T decisions $t \in \{0, \dots, T-1\}$. At decision t the policy selects one expert per region, $\mathbf{a}_t = (a_{t,1}, \dots, a_{t,R}) \in \{1, \dots, E\}^R$. Only the selected experts, $\mathcal{E}_t = \{a_{t,1}, \dots, a_{t,R}\}$, are advanced by K further Adam iterations on Eq. 1, resuming from their current state, so each decision also allocates the optimization budget. The reconstruction takes, in every region, the output of the expert selected there,

$$\mathbf{x}_{t+1} = \sum_{r=1}^{R} \mathbf{W}_r \, \mathbf{x}_{t+1}^{(a_{t,r})}, \tag{3}$$

where the fixed diagonal matrices $\mathbf{W}_r \in \mathbb{R}^{N \times N}$ restrict the image to region r with a smooth border; they sum to the identity and only avoid seams, so the policy makes a single discrete choice per region and never predicts mixing weights. The policy observes a state $\mathbf{s}_t$ that stacks the current reconstruction, the held-out residual map and the current output of every expert, how long ago each expert was last advanced, the previous selection, and an embedding of the operator family and noise level (Appendix A.3).

2.4 Policy with a Learned Risk Estimator

A region should be assigned the expert that will be best at the *end* of the reconstruction, not the one that is best now. Experts converge at different rates (pixel-domain TV within a few iterations, convolutional decoders after hundreds), and implicit priors begin to fit noise late in optimization, so the ranking of the experts in a region changes during the episode, while the held-out score of Eq. 2 only reflects the current, partially optimized state. We therefore define the *terminal risk* $\varepsilon_{e,r}^{\mathrm{T}}$ as the error of expert e in region r after it has received the full budget of all T decisions, and summarize the ranking of the experts in each region by the standardized log-risk

$$z_{e,r} = \frac{\log \varepsilon_{e,r}^{\mathrm{T}} - \mu_r}{\varsigma_r + \epsilon_z}, \tag{4}$$

with μ_r and ς_r the mean and standard deviation of $\{\log \varepsilon_{j,r}^{\mathrm{T}}\}_{j=1}^{E}$ and $\epsilon_z > 0$. The standardization removes the difficulty of the region and the scale of the problem, so $z_{e,r}$ retains only which prior is better where, which is what a selection needs and what transfers across operators, noise levels and images.

The policy network has three heads per region: a risk head $\hat{z}_{\boldsymbol{\psi}}(\mathbf{s}_t)_{r,e}$ that forecasts $z_{e,r}$ for all E experts from the current state, a correction head $g_{\boldsymbol{\psi}}(\mathbf{s}_t)_{r,e}$, and a critic $V_{\boldsymbol{\psi}}(\mathbf{s}_t)_r$. Experts are selected with

$$\pi_{\boldsymbol{\psi}}(a_{t,r} = e \mid \mathbf{s}_t) \propto \exp\big(g_{\boldsymbol{\psi}}(\mathbf{s}_t)_{r,e} - \beta \, \mathrm{sg}[\hat{z}_{\boldsymbol{\psi}}(\mathbf{s}_t)_{r,e}]\big), \tag{5}$$

where $\beta > 0$ is a learned inverse temperature and $\mathrm{sg}[\cdot]$ stops gradients. The policy is thus a Boltzmann distribution over the negated forecast risks: with $g = 0$ it assigns each region the expert expected to end with the lowest error, i.e. a greedy rule on a forecast of the *outcome* rather than on the current evidence. The correction g accounts for what the forecast ignores, namely that the current choice determines which experts are optimized and therefore the risk that is actually realized; it is trained by the policy gradient, while the stop-gradient keeps the risk head a pure predictor trained by regression only (Sec. 2.5).

Eq. 5 combines four established ideas. (i) Predicting, for every action, a measurable future outcome by supervised regression and acting on the prediction is the principle of Direct Future Prediction (Dosovitskiy & Koltun, 2017), which outperforms reward-only policy gradients because the

regression targets carry far denser information than a scalar return; here the outcome is the terminal regional error of each expert. (ii) Selecting actions by a softmax over action values scaled by a temperature is the policy form of maximum-entropy reinforcement learning and soft Q-learning (Haarnoja et al., 2017), with $-\hat{z}$ in the role of the action value; unlike Q-learning, $\hat{z}$ is regressed on measured outcomes rather than bootstrapped from successor states, because the terminal risk of every expert is observed directly. (iii) The additive correction g on top of a fixed decision rule is a residual policy (Silver et al., 2018; Johannink et al., 2019); the penalty on g in Eq. 7 keeps the policy close to the risk-driven base and admits deviations only where they increase the return. (iv) The ground truth that defines the targets is available only in simulation and never enters the state, as in asymmetric actor–critic and privileged-information training (Pinto et al., 2018; Chen et al., 2019). The essential difference from standard policy optimization is the feedback. A policy gradient observes only the return of the sampled expert in each region (bandit feedback), which has high variance and reinforces early preferences; the catch-up of Sec. 2.5 instead reveals the terminal risk of *every* expert in every region (full-information feedback), so the risk head is supervised on all E counterfactual actions in every episode.

2.5 Policy Learning from Synthetic Data

Each training episode draws a new procedural image (collages of textured dead leaves, $1/f^{\alpha}$ fields, gratings and mosaics with irregular zone boundaries) and a random operator and noise level; natural images are never used. Since $\mathbf{x}^{\star}$ is known, all regional errors are exact; it is used only for the reward and the risk targets, never by the experts, the state or the policy at test time. After the last decision every expert is advanced to its full budget, which makes the terminal risks $\varepsilon^{\mathrm{T}}_{e,r}$ exact for all experts whatever the policy selected.

With $e^{\star}_r = \arg\min_e \varepsilon^{\mathrm{T}}_{e,r}$ the regional oracle and $\bar{\varepsilon}_0$ the error of the initialization, region r receives at decision t

$$r_{t,r} = \frac{\varepsilon_r(\mathbf{x}_t) - \varepsilon_r(\mathbf{x}_{t+1})}{\bar{\varepsilon}_0} + \lambda_{\mathrm{reg}} \frac{\varepsilon^{\mathrm{T}}_{e^{\star}_r,r} - \varepsilon^{\mathrm{T}}_{a_{t,r},r}}{\varepsilon^{\mathrm{T}}_{e^{\star}_r,r}}, \tag{6}$$

the improvement of the reconstruction in that region plus a relative regret, which is zero when the region selects its oracle and negative otherwise.

Diversity follows from the optimum of the reward rather than from a regularizer. (i) The reward contains no term shared by all regions: a global term gives every region the same credit, favors the expert that is best on average, and empirically makes the policy collapse onto one expert per operator. (ii) The regret is tight: it is maximized exactly by the oracle map $(e^{\star}_r)_{r=1}^{R}$, so the optimal policy varies across regions wherever the best prior does and stays concentrated where one prior dominates. A comparison with the average alternative, by contrast, is satisfied by any above-average expert and cannot separate the best from the second best. (iii) The regret is relative, so every region counts equally regardless of its error level; otherwise high-error textured regions would dominate and impose their best expert on smooth regions.

The parameters $\boldsymbol{\psi}$ minimize

$$\mathcal{L} = \mathcal{L}_{\mathrm{AC}} + \frac{c_z}{TRE} \sum_{t,r,e} \left(\hat{z}_{\boldsymbol{\psi}}(\mathbf{s}_t)_{r,e} - z_{e,r} \right)^2 + \frac{c_g}{TRE} \sum_{t,r,e} g_{\boldsymbol{\psi}}(\mathbf{s}_t)^2_{r,e}, \tag{7}$$

where $\mathcal{L}_{\mathrm{AC}}$ is an advantage actor–critic loss on the returns of Eq. 6 with per-region advantages (Appendix A.5). The risk regression uses the full-information targets of Sec. 2.4, and the penalty on g keeps the risk-driven choice as the default. During training, we occasionally force one region onto the least recently advanced expert so no expert is starved. At test time the policy acts greedily with respect to Eq. 5 and advances only the selected experts, so the risk estimator replaces the evaluation of all experts that greedy selection by held-out scores requires. We show the inference algorithm at 1.

Table 1: All five benchmarks, 372 cells. "vs. fixed" is the mean paired margin over the best fixed expert, with the cells in which the method beats it. "Iter." is the mean number of expert Adam iterations per reconstruction and "Time" the mean wall-clock time on one RTX 3090. Both references use the ground truth and are not available at test time.

