ALPS: AnneaLed Posterior Sampling for Inverse Imaging via Diffusion-to-Energy Distillation

Anonymous authors
Paper under double-blind review

Abstract

Diffusion models have emerged as the dominant prior for inverse imaging problems. Existing diffusion posterior sampling algorithms are constrained to follow the reverse diffusion trajectory - i.e., the *prior path* from Gaussian noise to clean images - while enforcing data consistency. Methods with convergence guarantees typically rely on computationally expensive ordinary or stochastic differential solvers to map the current noisy latent to a denoised image at each iteration, resulting in a multi-kernel convergence analysis. We propose ALPS (AnneaLed Posterior Sampling), which departs from the above paradigm by defining an explicit family of *posterior distributions* indexed by noise scale. It then runs a preconditioned Metropolis-adjusted Langevin algorithm at a discrete sequence of noise scales along this posterior path. The key enabler of ALPS is the multi-scale energy-based model (EBM) distilled from a pre-trained diffusion model. The resulting energy formulation enables ALPS to unlock three capabilities: analytical composition of the negative log-prior and negative log-likelihood at each noise scale in the discrete annealing schedule, which eliminates latent-image transport at each iteration; a Metropolis-Hastings correction that guarantees convergence at each scale without multi-kernel analysis; and maximum-a-posteriori estimation without contraction assumptions on the score model. We establish guarantees on the geometric ergodicity of each annealed chain, convergence of the annealed sequence to the true posterior, and a non-asymptotic mixing bound in which preconditioning mitigates the dominant ill-conditioning induced by the forward operator. The pseudo-inverse solution initialization at the coarsest scale and geometry-aware preconditioner further accelerate convergence in practice. Experiments on image inpainting, Gaussian, motion deblurring, and accelerated MRI reconstruction demonstrate consistent improvements over diffusion-based sampling algorithms. Notably, these gains arise from the posterior formulation rather than a stronger prior, as the EBM is distilled from the same diffusion model used by the baselines.

1 Introduction

Inverse problems like image inpainting, deblurring, and MRI reconstruction involve recovering an unknown image $\boldsymbol{x}$ from degraded measurements $\mathbf{y}$. Under the Bayesian formulation, recovery is posed as sampling from or optimizing the posterior distribution $p(\boldsymbol{x}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{x})\, p(\boldsymbol{x})$, where $p(\mathbf{y}|\boldsymbol{x})$ denotes the likelihood induced by the forward model and $p(\boldsymbol{x})$ is the prior distribution over images. Diffusion models have emerged as powerful image priors by learning the score of the data distribution across a continuum of noise levels, defining a *prior path* that connects a Gaussian distribution at high noise to the data distribution at low noise Song et al. (2021); Karras et al. (2022). Modern diffusion-based posterior samplers navigate this prior trajectory, while incorporating data-consistency (DC) corrections Jalal et al. (2021); Chung et al. (2023); Song et al. (2023); Chung et al. (2024); Zhu et al. (2023); Mardani et al. (2024). Because these corrections perturb the prior-driven reverse dynamics, careful step-size calibrations is often required to avoid destabilizing the sampling process. A deeper obstacle is that the diffusion-based posterior algorithms typically rely on the score, which provides only local gradient information, without access to the log-prior or log-posterior itself. Consequently, they lack a principled criterion for determining whether a generated sample lies in a high-posterior region. More broadly, score-based samplers implement fine-step discretization of a continuous-time reverse diffusion

process, and the resulting bias from finite step sizes and truncated iteration counts is not corrected by the underlying dynamics. Principled approaches Xu & Chi (2024); Wu et al. (2024) alternate between denoising updates and DC correction with formal convergence guarantees. However, they incur high computational complexity due to ordinary differential equation (ODE) solves or multi-step denoising chains, and require a more involved multi-kernel convergence analysis. We briefly review these methods in Sec. 2.

We propose a complementary formulation based on *posterior path*. The resulting algorithm, ALPS (**A**nnea**L**ed **P**osterior **S**ampling), targets an explicit posterior distribution at each noise scale using a preconditioned Metropolis adjusted Langevin algorithm (MALA) Roberts & Tweedie (1996). The key enabler is a multi-scale energy-based model (EBM) whose scalar potential $E_\theta(\boldsymbol{x}; t)$ composes analytically with the negative log-likelihood at every scale, yielding an explicit negative log-posterior $C_t(\boldsymbol{x})$. We distill the multi-scale EBM from a pre-trained diffusion model Thornton et al. (2025), sidestepping the instabilities of contrastive divergence training. Our main contributions are described below; the computational efficiency, convergence guarantees, and performance improvements can be attributed to four design choices that distinguish it from prior work:

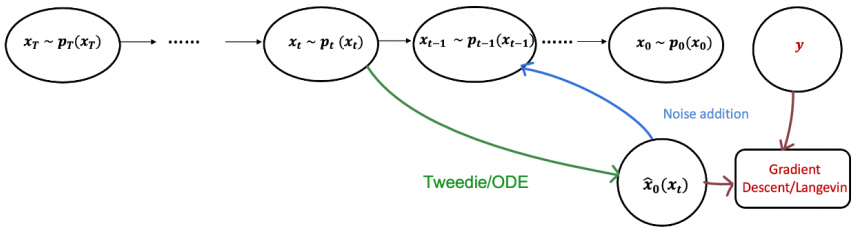
1. **Posterior path with pseudo-inverse initialization:** Unlike current diffusion posterior samplers, which discretize a continuous prior-driven trajectory with intermittent DC corrections, ALPS constructs an explicit discrete sequence of time-dependent posteriors $\{p_t(\boldsymbol{x}|\mathbf{y})\}$ and samples each one directly (see Fig. 1 for an illustration). This design eliminates the need to transport samples from noisy latent states to denoised image estimates, thereby significantly reducing the function evaluations and sidesteps the discretization errors associated with reverse-diffusion dynamics. At each scale, the Markov Chain Monte Carlo (MCMC) kernel converges geometrically from its initialization (see Sec. 4). Furthermore, the chain at the coarsest scale is initialized from the pseudo-inverse solution rather than from pure Gaussian noise, placing it closer to the target posterior from the outset and thereby further accelerating convergence.

2. **Geometry-aware preconditioner:** The proposed preconditioner $\boldsymbol{B}_t$ incorporates the curvature of the negative log-likelihood into the proposal geometry, thereby alleviating the ill-conditioning induced by the forward operator. As a result, the effective condition number is governed primarily by the anisotropy of the prior energy, enabling fast mixing even in severely ill-posed inverse problems.

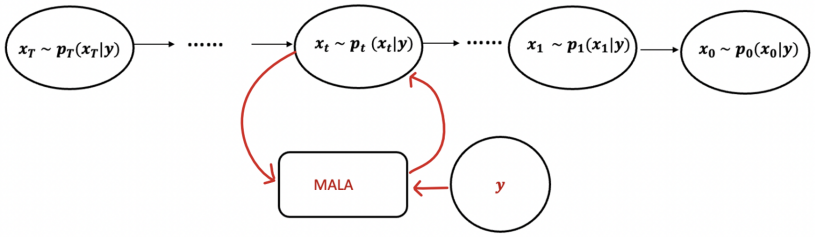
3. **Explicit energy formulation, enabling analytical composition:** Because the EBM composes analytically with the negative log-likelihood, the posterior energy is available in closed-form at every scale. This explicit energy enables a Metropolis-Hastings (MH) correction that guarantees convergence and corrects the discretization bias. The explicit energy also enables local maximum-a-posteriori (MAP) estimation via majorization minimization (MM) with guaranteed monotonic energy decrease, without contraction assumptions on the score network.

4. **Formal convergence guarantees:** Because ALPS targets a fixed posterior at each scale with a single MALA kernel, convergence analysis reduces to single-kernel theory - simpler than the multi-kernel composition required by prior-path methods Xu & Chi (2024); Wu et al. (2024). Asymptotically, the MH-corrected chain converges geometrically to the target posterior distribution at every scale (Theorem 4.1), and the annealed sequence converges to the true posterior distribution as $t \to 0$ (Theorem 4.2). A non-asymptotic bound (Theorem 4.3) quantifies how the total variation error decays with Langevin steps and accumulates from approximation errors. The geometry-aware preconditioner improves the effective conditioning of the target distribution, which enlarges the spectral gap of the MALA kernel and thereby accelerates mixing.

We evaluate ALPS on face inpainting, Gaussian deblurring, motion deblurring, and accelerated MRI at 320×320, demonstrating consistent improvements over diffusion-based samplers using the same pretrained prior. Sec. 2 reviews background; Sec. 3 presents the EBM and algorithm; Sec. 4 establishes convergence; Secs. 5 and 6 present results and discussion.



(a)



(b)

Figure 1: Illustration of posterior sampling strategies: (a) current diffusion posterior sampling methods either obtain the Tweedie estimate $\hat{\boldsymbol{x}}_0(x_t) = D_\theta(\boldsymbol{x}_t, \sigma_t)$ at each step and approximate the likelihood gradient $\nabla_{\boldsymbol{x}_t} \log p(\mathbf{y}|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$ Chung et al. (2023); Song et al. (2023); Chung et al. (2024); Zhu et al. (2023); Mardani et al. (2024), or alternate between a full ODE integration from $\boldsymbol{x}_t$ to $\hat{\boldsymbol{x}}_0$, a Langevin step that enforces DC at the clean-image level, and a re-noising step back to scale $t-1$ Xu & Chi (2024); Wu et al. (2024). The back and forth transport between the noisy latents and the image domain, which is needed to keep the latents in the prior path, is associated with high computational complexity. (b) By contrast, ALPS constructs a posterior path; it constructs an explicit negative log-posterior distribution at discrete noise scales along the path, and sample from them directly using preconditioned MALA with MH correction. Each sampling step converges geometrically to its corresponding posterior distribution, with the convergence rate governed by the spectral gap of the MALA kernel, which is improved by the geometry-aware preconditioner. In addition, pseudo-inverse initialization at the coarsest scale provides a DC warm start.

2 Background and Related Work

We consider the linear inverse problem

$$\mathbf{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{n}, \quad \boldsymbol{n} \sim \mathcal{N}(0, \eta^2\mathbf{I}) \tag{1}$$

where $\mathbf{A}$ encodes the forward physics. For natural image restoration tasks, we assume $\mathbf{y} \in \mathbb{R}^m$, $\boldsymbol{x} \in \mathbb{R}^n$, and $\mathbf{A} \in \mathbb{R}^{m\times n}$. For MRI reconstruction, the same formulation applies in the complex domain.

Score-based diffusion models Song et al. (2021) define a forward noising process $\boldsymbol{x}_t = \boldsymbol{x}_0 + t\boldsymbol{z}$, $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$, that corrupts a clean image $\boldsymbol{x}_0 \sim p(\boldsymbol{x})$ at noise level $t \in [\sigma_{\min}, \sigma_{\max}]$. A denoiser $D_\phi(\boldsymbol{x}_t; t)$ is trained to predict $\boldsymbol{x}_0$ given $\boldsymbol{x}_t$, implicitly learning the score $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = (D_\phi(\boldsymbol{x}_t; t) - \boldsymbol{x}_t)/t^2$. Clean samples are generated by integrating the probability flow ODE $d\boldsymbol{x}/dt = (\boldsymbol{x}_t - D_\phi(\boldsymbol{x}_t; t))/t$ from $\sigma_{\max}$ to $\sigma_{\min}$ - the prior path.

Diffusion models have been widely used in inverse problems, where the above prior path is modified by additional DC updates using the measurements $\mathbf{y}$. Posterior algorithms leveraging diffusion models can be grouped into two categories:

1. Tweedie-based posterior methods: Algorithms in this category use the Tweedie estimate $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = D_\phi(\boldsymbol{x}_t; t)$ to map a noisy latent to an estimate of the clean image and then enforce DC as a separate step. DPS Chung et al. (2023) injects a DC gradient $\nabla_{\boldsymbol{x}_t} \log p(\mathbf{y}|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$ into the reverse SDE. ΠGDM Song et al. (2023) and DDS Chung et al. (2024) refine this approach with pseudoinverse corrections; DiffPIR Zhu et al. (2023) uses half-quadratic splitting Geman & Yang (1995); RED-Diff Mardani et al. (2024) formulates the

problem as the minimization of a variational objective. However, all of these methods rely on discrete-time implementations; moreover, the Tweedie approximation can become inaccurate at high noise levels, and none of these approaches guarantees that the resulting sample distribution matches the true posterior.
2. ODE-based methods: These methods offer a more principled route by decoupling denoising and DC into alternating steps, and offer convergence guarantees. For example, DAPS Zhang et al. (2025) uses an ODE-based denoising step to obtain a clean estimate, applies a Langevin DC update, and then re-noises the sample back to the prior trajectory. DPnP Xu & Chi (2024) alternates between proximal likelihood sampler and a full DDPM/DDIM prior update, yielding non-asymptotic convergence bounds. PnP-DM Wu et al. (2024) uses a split Gibbs sampling scheme with EDM-based denoising. A challenge with these methods is that each prior update requires an ODE/SDE solve, translating to higher computational complexity. Moreover, because these methods alternate between DC updates in the image domain and prior updates in the latent domain, their convergence analysis requires multi-kernel composition.

We note that single-scale EBMs define $p(\boldsymbol{x}) \propto \exp(-E_\theta(\boldsymbol{x}))$ via an explicit scalar energy. Because energies compose analytically, the posterior energy $-\log p(\boldsymbol{x}|\mathbf{y}) = E_\theta(\boldsymbol{x}) + \|\mathbf{A}\boldsymbol{x} - \mathbf{y}\|^2/(2\eta^2) + \text{const}$ is available in closed-form, making EBMs natural for posterior sampling and MAP estimation Hurault et al. (2022); Chand & Jacob (2024). However, single-scale EBMs are not competitive with large-scale diffusion models: training via contrastive divergence requires persistent Markov chains that suffer from slow mixing and mode collapse at high resolution Nijkamp et al. (2019); Du & Mordatch (2019), and single-scale models lack the multi-resolution structure needed to capture the full distribution of images.

3 ALPS: AnneaLed Posterior Sampling algorithm

In this section, we formulate time-dependent posteriors $\{p_t(\boldsymbol{x}|\mathbf{y})\}$ using multi-scale EBMs. We implement the pre-conditioned multi-scale MALA algorithm to sample from the posterior distributions $p_t(\boldsymbol{x}_t|\mathbf{y})$. Finally, we also propose an algorithm that can provide the local MAP estimation.

3.1 Multi-scale Energy Based Model

To solve the inverse problem in (1), we first define the prior parameterized by the multi-scale EBM at noise level t as Li et al. (2023):

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{Z_{\boldsymbol{\theta}}(t)} \exp\left(-\frac{E_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{t^2}\right), \tag{2}$$

where $Z_{\boldsymbol{\theta}}(t)$ is the partition function and $E_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ is the EBM defined as:

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{1}{2}\|\boldsymbol{x}_t - D_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\|^2, \tag{3}$$

where $D_{\boldsymbol{\theta}} : \mathbb{R}^{n+1} \to \mathbb{R}^n$ is a convolutional neural network (CNN). The gradient of the energy defines a denoiser $\tilde{D}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \boldsymbol{x}_t - \nabla_{\boldsymbol{x}_t} E_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, which expands via the chain rule as $\boldsymbol{x}_t - [(\boldsymbol{x}_t - D_{\boldsymbol{\theta}}) - J_D^\top(\boldsymbol{x}_t - D_{\boldsymbol{\theta}})]$, where J_D is the Jacobian of $D_{\boldsymbol{\theta}}$ computed via PyTorch's `autograd`.

The EBM parameters are learned by distillation from a pre-trained diffusion model. An EBM requires a conservative curl-free score, which standard diffusion training does not guarantee. Following Thornton et al. (2025), we regress the EBM-induced denoiser onto the teacher denoiser; because the EBM denoiser is conservative by construction, this yields the closest energy-compatible approximation to the teacher score. The distilled EBM inherits the teacher's generative quality while providing the scalar energy required for posterior composition. Training details and the EDM preconditioning scheme are given in Appendix D, where we also validate unconditional generation quality.

3.2 Explicit Multi-scale Posterior Distribution

Given the multi-scale EBM prior $p_\theta(\boldsymbol{x}_t, t) \propto \exp(-E_\theta(\boldsymbol{x}_t, t)/t^2)$, the negative log-prior composes analytically with the DC term to yield an explicit negative log-posterior distribution at every noise scale. We solve the

inverse problem (1) by sampling from the family of posterior distributions $p_t(\boldsymbol{x}_t|\mathbf{y}) \propto p_t(\mathbf{y}|\boldsymbol{x}_t)\, p_\theta(\boldsymbol{x}_t, t)$, where the negative log-posterior is defined as:

$$-\log p_t(\boldsymbol{x}_t|\mathbf{y}) = C_t(\boldsymbol{x}_t) = \underbrace{\frac{\|\mathbf{A}\boldsymbol{x}_t - \mathbf{y}\|^2}{2\eta^2}}_{-\log p_t(\mathbf{y}|\boldsymbol{x}_t)} + \underbrace{\frac{E_\theta(\boldsymbol{x}_t, t)}{t^2}}_{-\log p_\theta(\boldsymbol{x}_t, t)} \tag{4}$$

The negative log-likelihood converges to the true likelihood of (1) as $t \to \sigma_{\min} \approx 0$.

