

SAMPLE-EFFICIENT EVIDENCE ESTIMATION OF SCORE BASED PRIORS FOR MODEL SELECTION

Frederic Wang & Katherine L. Bouman

Department of Computing and Mathematical Sciences
 Caltech
 Pasadena, CA 91125, USA
 {fzwang, klbouman}@caltech.edu

ABSTRACT

The choice of prior is central to solving ill-posed imaging inverse problems, making it essential to select one consistent with the measurements \mathbf{y} to avoid severe bias. In Bayesian inverse problems, this could be achieved by evaluating the model evidence $p(\mathbf{y} | M)$ under different models M that specify the prior and then selecting the one with the highest value. Diffusion models are the state-of-the-art approach to solving inverse problems with a data-driven prior; however, directly computing the model evidence with respect to a diffusion prior is intractable. Furthermore, most existing model evidence estimators require either many pointwise evaluations of the unnormalized prior density or an accurate clean prior score. We propose `DiME`, an estimator of the model evidence of a diffusion prior by integrating over the time-marginals of posterior sampling methods. Our method leverages the large amount of intermediate samples naturally obtained during the reverse diffusion sampling process to obtain an accurate estimation of the model evidence using only a handful of posterior samples (e.g., 20). We also demonstrate how to implement our estimator in tandem with recent diffusion posterior sampling methods. Empirically, our estimator matches the model evidence when it can be computed analytically, and it is able to both select the correct diffusion model prior and diagnose prior misfit under different highly ill-conditioned, non-linear inverse problems, including a real-world black hole imaging problem.

1 INTRODUCTION

In Bayesian inverse imaging problems across science and engineering, the chosen prior distribution $p(\mathbf{x})$ plays a pivotal role in shaping the posterior $p(\mathbf{x} | \mathbf{y})$ from which candidate solutions are drawn (Feng et al., 2024; Stuart, 2010). In ill-posed settings, the prior acts as a regularizer, steering reconstructions toward solutions consistent with expected image properties. If the ground truth image lies outside the support of the chosen prior, the resulting reconstruction can be severely biased.

A central challenge in Bayesian inference is how to select a prior to use when the true distribution is unknown. Consequently, the priors used in practice are often chosen somewhat arbitrarily. If the likelihood of a prior could be easily evaluated, it would provide a principled basis for prior selection. Specifically, given a set of potential models $\{M_i\}$ that specify the prior $p(\mathbf{x} | M_i)$, we would like to find the most likely candidate model by computing $p(M_i | \mathbf{y}) \propto p(\mathbf{y} | M_i)p(M_i)$. It is commonly assumed that $p(M_i)$ is uniform, so it suffices to compute the (*log*) *model evidence*, $\log p(\mathbf{y} | M_i)$. Model evidence is not only useful for selecting the best prior for reconstruction, but also for quantifying epistemic uncertainty and assessing the assumptions underlying a model or physical theory (Trotta, 2007).

Although the model evidence provides a principled way to select the best prior for reconstruction, computing it generally requires evaluating an intractable integral over the full prior: $\log p(\mathbf{y} | M_i) = \log \int p(\mathbf{y} | \mathbf{x}, M_i)p(\mathbf{x} | M_i)d\mathbf{x}$. This motivates the development of estimation methods, such as sequential Monte Carlo (Del Moral et al., 2006; Chen et al., 2024), nested sampling (Skilling, 2006; Cai et al., 2022), annealed importance sampling Neal (2001); Doucet et al. (2022), and harmonic mean estimators (Newton & Raftery, 1994; McEwen et al., 2021; Polanska et al., 2023;

Spurio Mancini et al., 2023). However, these methods rely on sampling using the clean prior score $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ or the unnormalized density $\log p(\mathbf{x})$, which may be inaccurate for out-of-distribution \mathbf{x} , expensive to evaluate, and can lead to slow mixing due to the ill-conditioned data geometry.

Bayesian inverse imaging solvers achieve state-of-the-art performance with diffusion-based priors (Song et al., 2020; Ho et al., 2020; Song et al., 2021b; Jalal et al., 2021; Mardani et al., 2023), yet an accurate model evidence remains intractable using existing methods. Diffusion models generate unconditional samples by denoising a sequence of intermediate noised images, each corresponding to a time-marginal distribution between pure noise and the clean image. Diffusion-based posterior sampling methods adhere to posterior time-marginals instead (Chung et al., 2022; Song et al., 2023; Coeurdoux et al., 2024; Li et al., 2024; Zhu et al., 2023; Wu et al., 2024; Zhang et al., 2025). While efficient diffusion density estimators have been proposed (Skreta et al., 2024), density-based estimation methods are still typically high-variance for high dimensional problems and require thousands of posterior samples, which is computationally challenging with diffusion models. Furthermore, as diffusion models learn the score of intermediate noised priors, their estimates of the clean prior score are usually less accurate and/or highly ill-conditioned, making it difficult for traditional methods to explore the full distribution, while using their score from a higher noise levels can add bias. As a result, existing estimators are not suitable for or give biased estimates for diffusion-based priors.

Our contribution: We propose the Diffusion Model Evidence (DiME) estimator, a method for estimating the model evidence under a diffusion model prior that does not require the prior score or density. We derive the estimator for posterior sampling methods that anneal along the *standard (posterior) marginals* $p(\mathbf{x}_t | \mathbf{y})$, as well as a generalized estimator for any arbitrary path of marginals that anneal to the true posterior. By integrating along these posterior time-marginals, DiME accurately estimates the model evidence using intermediate samples already generated during posterior sampling. DiME thus only requires a small number of posterior samples (e.g., 20) for accurate estimation. We also describe how to practically implement DiME alongside a state-of-the-art posterior sampling method Decoupled Annealing Posterior Sampling (DAPS).

We perform several experiments validating our method. We first test in a mixture of Gaussian setting where an analytic evidence can be derived, and find that DiME provides nearly unbiased estimates of the evidence and performs comparably to baselines sequential Monte Carlo, annealed importance sampling, and thermodynamic integration despite not using the prior score. We also test DiME on two non-convex inverse problems, Gaussian and Fourier phase retrieval. DiME consistently selects the correct prior from a set of 10 diffusion models trained on MNIST (Deng, 2012) digits given a single noisy measurement of one MNIST digit, while the baselines fail to do so.

Finally, we use our method to perform both model selection and validation on real M87* black hole observations from the Event Horizon Telescope (EHTC, 2019b;c). DiME indicates that a prior derived from synthetic black hole images generated using General-Relativistic Magnetohydrodynamics (GRMHD) (Mizuno, 2022) has higher likelihood compared to the Radiatively Inefficient Accretion Flow (RIAF) (Broderick et al., 2011) black hole prior, a prior trained on general space images (Alam et al., 2024), a prior trained on faces (Liu et al., 2015), and a prior trained on MNIST digit 0s. Furthermore, we diagnose the validity of the GRMHD prior by prior predictive checking on the model evidence. Our results suggest that the M87* observations are statistically in-distribution with respect to the GRMHD prior, while still leaving room in the model for refinement.

2 BACKGROUND

2.1 DIFFUSION MODELS

Given a clean image $\mathbf{x}_0 \sim p_0(\mathbf{x}_0)$, the diffusion forward process at time $t \in [0, T]$ is given by:

$$\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t, \quad \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I}) \quad (1)$$

where a_t is the signal scaling, σ_t is noise scaling, and a'_t, σ'_t are the time derivatives. We can sample from the clean image distribution by discretizing the reverse SDE:

$$d\mathbf{x}_t = \left[\frac{a'_t}{a_t} \mathbf{x}_t - g_t^2 \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \right] dt + g_t d\mathbf{w}_t \quad (2)$$

where $g_t^2 = 2 \left(\sigma_t \sigma'_t - \frac{a'_t}{a_t} \sigma_t^2 \right)$, \mathbf{w}_t is Brownian motion and

$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)$ is learned. We can obtain the posterior mean via Tweedie’s formula (Efron, 2011):

$$\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t] = \frac{1}{\alpha_t} \left(\mathbf{x}_t + \sigma_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \right). \quad (3)$$

2.2 DIFFUSION-BASED POSTERIOR SAMPLING METHODS

Most diffusion posterior sampling methods fall broadly under two categories: (1) explicitly approximating the intractable likelihood score $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{y} | \mathbf{x}_t)$ as guidance during reverse diffusion, and (2) methods that anneal along a path of specified intermediate marginals, using techniques such as sequential Monte Carlo (Cardoso et al., 2023; Wu et al., 2023), gradient-based optimization (Moufad et al., 2024), or Langevin dynamics (Zhang et al., 2025; Wu et al., 2024; Zhu et al., 2023). The approximation errors of (1) compound over the sampling process, resulting in intermediate time-marginals with no tractable form and a biased posterior, while (2) corrects these errors by adhering to a specified marginal path, allowing for their time-marginals to anneal to the true posterior. As a result, in this work we focus on two different sampling methods of (2) that have demonstrated strong performance for solving physical inverse problems (Zheng et al., 2025).

Concretely, given measurements \mathbf{y} from an inverse problem $\mathbf{y} = \mathbf{A}(\mathbf{x}) + \varepsilon$ we introduce two likelihood terms. The first, $p(\mathbf{y} | \mathbf{x}_t) = \int p(\mathbf{y} | \mathbf{x}_0)p(\mathbf{x}_0 | \mathbf{x}_t)d\mathbf{x}_0$ is the marginal likelihood of \mathbf{y} conditioned on the current diffusion state \mathbf{x}_t . The second, $q(\mathbf{y} | \mathbf{x}_t) := p(\mathbf{y} | \mathbf{x}_0 = \mathbf{x}_t)$ is a time-independent data likelihood term that assumes \mathbf{x}_t is the clean image.

The *Decoupled Annealing Posterior Sampling* (DAPS) method (Zhang et al., 2025) alternates between noising and sampling from the posterior. Given initial noise samples $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and N -step annealing schedule $[t_1, \dots, t_N]$, we sample from the clean image posterior using Langevin dynamics before re-adding noise according to the diffusion process:

$$\tilde{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_{t_k}, \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x}_0)p(\mathbf{x}_0 | \mathbf{x}_{t_k}) \quad (4)$$

$$\mathbf{x}_{t_{k-1}} \sim p(\mathbf{x}_{t_{k-1}} | \tilde{\mathbf{x}}_0) \quad (5)$$

Zhang et al. (2025) proposes two approaches: *Gaussian approximation DAPS*, where $p(\mathbf{x}_0 | \mathbf{x}_t)$ is approximated via a Gaussian $\mathcal{N}(\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t], \Sigma_{\mathbf{x}_0 | \mathbf{x}_t})$ and the covariance $\Sigma_{\mathbf{x}_0 | \mathbf{x}_t}$ is heuristically chosen as σ_t^2 , and *exact DAPS*, where the prior score is computed using the diffusion model at a tiny noise level $\nabla \log p(\mathbf{x}_0) \approx \nabla \log p(\mathbf{x}, \sigma_{min})$. The unconditional distribution of intermediate samples closely match the *standard marginals*: $\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x}_t)p(\mathbf{x}_t)$ for all times t .

The *Plug-and-Play Diffusion Models* (PnP-DM) sampling method (Wu et al., 2024) takes inspiration from the Split Gibbs Sampler (Vono et al., 2019) and related diffusion-based samplers (Coourdoux et al., 2024) by splitting each annealing iteration into a likelihood step and a prior step. In particular, given N -step annealing schedule $[t_1, \dots, t_N]$ and clean image initialization $\hat{\mathbf{x}}_0$, we alternate between running Langevin dynamics to obtain a sample from the next time-marginal posterior and using the diffusion model to sample a clean image. We rewrite the updates for one timestep t_k in terms of standard diffusion notation:

$$\mathbf{x}_{t_k} \sim q(\mathbf{x}_{t_k} | \mathbf{y}, \hat{\mathbf{x}}_0) \propto q(\mathbf{y} | \mathbf{x}_{t_k})p(\mathbf{x}_{t_k} | \hat{\mathbf{x}}_0) \quad (6)$$

$$\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_{t_k}) \quad (7)$$

where \mathbf{x}_{t_k} is sampled from a joint distribution of the noised $\hat{\mathbf{x}}_0$ and the time-independent likelihood q . We write $q(\mathbf{y} | \mathbf{x}_{t_k})$ to emphasize this proxy likelihood applies the measurement model directly to a noisy rather than clean image.

We refer to the resulting unconditional distributions of intermediate samples as *PnP-DM marginals*, denoted by $\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{y}) \propto q(\mathbf{y} | \mathbf{x}_t)p(\mathbf{x}_t)$ for all t .

2.3 EXISTING MODEL EVIDENCE ESTIMATORS

The simplest estimator approximates the evidence by averaging $p(\mathbf{y} | \mathbf{x})$ over samples $\mathbf{x} \sim p(\mathbf{x})$, though this requires exponentially many samples in high dimensions. The standard harmonic mean estimator (Newton & Raftery, 1994) mitigates this issue but has infinite variance. Learned harmonic

mean estimators (McEwen et al., 2021; Polanska et al., 2023; Spurio Mancini et al., 2023) stabilize the variance by importance sampling with a surrogate neural density, under the assumption that the prior density is known. However, surrogate training requires thousands of posterior samples and importance sampling requires a similar amount of density evaluations even in medium-dimensional settings, making these approaches infeasible for diffusion-based priors.