Method	PSNR	SSIM	vs. fixed	Cells won	Iter.	Time
Best fixed expert (SIREN)	24.28	0.661	—	—	300	0.9 s
SPIRA, $G = 8$	24.36	0.670	$+0.08$	179/372	859	4.0 s
SPIRA+P, $G = 8$	25.02	0.687	$+0.74$	207/372	859	4.0 s
SPIRA, $G = 16$	24.71	0.679	$+0.43$	205/372	1,125	5.5 s
SPIRA+P, $G = 16$	**25.15**	**0.691**	$\mathbf{+0.88}$	**213/372**	1,125	5.5 s
Greedy	25.49	0.710	$+1.22$	254/372	2,100	10.5 s
Per-cell oracle	25.64	0.719	$+1.36$	—	2,100	—

3 Results

3.1 Computational Imaging

We evaluate on five benchmarks and 372 cells, a cell being one image under one setting: Set14 at 128×128 RGB with inpainting, Gaussian blur and compressed sensing (126 cells); a harder version of the same three operators (84); Bayer demosaicing (42); sparse-view CT on LoDoPaB at 10, 20 and 30 views (60); and single-coil MRI on fastMRI knee slices at $4\times$, $6\times$ and $8\times$ Cartesian acceleration (60). Demosaicing and MRI are strict zero-shot tests: neither operator appears in training, and the policy is conditioned on the inpainting and compressed-sensing tokens respectively. Every method sees identical measurements, masks, noise realizations and fitting splits, with per-cell seeded expert initializations, so all comparisons are paired. The library has seven experts (DIP, DIP-TV, TV, SIREN, Deep Decoder, DIP-tiny, DIP-EMA). We report the policy at two decision resolutions, $G = 8$ and $G = 16$, with and without the unsupervised polish (+P), against greedy selection and two references drawn from the same library: the *best fixed expert*, the single expert with the highest mean PSNR over all cells, and the *per-cell oracle*, the best expert in each cell. Both references use the ground truth and are therefore unavailable at test time; the fixed one is what a practitioner who commits to a single prior could hope to achieve, and the per-cell oracle upper-bounds any selector over this library.

Adapting the prior beats a fixed choice. Table 1 gives the aggregate. No single prior serves all five benchmarks: the best fixed expert is SIREN at 24.28 dB, 1.36 dB below the per-cell oracle, and the identity of the best fixed expert changes with the benchmark (SIREN on Set14 and demosaicing, DIP-TV on Set14 hard, TV on CT and MRI, Table 2). The policy at $G = 16$ with the polish reaches 25.15 dB, 0.88 dB above that fixed choice, and wins 213 of the 372 cells, while using $1{,}125$ expert iterations against the $2{,}100$ that greedy needs to advance all seven experts at every step. Greedy remains the stronger selector, 1.22 dB above the fixed expert and within 0.14 dB of the oracle, but it verifies after optimizing and pays twice the iterations to do so; the policy commits before optimizing and recovers two thirds of the same gap at roughly half the cost. Resolution and polish both matter, and in the same direction: the finer grid adds 0.35 dB over $G = 8$ without the polish, and the polish adds 0.44 dB at $G = 16$ and 0.66 dB at $G = 8$. This margin comes from adapting across problems. Against a reference that already knows which prior suits each modality, the best fixed expert chosen separately for each benchmark, the policy leads on Set14 hard by 0.08 dB and trails by 0.11 to 0.63 dB elsewhere (Table 2), so what a single learned policy buys is the choice itself rather than a better prior in any one modality.

Where the policy helps. Per setting, the margins split cleanly, and Appendix C.1 lists all of them. The policy beats the per-setting oracle, not just the fixed expert, where the measurements are scarce and no expert dominates: inpainting with 20% of the pixels observed ($+0.44$ dB at $G = 16$, 14 of 14 images), inpainting at 30% ($+0.44$ dB with the polish, 13 of 14), sparse-view CT at 10 views ($+0.11$ dB at $G = 8$, 20 of 20 images), MRI at $4\times$ ($+0.47$ dB with the polish) and compressed sensing at $m/n = 0.02$ ($+0.16$ dB). It falls behind wherever one expert holds a wide lead that a pre-commitment cannot anticipate, and in these benchmarks that expert is almost always SIREN:

Table 2: Per benchmark, PSNR (dB). Here the reference is the best fixed expert *for that benchmark*, which differs from one to the next and is a stronger reference than the single expert used over all 372 cells in Table 1; margins in parentheses are relative to it. "Oracle" is the best expert in each cell.

Benchmark	Best fixed	Oracle	SPIRA+P G=8	SPIRA+P G=16	Greedy
Set14 (126)	SIREN 25.51	25.81	25.00 (-0.51)	25.40 (-0.11)	**25.82** ($+0.30$)
Set14 hard (84)	DIP-TV 20.58	21.05	20.42 (-0.16)	20.66 ($+0.08$)	**20.99** ($+0.41$)
Demosaicing (42)	SIREN 27.30	27.35	26.40 (-0.89)	26.67 (-0.63)	**27.18** (-0.12)
CT (60)	TV 30.10	30.10	29.91 (-0.19)	29.57 (-0.53)	29.87 (-0.22)
MRI (60)	TV 25.70	26.01	**25.62** (-0.08)	25.45 (-0.24)	25.57 (-0.13)

deblurring at every blur width (-0.5 to -0.7 dB), demosaicing at every noise level, dense inpainting with 50% and 70% observed, and compressed sensing at $m/n = 0.25$. The margin therefore does not follow difficulty monotonically, and MRI is the counterexample, since the policy wins at $4\times$ and loses at $6\times$ and $8\times$. It does follow how close the library's members are to one another. The polish is what keeps the loss bounded in the hardest cases, because it re-weights already fitted experts from held-out evidence: at $m/n = 0.02$ it turns -7.28 dB into $+0.16$, and at $8\times$ acceleration -3.36 into -0.97.

What the policy selects. Figure 2 shows the executed selections region by region and step by step. The allocation is neither constant nor arbitrary. On *zebra* under deblurring the policy assigns SIREN to the striped band and TV to the smooth background, and it grows that band as the episode advances; under inpainting on the same image it holds DIP-EMA almost everywhere and switches to SIREN over the stripes. On CT it starts with TV in every region and then carves out the air surrounding the patient. On MRI the map is organized in columns aligned with the phase-encoding direction of the Cartesian mask, so the policy is reacting to the geometry of the operator and not only to image content. Across the four cells of the figure, TV takes 52% of the regions, SIREN 30% and DIP-EMA 23%, and the two DIP capacity variants together take less than 1%, which is consistent with the ranking in Table 2.

Qualitative results. Figure 3 shows four of these cells. On *zebra* with 30% of the pixels observed, SIREN, the best expert for that setting, smooths the stripes into grey; the policy keeps them separated because it reserves SIREN for the flat background and fits the striped region with a different prior. On *monarch* with 20% observed the wing pattern survives only in the policy's reconstruction. On sparse-view CT the ten-view back-projection is dominated by streaks, TV alone returns the piecewise-constant reconstruction one expects from it, and the policy recovers more of the mediastinal boundary. The last row is a case we lose: under $\sigma_b = 3$ deblurring SIREN is uniformly better, and the regional composite of the policy leaves visible seams between regions, which is the failure mode of committing per region when one prior is best everywhere.

Run Time and Complexity. Table 3 compares the evaluated methods in terms of optimized capacity, computational cost (TFLOPs, computed as the per-iteration cost of each expert multiplied by the iterations it executes, probe included) and wall-clock inference time, all measured under a standardized setting (running on NVIDIA RTX 3090, input resolution $1 \times 128 \times 128$, seven experts, $T = 5$ decisions of $K = 60$ iterations). Per-iteration cost varies widely across the library, from 0.0 GFLOP for TV, a pixel-space expert with no dense layers, to 3.3 GFLOP for SIREN, so the method that selects cheap priors also spends less. Greedy advances the seven experts at every step and therefore pays for the full library on every cell (3.97 TFLOPs and 2.31 s), while a single expert is cheap (0.98 TFLOPs and 0.60 s) but requires knowing in advance which one to run. **Ours** spends 1.25 TFLOPs and 0.76 s on CT, where it commits to TV, and 3.50 TFLOPs and 1.68 s on demosaicing, where several convolutional experts stay active, yielding a $1.1\times$ to $3.2\times$ speedup over greedy. The policy network accounts for 0.14 M parameters and 1.42 GFLOP per decision, between 0.2 and 0.6% of the total, so the budget is set by the experts it chooses and not by the selector.

Complexity scales linearly with the input resolution for every expert and for the policy, all of which are fully convolutional or act in pixel space: from 64^2 to 128^2 to 256^2 the cost of one iteration goes from 0.80 to 3.20 to 12.79 GFLOP for DIP and from 0.36 to 1.42 to 5.66 GFLOP for the policy. Storing the library takes 10.2 MiB, 0.9 M parameters plus 1.8 M Adam moments, against 0.54 MiB