3.3 Preconditioned MALA

We generate samples from $p_t(\boldsymbol{x}_t|\mathbf{y})$ using preconditioned MALA, which consists of two steps: (a) a proposal drawn from the Unadjusted Langevin Algorithm (ULA), and (b) a Metropolis-Hastings (MH) accept/reject step. The ULA proposal at scale t is:

$$\tilde{\boldsymbol{x}}_{k+1,t} = \boldsymbol{x}_{k,t} - \frac{\tau^2}{2}\boldsymbol{B}_t \nabla_{\boldsymbol{x}_{k,t}} C_t(\boldsymbol{x}_{k,t}) + \tau\sqrt{\boldsymbol{B}_t}\epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, \mathbf{I}), \tag{5}$$

where $\boldsymbol{B}_t$ is the preconditioner:

$$\boldsymbol{B}_t = \left(\frac{\mathbf{A}^T\mathbf{A}}{\eta^2} + \frac{\mathbf{I}}{t^2}\right)^{-1} \tag{6}$$

We discuss the computation of noise covariance matrix for different imaging tasks in Appendix E.1. The preconditioner $\boldsymbol{B}_t$ approximates the inverse Hessian of C_t by replacing $\nabla^2 E_\theta/t^2$ with $\mathbf{I}/t^2$ (See Lemma C.4 in Appendix C.3 for the proof). We note that this approximation is asymptotically exact: at high noise t the denoiser Jacobian is small ($J_{\tilde{D}_\theta} \approx 0$), so $\nabla^2_{\boldsymbol{x}_t} \frac{E_\theta(\boldsymbol{x}_t, t)}{t^2} = \frac{\mathbf{I} - J_{\tilde{D}_\theta}}{t^2} \to \frac{\mathbf{I}}{t^2}$. Thus $\boldsymbol{B}_t$ captures the curvature of the DC term $\mathbf{A}^\top\mathbf{A}/\eta^2$ exactly while providing a good approximation to the prior curvature at large t, so the ill-conditioning of $\mathbf{A}$ does not affect mixing in this regime. To characterize the remaining approximation error, we define $\boldsymbol{\Delta}_t = \nabla^2_{\boldsymbol{x}_t} E_\theta(\boldsymbol{x}_t, t) - \mathbf{I}$ as the residual anisotropy of the prior Hessian. Then the preconditioned Hessian can be written as $\boldsymbol{B}_t^{1/2}\nabla^2 C_t \boldsymbol{B}_t^{1/2} = \mathbf{I} + 1/t^2(\boldsymbol{B}_t^{1/2}\boldsymbol{\Delta}_t\boldsymbol{B}_t^{1/2})$. At high t, $\boldsymbol{\Delta}_t \approx 0$ and hence the chain is well-conditioned and mixes rapidly (see Section 4.2). At low t, $\boldsymbol{\Delta}_t$ grows as the energy sharpens toward the data manifold, making the isotropic assumption less accurate and the preconditioner less effective. In this case, the annealing schedule and warm initialization enable faster convergence.

The update (5) is equivalent to solving a quadratic optimization problem followed by noise addition i.e.,

$$\tilde{\boldsymbol{x}}_{k+1,t} = \boldsymbol{x}_{k,t} + \frac{\tau^2}{2}(\hat{\boldsymbol{x}}_{k+1,t} - \boldsymbol{x}_{k,t}) + \tau\sqrt{\boldsymbol{B}_t}\epsilon_k, \tag{ULA}$$

where $\hat{\boldsymbol{x}}_{k+1,t}$ minimizes:

$$\hat{\boldsymbol{x}}_{k+1,t} = \arg\min_{\boldsymbol{x}} \frac{\|\mathbf{A}\boldsymbol{x} - \mathbf{y}\|^2}{2\eta^2} + \frac{\|\boldsymbol{x} - \boldsymbol{d}_{k,t}\|^2}{2t^2}, \tag{7}$$

with $\boldsymbol{d}_{k,t} = \boldsymbol{x}_{k,t} - \nabla_{\boldsymbol{x}_{k,t}} E_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,t}; t)$. For non-diagonal $\mathbf{A}$, (7) is solved via Conjugate Gradient (CG). The equivalence between (5) and (ULA) is shown in Appendix C.1.

The update in (ULA) introduces a step-size-dependent bias that vanishes only as $\tau \to 0$ (see Fig. 3.(b) under ablation studies of the results section). We note that ALPS has access to explicit posterior energy C_t, which enables the application of MH correction to accept or reject $\tilde{\boldsymbol{x}}_{k+1,t}$ based on the posterior ratio:

$$\alpha(\boldsymbol{x}_{k,t}, \tilde{\boldsymbol{x}}_{k+1,t}) = \min\left(1, \frac{\pi_t(\tilde{\boldsymbol{x}}_{k+1,t})\, q_t(\boldsymbol{x}_{k,t}|\tilde{\boldsymbol{x}}_{k+1,t})}{\pi_t(\boldsymbol{x}_{k,t})\, q_t(\tilde{\boldsymbol{x}}_{k+1,t}|\boldsymbol{x}_{k,t})}\right), \tag{8}$$

Here, $\pi_t(\cdot) = p_t(\cdot|\mathbf{y})$ and the proposal kernel is $q_t(\boldsymbol{x}'|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}'; \boldsymbol{\mu}_t(\boldsymbol{x}), \tau^2\boldsymbol{B}_t)$ with mean $\boldsymbol{\mu}_t(\boldsymbol{x}) = \boldsymbol{x} - \frac{\tau^2}{2}\boldsymbol{B}_t\nabla_{\boldsymbol{x}}C_t(\boldsymbol{x})$. The log acceptance ratio is (see Appendix C.2):

$$\log\alpha(\boldsymbol{x}_{k,t}, \tilde{\boldsymbol{x}}_{k+1,t}) = C_t(\boldsymbol{x}_{k,t}) - C_t(\tilde{\boldsymbol{x}}_{k+1,t}) - \frac{1}{2\tau^2}\left[\|\boldsymbol{x}_{k,t} - \boldsymbol{\mu}_t(\tilde{\boldsymbol{x}}_{k+1,t})\|^2_{\boldsymbol{B}_t^{-1}} - \|\tilde{\boldsymbol{x}}_{k+1,t} - \boldsymbol{\mu}_t(\boldsymbol{x}_{k,t})\|^2_{\boldsymbol{B}_t^{-1}}\right], \tag{9}$$

Algorithm 1 ALPS

Input: Data $\mathbf{y}$, forward operator $\mathbf{A}$, noise level η, schedule t, iterations K at each scale, score model $\nabla_{\boldsymbol{x}} E_{\boldsymbol{\theta}}(\boldsymbol{x}; t)$
Initialization: $\boldsymbol{x}_{0,t_0} \sim p_{t_0}(\boldsymbol{x}|\mathbf{y})$ from (11)
for $i = 0$ to N **do**
  for $k = 0$ to K **do**
    Denoising: $\boldsymbol{d}_{k,t_i} = \boldsymbol{x}_{k,t_i} - \nabla E_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,t_i}, t_i)$
    DC: $\hat{\boldsymbol{x}}_{k+1,t_i}$ as in (7)
    Noise addition: $\tilde{\boldsymbol{x}}_{k+1,t_i} = \boldsymbol{x}_{k,t_i} + \frac{\tau^2}{2}(\hat{\boldsymbol{x}}_{k+1,t_i} - \boldsymbol{x}_{k,t_i}) + \tau\sqrt{\boldsymbol{B}_{t_i}}\epsilon_k$
    MH accept/reject: draw $u \sim \text{Uniform}[0,1]$; if $u \leq \alpha(\boldsymbol{x}_{k,t_i}, \tilde{\boldsymbol{x}}_{k+1,t_i})$: $\boldsymbol{x}_{k+1,t_i} = \tilde{\boldsymbol{x}}_{k+1,t_i}$; else $\boldsymbol{x}_{k+1,t_i} = \boldsymbol{x}_{k,t_i}$
  end for
  Initialize next scale: $\boldsymbol{x}_{0,t_{i+1}} \leftarrow \tilde{\boldsymbol{x}}_{K,t_i}$
end for
Return $\boldsymbol{x}_{K,N}$

where $\|\boldsymbol{v}\|^2_{\boldsymbol{B}_t^{-1}} = \boldsymbol{v}^\top \boldsymbol{B}_t^{-1}\boldsymbol{v}$. The sample is accepted if $\log u \leq \log \alpha$ with $u \sim \text{Uniform}[0,1]$.

We note that at the highest scale, the negative log-posterior can be approximated as:

$$C_{\sigma_{max}} = \frac{\|\mathbf{A}\boldsymbol{x} - \mathbf{y}\|^2}{2\eta^2} + \frac{\|\boldsymbol{x}\|^2}{2t_{max}^2} \tag{10}$$

Then the samples from this distribution can be obtained by sampling from:

$$p_{\sigma_{max}}(\boldsymbol{x}|\mathbf{y}) = \mathcal{N}\left(\left(\frac{\mathbf{A}^T\mathbf{A}}{\eta^2} + \frac{\mathbf{I}}{t_{max}^2}\right)^{-1}\frac{\mathbf{A}^T\mathbf{y}}{\eta^2}, \left(\frac{\mathbf{A}^T\mathbf{A}}{\eta^2} + \frac{\mathbf{I}}{t_{max}^2}\right)^{-1}\right) \tag{11}$$

The initialization (11) is the exact posterior at $\sigma_{\max}$ under a Gaussian prior, incorporating the measurements $\mathbf{y}$ via the pseudo-inverse and placing the chain close to the posterior from the outset.

3.4 Local MAP estimation

The explicit posterior energy $C_t(\boldsymbol{x}_t)$ also enables local MAP estimation. We propose to sequentially minimize $C_t(\boldsymbol{x}_t)$ using the MM approach, which guarantees monotonic decrease of $C_t(\boldsymbol{x}_t)$ at each scale t and convergence to a stationary point of $C_t(\boldsymbol{x}_t)$.

Suppose $\nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x}, t)$ is L_t-Lipschitz, i.e., $\|\nabla^2_{\boldsymbol{x}} E_\theta(\boldsymbol{x}, t)\|_2 \leq L_t$ for all $\boldsymbol{x}$ (where L_t is the maximum eigenvalue of the Hessian of E_θ at scale t). By the descent lemma, this gives:

$$E_\theta(\boldsymbol{x}_t, t) \leq E_\theta(\boldsymbol{x}_{k,t}, t) + \nabla E_\theta(\boldsymbol{x}_{k,t}, t)^\top(\boldsymbol{x}_t - \boldsymbol{x}_{k,t}) + \frac{L_t}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{k,t}\|^2, \tag{12}$$

which provide a quadratic surrogate $Q(\boldsymbol{x}_t|\boldsymbol{x}_{k,t})$ for $C_t(\boldsymbol{x}_t)$:

$$Q(\boldsymbol{x}_t|\boldsymbol{x}_{k,t}) = \frac{\|\mathbf{A}\boldsymbol{x}_t - \mathbf{y}\|^2}{2\eta^2} + \frac{1}{t^2}\Big[\nabla E_\theta(\boldsymbol{x}_{k,t}, t)^\top \boldsymbol{x}_t + \frac{L_t}{2}\|\boldsymbol{x}_t - \boldsymbol{x}_{k,t}\|^2\Big]. \tag{13}$$

Minimizing $Q(\boldsymbol{x}_t|\boldsymbol{x}_{k,t})$ over $\boldsymbol{x}_t$ gives the MM update:

$$\boldsymbol{x}_{k+1,t} = \left(\frac{\mathbf{A}^H\mathbf{A}}{\eta^2} + \frac{L_t}{t^2}\mathbf{I}\right)^{-1}\left(\frac{\mathbf{A}^H\mathbf{y}}{\eta^2} + \frac{L_t\boldsymbol{x}_{k,t} - \nabla_{\boldsymbol{x}_{k,t}} E_\theta(\boldsymbol{x}_{k,t}, t)}{t^2}\right). \tag{14}$$

Lemma C.4 in Appendix C.3 shows that $\lambda_{\max}\big(\nabla^2_{\boldsymbol{x}_t} E_\theta(\boldsymbol{x}_t, t)/t^2\big) \leq 1/t^2$. Nonetheless, to account for imperfect training we use the conservative bound L_t/t^2 with $L_t \geq 1$. When $L_t = 1$, (14) reduces to (7). Overestimating L_t is conservative, but correct: it produces smaller steps without affecting the monotone descent guarantee. The MAP algorithm is Algorithm 1 with the noise addition and MH update omitted. Convergence is established in Proposition 4.1 in the next section.

4 Convergence Results

We establish convergence of ALPS at three levels of increasing precision. First, at each fixed noise scale t, the preconditioned MALA kernel converges geometrically to the annealed posterior π_t (Theorem 4.1). Second, as the scale is annealed to zero, the sequence of intermediate posteriors π_t converges to the true posterior π (Theorem 4.2). Third, we provide a non-asymptotic bound that provides insights on the convergence rate and how implementation errors (e.g from imperfect score/energy and finite CG solves) accumulate (Theorem 4.3).

4.1 Asymptotic convergence

Throughout this section we assume:

(A1) *Bounded Hessian.* For each $t > 0$, the negative log-prior has bounded curvature: $L_t = \sup_{\boldsymbol{x}} \lambda_{\max}(\nabla^2_{\boldsymbol{x}} E_\theta(\boldsymbol{x}; t)/t^2) < \infty$. When ∇E_θ is the exact score, Lemma C.4 implies $L_t \leq 1/t^2$.

(A2) *Bounded forward operator and positive-definite preconditioner.* The forward operator $\mathbf{A}$ is bounded ($\|\mathbf{A}\| < \infty$) and the measurement noise level $\eta > 0$. Consequently, the preconditioner $\boldsymbol{B}_t = \left(\frac{\mathbf{A}^\top \mathbf{A}}{\eta^2} + \frac{\mathbf{I}}{t^2}\right)^{-1}$ is symmetric positive-definite for all $t > 0$. Note that even when $\mathbf{A}$ is rank-deficient (e.g., the masking operator in inpainting), the term $\frac{\mathbf{I}}{t^2}$ ensures all eigenvalues of $\boldsymbol{B}_t^{-1}$ are at least $1/t^2 > 0$, so $\boldsymbol{B}_t \succ 0$ for every $t > 0$.

(A3) *Distillation quality / annealing bias.* The distilled energy approximates the true negative log-prior accurately enough such that the posterior $\pi_t = p_t(\cdot|\mathbf{y}) \propto \exp(-C_t(\boldsymbol{x}))$ converges weakly to the true posterior $\pi = p(\cdot|\mathbf{y})$ as $t \to 0$. A sufficient condition is that the distillation error $\sup_{\boldsymbol{x}} \|\nabla E_\theta(\boldsymbol{x}; t) + t^2 \nabla \log p_t(\boldsymbol{x})\|$ is small and vanishes as $t \to 0$, which we assume holds for the trained model.

(A4) *Tail coercivity of the posterior energy.* There exist $\upsilon > 0$ and $W > 0$ such that posterior modes at all t lie inside $\{\|\boldsymbol{x}\| \leq W/2\}$ and $\nabla^2_{\boldsymbol{x}} C_t(\boldsymbol{x}) \succeq 16\upsilon \mathbf{I}$ for all $\|\boldsymbol{x}\| \geq W$. The first condition is satisfied as training images are normalized, while the latter is a mild technical condition that can be enforced by construction, if needed; see Appendix C.5.

We measure convergence using the V-norm for a finite signed measure μ:

$$\|\mu\|_V = \sup_{\substack{f:\mathbb{R}^d \to \mathbb{R} \\ |f(\boldsymbol{x})| \leq V(\boldsymbol{x})}} \int_{\mathbb{R}^d} f(\boldsymbol{x})\, d\mu(\boldsymbol{x}). \tag{15}$$

We choose $V : \mathbb{R}^d \to [1, \infty)$ as the Lyapunov weighting function:

$$V_\kappa(\boldsymbol{x}) = \exp\big(\kappa \|\boldsymbol{x}\|^2\big), \tag{16}$$

The choice $V_\kappa(\boldsymbol{x}) = \exp(\kappa\|\boldsymbol{x}\|^2)$ penalizes tail deviations more heavily than standard total variation (TV), giving a tighter bound on mixing and annealing bias. We note that $V_\kappa(\boldsymbol{x}) = 1$ results in the TV norm.

The convergence of ALPS, which applies MALA at each scale, follows from two key properties.

1. **Fixed-scale convergence.** At each noise level t, the iterates converge geometrically to the intermediate posterior $\pi_t = p_t(\boldsymbol{x}_t \mid \mathbf{y})$.

   Theorem 4.1 *Under assumptions (A1), (A2), and (A4), for each scale $t > 0$ there exist constants $C_{\tau,t} \geq 0$ and $\rho_{\tau,t} \in (0, 1)$, and a threshold $\bar{\Gamma}_t > 0$, such that for any step size $\tau \in (0, \bar{\Gamma}_t]$ and all $\boldsymbol{x} \in \mathbb{R}^d$:*

   $$\|\delta_{\boldsymbol{x}} R^K_{\tau,t} - \pi_t\|_{V_\kappa} \leq C_{\tau,t}\, \rho^K_{\tau,t}\, V_\kappa(\boldsymbol{x}), \tag{17}$$

   where $R_{\tau,t}$ is the MALA kernel associated with the preconditioned proposal at scale t, i.e., $q_t(\boldsymbol{x}' \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}'; \mu_t(\boldsymbol{x}), \tau^2 \boldsymbol{B}_t)$, $V_{\bar{\eta}}(\boldsymbol{x}) = \exp(\bar{\eta}\|\boldsymbol{x}\|^2)$ with $\bar{\eta} > 0$, and $\delta_{\boldsymbol{x}}$ denotes the Dirac measure at $\boldsymbol{x}$. The notation $\delta_{\boldsymbol{x}} R^K_{\tau,t}$ denotes the distribution of the chain after K steps of $R_{\tau,t}$, with initialization as $\boldsymbol{x}$.

Proof 4.1 *See Appendix C.3.*

We note that the proof relies on single-kernel analysis when compared with prior-path methods such as DPnP Xu & Chi (2024) and PnP-DM Wu et al. (2024) which requires analyzing the composition of two kernels per scale - one for the denoising step and one for the likelihood step. Because ALPS targets a fixed posterior distribution at each scale with a single MALA kernel, the standard geometric ergodicity result applies directly.

2. **Annealed convergence.** To overcome slow mixing at $t \to 0$, ALPS anneals through a decreasing sequence $t_0 > t_1 > \cdots > t_N \to 0$, targeting the true posterior $\pi = p(\boldsymbol{x} \mid \mathbf{y})$ in the limit. The following result establishes convergence of this annealed chain to π.

Theorem 4.2 *Let $p_{i,k} = p_{i,0} R^K_{\tau,t_i}$ denote the distribution of the iterates after K MALA steps at scale t_i, where $p_{i,0}$ is the initialization distribution at scale t_i. Then we have:*

$$\|p_{i,K} - \pi\|_{V_\kappa} \le C_{\tau,t_i} \rho^K_{\tau,t_i} \, \mathbb{E}_{p_{i,0}}[V_\kappa(\boldsymbol{X})] + \|\pi_{t_i} - \pi\|_{V_\kappa}. \tag{18}$$

The first term is the finite-iteration sampling error at scale t_i, which decays geometrically in K. The second term is the annealing bias *measuring how far the intermediate target π_{t_i} is from the final posterior π; under assumption (A3), $\|\pi_{t_i} - \pi\|_{V_\kappa} \to 0$ as $t_i \to 0$.*

Proof 4.2 *See Appendix C.4 for the proof.*

Theorem 4.2 decomposes the total error into two reducible parts. The first part is the sampling error that measures the deviation of the samples from π_{t_i}, which decays geometrically in K as shown in Theorem 4.1. The second term is the annealing bias, which vanishes as $t_i \to 0$. For any $\varepsilon > 0$, a sufficiently fine schedule and sufficiently large K gives $\|p_{i,K} - \pi\|_{V_\kappa} < \varepsilon$, thereby justifying both the multi-scale design and the annealing to zero. We note that annealing is necessary for mixing. At small t, mixing is slow, but the chain is already localized near the posterior modes from the previous scale. Warm initialization keeps $\mathbb{E}_{p_{i,0}}[V_\kappa]$ small, so the sampling error remains controlled at every scale. A cold start at $t = 0$ would face slow mixing and a large V_κ-moment simultaneously.

