WHEN THE PRIOR IS WRONG: PROBING MISSPECIFIED DIFFUSION PRIORS IN INVERSE PROBLEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

While diffusion priors generate high-quality posterior samples across many scientific inverse problems, they are often trained on purely simulated data, thus inheriting the errors and biases of these underlying sources. Many approaches to adapting diffusion priors from a single observation require sufficient conflict between the data and the prior to successfully adapt and do not allow specific misspecification hypotheses to be tested. On the other hand, existing out-of-distribution (OOD) detection methods determine whether the observation is compatible with a fixed prior, rather than specifying which assumptions to relax or by how much. To address these issues, we propose PIVOT: Prior Inference Via Observation-driven Transformations, which formulates diffusion prior misspecification as inference over an underlying latent variable representing a scientifically plausible discrepancy. By expanding the prior along these plausible directions of misspecification, PIVOT infers a posterior over the strength of each proposed discrepancy given the observation. This process can be efficiently amortized with just one posterior sampling pass using a conditional diffusion model. We also demonstrate how to determine if a proposed discrepancy is inconclusive because the measurement is too noisy or ill-posed. We validate PIVOT on seismic imaging and real-world black hole imaging problems, where we find that OOD observations generally lead to larger perturbation strengths when the proposed discrepancy family captures relevant components of the mismatch, as long as the discrepancy is visible to the measurement. PIVOT thereby provides an interpretable probe into how real-world priors could be misspecified.

1 INTRODUCTION

In ill-posed imaging inverse problems, our reconstructions are only as good as the assumptions we make. These assumptions, often encoded as prior probability distributions $p(\boldsymbol{x})$, describe what we believe plausible image solutions should look like before seeing the data. Yet in practice, the true prior distribution that governs nature's variability is never known. Instead, we rely on approximations: models that are convenient, interpretable, or empirically motivated, but almost certainly wrong in subtle ways. These mismatches can bias the resulting reconstructions, leading us to underestimate uncertainty or to impose features that reflect our model more than reality.

Diffusion models (Song et al., 2021b; Ho et al., 2020) are state-of-the-art data-driven priors for imaging inverse problems, and most existing posterior sampling methods (Chung et al., 2022; Coeurdoux et al., 2024; Zhu et al., 2023; Wu et al., 2024; Zhang et al., 2025a) treat the pretrained diffusion models as properly specified. However, many diffusion models used for scientific priors are trained on images from simulations or highly limited training sets, inheriting potentially incorrect physical assumptions about the world, simulation constraints, missing coverage of rare phenomena, and other modeling biases.

Robust Bayesian analysis provides a natural framework for studying prior misspecification by asking how posterior conclusions change under perturbations of the prior (Berger et al., 1994). For analytic priors, these perturbations can often be expressed through hyperparameters or explicit prior expansions (Nott et al., 2021). Extending this idea to data-driven priors is less straightforward: their assumptions are encoded implicitly in a high-dimensional training distribution, and generic perturbations need not preserve realistic image structure. For example, tempering a high-dimensional

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

data-driven prior can shift probability mass away from its typical set, producing unrealistic samples; see Appendix A.1. This motivates considering perturbations that correspond to plausible ways the learned prior could be misspecified. Data augmentation during training (Karras et al., 2022; Zhang et al., 2025b) can be viewed as a form of prior expansion, but does not probe specific misspecification directions nor provide a way to determine if the expansion is detectable by our observations.

One line of work attempts to quantify how OOD a measurement is with respect to a given diffusion prior, using statistics such as evidence (Wang & Bouman, 2026), conditional evidence (Sprunck et al., 2026), and KL divergence (Shoushtari et al., 2025). While these approaches are suitable for diagnosing OOD, they do not provide any insight as to how a mismatched prior distribution can be perturbed to best fit our observation. On the other hand, test-time diffusion prior adaptation methods finetune the network weights during sampling to better fit the measurements (Barbano et al., 2025; Chung & Ye, 2024; Hu et al., 2024a). These methods require careful tuning to prevent overfitting to the measurement, and perform best when there is conflict between the prior and the observation. However, after finetuning, the diffusion model may no longer represent the score of a coherent prior, and the adaptation is driven largely by inductive biases in the network architecture rather than by an explicitly specified, scientifically plausible direction of misspecification, making the resulting reconstructions difficult to interpret. Patch-based diffusion priors have been shown to perform better on OOD measurements (Hu et al., 2024b;a) by weakening global correlations in images; however, patch size controls how strongly the original prior is relaxed, and existing methods do not infer this.

Our contributions: We propose PIVOT: Prior Inference Via Observation-driven Transformations, a framework for probing different hypotheses of prior misspecification from a single observation. PIVOT expands the base diffusion prior with a hierarchical family of perturbed priors, and computes the posterior over perturbation strengths supported by the observation according to the Bayesian evidence. This discrepancy family, i.e. an explicit perturbation retroactively applied to simulated data or an implicit simulator variable, encodes a plausible scientific hypothesis and provides a probe into how likely these discrepancies are given our observation relative to synthetic in-distribution and OOD observations. Rather than performing posterior sampling separately for many different perturbation strengths, PIVOT efficiently amortizes this using a conditional diffusion prior trained with label dropout. While some discrepancies are either invisible or confounded under evidence (a fundamental limitation arising from the forward operator used to collect observations), we find that when the data likelihood is sensitive to the perturbation family, then the computed evidence recovers whether the proposed discrepancy is supported by the observation and at what strength. We also find that empirically, the reconstruction quality under the expanded prior can exceed the strong prior, even if the true discrepancy is not captured by the family. In doing so, PIVOT reframes prior selection not as a fixed choice but as an inference problem in its own right.

We validate PIVOT on a toy Gaussian model where the evidence and diffusion score can be computed analytically. Here, we find that despite the approximate diffusion posterior sampling and approximate density evaluation, we still achieve highly accurate perturbation strength estimates. We also demonstrate an example on the travel-time tomography seismic imaging problem, where we design a discrepancy family of elliptical subsurface structures, inspired by magma chambers, to extend a narrow data-driven OpenFWI prior (Deng et al., 2022). We find that our method identifies additional elliptical subsurface features when they are present, and favors such structures for completely OOD measurements when they better explain the observations.

Finally, we use PIVOT on M87* black hole imaging, where in the case of real data the true prior is unknown. Priors trained on physical simulations of black hole images (Mizuno, 2022) have been successfully used to reconstruct real M87* observations (Wu et al., 2024; Feng et al., 2024). To demonstrate how our approach can probe different hypotheses on real scientific data, we design two discrepancy families: additional compact emission and elliptical deformations. We find that the black hole forward model is not very sensitive to elliptical deformations, and that the real M87* observations do not favor ellipticity beyond what is already seen in undeformed simulations, in agreement with previous results (Dahale et al., 2025). In contrast, the real M87* observations on April 5 and 6, 2017 (EHTC, 2019a) demonstrate consistently larger compact-emission discrepancies than most in-distribution observations. This feature is consistently located southwest of the ring across all four observation days, suggesting the presence of additional compact emission that is not reconstructed under the strong simulation prior and could be associated with material flowing outward from the black hole in a jet.

2 BACKGROUND

2.1 DIFFUSION INVERSE PROBLEM SOLVERS

Given a ground truth image or signal $\boldsymbol{x}$, forward operator $A(\cdot)$, and measurement noise $\boldsymbol{\varepsilon}$, we can obtain a measurement $\boldsymbol{y} = A(\boldsymbol{x}) + \boldsymbol{\varepsilon}$. Classical inverse problem solvers use handcrafted prior assumptions, such as total variation or ℓ_1 norm, to narrow down the ill-posedness of the problem. Diffusion models can be used as data-driven priors for posterior sampling. Early approaches to diffusion posterior sampling (Chung et al., 2022; Song et al., 2023; Boys et al., 2023) guide the sampling trajectory with approximations of the intractable $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y} \mid \boldsymbol{x}_t)$. Recently, posterior samplers that anneal along specified marginals using invariant transition kernels (Zhu et al., 2023; Wu et al., 2024; Zhang et al., 2025a; Zhu et al., 2024) have led to more accurate posterior samples.