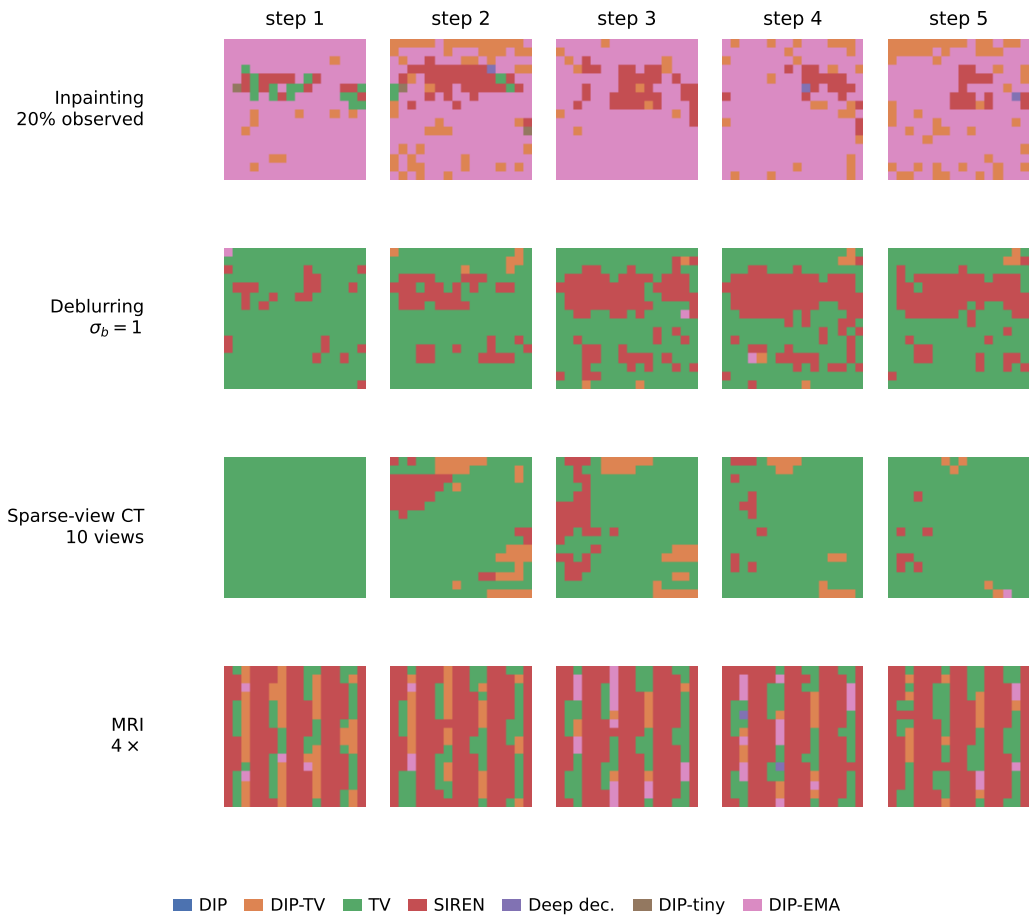

Figure 2: Expert selected in each region at each decision step, policy at $G = 16$ (256 regions). Rows: inpainting with 20% observed and deblurring with $\sigma_b = 1$ on *zebra*, sparse-view CT at 10 views, and MRI at $4\times$. The selection follows image structure on Set14, the support of the object on CT, and the undersampling direction on MRI.

Table 3: Runtime and complexity comparison (test on size $1 \times 128 \times 128$, seven experts). "#Params" counts the parameters optimized during one reconstruction, averaged over cells. **Ours** is reported on its cheapest and its most expensive benchmark; reconstruction quality is given in Table 1.

Method	#Params [M]	TFLOPs	Time [s]
Best fixed expert (SIREN)	0.03	0.98	0.60
Greedy	0.90	3.97	2.31
Ours (CT)	**0.35**	**1.25**	**0.76**
Ours (demosaicing)	0.77	3.50	1.68

for the policy, and the peak activation memory of a single expert, 18 to 82 MiB, exceeds both. Of those 10.2 MiB the policy updates 7.0 to 7.2 MiB per reconstruction on Set14 and CT and all of them on demosaicing and MRI, whereas greedy always updates all of them. Training follows the same accounting: one episode costs 2.2 TFLOPs and the three 700-episode stages of the reported checkpoint cost 4.6 PFLOPs, which the 1.8 TFLOPs saved per reconstruction repay after about 2,500 reconstructions.

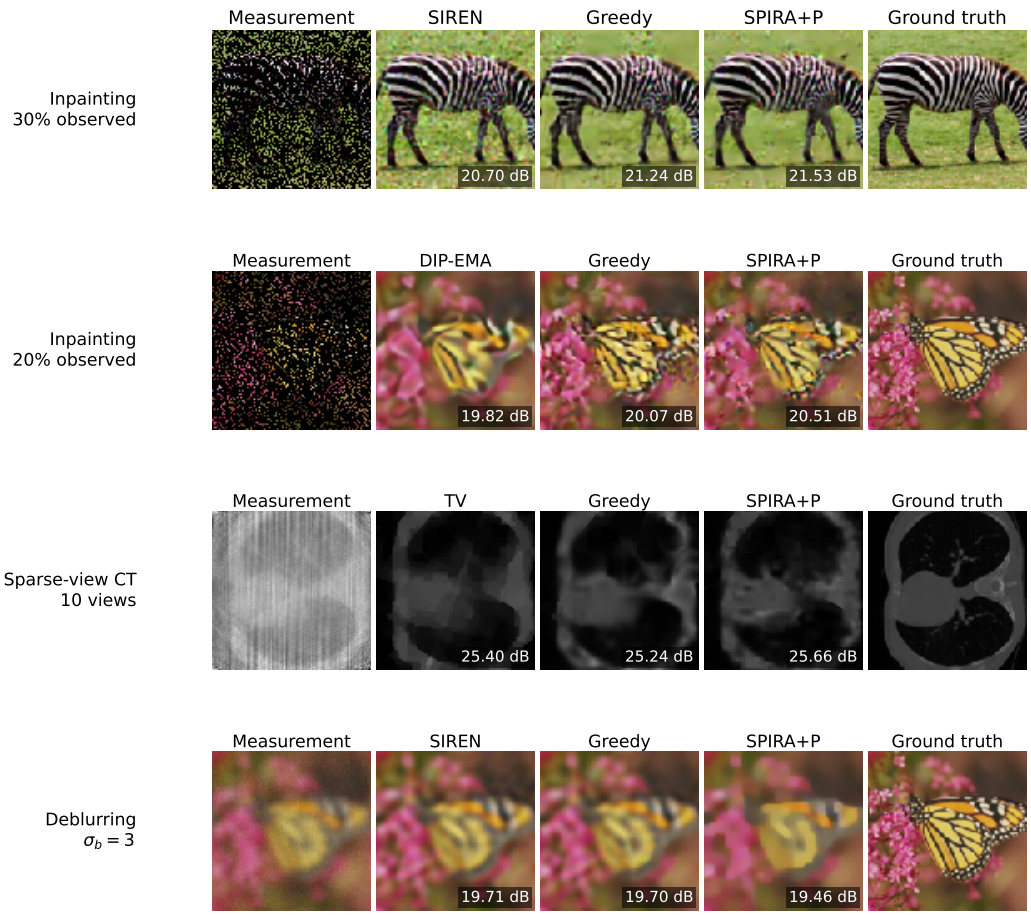

Figure 3: Reconstructions on four cells of the benchmark, with the measurement, the best single expert for that setting, greedy and the policy at $G = 16$ with the polish. The measurement panel shows the observed image for inpainting and deblurring and the ten-view back-projection for CT. PSNR was recomputed on one RTX 5090 for these four cells; the deterministic experts reproduce the values of Table 8 exactly, while the DIP-type experts and the policy vary by up to 0.2 dB across GPUs.

4 Conclusion

We presented SPIRA, a spatial policy that selects, for every region of the reconstruction, which unsupervised expert to use and which experts to optimize. Its core component is a risk estimator trained on synthetic images to forecast the terminal regional error of every expert, so that decisions follow the predicted outcome of optimization rather than the current, biased measurement evidence. Combined with a purely regional reward and a relative regret against the regional oracle, this turns spatial diversity into a property of the optimal policy instead of a regularizer, and it replaces the evaluation of every expert at every step with a single forward pass. Without knowing the best expert in advance, the policy outperforms any single expert used everywhere at one-third to one-half of the expert iterations of greedy selection, and it surpasses even the per-setting best expert in sparse inpainting, extremely undersampled compressed sensing, few-view CT and moderately accelerated MRI, including operators never seen in training. Its main limitation is transfer from synthetic to natural images: priors that rarely win on the synthetic collages, such as the sinusoidal representation, remain under-selected in deblurring and demosaicing, where greedy selection retains an advantage.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based coding and writing assistants) to help implement and debug the evaluation, plotting, and data-processing scripts and the accompanying demo website, to draft and edit prose, and to typeset tables and figures. We have not used generative AI tools to generate research ideas, to design the method or the experiments, or to produce, select or alter experimental results; every controller, rollout and measurement reported here was produced by the code. We have reviewed all AI-assisted work: LLM-generated code was checked and tested by the authors, and all AI-edited text was verified against the underlying results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Daniel Otero Baguer, Johannes Leuschner, and Maximilian Schmidt. Computed tomography reconstruction using deep image prior and learned reconstruction methods. *Inverse Problems*, 36(9): 094004, 2020.

Manel Baradad, Jonas Wulff, Tongzhou Wang, Phillip Isola, and Antonio Torralba. Learning to see by looking at noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Joshua Batson and Loic Royer. Noise2Self: Blind denoising by self-supervision. In *International Conference on Machine Learning (ICML)*, pp. 524–533, 2019.

Mario Bertero, Patrizia Boccacci, and Christine De Mol. *Introduction to inverse problems in imaging*. CRC press, 2021.

Ashish Bora, Ajil Jalal, Eric Price, and Alexandros G Dimakis. Compressed sensing using generative models. In *International conference on machine learning*, pp. 537–546. PMLR, 2017.

E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008. doi: 10.1109/MSP.2007.914731.

Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating. In *Conference on Robot Learning (CoRL)*, pp. 66–75, 2019.