4.2 Non-asymptotic convergence rate and robustness

Theorems 4.1 and 4.2 establish asymptotic convergence as $K \to \infty$ and $t \to 0$. The following result complements them with a finite-K bound in TV norm $\mathrm{TV}(p, \pi_t) = \|p - \pi_t\|_{V \equiv 1}$ and characterizes how approximation errors from imperfect scores and finite CG solves accumulate. The proof follows the non-asymptotic analysis of Xu & Chi (2024), adapted to the single-kernel setting with the preconditioned MALA kernel $R_{\tau,t}$.

Spectral gap and its dependence on noise scale. The key quantity governing the mixing rate is the *spectral gap* $1 - \lambda_t$ of the ideal MALA kernel $R_{\tau,t}$, defined as the gap between its largest eigenvalue (unity, by MH reversibility) and its second-largest eigenvalue in $L^2(\pi_t)$. As discussed in Sec. 3.3, at high t the chain is well-conditioned, so the spectral gap is large and mixing is rapid. At low t, the preconditioner becomes less effective, the spectral gap shrinks, and mixing slows down. We use the spectral gap to bound the TV error, where we assume

$$\varepsilon_t := \sup_{\boldsymbol{x}} \mathrm{TV}\big(\tilde{R}_{\tau,t}(\boldsymbol{x}, \cdot),\, R_{\tau,t}(\boldsymbol{x}, \cdot)\big), \tag{19}$$

to be the per-step implementation error, which captures score approximation error and finite CG precision.

Theorem 4.3 (Non-asymptotic robustness) *Under assumptions* (A1)–(A4)*, let p_K denote the law after K steps of the implemented kernel $\tilde{R}_{\tau,t}$ that approximates $R_{\tau,t}$, and let $\chi^2(p_1 \| \pi_t) = \int (dp_1/d\pi_t - 1)^2 \, d\pi_t$ denote the χ^2-divergence with respect to π_t; in particular, $\chi^2(p_1 \| \pi_t)$ measures the post-initialization mismatch with the posterior π_t. Then for any $K \ge 1$,*

$$\mathrm{TV}(p_K,\, \pi_t) \;\le\; \lambda_t^{K-1} \sqrt{\tfrac{1}{2} \chi^2(p_1 \| \pi_t)} \;+\; (K-1)\, \varepsilon_t. \tag{20}$$

The first term decays geometrically in K; the second accumulates linearly from per-step implementation errors (19).

Proof 4.3 *The proof is given in Appendix C.5.*

Implications. The reversibility of $R_{\tau,t}$ (Lemma C.2) gives χ^2-contractivity in terms of the second-largest eigenvalue λ_t from the spectral expansion Levin & Peres (2017), which bounds TV via TV $\leq \sqrt{\chi^2/2}$. The bound (20) has two terms. The first is the mixing term, controlled by initialization quality and the spectral gap $1-\lambda_t$: to reduce it below ε requires $K \sim \frac{1}{1-\lambda_t} \log \frac{1}{\varepsilon}$ steps. The second is the accumulated implementation error which grows linearly as $(K-1)\epsilon$. Evaluating at the iteration budget required for mixing, this term is of order $\frac{\varepsilon_t}{1-\lambda_t} \log \frac{1}{\varepsilon}$. At high t, the spectral gap is large ($\lambda_t \ll 1$), so both terms are small: the mixing term contracts in few steps and the accumulated error remains low. At low t, the spectral gap shrinks ($\lambda_t \to 1$), inflating both the iteration budget and the accumulated error. To overcome the slow mixing, ALPS relies on warm initialization that keeps initialization mismatch with π_t small even when mixing is slow.

4.3 Convergence of local MAP estimation

Proposition 4.1 (Monotonic convergence of MM) *The iterates $\{\boldsymbol{x}_{k,t}\}$ produced by (14) satisfy:*

$$C_t(\boldsymbol{x}_{k+1,t}) \leq C_t(\boldsymbol{x}_{k,t}) - \frac{1}{2}\left(\frac{L_t}{t^2} + \lambda_{\min}\left(\frac{\mathbf{A}^H\mathbf{A}}{\eta^2}\right)\right)\|\boldsymbol{x}_{k+1,t} - \boldsymbol{x}_{k,t}\|^2, \tag{21}$$

where $\lambda_{\min}(\mathbf{A}^H\mathbf{A}/\eta^2) \geq 0$. In particular, $C_t(\boldsymbol{x}_{k+1,t}) \leq C_t(\boldsymbol{x}_{k,t})$ at every iteration, and the sequence converges to a stationary point of C_t - a local MAP estimate of the annealed posterior.

Proof 4.4 *Since $Q(\boldsymbol{x}_t|\boldsymbol{x}_{k,t})$ majorizes $C_t(\boldsymbol{x}_t)$ by (12), and $\boldsymbol{x}_{k+1,t}$ minimizes Q, we have $C_t(\boldsymbol{x}_{k+1,t}) \leq Q(\boldsymbol{x}_{k+1,t}|\boldsymbol{x}_{k,t}) \leq Q(\boldsymbol{x}_{k,t}|\boldsymbol{x}_{k,t}) = C_t(\boldsymbol{x}_{k,t})$. The energy gap follows from the strong convexity of Q in $\boldsymbol{x}$ with modulus $L_t/t^2 + \lambda_{\min}(\mathbf{A}^H\mathbf{A}/\eta^2)$, which bounds the decrease from below. Since $C_t(\boldsymbol{x}_t)$ is bounded below by a finite value and the surrogate function satisfies: (a) $Q(\boldsymbol{x}_{k,t}|\boldsymbol{x}_{k,t}) = C_t(\boldsymbol{x}_{k,t})$ and, (b) $Q(\boldsymbol{x}_t|\boldsymbol{x}_{k,t}) \geq C_t(\boldsymbol{x}_t)$, the descent at every step ensures convergence to a point where $\nabla C_t = 0$, i.e., a stationary point of C_t Razaviyayn et al. (2013).*

5 Results

5.1 Experimental Setup

We evaluate ALPS on five inverse problems: mask inpainting, random inpainting, Gaussian deblurring, motion deblurring, and MRI reconstruction. The first four are evaluated on the FFHQ face images (64×64); MRI reconstruction is evaluated on the brain slices from the fastMRI dataset (320×320). We add zero-mean Gaussian measurement noise with $\eta = 0.05$ for natural image tasks, while the inherently noisy k-space samples are used in MRI. Each quantitative result is averaged over ten test images.

Metrics. For sampling algorithms, we report (a) the sample avg. peak signal-to-noise ratio (PSNR), averaged over five generated samples, (b) MMSE PSNR i.e., PSNR of the posterior mean, and (c) SSIM of the posterior mean. For point-estimate methods (ALPS-MAP, PnP-ISTA) we report MAP/fixed-point PSNR and SSIM.

Baselines. We compare three ALPS variants: *ALPS-ULA* (Algorithm 1 without MH correction), *ALPS-MALA* (with MH correction), and *ALPS-MAP* (deterministic MAP via (14)) - against the diffusion-based samplers DPS (Chung et al., 2023), DAPS (Zhang et al., 2025), and PnP-ULA (Laumont et al., 2022), and the fixed-point method PnP-ISTA (Ryu et al., 2019). Hyperparameters and implementation details are in the Appendix.

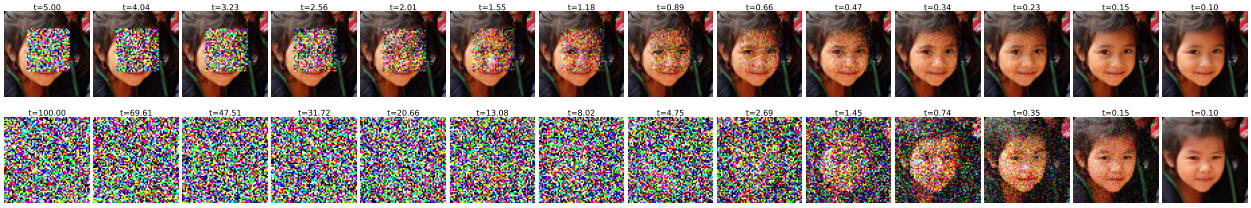


Figure 2: Sampling trajectories for mask inpainting inverse problem. *Top:* ALPS evolves from a pseudo-inverse initialization, adding noise predominantly in the null-space of $\mathbf{A}$. *Bottom:* We choose a representative posterior approach (DAPS), which evolves from Gaussian noise, remaining in the diffusion prior path throughout. The geometry-aware initialization of ALPS produces a shorter, more direct path to the posterior, explaining its efficiency advantage.

5.2 Sampling Trajectory Analysis

In this section, we compare the sampling trajectory of ALPS with that of DAPS. Fig. 2 illustrates the evolution of samples for the mask inpainting inverse problem. As shown in the figure, DAPS begins from a Gaussian noise initialization, and the guidance term is carefully engineered such that the samples stay in p_t. Moreover, the samples become data-consistent only as $t \to 0$. In contrast, ALPS adopts a fundamentally different strategy. It uses an initialization based on the pseudo-inverse solution, which provides a data-consistent starting point. Noise is then injected predominantly in the null-space of the forward operator. This aligns the trajectory with the problem geometry, yielding a shorter, more direct path to the posterior and reducing both the NFE and the DC error at convergence. The conceptual differences are illustrated in Fig. 1, and a 2D comparison on the `moons` dataset is given in the Appendix.

5.3 Ablation studies

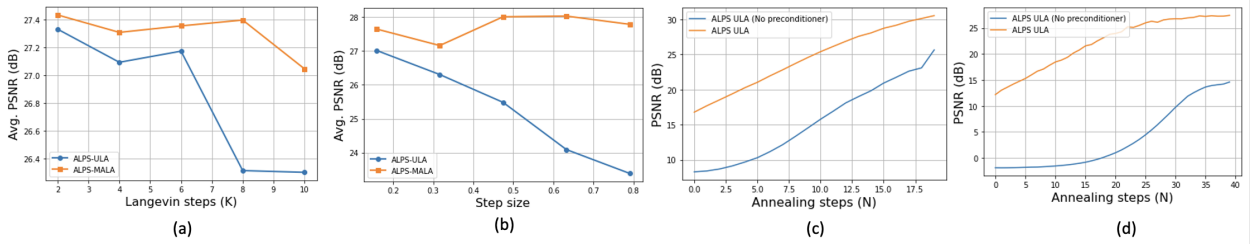


Figure 3: Performance comparison of ALPS-ULA and ALPS-MALA for image inpainting inverse problem as a function of (a) Langevin steps and (b) step-size; (c) and (d) shows the convergence plot of ALPS-ULA with and without the preconditioner for Gaussian and image inpainting tasks, respectively.

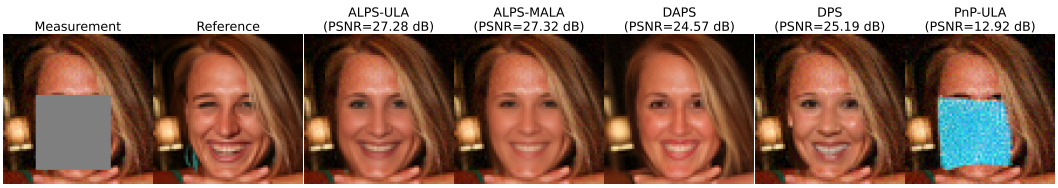
In this section, we study the effect of Langevin steps, step-size, and preconditioner on the performance of the ALPS algorithm. Fig. 3.(a) compares the Avg. PSNR (computed over two test images) of ALPS-ULA and ALPS-MALA as a function of Langevin steps, with the annealing steps fixed at $N = 40$ for the image inpainting inverse problem. From the figure, it can be observed that as the number of Langevin steps increases, the performance of ALPS-ULA drops by about 1 dB compared to that of ALPS-MALA. This is a direct consequence of error accumulation, which is corrected by the MH step in ALPS-MALA.

The effect of MH correction is more striking in Fig.3.(b), which shows the Avg. PSNR of ALPS-ULA and ALPS-MALA algorithm for different step-sizes. Compared to ALPS-MALA, the performance of ALPS-ULA drops by about 3 dB as the step size increases. This shows that one needs to carefully tune the step-size of ALPS-ULA to achieve optimum performance. Nonetheless, in the case of ALPS-MALA, the MH correction rejects proposals with low posterior probability, making the algorithm robust to step sizes.

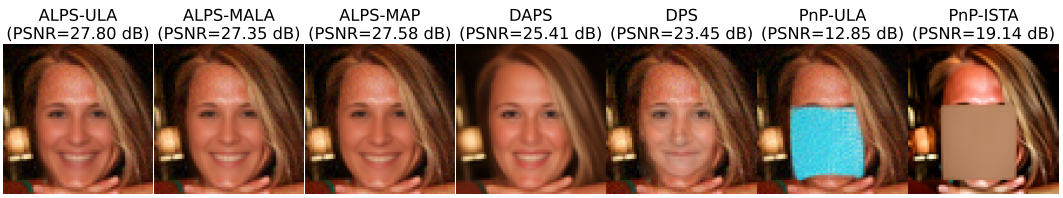
Figs. 3.(c) and (d) compare the PSNR evolution of ALPS-ULA, with and without the preconditioner, for the Gaussian deblurring and image inpainting tasks, respectively, with $K = 25$. We observe that with preconditioner, the convergence is faster for both the inverse problems. We also observe that the performance gap is larger (by about 10 dB) in the severely ill-posed problem i.e., image painting when compared to a Gaussian deblurring. This is a direct consequence of the preconditioner, which mitigates the poor conditioning induced by the forward model and improves performance.

5.4 Reconstruction Quality

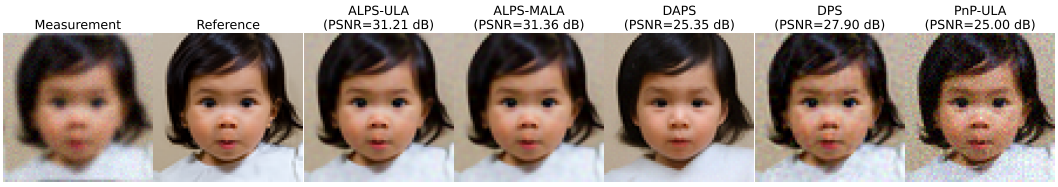
Quality of posterior samples: Tables 1 and 2 report the quantitative results on the face and MRI inverse problems, respectively. Figs. 4, 5, and 6 shows sample and MMSE/MAP comparisons for mask inpainting, Gaussian deblurring, and FLAIR-weighted MRI brain recovery. Figs. 11 - 14 in Appendix B.1 shows the visual comparisons for the remaining tasks. Top row of each task shows the sample that yields the best PSNR for each method, while the bottom row shows MMSE and MAP estimates. The MMSE estimate is obtained by computing the average of five posterior samples. From Tables 1 and 2, it can be observed that all three ALPS variants have improved performance when compared to the baseline algorithms. We observe that the gain is largest on the highly ill-posed problem. For example, in the case of 8x FLAIR-weighted MRI image recovery, ALPS-MALA achieves about 4.52 dB and 6.95 dB higher PSNR when compared to DAPS



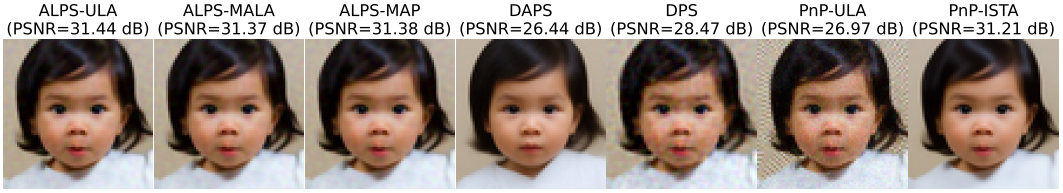
(a) Mask inpainting - samples



(b) Mask inpainting - MMSE and MAP



(c) Gaussian deblurring ($\sigma = 1$) - samples



(d) Gaussian deblurring ($\sigma = 1$) - MMSE and MAP

Figure 4: FFHQ (64×64) reconstruction results: (a,b) mask inpainting and (c,d) Gaussian deblurring. For each task, the top row shows the samples (highest PSNR per method shown); the bottom row shows MMSE and MAP estimates. Results for random inpainting and motion deblurring are shown in Figs. 11 and 12 in Appendix B.1.

Table 1: Reconstruction performance of different algorithms on FFHQ dataset.