2.2 CONDITIONAL DIFFUSION MODELS

Given a condition $\boldsymbol{a}$ and a conditioned image $\boldsymbol{x}_0 \sim p(\boldsymbol{x}_0 \mid \boldsymbol{a})$, the variance exploding (VE) forward diffusion process at time $t \in [0, T]$ and noise schedule σ_t is given by:

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

We can sample from the diffusion by discretizing the probability flow ODE (PF-ODE, Eq. 2):

$$d\boldsymbol{x}_t = -\sigma_t \sigma_t' \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t \mid \boldsymbol{a}) dt \tag{2}$$

where $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t \mid \boldsymbol{a})$ is learned. We use the VE process and PF-ODE throughout this paper for ease of notation, but our method can be extended to other diffusion noise schedules or the reverse SDE formulation sampling under a reparameterization. Given a noisy image $\boldsymbol{x}_t \sim p(\boldsymbol{x}_t \mid \boldsymbol{a})$, Tweedie's formula gives us the posterior mean:

$$\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t, \boldsymbol{a}] = \boldsymbol{x}_t + \sigma_t^2 \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t \mid \boldsymbol{a}). \tag{3}$$

To strengthen the conditioning signal, Bernoulli dropout with a null condition $\varnothing$ is used during training (Ho & Salimans, 2022); thus, the null condition is trained on the marginalized distribution:

$$p(\boldsymbol{x} \mid \varnothing) = \int p(\boldsymbol{x} \mid \boldsymbol{a}) p(\boldsymbol{a}) d\boldsymbol{a} \tag{4}$$

The evidence lower bound (ELBO) (Song et al., 2021a) provides an efficient surrogate log-density:

$$L_{ELBO}(\boldsymbol{x}_0 \mid \boldsymbol{a}) = -\int_0^T \mathbb{E}_{\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I})} \left[\frac{\dot{\sigma}_t}{\sigma_t^3} \left\| D_\theta(\boldsymbol{x}_0 + \sigma_t \boldsymbol{\varepsilon}, \sigma_t, \boldsymbol{a}) - \boldsymbol{x}_0 \right\|_2^2 \right] dt \tag{5}$$

$$\leq \log p(\boldsymbol{x}_0 \mid \boldsymbol{a}) - C(\boldsymbol{x}_0) \tag{6}$$

As $C(\boldsymbol{x}_0)$ is independent of the condition $\boldsymbol{a}$, the ELBO has been used for classification (Li et al., 2023) and for providing accurate gradients for sampling conditions of clean images (Mudur et al., 2024), despite the remaining lower bound gap.

2.3 DIFFUSION OOD ADAPTATION

When a prior $p(\boldsymbol{x})$ is mismatched to an observation $\boldsymbol{y}$, one approach is to relax the prior rather than use it as-is. Classical robust Bayesian methods do this by, for example, mixing a strong prior with a weaker contamination prior (Berger & Berliner, 1986), or introducing hyperparameters that allow the prior to vary over different assumptions (Tiede et al., 2026). Recent work has explored analogous strategies for diffusion priors. Deep Diffusion Image Prior (DDIP) methods (Barbano et al., 2025; Chung & Ye, 2024) adapt a pretrained diffusion model to a single observation at test time by fine-tuning the denoiser as a Deep Image Prior (Ulyanov et al., 2018). Alternatively, patch-based diffusion priors relax global image structure by modeling local patches, improving robustness to distribution shift (Hu et al., 2024b), and can also be combined with DDIP (Hu et al., 2024a). DDIP requires careful tuning to avoid overfitting and adapts the prior in directions determined implicitly by the network, while patch-based methods can only encode the specific hypothesis that global image structure is the source of mismatch.

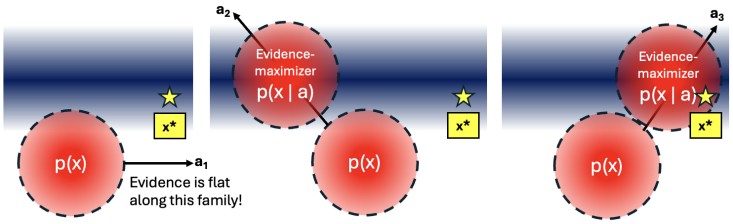

Figure 1: Three examples of maximizing evidence, where each perturbation family i probes the mean of the Gaussian prior $p(\boldsymbol{x})$ along $\boldsymbol{a}_i$. **Left:** The perturbation may be useful, but is unsupported by the measurement, so the evidence is flat; with one measurement, this is unavoidable. **Middle:** A failure case. The perturbation family is poor, so the evidence is maximized at a prior that the ground truth $\boldsymbol{x}^*$ is less likely under. Therefore, the perturbation family which determines the coupling between measurement and null space must be selected well a priori and calibrated. **Right:** The perturbation family is specified well, and the evidence-maximizing prior in this family captures $\boldsymbol{x}^*$.

Another related line of work uses a large bank of measurements with different forward operators to train or finetune a diffusion model. Ambient diffusion methods (Daras et al., 2023; Aali et al., 2024; Daras et al., 2024) train a diffusion model using many corrupted observations, while expectation-maximization methods finetune an initial diffusion model prior (Bai et al., 2024; Rozet et al., 2024; Barco et al., 2025). However, these works assume many observations while we target situations where only a single observation may be available. EM methods have been used to optimize the input condition in latent diffusion models for solving inverse problems (Chung et al., 2023), offering an alternative approach of adapting the prior for large pretrained latent diffusion models.

3 Method

PIVOT has three steps. First, we define a family of priors $p(\boldsymbol{x} \mid \boldsymbol{a})$, where $\boldsymbol{a} = 0$ is the original prior and increasing $\boldsymbol{a}$ introduces more of a user-specified discrepancy. Second, we train a conditional diffusion model over this family, with label dropout so that the unconditional dropout condition represents the family marginalized over $\boldsymbol{a}$. Finally, for a new observation $\boldsymbol{y}$, we perform posterior sampling once under this marginalized prior and use the resulting samples and conditional diffusion model to estimate which discrepancy strengths $\boldsymbol{a}$ are best supported by the observation.

3.1 Posterior Sampling over an Expanded Prior Family

We begin by specifying a plausible way in which the base prior could be misspecified. Importantly, this discrepancy family must be chosen *a priori* to represent a scientifically meaningful hypothesis. Bayesian evidence can determine whether a perturbation improves agreement with the measurements, but it cannot determine whether the associated changes in the unobserved null space are physically meaningful. Restricting the perturbation to a scientifically plausible direction therefore helps ensure that improvements in measurement fit are accompanied by changes we are willing to believe in the unconstrained parts of the solution. Figure 1 (middle) illustrates what can happen when this direction is poorly chosen.

Let $\boldsymbol{a} \in [0, 1]^d$ denote the strengths of d different discrepancies, with $\boldsymbol{a} = \boldsymbol{0}$ as the base prior and $a_i = 1$ as the maximum perturbation considered for discrepancy i. Given a sample $\boldsymbol{x}' \sim p_0(\boldsymbol{x})$ from the base prior p_0, let $p(\boldsymbol{x} \mid \boldsymbol{x}', \boldsymbol{a})$ represent a distribution over perturbations of $\boldsymbol{x}'$ with strength $\boldsymbol{a}$, such that the non-perturbed case gives $p(\boldsymbol{x} \mid \boldsymbol{x}', \boldsymbol{0}) = \delta(\boldsymbol{x} - \boldsymbol{x}')$. This defines the family of priors: $p(\boldsymbol{x} \mid \boldsymbol{a}) = \int p(\boldsymbol{x} \mid \boldsymbol{x}', \boldsymbol{a}) p_0(\boldsymbol{x}')\, d\boldsymbol{x}'$. We train a conditional diffusion model on samples from this family. During training, we sample $\boldsymbol{a} \sim \mathrm{Unif}([0, 1]^d)$ and use label dropout so that the null condition $\varnothing$ learns the prior marginalized over discrepancy strength $\boldsymbol{a}$:

$$p(\boldsymbol{x} \mid \varnothing) = \int p(\boldsymbol{x} \mid \boldsymbol{a}) p(\boldsymbol{a} \mid \varnothing)\, d\boldsymbol{a}, \qquad p(\boldsymbol{a} \mid \varnothing) = \mathrm{Unif}([0, 1]^d). \tag{7}$$

The conditional model therefore represents each individual prior $p(\boldsymbol{x} \mid \boldsymbol{a})$, while the unconditional head represents the entire discrepancy family at once. Although we use a uniform distribution over $\boldsymbol{a}$ during training, different hyperpriors $p(\boldsymbol{a})$ can be imposed at inference time through reweighting, as described in Section 3.2. See Algorithm 1 for an example training iteration.

Algorithm 1 One training iteration of conditional diffusion

Require: Base prior $p_0(\boldsymbol{x})$, perturbation kernel $p(\boldsymbol{x} \mid \boldsymbol{x}', \boldsymbol{a})$, label dropout $p_{dropout}$, conditional diffusion net $D_\theta(\boldsymbol{x}_t, \sigma_t, \boldsymbol{a})$, diffusion loss weighting $w(\sigma_t)$.

1: Sample $\boldsymbol{a} \sim Unif([0,1]^d)$, $\boldsymbol{x}' \sim p_0(\boldsymbol{x})$, $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{x}', \boldsymbol{a})$.
2: Sample t from noise schedule, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \boldsymbol{I})$, $\boldsymbol{x}_t \leftarrow \boldsymbol{x} + \sigma_t \boldsymbol{\varepsilon}$
3: **if** rand $< p_{dropout}$ **then**
4: Take a gradient step on the diffusion objective $w(\sigma_t) \left\| D_\theta(\boldsymbol{x}_t, \sigma_t, \varnothing) - \boldsymbol{x} \right\|_2^2$
5: **else**
6: Take a gradient step on the diffusion objective $w(\sigma_t) \left\| D_\theta(\boldsymbol{x}_t, \sigma_t, \boldsymbol{a}) - \boldsymbol{x} \right\|_2^2$
7: **end if**

The unconditional head represents the prior marginalized over all discrepancy strengths. Therefore, conditioning it on an observation $\boldsymbol{y}$, whose likelihood is independent of the condition $\boldsymbol{a}$, gives

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing) = \int p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing, \boldsymbol{a}) p(\boldsymbol{a} \mid \boldsymbol{y}, \varnothing) d\boldsymbol{a} = \int p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{a}) p(\boldsymbol{a} \mid \boldsymbol{y}, \varnothing) d\boldsymbol{a} \tag{8}$$

The marginalized posterior combines the posterior associated with each discrepancy strength $\boldsymbol{a}$, weighted by how well that prior explains the observation. Since we assume a uniform prior over discrepancy strengths, the posterior weight assigned to each $\boldsymbol{a}$ is proportional to its evidence $p(\boldsymbol{y} \mid \boldsymbol{a})$. Thus, discrepancy strengths that make the observed measurement more likely receive more weight in the marginalized posterior. This allows a single posterior sampling run to explore the full discrepancy family, rather than separately solving the inverse problem for each value of $\boldsymbol{a}$. To perform posterior sampling, we use the DPMC sampler (Zhu et al., 2024) which anneals along $p(\boldsymbol{x}_t) p(\boldsymbol{y} \mid \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t, \varnothing])^{\beta_t}$ for $t \to 0$ in a predictor-corrector fashion; this sampler is asymptotically exact, allows for high-noise mixing and mode exploration, and demonstrates strong empirical performance for scientific inverse problems as shown in Section 4.2 and 4.3.