Dongdong Chen, Julián Tachella, and Mike E Davies. Equivariant imaging: Learning beyond the range space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4379–4388, 2021.

Dongdong Chen, Julián Tachella, and Mike E Davies. Robust equivariant imaging: a fully unsupervised framework for learning to image from noisy and partial measurements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5647–5656, 2022.

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

Yuning Cui, Syed Waqas Zamir, Salman Khan, Alois Knoll, Mubarak Shah, and Fahad Shahbaz Khan. AdaIR: Adaptive all-in-one image restoration via frequency mining and modulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=M5t0WvjfCg.

Mauricio Delbracio, Damien Kelly, Michael S. Brown, and Peyman Milanfar. Mobile computational photography: A tour. *Annual Review of Vision Science*, 7:571–604, 2021.

Laura C Diaz-Delgado, Emmanuel Martinez, and Henry Arguello. Frozen clip priors for robust self-supervised poisson inverse problems. In *European Conference on Computer Vision*, pp. 588–604. Springer, 2026.

Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. In *International Conference on Learning Representations (ICLR)*, 2017.

Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008.

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

Maryam Fazel, E Candes, Benjamin Recht, and P Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1043–1047. IEEE, 2008.

Simon Foucart and Holger Rauhut. *An Invitation to Compressive Sensing*, pp. 1–39. Springer New York, New York, NY, 2013. ISBN 978-0-8176-4948-7. doi: 10.1007/978-0-8176-4948-7_1. URL https://doi.org/10.1007/978-0-8176-4948-7_1.

Ryosuke Furuta, Naoto Inoue, and Toshihiko Yamasaki. Fully convolutional network with multi-step reinforcement learning for image processing. In *AAAI Conference on Artificial Intelligence*, pp. 3598–3605, 2019.

Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, pp. 399–406, 2010.

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning (ICML)*, pp. 1352–1361, 2017.

Reinhard Heckel and Paul Hand. Deep decoder: Concise image representations from untrained non-convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, and Kari Pulli. FlexISP: A flexible camera image processing framework. *ACM Transactions on Graphics*, 33(6):1–13, 2014.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6023–6029, 2019.

Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2Void: Learning denoising from single noisy images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2129–2137, 2019.

Ann B. Lee, David Mumford, and Jinggang Huang. Occlusion models for natural images: A comparative study of a new variational model and the dead leaves model. *International Journal of Computer Vision*, 41(1):35–59, 2001.

Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. Noise2Noise: Learning image restoration without clean data. In *International Conference on Machine Learning (ICML)*, pp. 2965–2974, 2018.

Johannes Leuschner, Maximilian Schmidt, Daniel Otero Baguer, and Peter Maass. LoDoPaB-CT, a benchmark dataset for low-dose computed tomography reconstruction. *Sci. Data*, 8(1):109, April 2021.

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

Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

Segolene Martin, Samuel Hurault, Anne Gaubert, Matthieu Kowalski, and Jean-Christophe Pesquet. PnP-Flow: Plug-and-play image restoration with flow matching. In *International Conference on Learning Representations (ICLR)*, 2025.

Michael T McCann, Kyong Hwan Jin, and Michael Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

Christopher A. Metzler, Ali Mousavi, Reinhard Heckel, and Richard G. Baraniuk. Unsupervised learning with Stein's unbiased risk estimator. *arXiv preprint arXiv:1805.10531*, 2018.

Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

Brayan Monroy, Jorge Bacca, and Julián Tachella. Generalized recorrupted-to-recorrupted: Self-supervised learning beyond gaussian noise. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28155–25165. IEEE, 2025.

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

Ryo Nakamura, Ryu Tadokoro, Ryosuke Yamada, Yuki M. Asano, Iro Laina, Christian Rupprecht, Nakamasa Inoue, Rio Yokota, and Hirokatsu Kataoka. Pre-training vision transformers with very limited synthesized images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20360–20369, 2023.

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*, 2020.

Yigang Peng, Jinli Suo, Qionghai Dai, and Wenli Xu. Reweighted low-rank matrix recovery and its application in image restoration. *IEEE transactions on cybernetics*, 44(12):2418–2430, 2014.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems (RSS)*, 2018.

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.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1):259–268, 1992. ISSN 0167-2789. doi: https://doi.org/10.1016/0167-2789(92)90242-F. URL https://www.sciencedirect.com/science/article/pii/016727899290242F.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations (ICLR)*, 2017.

Tom Silver, Kelsey Allen, Josh Tenenbaum, and Leslie Kaelbling. Residual policy learning. *arXiv preprint arXiv:1812.06298*, 2018.

Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Julian Tachella, Dongdong Chen, and Mike E. Davies. Sensing theorems for unsupervised learning in linear inverse problems. *Journal of Machine Learning Research*, 24(39):1–45, 2023a.

Julian Tachella, Dongdong Chen, Samuel Hurault, Matthieu Terris, and Andrew Wang. DeepInverse: A deep learning framework for inverse problems in imaging, June 2023b. URL https://github.com/deepinv/deepinv.

Matthieu Terris, Ulugbek S Kamilov, and Thomas Moreau. Fire: Fixed-points of restoration priors for solving inverse problems. *arXiv preprint arXiv:2411.18970*, 2024a.

Matthieu Terris, Thomas Moreau, Nelly Pustelnik, and Julian Tachella. Equivariant plug-and-play image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25255–25264, 2024b.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28165–28175, 2025.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9446–9454, 2018.

Singanallur V. Venkatakrishnan, Charles A. Bouman, and Brendt Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948, 2013. doi: 10.1109/GlobalSIP.2013.6737048.

Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Carola-Bibiane Schonlieb, and Hua Huang. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *International Conference on Machine Learning (ICML)*, pp. 10158–10169, 2020.

Martin J Willemink, Pim A de Jong, Tim Leiner, Linda M de Heer, Rutger AJ Nievelstein, Ricardo PJ Budde, and Arnold MR Schilham. Iterative reconstruction techniques for computed tomography part 1: technical principles. *European radiology*, 23:1623–1631, 2013.

Jinxi Xiang, Yonggui Dong, and Yunjie Yang. Fista-net: Learning a fast iterative shrinkage thresholding network for inverse problems in imaging. *IEEE Transactions on Medical Imaging*, 40(5): 1329–1339, 2021. doi: 10.1109/TMI.2021.3054167.

Burhaneddin Yaman, Seyed Amir Hossein Hosseini, Steen Moeller, Jutta Ellermann, Kamil Ugurbil, and Mehmet Akcakaya. Self-supervised learning of physics-guided reconstruction neural networks without fully sampled reference data. *Magnetic Resonance in Medicine*, 84(6):3172–3191, 2020.

Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2443–2452, 2018.

Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pp. 711–730, 2010.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26 (7):3142–3155, 2017.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1219–1229, June 2023.

Algorithm 1 SPIRA inference with unsupervised polish (+P). Ground truth is never used and only the selected experts are optimized. $v(\mathbf{x}) = \|\mathbf{M}_{\mathrm{val}}(\mathbf{H}\mathbf{x} - \mathbf{y})\|_2^2 / \operatorname{tr}(\mathbf{M}_{\mathrm{val}})$ is the held-out fidelity and $\delta(\mathbf{x}) = |\log d(\mathbf{x})|$ the log-distance of the discrepancy ratio $d(\mathbf{x}) = \|\mathbf{M}_{\mathrm{sup}}(\mathbf{H}\mathbf{x} - \mathbf{y})\|_2^2 / [\sigma^2 \operatorname{tr}(\mathbf{M}_{\mathrm{sup}})]$ from one; both depend on $\mathbf{y}$ and σ only.

Require: measurements $\mathbf{y}$, operator $\mathbf{H}$, noise level σ; experts $\{f_e\}_{e=1}^{E}$ with probe budgets $\{n_e\}$; trained policy $\boldsymbol{\psi}$; horizon T, iterations per step K; polish flag