Nested sampling (Skilling, 2006; Cai et al., 2022) re-parametrizes the model evidence into a one-dimensional integral over the prior volume, but still requires the score or density and is limited to log-convex likelihoods. Thermodynamic integration (TI) (Lartillot & Philippe, 2006), annealed importance sampling (AIS) (Neal, 2001) and sequential Monte Carlo (SMC) (Del Moral et al., 2006) compute the evidence by integrating over a path of distributions p_β , with $0 \leq \beta \leq 1$. For diffusion priors, accurately sampling from arbitrary $p_\beta(\mathbf{x})$ is computationally difficult as we lack a clean prior score. Reverse diffusion Monte Carlo (RDMC) (Huang et al., 2023) methods simulate the diffusion process to sample from unnormalized densities, using a Monte Carlo approach to approximate the noised scores. Guo et al. (2025) independently derived an evidence estimator for the reverse diffusion process under very different assumptions: they rely on a closed-form posterior score, obtain exact posterior marginals directly from an analytic diffusion that is not learned, and provide only theoretical convergence results without experimental validation.

3 ESTIMATING THE MODEL EVIDENCE ALONG THE STANDARD MARGINALS

We introduce `DIME` to estimate model evidence from the standard marginals $p(\mathbf{x}_t | \mathbf{y})$, aligning closely with the annealing path of DAPS. All proofs in this section are shown in Appendix A. A general model evidence estimator for arbitrary marginals annealing to the true posterior is derived in Appendix B, and an unnormalized model evidence estimator that instead uses the PnP-DM marginals (`DIME-PnPDM`) is derived and implemented in Appendix B.1.

Proposition 1 (`DIME`: Model evidence estimation along the standard marginals). *Given diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$, posterior marginals $p(\mathbf{x}_t | \mathbf{y}) \propto p(\mathbf{x}_t)p(\mathbf{y} | \mathbf{x}_t)$, and timesteps $0 = t_0 < \dots < t_N = T$, the model evidence can be estimated via:*

$$\log p(\mathbf{y}) = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{y} | \mathbf{x}_0)] - D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)). \quad (8)$$

The log-likelihood term can be estimated with posterior samples¹, and the KL divergence with:

$$D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)) \approx \sum_{i=1}^N c_{t_i} \Delta t_i \mathbb{E}_{\mathbf{x}_{t_i} \sim p(\mathbf{x}_{t_i} | \mathbf{y})} \|\nabla_{\mathbf{x}_{t_i}} \log p(\mathbf{y} | \mathbf{x}_{t_i})\|^2 \quad (9)$$

where $c_{t_i} = \sigma'_{t_i} \sigma_{t_i} - \sigma_{t_i}^2 \frac{a'_{t_i}}{a_{t_i}}$ and $\Delta t_i = t_i - t_{i-1}$ are derived from the diffusion schedule.

The sum across each sample-path $\{\mathbf{x}_{t_i}\}_{i=0, \dots, N}$ can be viewed as the distance of the resulting posterior sample \mathbf{x}_0 to the prior. In the remainder of the section, we showcase how to practically implement `DIME` in tandem with the DAPS method but with two modifications: in Section 3.1 we design an improved covariance approximation² for $p(\mathbf{x}_0 | \mathbf{x}_t)$ if the Gaussian approximation DAPS is used, and in Section 3.2 we propose to sample two $\tilde{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y})$ for each \mathbf{x}_t to estimate $\|\nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t)\|^2$ needed for Eq. 9. The full algorithm can be seen in Algorithm 1.

3.1 AN IMPROVED APPROXIMATION FOR THE POSTERIOR COVARIANCE

While the exact DAPS sampler can be used for accurate sampling of $\tilde{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}) \propto p(\mathbf{y} | \mathbf{x}_0)p(\mathbf{x}_0 | \mathbf{x}_t)$ (Eq. 4), Gaussian approximation DAPS provides a significantly cheaper alternative by approximating $p(\mathbf{x}_0 | \mathbf{x}_t) \approx \mathcal{N}(\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t], \sigma_t^2)$. However, this covariance heuristic only accounts for $p(\mathbf{x}_t | \mathbf{x}_0)$ and ignores $p(\mathbf{x}_0)$. While this heuristic is accurate at low noise levels, it overestimates the variance in the high noise regimes: at the terminal noise level $t = T$, we expect $p(\mathbf{x}_0 | \mathbf{x}_T) \approx p(\mathbf{x}_0)$ with a covariance that should match the prior $\text{Cov}(\mathbf{x}_0)$, but the heuristic gives

¹For likelihoods with Gaussian noise, if the posterior and prior have enough overlap such that posterior samples have a data fit of $\chi^2 \approx 1$, this term essentially becomes constant under the same likelihood function.

²While approximating $p(\mathbf{x}_0 | \mathbf{x}_t)$ is necessary for the standard marginals annealing path, it is not necessary for many annealing paths under the generalized framework introduced in Appendix B.

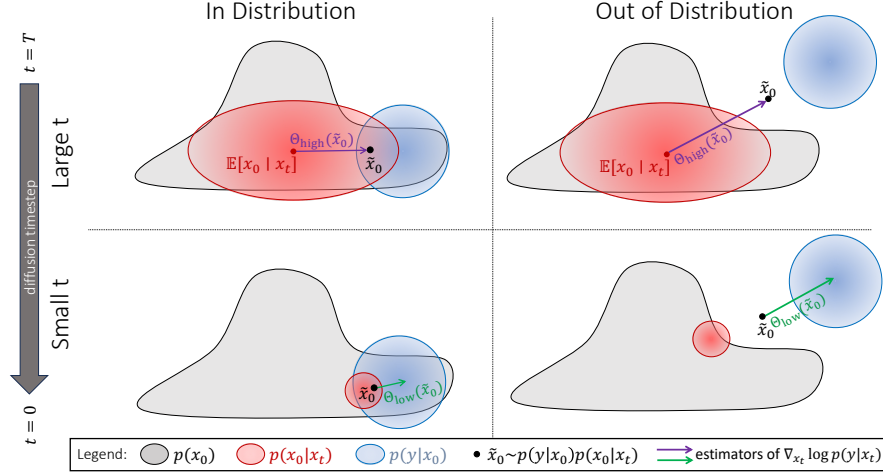


Figure 1: Visualization of the unbiased estimators $\Theta_{high}, \Theta_{low}$ of the likelihood score $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t)$ described in Lemma 2 for both in-distribution (left) and out-of-distribution (right) measurements \mathbf{y} . At large diffusion time steps, the high noise estimator (top) uses the distance between $\tilde{\mathbf{x}}_0$ and $\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$. At small diffusion time steps, the low noise estimator (bottom) uses the likelihood score at $\tilde{\mathbf{x}}_0$. The estimator is larger (longer arrows) when \mathbf{y} is out-of-distribution, as there greater KL divergence from the posterior to the prior; this results in DiME computing a lower evidence. Note the scaling factors $\frac{a_t}{\sigma_t^2}$ and $\Sigma_{\mathbf{x}_0|\mathbf{x}_t}$ of the estimator are not shown.

$\text{Cov}(\mathbf{x}_0 | \mathbf{x}_T) \approx \sigma_T^2 \gg \text{Cov}(\mathbf{x}_0)$. This mismatch at high noise was also observed empirically in Zhang et al. (2025). Since DiME uses all time-marginals it is important to accurately estimate $p(\mathbf{x}_0 | \mathbf{x}_t)$ for all t . Therefore, we include knowledge of $p(\mathbf{x}_0)$ by approximating it as Gaussian $\mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$, using an empirical covariance computed from the training data similar to the covariance approximation introduced in Linhart et al. (2024). Introducing an approximate prior provides a more accurate representation of $p(\mathbf{x}_0 | \mathbf{x}_t)$, which remains a Gaussian approximation as it is now the product of two Gaussians. The resulting covariance is given by:

Lemma 1. *Suppose we have diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t, \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$. Let $\mathbf{x}_t \sim p(\mathbf{x}_t)$ and suppose $p(\mathbf{x}_0) \approx \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$. Then we can express the posterior covariance $\Sigma_{\mathbf{x}_0|\mathbf{x}_t}$ as*

$$\Sigma_{\mathbf{x}_0|\mathbf{x}_t} = \left[\Sigma_0^{-1} + \frac{a_t^2}{\sigma_t^2} \mathbf{I} \right]^{-1}. \quad (10)$$

For large images where computing a full covariance matrix is intractable, a diagonal approximation in a transform basis has also been demonstrated to work well (Peng et al., 2024), and for linear inverse problems, an analytic local covariance is computable (Boys et al., 2023). Our approximation aligns the annealing path more closely with $p(\mathbf{x}_t | \mathbf{y})$ for all t . It is worth noting that DAPS re-samples at every annealing step, ensuring that these Gaussian approximation errors are never compounded. As shown in Sections 4.1 and 4.3.1, using the proposed improved approximation leads to nearly unbiased evidence estimates for both a multimodal Gaussian toy example with varying local curvature as well as real-world scientific priors.

3.2 ESTIMATING THE SQUARED LIKELIHOOD SCORE

While directly computing $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t)$ is intractable, DiME only requires an unbiased estimate as seen in Eq. 9. Estimators using unconditional samples of $\mathbf{x}_0 | \mathbf{x}_t$ are high-variance: these samples usually have low likelihood $p(\mathbf{y} | \mathbf{x}_0)$ resulting in volatile scores. Thankfully, DAPS provides samples of $\tilde{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y})$, which we can use to create estimators with lower variance:

Lemma 2. *Suppose we have diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t, \mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$ and sample $\tilde{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y})$. Under the assumption that $p(\mathbf{x}_0 | \mathbf{x}_t) \approx \mathcal{N}(\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t], \Sigma_{\mathbf{x}_0|\mathbf{x}_t})$, both $\Theta_{high}(\tilde{\mathbf{x}}_0)$*

and $\Theta_{low}(\tilde{\mathbf{x}}_0)$ are unbiased estimators of $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t)$ with

$$\Theta_{high}(\tilde{\mathbf{x}}_0) = \frac{a_t}{\sigma_t^2} (\tilde{\mathbf{x}}_0 - \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]). \quad (11)$$

$$\Theta_{low}(\tilde{\mathbf{x}}_0) = \frac{a_t}{\sigma_t^2} (\boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t} \nabla_{\tilde{\mathbf{x}}_0} \log p(\mathbf{y} | \tilde{\mathbf{x}}_0)). \quad (12)$$

Under low diffusion noise ($a_t \rightarrow 1, \sigma_t \rightarrow 0$), $\text{Var}(\Theta_{low}) \rightarrow 0$ but $\text{Var}(\Theta_{high}) = O(\frac{1}{\sigma_t^2})$.

Under high noise ($a_t \rightarrow 0, \sigma_t \rightarrow 1$), $\text{Var}(\Theta_{high}) \rightarrow 0$ but $\text{Var}(\Theta_{low}) = O(\frac{\sigma_t^4}{\sigma_t^2} \text{Var}(\Theta_{high}))$.

A proof of these bounds is displayed in Appendix A.1. The lack of the measurement \mathbf{y} in the high noise estimator bounds the worst possible case, where the data is weak; when the measurement contains a lot of information, the variance would be much lower. As such, Θ_{high} is used more towards the beginning of the sampling process while Θ_{low} is used towards the end. As both estimators are cheap to compute, we can simply evaluate both and choose the lower variance estimator at each timestep. Finally, as we want an unbiased estimate of the *squared* likelihood score, squaring the estimator is biased: $\mathbb{E}\|\Theta\|^2 = \|\mathbb{E}\Theta\|^2 + \text{Tr}(\text{Cov}(\Theta))$. Therefore, for each intermediate sample \mathbf{x}_t , we propose to sample two i.i.d $\tilde{\mathbf{x}}_0^{(1)}, \tilde{\mathbf{x}}_0^{(2)} \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y})$, giving us the following unbiased estimator: $\Theta(\tilde{\mathbf{x}}_0^{(1)})^T \Theta(\tilde{\mathbf{x}}_0^{(2)})$. The full method is displayed in Algorithm 1, and a visualization for both in-distribution and out-of-distribution measurements is visualized in Figure 1.

Algorithm 1 One posterior sample-path of DiME implemented with DAPS

Require: Score-based model s_θ , measurement \mathbf{y} , noise schedule $\sigma_t, (t_i)_{i \in \{1, \dots, N\}}$

- 1: Sample $\mathbf{x}_{t_N} \sim \mathcal{N}(0, \sigma_{t_N}^2 \mathbf{I})$.
 - 2: Initialize the integral sum, $f = 0$.
 - 3: **for** $i = N, N - 1, \dots, 1$ **do**
 - 4: Compute $\hat{\mathbf{x}} = \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$ using Tweedie’s formula and s_θ .
 - 5: Sample $\tilde{\mathbf{x}}_0^{(1)}, \tilde{\mathbf{x}}_0^{(2)} \sim p(\mathbf{x}_0 | \mathbf{x}_{t_i}, \mathbf{y})$ using Langevin with $p(\mathbf{x}_0 | \mathbf{x}_t) = \mathcal{N}(\hat{\mathbf{x}}, \boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t})$. (3.1)
 - 6: Sample $\mathbf{x}_{t_{i-1}} \sim \mathcal{N}(a_{t_{i-1}} \tilde{\mathbf{x}}_0^{(1)}, \sigma_{t_{i-1}}^2 \mathbf{I})$.
 - 7: $f_i = \arg \min_{\Theta \in \{\Theta_{low}, \Theta_{high}\}} \text{Var}(\Theta(\tilde{\mathbf{x}}_0^{(1)})^T \Theta(\tilde{\mathbf{x}}_0^{(2)}))$ computed using all sample-paths. (3.2)
 - 8: Update the integral sum $f \leftarrow f + (t_i - t_{i-1}) \left(\sigma_t' \sigma_t - \sigma_t^2 \frac{a_t'}{a_t} \right) f_i$.
 - 9: **end for**
 - 10: **return** $\mathbf{x}_0, -f + \log p(\mathbf{y} | \mathbf{x}_0)$
-

4 EXPERIMENTS

We compare DiME and DiME-PnPDM with five baselines: naive Monte Carlo (Naive MC), thermodynamic integration (TI), annealed importance sampling (AIS), Sequential Monte Carlo (SMC), and the original DAPS covariance heuristic (Original DAPS heuristic). Baseline methods are described in Appendix C and all tuning and implementation details are provided in Appendix D.