3.2 Inferring the Discrepancy Posterior

Given samples from the marginalized posterior $p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)$, we now want to determine which discrepancy strengths $\boldsymbol{a}$ are supported by the observation. The key observation is that each posterior sample $\boldsymbol{x}$ contains information about which member of the prior family $p(\boldsymbol{x} \mid \boldsymbol{a})$ could have generated it. Averaging this information across posterior samples allows us to recover the discrepancy posterior $p(\boldsymbol{a} \mid \boldsymbol{y})$ without rerunning posterior sampling for each value of $\boldsymbol{a}$.

$$\begin{aligned} p(\boldsymbol{a} \mid \boldsymbol{y}) &\propto p(\boldsymbol{a}) p(\boldsymbol{y} \mid \boldsymbol{a}) && (9) \\ &\propto p(\boldsymbol{a}) p(\boldsymbol{a} \mid \boldsymbol{y}, \varnothing) && (10) \\ &\propto p(\boldsymbol{a}) \int p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{y}, \varnothing) p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing) d\boldsymbol{x} && (11) \\ &= p(\boldsymbol{a}) \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)} [p(\boldsymbol{a} \mid \boldsymbol{x}, \varnothing)] && (12) \end{aligned}$$

The second line follows because the training distribution $p(\boldsymbol{a} \mid \varnothing)$ is uniform. For each posterior sample $\boldsymbol{x}$, we use the conditional prior $p(\boldsymbol{x} \mid \boldsymbol{a})$ to determine which discrepancy strengths are most compatible with that sample, yielding $p(\boldsymbol{a} \mid \boldsymbol{x}, \varnothing)$. Averaging these distributions across posterior samples then gives $p(\boldsymbol{a} \mid \boldsymbol{y})$. We consider two ways to estimate this distribution: MCMC sampling over $\boldsymbol{a}$, or direct evaluation on a grid if $\boldsymbol{a}$ is low-dimensional ($d = 1, 2$). Using the MCMC approach, for each posterior sample $\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)$, we can sample $\boldsymbol{a}$ using reflected Langevin dynamics and score $\nabla_{\boldsymbol{a}} \log p(\boldsymbol{a} \mid \boldsymbol{x}, \varnothing) = \nabla_{\boldsymbol{a}} \log p(\boldsymbol{x} \mid \boldsymbol{a})$. The result can be reweighted by the desired $p(\boldsymbol{a})$ and combined across all samples of $p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)$. In practice, we approximate $\log p(\boldsymbol{x} \mid \boldsymbol{a})$ using the discretized ELBO (Eq. 5); we directly validate this approximation and show it recovers the analytic posterior in Section 4.1 and the true discrepancy strength in Appendix A.4 despite the lower bound gap. Alternatively, for low-dimensional $\boldsymbol{a}$, we can evaluate its (unnormalized) posterior on a grid:

$$\begin{aligned} p(\boldsymbol{a} \mid \boldsymbol{y}) &\propto p(\boldsymbol{a}) \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)} [p(\boldsymbol{a} \mid \boldsymbol{x}, \varnothing)] && (13) \\ &= p(\boldsymbol{a}) \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)} [\frac{p(\boldsymbol{x} \mid \boldsymbol{a})}{p(\boldsymbol{x} \mid \varnothing)}] && (14) \\ &= p(\boldsymbol{a}) \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x} \mid \boldsymbol{y}, \varnothing)} [\frac{p(\boldsymbol{x} \mid \boldsymbol{a})}{\mathbb{E}_{\boldsymbol{a} \sim Unif([0,1]^d)} [p(\boldsymbol{x} \mid \boldsymbol{a})]}] && (15) \end{aligned}$$

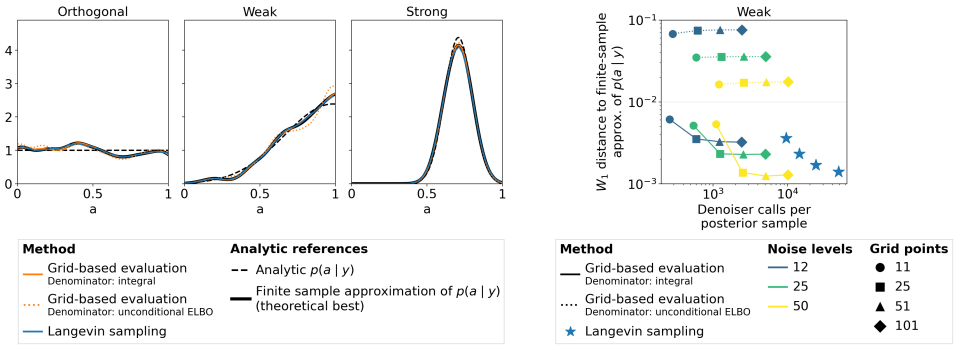

Figure 2: **(Left)** Discrepancy posterior $p(\boldsymbol{a} \mid \boldsymbol{y})$ results on a Gaussian toy experiment where the forward operator has an increasing level of sensitivity to the perturbation (orthogonal, weak, strong). The ideal analytic $p(\boldsymbol{a} \mid \boldsymbol{y})$, analytic finite-sample (50) approximation of $p(\boldsymbol{a} \mid \boldsymbol{y})$ which is the theoretical best we can achieve, and ELBO-based discrepancy methods are shown. The shape of the evidence curve is correctly recovered up to the Monte Carlo noise of the finite-sample approximation despite the approximate sampler and ELBO. **(Right)** computation vs. accuracy plot. With $O(10^3)$ denoiser calls per sample, the analytic finite-sample $p(\boldsymbol{a} \mid \boldsymbol{y})$ is achieved with $O(10^{-3})$ W_1 distance.

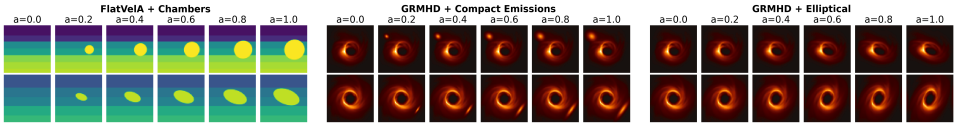

Figure 3: Two prior samples from each discrepancy family at different perturbation (i.e., $\boldsymbol{a}$) levels.

where the numerator can be computed with ELBO for condition $\boldsymbol{a}$. The denominator can be computed by averaging over the rest of the grid, which gives a lower variance estimate than performing a second ELBO on the unconditional head as demonstrated in Section 4.1. Note that the constant $C(\boldsymbol{x})$ from Eq. 5 appears in both the numerator and denominator after the ELBO and cancels out.

Finally, $p(\boldsymbol{a} \mid \boldsymbol{y})$ should not be interpreted in isolation, as it depends on how strongly the measurement operator constrains the proposed discrepancy (Fig. 1, left or middle). We therefore compare the inferred discrepancy posterior against those obtained from synthetic in-distribution, in-family perturbed, and OOD observations to calibrate what discrepancies can meaningfully be distinguished.

4 Experiments

We evaluate PIVOT on an analytic Gaussian, travel time tomography, and black hole imaging. We use $d = 1$ in our experiments to isolate scientific hypotheses and make calibration straightforward; a $d = 2$ ablation is shown in Appendix A.2. Experimental details are described in Appendix A.5.

4.1 Gaussian Ablation

We first validate PIVOT on a multimodal 100-component Gaussian mixture prior and linear inverse problem $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$, $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma_y^2 \boldsymbol{I})$ with $m = 250$, $n = 1000$ and $\sigma_y = 0.1$, whose analytic evidence can help inform us how many posterior samples and ELBO evaluations are necessary to accurately model $p(\boldsymbol{a} \mid \boldsymbol{y})$ for high-dimensional images. We consider a discrepancy family where the mean of each Gaussian component is shifted independently in a way that is (1) orthogonal to the forward operator, so evidence should be flat (Fig. 1, left), (2) weakly constrained by the forward operator, and (3) strongly constrained (Fig. 1, middle and right). We generate posterior samples using the analytic diffusion score of the Gaussian mixture $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_t)$ and the DPMC posterior sampler, which while asymptotically exact, accounts for finite-time sampling errors.