split the measurements once: $\mathbf{M}_{\mathrm{fit}}$, $\mathbf{M}_{\mathrm{val}}$
$\mathbf{x}_0 \leftarrow \mathrm{clip}_{[0,1]}\left(\mathbf{H}^{\dagger}\mathbf{M}_{\mathrm{fit}}\mathbf{y}\right)$
for $e = 1, \ldots, E$ **do** ▷ probe: every expert once
    initialize $\boldsymbol{\theta}^{(e)}$ from $\mathbf{x}_0$; run n_e Adam iterations on $\mathcal{J}_e$ (Eq. 1)
    $\mathbf{x}_0^{(e)} \leftarrow f_e(\boldsymbol{\theta}^{(e)})$; $\widetilde{\boldsymbol{\rho}}^{(e)} \leftarrow \widetilde{\boldsymbol{\rho}}\big(\mathbf{x}_0^{(e)}\big)$ ▷ held-out residual map
end for
for $t = 0, \ldots, T-1$ **do**
    form $\mathbf{s}_t$ from $\mathbf{x}_t$, $\{\mathbf{x}_t^{(e)}, \widetilde{\boldsymbol{\rho}}^{(e)}\}_e$, the staleness of each expert, $\mathbf{a}_{t-1}$ and $(\mathbf{H}, \sigma)$
    $\hat{\mathbf{z}} \leftarrow \hat{z}_{\boldsymbol{\psi}}(\mathbf{s}_t)$, $\mathbf{g} \leftarrow g_{\boldsymbol{\psi}}(\mathbf{s}_t)$ ▷ risk and correction heads, Eq. 5
    $a_{t,r} \leftarrow \arg\max_e \left(g_{r,e} - \beta\, \hat{z}_{r,e}\right)$ for $r = 1, \ldots, R$
    $\mathcal{E}_t \leftarrow \{a_{t,1}, \ldots, a_{t,R}\}$
    for $e \in \mathcal{E}_t$ **do** ▷ only selected experts advance
        run K Adam iterations on $\mathcal{J}_e$; $\mathbf{x}_{t+1}^{(e)} \leftarrow f_e(\boldsymbol{\theta}^{(e)})$; $\widetilde{\boldsymbol{\rho}}^{(e)} \leftarrow \widetilde{\boldsymbol{\rho}}\big(\mathbf{x}_{t+1}^{(e)}\big)$
    end for
    $\mathbf{x}_{t+1}^{(e)} \leftarrow \mathbf{x}_t^{(e)}$ for $e \notin \mathcal{E}_t$ ▷ unselected experts keep their state
    $\mathbf{x}_{t+1} \leftarrow \sum_{r=1}^{R} \mathbf{W}_r\, \mathbf{x}_{t+1}^{(a_{t,r})}$ ▷ Eq. 3
end for
if polish **then**
    $a_r^{\mathrm{P}} \leftarrow \arg\min_e\ s_{T,e,r}$ for $r = 1, \ldots, R$ ▷ latest held-out score, Eq. 2
    $\mathbf{x}_{\mathrm{alt}} \leftarrow \sum_{r=1}^{R} \mathbf{W}_r\, \mathbf{x}_T^{(a_r^{\mathrm{P}})}$ ▷ no further optimization
    if $v(\mathbf{x}_{\mathrm{alt}}) < v(\mathbf{x}_T)$ **and** $\delta(\mathbf{x}_{\mathrm{alt}}) \le \delta(\mathbf{x}_T) + 0.2$ **then**
        $\mathbf{x}_T \leftarrow \mathbf{x}_{\mathrm{alt}}$ ▷ held-out fidelity improves, discrepancy stays consistent
    end if
end if
return $\mathbf{x}_T$

ETHICS STATEMENT

This work proposes a reconstruction method for imaging inverse problems and involves no human subjects, personal data or user studies. The policy is trained exclusively on procedurally generated images (Section 2.5); no natural or medical image is used for training. Evaluation uses four standard test photographs distributed with scikit-image, the public Set14 benchmark (Zeyde et al., 2010), the ground-truth test split of the public LoDoPaB-CT dataset (Leuschner et al., 2021), released under a CC BY 4.0 license and derived from anonymized clinical scans, and synthetic gravimetric volumes; all are used within their license terms and only for computing evaluation metrics. The medical images are used to test reconstruction from simulated sparse-view measurements, not for diagnosis, and we make no clinical claim. As with any reconstruction prior, the experts can hallucinate plausible content in the null space of the operator; the method mitigates this by exposing only unsupervised experts at test time and by reporting the cost and the failure modes explicitly (e.g. compressed sensing, where the learned selector does not surpass the held-out baseline), but reconstructions should not be interpreted as measurements in safety-critical use. We foresee no direct negative societal impact beyond those shared by all learned reconstruction methods, and we declare no conflict of interest.

A DETAILED FORMULATION OF SPIRA

This appendix states the method of Section 2 with all dimensions, index ranges and implementation-level definitions.

A.1 NOTATION, EXPERTS AND MEASUREMENT EVIDENCE

Images, measurements and regions. The unknown image $\mathbf{x}^\star \in \mathbb{R}^N$ is vectorized from C channels over N_p pixels, $N = CN_p$ ($C = 1$ for grayscale, CT and MRI magnitude images, $C = 3$ for RGB), and $x_{c,i}$ denotes the entry of $\mathbf{x} \in \mathbb{R}^N$ in channel $c \in \{1, \dots, C\}$ at pixel $i \in \{1, \dots, N_p\}$. The measurements are $\mathbf{y} = \mathbf{H}\mathbf{x}^\star + \boldsymbol{\omega} \in \mathbb{R}^M$ with $\mathbf{H} \in \mathbb{R}^{M \times N}$, $\boldsymbol{\omega} \sim \mathcal{N}(\mathbf{0}_M, \sigma^2 \mathbf{I}_M)$ and $\sigma > 0$. The regions $r \in \{1, \dots, R\}$ have disjoint pixel sets $B_r \subset \{1, \dots, N_p\}$ covering the grid, and smooth windows $\mathbf{w}_r \in [0,1]^{N_p}$ with $\sum_{r=1}^R \mathbf{w}_r = \mathbf{1}_{N_p}$. The regional squared error of $\mathbf{x} \in \mathbb{R}^N$ is

$$\varepsilon_r(\mathbf{x}) = \frac{1}{C\,|B_r|} \sum_{c=1}^{C} \sum_{i \in B_r} \left(x_{c,i} - x^\star_{c,i}\right)^2 \in \mathbb{R}_{\geq 0}. \tag{8}$$

Experts. Expert $e \in \{1, \dots, E\}$ has a generator $f_e : \mathbb{R}^{D_e} \to \mathbb{R}^N$ with D_e parameters and a regularizer $h_e : \mathbb{R}^N \to \mathbb{R}_{\geq 0}$ (Eq. 1). The diagonal masks $\mathbf{M}_{\text{fit}}, \mathbf{M}_{\text{val}} \in \{0,1\}^{M \times M}$ satisfy $\mathbf{M}_{\text{fit}} + \mathbf{M}_{\text{val}} = \mathbf{M}_{\text{sup}}$ and $\mathbf{M}_{\text{fit}}\mathbf{M}_{\text{val}} = \mathbf{0}_{M \times M}$, where $\mathbf{M}_{\text{sup}} \in \{0,1\}^{M \times M}$ selects the valid entries of $\mathbf{y}$ (e.g. observed pixels in inpainting, acquired k-space locations in MRI) and $\operatorname{tr}(\cdot)$ counts selected entries. All experts start from $\mathbf{x}_0 = \operatorname{clip}_{[0,1]}(\mathbf{H}^\dagger \mathbf{M}_{\text{fit}} \mathbf{y}) \in \mathbb{R}^N$, with $\mathbf{H}^\dagger \in \mathbb{R}^{N \times M}$ the pseudo-inverse of $\mathbf{H}$ (or $\mathbf{H}^\top$ when none is available). $\mathcal{U}_K^{(e)} : \mathbb{R}^{D_e} \to \mathbb{R}^{D_e}$ applies $K \in \mathbb{N}$ Adam iterations on Eq. 1, resuming from the current optimizer state, and every expert receives $n_e^{\text{probe}} \in \mathbb{N}$ probe iterations before the first decision.

Measurement evidence. For $\mathbf{x} \in \mathbb{R}^N$, the held-out residual energy map $\boldsymbol{\rho}(\mathbf{x}) \in \mathbb{R}^{N_p}_{\geq 0}$ has entries

$$\rho_i(\mathbf{x}) = \frac{1}{C} \sum_{c=1}^{C} \left(\left[\mathbf{H}^\top \mathbf{M}_{\text{val}} (\mathbf{H}\mathbf{x} - \mathbf{y})\right]_{c,i} \right)^2, \qquad i \in \{1, \dots, N_p\}, \tag{9}$$

and $\widetilde{\boldsymbol{\rho}}(\mathbf{x}) = \boldsymbol{\rho}(\mathbf{x}) / \bar{\rho}_0$ with $\bar{\rho}_0 = N_p^{-1} \mathbf{1}_{N_p}^\top \boldsymbol{\rho}(\mathbf{x}_0) > 0$, a scale that is fixed for the whole reconstruction. The fitting counterpart $\widetilde{\boldsymbol{\rho}}_{\text{fit}}(\mathbf{x}) \in \mathbb{R}^{N_p}_{\geq 0}$ uses $\mathbf{M}_{\text{fit}}$ in Eq. 9 and is normalized by its own mean. The regional score $s_{t,e,r} \in \mathbb{R}_{\geq 0}$ of Eq. 2 averages $\widetilde{\boldsymbol{\rho}}(\mathbf{x}_t^{(e)})$ over B_r, where $\mathbf{x}_t^{(e)} \in \mathbb{R}^N$ is the output of expert e before decision $t \in \{0, \dots, T\}$; the probe makes $s_{0,e,r}$ defined for every e and r.

A.2 TRANSITIONS, COST AND COMPOSITION

Let $\boldsymbol{\theta}_t^{(e)} \in \mathbb{R}^{D_e}$ be the parameters of expert e before decision $t \in \{0, \dots, T-1\}$, with $\boldsymbol{\theta}_0^{(e)}$ the parameters after the probe, and $\mathcal{E}_t = \{a_{t,r} : r \in \{1, \dots, R\}\} \subseteq \{1, \dots, E\}$. The transition is

$$\boldsymbol{\theta}_{t+1}^{(e)} = \begin{cases} \mathcal{U}_K^{(e)}(\boldsymbol{\theta}_t^{(e)}), & e \in \mathcal{E}_t, \\ \boldsymbol{\theta}_t^{(e)}, & e \notin \mathcal{E}_t, \end{cases} \qquad \mathbf{x}_{t+1}^{(e)} = f_e(\boldsymbol{\theta}_{t+1}^{(e)}). \tag{10}$$

With $n_{t,e} \in \mathbb{N}$ the iterations of expert e before decision t ($n_{0,e} = n_e^{\text{probe}}$), an episode costs $\sum_{e=1}^E n_{T,e} = \sum_{e=1}^E n_e^{\text{probe}} + K \sum_{t=0}^{T-1} |\mathcal{E}_t|$ expert iterations, and $n_e^{\text{full}} = n_e^{\text{probe}} + TK$ is the budget of an expert advanced at every decision.