4.1 BENCHMARKING ON GAUSSIAN MIXTURE PRIOR WITH GROUND TRUTH EVIDENCE

To verify DiME, we first test on a multimodal 1000-D Gaussian mixture prior and linear forward model $\mathbf{y} = \mathbf{A}\mathbf{x} + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ where the model evidence has an analytic form to compare to. We use two Gaussian components with means at $(-0.75, 0.75)$ and identical covariances of $0.25 \mathbf{I}$. The entries of $\mathbf{A} \in \mathbb{R}^{200 \times 1000}$ are i.i.d. $\mathcal{N}(0, \frac{1}{200})$, and $\sigma = 0.1$. We run three experiments where the ground truth \mathbf{x}^* has varying model evidence to test the robustness of our method: in-distribution \mathbf{x}^* , out-of-distribution \mathbf{x}^* , and $\mathbf{x}^* = 0$ located at a saddle point. The analytic $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ is used to sample the intermediate distributions for TI, AIS, and SMC. In contrast, for DiME the analytic $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ is used to compute $\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]$ for $\sigma_t > 0.05$ and the analytic score is never used. All methods are tested using 100 discretization steps (equivalently, annealing time steps for DiME) and 20 sample paths for 50 trials. The mean and standard deviation is reported.

	In-distribution \boldsymbol{x}^*		Out-of-distribution \boldsymbol{x}^*		Saddle point \boldsymbol{x}^*	
	Estimate	Rel. Error ↓	Estimate	Rel. Error ↓	Estimate	Rel. Error ↓
Ground truth $\log p(\boldsymbol{y})$	-128.7 ± 0.0	0.0	-1680.2 ± 0	0.0	-251.3 ± 0.0	0.0
Naive MC (1000 samples)	-3283 ± 130	2451%	-41289 ± 577	2357%	-6029 ± 170	2299%
Naive MC (10000 samples)	-3041 ± 113	2263%	-40123 ± 487	2288%	-5666 ± 144	2155%
Original DAPS heuristic	-316.2 ± 79.5	146%	-1735.4 ± 4.7	3.3%	-269.6 ± 3.3	7.3%
TI	-132.8 ± 1.3	3.2%	-1773.8 ± 11.8	5.6%	-254.2 ± 3.2	1.2%
AIS	-135.9 ± 2.5	5.6%	-1754.6 ± 7.1	4.4%	-259.1 ± 2.9	3.1%
SMC	-132.1 ± 1.0	2.6%	-1701.1 ± 5.8	1.2%	-253.0 ± 3.4	0.7%
DiME	-126.8 ± 2.8	1.5%	-1674.8 ± 4.7	0.3%	-253.3 ± 3.1	0.8%
DiME-PnPDM*	-131.8 ± 2.0	2.4%	-1912.1 ± 6.5	13.8%	-245.9 ± 3.6	2.1%

Table 1: Evidence estimates on Mixture of Gaussians study with analytic $\log p(\boldsymbol{y})$ and ground truth \boldsymbol{x}^* in-distribution (top), out-of-distribution (middle), and at a saddle point (bottom). Mean and standard deviation of 50 trials displayed each with 20 sample paths. When the analytic prior score is available, DiME performs comparably to the baselines for all cases, while DiME-PnPDM performs well when the ground truth is roughly in-distribution. ***Important:** as DiME-PnPDM outputs an unnormalized estimate, we normalize via the analytic constant factor for ease of comparison.

Results are displayed in Table 1. DiME gives nearly unbiased estimates of the evidence in all cases, with comparable performance to the strongest baseline (SMC), despite never using the true prior score. To probe the effect of the covariance choice discussed in Sec. 3.1, we first consider the case where \boldsymbol{x}^* lies within a prior mode (in-distribution). Here the original DAPS heuristic consistently overestimates the variance of the marginals, pushing posterior samples into the wrong mode and inducing a large bias in the evidence estimate. Even when \boldsymbol{x}^* is outside any single mode, the original heuristic continues to overestimate the evidence. In contrast, our improved covariance heuristic, incorporated in DiME, always generates samples from the correct mode and eliminates this bias, even when the true \boldsymbol{x}^* lies in regions of local curvature far from the unimodal Gaussian assumption.

Unlike DiME, DiME-PnPDM only performs well for roughly in-distribution tasks. For out-of-distribution measurements \boldsymbol{y} , while the PnP-DM marginals asymptotically anneal to the correct posterior, we find that in the finite-step case this annealing path results in greater bias than DAPS. Because out-of-distribution measurements frequently arise in model selection, we conclude that PnP-DM is less well suited for this task.

4.2 MODEL SELECTION USING LEARNED SCORE ON NON-CONVEX INVERSE PROBLEMS

We test both DiME and the SMC baseline on two non-convex inverse problems, Gaussian phase retrieval and Fourier phase retrieval. In Gaussian phase retrieval, the forward model is $\boldsymbol{y} = |\boldsymbol{A}\boldsymbol{x}| + \boldsymbol{\varepsilon}$, where $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ consists of i.i.d Gaussian $\mathcal{N}(0, \frac{1}{m})$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$. In Fourier phase retrieval, the forward model is $\boldsymbol{y} = |\boldsymbol{F}\boldsymbol{x}| + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ where \boldsymbol{F} is the discrete Fourier transform. In our experiments, we use $m = 0.1n$ and $\sigma = 0.3$. While Fourier phase retrieval has many more measurements, the structured forward operator means that both translation and flips of \boldsymbol{x} are invariances under the forward model. We test with ten diffusion model priors, one on each MNIST (Deng, 2012) digit, and compute the evidence for each prior given a measurement of each digit. For DiME, we use 50 annealing iterations and 20 generated sample-paths for each entry. For SMC, we use the learned $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t)$ as a surrogate prior score at different low noise levels.

DiME results are shown in Figure 2, and SMC results are displayed in Appendix E. For both inverse problems, DiME always selects the correct model given a single noisy measurement \boldsymbol{y} , while SMC often fails to do so, demonstrating that methods relying on a clean prior score are not suitable for diffusion-based model selection. For Gaussian phase retrieval, DiME predicts higher likelihoods of digits that look similar, such as 4 and 9 (Figure 2, bottom left, third row). For Fourier phase retrieval, DiME predicts high evidence for the MNIST 9 prior when the ground truth image is a 6 (and vice versa), but as both vertical and horizontal flips are invariances under Fourier phase retrieval, performing these operations on a 6 results in a 9 (Figure 2, bottom right, second row). The normalized evidence can also provide useful insights: as the MNIST digit 1 prior has higher evidence for in-distribution \boldsymbol{y} and lower evidence for out-of-distribution \boldsymbol{y} , it must be a narrower prior. On the other hand, the MNIST 3, 5 and 8 priors have higher out-of-distribution model evidence, suggesting that they are wider priors, likely due to the variation in which humans write these digits.

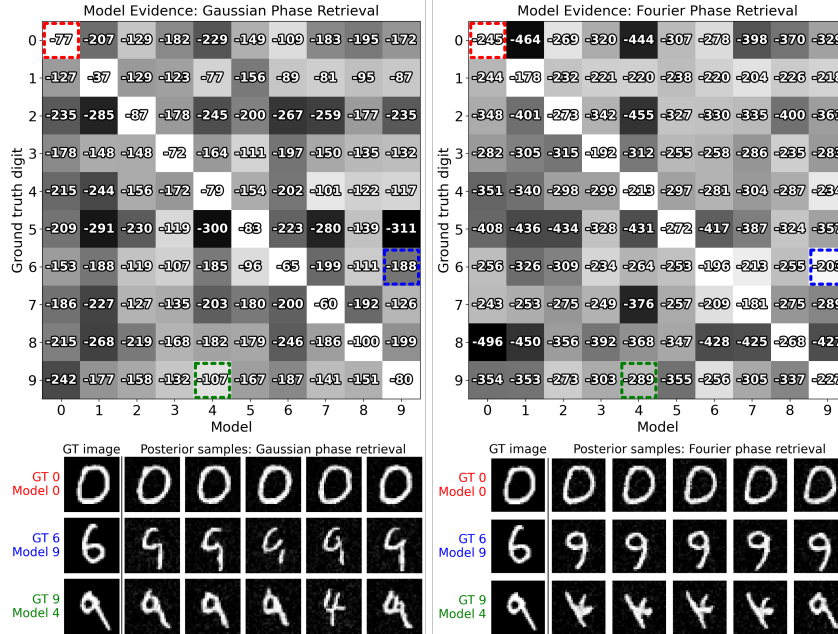


Figure 2: Model evidence confusion matrix for Gaussian phase retrieval (**left**) and Fourier phase retrieval (**right**) for each (ground truth measurement, model) pair of MNIST digits. Our method selects the correct model for all cases. Posterior samples shown for dotted matrix entries below. For Gaussian phase retrieval, DiME estimates higher model likelihood for visually similar digits, such as 4 and 9. For Fourier phase retrieval, both translations and reflections are invariances as seen in the posterior samples, so DiME estimates a high likelihood of model 9 given a measurement of a 6.

4.3 EVALUATING PRIOR MODEL EVIDENCE FOR REAL M87* BLACK HOLE DATA

In this section, we perform model selection and validation on real black hole M87* observations from April 6, 2017 (EHTC, 2019a;b;c). The need for such analysis was first emphasized during the release of the initial M87* images by the Event Horizon Telescope Collaboration (EHTC, 2019e), where the goal was not only to reconstruct images but also to test which simple parametric models best matched the data—particularly whether a ring provided the most suitable description (EHTC, 2019e; Broderick et al., 2020). In addition, the team directly compared the M87* data to a discrete set of synthetic black hole simulations (GRMHD) to assess which physical model parameters were most consistent with the observations (Fromm et al., 2019; EHTC, 2019e;d). However, as noted in EHTC (2019e), these conclusions are only valid to the extent that the selected simulations provide an adequate description of M87* – a question that can be answered by our proposed method.

Measurements, known as *visibilities*, are obtained by combining signals collected by telescopes observing simultaneously at different locations around the world. The forward model for complex visibilities between two telescopes i, j is $v_{i,j} = g_i g_j e^{i(\phi_i - \phi_j)} \mathbf{v}_{i,j}^* + \varepsilon_{i,j}$, where $\mathbf{v}_{i,j}^*$ is the true visibilities measurements, g_i is a station-dependent gain error, ϕ_i is a station-dependent phase error arising from atmospheric noise and $\varepsilon_{i,j}$ is Gaussian thermal noise. To become invariant to the phase error, we use the closure phase $\mathbf{y}_{cp} = \mathbf{v}_{i,j} \mathbf{v}_{j,k} \mathbf{v}_{k,i}$ for a minimal set of 3 telescopes (i, j, k) and for similar invariance to the gain error, we use the closure amplitude $\mathbf{y}_{ca} = \frac{\mathbf{v}_{i,j} \mathbf{v}_{k,l}}{\mathbf{v}_{i,k} \mathbf{v}_{j,l}}$ for a minimal set of 4 telescopes (i, j, k, l) (Thompson et al., 2017; Blackburn et al., 2020). We use the InverseBench (Zheng et al., 2025) library to compute the likelihood. We generate 20 posterior samples using 50 annealing steps and 1000 Langevin dynamics steps per iteration. Gaussian approximation DAPS takes around 4 minutes to compute, while exact DAPS takes around 30 minutes to compute for the same parameters, a speedup of around 7x.

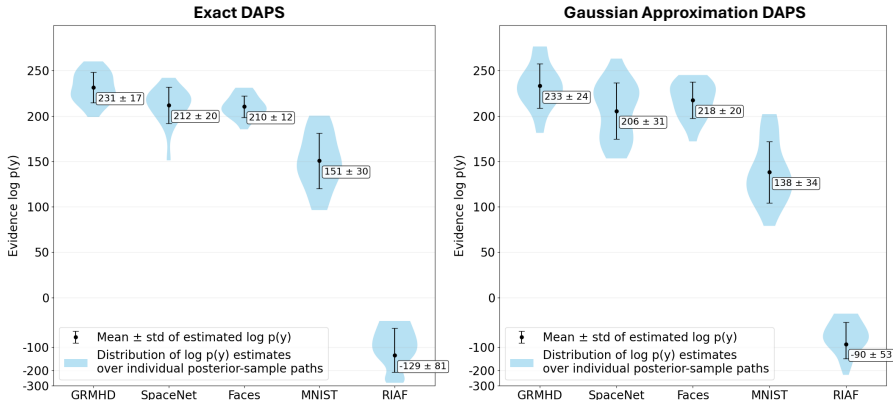


Figure 3: Model evidence estimates on real M87* observations across 5 different priors using exact DAPS (left) and Gaussian approximation DAPS (right). Our method concludes that, of these prior models, GRMHD is the most likely model. The violin plots show the distribution of evidences from 20 different posterior-sample paths from the DAPS sampling process, and the mean is the overall evidence estimate. Gaussian approximation DAPS gives highly accurate evidence estimates with 7x less compute than exact DAPS, but with slightly higher variance.