We perform the ablation sweep for both MCMC (Eq. 12) via reflected Langevin, and the 1D grid-based evaluation (Eq. 15) using both the unconditional ELBO and grid-based integration to evaluate the denominator. Results are shown in Fig. 2. The shape of analytic $p(\boldsymbol{a} \mid \boldsymbol{y})$ is recovered up to the finite-sample Monte Carlo error with $O(10^3)$ denoiser calls per sample, indicating that the

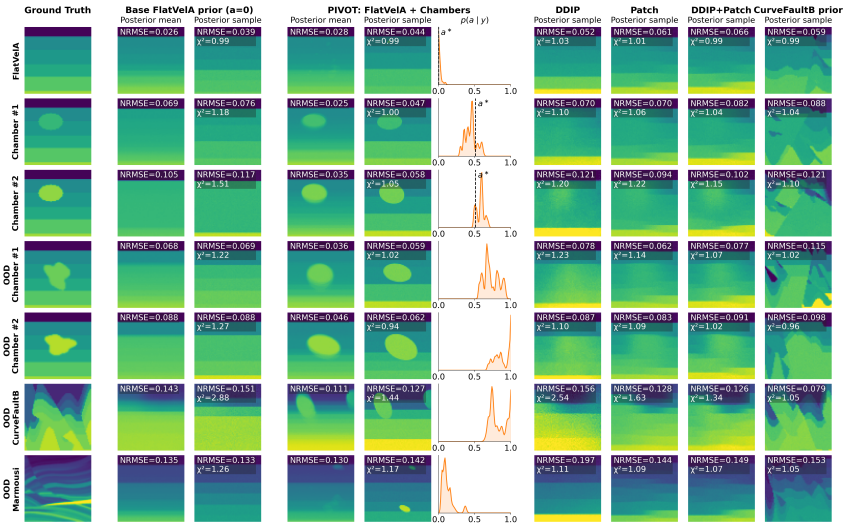

Figure 4: Posterior mean and samples of FlatVelA (in-distribution), Chamber 1 and 2 (in-family), and OOD measurements. For the in-family measurements, PIVOT detects an elliptical feature and places substantial posterior mass near the true discrepancy $\boldsymbol{a}^*$, while detecting similar subsurface features of OOD measurements. The adaptation baselines (DDIP, DDIP+Patch) are not able to distinguish the chamber from a continuous blur, while the CurveFaultB prior imposes extra structure.

finite-time DPMC error and ELBO gap are small in this regime. Furthermore, using the integrated grid to compute the denominator of Eq. 15 results in lower error than ELBO on $p(\boldsymbol{x} \mid \varnothing)$.

4.2 Travel Time Tomography

In seismic imaging, the arrival time of earthquakes at different receivers provides information about the structure under the Earth's surface. Given earthquake location s and receiver location r, the velocity field V is related to travel-time T by the eikonal equation $\|\nabla_r T(s,r)\|^2 = V(r)^{-2}$. Given measurements $y_{ij} = T(s_i, r_j) + \varepsilon_{ij}$, the goal is to recover the velocity field $V(r)$ at all locations r. We use the OpenFWI (Deng et al., 2022) FlatVelA dataset as a strong prior, which imposes discrete layers of increasing velocity. A natural scientific question is whether the subsurface contains previously unmodeled structures, such as magma chambers, that are absent from the prior. To probe this, we design a family of chamber discrepancies (Fig. 3, left). We find that the observations are sensitive to the discrepancy: using the posterior mean $\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{y}]$ as a summary statistic, the in-distribution observations have low discrepancy means while the in-family discrepancy means are high (Fig. 18, in Appendix), making this a case where the perturbation is observable (Fig. 1, right).

We compare PIVOT to adapted FlatVelA via DDIP (Chung & Ye, 2024), FlatVelA patch prior (Hu et al., 2024b), adapted FlatVelA via DDIP+Patch (Hu et al., 2024a), and the more complex OpenFWI CurveFaultB prior in Fig. 4. PIVOT correctly determines that there are no additional chamber features for in-distribution measurements while detecting the chamber for in-family measurements and accurately inferring the true discrepancy $\boldsymbol{a}^*$. DDIP and Patch detect some discrepancy when there is a chamber present, but blur the recovered feature into a diffuse structure; using the CurveFaultB prior imposes substantial additional geological features. For fully OOD CurveFaultB measurements, although the ground truth does not come from the chamber family, a few local features align with the elliptical shape; PIVOT detects mismatch and adds those features. In contrast, for the OOD Marmousi measurement, the fine-scale structure is invisible to the forward operator and there are no chamber-like features, so the data provide little support for increasing the discrepancy strength.

4.3 Case Study: M87* Black Hole Imaging

Black hole images cannot be observed directly; instead they must be inferred from sparse measurements in a highly ill-posed inverse problem whose solutions depend critically on prior assumptions. Physical models, such as General-Relativistic Magnetohydrodynamics (GRMHD) (Mizuno, 2022), can be used to generate black hole simulations from which priors can be learned. However, these models reflect our current understanding of black hole physics, which is likely incorrect.

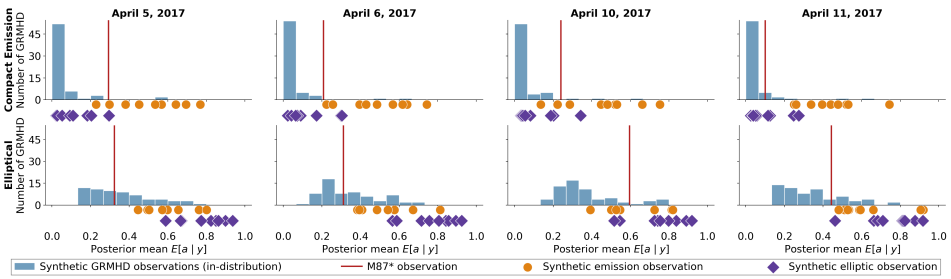

Figure 5: Discrepancy mean $\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{y}]$ for in-distribution synthetic GRMHD observations, in-family perturbed GRMHD observations, and OOD observations for the compact emission (**top**) and elliptic (**bottom**) discrepancies, using the M87* forward model on April 5, 6, 10, and 11. The forward model is sensitive to emission and accurately separates in-distribution and perturbed samples, but poorly constrains ellipticity, often assigning flat or broad posteriors to the elliptic discrepancy prior for in-distribution GRMHD observations. The M87* observation has a consistently high emission discrepancy mean compared to in-distribution GRMHD, suggesting a likely mismatch, while its elliptic discrepancy falls under standard variability of synthetic GRMHD observations.

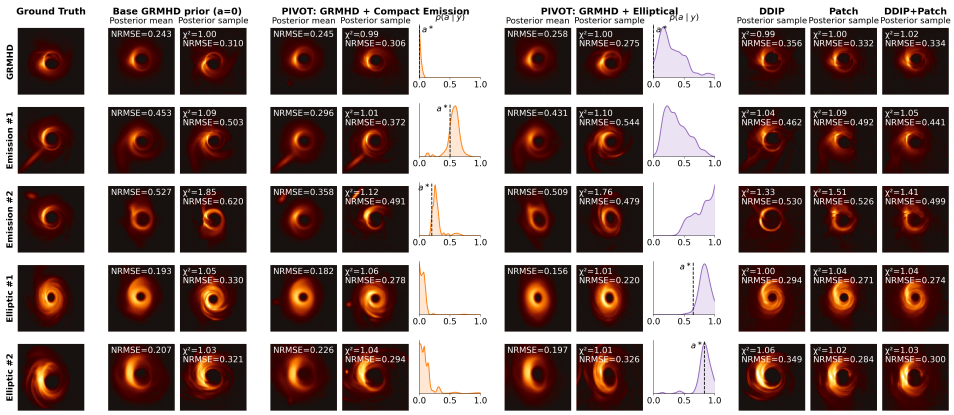

Figure 6: Posterior samples of synthetic in-distribution and OOD measurements under the base GRMHD prior (left), compact emission and ellipticity (middle) discrepancy priors, and baselines (right). The emission prior places substantial posterior mass near the true discrepancy $\boldsymbol{a}^*$ of the two jet observations, while identifying the elliptical observations as not discrepant. The elliptical prior gives uncertain results even for in-distribution GRMHD, indicating that the forward model is too noisy or ill-posed to constrain ellipticity. The baselines struggle to detect either discrepancy.

We obtain measurements from the Event Horizon Telescope network (EHT) (EHTC, 2019b). Given true frequency-domain measurements $\boldsymbol{v}^*$ of the image, we obtain measurements from pairs of telescopes i, j through the forward model $v_{i,j} = g_i g_j e^{i(\phi_i - \phi_j)} v^*_{i,j} + \varepsilon_{i,j}$, where g_i is a station-dependent gain error, ϕ_i is a station-dependent phase error from atmospheric noise and $\varepsilon_{i,j}$ is thermal noise. We define two quantities to decouple the noise sources into approximately independent Bayesian likelihoods: the closure phase $y_{cp} = \angle v_{i,j} v_{j,k} v_{k,i}$ for all minimal sets of 3 telescopes (i, j, k), and the closure amplitude $y_{ca} = \frac{|v_{i,j}||v_{k,l}|}{|v_{i,k}||v_{j,l}|}$ for all minimal sets of 4 telescopes (i, j, k, l) (Thompson et al., 2017; Blackburn et al., 2020). Two major open questions of the M87* black hole data collected in 2017 (EHTC, 2019a) concern the ring ellipticity and the possible presence of additional emission to the southwest (i.e., bottom right) that could correspond to a jet feature (Dahale et al., 2025; Saurabh et al., 2026). Most existing analysis (EHTC, 2019c) computes the likelihood of these features under simplified priors; however, these lack the high-resolution features of GRMHD, which could lead to alternate conclusions. We introduce two families of perturbations to answer these questions: additional compact **emission** and elliptical **deformations** (Fig. 3, middle and right).