Task	Algorithm	Sample Avg. PSNR (dB)	MMSE/MAP/FP PSNR (dB)	SSIM
Inpaint (Random)	ALPS-ULA (ours)	32.29± 1.21	32.58±1.15	0.94±0.01
$p = 0.3$	ALPS-MALA (ours)	32.51± 1.15	32.57±1.15	0.94±0.03
	ALPS-MAP (ours)	-	32.64±1.19	0.94±0.01
	DAPS	27.02 ± 1.11	28.76±1.21	0.89±0.02
	DPS	29.76 ± 1.46	30.84± 1.42	0.92 ±0.01
	PnP-ULA	26.01 ± 1.8	29.31 ± 2.4	0.90 ± 0.01
	PnP-ISTA	-	31.13 ± 2.82	0.95 ± 0.01
Inpaint (Random)	ALPS-ULA (ours)	27.23 ± 2.05	27.81±2.1	0.88±0.02
$p = 0.7$	ALPS-MALA (ours)	27.58 ± 2.09	27.67±2.1	0.88±0.02
	ALPS-MAP (ours)	-	28.13±2.4	0.89±0.03
	DAPS	23.31 ± 1.34	24.66±1.44	0.78±0.03
	DPS	21.55 ± 1.74	23.52± 1.85	0.76 ±0.05
	PnP-ULA	18.18 ±2.89	21.29 ± 3.4	0.69 ± 0.09
	PnP-ISTA	-	25.91 ± 5.9	0.85 ± 0.14
Inpaint (Box)	ALPS-ULA (ours)	26.08 ± 2.43	26.71±2.43	0.85±0.03
Mask size = $(32, 33)$	ALPS-MALA (ours)	25.75 ± 3.03	25.81±3.03	0.83±0.04
	ALPS-MAP (ours)	-	26.01 ±2.71	0.83±0.03
	DAPS	23.64 ± 2.19	25.28±2.46	0.81±0.05
	DPS	21.11 ± 2.29	23.23± 2.83	0.74 ±0.06
	PnP-ULA	15.44 ± 1.11	17.54 ± 1.74	0.63 ± 0.02
	PnP-ISTA	-	18.19 ± 1.53	0.55 ± 0.05
Inpaint (Box)	ALPS-ULA (ours)	29.19 ± 2.53	29.78±2.45	0.91±0.02
Mask size = $(22, 25)$	ALPS-MALA (ours)	29.32 ± 2.75	29.38±2.75	0.91±0.02
	ALPS-MAP (ours)	-	29.39 ±2.69	0.91±0.02
	DAPS	25.97 ± 1.82	27.78±2.03	0.87±0.02
	DPS	24.79 ± 1.54	26.31± 1.72	0.84 ±0.02
	PnP-ULA	18.83 ± 3.15	20.41 ± 4.07	0.79 ± 0.03
	PnP-ISTA	-	19.57 ± 1.65	0.68 ± 0.05
Gaussian Deblurring	ALPS-ULA (ours)	30.68 ± 1.34	31.01 ± 1.35	0.94 ± 0.01
Sigma = 1	ALPS-MALA (ours)	30.96 ± 1.39	31.01 ± 1.39	0.94 ± 0.01
	ALPS-MAP (ours)	-	30.99 ± 1.39	0.94 ± 0.01
	DAPS	25.14 ± 1.32	26.54 ± 1.52	0.85 ±0.03
	DPS	27.53 ± 1.11	28.44 ±1.15	0.88 ± 0.01
	PnP-ULA	23.36 ± 1.87	27.72 ± 1.71	0.87 ± 0.02
	PnP-ISTA	-	30.43 ± 1.71	0.94 ±0.02
Gaussian Deblurring	ALPS-ULA (ours)	26.64 ± 1.69	27.03 ± 1.75	0.88 ± 0.02
Sigma = 2	ALPS-MALA (ours)	26.91 ± 1.75	26.97 ± 1.77	0.88 ± 0.02
	ALPS-MAP (ours)	-	26.98 ± 1.79	0.88 ±0.02
	DAPS	22.29 ± 1.34	23.45 ± 1.47	0.74 ± 0.05
	DPS	22.35 ± 1.23	24.06 ± 1.41	0.79 ± 0.03
	PnP-ULA	20.85 ±1.78	24.65 ± 1.71	0.78 ± 0.03
	PnP-ISTA	-	25.46 ± 1.71	0.84 ± 0.03
Motion Deblurring	ALPS-ULA (ours)	29.78 ± 1.18	30.15 ±1.18	0.91 ± 0.01
Length = 6, Deg = 30	ALPS-MALA (ours)	30.04 ±1.21	30.01 ± 1.21	0.91 ± 0.01
	ALPS-MAP (ours)	-	30.01 ± 1.21	0.91 ± 0.01
	DAPS	24.47 ± 1.52	25.88 ± 1.73	0.83 ± 0.04
	DPS	25.29 ± 1.29	26.31 ±1.32	0.83 ± 0.02
	PnP-ULA	24.38 ± 0.92	28.14 ± 0.90	0.87 ± 0.02
	PnP-ISTA	-	29.75 ± 1.61	0.92 ± 0.02
Motion Deblurring	ALPS-ULA (ours)	27.26 ± 1.45	27.71 ± 1.47	0.87 ± 0.02
Length = 15, Deg = 30	ALPS-MALA (ours)	27.53 ± 1.45	27.59 ± 1.45	0.87 ± 0.02
	ALPS-MAP (ours)	-	27.61 ± 1.48	0.87 ± 0.02
	DAPS	21.31 ±1.54	22.49 ± 1.77	0.69 ± 0.06
	DPS	20.51 ± 1.09	22.12 ± 1.38	0.67 ± 0.04
	PnP-ULA	21.81 ± 1.61	25.18 ± 1.38	0.78 ± 0.03
	PnP-ISTA	-	25.92 ± 1.61	0.83 ± 0.04

Table 2: Reconstruction performance of different algorithms on MRI dataset.

Task	Algorithm	Sample Avg. PSNR (dB)	MMSE/MAP/FP PSNR (dB)	SSIM
MRI (T2-weighted) Acc = 4x 1D sampling	ALPS-ULA (ours)	36.44 ± 0.61	36.75±0.62	0.95±0.01
	ALPS-MALA (ours)	36.71 ± 0.55	36.71±0.55	0.94±0.01
	ALPS-MAP (ours)	-	35.57 ±1.78	0.95±0.01
	DAPS	31.24 ± 0.66	32.75±0.71	0.89±0.02
	DPS	33.06 ± 1.03	34.92± 0.99	0.93 ±0.01
	PnP-ULA	26.08 ± 0.59	28.99 ± 0.72	0.78 ± 0.03
	PnP-ISTA	-	26.53± 0.78	0.71 ± 0.06
MRI (T2-weighted) Acc = 8x 1D sampling	ALPS-ULA (ours)	31.52 ± 0.45	32.12±0.47	0.91±0.01
	ALPS-MALA (ours)	31.68 ± 0.66	31.77±0.63	0.91±0.01
	ALPS-MAP (ours)	-	28.42 ±1.9	0.88±0.01
	DAPS	28.87 ± 0.66	29.98±0.71	0.85±0.02
	DPS	24.97 ± 1.17	27.35± 1.27	0.81 ±0.04
	PnP-ULA	24.66 ± 0.55	26.99 ± 0.67	0.74 ± 0.04
	PnP-ISTA	-	25.04± 0.97	0.66 ± 0.07
MRI (FLAIR-weighted) Acc = 4x 1D sampling	ALPS-ULA (ours)	37.21 ± 0.76	37.62±0.79	0.96±0.01
	ALPS-MALA (ours)	37.63 ± 0.71	37.63±0.71	0.96±0.01
	ALPS-MAP (ours)	-	37.35 ±2.01	0.96±0.01
	DAPS	31.64 ± 0.33	33.21 ±0.38	0.91±0.01
	DPS	32.89 ± 0.67	34.68± 0.71	0.93 ±0.01
	PnP-ULA	25.95 ± 0.53	29.12 ± 0.54	0.81 ± 0.02
	PnP-ISTA	-	26.27± 1.48	0.72 ± 0.03
MRI (FLAIR-weighted) Acc = 8x 1D sampling	ALPS-ULA (ours)	32.85 ± 0.73	33.35±0.81	0.92±0.01
	ALPS-MALA (ours)	32.82 ± 0.84	32.86±0.82	0.92±0.01
	ALPS-MAP (ours)	-	32.31 ±2.34	0.92±0.01
	DAPS	28.33 ± 1.27	29.81 ±0.98	0.86±0.02
	DPS	25.87 ± 1.63	28.25± 1.77	0.82 ±0.06
	PnP-ULA	24.57 ± 0.53	27.44± 0.52	0.76 ± 0.03
	PnP-ISTA	-	24.98± 1.92	0.69 ± 0.04

and DPS, respectively. We note from FFHQ reconstructions in Fig. 4 that DAPS offers visually sharper reconstructions. However, it can be observed that the generated image is less measurement consistent; indicating strong consistency with the diffusion prior but weaker posterior consistency.

From Tables 1 and 2 we note that ALPS-MALA achieves higher sample avg. PSNR, while ALPS-ULA achieves marginally higher MMSE PSNR. ALPS-ULA produces a slightly broader sample distribution whereas ALPS-MALA yields samples that is more posterior consistent (see Figs. 7 and 8 for examples from mask inpainting and MRI recovery). Averaging over a slightly broader sample distribution often offers improved MMSE PSNR. The MH correction rejects proposals that deviate from target posterior distribution, making ALPS-MALA samples more posterior consistent. The ablation study in Fig. 3 further quantifies the effect of the MH correction and the preconditioner on convergence. Tables 1 and 2 also show that ALPS-MAP outperforms PnP-ISTA across all tasks. For example, for motion deblurring, ALPS-MAP achieves a 1.69 dB PSNR improvement over PnP-ISTA.

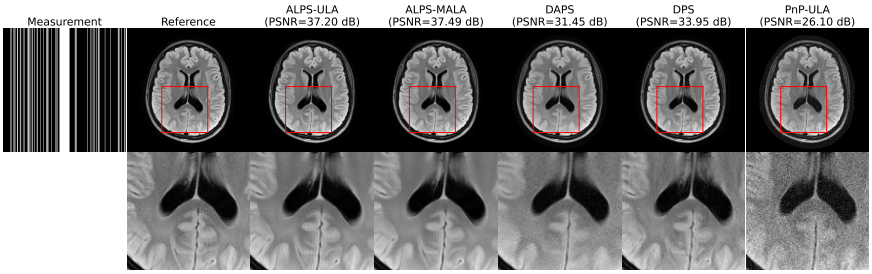
Posterior calibration: Figs. 7 and 8 show per-sample variance maps and DC values. ALPS-MALA concentrates uncertainty in regions where the posterior is genuinely broad: within the masked region for inpainting, and around edges and fine structures for MRI. Variance is low in measured regions, where the likelihood strongly constrains the posterior. ALPS-MALA also achieves the lowest DC error among all methods. In comparison, DAPS and PnP-ULA exhibit high variance even in measured regions, a sign of DC violations rather than legitimate posterior diversity. Consequently, DAPS and PnP-ULA generates samples that have high DC value as shown in Figs. 7 and 8.

Table 3: NFEs to generate one sample and inference time in seconds (in brackets) per method per inverse problem. For sampling methods, inference time reflects generating five samples; for ALPS-MAP and PnP-ISTA it reflects a single point estimate.

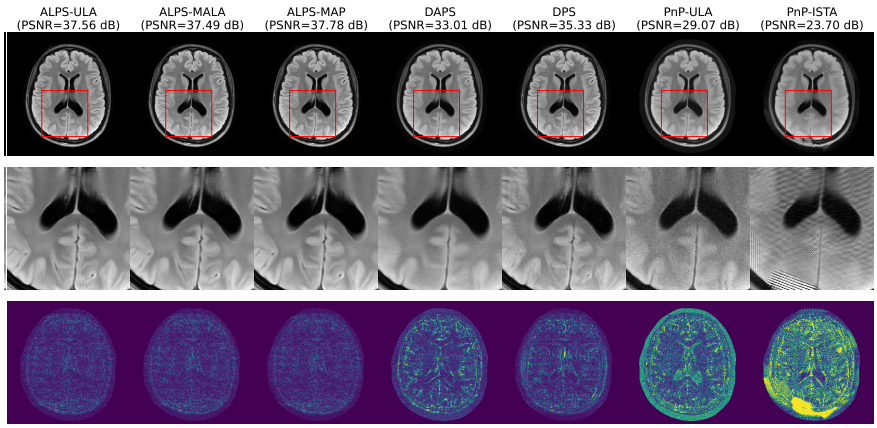
	ALPS-ULA	ALPS-MALA	ALPS-MAP	DAPS	DPS	PnP-ULA	PnP-ISTA
Inpaint (random)	100 (47.7)	≤200 (200.76)	100 (7.55)	1000 (140.49)	300 (76.88)	5000 (646.91)	300 (7.92)
Inpaint (mask)	200 (91.07)	≤400 (394.32)	200 (15.68)	1000 (138.67)	300 (76.24)	5000 (645.87)	300 (7.53)
Gaussian deblurring	100 (49.19)	≤200 (244.92)	100 (7.95)	1000 (158.38)	300 (77.24)	5000 (663.09)	300 (8.15)
Motion deblurring	100 (47.25)	≤200 (230.15)	100 (7.89)	1000 (164.64)	300 (85.69)	5000 (678.32)	300 (7.87)
MRI	100 (64.53)	≤200 (367.01)	100 (12.41)	1000 (312.95)	300 (124.54)	5000 (1150.21)	300 (14.66)

5.5 Computational Efficiency

Table 3 compares NFEs and wall-clock time across methods. ALPS-MAP and PnP-ISTA are the fastest methods, since both compute point estimates without sampling. However, ALPS-MAP achieves improved reconstruction quality than PnP-ISTA (see Tables 1 and 2). Among the sampling-based methods, ALPS-ULA is the fastest: it requires 10× fewer NFEs than DAPS, 3× fewer than DPS, and 50× fewer than PnP-ULA. This efficiency stems from requiring only a single score evaluation and CG solve, thereby avoiding the ODE integration required by DAPS. ALPS-MALA uses at most 2× as many NFEs as ALPS-ULA (since rejected proposals still cost one score evaluation), and has higher wall-clock time due to the MH step. In return, it yields more data consistent samples (see Fig. 7) and enjoys the geometric ergodicity guarantee of Theorem 4.1.



(a) FLAIR-weighted, 4x acceleration — samples



(b) FLAIR-weighted, 4x acceleration — MMSE and MAP

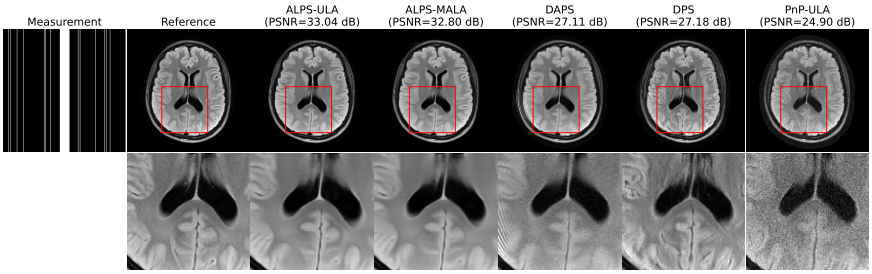
Figure 5: FastMRI (320 × 320) FLAIR-weighted reconstruction at 4x acceleration. The top panel shows samples (highest PSNR per method); the bottom panel shows MMSE and MAP estimates with error maps. Results at 8x acceleration are shown in Fig. 6.

6 Discussion

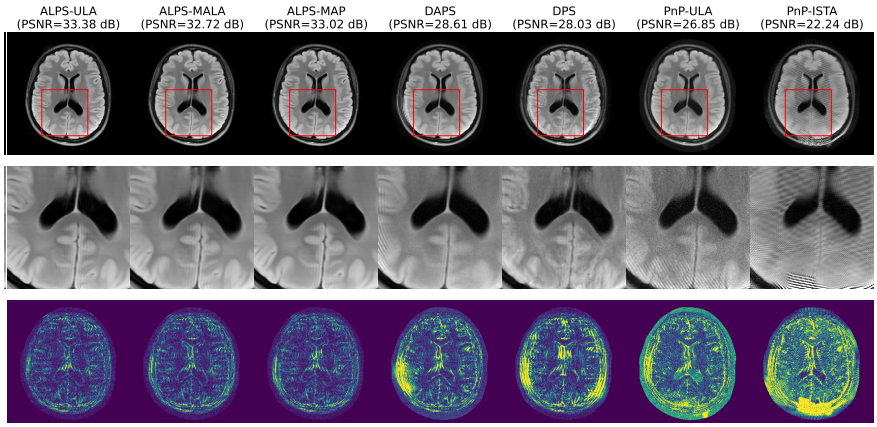
6.1 Sample Quality and Posterior Fidelity

All three ALPS variants outperform prior-path baselines across all tasks. We attribute this to three factors: (i) improved posterior consistency, enabled by the posterior path and the absence of latent-to-image transport; (ii) a preconditioner that mitigates the ill-conditioning of $\mathbf{A}$, thereby preventing slow mixing; and (iii) an explicit EBM that enables MH correction, converting discretization errors into rejected proposals rather than bias.

From Figs. 7 and 8, we observe that both ALPS-MALA and ALPS-ULA suppress spurious variance in measured regions relative to DAPS and PnP-ULA. ALPS-MALA sharpens this further by rejecting posterior inconsistent samples. In contrast, DAPS exhibits high variance even in directly observed pixels, a signature of DC violations that aligns with the trade-off reported in the DAPS paper. ALPS-MALA samples achieves lower DC error and higher sample PSNR than ALPS-ULA, as seen in Figs. 7 and 8. By contrast, ALPS-ULA's slightly broader distribution translates to improved averaging, giving marginally higher MMSE PSNR on several tasks. Since the MMSE estimate is approximated by a sample average, a chain that spreads mass more broadly across the posterior support produces a mean that is closer to the true $\mathbb{E}[\boldsymbol{x}|\mathbf{y}]$. Conversely, ALPS-MALA's tighter concentration near high-posterior regions improve individual sample quality, but offers less diversity for averaging. Overall, the key comparison is between the ALPS variants and the prior-path baselines, rather than between ALPS-MALA and ALPS-ULA.



(a) FLAIR-weighted, 8x acceleration — samples



(b) FLAIR-weighted, 8x acceleration — MMSE and MAP

Figure 6: FastMRI (320 × 320) FLAIR-weighted reconstruction at 8x acceleration. The top row shows samples (highest PSNR per method); the bottom row shows MMSE and MAP estimates with error maps (reference omitted for space; see Table 2 for Avg. PSNR). T2-weighted results are shown in Figs. 13 and 14 in Appendix B.1.

Fig. 7 shows ALPS-MALA reduces variance in both the measured foreground and the masked region. The latter is not mode collapse: DC values are tightly clustered (238.5–239.4), consistently lower than ULA (247.8–251.3), and PSNR varies across samples (27.85–28.15 dB), confirming the chain is mixing.

6.2 Consistent advantage across conditioning regimes

ALPS maintains a consistent performance advantage over the baseline algorithms across both well-posed and ill-posed settings. For random inpainting with p=0.3 (70% of pixels observed), the MMSE PSNR improvement over DAPS is approximately 3–4 dB. At 8x MRI acceleration (12.5% of k-space sampled), the improvement remains 3.5 dB on FLAIR and 2.1 dB on T2-weighted MRI. Notably, this gap does not diminish as the inverse problem becomes more challenging. We attribute this improvement to the posterior-path formulation, which enforces DC structurally from the first iterate, thereby maintaining performance even when measurements are sparse or the forward operator is ill-conditioned.

6.3 Score Evaluation Under Posterior Shift

The EBM is trained on samples from p_t, but posterior sampling explores $p_t(\boldsymbol{x}_t|\mathbf{y})$. However, we note that the preconditioned posterior score reduces the impact of any mismatch. In the measurement subspace, the data term dominates the gradient and limits the influence of the score. In the nullspace of $\mathbf{A}$, the data term vanishes and the score drives the update - but the injected noise has variance $\tau^2 t^2$, consistent with the EBM's training distribution. Out-of-distribution (OOD) exposure is thus concentrated where the score has limited influence, while the score receives in-distribution samples in the directions where it has significant influence.