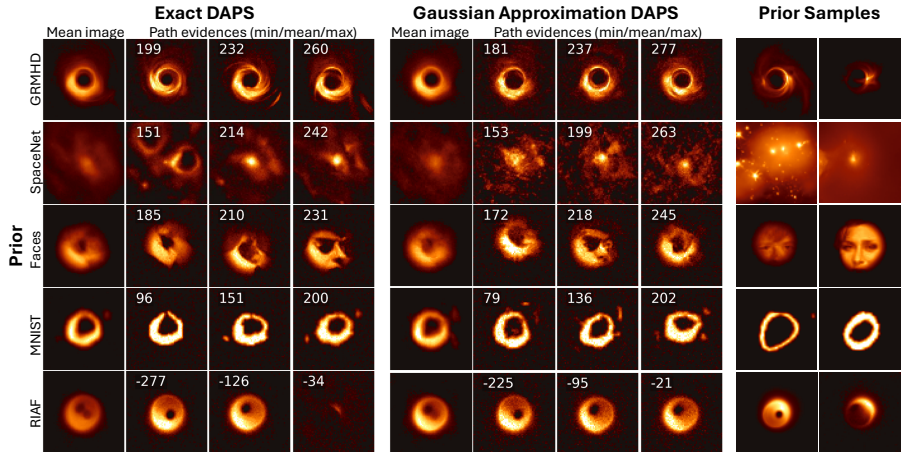


Figure 4: M87* reconstructions using the 5 priors using exact DAPS (left) and Gaussian approximation DAPS (middle). While all posterior samples have a data fit of around reduced $\chi^2 \approx 1$, they have varying path evidence estimates (shown in the top left of each image). Right: Example unconditional samples from each prior.

4.3.1 MODEL SELECTION

We trained five candidate diffusion priors: (1) on 48000 General-Relativistic Magnetohydrodynamic (GRMHD) simulations (Mizuno, 2022), (2) on 9070 Radiatively Inefficient Accretion Flow (RIAF) simulations (Broderick et al., 2011), (3) on 11000 general space images (SpaceNet) (Alam et al., 2024), (4) on 5923 MNIST (Deng, 2012) digit 0 images, and (5) on 48000 radially tapered CelebA Liu et al. (2015) images (Faces). Additionally, we augment our data during training to improve generality and robustness of our priors. The GRMHD images are randomly flipped horizontally and/or vertically and zoomed in and out in the range $[0.75, 1.25]$. The RIAF images were randomly flipped, zoomed in and out in the range $[0.75, 1.25]$, and rotated uniformly in the range $[0, 2\pi]$. The MNIST digit 0 images were zoomed in and out in the range $[0.5, 1]$.

Model selection results for the real M87* data can be seen in Figure 3. The mean reconstruction and posterior samples with the minimum, mean and maximum path evidence are shown on in Figure 4. Synthetic tests validating the method on this problem are shown in Appendix F. DiME determines that, for static imaging, GRMHD is the most likely prior for the M87* observations, followed by

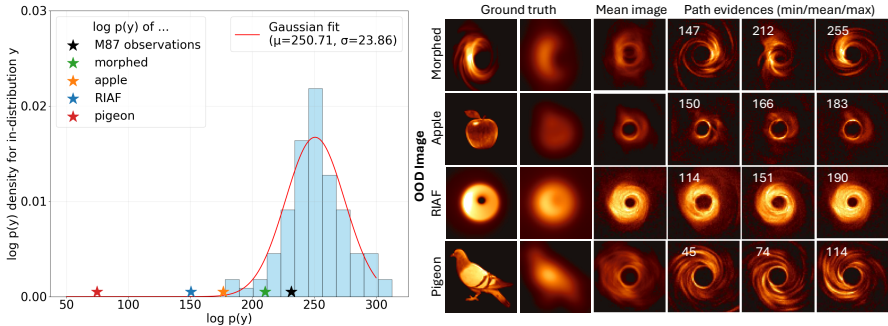


Figure 5: **Left:** GRMHD model validation results on M87* observations by comparing to evidence of in-distribution measurements y . Our method shows that the evidence of M87* observations have a z -score of about -0.81 compared to the evidence distribution of GRMHD measurements, indicating that M87* is statistically in-distribution of GRMHD. The evidence of simulated measurements of out-of-distribution images are also shown, demonstrating that measurements from out-of-distribution (OOD) images have OOD evidence. **Right:** Mean reconstruction and posterior samples of the OOD images with the highest and lowest distance to the GRMHD prior. The blurred ground truth image, or the maximum resolution that can be obtained from measurements, is also displayed.

SpaceNet, Faces, MNIST, and finally RIAF. While SpaceNet contains a few black hole images in its training data, our method penalizes it for being too general, as demonstrated with its wide range of unconditional samples (Figure 3, right, second row). The RIAF model is a very narrow prior that enforces a strong assumption on the structure of black holes, so any out-of-distribution images (such as M87* as determined by our method) will have low likelihood. We also find that Gaussian approximation DAPS gives nearly identical estimates as exact DAPS with 7x less compute.

4.3.2 MODEL VALIDATION

We also diagnose the validity of the GRMHD model for static imaging of the M87* observations by computing the evidence $\log p(y)$. We simulated measurements from 100 GRMHD test images using the same forward model as the EHT observations of M87*, and then computed the evidence for these measurements. Under a central limit theorem argument, we can approximate this distribution as Gaussian. We then compared the evidence of the true M87* observation to this Gaussian as well as the evidence of out-of-distribution measurements under the same forward model to assess if the GRMHD prior is valid for M87*.

Results are shown in Figure 5. The fitted Gaussian has $\mu = 250.71$ and $\sigma = 23.86$. M87* has an evidence of 231.46, giving it a z -score of -0.81 and p -value of 0.209 with respect to the previous Gaussian fit. Therefore, M87* likely lies within the support of the GRMHD image prior, offering credence to our current physical model of black holes through the images it produces, while still leaving open the possibility of an improved model. On the other hand, the morphed GRMHD image has evidence 210.37, giving it a z -score of -1.69 and p -value of 0.046, suggesting that it is not in the true GRMHD distribution. The apple, RIAF, and pigeon images have even lower evidences of 176.52, 150.41, and 74.68 respectively and are all considered out-of-distribution.

5 CONCLUSION

We introduce DiME, a method that estimates the Bayesian model evidence when solving inverse problems using diffusion model priors. DiME computes the KL divergence between the posterior and prior by integrating along posterior time-marginals, a task made easy by using the intermediate samples of recent diffusion posterior sampling methods. Empirically, DiME outperforms all baseline methods and we demonstrate how DiME can be used on a real-world black hole imaging problem. Altogether, DiME makes it possible to harness diffusion priors not only for reconstruction but also for principled model selection and validation, laying the foundation for more reliable inference in scientific imaging.

ETHICS STATEMENT

Our method is a statistical estimator and can occasionally select the incorrect model, potentially leading to incorrect conclusions. Furthermore, while it could in principle be applied to sensitive characteristics (e.g., facial recognition), we do not support its use in this context.

REPRODUCIBILITY STATEMENT

All code and trained models are available at <https://github.com/fredwang25/DiME>. All mathematical derivations are detailed in the appendix. The MNIST dataset is available through torchvision (Marcel & Rodriguez, 2010), the CelebA dataset is available at <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>, and the SpaceNet dataset is available at <https://www.kaggle.com/datasets/razaimam45/spacenet-an-optimally-distributed-astronomy-data/>. The RIAF and GRMHD datasets are not currently available for public use.

ACKNOWLEDGMENTS

This work was supported by NSF Award 2048237, an Amazon AI4Science Discovery Award, OpenAI, and a Sloan Research Fellowship. FW is supported by a Kortschak Fellowship. The authors would like to thank Angela Gao for helpful discussions.

REFERENCES

- Mohammed Talha Alam, Raza Imam, Mohsen Guizani, and Fakhri Karray. Flare up your data: Diffusion-based augmentation method in astronomical imaging, 2024.
- Lindy Blackburn, Dominic W Pesce, Michael D Johnson, Maciek Wielgus, Andrew A Chael, Pierre Christian, and Sheperd S Doeleman. Closure statistics in interferometric data. *The Astrophysical Journal*, 894(1):31, 2020.
- Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *arXiv preprint arXiv:2310.06721*, 2023.
- Avery E Broderick, Vincent L Fish, Sheperd S Doeleman, and Abraham Loeb. Evidence for low black hole spin and physically motivated accretion models from millimeter-vlbi observations of sagittarius a. *The Astrophysical Journal*, 735(2):110, 2011.
- Avery E Broderick, Roman Gold, Mansour Karami, Jorge A Preciado-Lopez, Paul Tiede, Hung-Yi Pu, et al. Themis: A parameter estimation framework for the event horizon telescope. *The Astrophysical Journal*, 897(2):139, 2020.
- Xiaohao Cai, Jason D McEwen, and Marcelo Pereyra. Proximal nested sampling for high-dimensional bayesian model selection. *Statistics and Computing*, 32(5):87, 2022.
- Gabriel Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided diffusion for bayesian linear inverse problems. *arXiv preprint arXiv:2308.07983*, 2023.
- Junhua Chen, Lorenz Richter, Julius Berner, Denis Blessing, Gerhard Neumann, and Anima Anandkumar. Sequential controlled langevin diffusions. *arXiv preprint arXiv:2412.07081*, 2024.
- Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.
- Florentin Coeurdoux, Nicolas Dobigeon, and Pierre Chainais. Plug-and-play split gibbs sampler: embedding deep generative priors in bayesian inference. *IEEE Transactions on Image Processing*, 33:3496–3507, 2024.
- Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3):411–436, 2006.
- Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

- Arnaud Doucet, Will Grathwohl, Alexander G Matthews, and Heiko Strathmann. Score-based diffusion meets annealed importance sampling. *Advances in Neural Information Processing Systems*, 35:21482–21494, 2022.
- Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.
- EHTC. First m87 eht results: Calibrated data, 2019a.
- EHTC. First m87 event horizon telescope results. i. the shadow of the supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L1, 2019b.
- EHTC. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875(1):L4, 2019c.
- EHTC. First m87 event horizon telescope results. v. physical origin of the asymmetric ring. *The Astrophysical Journal Letters*, 875(1):L5, 2019d.
- EHTC. First m87 event horizon telescope results. vi. the shadow and mass of the central black hole. *The Astrophysical Journal Letters*, 875(1):L6, 2019e.
- Berthy T Feng, Katherine L Bouman, and William T Freeman. Event-horizon-scale imaging of m87* under different assumptions via deep generative image priors. *The Astrophysical Journal*, 975(2):201, 2024.
- CM Fromm, Z Younsi, A Baczko, Y Mizuno, O Porth, M Perucho, H Olivares, A Nathanail, E Angelakis, E Ros, et al. Using evolutionary algorithms to model relativistic jets-application to ngc 1052. *Astronomy & Astrophysics*, 629:A4, 2019.
- Wei Guo, Molei Tao, and Yongxin Chen. Complexity analysis of normalizing constant estimation: from jarzynski equality to annealed importance sampling and beyond. *arXiv preprint arXiv:2502.04575*, 2025.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Xunpeng Huang, Hanze Dong, Yifan Hao, Yi-An Ma, and Tong Zhang. Reverse diffusion monte carlo. *arXiv preprint arXiv:2307.02037*, 2023.
- 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.
- Nicolas Lartillot and Hervé Philippe. Computing bayes factors using thermodynamic integration. *Systematic biology*, 55(2):195–207, 2006.
- Xiang Li, Soo Min Kwon, Shijun Liang, Ismail R Alkhouri, Saiprasad Ravishankar, and Qing Qu. Decoupled data consistency with diffusion purification for image restoration. *arXiv preprint arXiv:2403.06054*, 2024.
- Julia Linhart, Gabriel Victorino Cardoso, Alexandre Gramfort, Sylvain Le Corff, and Pedro LC Rodrigues. Diffusion posterior sampling for simulation-based inference in tall data settings. *arXiv preprint arXiv:2404.07593*, 2024.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Sébastien Marcel and Yann Rodriguez. Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010.
- Morteza Mardani, Jiaming Song, Jan Kautz, and Arash Vahdat. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

- Jason D McEwen, Christopher GR Wallis, Matthew A Price, and Alessio Spurio Mancini. Machine learning assisted bayesian model comparison: learnt harmonic mean estimator. *arXiv preprint arXiv:2111.12720*, 2021.
- Yosuke Mizuno. Grmhd simulations and modeling for jet formation and acceleration region in agns. *Universe*, 8(2):85, 2022.
- Badr Moufad, Yazid Janati, Lisa Bedin, Alain Durmus, Randal Douc, Eric Moulines, and Jimmy Olsson. Variational diffusion posterior sampling with midpoint guidance. *arXiv preprint arXiv:2410.09945*, 2024.
- Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.
- Michael A Newton and Adrian E Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 56(1): 3–26, 1994.
- Xinyu Peng, Ziyang Zheng, Wenrui Dai, Nuoqian Xiao, Chenglin Li, Junni Zou, and Hongkai Xiong. Improving diffusion models for inverse problems using optimal posterior covariance. *arXiv preprint arXiv:2402.02149*, 2024.
- Alicja Polanska, Matthew A Price, Alessio Spurio Mancini, and Jason D McEwen. Learned harmonic mean estimation of the marginal likelihood with normalizing flows. In *Physical sciences forum*, volume 9, pp. 10. MDPI, 2023.
- John Skilling. Nested sampling for general bayesian computation. 2006.
- Marta Skreta, Lazar Atanackovic, Avishek Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the it^o density estimator. *arXiv preprint arXiv:2412.17762*, 2024.
- 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, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in neural information processing systems*, 34:1415–1428, 2021a.
- Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021b.
- A Spurio Mancini, MM Docherty, MA Price, and JD McEwen. Bayesian model comparison for simulation-based inference. *RAS Techniques and Instruments*, 2(1):710–722, 2023.
- Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.
- A Richard Thompson, James M Moran, and George W Swenson. *Interferometry and synthesis in radio astronomy*. Springer Nature, 2017.
- Roberto Trotta. Applications of bayesian model selection to cosmological parameters. *Monthly Notices of the Royal Astronomical Society*, 378(1):72–82, 2007.
- Maxime Vono, Nicolas Dobigeon, and Pierre Chainais. Split-and-augmented gibbs sampler—application to large-scale inference problems. *IEEE Transactions on Signal Processing*, 67(6):1648–1661, 2019.
- Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *Advances in Neural Information Processing Systems*, 36:31372–31403, 2023.