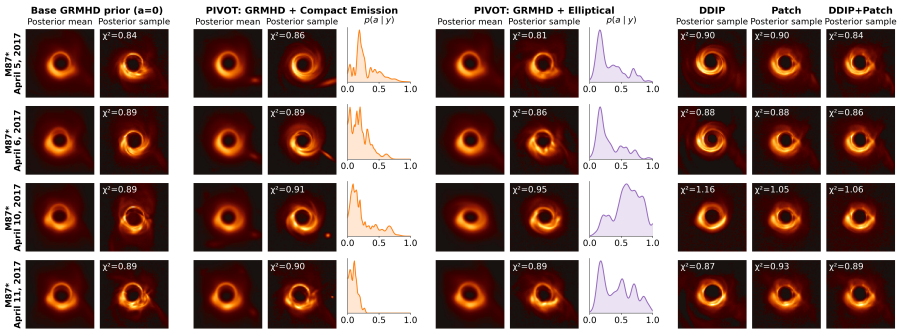

Figure 7: M87* posterior samples across four different days under the base GRMHD prior (left), compact emission and ellipticity (middle) discrepancy priors, and baselines (right). The M87* has a non-trivial peak in the emission discrepancy posterior $p(\boldsymbol{a} \mid \boldsymbol{y})$. Furthermore, the posterior samples indicate that this emission exists only on the bottom right across observations from four different days, indicating that this is likely a feature supported by data rather than posterior uncertainty. On the other hand, because the forward model is not sensitive to ellipticity (Fig. 5, bottom), $p(\boldsymbol{a} \mid \boldsymbol{y})$ is uncertain and any ellipticity conclusions are limited by the ill-posedness of the forward model.

We display the distribution of $\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{y}]$, or mean discrepancy supported by data, (Fig. 5) for 64 in-distribution synthetic observations of GRMHD, 9 synthetic in-family observations across a wide range of discrepancy $\boldsymbol{a}^*$, 9 synthetic observations of the other perturbation family across a wide range of $\boldsymbol{a}^*$, and M87* observations using each day's forward models across all four observation days. We also show posterior samples for selected synthetic observations (Fig. 6) and M87* observations (Fig. 7) under the base GRMHD prior, expanded priors, and baselines; see Appendix A.4 for more examples across a range of discrepancy strengths. For the emission discrepancy, the synthetic in-distribution GRMHD posterior and OOD family have low discrepancy, while the in-family posteriors recover the true discrepancy $\boldsymbol{a}^*$, indicating that the forward model is sensitive to extra emissions under all forward models from all four observation days (Fig. 1, right). The M87* observations also show consistently higher emission discrepancy than most in-distribution observations, at percentiles 96.9, 96.9, 95.3, 89.1 respectively over the four days, consistent with an additional compact emission feature that could be associated with a jet. The few in-distribution samples that have higher discrepancy means than M87* have uncertain posteriors even under the base prior, resulting in emissions across the whole image as a result of ill-posedness (Appendix A.3). On the other hand, the M87* posterior is stable and emissions consistently appear at the bottom right of the image across all four days, with fading intensity over time.

For the elliptical family, while in-family observations have high discrepancy, so do many in-distribution observations and emission family observations, indicating that the forward model is insensitive to ellipticity. The discrepancy posterior is broad and often overestimated at lower true discrepancies (Appendix A.4) and thus not as well suited for hypothesis testing (Fig. 1, left/middle). Furthermore, the M87* elliptic discrepancies are not unusual compared to the ellipticity posterior of in-distribution GRMHD, at percentiles 51.6, 50.0, 87.5, 73.4 over the four days respectively. Thus, we can only conclude that the ellipticity of M87* lies within the natural variability of GRMHD, as is shown in Dahale et al. (2025). For both perturbations, the adaptation baselines are not able to recover the missing features, as the posterior samples achieve reasonable datafit under the base prior; in Appendix A.4 we find that more prominent features are necessary for accurate adaptation.

5 Conclusion

Rather than treating diffusion priors as fixed truths, our framework PIVOT turns prior misspecification into a testable scientific question. By asking which plausible departures are supported by the measurements, it allows the data to determine how prior assumptions should be relaxed. We demonstrate this on controlled problems and on real black hole observations, where PIVOT identifies features or latent variables supported by the data but missing from a narrow simulation prior. More broadly, this offers a path toward using learned priors not only to reconstruct the world, but to reveal where our models of it are incomplete.

AI USE STATEMENT

In this work, we used generative AI tools for [formulate mathematical claims, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks, propose or refine hypotheses, assist with translation, or interpret results] and [generating synthetic data sets, provide critical ingredients for proving mathematical claims, assist in the writing of proofs] are not applicable to this work. Additionally, we used generative AI tools for [create or modify scientific figures or images, create or edit software code, summarize or analyse existing literature, brainstorming, edit a research paper to improve readability, identify relevant literature]. We have reviewed all AI-assisted work. All generated math, text and code was manually verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our method could be used to identify sensitive characteristics from noisy observations, but we do not support that use case.

REPRODUCIBILITY STATEMENT

All code is attached in the supplementary material, all trained models will be released at publication, and all hyperparameters and experimental details are described in Appendix A.5. The OpenFWI dataset is publicly available (Deng et al., 2022). The GRMHD dataset is not available for public use.

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. *arXiv preprint arXiv:2403.08728*, 2024.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. *Advances in Neural Information Processing Systems*, 37:19447–19471, 2024.

Riccardo Barbano, Alexander Denker, Hyungjin Chung, Tae Hoon Roh, Simon Arridge, Peter Maass, Bangti Jin, and Jong Chul Ye. Steerable conditional diffusion for out-of-distribution adaptation in medical image reconstruction. *IEEE Transactions on Medical Imaging*, 44(5):2093–2104, 2025.

Gabriel Missael Barco, Alexandre Adam, Connor Stone, Yashar Hezaveh, and Laurence Perreault-Levasseur. Tackling the problem of distributional shifts: Correcting misspecified, high-dimensional data-driven priors for inverse problems. *The Astrophysical Journal*, 980(1):108, 2025.

James Berger and L Mark Berliner. Robust bayes and empirical bayes analysis with ε-contaminated priors. *the annals of statistics*, pp. 461–486, 1986.

James O Berger, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, et al. An overview of robust bayesian analysis. *Test*, 3(1):5–124, 1994.

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.

Hyungjin Chung and Jong Chul Ye. Deep diffusion image prior for efficient ood adaptation in 3d inverse problems. In *European Conference on Computer Vision*, pp. 432–455. Springer, 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.

Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. *arXiv preprint arXiv:2310.01110*, 2023.

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.

Rohan Dahale, Ilje Cho, Kotaro Moriyama, Kaj Wiik, Paul Tiede, José L Gómez, Chi-kwan Chan, Roman Gold, Vadim Y Bernshteyn, Marianna Foschi, et al. Origin of the ring ellipticity in the black hole images of m87. *Astronomy & Astrophysics*, 699:A279, 2025.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36:288–313, 2023.

Giannis Daras, Alexandros G Dimakis, and Constantinos Daskalakis. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data. *arXiv preprint arXiv:2404.10177*, 2024.

Chengyuan Deng, Shihang Feng, Hanchen Wang, Xitong Zhang, Peng Jin, Yinan Feng, Qili Zeng, Yinpeng Chen, and Youzuo Lin. Openfwi: Large-scale multi-structural benchmark datasets for full waveform inversion. *Advances in Neural Information Processing Systems*, 35:6007–6020, 2022.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

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), 2019b.

EHTC. First m87 event horizon telescope results. vi. the shadow and mass of the central black hole. *The Astrophysical Journal Letters*, 875(1):L6, 2019c.

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.

Angela Gao, Jorge Castellanos, Yisong Yue, Zachary Ross, and Katherine Bouman. Deepgem: Generalized expectation-maximization for blind inversion. *Advances in Neural Information Processing Systems*, 34:11592–11603, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jason Hu, Bowen Song, Jeffrey A Fessler, and Liyue Shen. Patch-based diffusion models beat whole-image models for mismatched distribution inverse problems. *arXiv preprint arXiv:2410.11730*, 2024a.

Jason Hu, Bowen Song, Xiaojian Xu, Liyue Shen, and Jeffrey A Fessler. Learning image priors through patch-based diffusion models for solving inverse problems. *Advances in Neural Information Processing Systems*, 37:1625–1660, 2024b.

Sanyam Kapoor, Wesley J Maddox, Pavel Izmailov, and Andrew G Wilson. On uncertainty, tempering, and data augmentation in bayesian classification. *Advances in neural information processing systems*, 35:18211–18225, 2022.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2206–2217, 2023.

Yosuke Mizuno. Grmhd simulations and modeling for jet formation and acceleration region in agns. *Universe*, 8(2):85, 2022.