In Eq. 3, $\mathbf{W}_r = \mathbf{I}_C \otimes \operatorname{Diag}(\mathbf{w}_r) \in \mathbb{R}^{N \times N}$ applies window r to every channel ($\otimes$ is the Kronecker product) and $\mathbf{x}_{t+1}^{(a_{t,r})} \in \mathbb{R}^N$ is the output of expert $a_{t,r}$ after Eq. 10. For imaging, the regions are square tiles of side ℓ pixels; along each axis $\mathbf{w}_r$ equals one on its tile and decays to zero over a raised-cosine taper of length $\ell/4$ beyond each tile edge, after which the windows are normalized to sum to one at every pixel. Hence: (i) on the core of each tile, $\ell/2$ pixels wide along each axis, the reconstruction equals the output of the selected expert; (ii) in the $\ell/2$-wide bands centered on tile boundaries it interpolates between the experts selected in adjacent tiles, with weights fixed by the window profiles, and equals a single expert output wherever adjacent tiles select the same expert; (iii) $\mathbf{x}_{t+1} = \mathbf{x}_{t+1}^{(e)}$ when all regions select e.

A.3 STATE

For every expert, the staleness $\kappa_{t,e} = (t - t_{t,e}^{\text{last}})/T \in [0,1]$ records how long ago it was last advanced, with $t_{t,e}^{\text{last}} \in \{0, \dots, t\}$ the last decision before t at which e was advanced (0 if never),

and the progress $\varpi_{t,e} = n_{t,e}/n_e^{\text{full}} \in (0,1]$ its fraction of the full budget. Let $\mathbf{L} \in \mathbb{R}^{N_p \times N}$ map an image to its luminance ($\mathbf{L} = \mathbf{I}_{N_p}$ for $C = 1$ and $[0.299, 0.587, 0.114] \otimes \mathbf{I}_{N_p}$ for $C = 3$). The state is a stack of d_s planes over the pixel grid, $\mathbf{s}_t \in \mathbb{R}^{N_p \times d_s}$,

$$\mathbf{s}_t = \Big[\mathbf{L}\mathbf{x}_t,\ \widetilde{\boldsymbol{\rho}}_{\text{fit}}(\mathbf{x}_t),\ \mathbf{A}_{t-1},\ \{\widetilde{\boldsymbol{\rho}}_t^{(e)},\ \boldsymbol{\Delta}_t^{(e)},\ \kappa_{t,e}\mathbf{1}_{N_p},\ \varpi_{t,e}\mathbf{1}_{N_p},\ \mathbf{L}\mathbf{x}_t^{(e)}\}_{e=1}^{E},\ \tfrac{t}{T}\mathbf{1}_{N_p},\ \mathbf{1}_{N_p}\boldsymbol{\phi}^\top\Big], \quad (11)$$

with the luminance of the reconstruction $\mathbf{L}\mathbf{x}_t \in \mathbb{R}^{N_p}$; the normalized fitting residual $\widetilde{\boldsymbol{\rho}}_{\text{fit}}(\mathbf{x}_t) \in \mathbb{R}_{\geq 0}^{N_p}$; the previous selection map $\mathbf{A}_{t-1} \in [0,1]^{N_p \times E}$, whose column e is $\sum_{r=1}^{R} \mathbb{1}[a_{t-1,r} = e]\, \mathbf{w}_r$ ($\mathbb{1}[\cdot]$ is the indicator function; $\mathbf{A}_{-1} = \mathbf{0}_{N_p \times E}$); for each expert, its normalized held-out residual map $\widetilde{\boldsymbol{\rho}}_t^{(e)} = \widetilde{\boldsymbol{\rho}}(\mathbf{x}_t^{(e)}) \in \mathbb{R}_{\geq 0}^{N_p}$, the change $\boldsymbol{\Delta}_t^{(e)} \in \mathbb{R}^{N_p}$ of this map produced by the last advancement of expert e before decision t (for the probe, $\boldsymbol{\Delta}_0^{(e)} = \widetilde{\boldsymbol{\rho}}(\mathbf{x}_0) - \widetilde{\boldsymbol{\rho}}_0^{(e)}$), its staleness and progress broadcast over the pixels, and the luminance of its current output $\mathbf{L}\mathbf{x}_t^{(e)} \in \mathbb{R}^{N_p}$; the time plane; and the problem embedding $\boldsymbol{\phi} = [\boldsymbol{\tau}_H^\top, 10\sigma, \eta]^\top \in \mathbb{R}^{d_\phi}$ broadcast over the pixels, where $\boldsymbol{\tau}_H \in \{0,1\}^F$ one-hot encodes the operator family among F training families, $\eta \in (0,1]$ is a measurement-rate proxy (observed fraction for inpainting, $1/(1+\sigma_b)$ for a Gaussian blur of width σ_b, M/N for compressed sensing, number of views over 128 for CT), and $d_\phi = F + 2$. In total $d_s = 3 + 6E + d_\phi$. Operators outside the training families are handled zero-shot through the embedding of the closest family.

A.4 Policy network and risk estimator

The terminal risk is $\varepsilon_{e,r}^{\text{T}} = \varepsilon_r(\mathbf{x}^{(e)}) \in \mathbb{R}_{>0}$, evaluated at the output of expert e after n_e^{full} iterations. In Eq. 4, $\mu_r = E^{-1}\sum_{j=1}^{E} \log \varepsilon_{j,r}^{\text{T}} \in \mathbb{R}$ and $\varsigma_r = \big((E-1)^{-1}\sum_{j=1}^{E}(\log \varepsilon_{j,r}^{\text{T}} - \mu_r)^2\big)^{1/2} \in \mathbb{R}_{\geq 0}$ are the mean and sample standard deviation of the log-risks in region r, and the targets form $\mathbf{Z} = [z_{e,r}] \in \mathbb{R}^{E \times R}$.

The policy $\pi_{\boldsymbol{\psi}}$, with parameters $\boldsymbol{\psi}$, is a fully convolutional network. A trunk maps $\mathbf{s}_t$ to features pooled onto the regions, $\boldsymbol{\Phi}_{\boldsymbol{\psi}}(\mathbf{s}_t) \in \mathbb{R}^{R \times d_f}$. These are concatenated with a score shortcut that exposes the measurement evidence of region r directly: the standardized log held-out scores $\tilde{z}_{t,e,r} \in \mathbb{R}$ (Eq. 4 applied to $s_{t,e,r}$), their log-trend over the last advancement $\zeta_{t,e,r} = \log\big((s_{t,e,r} + \bar{\Delta}_{t,e,r})/s_{t,e,r}\big) \in \mathbb{R}$ with $\bar{\Delta}_{t,e,r} = |B_r|^{-1}\sum_{i \in B_r} \Delta_{t,i}^{(e)}$, and $\kappa_{t,e}, \varpi_{t,e}$, giving $\mathbf{f}_{t,r} \in \mathbb{R}^{d_f + 4E}$ per region. Three linear heads shared across regions read $\mathbf{f}_{t,r}$ and output the risk estimate $\hat{\mathbf{z}}_{\boldsymbol{\psi}}(\mathbf{s}_t)_r \in \mathbb{R}^E$, the correction $\mathbf{g}_{\boldsymbol{\psi}}(\mathbf{s}_t)_r \in \mathbb{R}^E$ and the value $V_{\boldsymbol{\psi}}(\mathbf{s}_t)_r \in \mathbb{R}$. The action distribution factorizes over regions, $\pi_{\boldsymbol{\psi}}(\mathbf{a}_t \mid \mathbf{s}_t) = \prod_{r=1}^{R} \pi_{\boldsymbol{\psi}}(a_{t,r} \mid \mathbf{s}_t)$, with the softmax of Eq. 5, $\beta = \exp(b)$ for a learnable $b \in \mathbb{R}$, and $\text{sg}[\cdot]$ the identity in the forward pass and zero in the backward pass. For $\mathbf{g} \equiv \mathbf{0}$ the policy tends to $\arg\min_e \hat{z}_{\boldsymbol{\psi}}(\mathbf{s}_t)_{r,e}$ as $\beta \to \infty$.