Zihui Wu, Yu Sun, Yifan Chen, Bingliang Zhang, Yisong Yue, and Katherine Bouman. Principled probabilistic imaging using diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 37:118389–118427, 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 Computer Vision and Pattern Recognition Conference*, pp. 20895–20905, 2025.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy T Feng, Caifeng Zou, Yu Sun, Nikola Kovachki, Zachary E Ross, et al. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. *arXiv preprint arXiv:2503.11043*, 2025.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhong 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*, pp. 1219–1229, 2023.

A PROOFS

Although the main text uses the diffusion SDE formulation, we introduce the Probability Flow ODE (PF-ODE) for the proofs and for Appendix B. This ODE has identical marginals as the SDE under ideal conditions. Given diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$, the PF-ODE is:

$$d\mathbf{x}_t = \left[\frac{a'_t}{a_t} \mathbf{x}_t - \left(\sigma_t \sigma'_t - \frac{a'_t}{a_t} \sigma_t^2 \right) \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \right] dt = \mathbf{v}_p(\mathbf{x}_t) dt \quad (13)$$

where $\mathbf{v}_p(\mathbf{x}_t)$ is the unconditional flow. The conditional flow $\mathbf{v}_p(\mathbf{x}_t | \mathbf{y})$ can be derived via Bayes:

$$\mathbf{v}_p(\mathbf{x}_t | \mathbf{y}) = \frac{a'_t}{a_t} \mathbf{x}_t - \left(\sigma_t \sigma'_t - \frac{a'_t}{a_t} \sigma_t^2 \right) \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{y}) \quad (14)$$

$$= \mathbf{v}_p(\mathbf{x}_t) - \left(\sigma'_t \sigma_t - \sigma_t^2 \frac{a'_t}{a_t} \right) \nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t). \quad (15)$$

For the following proof, we impose the regularity and smoothness assumptions listed in Appendix A of Song et al. (2021a) for both the prior marginals $p(\mathbf{x}_t)$ and the posterior marginals $\pi(\mathbf{x}_t | \mathbf{y})$.

Lemma 3. *Let $p(\mathbf{x}_t)$ be the unconditional diffusion marginal and $\pi(\mathbf{x}_t | \mathbf{y}) \propto p(\mathbf{x}_t) \pi(\mathbf{y} | \mathbf{x}_t)$ be a posterior time-marginal obtained at diffusion timestep t . Denote $\mathbf{v}_p(\mathbf{x}_t)$ and $\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y})$ as their corresponding velocity vector fields. We have the following identity:*

$$\frac{d}{dt} D_{\text{KL}}(\pi(\mathbf{x}_t | \mathbf{y}) \| p(\mathbf{x}_t)) = \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_t | \mathbf{y})} \left[(\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y}) - \mathbf{v}_p(\mathbf{x}_t))^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) \right]. \quad (16)$$

Proof. Simplifying the derivative gives:

$$\begin{aligned} \frac{d}{dt} D_{\text{KL}}(\pi(\mathbf{x}_t | \mathbf{y}) \| p(\mathbf{x}_t)) &= \frac{d}{dt} \int \pi(\mathbf{x}_t | \mathbf{y}) [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t \\ &= \int \left[\frac{d}{dt} \pi(\mathbf{x}_t | \mathbf{y}) \right] [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t \\ &\quad + \int \pi(\mathbf{x}_t | \mathbf{y}) \frac{d}{dt} [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t. \quad (\text{Product rule}) \end{aligned}$$

Now we expand the first term. Let $\mathbf{v}_p(\mathbf{x}_t)$, $\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y})$ be the flows at $p(\mathbf{x}_t)$, $\pi(\mathbf{x}_t | \mathbf{y})$ respectively:

$$\begin{aligned} &\int \left[\frac{d}{dt} \pi(\mathbf{x}_t | \mathbf{y}) \right] [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t \\ &= - \int \nabla \cdot (\pi(\mathbf{x}_t | \mathbf{y}) \mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y})) [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t \quad (\text{Continuity equation}) \\ &= \int \pi(\mathbf{x}_t | \mathbf{y}) \mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y})^T \nabla_{\mathbf{x}_t} [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t \quad (\text{Integration by parts}) \\ &= \int \pi(\mathbf{x}_t | \mathbf{y}) \mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y})^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) d\mathbf{x}_t. \quad (\text{Bayes}) \end{aligned}$$

For the second term:

$$\begin{aligned}
& \int \pi(\mathbf{x}_t | \mathbf{y}) \frac{d}{dt} [\log \pi(\mathbf{x}_t | \mathbf{y}) - \log p(\mathbf{x}_t)] d\mathbf{x}_t \\
&= \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_t | \mathbf{y})} \left[\frac{d}{dt} \log \pi(\mathbf{x}_t | \mathbf{y}) \right] - \int \pi(\mathbf{x}_t | \mathbf{y}) \frac{d}{dt} \log p(\mathbf{x}_t) d\mathbf{x}_t \\
&= - \int \pi(\mathbf{x}_t | \mathbf{y}) \frac{d}{dt} \log p(\mathbf{x}_t) d\mathbf{x}_t \\
&= \int \pi(\mathbf{x}_t | \mathbf{y}) [\nabla \cdot \mathbf{v}_p(\mathbf{x}_t) + \mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)] d\mathbf{x}_t \quad (\text{Log-continuity equation}) \\
&= \int \pi(\mathbf{x}_t | \mathbf{y}) [\nabla \cdot \mathbf{v}_p(\mathbf{x}_t) + \mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{x}_t | \mathbf{y})] d\mathbf{x}_t \\
&\quad - \int \pi(\mathbf{x}_t | \mathbf{y}) \mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) d\mathbf{x}_t \quad (\text{Bayes}) \\
&= - \int \pi(\mathbf{x}_t | \mathbf{y}) \mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) d\mathbf{x}_t. \quad (\text{Stein's identity})
\end{aligned}$$

Adding the two terms back together gives us:

$$\begin{aligned}
\frac{d}{dt} D_{\text{KL}}(\pi(\mathbf{x}_t | \mathbf{y}) || p(\mathbf{x}_t)) &= \int \pi(\mathbf{x}_t | \mathbf{y}) (\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y}) - \mathbf{v}_p(\mathbf{x}_t))^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) d\mathbf{x}_t \\
&= \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_t | \mathbf{y})} \left[(\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y}) - \mathbf{v}_p(\mathbf{x}_t))^T \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) \right].
\end{aligned}$$

□

Proposition 1 (DIME: Model evidence estimation along the standard marginals). *Given diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$, posterior marginals $p(\mathbf{x}_t | \mathbf{y}) \propto p(\mathbf{x}_t) p(\mathbf{y} | \mathbf{x}_t)$, and timesteps $0 = t_0 < \dots < t_N = T$, the model evidence can be estimated via:*

$$\log p(\mathbf{y}) = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{y} | \mathbf{x}_0)] - D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)). \quad (8)$$

The log-likelihood term can be estimated with posterior samples³, and the KL divergence with:

$$D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)) \approx \sum_{i=1}^N c_{t_i} \Delta t_i \mathbb{E}_{\mathbf{x}_{t_i} \sim p(\mathbf{x}_{t_i} | \mathbf{y})} \|\nabla_{\mathbf{x}_{t_i}} \log p(\mathbf{y} | \mathbf{x}_{t_i})\|^2 \quad (9)$$

where $c_{t_i} = \sigma'_{t_i} \sigma_{t_i} - \sigma_{t_i}^2 \frac{a'_{t_i}}{a_{t_i}}$ and $\Delta t_i = t_i - t_{i-1}$ are derived from the diffusion schedule.

Proof. Expanding out the KL divergence gives us an expression for the model evidence:

$$\begin{aligned}
D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)) &= \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{x}_0 | \mathbf{y}) - \log p(\mathbf{x}_0)] \\
&= \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{y} | \mathbf{x}_0) - \log p(\mathbf{y})] \\
&= \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{y} | \mathbf{x}_0)] - \log p(\mathbf{y}) \\
&\implies \log p(\mathbf{y}) = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{y} | \mathbf{x}_0)] - D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)).
\end{aligned}$$

³For likelihoods with Gaussian noise, if the posterior and prior have enough overlap such that posterior samples have a data fit of $\chi^2 \approx 1$, this term essentially becomes constant under the same likelihood function.

At the final diffusion step, the forward process has destroyed all information about the measurements, so we have $p(\mathbf{x}_T | \mathbf{y}) = p(\mathbf{x}_T)$, or $D_{\text{KL}}(p(\mathbf{x}_T | \mathbf{y}) || p(\mathbf{x}_T)) = 0$. Therefore, we have:

$$\begin{aligned}
D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)) &= - \int_0^T \frac{d}{dt} D_{\text{KL}}(p(\mathbf{x}_t | \mathbf{y}) || p(\mathbf{x}_t)) dt \\
&\stackrel{(i)}{=} - \int_0^T \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{y})} \left[(\mathbf{v}_p(\mathbf{x}_t | \mathbf{y}) - \mathbf{v}_p(\mathbf{x}_t))^\top \nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t) \right] dt \\
&\stackrel{(ii)}{=} \int_0^T \left(\sigma'_t \sigma_t - \sigma_t^2 \frac{a'_t}{a_t} \right) \mathbb{E}_{\mathbf{x}_t \sim p(\mathbf{x}_t | \mathbf{y})} \left\| \nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t) \right\|^2 dt. \\
&\approx \sum_{i=1}^N \Delta t_i \left(\sigma'_{t_i} \sigma_{t_i} - \sigma_{t_i}^2 \frac{a'_{t_i}}{a_{t_i}} \right) \mathbb{E}_{\mathbf{x}_{t_i} \sim p(\mathbf{x}_{t_i} | \mathbf{y})} \left\| \nabla_{\mathbf{x}_{t_i}} \log p(\mathbf{y} | \mathbf{x}_{t_i}) \right\|^2.
\end{aligned}$$

where (i) uses Lemma 3 but with $\pi(\mathbf{x}_t | \mathbf{y}) = p(\mathbf{x}_t | \mathbf{y})$ and (ii) uses the PF-ODE of the conditional flow (Eq. 15). \square

A.1 PROOFS AND DERIVATIONS FOR THE DAPS-BASED IMPLEMENTATION

Lemma 1. *Suppose we have diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$. Let $\mathbf{x}_t \sim p(\mathbf{x}_t)$ and suppose $p(\mathbf{x}_0) \approx \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Then we can express the posterior covariance $\boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t}$ as*

$$\boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t} = \left[\boldsymbol{\Sigma}_0^{-1} + \frac{a_t^2}{\sigma_t^2} \mathbf{I} \right]^{-1}. \quad (10)$$

Proof. Our diffusion corruption process gives us:

$$\begin{aligned} \text{Cov}(\mathbf{x}_t) &= a_t^2 \boldsymbol{\Sigma}_0 + \sigma_t^2 \mathbf{I} \\ \text{Cov}(\mathbf{x}_0, \mathbf{x}_t) &= \text{Cov}(\mathbf{x}_0, a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t) = a_t \text{Cov}(\mathbf{x}_0) = a_t \boldsymbol{\Sigma}_0. \end{aligned}$$