Nayantara Mudur, Carolina Cuesta-Lazaro, and Douglas P Finkbeiner. Diffusion-hmc: Parameter inference with diffusion-model-driven hamiltonian monte carlo. *The Astrophysical Journal*, 978 (1):64, 2024.

David J Nott, Max Seah, Luai Al-Labadi, Michael Evans, Hui Khoon Ng, and Berthold-Georg Englert. Using prior expansions for prior-data conflict checking. 2021.

François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682, 2024.

Saurabh, Hendrik Müller, Sebastiano D. von Fellenberg, Paul Tiede, Michael Janssen, Lindy Blackburn, Avery E. Broderick, et al. Probing jet base emission of m87* with the 2021 event horizon telescope observations. *Astronomy & Astrophysics*, 706:A27, 2026. doi: 10.1051/0004-6361/202557022.

Shirin Shoushtari, Edward P Chandler, Yuanhao Wang, M Salman Asif, and Ulugbek S Kamilov. Unsupervised detection of distribution shift in inverse problems using diffusion models. *arXiv preprint arXiv:2505.11482*, 2025.

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, 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, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Tom Sprunck, Marcelo Pereyra, and Tobias Liaudat. Bayesian model selection and misspecification testing in imaging inverse problems only from noisy and partial measurements. *ICML*, 2026.

A Richard Thompson, James M Moran, and George W Swenson. *Interferometry and synthesis in radio astronomy*. Springer Nature, 2017.

Paul Tiede, William Moses, Valentin Churavy, Michael D Johnson, Dominic W Pesce, Lindy Blackburn, and Peter Galison. Hierarchical interferometric bayesian imaging. *The Astrophysical Journal*, 997(2):262, 2026.

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

Frederic Wang and Katherine Bouman. Sample-efficient evidence estimation of score based priors for model selection. In *International Conference on Learning Representations*, volume 2026, pp. 74960–74987, 2026.

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, 2025a.

Zhaoyu Zhang, Yang Hua, Guanxiong Sun, Hui Wang, and Seán McLoone. Training diffusion-based generative models with limited data. In *Forty-second International Conference on Machine Learning*, 2025b.

Yaxuan Zhu, Zehao Dou, Haoxin Zheng, Yasi Zhang, Ying Nian Wu, and Ruiqi Gao. Think twice before you act: Improving inverse problem solving with mcmc. *URL https://arxiv.org/abs/2409.08551*, 1(2), 2024.

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*, pp. 1219–1229, 2023.

A Appendix

A.1 Why tempering diffusion priors is problematic

Tempering the prior has been a common nonparametric approach for weakening prior assumptions for both classical Bayesian analysis and deep learning (Kapoor et al., 2022). However, for high-dimensional image priors, tempering changes the expected energy distribution of its samples. As an example, for the standard multivariate normal $p(\boldsymbol{x}) = \mathcal{N}(0, \boldsymbol{I})$ in $\mathbb{R}^d$, its samples $\boldsymbol{x}$ have expected energy $\mathbb{E}\|\boldsymbol{x}\|^2 = d$; while the tempered $p(\boldsymbol{x})^{0.5} \propto \mathcal{N}(0, 2\boldsymbol{I})$ has samples $\boldsymbol{x}$ with expected energy $\mathbb{E}\|\boldsymbol{x}\|^2 = 2d$.

To empirically demonstrate the problem on diffusion priors, we use annealed MCMC (Du et al., 2023) to sample from the tempered GRMHD diffusion prior. See Fig. 8 for a visual example. The energy distribution changes and the samples end up all concentrated at increasingly high energy as the tempering exponent β decreases, rather than becoming more general prior as expected.

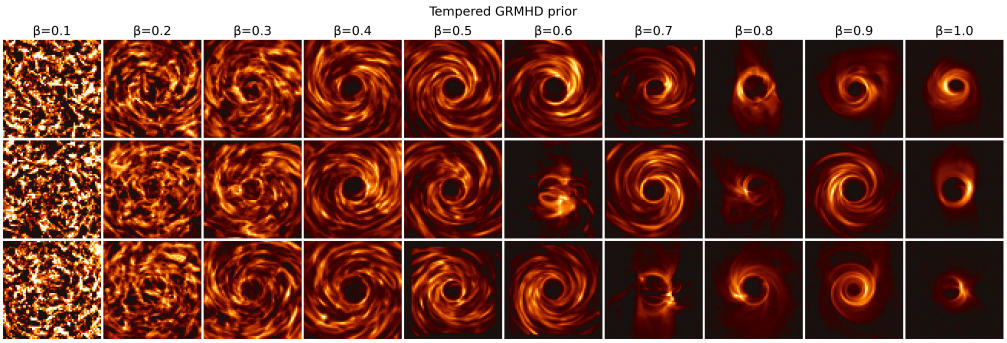

Figure 8: Samples of the tempered GRMHD prior, $p(\boldsymbol{x})^{\beta}$ with $\beta \in [0.1, 1.0]$.

A.2 Gaussian Ablation on 2D Discrepancy

In this section, we perform a Gaussian ablation but with a 2D discrepancy $\boldsymbol{a}$, demonstrating that PIVOT can be used to jointly infer multi-dimensional discrepancy hypotheses. We use the same multimodal Gaussian mixture prior, linear inverse problem, and mean-shift perturbation setup described in 4.1. We also still use the analytic diffusion score, DPMC sampling, and ELBO evaluation to match real use cases. With $\boldsymbol{a} \in [0, 1]^2$, we consider three cases: (1) the discrepancy is completely orthogonal to the observations (orthogonal), (2) one component of the discrepancy is visible to the observations (Weak), and (3) both components are visible to the observations (Strong).

Results are displayed in Fig. 9. More posterior samples are required to reduce the finite-sample Monte Carlo error, with the orthogonal case remaining the most difficult similar to the 1D result. Furthermore, MCMC becomes more compute-efficient than grid-based evaluation.

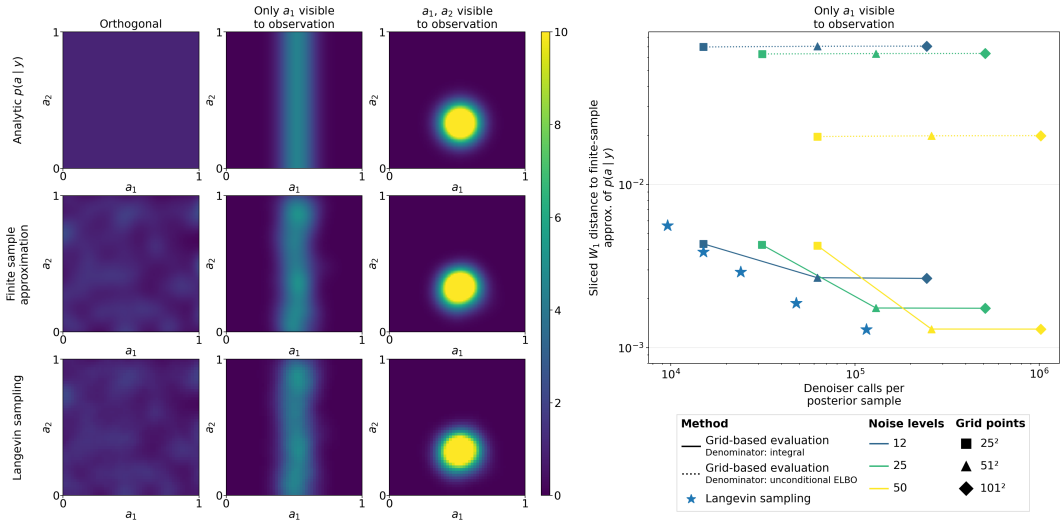

Figure 9: **(Left)** Discrepancy posterior $p(\boldsymbol{a} \mid \boldsymbol{y})$ results on a Gaussian toy experiment where the forward operator has an increasing level of sensitivity to the perturbation (orthogonal, weak, strong) and $\boldsymbol{a} \in [0, 1]^2$. The ideal analytic $p(\boldsymbol{a} \mid \boldsymbol{y})$, analytic finite-sample (50) approximation of $p(\boldsymbol{a} \mid \boldsymbol{y})$, and Langevin sampled discrepancy are shown. The density shape is correctly recovered up to the Monte Carlo error of the finite-sample approximation despite the approximate sampler and ELBO. In comparison to the 1D $\boldsymbol{a}$ case, the finite-sample Monte Carlo error is higher. **(Right)** computation vs. accuracy plot for the partial sensitivity case. In comparison to the 1D $\boldsymbol{a}$ case, MCMC becomes a more efficient alternative compared to grid-based integration.

A.3 SELECTED POSTERIORS FOR BLACK HOLE IMAGING

In Section 4.3, we show that on April 5 and 6, 2017, the M87* observations have higher compact emission discrepancy than in-distribution synthetic observations, except for two GRMHD observations (which were the same two GRMHD samples on both days). We display the posterior mean, standard deviation, and 10 randomly selected posterior samples for these two in-distribution GRMHD under both the base GRMHD prior as well as the expanded prior with emissions in Fig. 10. We also display the posterior mean, standard deviation, and 10 randomly selected posterior samples of M87* on April 5 and 6 in Fig. 11.