A.5 Reward, returns and objective

In Eq. 6, $\bar{\varepsilon}_0 = N^{-1}\|\mathbf{x}_0 - \mathbf{x}^\star\|_2^2 > 0$, $e_r^\star = \arg\min_{e \in \{1,\dots,E\}} \varepsilon_{e,r}^{\text{T}}$, $\lambda_{\text{reg}} \geq 0$, and the regret term is clipped from below at $-c_{\text{clip}}$, $c_{\text{clip}} > 0$, to bound outlier ratios. With $\tilde{r}_{t,r} = r_{t,r}$ for $t < T-1$ and $\tilde{r}_{T-1,r} = \gamma_T\, r_{T-1,r}$, $\gamma_T \geq 1$, the undiscounted return is $R_{t,r} = \sum_{k=t}^{T-1} \tilde{r}_{k,r}$; the improvement terms telescope to the regional gain of the final reconstruction over $\mathbf{x}_0$, so the return rewards final quality rather than speed of convergence. The advantages $A_{t,r} = R_{t,r} - V_{\boldsymbol{\psi}}(\mathbf{s}_t)_r$ are standardized over the $T \times R$ entries of each episode, $\hat{A}_{t,r} = (A_{t,r} - \bar{A})/(\text{std}(A) + 10^{-6})$, and treated as constants. The actor–critic loss of Eq. 7 is

$$\mathcal{L}_{\text{AC}} = \frac{1}{TR}\sum_{t=0}^{T-1}\sum_{r=1}^{R}\Big[-m_{t,r}\hat{A}_{t,r}\log \pi_{\boldsymbol{\psi}}(a_{t,r} \mid \mathbf{s}_t) + c_v\big(V_{\boldsymbol{\psi}}(\mathbf{s}_t)_r - R_{t,r}\big)^2 - \beta_H\, \mathcal{H}\big(\pi_{\boldsymbol{\psi}}(\cdot \mid \mathbf{s}_t)_r\big)\Big], \quad (12)$$

with $\mathcal{H}(\pi(\cdot \mid \mathbf{s}_t)_r) = -\sum_{e=1}^{E} \pi(e \mid \mathbf{s}_t)_r \log \pi(e \mid \mathbf{s}_t)_r$ and weights $c_v, \beta_H \geq 0$; $c_z, c_g \geq 0$ weight the risk regression and the correction penalty in Eq. 7. The expectation is taken over synthetic images $\mathbf{x}^\star \sim p_{\text{syn}}$, operators and noise levels $(\mathbf{H}, \sigma) \sim p_{\text{op}}$, and actions $\mathbf{a}_t \sim \pi_{\boldsymbol{\psi}}$.

Exploration. With probability $p_{\text{s}} \in [0,1]$ at each decision, one region drawn uniformly from $\{1,\dots,R\}$ is assigned to the least recently advanced expert, $\arg\max_e \kappa_{t,e}$. This entry is excluded

from the policy-gradient term ($m_{t,r} = 0$; $m_{t,r} = 1$ otherwise) but still advances the expert and contributes to the critic and risk terms.

Curriculum and inference. The policy is first trained at a low resolution on a coarse region grid, then warm-started at the target resolution and on finer grids; the fully convolutional architecture shares all parameters across resolutions and grids. At test time, $a_{t,r} = \arg\max_e \pi_{\boldsymbol{\psi}}(a_{t,r} = e \mid \mathbf{s}_t)$ for all $t \in \{0, \dots, T-1\}$ and $r \in \{1, \dots, R\}$, only the experts in $\mathcal{E}_t$ are advanced, and $\mathbf{x}_T$ is returned.

B Architecture and Training Details

This appendix lists the expert library, the policy network, the synthetic training distribution and all hyperparameters of the imaging policy used in our experiments. Values refer to 128×128 reconstructions unless stated otherwise.

B.1 Expert library

The library contains $E = 7$ experts (Table 4). Every expert minimizes Eq. 1 with Adam, starting from the fitting-only initialization $\mathbf{x}_0$, and keeps its parameters and optimizer state for the whole episode. The convolutional generators follow the deep image prior (Ulyanov et al., 2018): a fixed random code $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, 0.1^2\mathbf{I})$ of spatial size $H/2^S \times W/2^S$ is decoded by S stages, each consisting of $2\times$ bilinear upsampling followed by two 3×3 convolutions with batch normalization and LeakyReLU (slope 0.1), then a 1×1 convolution and a sigmoid. The deep decoder (Heckel & Hand, 2019) uses only 1×1 convolutions: a fixed code $\mathbf{z} \sim \mathcal{U}(0, 1)$ with $k = 64$ channels at $H/16 \times W/16$ passes through four stages of 1×1 convolution, $2\times$ bilinear upsampling, ReLU and batch normalization, followed by a 1×1 convolution, ReLU, batch normalization, a final 1×1 convolution and a sigmoid; it has far fewer parameters than pixels. DIP-EMA returns the exponential moving average of its iterates with decay 0.97 instead of the last iterate, a cheap surrogate for early stopping. The sinusoidal representation (Sitzmann et al., 2020) maps pixel coordinates in $[-1, 1]^2$ through three sine layers of width 128 with $\omega_0 = 15$ and a linear output with a sigmoid; the reduced frequency and a weight decay of 10^{-4} prevent it from fitting high-frequency content that the operator cannot observe. The total variation is the anisotropic Charbonnier penalty $\mathrm{TV}(\mathbf{x}) = \overline{(\nabla_v \mathbf{x})^2 + \epsilon}^{1/2} + \overline{(\nabla_h \mathbf{x})^2 + \epsilon}^{1/2}$, with $\bar{\cdot}$ the mean over all entries, ∇_v, ∇_h vertical and horizontal finite differences, and $\epsilon = 10^{-6}$. Before the first decision, each expert receives a probe of 20 base iterations multiplied by a family-specific factor, so that probe scores reflect comparable optimization progress rather than raw convergence speed.

Table 4: Expert library. "Params." is the number of optimized parameters for a 128×128 grayscale image; "Probe" the number of probe iterations. All experts use Adam.

Expert	Parameterization f_e	$\lambda_e h_e$	Params.	Adam lr	Probe
DIP	conv. decoder, width 64, $S = 4$ stages, code 16 ch.	—	269,889	10^{-2}	60
DIP-TV	same as DIP	$0.02\,\mathrm{TV}$	269,889	10^{-2}	60
DIP-tiny	conv. decoder, width 16, $S = 3$ stages, code 8 ch.	—	15,025	10^{-2}	60
DIP-EMA	same as DIP, output = EMA of iterates (decay 0.97)	—	269,889	10^{-2}	60
Deep decoder	1×1 conv. decoder, $k = 64$, 4 stages	—	25,281	10^{-2}	60
TV	pixels, initialized at $\mathbf{x}_0$	$0.03\,\mathrm{TV}$	16,384	5×10^{-2}	20
SIREN	sine MLP 3×128, $\omega_0 = 15$, weight decay 10^{-4}	—	33,537	10^{-3}	30

B.2 Policy network

Table 5 summarizes the policy. For grayscale observations and $E = 7$, the state of Eq. 11 has $d_s = 3 + 6E + d_\phi = 51$ planes ($d_\phi = 6$: four operator families, the scaled noise level and the rate proxy). On RGB problems the reconstruction and the expert outputs are observed as luminance, so the same weights apply. The trunk is fully convolutional and its features are average-pooled onto the $G \times G$ region grid, so a policy trained at one resolution and grid runs on others; the three heads

are 1×1 convolutions on the $64 + 4E = 92$ features per region (pooled trunk features and the score shortcut of Appendix A.4). The correction head is zero-initialized and β is initialized to 3, so the untrained policy is driven by the risk estimate alone.

Table 5: Policy network ($E = 7$, grayscale observations).

Component	Setting
Input	state $\mathbf{s}_t \in \mathbb{R}^{N_p \times 51}$
Trunk	4 blocks: 3×3 conv. (64 ch., strides $1, 2, 2, 1$) + GroupNorm (8 groups) + SiLU
Region pooling	average pooling of trunk features onto the $G \times G$ grid
Score shortcut	$\tilde{z}_{t,e,r}$, $\zeta_{t,e,r}$, $\kappa_{t,e}$, $\varpi_{t,e}$ per expert ($4E = 28$ features)
Heads	1×1 conv. on 92 features: risk $\hat{\mathbf{z}} \in \mathbb{R}^E$, correction $\mathbf{g} \in \mathbb{R}^E$, value $V \in \mathbb{R}$
Initialization	correction head zero-initialized; $\beta = \exp(b)$ with $\beta_{\text{init}} = 3$
Parameters	142,132

B.3 Training distribution

Synthetic images. Each episode draws a new grayscale image from a procedural generator. A base canvas and zero to three additional zones are filled by independently drawn families with probabilities: textured dead leaves with power-law radii (0.30), $1/f^\alpha$ Gaussian random fields with $\alpha \sim \mathcal{U}(1.4, 3.2)$ (0.20), flat dead leaves (Lee et al., 2001) (0.10), multi-octave value noise (0.10), smooth ramps and gratings (0.12), Voronoi mosaics (0.13) and sparse impulses (0.05). Zone masks are thresholded smooth random fields (probability 0.40), unions of discs (0.25), oblique half-planes (0.20) or axis-aligned rectangles (0.15), so content boundaries cross the region grid at arbitrary positions and orientations. Each zone is defocused with probability 0.2 and the whole image with probability 0.25; finally a random affine intensity map and a gamma correction $\gamma = \exp(u)$, $u \sim \mathcal{U}(-0.5, 0.5)$, are applied.