Plugging these into the Gaussian posterior covariance formula gives:

$$\begin{aligned} \text{Cov}(\mathbf{x}_0 | \mathbf{x}_t) &= \text{Cov}(\mathbf{x}_0) - \text{Cov}(\mathbf{x}_0, \mathbf{x}_t) \text{Cov}(\mathbf{x}_t)^{-1} \text{Cov}(\mathbf{x}_t, \mathbf{x}_0) \\ &= \boldsymbol{\Sigma}_0 - a_t \boldsymbol{\Sigma}_0 \left[a_t^2 \boldsymbol{\Sigma}_0 + \sigma_t^2 \mathbf{I} \right]^{-1} a_t \boldsymbol{\Sigma}_0 \\ &= \boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0 \left[\boldsymbol{\Sigma}_0 + \frac{\sigma_t^2}{a_t^2} \mathbf{I} \right]^{-1} \boldsymbol{\Sigma}_0 \\ &= \left[\boldsymbol{\Sigma}_0^{-1} + \frac{a_t^2}{\sigma_t^2} \mathbf{I} \right]^{-1}. \end{aligned} \quad (\text{Woodbury identity}).$$

□

Corollary 1. *Suppose the training data has $\lambda_{\max}(\boldsymbol{\Sigma}_0) = 1$. Then $\|\boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t}\| = \frac{\sigma_t^2}{\sigma_t^2 + a_t^2}$.*

Proof. Plugging in the covariance from Lemma 1 gives:

$$\|\boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t}\| = \lambda_{\max} \left[\boldsymbol{\Sigma}_0^{-1} + \frac{a_t^2}{\sigma_t^2} \mathbf{I} \right]^{-1} = \frac{1}{\lambda_{\min} \left(\boldsymbol{\Sigma}_0^{-1} + \frac{a_t^2}{\sigma_t^2} \mathbf{I} \right)} = \frac{1}{1 + \frac{a_t^2}{\sigma_t^2}} = \frac{\sigma_t^2}{\sigma_t^2 + a_t^2}.$$

□

Lemma 2. *Suppose we have diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$ and sample $\tilde{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y})$. Under the assumption that $p(\mathbf{x}_0 | \mathbf{x}_t) \approx \mathcal{N}(\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t], \boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t})$, both $\Theta_{\text{high}}(\tilde{\mathbf{x}}_0)$ and $\Theta_{\text{low}}(\tilde{\mathbf{x}}_0)$ are unbiased estimators of $\nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t)$ with*

$$\Theta_{\text{high}}(\tilde{\mathbf{x}}_0) = \frac{a_t}{\sigma_t^2} (\tilde{\mathbf{x}}_0 - \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]). \quad (11)$$

$$\Theta_{\text{low}}(\tilde{\mathbf{x}}_0) = \frac{a_t}{\sigma_t^2} (\boldsymbol{\Sigma}_{\mathbf{x}_0 | \mathbf{x}_t} \nabla_{\tilde{\mathbf{x}}_0} \log p(\mathbf{y} | \tilde{\mathbf{x}}_0)). \quad (12)$$

Under low diffusion noise ($a_t \rightarrow 1, \sigma_t \rightarrow 0$), $\text{Var}(\Theta_{\text{low}}) \rightarrow 0$ but $\text{Var}(\Theta_{\text{high}}) = O(\frac{1}{\sigma_t^2})$.

Under high noise ($a_t \rightarrow 0, \sigma_t \rightarrow 1$), $\text{Var}(\Theta_{\text{high}}) \rightarrow 0$ but $\text{Var}(\Theta_{\text{low}}) = O(\frac{\sigma_t^4}{\sigma_t^2} \text{Var}(\Theta_{\text{high}}))$.

Proof. Using Tweedie's rule, we can show that Θ_{high} is unbiased for the likelihood score:

$$\begin{aligned} \nabla_{\mathbf{x}_t} \log p(\mathbf{y} | \mathbf{x}_t) &= \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t | \mathbf{y}) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \\ &= \frac{a_t}{\sigma_t^2} \left(\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}] - \frac{\mathbf{x}_t}{a_t} \right) - \frac{a_t}{\sigma_t^2} \left(\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t] - \frac{\mathbf{x}_t}{a_t} \right) \\ &= \frac{a_t}{\sigma_t^2} (\mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}] - \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]). \end{aligned} \quad (\star)$$

Our Gaussian approximation of $p(\mathbf{x}_0 | \mathbf{x}_t)$ gives us $\nabla_{\mathbf{x}_0} \log p(\mathbf{x}_0 | \mathbf{x}_t) = -\Sigma_{\mathbf{x}_0 | \mathbf{x}_t}^{-1} (\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t])$. Using the fact that the expectation of the score is zero, we have:

$$\begin{aligned} \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}] &= \mathbb{E}[\mathbf{x}_0 + \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} \nabla_{\mathbf{x}_0} \log p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}) | \mathbf{x}_t, \mathbf{y}] \\ &= \mathbb{E}[\mathbf{x}_0 + \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} (\nabla_{\mathbf{x}_0} \log p(\mathbf{x}_0 | \mathbf{x}_t) + \nabla_{\mathbf{x}_0} \log p(\mathbf{y} | \mathbf{x}_0)) | \mathbf{x}_t, \mathbf{y}] \\ &= \mathbb{E}[\mathbf{x}_0 - (\mathbf{x}_0 - \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t]) + \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} \nabla_{\mathbf{x}_0} \log p(\mathbf{y} | \mathbf{x}_0) | \mathbf{x}_t, \mathbf{y}] \\ &= \mathbb{E}[\mathbf{x}_0 | \mathbf{x}_t] + \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y})} [\nabla_{\mathbf{x}_0} \log p(\mathbf{y} | \mathbf{x}_0)]. \end{aligned} \quad (**)$$

Substituting (**) into (*) shows that Θ_{low} is unbiased. To compute a variance bound for Θ_{high} ,

$$\begin{aligned} \text{Var}(\Theta_{high}) &= \frac{a_t^2}{\sigma_t^4} \text{Cov}(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}) \\ &\leq \frac{a_t^2}{\sigma_t^4} \text{Cov}(\mathbf{x}_0 | \mathbf{x}_t) \quad (\text{Gaussian conditional covariance}) \\ &\leq \frac{a_t^2}{\sigma_t^2(\sigma_t^2 + a_t^2)}. \quad (\text{Corollary 1}) \end{aligned}$$

The term $\sigma_t^2 + a_t^2$ is of constant order over the diffusion process, resulting in a variance of a_t^2/σ_t^2 .

To compute a bound for Θ_{low} , we linearize the likelihood f at a fixed $\bar{\mathbf{x}}$ with $\mathbf{A} = \left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\bar{\mathbf{x}}}$ and get

$$\begin{aligned} \text{Var}(\Theta_{low}) &= \frac{a_t^2}{\sigma_t^4} \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} \text{Cov}(\nabla_{\mathbf{x}_0} \log p(\mathbf{y} | \mathbf{x}_0) | \mathbf{x}_t, \mathbf{y}) \Sigma_{\mathbf{x}_0 | \mathbf{x}_t}^T \\ &= \frac{a_t^2}{\sigma_t^4} \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} \text{Cov}\left(\frac{1}{\sigma_{\mathbf{y}}^2} \mathbf{A}^T (\mathbf{y} - \mathbf{A} \mathbf{x}_0) | \mathbf{x}_t, \mathbf{y}\right) \Sigma_{\mathbf{x}_0 | \mathbf{x}_t}^T \\ &= \frac{1}{\sigma_{\mathbf{y}}^4} \Sigma_{\mathbf{x}_0 | \mathbf{x}_t} \mathbf{A}^T \mathbf{A} \left(\frac{a_t^2}{\sigma_t^4} \text{Cov}(\mathbf{x}_0 | \mathbf{x}_t, \mathbf{y}) \right) \mathbf{A}^T \mathbf{A} \Sigma_{\mathbf{x}_0 | \mathbf{x}_t}^T \\ &\leq \frac{\|\mathbf{A}\|^4}{\sigma_{\mathbf{y}}^4} \frac{\sigma_t^4}{(\sigma_t^2 + a_t^2)^2} \text{Var}(\Theta_{high}). \quad (\text{Corollary 1}) \\ &\leq \frac{\|\mathbf{A}\|^4}{\sigma_{\mathbf{y}}^4} \frac{\sigma_t^2 a_t^2}{(\sigma_t^2 + a_t^2)^3}. \end{aligned}$$

Unlike Θ_{high} , the variance of this estimator goes to 0 as $\sigma_t \rightarrow 0$, but for the rest of diffusion process the constant factor makes it a higher variance estimator than Θ_{high} in practice. \square

B GENERALIZING DiME TO ARBITRARY POSTERIOR MARGINALS

While we only present DiME along the standard marginals in the main text, DiME can be extended to any arbitrary diffusion posterior marginals that anneal to the true posterior as follows:

Corollary 2 (Generalized DiME). *Let $p(\mathbf{x}_t)$ be the unconditional diffusion marginal and $\pi(\mathbf{x}_t | \mathbf{y}) \propto p(\mathbf{x}_t)\pi(\mathbf{y} | \mathbf{x}_t)$ be a path of marginals that anneals to the true posterior, or $\pi(\mathbf{x}_0 | \mathbf{y}) = p(\mathbf{x}_0 | \mathbf{y})$. Denote $\mathbf{v}_p(\mathbf{x}_t)$ and $\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y})$ as their corresponding velocity fields. Given reverse diffusion timesteps $0 = t_0 < \dots < t_N = T$, we have:*

$$\log p(\mathbf{y}) = \mathbb{E}_{\mathbf{x}_0 \sim p(\mathbf{x}_0 | \mathbf{y})} [\log p(\mathbf{y} | \mathbf{x}_0)] - D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0))$$

where

$$D_{\text{KL}}(p(\mathbf{x}_0 | \mathbf{y}) || p(\mathbf{x}_0)) \approx - \sum_{i=1}^N \Delta t_i \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_t | \mathbf{y})} \left[(\mathbf{v}_\pi(\mathbf{x}_t | \mathbf{y}) - \mathbf{v}_p(\mathbf{x}_t))^\top \nabla_{\mathbf{x}_t} \log \pi(\mathbf{y} | \mathbf{x}_t) \right] + D_{\text{KL}}(\pi(\mathbf{x}_T | \mathbf{y}) || p(\mathbf{x}_T)).$$

Proof. Follows directly from the proof of Prop. 1. \square

While \mathbf{v}_p can be directly computed using the PF-ODE (Eq. 13), computing \mathbf{v}_π can be difficult for arbitrary time-marginals π . By crafting a sequence of diffusion marginals with known velocity, we can avoid any approximation like the one used for $p(\mathbf{x}_0 | \mathbf{x}_t)$ in the DAPS implementation. The marginals $\pi(\mathbf{x}_t | \mathbf{y})$ should also ideally have high overlap with the unconditional marginal $p(\mathbf{x}_t)$ to ensure accurate computation of $\mathbf{v}_p(\mathbf{x}_t)$ when the score is learned. If $\pi(\mathbf{x}_T | \mathbf{y}) \neq p(\mathbf{x}_T)$, then we must use other estimation methods to compute $D_{\text{KL}}(\pi(\mathbf{x}_T | \mathbf{y}) || p(\mathbf{x}_T))$ or settle for an unnormalized model evidence estimate. The following section discusses two estimators that does not require any approximation, but are not normalized.

B.1 ESTIMATING THE MODEL EVIDENCE FROM ALTERNATIVE SAMPLING METHODS

Algorithm 2 One posterior sample-path of DiME-PnPDM

Require: Score-based model s_θ , measurement \mathbf{y} , noise schedule σ_t , $(t_i)_{i \in \{1, \dots, N\}}$,

- 1: Sample $\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0)$.
 - 2: **Initialize the integral sum**, $f = 0$.
 - 3: **for** $i = N, N - 1 \dots, 0$ **do**
 - 4: Sample $\mathbf{x}_{t_i} \sim q(\mathbf{x}_{t_i} | \mathbf{y}, \hat{\mathbf{x}}_0) \propto q(\mathbf{y} | \mathbf{x}_{t_i})p(\mathbf{x}_{t_i} | \hat{\mathbf{x}}_0)$.
 - 5: Sample $\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0 | \mathbf{x}_{t_i})$ using s_θ .
 - 6: **Update the integral sum** $f \leftarrow f - (t_i - t_{i-1})\mathbf{v}_p(\mathbf{x}_{t_i})^\top \nabla_{\mathbf{x}_t} \log q(\mathbf{y} | \mathbf{x}_{t_i})$
 - 7: **end for**
 - 8: **return** $x^{(0)}, f$
-

The PnP-DM marginals follow $q(\mathbf{x}_t | \mathbf{y}) \propto q(\mathbf{y} | \mathbf{x}_t)p(\mathbf{x}_t)$ where $q(\mathbf{y} | \mathbf{x}_t)$ is a time-independent likelihood term. As a result, the initial distribution $q(\mathbf{x}_T | \mathbf{y}) \neq p(\mathbf{x}_T)$ and the KL divergence $D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{y}) || p(\mathbf{x}_T))$ is non-zero, so we can only compute the unnormalized model evidence. However, the unknown constant is dependent on \mathbf{y} and the forward model, so this unnormalized quantity can still be used for model selection. We derive the estimator in Proposition 2 and implement it with PnP-DM in Algorithm 2.