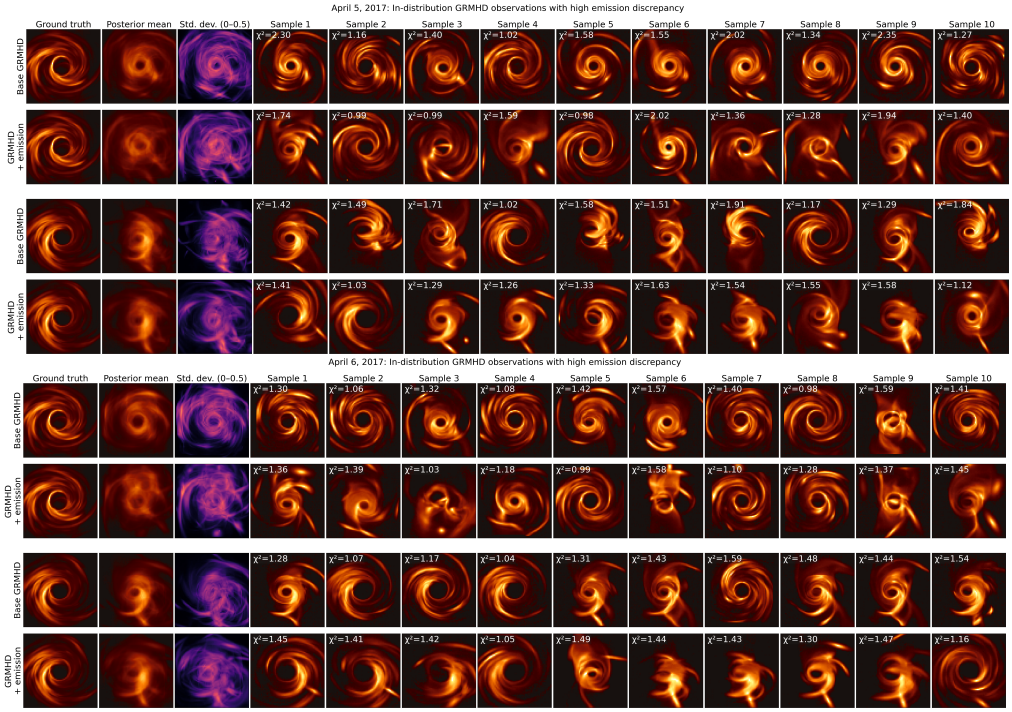

Figure 10: Posterior mean, standard deviation, and samples of the two GRMHD with high emission discrepancy, using forward operators from April 5/6 and both the base GRMHD prior and the emission expanded prior. The posterior is highly multimodal and uncertain, resulting in random emission-like features all over the image even under the base prior.

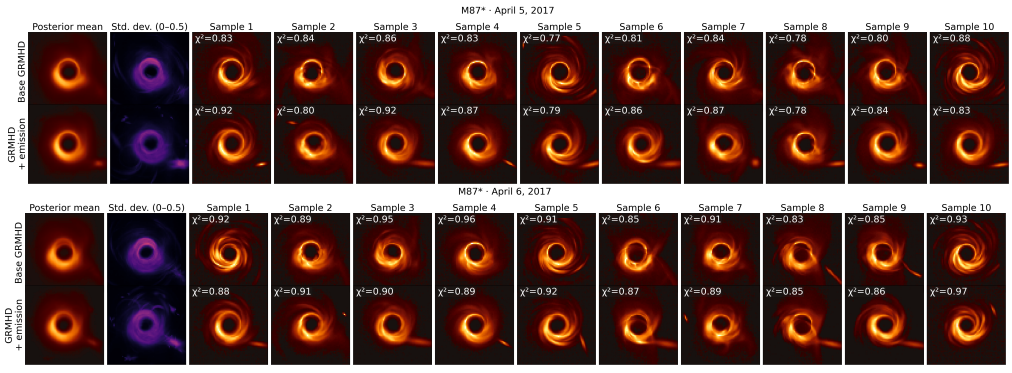

Figure 11: Posterior mean, standard deviation, and samples of M87* from April 5/6 under both the base GRMHD prior and the emission expanded prior. The posterior is concentrated, with much lower variance and the emission is almost always located in the bottom-right.

A.4 Sensitivity under Measurement Noise and Perturbation Strengths

In this section, we perform an ablation over decreasing measurement noise levels in the travel time tomography in-family example (Fig. 12), where we observe that there is much lower discrepancy posterior spread, or uncertainty, at lower measurement noise levels. We also demonstrate how increasing perturbation strength (feature prominence) in black hole emissions (Fig. 13 and 14) and ellipticity (Fig. 15 and 16) affects the estimate. For the emission discrepancy, PIVOT is able to infer posteriors that assign large mass to the true discrepancy $\boldsymbol{a}^*$ across a large range of measurement noise or feature prominence, while the elliptical discrepancy consistently overestimates the true discrepancy due to confounding effects in the forward model.

We also demonstrate that DDIP methods perform the best when there is significant conflict between the observations and the prior; this can be achieved by having low observation noise, a less ill-posed observation, or a stronger mismatched feature. In the travel time tomography example, decreasing the measurement noise σ_y results in monotonically improved reconstruction of the subsurface chamber. For the black hole imaging example, we see that increasing the feature's prominence can result in more accurate adaptations, such as the elongated photon ring in Fig. 16, bottom. However, in the second example, for the adaptation to be successful the feature has to be strong enough that it is even recovered under the base GRMHD prior, as shown in the bottom row of Fig. 14.

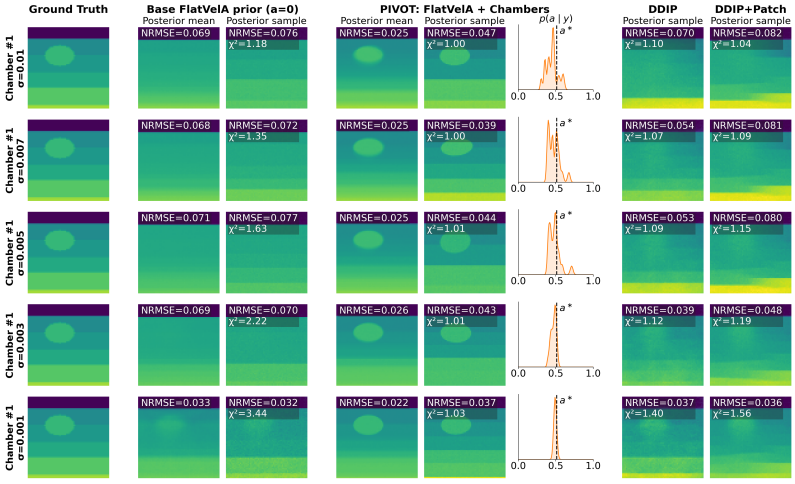

Figure 12: Travel Time Tomography results with decreasing observation noise. Lower noise results in a more accurate adaptation as well as lower uncertainty in the discrepancy posterior $p(a \mid y)$.

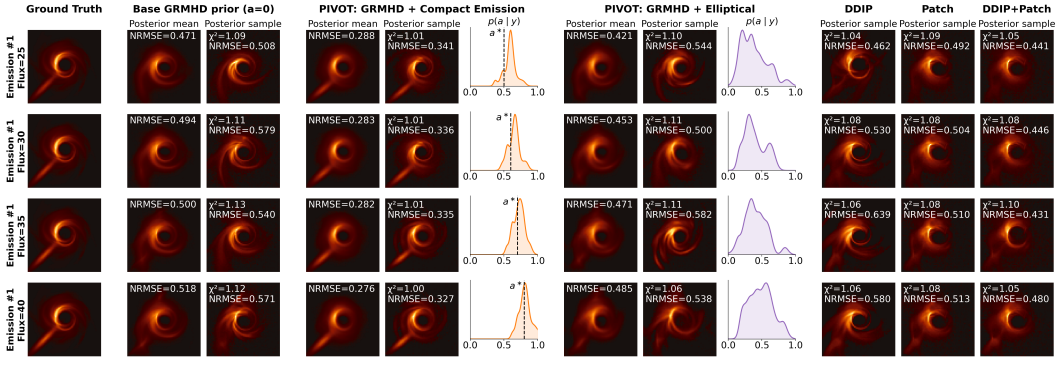

Figure 13: Black hole imaging results with increasing emission feature prominence, in flux. A more prominent feature results in a clearer emission in the adaptation, while the emission discrepancy posterior correctly tracks the ground truth discrepancy $\boldsymbol{a}^*$, and the elliptic discrepancy remains low.

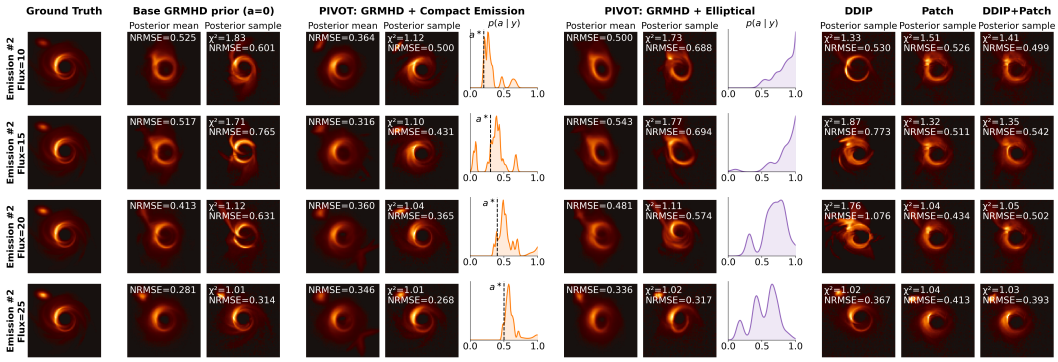

Figure 14: Black hole imaging results with increasing emission feature prominence, in flux. The adaptation is only successful when the feature is so strong that it gets reconstructed under the base GRMHD prior. Meanwhile, the emission discrepancy posterior correctly tracks the ground truth discrepancy $\boldsymbol{a}^*$, and the elliptic discrepancy remains low.