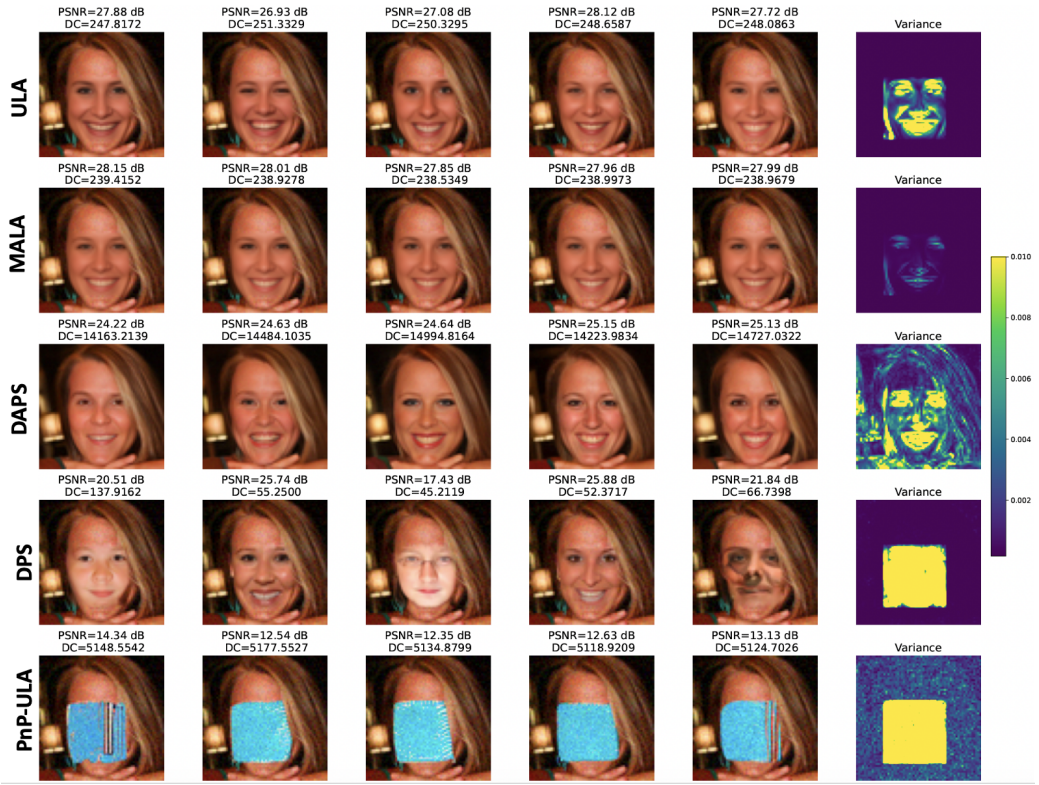


Figure 7: Posterior samples and uncertainty maps for the mask inpainting inverse problem. Per-sample PSNR and DC values are reported above each sample. ALPS-ULA and ALPS-MALA produces high PSNR samples, with uncertainty concentrated primarily within the masked region. ALPS-MALA attains lowest DC values due to the MH correction step, indicating stronger measurement consistency. MMSE and MAP comparisons are shown in Fig. 4(b).

The MH correction provides an additional safeguard: miscalibrated proposals in the measurement subspace are more likely to be rejected, converting any residual OOD error into elevated rejection rates rather than bias in the stationary distribution.

6.4 Design Choices Are Independently Justified

The ablation study in Fig. 3 shows that the geometry-aware preconditioner and the MH correction contributes independently and measurably to performance, and that each degrades in a predictable way when removed.

We note that the PSNR of ALPS-ULA matches that of ALPS-MALA at small step sizes, but degrades as the step size increases. In particular, large step sizes can cause ULA to generate samples in regions of low posterior probability. The MH correction rejects such proposals, making ALPS-MALA robust to the choice of step size (see Fig. 3.b). The correction is therefore a mechanism that makes the method robust to step size choice. A practitioner who cannot tune step size precisely retains reliable performance with ALPS-MALA but not with ALPS-ULA.

Removing the preconditioner slows convergence sharply on mask inpainting when compared to Gaussian deblurring (see Figs. 3.c and d). This observation is consistent with Theorem 4.3, which shows that the convergence rate is governed by the spectral gap of the MALA kernel. By mitigating the dominant ill-conditioning introduced by the forward operator, the geometry-aware preconditioner facilitates faster mixing. Consequently, the advantage of preconditioning becomes increasingly significant as the inverse problem becomes more ill-conditioned, consistent with the larger performance gap observed for mask inpainting in Fig. 3.d.

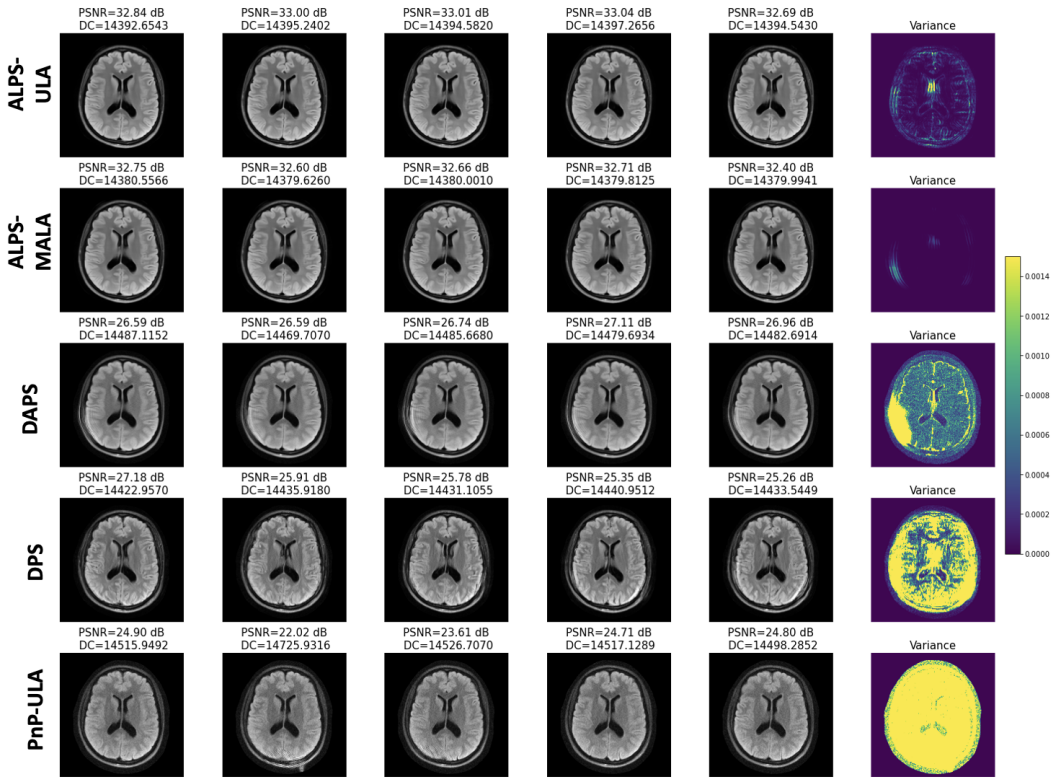


Figure 8: Posterior samples and uncertainty maps for 8x accelerated FLAIR-weighted MRI inverse problem. Per-sample PSNR and DC values are reported. ALPS-ULA and ALPS-MALA produce high PSNR samples; while uncertainty is concentrated around edges and fine anatomical structures that are weakly constrained by the undersampled measurements, consistent with a calibrated posterior. ALPS-MALA achieves the lowest DC values. MMSE and MAP comparisons are shown in Fig. 6(b).

Taken together, the ablation studies show that ALPS's performance advantage over prior-path baselines is not a consequence of any single design choice. Removing any one of them degrades performance in the direction and magnitude that theory predicts.

7 Conclusion

We proposed ALPS, a framework for posterior sampling that departs from the prior-path paradigm. Rather than imposing data consistency via corrective updates along an unconditional diffusion trajectory, ALPS distills a multi-scale EBM from the diffusion prior and constructs a sequence of annealed posteriors indexed by noise scale. Consequently, DC is a structural property of the chain from the first iterate, not an approximation enforced at the end.

The key enabler is the scalar energy recovered by distillation. Score-only methods discard this quantity, retaining only its gradient. Having the energy directly unlocks three capabilities. First, the negative log-likelihood composes analytically with the multi-scale EBM, so the negative log-posterior at each scale is available in closed-form without approximation. Second, the MH acceptance ratio is tractable, turning discretization errors into rejections rather than accumulated bias. Third, MAP estimation via MM admits a monotonic convergence guarantee without contraction assumptions on the score network.

A geometry-aware preconditioner, derived from an isotropic approximation of the energy Hessian, mitigates the dominant ill-conditioning introduced by the forward operator. This approximation is accurate at high noise scales, where it substantially improves the sampling geometry and facilitates faster mixing. Consequently, the benefits of preconditioning become increasingly pronounced for severely ill-conditioned inverse problems, consistent with the convergence analysis and the empirical results.

The multi-scale EBM is distilled from the same diffusion model used by the baselines, so no stronger prior is introduced. The PSNR improvements over diffusion-based posterior sampling algorithms therefore reflect the benefits of the proposed algorithm rather than a strong learned prior.

References

Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alexandros G. Dimakis, and Jonathan I. Tamir. Ambient diffusion posterior sampling: Solving inverse problems with diffusion models trained on corrupted data. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://github.com/utcsilab/ambient-diffusion-mri`.

Jyothi Rikhab Chand and Mathews Jacob. Multi-scale energy (muse) framework for inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 10:1250–1265, 2024. doi: 10.1109/TCI.2024.3449101.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. URL `https://arxiv.org/abs/2209.14687`. Spotlight.

Hyungjin Chung, Suhyeon Lee, and Jong Chul Ye. Decomposed diffusion sampler for accelerating large-scale inverse problems. In *International Conference on Learning Representations*, 2024. URL `https://arxiv.org/abs/2303.05754`.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/378a063b8fdb1db941e34f4bde584c7d-Abstract.html`.

Alain Durmus and Éric Moulines. On the geometric convergence for mala under verifiable conditions. *arXiv preprint arXiv:2201.01951*, 2022.

Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE Transactions on Image Processing*, 4(7):932–946, 1995.

Samuel Hurault, Arthur Leclaire, and Nicolas Papadakis. Gradient step denoiser for convergent plug-and-play. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2110.03220.

Ajil Jalal, Marius Arvinte, Giannis Daras, Eric Price, Alexandros G Dimakis, and Jon Tamir. Robust compressed sensing mri with deep generative priors. *Advances in neural information processing systems*, 34:14938–14954, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2206.00364.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in neural information processing systems*, 34:21757–21769, 2021.

Rémi Laumont, Valentin De Bortoli, Andrés Almansa, Julie Delon, Alain Durmus, and Marcelo Pereyra. Bayesian imaging using plug & play priors: when langevin meets tweedie. *SIAM Journal on Imaging Sciences*, 15(2):701–737, 2022.

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2nd edition, 2017.

Zengyi Li, Yubei Chen, and Friedrich T. Sommer. Learning energy-based models in high-dimensional spaces with multiscale denoising-score matching. *Entropy*, 25(10):1367, 2023. doi: 10.3390/e25101367. URL https://www.mdpi.com/1099-4300/25/10/1367.

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

Koichi Miyasawa. An empirical Bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38(4):181–188, 1961.

Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. On learning non-convergent short-run mcmc toward energy-based model. *arXiv preprint arXiv:1904.09770*, 2019.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Ernest Ryu, Jialin Liu, Sicheng Wang, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pp. 5546–5557. PMLR, 2019.

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

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12438–12448, 2020. URL https://proceedings.neurips.cc/paper/2020/file/92c3b916311a5517d9290576e3ea37ad-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2011.13456.

James Thornton, Louis Bethune, Ruixiang Zhang, Arwen Bradley, Preetum Nakkiran, and Shuangfei Zhai. Composition and control with distilled energy diffusion models and sequential monte carlo. *arXiv preprint arXiv:2502.12786*, 2025. URL https://arxiv.org/abs/2502.12786.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine L. Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*, 2024.

Xingyu Xu and Yuejie Chi. Provably robust score-based diffusion posterior sampling for plug-and-play image reconstruction. *arXiv preprint arXiv:2403.17042*, 2024.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL `https://arxiv.org/abs/2407.01521`.

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 Workshops*, pp. 1219–1229, 2023.

A Illustrative 2D Example

A.1 2D example

The main focus of this section is to illustrate the proposed ALPS approach using a toy dataset. We trained the multi-scale EBM on the 2D `moons` dataset. The top row of Fig.9 visualizes the learned negative log-prior (NLPr) or energy $E_\theta(\boldsymbol{x}; t)$ over several noise levels t. The blue points represent the data samples used to learn the EBMs, and the iso-contours indicate the learned energy landscape. The top row shows the energy at different noise scales $t \in \{0.750, 0.350, 0.200, 0.100\}$. The illustration indicates that the multi-scale EBM captures the manifold geometry at different scales. At the largest noise level $t = \sigma_{\max}$, the energy is approximately quadratic, corresponding to a Gaussian prior. As $t \to 0$, the energy landscape becomes increasingly representative of the true data distribution, capturing its multi-modal structure. This progressive refinement is a direct consequence of the annealed score-matching approach, which learns the score function across multiple noise scales.

A.2 Modeling the posterior

In inverse problems, the true posterior distribution often exhibits complex, multi-modal behavior due to measurement uncertainty and prior constraints. Here, we consider a linear measurement model $\mathbf{y} = \mathbf{a}^\top \mathbf{x} + \boldsymbol{b} + \eta\boldsymbol{\epsilon}; \eta = 0.2; \epsilon \sim \mathcal{N}(0, \mathbf{I})$, denoted by the red line in Fig. 9. The measurement line intersects the manifold at multiple locations, inducing a multi-modal posterior. Directly sampling from the true posterior is challenging because of the complex multi-modal nature. The composition property of EBMs allows us to combine the learned energy at time t with the negative-log likelihood term to form a family of posterior energies:

$$E_{\text{posterior}}(\boldsymbol{x}; \boldsymbol{t}) = E_{\text{prior}}(\boldsymbol{x}; t) + E_{\text{likelihood}}(\boldsymbol{x}),$$

where $E_{\text{likelihood}}$ encodes the measurement model. The time-dependent posterior energies at different noise scales are shown in the bottom row of Fig. 9.

A.3 ALPS recovery of samples

ALPS uses pre-conditioned Langevin dynamics with K steps to derive samples from the posterior at each scale. It also uses an annealing strategy, which gradually decreases t, starting from $t_0 = \sigma_{\max}$. This approach mitigates the impact of local minima and facilitates exploration of the energy landscape. At high noise levels, the energy surface is smooth and unimodal, allowing easy mixing. As t decreases, the model transitions towards the true multi-modal posterior, while the initialization from previous scales helps avoid getting trapped in spurious modes. This approach is particularly powerful for inverse problems where the posterior is highly non-convex and traditional optimization methods fail. Overall, Fig. 9 demonstrates how

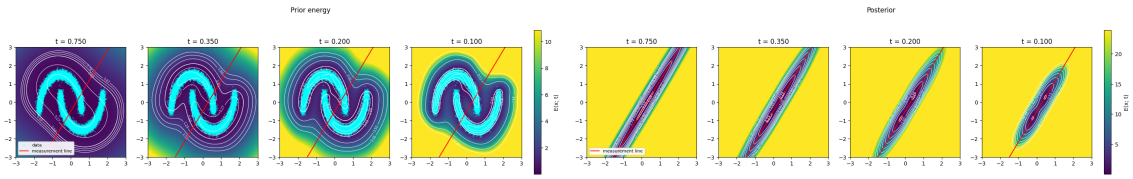


Figure 9: Learned EBM using score-based learning from the moons dataset. In the top row we show the data samples (blue points) and the score-learned energy at different time-scales. At $t = \sigma_{max}$, the energy is roughly quadratic corresponding to a Gaussian prior. As time approaches to 0, the energy becomes more representative of the true distribution. We also assume the measurements to be defined by the red straight line. The posterior energies at different time scales are shown in the bottom row; the maximum of the negative log-posterior is clipped for improved visualization. As time $t \to 0$, the posterior evolves from the likelihood to the true multi-modal posterior. The evolution of the samples offered by the ALPS algorithm is shown in Fig. 10

score-based EBMs, combined with annealed Langevin sampling, provide a principled framework for tackling challenging inverse problems with complex posterior distributions.

ALPS loop essentially alternates denoising (via the conservative score of the EBM), a quadratic DC update, and noise addition with a scale-aware preconditioner; samples are carried across scales to progressively sharpen the posterior. Figure 10 compares ALPS to the state-of-the-art DAPS. Unlike ALPS that anneals the posterior, most diffusion based inverse solvers (including DAPS) move through samples in time-dependent prior distributions, specified by the diffusion model. At each time point, DAPS uses an ODE solver to determine the true data samples corresponding to the specific latent, which is used to enforce the DC. By contrast, diffusion posterior sampling (DPS) relies on the score to map the latents to the true data samples. As seen from the bottom row of Fig. 10, the samples at each time point closely approximate the prior distributions. By contrast, ALPS does not require these mappings between latents and data samples as it works directly with time-dependent posteriors. Note from the top row of Fig. 10 that the samples are initialized approximately in the null-space of the measurement operator (the line). The noise update in ALPS ensures that they approximately stay in the null space, while evolving to the multimodal posterior distribution. This forward-model-aware update enables ALPS to obtain smoother trajectories across t, translating to faster convergence. The smoother nature of the convergence can also be seen from Fig. 2.

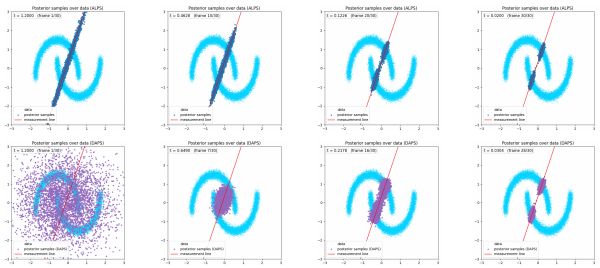


Figure 10: Comparison of posterior samples from ALPS (top row) and DAPS (bottom row), as a function of time. The columns correspond to posterior samples as a function of time, with the left most columns corresponding to the initialization and the right most ones as the final solution. The blue points indicate the data samples on the moons dataset, which is used to learn the prior distributions. We note that ALPS initializes the samples closer to the measurement manifold; the annealing of the posterior distributions smoothly transforms them to the desired posterior samples. By contrast, DAPS are initialized with the prior distribution $p_{\sigma_{\max}}$. The guidance term is carefully engineered such that the samples stay in the time dependent prior distributions p_t, while becoming data consistent at $t \to 0$. The smoother evolution of the samples in ALPS translates to faster sampling, and hence improved performance.

B Additional Inverse Problem Results

B.1 Remaining FFHQ and MRI Comparisons

Figs. 11 and 13 extend the comparisons in Fig. 4 (mask inpainting, Gaussian deblurring) and Figs. 5, 6 (FLAIR 4x, FLAIR 8x) to the remaining FFHQ tasks (random inpainting, motion deblurring) and T2-weighted MRI (4x, 8x). Each figure pairs best-sample rows with MMSE/MAP rows for direct comparison. The same pattern holds throughout: ALPS-MALA achieves the lowest DC values among sampling methods, ALPS-MAP matches or exceeds all baselines, and ALPS-ULA achieves marginally higher MMSE PSNR than ALPS-MALA in several settings (Tables 1, 2).