Proposition 2 (DiME-PnPDM: model evidence estimation along the PnP-DM marginals). *Given diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$, PnP-DM marginals $q(\mathbf{x}_t | \mathbf{y}) \propto p(\mathbf{x}_t)q(\mathbf{y} | \mathbf{x}_t)$, and timesteps $0 = t_0 < \dots < t_N = T$, the model evidence can be estimated via:*

$$\log p(\mathbf{y}) \approx C(q, \mathbf{y}) - \sum_{i=1}^N \Delta t_i \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_{t_i} | \mathbf{y})} [\mathbf{v}_p(\mathbf{x}_{t_i})^\top \nabla_{\mathbf{x}_{t_i}} \log q(\mathbf{y} | \mathbf{x}_{t_i})] \quad (17)$$

where $C(q, \mathbf{y}) = \log \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})} [q(\mathbf{y} | \mathbf{x})]$ is constant for fixed measurement \mathbf{y} and forward model, and $\mathbf{v}_p(\mathbf{x}_t) = \frac{a_t'}{a_t} \mathbf{x}_t - (\sigma_t' \sigma_t - \sigma_t^2 \frac{a_t'}{a_t}) \nabla_{\mathbf{x}} \log p(\mathbf{x}_t)$ is the PF-ODE velocity at \mathbf{x}_t (Eq. 13).

Proof. While we can apply Corollary 2, we can take a shortcut to cancel some extra terms. Under the PnP-DM marginals, $\log q(\mathbf{y} \mid t)$ is time-dependent while the likelihood $q(\mathbf{y} \mid \mathbf{x}_t)$ is time-independent, so we can directly compute the derivative of the evidence:

$$\begin{aligned}
\frac{d}{dt} \log q(\mathbf{y} \mid t) &= \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_t \mid \mathbf{y})} \left[\frac{d}{dt} \log p(\mathbf{x}_t) \right] && \text{(Fisher's identity)} \\
&= -\mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_t \mid \mathbf{y})} \left[\nabla \cdot \mathbf{v}_p(\mathbf{x}_t) + \mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) \right] && \text{(Log-cont. equation)} \\
&= -\mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_t \mid \mathbf{y})} \left[\nabla \cdot \mathbf{v}_p(\mathbf{x}_t) + \mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t \mid \mathbf{y}) \right] \\
&\quad + \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_t \mid \mathbf{y})} \left[\mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log q(\mathbf{y} \mid \mathbf{x}_t) \right] && \text{(Bayes)} \\
&= \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_t \mid \mathbf{y})} \left[\mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log q(\mathbf{y} \mid \mathbf{x}_t) \right]. && \text{(Stein's identity)}
\end{aligned}$$

Integrating over all t gives us the estimator:

$$\begin{aligned}
\log p(\mathbf{y}) &= \log q(\mathbf{y} \mid t = 0) \\
&= \log q(\mathbf{y} \mid t = T) - \int_0^T \frac{d}{dt} \log q(\mathbf{y} \mid t) dt \\
&= \log q(\mathbf{y} \mid t = T) - \sum_{i=1}^N \Delta t_i \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x}_{t_i} \mid \mathbf{y})} \left[\mathbf{v}_p(\mathbf{x}_{t_i})^T \nabla_{\mathbf{x}_{t_i}} \log q(\mathbf{y} \mid \mathbf{x}_{t_i}) \right]
\end{aligned}$$

where $\log q(\mathbf{y} \mid t = T) = \log \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \sigma_T^2)} [q(\mathbf{y} \mid \mathbf{x})]$ is the model evidence under a Gaussian prior, and $\mathbf{v}_p(\mathbf{x}_t) = \frac{a'_t}{a_t} \mathbf{x}_t - (\sigma'_t \sigma_t - \sigma_t^2 \frac{a'_t}{a_t}) \nabla_{\mathbf{x}} \log p(\mathbf{x}_t)$ is the PF-ODE from Equation 13. \square

We can also do a similar derivation for the marginals used for the Twisted Diffusion Sampler (Wu et al., 2023), which follow $\pi(\mathbf{x}_t \mid \mathbf{y}) \propto p(\mathbf{x}_t) p(\mathbf{y} \mid \mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t])$; note that this evidence gradient can be computed before resampling using SMC weights or after resampling without weights.

Proposition 3 (DIME-TDS: model evidence estimation along the TDS marginals). *Given diffusion process $\mathbf{x}_t = a_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$, $\mathbf{z}_t \sim \mathcal{N}(0, \mathbf{I})$, TDS marginals $\pi(\mathbf{x}_t \mid \mathbf{y}) \propto p(\mathbf{x}_t) p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t))$, and timesteps $0 = t_0 < \dots < t_N = T$, the model evidence can be estimated via:*

$$\log p(\mathbf{y}) \approx C(q, \mathbf{y}) - \sum_{i=1}^N \Delta t_i \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_{t_i} \mid \mathbf{y})} \left[\mathbf{v}_p(\mathbf{x}_{t_i})^T \nabla_{\mathbf{x}_{t_i}} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_{t_i})) + \frac{d}{dt} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_{t_i})) \right] \quad (18)$$

where $C(q, \mathbf{y}) = \log \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \sigma_T^2)} [p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_T))]$ is constant for fixed measurement \mathbf{y} and forward model, and $\mathbf{v}_p(\mathbf{x}_t) = \frac{a'_t}{a_t} \mathbf{x}_t - (\sigma'_t \sigma_t - \sigma_t^2 \frac{a'_t}{a_t}) \nabla_{\mathbf{x}} \log p(\mathbf{x}_t)$ is the PF-ODE velocity at \mathbf{x}_t (Eq. 13).

Proof.

$$\begin{aligned}
\frac{d}{dt} \log \pi(\mathbf{y} \mid t) &= \frac{1}{\pi(\mathbf{y} \mid t)} \frac{d}{dt} \int p(\mathbf{x}_t) p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t)) d\mathbf{x}_t \\
&= \frac{1}{\pi(\mathbf{y} \mid t)} \int \left[\frac{d}{dt} p(\mathbf{x}_t) \right] p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t)) d\mathbf{x}_t + \frac{1}{\pi(\mathbf{y} \mid t)} \int p(\mathbf{x}_t) \left[\frac{d}{dt} p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t)) \right] d\mathbf{x}_t \\
&= \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_t \mid \mathbf{y})} \left[\frac{d}{dt} \log p(\mathbf{x}_t) + \frac{d}{dt} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t)) \right] \\
&= \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_t \mid \mathbf{y})} \left[\mathbf{v}_p(\mathbf{x}_t)^T \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t)) + \frac{d}{dt} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_t)) \right]
\end{aligned}$$

where the last line is derived in a similarly to Proposition 2. Integrating over all t gives the estimator:

$$\begin{aligned}
\log p(\mathbf{y}) &= \log \pi(\mathbf{y} \mid t = T) - \int_0^T \frac{d}{dt} \log \pi(\mathbf{y} \mid t) dt \\
&\approx \log \pi(\mathbf{y} \mid t = T) - \sum_{i=1}^N \Delta t_i \mathbb{E}_{\mathbf{x}_t \sim \pi(\mathbf{x}_{t_i} \mid \mathbf{y})} \left[\mathbf{v}_p(\mathbf{x}_{t_i})^T \nabla_{\mathbf{x}_{t_i}} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_{t_i})) + \frac{d}{dt} \log p(\mathbf{y} \mid \boldsymbol{\mu}(\mathbf{x}_{t_i})) \right]
\end{aligned}$$

where $\log \pi(\mathbf{y} | T) \approx \log \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, \sigma_T^2)} [p(\mathbf{y} | \mu(\mathbf{x}_T))]$.

□

C BASELINE METHODS

In Section 4.1, we compared to five baseline methods, described here. Define p_β as an intermediate distribution of the *power-posterior path*, used in TI, AIS, and SMC: $p_\beta(\mathbf{x}) \propto p(\mathbf{x})p(\mathbf{y} | \mathbf{x})^\beta$, $\beta \in [0, 1]$. Note that this path is equivalent to the *geometric path* under a different parametrization.

Naive Monte Carlo (Naive MC): We generate many unconditional prior samples and compute the expectation of the likelihood as follows: $\log p(\mathbf{y}) = \log \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}_0)}[p(\mathbf{y} | \mathbf{x})]$.

Original DAPS covariance heuristic (Original heuristic): We perform the same steps as described in Algorithm 1 but using the originally proposed covariance heuristic $\Sigma_{\mathbf{x}_0 | \mathbf{x}_t} = \sigma_t^2$.

Thermodynamic integration (TI) (Lartillot & Philippe, 2006): Originally from statistical mechanics, TI can also compute expectations by integrating the evidence over the power-posterior path:

$$\begin{aligned} \frac{d}{d\beta} \log p_\beta(\mathbf{y}) &= \frac{1}{p_\beta(\mathbf{y})} \frac{d}{d\beta} p_\beta(\mathbf{y}) = \frac{1}{p_\beta(\mathbf{y})} \int p(\mathbf{x}) \frac{d}{d\beta} p(\mathbf{y} | \mathbf{x})^\beta d\mathbf{x} \\ &= \frac{1}{p_\beta(\mathbf{y})} \int p(\mathbf{x}) p(\mathbf{y} | \mathbf{x})^\beta \log p(\mathbf{y} | \mathbf{x}) d\mathbf{x} = \mathbb{E}_{\mathbf{x} \sim p_\beta}[\log p(\mathbf{y} | \mathbf{x})]. \end{aligned}$$

Using the fact that $\log p_{\beta=0}(\mathbf{y}) = \log 1 = 0$, using K integration steps we have the estimator:

$$\log p_{\beta=1}(\mathbf{y}) \approx \sum_{k=1}^K \Delta\beta_k \mathbb{E}_{\mathbf{x} \sim p_{\beta_{k-1}}}[\log p(\mathbf{y} | \mathbf{x})].$$

Samples from each p_β were generated using Sequential Monte Carlo (details below).

Annealed importance sampling (AIS) (Neal, 2001): We can compute the ratio of evidence:

$$\begin{aligned} \frac{p_{\beta_k}(\mathbf{y})}{p_{\beta_{k-1}}(\mathbf{y})} &= \frac{\int p(\mathbf{x}) p(\mathbf{y} | \mathbf{x})^{\beta_{k-1}} p(\mathbf{y} | \mathbf{x})^{\Delta\beta_k} d\mathbf{x}}{p_{\beta_{k-1}}(\mathbf{y})} = \mathbb{E}_{\mathbf{x} \sim p_{\beta_{k-1}}}[p(\mathbf{y} | \mathbf{x})^{\Delta\beta_k}] \\ &= \mathbb{E}_{p_{\mathbf{x} \sim \beta_{k-1}}}[\exp(\Delta\beta_k \log p(\mathbf{y} | \mathbf{x}))]. \end{aligned}$$

Using the fact that $\log p_{\beta=0}(\mathbf{y}) = \log 1 = 0$, telescoping over all β_k gives:

$$\begin{aligned} p_{\beta=1}(\mathbf{y}) &= \prod_{k=1}^K \mathbb{E}_{p_{\beta_{k-1}}}[\exp(\Delta\beta_k \log p(\mathbf{y} | \mathbf{x}))] \\ &= \mathbb{E}_{\mathbf{x}_{0:k-1} \sim p_{\beta_{0:k-1}}} \exp\left(\sum_{k=1}^K \Delta\beta_k \log p(\mathbf{y} | \mathbf{x}_{k-1})\right). \end{aligned}$$

Note that $\mathbf{x}_{0:k-1} \sim p_{\beta_{0:k-1}}$ represents the sample-path of one particle, where each transition kernel from $k-1$ to k was performed using Langevin dynamics.

Sequential Monte Carlo (SMC) (Del Moral et al., 2006): An extension of AIS that uses the resampling step from Sequential Monte Carlo methods to replace particles with degenerate weights (i.e. particles stuck in the wrong mode) while sampling the next intermediate distribution, potentially reducing both bias and variance. These resampling steps are performed either using a fixed schedule or adaptively when the effective sample size falls below a threshold.

D IMPLEMENTATION DETAILS

D.1 DIME

As DIME is implemented alongside DAPS, we have a similar set of hyperparameters.

Number of diffusion steps for sampling $p(x_0 | x_t)$: While the original DAPS method uses many diffusion steps to estimate the denoised image at each annealing iteration, we use 1 diffusion step to ensure an exact estimate of $\mathbb{E}[x_0 | x_t]$ by Tweedie’s formula.

Langevin dynamics: The learning rate of Langevin dynamics was tuned so that posterior samples $\tilde{x}_0 \sim p(x_0 | x_t, \mathbf{y})$ have a mean reduced χ^2 data likelihood fit that converges as low as possible (ideally around 1, which is indicative of a good posterior sample for likelihoods with Gaussian noise). We also use a linear learning rate decay as is done in Zhang et al. (2025). The table below lists the learning rates and maximum step counts used for each experiment. In practice, sample convergence usually happens much sooner than the maximum step count.

Task	lr	Max steps
Mixture of Gaussians (in-distribution)	5×10^{-4}	2,000
Mixture of Gaussians (out-of-distribution)	2×10^{-5}	10,000
Mixture of Gaussians (saddle point)	2×10^{-5}	10,000
Fourier phase retrieval	1×10^{-3}	1,000
Gaussian phase retrieval	5×10^{-4}	1,000
M87* black hole imaging	1×10^{-5}	1,000

Selection of σ_{min} : This parameter can be viewed as the minimum noise level that the learned score can still be trusted. For the analytic Mixture of Gaussians experiment, we use a minimum annealing noise level $\sigma_{min} = 0.05$, and for the diffusion model experiments, we use $\sigma_{min} = 0.1$ as is done in Zhang et al. (2025). The remaining integral from $t = 0$ to $t = t_{min}$ in Eq. 9 is approximated by a trapezoid, using the fact that the integrand equals 0 at $t = 0$.