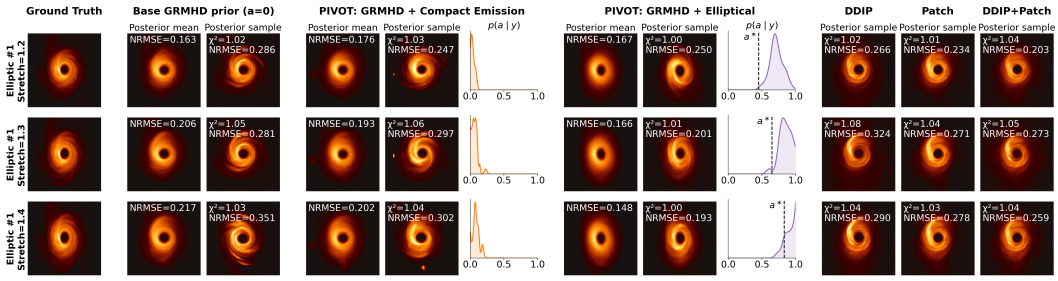

Figure 15: Black hole imaging results with increasing ellipticity. The adaptation is not successful and the photon ring remains circular. Meanwhile, the emission discrepancy posterior correctly determines that the mismatch is not well explained by additive compact emission, and the elliptic discrepancy is overestimated at lower discrepancies compared to the true a^* due to the limitations of the forward model.

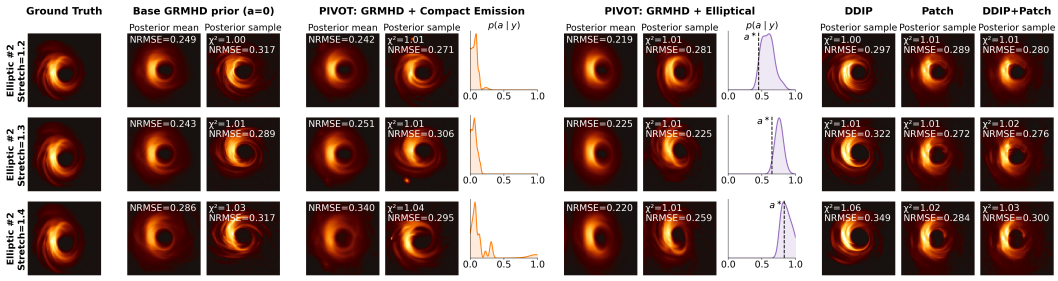

Figure 16: Black hole imaging results with increasing ellipticity. For this example, both the patch prior and the patch adaptation (DDIP+Patch) are successful only at the high 1.4x stretch, and the photon ring is elongated. Meanwhile, the emission discrepancy posterior correctly determines that the mismatch is not well explained by additive compact emission, and the elliptic discrepancy is overestimated at lower discrepancies compared to the true a^* due to the limitations of the forward model.

A.5 Experimental Details

A.5.1 Hyperparameters

To train our conditional diffusion models, we used a SongUNet with standard EDM (Karras et al., 2022) preconditioning and noise sampling. The OpenFWI priors were resized to 64×64, and the GRMHD prior is resized to 64×64 and then augmented with random zoom between $[0.75, 1.25]$. Both the travel time tomography and black hole imaging diffusion models used 128 channels, channel mult $[1, 2, 2]$, 1 residual block, and 16×16 attention. We use $p_{dropout} = 0.2$ and train for 50000 optimizer steps, which took 2.5 hours per model on an NVIDIA A100.

During inference, we used DPMC (Zhu et al., 2024) for posterior sampling, which performs annealed MCMC along $p(\boldsymbol{x}_t)p(\boldsymbol{y} \mid \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t])^{\beta_t}$. For all methods, we use $\sigma_{max} = 50, \sigma_{min} = 0.01$, and the Karras $\rho = 7$ noise schedule for sampling. We also parameterize $\beta_t = C/(C + \sigma_t^2)$ as a surrogate modeling the uncertainty of $p(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)$ that is missing from using the point estimate $\mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_t]$. We sample 25 posterior samples for each observation, and for ELBO computation, we use 101 gridpoints and 25 noise levels with $\sigma_{max} = 5$. We list the shared hyperparameters and runtime on one NVIDIA A100 here:

Hyperparameter	**Gaussian**	**Seismic Imaging**	**Black Hole Imaging**
Annealing steps	200	200	300
Corrector steps	20	20	20
Corrector LR	$0.3\sigma_t^2$	$0.005\sigma_t^2$	$0.05\sigma_t^2$
C	0.1	10	0.02
Sampling $x \mid y$ runtime (25 samples)	<0.1 min	9.3 min	6.2 min
ELBO grid $a \mid y$ runtime (25 samples)	<0.1 min	5.3 min	5.3 min

The Gaussian ablation MCMC used Langevin dynamics lr=0.0005 with 200 burn-in steps.

We tuned DDIP and DDIP+Patch independently on synthetic observations for each task, but we found that the same task-specific parameters work well with both methods. We chose to finetune all network weights during adaptation which has been demonstrated to perform better (Hu et al., 2024a). For patch-based methods, we choose a random tiling offset at each sampling step as is done in Hu et al. (2024b). We report tuned hyperparameters and runtimes on one A100 below:

Hyperparameter	**Seismic Imaging**	**Black Hole Imaging**
Network LR	0.0001	0.00003
Network updates	1	3
Reconstruction likelihood LR	0.01	0.01
Reconstruction likelihood updates	3	3
Adaptation count	10	30
DDIP runtime (1 recon)	4.4 min	2.7 min
DDIP+Patch patch size	16	16
DDIP+Patch runtime (1 recon)	4.4 min	2.8 min

Finally, in black hole imaging, closure quantities remove absolute phase information, the reconstruction is invariant to image space translations. Therefore, we align posterior samples via normalized cross-correlation before computing reconstruction metrics or posterior mean / standard deviation. We do use the potentially offset original images for ELBO computation, for consistency with the diffusion model's distribution.

A.5.2 DISCREPANCY PARAMETERIZATIONS

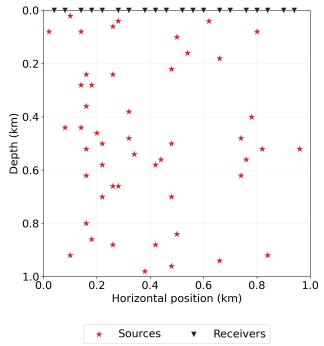

Figure 17: Locations of the 49 earthquake sources and 20 receivers used in the synthetic Travel Time Tomography experiments.

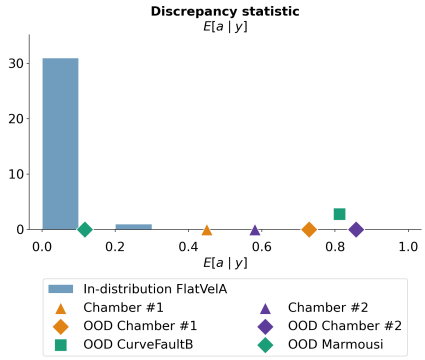

Figure 18: Mean discrepancy $\mathbb{E}[\boldsymbol{a} \mid \boldsymbol{y}]$ for in-distribution observations $\boldsymbol{y}$, in-family synthetic chamber observations $\boldsymbol{y}$, and OOD observations $\boldsymbol{y}$. The forward model is sensitive to the discrepancy.

For the travel time tomography experiment, we simulate 20 receivers and 49 earthquake sources in a 1 km x 1 km area with observation noise of $\sigma = 0.01$ seconds, and use factored fast marching to compute likelihood gradients: the same setup described in Gao et al. (2021); see Fig. 17. We parameterize our chamber perturbation $\boldsymbol{a} \in [0, 1]$ as an ellipse with uniform random center, axis ratio sampled uniformly from $[0.4, 1.0]$, velocity uniformly between at least 0.5 km/s higher than the background and 4.5 km/s, and the conditioning $\boldsymbol{a} \in [0, 1]$ represents the area $600\boldsymbol{a}$ pixels^2. We also impose a uniform $p(\boldsymbol{a})$. This perturbation is well-detected by observations; see Fig. 18.

For black hole imaging, our emission discrepancy family consists of off-ring Gaussians with centers located between a radial distance of 25 and 30 pixels from the image center, axis ratio sampled uniformly from $[0.2, 1.0]$, rotation sampled uniformly from $[0, 2\pi]$, peak amplitude value sampled uniformly from $[0.5, 1.0]$, and the condition $\boldsymbol{a} \in [0, 1]$ represents added flux of $50a$ in normalized diffusion coordinates. The elliptical family maps $\boldsymbol{a} \in [0, 1]$ to an area-preserving anisotropic deformation of ratio $1.5^{\boldsymbol{a}}$ along a random axis. We also impose a uniform $p(\boldsymbol{a})$ for both families.