C Theoretical Results and Proofs

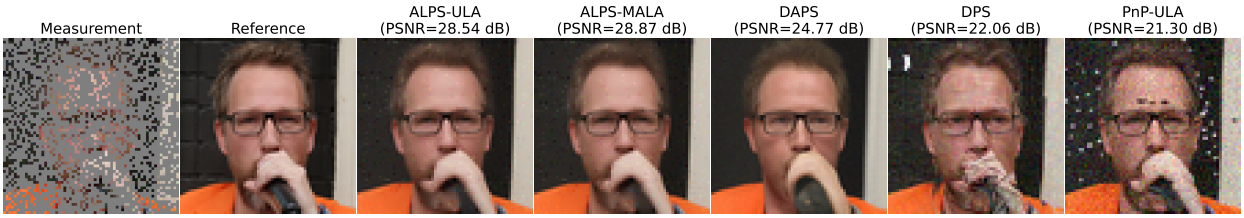
C.1 Equivalence of (5) and (ULA)

We begin with (5):

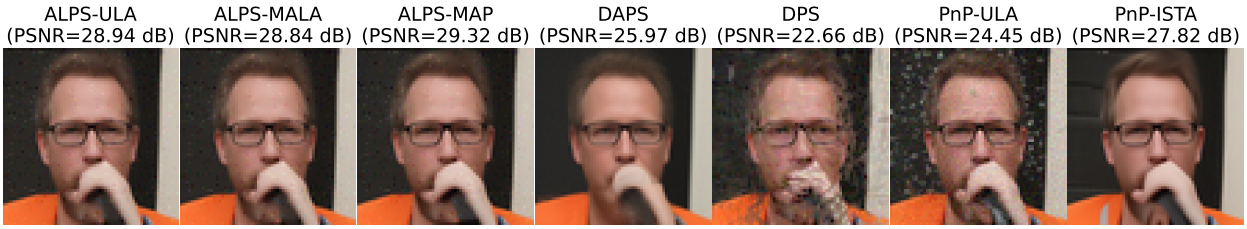
$$\tilde{\boldsymbol{x}}_{k+1,t} = \boldsymbol{x}_{k,t} - \frac{\tau^2}{2}\boldsymbol{B}_t\left(\frac{\mathbf{A}^T(\mathbf{A}\boldsymbol{x}_{k,t}-\mathbf{y})}{\eta^2} + \frac{\nabla_{\boldsymbol{x}_{k,t}}E_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,t})}{t^2}\right) + \tau\sqrt{\boldsymbol{B}_t}\epsilon_k \qquad \epsilon_k \sim \mathcal{N}(0,\mathbf{I}) \tag{22}$$

Using (6) we get:

$$\boldsymbol{x}_{k,t} - \frac{\tau^2}{2}\boldsymbol{B}_t\frac{\mathbf{A}^T\mathbf{A}}{\eta^2}\boldsymbol{x}_{k,t} = \boldsymbol{x}_{k,t}(\mathbf{I} - \frac{\tau^2\mathbf{I}}{2}) + \frac{\tau^2}{2t^2}\boldsymbol{B}_t\boldsymbol{x}_{k,t} \tag{23}$$



(a) Random inpainting ($p = 0.7$) — samples



(b) Random inpainting ($p = 0.7$) — MMSE and MAP

Figure 11: FFHQ (64×64) random inpainting ($p = 0.7$): samples (a) and MMSE/MAP (b). Results consistent with Fig. 4: ALPS-MALA achieves the highest sample PSNR and ALPS-MAP matches or exceeds all baselines.

Substituting the above equation in (22), we get:

$$\tilde{\boldsymbol{x}}_{k+1,t} = \boldsymbol{x}_{k,t}(\mathbf{I} - \frac{\tau^2\mathbf{I}}{2}) + \frac{\tau^2}{2}\boldsymbol{B}_t\left(\frac{\mathbf{A}^T\mathbf{y}}{\eta^2} + \frac{1}{t^2}\left(\boldsymbol{x}_{k,t} - \nabla_{\boldsymbol{x}_{k,t}}E_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,t})\right)\right) + \tau\sqrt{\boldsymbol{B}_t}\epsilon_k \tag{24}$$

$$\tilde{\boldsymbol{x}}_{k+1,t} = \boldsymbol{x}_{k,t}(\mathbf{I} - \frac{\tau^2\mathbf{I}}{2}) + \frac{\tau^2}{2}\underbrace{\boldsymbol{B}_t\left(\frac{\mathbf{A}^T\mathbf{y}}{\eta^2} + \frac{d_{k,t}}{t^2}\right)}_{\hat{\boldsymbol{x}}_{k+1,t}} + \tau\sqrt{\boldsymbol{B}_t}\epsilon_k$$

$$\tilde{\boldsymbol{x}}_{k+1,t} = \boldsymbol{x}_{k,t} + \frac{\tau^2}{2}(\hat{\boldsymbol{x}}_{k+1,t} - \boldsymbol{x}_{k,t}) + \tau\sqrt{\boldsymbol{B}_t}\epsilon_k$$

The above update step is a convex combination of current update $\boldsymbol{x}_{k,t}$ and the next update $\hat{\boldsymbol{x}}_{k+1,t}$ followed by noise addition. Note that $\hat{\boldsymbol{x}}_{k+1}$ is the solution of the following optimization problem:

$$\hat{\boldsymbol{x}}_{k+1,t} = \arg\min_{\boldsymbol{x}} \frac{\|\mathbf{A}\boldsymbol{x} - \mathbf{y}\|^2}{2\eta^2} + \frac{\|\boldsymbol{x} - \boldsymbol{d}_{k,t}\|^2}{2t^2} \tag{25}$$

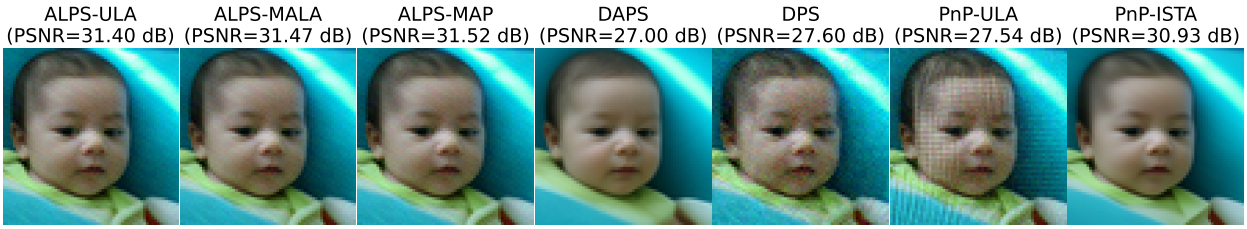
C.2 Implementation of $\log\alpha$ defined in (8)

$$\log\alpha(\boldsymbol{x}_{k,t}, \tilde{\boldsymbol{x}}_{k+1,t}) = min\{0, \log\left(\frac{\pi_t(\tilde{\boldsymbol{x}}_{k+1,t})q_t(\boldsymbol{x}_{k,t}|\tilde{\boldsymbol{x}}_{k+1,t})}{\pi_t(\boldsymbol{x}_{k,t})q_t(\tilde{\boldsymbol{x}}_{k+1,t}|\boldsymbol{x}_{k,t})}\right)\} = min\{0, \log r\} \tag{26}$$

where

$$\begin{aligned}\log r &= \log\pi_t(\tilde{\boldsymbol{x}}_{k+1,t}) - \log\pi_t(\boldsymbol{x}_{k,t}) + \log q_t(\boldsymbol{x}_{k,t} \mid \tilde{\boldsymbol{x}}_{k+1,t}) - \log q_t(\tilde{\boldsymbol{x}}_{k+1,t} \mid \boldsymbol{x}_{k,t}) \\ &= \log\pi_t(\tilde{\boldsymbol{x}}_{k+1,t}) - \log\pi_t(\boldsymbol{x}_{k,t}) \\ &\quad - \frac{1}{2\tau^2}\Big[\big((\boldsymbol{x}_{k,t} - \boldsymbol{\mu}_t(\tilde{\boldsymbol{x}}_{k+1,t}))^T\boldsymbol{B}_t^{-1}(\boldsymbol{x}_{k,t} - \boldsymbol{\mu}_t(\tilde{\boldsymbol{x}}_{k+1,t}))\big) \\ &\qquad - \big((\tilde{\boldsymbol{x}}_{k+1,t} - \boldsymbol{\mu}_t(\boldsymbol{x}_{k,t}))^T\boldsymbol{B}_t^{-1}(\tilde{\boldsymbol{x}}_{k+1,t} - \boldsymbol{\mu}_t(\boldsymbol{x}_{k,t}))\big)\Big],\end{aligned} \tag{27}$$

(a) Motion deblurring (length= 6, angle= 30°) — samples



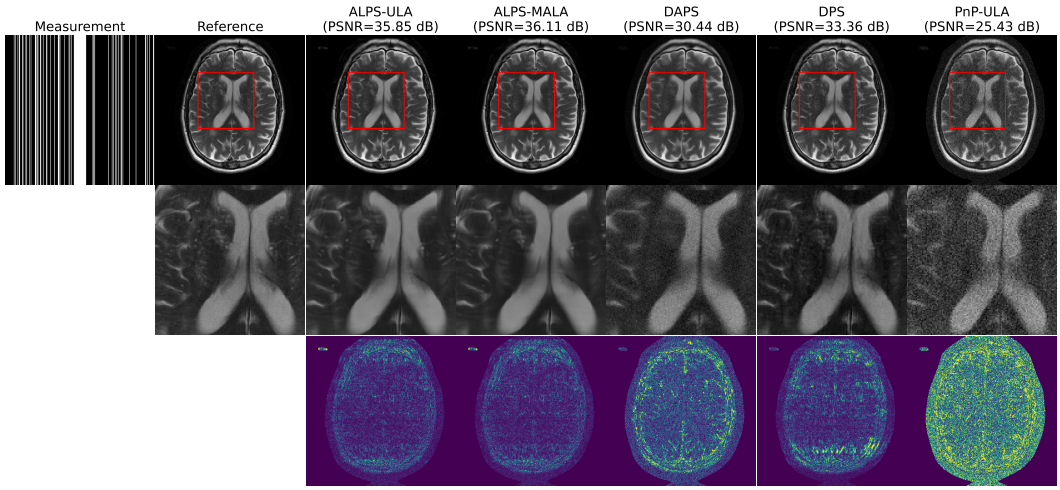
(b) Motion deblurring (length= 6, angle= 30°) — MMSE and MAP

Figure 12: FFHQ (64×64) motion deblurring (length= 6, angle= 30°): samples (a) and MMSE/MAP (b). Results consistent with Fig. 4: ALPS-MALA achieves the highest sample PSNR and ALPS-MAP matches or exceeds all baselines.

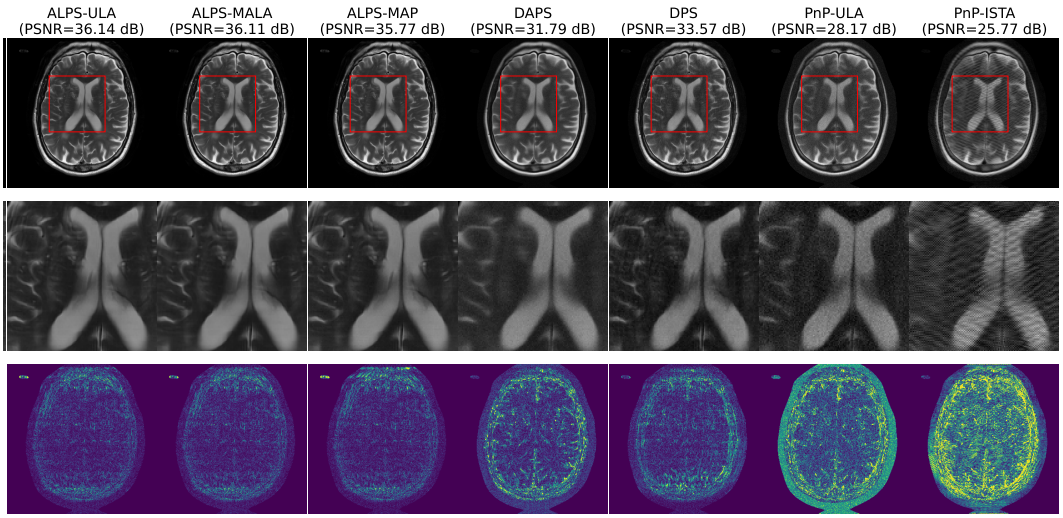
where $\boldsymbol{\mu}_t(\boldsymbol{x}) = \boldsymbol{x} - \frac{\tau^2}{2}\boldsymbol{B}_t\nabla_{\boldsymbol{x}}C_t(\boldsymbol{x})$. Next, we simplify the last two terms. Let $\boldsymbol{v}_1 = \boldsymbol{x}_{k,t} - \boldsymbol{\mu}_t(\tilde{\boldsymbol{x}}_{k+1,t})$ and $\boldsymbol{v}_2 = \tilde{\boldsymbol{x}}_{k+1,t} - \boldsymbol{\mu}_t(\boldsymbol{x}_{k,t})$. Then (27) simplifies as:

$$\begin{aligned} \log r &= \log \pi_t(\tilde{\boldsymbol{x}}_{k+1,t}) - \log \pi_t(\boldsymbol{x}_{k,t}) - \frac{1}{2\tau^2}(\boldsymbol{v}_1^T\boldsymbol{B}_t^{-1}\boldsymbol{v}_1 - \boldsymbol{v}_2^T\boldsymbol{B}_t^{-1}\boldsymbol{v}_2) \\ &= -C_t(\tilde{\boldsymbol{x}}_{k+1,t}) + C_t(\boldsymbol{x}_{k,t}) - \frac{1}{2\tau^2}\left(\left(\frac{\|\mathbf{A}\boldsymbol{v}_1\|^2}{\eta^2} + \frac{\|\boldsymbol{v}_1\|^2}{t^2}\right) - \left(\frac{\|\mathbf{A}\boldsymbol{v}_2\|^2}{\eta^2} + \frac{\|\boldsymbol{v}_2\|^2}{t^2}\right)\right) \end{aligned} \tag{28}$$

where in the second step we used the definition of the preconditioner.



(a) T2-weighted, 4x acceleration — samples



(b) T2-weighted, 4x acceleration — MMSE and MAP

Figure 13: fastMRI (320×320) T2-weighted, 4x acceleration: samples (a) and MMSE/MAP (b). Results consistent with FLAIR results in Figs. 5 and 6. The second rows in (a) and (b) shows zoomed sections, while the third rows show the error. ALPS sampling variants consistently outperform prior-path baselines.

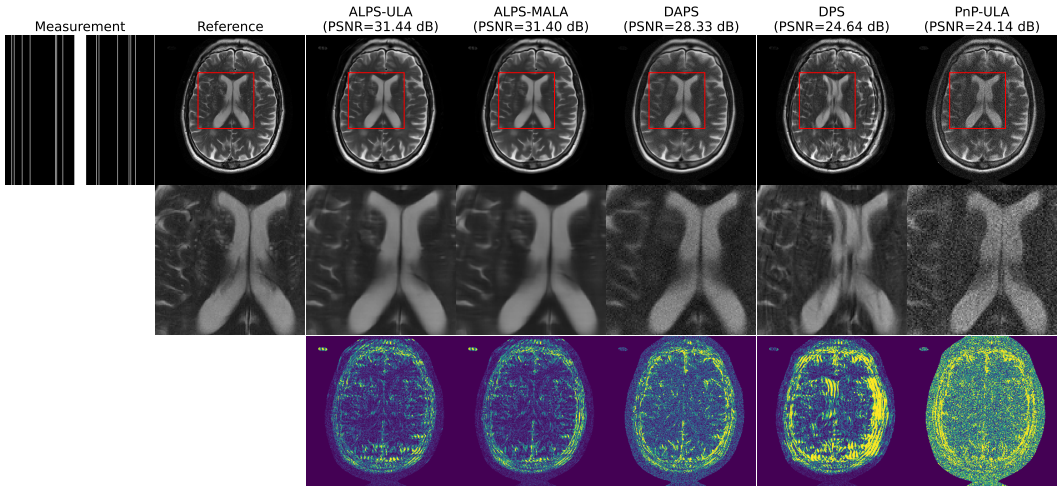
C.3 Proof of Theorem 4.1

For completeness, we first present the assumptions and Theorem 1 from Durmus & Moulines (2022), which, together with reversibility of the kernel, imply geometric ergodicity. We then verify that the ALPS algorithm satisfies these assumptions.

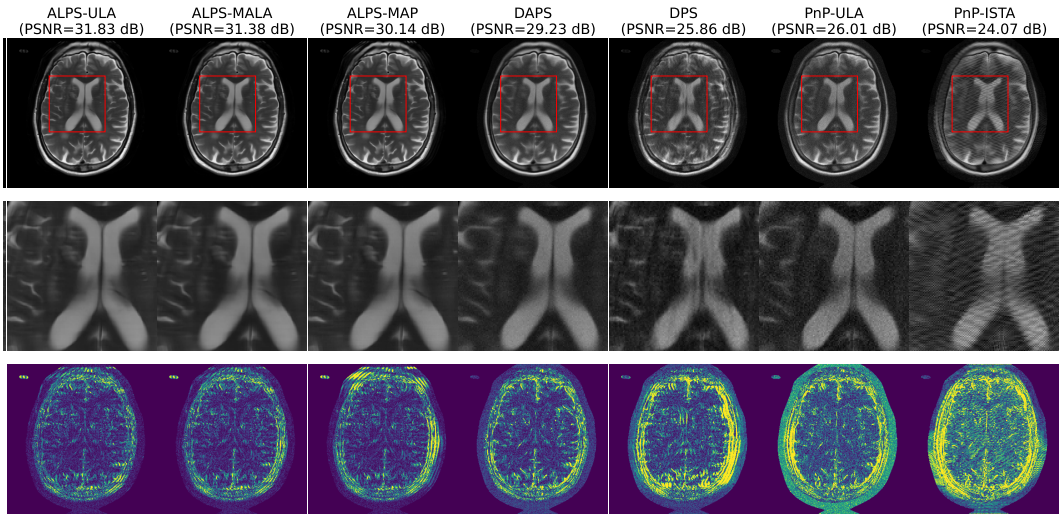
Durmus & Moulines (2022) establishes geometric ergodicity of MALA under the following two conditions on the potential U.