Computing the precision matrix Σ_0^{-1} : for some datasets, the covariance of specific pixel locations is nearly 0 (such as the corner regions of MNIST), so we added a jitter of 1e-2 for stability.

D.2 BASELINES

We use TI, AIS, and SMC as baseline methods for the mixture of Gaussians experiment (Section 4.1) and SMC for the MNIST experiment (Section 4.2; Appendix E). We used Langevin dynamics with 2000 steps and learning rate 0.0002 for our transitions between distributions. For SMC, we resample when the effective sample size (ESS) is less than 0.6.

We test with fixed linear, exponential, sigmoidal β as well as adaptive β schedules that decrease the Conditional ESS at a constant rate of [0.5, 0.6, 0.7]. For the mixture of Gaussians study, an exponential schedule with $\lambda = 6$ was found to be optimal. For the MNIST experiment, the fixed schedules use 50 annealing steps, while the adaptive schedule had a minimum β step size of 0.002 which in practice results in 150 steps at most.

E BASELINE RESULTS ON MNIST MODEL SELECTION

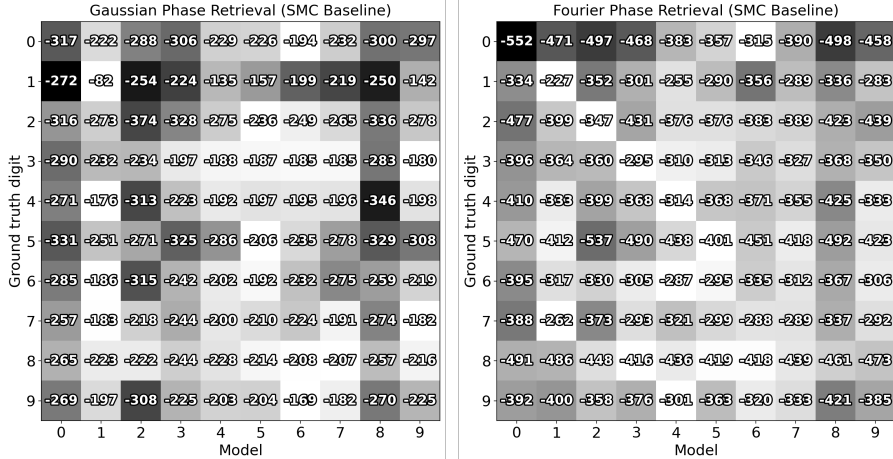


Figure 6: Model evidence confusion matrix using sequential Monte Carlo (SMC) for Gaussian phase retrieval (left) and Fourier phase retrieval (right) for each (ground truth measurement, model) pair of MNIST digits. When SMC uses the learned diffusion score $\nabla_x \log p(x_t)$ at $t = 0.1$ as the prior score, it often fails to select the correct model.

Results of the MNIST experiment (Section 4.2) using SMC and the learned $\nabla_x \log p(x_t)$ with $\sigma = 0.1$ as a prior score are shown in Figure 6. SMC is often unable to determine the correct MNIST model. Each SMC annealing step also takes about $10\times$ longer than an equivalent annealing step for DiME, due to having to query the diffusion model many times for the prior score. Interestingly, narrower priors (i.e. MNIST digit 1) leads to *higher* model evidence estimates here; since the training data is more clumped together, the learned score becomes more accurate at low noise levels.

For completeness, the model evidence estimates of a few runs over all temperature annealing schedules as well as for a second noise level, $\sigma = 0.05$, are reported in Tables 2 and 3. Mean and standard deviation over 5 runs is shown here. As can be seen, lowering the noise level only further increases overfitting to training data, causing the estimated evidence to drop.

SMC schedule	$\sigma_{\min} = 0.1$			$\sigma_{\min} = 0.05$		
	GT 8 Model 6	GT 8 Model 7	GT 8 Model 8	GT 8 Model 6	GT 8 Model 7	GT 8 Model 8
Linear	-182 ± 9	-212 ± 9	-210 ± 11	-279 ± 26	-352 ± 3	-426 ± 40
Exponential	-183 ± 7	-224 ± 5	-239 ± 6	-292 ± 20	-383 ± 9	-468 ± 12
Sigmoidal	-175 ± 5	-219 ± 13	-191 ± 2	-273 ± 15	-374 ± 9	-480 ± 17
Adaptive 0.5	-195 ± 31	-202 ± 23	-261 ± 1	-296 ± 52	-346 ± 16	-489 ± 33
Adaptive 0.6	-184 ± 6	-189 ± 7	-259 ± 4	-279 ± 45	-373 ± 17	-485 ± 26
Adaptive 0.7	-192 ± 14	-192 ± 2	-255 ± 11	-296 ± 26	-384 ± 18	-522 ± 30

Table 2: SMC model evidence estimates on Gaussian phase retrieval.

SMC schedule	$\sigma_{\min} = 0.1$			$\sigma_{\min} = 0.05$		
	GT 8 Model 6	GT 8 Model 7	GT 8 Model 8	GT 8 Model 6	GT 8 Model 7	GT 8 Model 8
Linear	-416 ± 11	-431 ± 4	-409 ± 1	-485 ± 25	-577 ± 2	-563 ± 2
Exponential	-422 ± 20	-427 ± 2	-467 ± 11	-527 ± 10	-582 ± 12	-585 ± 9
Sigmoidal	-415 ± 9	-451 ± 3	-411 ± 7	-520 ± 13	-580 ± 7	-575 ± 5
Adaptive 0.5	-409 ± 4	-409 ± 6	-460 ± 8	-524 ± 33	-563 ± 18	-567 ± 10
Adaptive 0.6	-431 ± 20	-437 ± 11	-458 ± 8	-510 ± 9	-566 ± 14	-562 ± 13
Adaptive 0.7	-442 ± 9	-437 ± 8	-453 ± 1	-516 ± 6	-584 ± 3	-570 ± 6

Table 3: SMC model evidence estimates on Fourier phase retrieval.

F SYNTHETIC MODEL SELECTION

In this section, we perform model selection on the same set of five priors as in Section 4.3.1, but on simulated closure quantities measurements of test images from each prior as verification. The correct model is chosen for all measurements. The posterior samples with the lowest, mean, and highest path evidences are also shown in each case.

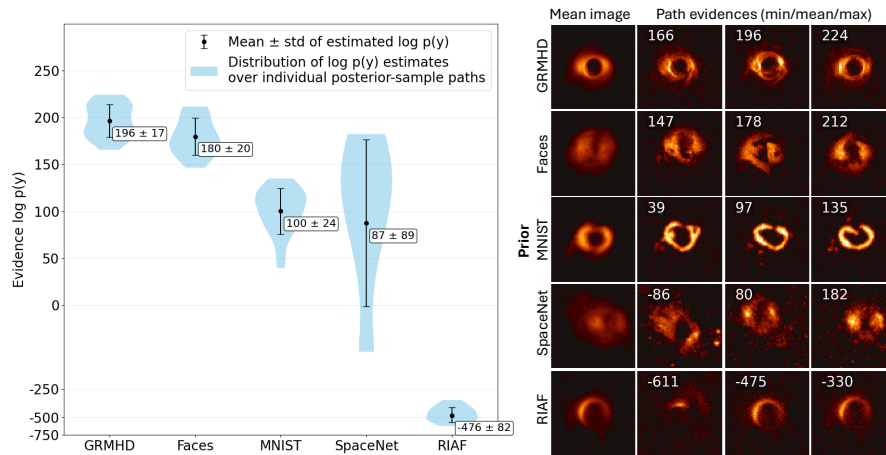


Figure 7: Ground truth: GRMHD.

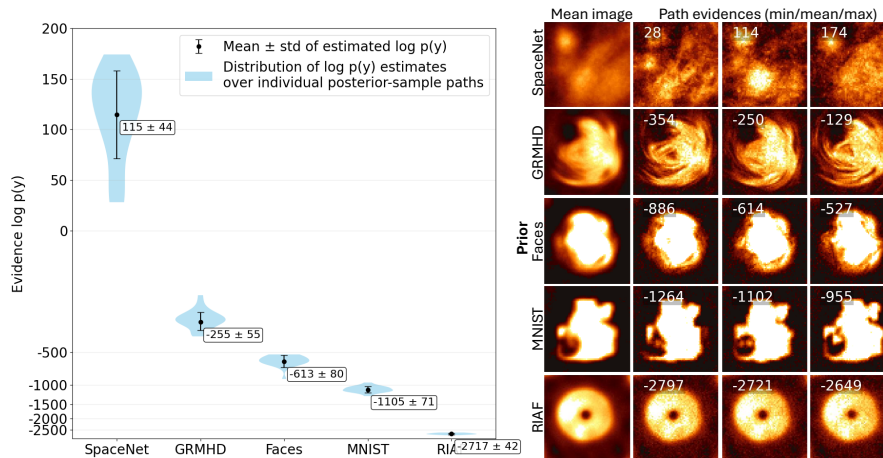


Figure 8: Ground truth: SpaceNet.

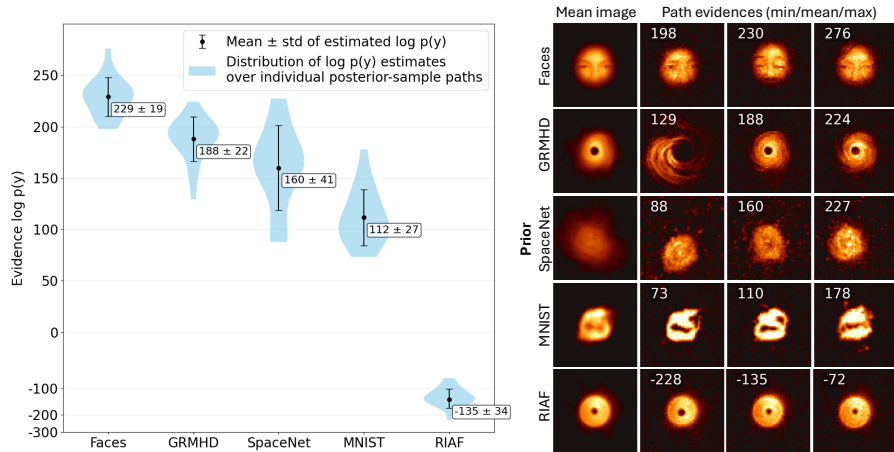


Figure 9: Ground truth: CelebA.

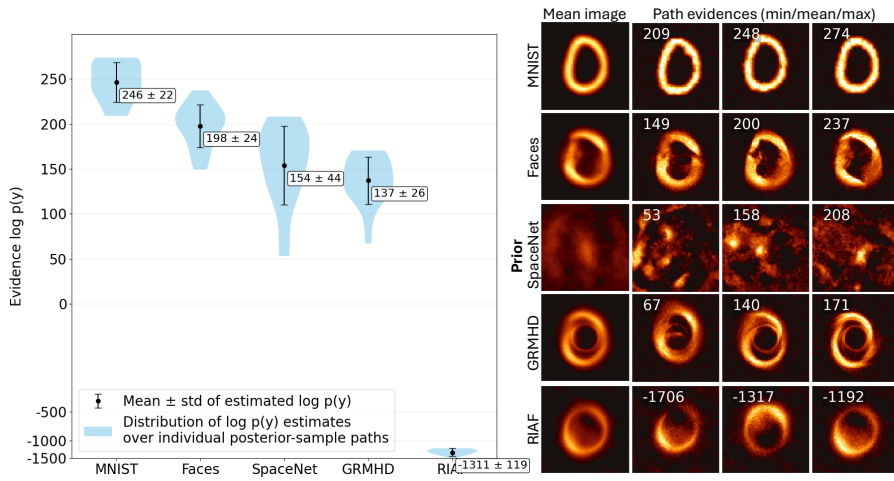


Figure 10: Ground truth: MNIST.

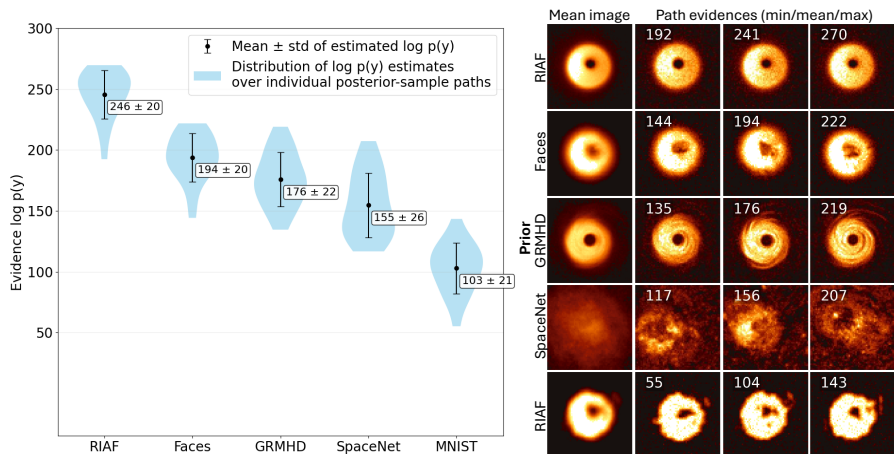


Figure 11: Ground truth: RIAF.

G LARGE LANGUAGE MODEL USAGE

We used ChatGPT for coding assistance, polishing text and LaTeX, and checking mathematical derivations.