Operators. The operator family is drawn uniformly from four families, and its parameters as follows:

- inpainting: observed fraction $\sim \mathcal{U}(0.3, 0.7)$, $\sigma \sim \mathcal{U}(0.01, 0.10)$;
- Gaussian deblurring (FFT): $\sigma_b \sim \mathcal{U}(1, 3)$, $\sigma \sim \mathcal{U}(0.01, 0.05)$;
- compressed sensing (dense Gaussian matrix at 64×64, random signs followed by a subsampled orthogonal transform at 128×128): $M/N \sim \mathcal{U}(0.15, 0.5)$, $\sigma \sim \mathcal{U}(0.005, 0.03)$;
- sparse-view CT (parallel beam): $\mathrm{round}(\mathcal{U}(10, 40))$ of 1000 views, $\sigma \sim \mathcal{U}(0.005, 0.03)$.

The measurements are split once into 85% fitting and 15% held-out entries: Bernoulli over the supported entries for inpainting and compressed sensing, over 4×4 blocks for deblurring (where isolated held-out pixels could be matched by interpolation), and over complete projection views for CT. All operators are implemented with DeepInverse (Tachella et al., 2023b).

B.4 Hyperparameters and training schedule

Table 6 lists the hyperparameters of the final stage. Training follows a three-stage curriculum (Table 7); each stage runs 700 episodes, with every episode drawing a new image and a new problem, and is warm-started from the previous one. The first stage uses the earlier reward that combines a regional improvement (weight 0.4), an image-level improvement (weight 0.6) and a contrast against the mean alternative (weight 0.3), and advances all experts at every decision; the second stage switches to on-policy advancement with catch-up at the target resolution; the third stage introduces the purely regional reward with relative regret of Eq. 6 and refines the region grid. Problems, splits and expert initializations are derived deterministically from the stage seed.

Inference and evaluation. At test time the policy acts deterministically ($\arg\max$ of Eq. 5) for $T = 5$ decisions with $K = 60$ iterations per selected expert, and only the selected experts are advanced. Single-expert baselines receive $T \cdot K = 300$ iterations, and greedy selection advances all seven experts at every decision ($E \cdot T \cdot K = 2{,}100$ iterations). All methods share the measurements,

Table 6: Hyperparameters of the final training stage (symbols as in Section 2 and Appendix A).

Group	Hyperparameter	Value
Episode	decisions T / iterations per decision K	5 / 60
	probe iterations (total over experts)	350
	image size / region grid	128×128 / $G \in \{4, 8, 16\}$
	fit / held-out split	0.85 / 0.15
Risk	ϵ_z / β_{init}	0.1 / 3
Reward	regret weight λ_{reg} / clip c_{clip}	0.3 / 3
	terminal scale γ_T / discount	3 / 1 (undiscounted)
Loss	critic c_v / risk c_z / correction c_g	0.5 / 1.0 / 0.01
	entropy β_H	$0.006 \rightarrow 0.003$ (linear)
	forced-exploration probability p_{s}	0.5
Optimizer	Adam, learning rate	5×10^{-4}
	episodes per update / gradient-norm clip	4 / 1.0
	episodes per stage	700

Table 7: Training curriculum. Every stage runs 700 episodes and is warm-started from the previous one.

Stage	Resolution / grid	Advanced experts	Reward	c_g	Ep./update
1	64^2 / 4×4	all, every decision	regional + global + contrast	—	2
2	128^2 / 4×4	selected + catch-up	regional + global + contrast	0.05	4
3a	128^2 / 8×8	selected + catch-up	regional + regret (Eq. 6)	0.01	4
3b	128^2 / 16×16	selected + catch-up	regional + regret (Eq. 6)	0.01	4

splits and per-expert initializations of each test cell; wall-clock times are measured on a single NVIDIA RTX 3090 with CUDA synchronization, one method at a time.

C Additional Results

C.1 Per-setting results on the five-benchmark suite

Table 8 expands Table 2 to the 24 individual settings. The reference is the best single expert *for that setting*, chosen with the ground truth and therefore a stronger oracle than the fixed expert of Table 1: it may be a different expert in every row. Margins are mean paired differences in PSNR, with the number of cells in which the method beats that reference.

C.2 Outside the training regime

Table 9 evaluates the same hard-selection policy on settings harder than those seen during training: compressed sensing at a 0.10 rate (training sampled 0.15 to 0.50), inpainting with 15% of the pixels observed (training sampled 0.30 to 0.70), and Gaussian blur of width $\sigma_b = 4$ (training sampled 1 to 3). The policy keeps selecting TV, so its accuracy follows that of TV alone: it matches greedy under the strongest blur and falls behind on the two sampling operators, sharply so for compressed sensing on the natural benchmark, where TV itself degrades to 10.88 dB. The terminal polish, which re-weights the already fitted experts from their held-out scores, recovers most of that loss (21.13 dB).

Table 8: All 24 settings. Reference: best single expert per setting (PSNR in dB, oracle). **Bold** marks a positive margin.

Benchmark	Task	Setting	Best single	Greedy	SPIRA+P $G{=}8$	SPIRA+P $G{=}16$
Set14	CS	$m/n = 0.1$	DIP-tiny 23.41	$+0.15$ **(11/14)**	-1.44 (3/14)	-0.19 (6/14)
	CS	$m/n = 0.25$	SIREN 27.22	$+0.01$ **(1/14)**	-1.54 (1/14)	-1.25 (1/14)
	CS	$m/n = 0.5$	TV 29.55	$+0.39$ **(7/14)**	-0.00 (5/14)	-0.09 (2/14)
	Deblurring	$\sigma_b = 1$	SIREN 26.46	-0.18 (4/14)	-0.23 (4/14)	$+0.11$ **(9/14)**
	Deblurring	$\sigma_b = 2$	SIREN 23.80	-0.09 (3/14)	-0.70 (0/14)	-0.53 (0/14)
	Deblurring	$\sigma_b = 3$	SIREN 22.30	-0.03 (4/14)	-0.73 (0/14)	-0.62 (0/14)
	Inpainting	30% obs.	DIP-tiny 24.00	$+0.52$ **(13/14)**	$+0.37$ **(11/14)**	$+0.44$ **(13/14)**
	Inpainting	50% obs.	SIREN 26.50	$+0.20$ **(12/14)**	-0.78 (2/14)	-0.18 (3/14)
	Inpainting	70% obs.	SIREN 28.04	$+0.10$ **(8/14)**	-1.20 (0/14)	-0.37 (2/14)
Set14 hard	CS	$m/n = 0.02$	DIP-TV 18.50	-0.04 (6/14)	$+0.16$ **(10/14)**	-0.36 (4/14)
	CS	$m/n = 0.05$	DIP-TV 21.40	$+0.21$ **(10/14)**	-2.19 (1/14)	$+0.00$ **(7/14)**
	Deblurring	$\sigma_b = 4$	SIREN 21.26	$+0.00$ **(6/14)**	-0.70 (0/14)	-0.63 (0/14)
	Deblurring	$\sigma_b = 5$	SIREN 20.56	-0.10 (3/14)	-0.74 (0/14)	-0.67 (0/14)
	Inpainting	10% obs.	DIP-EMA 21.21	-0.26 (5/14)	-0.11 (4/14)	-0.38 (0/14)
	Inpainting	20% obs.	DIP-EMA 22.81	$+0.39$ **(13/14)**	$+0.37$ **(12/14)**	$+0.27$ **(8/14)**
Demosaicing	Demosaicing	$\sigma = 0.01$	SIREN 28.67	-0.29 (2/14)	-0.93 (2/14)	-0.82 (0/14)
		$\sigma = 0.03$	SIREN 27.54	-0.30 (2/14)	-1.45 (1/14)	-0.91 (1/14)
		$\sigma = 0.05$	SIREN 25.68	$+0.23$ **(8/14)**	-0.30 (6/14)	-0.16 (5/14)
CT	CT	10 views	TV 28.71	-0.24 (10/20)	-0.04 (13/20)	-0.27 (12/20)
		20 views	TV 30.88	-0.29 (0/20)	-0.25 (1/20)	-0.51 (0/20)
		30 views	TV 30.69	-0.14 (5/20)	-0.27 (2/20)	-0.81 (0/20)
MRI	MRI	$4\times$	TV 27.81	$+0.64$ **(9/20)**	$+0.41$ **(9/20)**	$+0.47$ **(7/20)**
		$6\times$	TV 25.95	-0.01 (5/20)	-0.30 (6/20)	-0.23 (7/20)
		$8\times$	TV 23.33	-1.01 (0/20)	-0.36 (7/20)	-0.97 (3/20)

Table 9: Settings outside the training regime, PSNR (dB). CS at a 0.10 rate, inpainting with 15% observed pixels, Gaussian blur $\sigma_b = 4$; five experts, hard selection.

Benchmark	Setting	SPIRA	SPIRA+P	Greedy	TV alone	SIREN alone
Natural	CS 0.10	11.20	21.13	**21.92**	10.88	20.85
	Inpainting 15%	20.07	20.66	**21.21**	19.98	18.86
	Blur $\sigma_b = 4$	19.57	19.57	19.23	**19.64**	17.90
Set14	CS 0.10	21.88	21.88	**23.12**	21.84	21.61
	Inpainting 15%	20.80	21.26	**22.33**	20.70	20.22
	Blur $\sigma_b = 4$	20.49	20.49	20.61	20.45	**20.98**