Assumption H1 Durmus & Moulines (2022). U is twice continuously differentiable and there exists $L \geq 0$ such that $\sup_{\boldsymbol{x}} \|D^2 U(\boldsymbol{x})\| \leq L$, i.e., ∇U is globally L-Lipschitz:

$$\|\nabla U(\boldsymbol{x}) - \nabla U(\mathbf{y})\| \leq L\|\boldsymbol{x} - \mathbf{y}\| \quad \text{for all } \boldsymbol{x}, \mathbf{y} \in \mathbb{R}^d.$$



(a) T2-weighted, 8x acceleration — samples



(b) T2-weighted, 8x acceleration — MMSE and MAP

Figure 14: fastMRI (320×320) T2-weighted, 8x acceleration: samples (a) and MMSE/MAP (b). Results consistent with FLAIR results in Figs. 5 and 6: ALPS sampling variants consistently outperform prior-path baselines across both accelerations.

Assumption H3 Durmus & Moulines (2022). There exist $\mathsf{m} > 0$ and $K \geq 0$ such that for any $\boldsymbol{x} \in \mathbb{R}^d$ with $\|\boldsymbol{x}\| \geq K$ and any unit vector $\mathbf{y}$,

$$\mathbf{y}^T \nabla^2 U(\boldsymbol{x}) \mathbf{y} \geq \mathsf{m},$$

i.e., U is strongly convex outside the ball $B(0, K)$.

Under **H1** and **H3**, Theorem 1 in Durmus & Moulines (2022) establishes V_κ-geometric ergodicity. We state this result:

Theorem C.1 *Durmus & Moulines (2022) Let $U : \mathbb{R}^d \to \mathbb{R}$ satisfy **H1** and **H3**. Let $\pi \propto e^{-U}$ and let R_τ be the MALA kernel with step size τ. Set $V_\kappa(\boldsymbol{x}) = \exp(\kappa \|\boldsymbol{x}\|^2)$ with $\kappa = \mathsf{m}/16$. Then there exists $\bar{\Gamma} > 0$ such that for any $\tau \in (0, \bar{\Gamma}]$, there exist constants $C_\tau \geq 0$ and $\rho_\tau \in (0, 1)$ such that for all $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\|\delta_{\boldsymbol{x}} R_\tau^K - \pi\|_{V_\kappa} \leq C_\tau \, \rho_\tau^K \, V_\kappa(\boldsymbol{x}), \tag{29}$$

*where $\delta_{\boldsymbol{x}} R_\tau^K$ denotes the distribution of the chain after K steps started at $\boldsymbol{x}$. The constants C_τ and ρ_τ depend on U through L and m but are non-quantitative (their explicit form requires the additional assumption **H2** of bounded third derivative).*

We now verify that the proposed MALA kernel satisfy the reversibility condition. Throughout this section, we treat t as fixed and suppress its dependence in the Markov kernel and proposal distribution for notational simplicity.

Lemma C.2 (Reversibility of preconditioned MALA) *The preconditioned MALA kernel $R_{\tau,t}$ with proposal*

$$q_t(\boldsymbol{x}' \mid \boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{x}';\, \boldsymbol{x} - \tfrac{\tau^2}{2} \boldsymbol{B}_t \nabla C_t(\boldsymbol{x}),\, \tau^2 \boldsymbol{B}_t\big)$$

and MH acceptance probability $\alpha(\boldsymbol{x}, \boldsymbol{x}') = \min\big(1,\, \frac{\pi_t(\boldsymbol{x}')\, q_t(\boldsymbol{x}|\boldsymbol{x}')}{\pi_t(\boldsymbol{x})\, q_t(\boldsymbol{x}'|\boldsymbol{x})}\big)$ is reversible with respect to $\pi_t \propto e^{-C_t}$ for any $\tau > 0$.

Proof C.1 *We re-parametrize the preconditioned MALA to standard MALA. Let $\boldsymbol{z} = \boldsymbol{B}_t^{-1/2} \boldsymbol{x}$ and $\tilde{C}_t(\boldsymbol{z}) = C_t(\boldsymbol{B}_t^{1/2} \boldsymbol{z})$. Since $\boldsymbol{B}_t \succ 0$, this is a bijection with constant Jacobian $(\det \boldsymbol{B}_t)^{1/2}$, so $\tilde{\pi}_t(\boldsymbol{z}) \propto e^{-\tilde{C}_t(\boldsymbol{z})}$.*

Multiplying the proposal mean by $\boldsymbol{B}_t^{-1/2}$ and using the chain rule $\nabla_{\boldsymbol{z}} \tilde{C}_t(\boldsymbol{z}) = \boldsymbol{B}_t^{1/2} \nabla_{\boldsymbol{x}} C_t(\boldsymbol{B}_t^{1/2} \boldsymbol{z})$ gives the $\boldsymbol{z}$-space mean $\boldsymbol{z} - \frac{\tau^2}{2} \nabla_{\boldsymbol{z}} \tilde{C}_t(\boldsymbol{z})$. The proposal covariance transforms as $\boldsymbol{B}_t^{-1/2} (\tau^2 \boldsymbol{B}_t) \boldsymbol{B}_t^{-1/2} = \tau^2 \mathbf{I}$. So in $\boldsymbol{z}$-space the proposal is $q(\boldsymbol{z}' \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}';\, \boldsymbol{z} - \frac{\tau^2}{2} \nabla_{\boldsymbol{z}} \tilde{C}_t(\boldsymbol{z}),\, \tau^2 \mathbf{I})$, which is the standard isotropic MALA proposal for $\tilde{C}_t$.

The MH ratio depends only on density ratios. The Jacobian $(\det \boldsymbol{B}_t)^{1/2}$ appears in both the numerator and denominator and therefore cancels, so $\alpha(\boldsymbol{x}, \boldsymbol{x}')$ is unchanged by the reparameterization. It therefore suffices to verify detailed balance for $\boldsymbol{z} \neq \boldsymbol{z}'$ in the $\boldsymbol{z}$-coordinates (the diagonal case $\boldsymbol{z}' = \boldsymbol{z}$ is trivial). Let R denote the MALA transition kernel in the $\boldsymbol{z}$-coordinates, with target density $\tilde{\pi}_t$ and proposal $q(\boldsymbol{z}' \mid \boldsymbol{z})$. Let $r(\boldsymbol{z}, \boldsymbol{z}') = \frac{\tilde{\pi}_t(\boldsymbol{z}') q(\boldsymbol{z}|\boldsymbol{z}')}{\tilde{\pi}_t(\boldsymbol{z}) q(\boldsymbol{z}'|\boldsymbol{z})}$. We check both cases for $\boldsymbol{z} \neq \boldsymbol{z}'$.

Case $r \leq 1$: *$\alpha(\boldsymbol{z}, \boldsymbol{z}') = r$ and $\alpha(\boldsymbol{z}', \boldsymbol{z}) = 1$, so*

$$\tilde{\pi}_t(\boldsymbol{z})\, R(\boldsymbol{z}, \boldsymbol{z}') = \tilde{\pi}_t(\boldsymbol{z})\, q(\boldsymbol{z}' \mid \boldsymbol{z}) \cdot r = \tilde{\pi}_t(\boldsymbol{z}')\, q(\boldsymbol{z} \mid \boldsymbol{z}') = \tilde{\pi}_t(\boldsymbol{z}')\, R(\boldsymbol{z}', \boldsymbol{z}).$$

Case $r > 1$: *$\alpha(\boldsymbol{z}, \boldsymbol{z}') = 1$ and $\alpha(\boldsymbol{z}', \boldsymbol{z}) = 1/r$, so*

$$\tilde{\pi}_t(\boldsymbol{z})\, R(\boldsymbol{z}, \boldsymbol{z}') = \tilde{\pi}_t(\boldsymbol{z})\, q(\boldsymbol{z}' \mid \boldsymbol{z}) = \tilde{\pi}_t(\boldsymbol{z}')\, q(\boldsymbol{z} \mid \boldsymbol{z}') \cdot \frac{1}{r} = \tilde{\pi}_t(\boldsymbol{z}')\, R(\boldsymbol{z}', \boldsymbol{z}).$$

Detailed balance holds in $\boldsymbol{z}$-space. The constant Jacobian cancels on both sides when transforming back, so $\pi_t(\boldsymbol{x}) R_{\tau,t}(\boldsymbol{x}, \boldsymbol{x}') = \pi_t(\boldsymbol{x}') R_{\tau,t}(\boldsymbol{x}', \boldsymbol{x})$ for all $\boldsymbol{x} \neq \boldsymbol{x}'$.

Note that Durmus & Moulines (2022) states that reversibility of R_τ with respect to π is a known consequence of the MH construction and does not impose it as an additional assumption. Lemma C.2 confirms this holds for the preconditioned kernel $R_{\tau,t}$ used in ALPS.

Lemma C.3 *The preconditioned MALA kernel is π_t-irreducible for any step size τ.*

Proof C.2 *To establish irreducibility, both the proposal and the target must have full support: a degenerate proposal cannot reach all regions of the state space, and a target with zero mass on part of the space causes systematic rejection there. Thus, the chain must be able to reach any set of positive π_t-mass from any starting point. Since $\boldsymbol{B}_t \succ 0$, the proposal is a non-degenerate Gaussian with $q_t(\boldsymbol{x}' \mid \boldsymbol{x}) > 0$ for all $\boldsymbol{x}, \boldsymbol{x}'$. Since C_t is finite everywhere, $\pi_t \propto e^{-C_t} > 0$ everywhere. Hence $R_{\tau,t}(\boldsymbol{x}, A) > 0$ for any measurable set A with $\pi_t(A) > 0$, establishing π_t-irreducibility.*

The following lemma verifies assumption H1 for the $C_t(\boldsymbol{x}) = \frac{\|\mathbf{A}\boldsymbol{x}-\mathbf{y}\|^2}{2\eta^2} + \frac{E_\theta(\boldsymbol{x};t)}{t^2}$

Lemma C.4 (Maximum eigenvalue of $\nabla^2 E_\theta$ via Miyasawa) *Suppose $E_\theta(\,\cdot\,;t)$ is learned by denoising score matching, so that its induced denoiser $\tilde{D}_\theta(\boldsymbol{x}_t, t) = \boldsymbol{x}_t - \nabla_{\boldsymbol{x}_t} E_\theta(\boldsymbol{x}_t, t)$ approximates the MMSE denoiser $D^*(\boldsymbol{x}_t, t) = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$. Then:*

$$\lambda_{\max}\left(\frac{\nabla^2_{\boldsymbol{x}_t} E_\theta(\boldsymbol{x}_t, t)}{t^2}\right) \le \frac{1}{t^2}.$$

Proof C.3 *Differentiating Miyasawa's identity $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t, t) = (D^*(\boldsymbol{x}_t, t) - \boldsymbol{x}_t)/t^2$ gives $\nabla^2_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t, t) = (J_{D^*} - \mathbf{I})/t^2$. Since $\dfrac{E_\theta(\boldsymbol{x}_t, t)}{t^2} = -\log p_\theta(\boldsymbol{x}_t, t) + \text{const}$:*

$$\nabla^2_{\boldsymbol{x}_t} \frac{E_\theta(\boldsymbol{x}_t, t)}{t^2} = \frac{\mathbf{I} - J_{\tilde{D}_\theta}}{t^2}.$$

The MMSE denoiser satisfies $J_{D^*} \succeq 0$ Miyasawa (1961), hence $\mathbf{I} - J_{D^*} \preceq \mathbf{I}$. Assuming the learned denoiser inherits this and dividing by t^2, we have $\lambda_{\max}(\nabla^2_{\boldsymbol{x}_t} \dfrac{E_\theta(\boldsymbol{x}_t, t)}{t^2}) \le 1/t^2$.

Finally, the Hessian of the negative log-likelihood is $\dfrac{\mathbf{A}^T\mathbf{A}}{\eta^2}$. Combining this with the bound $\lambda_{\max}\left(\dfrac{\nabla^2_{\boldsymbol{x}_t} E_\theta(\boldsymbol{x}_t, t)}{t^2}\right) \le \frac{1}{t^2}$, we obtain $\lambda_{\max}(\nabla^2_{\boldsymbol{x}_t} C(\boldsymbol{x}_t)) \le \dfrac{\mathbf{A}^T\mathbf{A}}{\eta^2} + \dfrac{1}{t^2}$. Therefore, ∇C_t is Lipschitz continuous with constant $\dfrac{\|\mathbf{A}\|^2}{\eta^2} + \dfrac{1}{t^2}$.

We next verify Assumption H3.

Lemma C.5 *Under assumption (A4), $\nabla^2_{\boldsymbol{x}} C_t(\boldsymbol{x}) \succeq 16\upsilon\mathbf{I}$ for all $\|\boldsymbol{x}\| \ge W$, and hence C_t is coercive.*

Proof C.4 *By assumption (A4), there exist $W > 0$ and $\upsilon > 0$ such that*

$$\nabla^2_{\boldsymbol{x}} C_t(\boldsymbol{x}) \succeq 16\upsilon\mathbf{I} \qquad \text{for all } \|\boldsymbol{x}\| \ge W.$$

Thus, C_t is uniformly strongly convex outside the ball $B(0, W)$. In particular, for any unit vector $u \in \mathbb{R}^d$ and any $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \ge W$,

$$u^\top \nabla^2_{\boldsymbol{x}} C_t(\boldsymbol{x})\, u \ge 16\upsilon.$$

Integrating this lower bound along rays shows that $C_t(\boldsymbol{x})$ grows at least quadratically once $\|\boldsymbol{x}\|$ is sufficiently large. Therefore,

$$C_t(\boldsymbol{x}) \to \infty \qquad \text{as } \|\boldsymbol{x}\| \to \infty,$$

and hence C_t is coercive.

Remark C.1 *Assumption (A4) - namely, that $\nabla^2_{\boldsymbol{x}} C_t(\boldsymbol{x}) \succeq 16\upsilon \mathbf{I}$ for all sufficiently large $\|\boldsymbol{x}\|$ - is the technical condition needed for the Lyapunov drift argument underlying Theorem 4.1. Intuitively, it ensures that the posterior energy grows in all directions in the tails, preventing the chain from drifting to infinity.*

Although this condition may be difficult to verify from first principles for general neural-network denoisers with SiLU or GELU activations (whose Hessians need not vanish in the tails), it is mild in practice. Posterior distributions over natural images typically concentrate on a bounded region of the image space, so the tail behavior of C_t is of limited practical relevance. In our experiments, we did not need to enforce (A4) explicitly.

When (A4) does not hold a priori, it can be enforced by replacing $E_\theta(\boldsymbol{x}; t)$ with a modified energy that suppresses the network contribution in the tails and replaces it with a quadratic:

$$\tilde{E}_\theta(\boldsymbol{x}; t) = \chi\Big(\tfrac{2\|\boldsymbol{x}\|}{W}\Big)\, E_\theta(\boldsymbol{x}; t) + \Big(1 - \chi\Big(\tfrac{2\|\boldsymbol{x}\|}{W}\Big)\Big)\, \epsilon \|\boldsymbol{x}\|^2,$$

where $\chi : \mathbb{R}_{\geq 0} \to [0, 1]$ is a smooth cutoff satisfying $\chi(s) = 1$ for $s \leq 1$ and $\chi(s) = 0$ for $s \geq 2$, $W > 0$ is chosen large enough that all posterior modes lie inside $\{\|\boldsymbol{x}\| \leq W/2\}$, and $\epsilon > 0$. For $\|\boldsymbol{x}\| \geq W$, the network contribution is completely suppressed and $\tilde{E}_\theta(\boldsymbol{x}; t) = \epsilon\|\boldsymbol{x}\|^2$, so

$$\nabla^2_{\boldsymbol{x}} \tilde{C}_t(\boldsymbol{x}) \succeq \frac{1}{\eta^2}\mathbf{A}^\top \mathbf{A} + \frac{2\epsilon}{t^2}\mathbf{I} \succeq \frac{2\epsilon}{t^2}\mathbf{I}.$$

Thus, $\tilde{C}_t$ satisfies (A4) with $\upsilon = \epsilon/(8t^2)$. Since this modification only affects the region $\|\boldsymbol{x}\| > W/2$, it leaves the posterior near the modes unchanged.

Proof of Theorem 4.1. From Lemma C.2, $R_{\tau,t}$ is reversible with respect to π_t. By Lemmas C.4 and C.5, C_t satisfies assumptions H1 and H3, respectively. To apply Theorem C.1, it remains to verify the condition $\nabla U(0) = 0$ required by H1. Since C_t is coercive and smooth it attains a global minimum $\boldsymbol{x}_t^*$. Defining the shifted potential $\tilde{C}_t(\boldsymbol{x}) = C_t(\boldsymbol{x} + \boldsymbol{x}_t^*)$ we have $\nabla \tilde{C}_t(0) = 0$ while H1 and H3 continue to hold with the same constants. Moreover, MALA targeting e^{-C_t} is identical to MALA targeting $e^{-\tilde{C}_t}$ up to a translation of the state space. Applying Theorem C.1 with $U = \tilde{C}_t$ and $\pi = \pi_t$, the kernel $R_{\tau,t}$ is V_κ-uniformly geometrically ergodic for any $\tau \in (0, \bar{\Gamma}_t]$, i.e., there exist $C_{\tau,t} \geq 0$ and $\rho_{\tau,t} \in (0, 1)$ such that

$$\|\delta_{\boldsymbol{x}} R^K_{\tau,t} - \pi_t\|_{V_\kappa} \leq C_{\tau,t}\, \rho^K_{\tau,t}\, V_\kappa(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Since ALPS applies this kernel at each scale t, the iterates converge geometrically to π_t at every fixed scale. □

It is instructive to compare the proof structure here with the analogous results in DPnP Xu & Chi (2024) and PnP-DM Wu et al. (2024). Both of those methods require analyzing the composition of two Markov kernels per scale - one for the denoising or ODE step and one for the likelihood step - because their chains return to the prior path between scales, and the joint convergence must account for both kernels simultaneously. ALPS targets a fixed posterior π_t at each scale with a single MALA kernel, so the standard geometric ergodicity result applies directly without multi-kernel composition. The theoretical simplification is not an artifact of the proof strategy; it is a consequence of the posterior-path design.

C.4 Proof of Theorem 4.2

Let $p_{i,K} = p_{i,0} R^K_{\tau,t_i}$. Since $p_{i,K}$ can be written as a mixture of Dirac measures, we have:

$$p_{i,K} = \int \delta_{\boldsymbol{x}} R^K_{\tau,t_i}\, p_{i,0}(d\boldsymbol{x}). \tag{30}$$

Subtracting π_{t_i} on both sides and applying the V_κ-norm:

$$\|p_{i,K} - \pi_{t_i}\|_{V_\kappa} \leq \int \|\delta_{\boldsymbol{x}} R^K_{\tau,t_i} - \pi_{t_i}\|_{V_\kappa}\; p_{i,0}(d\boldsymbol{x}). \tag{31}$$

where the inequality follows from the triangle inequality and the linearity of integration. Applying Theorem 4.1, we obtain

$$\begin{aligned}\|p_{i,K} - \pi_{t_i}\|_{V_\kappa} &\leq \int C_\tau \rho_\tau^K V_\kappa(\boldsymbol{x})\, p_{i,0}(d\boldsymbol{x}) \\ &= C_\tau \rho_\tau^K \mathbb{E}_{p_{i,0}}[V_\kappa(\boldsymbol{X})]\end{aligned} \tag{32}$$

By the triangle inequality we have:

$$\|p_{i,K} - \pi\|_{V_\kappa} \leq \underbrace{\|p_{i,K} - \pi_{t_i}\|_{V_\kappa}}_{\text{finite-iteration sampling error}} + \underbrace{\|\pi_{t_i} - \pi\|_{V_\kappa}}_{\text{annealing bias}}. \tag{33}$$

As $t_i \to 0$, the intermediate distribution π_{t_i} converges weakly to the target π, so the annealing bias vanishes. Moreover, as $K \to \infty$, $p_{i,K}$ converges to π_{t_i} by Theorem 4.1. Therefore, $p_{i,K}$ approaches the target distribution π as $K \to \infty$ and $t_i \to 0$.

C.5 Proof of Theorem 4.3

The theorem statement and its interpretation appear in Section 4.2. Recall that $1 - \lambda_t$ denotes the $L^2(\pi_t)$ spectral gap of $R_{\tau,t}$, where λ_t is its second eigenvalue. The per-step TV error of the implemented kernel $\tilde{R}_{\tau,t}$ is specified by $\varepsilon_t = \sup_{\boldsymbol{x}} \mathrm{TV}(\tilde{R}_{\tau,t}(\boldsymbol{x}, \cdot), R_{\tau,t}(\boldsymbol{x}, \cdot))$, and $\chi^2(p\|\pi_t) = \int (dp/d\pi_t - 1)^2\, d\pi_t$ is the Pearson's chi-square divergence. The proof follows the non-asymptotic analysis of Xu & Chi (2024), adapted to the single-scale reversible MALA kernel $R_{\tau,t}$.

Proof C.5 (Proof of Theorem 4.3) *Let p_k denote the law after k steps of $\tilde{R}_{\tau,t}$. By the triangle inequality and data-processing,*

$$\mathrm{TV}(p_{k+1}, \pi_t) \leq \mathrm{TV}(p_k \circ R_{\tau,t},\, \pi_t) + \varepsilon_t. \tag{34}$$

By Lemma C.2, $R_{\tau,t}$ is reversible with respect to π_t, which implies it is self-adjoint on $L^2(\pi_t)$ (see Levin & Peres (2017), Chapter 12). A self-adjoint operator has real spectrum; the largest eigenvalue is 1 (corresponding to the constant function, since π_t is stationary), and the second-largest eigenvalue $\lambda_t \in (0, 1)$ determines the $L^2(\pi_t)$ spectral gap $1 - \lambda_t$. This gives the χ^2-contractivity:

$$\chi^2(p \circ R_{\tau,t}^n \|\pi_t) = \left\| R_{\tau,t}^n \left(\tfrac{p}{\pi_t} - 1 \right) \right\|_{L^2(\pi_t)}^2 \leq \lambda_t^{2n} \left\| \tfrac{p}{\pi_t} - 1 \right\|_{L^2(\pi_t)}^2 = \lambda_t^{2n} \chi^2(p\|\pi_t).$$

Applying $\mathrm{TV} \leq \sqrt{\chi^2/2}$ and unrolling (34) over K steps gives (20).

D Score Distillation of Multi-scale EBMs

D.1 Training via Denoising Score Matching and EDM Preconditioning

The parameters of the multi-scale EBM $E_{\boldsymbol{\theta}}$ may be learned from data using Denoising Score Matching (DSM) Song & Ermon (2020). For a noisy input $\tilde{\boldsymbol{x}} = \boldsymbol{x} + t\boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}(0, \mathbf{I})$, the DSM objective is:

$$\theta^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{z}} \Big[w(t) \big\| \nabla_{\tilde{\boldsymbol{x}}} E_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}; t) - t\boldsymbol{z} \big\|^2 \Big], \tag{35}$$

where $w(t)$ are scale-dependent weights. We adopt the EDM preconditioning of Karras et al. (2022) for numerical stability, giving the preconditioned loss:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{z}} \left[w(t) \|\tilde{\boldsymbol{D}}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}}; t) - \boldsymbol{x}\|^2 \right]. \tag{36}$$

D.2 Distillation from a Pre-trained Diffusion Model

Direct training via (36) is expensive due to the auto-grad step, and the high variance of DSM makes it difficult to train competitive EBMs at scale. Following Thornton et al. (2025), we use distillation. An EBM requires a conservative (curl-free) score — a property not enforced during diffusion training. Rather than training from scratch, we match the EBM's induced denoiser to a pre-trained diffusion model's denoiser. The EBM denoiser is conservative by construction, so this regression finds the closest energy-compatible approximation to the teacher:

$$\arg\min_{\theta} \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{z}}\left[w(t)\|\tilde{\boldsymbol{D}}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{x}};t) - \boldsymbol{D}_{\text{teacher}}(\tilde{\boldsymbol{x}};t)\|^2\right]. \tag{37}$$

As shown in Thornton et al. (2025), this recovers the conservative component of the teacher's score, which is the component relevant for sampling. The distilled EBM inherits the teacher's generative quality while providing the scalar energy needed for posterior composition.

E Implementation Details

E.1 Computation of noise covariance matrix for different inverse problems

1. Inpainting: In inpainting, the forward operator $\mathbf{A} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with values 1's and 0's. This implies that $\mathbf{A}^T = \mathbf{A}$, $\mathbf{A}^T\mathbf{A} = \mathbf{A}^2 = \mathbf{A}$. Then the noise covariance at each scale will be:

$$\tau^2\boldsymbol{B}[i,i] = \begin{cases} \tau^2\left(\dfrac{1}{\eta^2} + \dfrac{1}{t^2}\right)^{-1} & \text{if } \mathbf{A}[i,i] = 1 \\ t^2\tau^2 & \text{if } \mathbf{A}[i,i] = 0 \end{cases} \tag{38}$$

2. Gaussian Deblurring: In this case, the forward operator can be diagonalized in the Fourier domain: $\mathbf{A} = \mathbf{F}^{-1}\text{diag}(\boldsymbol{H})\mathbf{F}$, where $\mathbf{F}$ is the 2D DFT matrix and $\boldsymbol{H}$ is a diagonal matrix with the diagonal elements containing the Fourier transform of the Gaussian kernel. Then the noise covariance matrix is given as:

$$\tau^2\boldsymbol{B} = \tau^2\mathbf{F}^{-1}\left(\frac{\text{diag}(|\boldsymbol{H}|^2)}{\eta^2} + \frac{1}{t^2}\right)^{-1}\mathbf{F} \tag{39}$$

   Note that when $\boldsymbol{H} \approx 0$ i.e., at high frequencies, noise covariance is dominated by the noise scale t while at low frequencies, the noise covariance matrix is dependent on both the measurement noise and scale t.

3. Motion Deblurring: Similar to Gaussian deblurring, the forward operator in motion deblurring can also be diagonalized in the Fourier domain: $\mathbf{A} = \mathbf{F}^{-1}\text{diag}(\boldsymbol{H})\mathbf{F}$, where $\boldsymbol{H}$ is a diagonal matrix with the diagonal elements containing the Fourier transform of the motion blur point spread function.

4. Multi Channel MRI: For multichannel MRI with Cartesian undersampling, the forward operator takes the form:

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{S}\cdot\mathcal{F}(\boldsymbol{C}_1\cdot x) \\ \vdots \\ \boldsymbol{S}\cdot\mathcal{F}(\boldsymbol{C}_N\cdot x) \end{bmatrix},$$

   where $\mathcal{F}$ denotes the discrete Fourier transform, $\boldsymbol{S}$ is a binary sampling mask in k-space, and $\boldsymbol{C}_c$ are the coil sensitivity maps. We can approximate $\mathbf{A}^T\mathbf{A} \approx \mathcal{F}^{-1}(S)\mathcal{F}$. In this case, we choose a diagonal preconditioner in the Fourier domain:

$$\tau^2\boldsymbol{B} = \tau^2\mathcal{F}^{-1}\cdot\text{diag}\left(\frac{1}{\frac{S(k)}{\eta^2} + \frac{1}{\sigma_t^2}}\right)\cdot\mathcal{F},$$

   where $S(k)$ is the sampling mask in k-space.

Note that SNIPS Kawar et al. (2021) tries to make the noise covariance dependent on the measurement noise. To do so, they use SVD. Since, we are using a preconditioner in the update, the synthetic noise is indeed dependent on both measurement noise and synthetic noise.

E.2 Implementation discussion

E.2.1 Training Protocol

All experiments were conducted on A100 GPUs on a single node of up to 4 GPUs. The optimal weights of the energy models were obtained using the Adam optimizer. We did not perform augmentation on the training dataset when using score distillation technique to train the EBM. For natural images, we used the *SongNet* NCSNPP architecture for the denoiser $D_\theta(\boldsymbol{x}; t)$; while for MRI we used *SongNet* DDPM architecture. The details of the network are listed in Table 4.

Table 4: Network parameters for CIFAR-10 (NCSN++), FFHQ-64 (NCSN++), and MRI-384 (DDPM++).

Parameter	**CIFAR-10 (NCSN++)**	**FFHQ-64 (NCSN++)**	**MRI-384 (DDPM++)**
normalization	GroupNorm	GroupNorm	GroupNorm
nonlinearity	swish	swish	swish
nf	128	128	128
ch_mult	[1,2,2]	[1,2,2,2]	[1,2,2,2,2,2]
num_res_blocks	4	4	2
attn_resolutions	(16)	(16)	(12)
resamp_with_conv	True	True	True
fir	True	True	True
fir_kernel	[1,3,3,1]	[1,3,3,1]	[1,1]
skip_rescale	True	True	True
resblock_type	biggan	biggan	ddpmpp
progressive	none	none	none
progressive_input	residual	residual	standard
progressive_combine	sum	sum	sum
attention_type	ddpm	ddpm	ddpm
embedding_type	fourier	fourier	positional
channel_mult_emb	4	4	4
channel_mult_noise	2	2	1
init_scale	0.0	0.0	0.0
fourier_scale	16	16	—
conv_size	3	3	3
num_scales	15	18	30
dropout	0.10	0.10	0.10
label_dropout	0.0	0.0	0.0
encoder_type	residual	residual	standard
decoder_type	standard	standard	standard
resample_filter	[1,3,3,1]	[1,3,3,1]	[1,1]

E.3 ALPS implementation details

Noise schedule: Similar to Karras et al. (2022), we use the following schedule in Algorithm 1

$$t_i = \left(\sigma_{\max}^{1/\rho} + \frac{i}{N-1}\left(\sigma_{\min}^{1/\rho} - \sigma_{\max}^{1/\rho}\right)\right)^{\rho}, \qquad i = 0, \ldots, N.$$

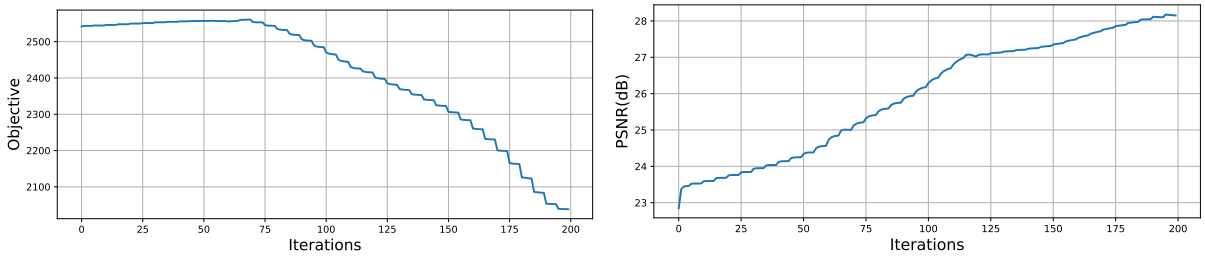
Table 5 shows values of σ_{max}, σ_{min}, and ρ for the different inverse problems.

Table 5: ALPS Hyperparameter

	Random inpaint	Mask inpaint	Motion/ Gaussian deblur	MRI
ρ	5	5	5	7
σ_{max}	1	5	1	1
σ_{min}	0.1	0.1	0.1	0.002

ALPS-MALA: To implement Algorithm 1 with the MH correction, we replace the *for* loop over K iterations

with *while* loop. The while loop exits either when either K proposals are accepted or 2K proposal attempts have been made. This compensates for samples rejected by the MH correction step. Furthermore, the step size τ is tuned so that the acceptance probability is neither always equal to one nor excessively small. Fig. 16 shows the acceptance rate vs. annealing step for the mask inpainting problem. We also found it beneficial to initialize from $\mathcal{N}(\hat{\boldsymbol{x}}_{0,t_{i+1}}, \tau^2 \boldsymbol{B}_{t_{i+1}})$. Note that the convergence proof still holds for this initialization.
Monotonic convergence of ALPS-MAP: Fig. 15. (a) and Fig. 15. (b) illustrates the convergence plot at each scale and the PSNR evolution for the mask inpainting problem. From Fig. 15.(a) it can be observed that MM algorithm is decreasing the objective monotonically at a particular scale. Furthermore, Fig. 15. (b) shows that PSNR gradually increases as annealing progresses towards $t \to 0$.



(a) Objective vs. Iterations

(b) PSNR vs. Iterations

Figure 15: Illustration of convergence plot and PSNR evolution of ALPS-MAP algorithm

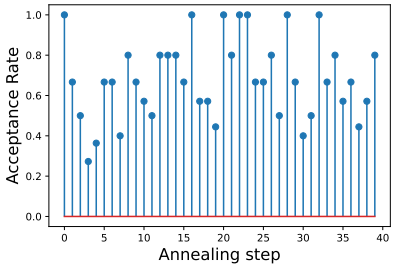


Figure 16: Acceptance rate of ALPS-MALA algorithm. The step size is tuned such that the algorithm does not always accept/reject a sample.

E.4 Baseline implementation details

For all the diffusion based algorithms we used the teacher diffusion model for the inference.

1. DAPS: We adopt the same settings suggested in Zhang et al. (2025) for the linear inverse problems.

2. DPS: We used the annealing schedule suggested in Karras et al. (2022) and Aali et al. (2025) for the natural image and MRI inverse problems, respectively. The total number of iterations is set to 300 in both cases. The step size is set to $\dfrac{0.95}{\|\mathbf{y} - \mathbf{A}\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)\|}$ for all inverse problems.

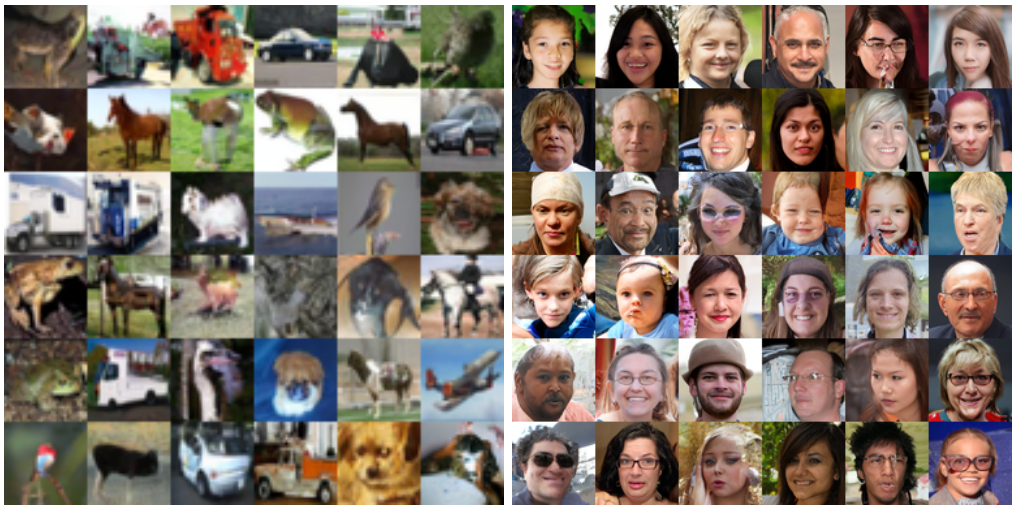
3. PnP-ULA: We used the EDM diffusion denoiser with $\sigma = 0.1$, step size $= 10^{-3}$, and number of iterations $= 5000$.

4. PnP-ISTA: We used the EDM diffusion denoiser with $\sigma = 0.1$, step size $= 0.005$, and number of iterations $= 300$. The algorithm was initialized with $\mathbf{A}^\top \mathbf{y}$.

F Unconditional Generation Quality

F.1 Prior samples

We show the prior samples generated by training the distilled multi-scale EBMs on the CIFAR-10, FFHQ dataset (64×64), and MRI images in Fig. 17 and Fig. 18, respectively. The samples are generated using Heun's algorithm. We note that on the CIFAR-10 dataset, the distilled EBM achieves competitive FID.



(a) CIFAR-10 (32×32), FID: 3.35, NFE: 35 (b) FFHQ (64×64), FID: 3.76, NFE: 79

Figure 17: Unconditional generation of prior samples using the distilled multi-scale energy-based model.

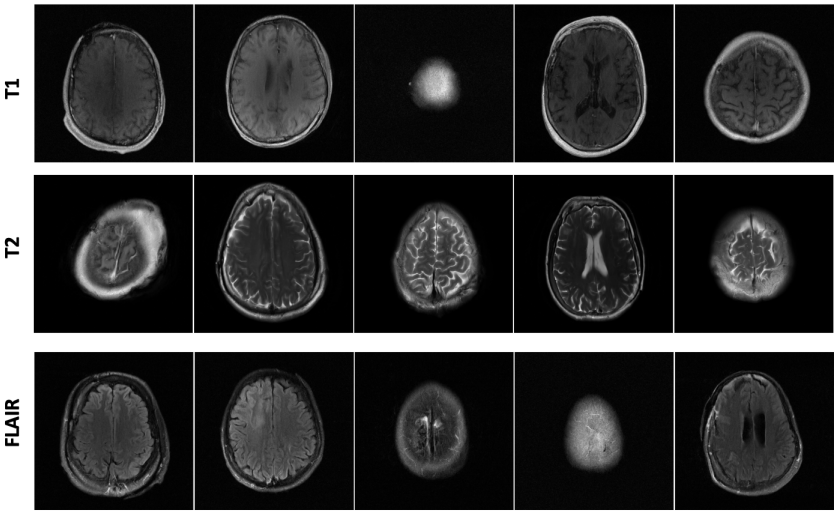


Figure 18: MRI prior samples with dimension 384×384 generated by the distilled EBM with 35 NFE.