Outlier-robust Diffusion Posterior Sampling for Bayesian Inverse Problems

Yiming Yang[1*] **Xiaoyuan Cheng**[1] **Yi He**[1] **Kaiyu Li**[1] **Wenxuan Yuan**[2] **Zhuo Sun**[3,4†]
[1]University College London, [2]Imperial College London,
[3]School of Statistics and Data Science, Shanghai University of Finance and Economics,
[4]Institute of Big Data Research, Shanghai University of Finance and Economics

Abstract

Diffusion models have emerged as powerful learned priors for Bayesian inverse problems (BIPs). Diffusion-based solvers rely on a presumed likelihood for the observations in BIPs to guide the generation process. Likelihood misspecification is common in practical BIPs and is known to degrade recovery performance, particularly under outlier contamination. We investigate this problem by first characterizing the induced posterior deviation and proving the *stability* of diffusion-based solvers for linear BIPs. Our stability analysis further reveals potential robustness deficiencies of existing diffusion-based solvers under outlier-contaminated measurements. To address this issue, we propose a simple yet effective solution: *robust diffusion posterior sampling*, which is provably *outlier-robust* for linear BIPs and compatible with existing gradient-based posterior samplers. Empirical results from scientific inverse problems and natural image tasks demonstrate the effectiveness and robustness of our method, with consistent performance gains in challenging scenarios involving outlier contamination for both linear and nonlinear tasks. Our implementation is provided at `https://github.com/yyiminguc l/RDP`.

1 Introduction

The goal of Inverse Problems (IPs) is to recover the unknown variable $\boldsymbol{x}$ given noisy measurements $\boldsymbol{y} = F(\boldsymbol{x}) + \boldsymbol{\epsilon}$. Because measurements are limited and noisy, IPs are often ill-posed [De Vito et al., 2005]. IPs appear in various domains, such as medical imaging [Song et al., 2021c], geophysics [Sambridge and Mosegaard, 2002], and beyond. Bayesian inference mitigates the ill-posedness via a prior $p(\boldsymbol{x})$ to regularize the solution based on prior knowledge. The likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ quantifies how well $F(\boldsymbol{x})$ explains the measurements $\boldsymbol{y}$. Applying Bayes' rule yields the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$. This formulation is referred to as a Bayesian inverse problem (BIP), and the posterior is viewed as the solution [Sullivan, 2015].

Despite the simple formulation of BIPs, determining a suitable prior $p(\boldsymbol{x})$ is both critical and challenging, particularly for high-dimensional complex data distributions. Apart from prior choices in existing works [Knapik et al., 2011, Wang et al., 2017, Hosseini and Nigam, 2017, Chung et al., 2023, Cardoso et al., 2024], probabilistic diffusion models (DMs) [Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Ho et al., 2020, Song et al., 2021d] have emerged as powerful data-driven priors for solving inverse problems across various scenarios and applications [Zheng et al., 2025]. DMs identify the target data distribution by reversing a forward diffusion process via a learned time-dependent score function $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$, which progressively transports samples from Gaussian noise at time T to data samples from the target distribution $p(\boldsymbol{x}_0)$. Unlike standard BIPs, applying diffusion priors requires estimating intermediate posteriors $p(\boldsymbol{x}_t|\boldsymbol{y})$ along the reverse diffusion process. These intermediate

*First Author: Email Yiming Yang to `zcahyy1@ucl.ac.uk`.
†Corresponding Author. Correspondence to Zhuo Sun: `sunzhuo@mail.shufe.edu.cn`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

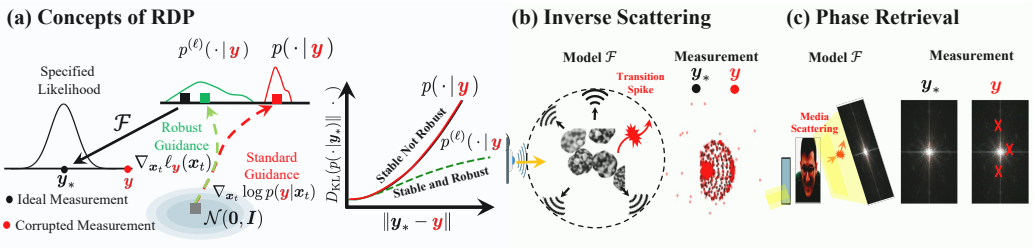


Figure 1: Overview. (a) Our proposed method achieves robust recovery under corruption. Visualization of the measurement model and data corruption, showing the sources of measurement corruption in (b) inverse scattering and (c) phase retrieval.

distributions are generally intractable. A widely adopted approach to address the intractability relies on *Tweedie's formula* [Miyasawa et al., 1961] to obtain an estimate of the denoised conditional $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$, effectively providing a point estimate (or a local Gaussian approximation) [Song et al., 2021a, Meng et al., 2021, Chung et al., 2023, Song et al., 2023a, Boys et al., 2024]. This allows for a tractable likelihood approximation that combines with the diffusion prior to estimate intermediate posteriors and progressively solves the IP; see [Daras et al., 2024] for a review.

On Stability for Diffusion-Based Solvers. Diffusion-based solvers have shown strong empirical performance for BIPs, but their stability with respect to measurement error remains less explicitly characterized. Prior works analyze surrogate likelihoods based on Tweedie approximations [Chung et al., 2023, Song et al., 2023b], while others study stability under forward-model shifts in linear BIPs [Renaud et al., 2024] or sampling stability for inexact Langevin-type algorithms [Renaud et al., 2026]. Complementary to these results, we study stability under measurement perturbations. Specifically, we show that the posterior induced by DPS [Chung et al., 2023] is stable under measurement error and quantify the posterior deviation, which scales with the magnitude of measurement noise.

On Robustness for Diffusion-Based Solvers under Outlier Contamination. A finding from our stability results indicates the potential lack of robustness of existing diffusion-based methods for BIPs under outlier contamination. Prior work has studied sensitivity to denoiser shifts and forward-model shifts in linear BIPs [Renaud et al., 2024], while robustness to outlier contamination in measurement noise is also receiving increasing attention [Zheng et al., 2026]. However, robustness to outlier contamination in the measurement noise remains largely underexplored. In BIPs, the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ is determined by the measurement model F and an assumed noise distribution $\boldsymbol{\epsilon}$ which can be misspecified in practice. For example, in optical tasks such as inverse scattering and phase retrieval, speckle noise or unexpected scattering from inhomogeneous media can produce heavy-tailed noise or outlier contamination, thereby violating the common Gaussian assumptions (see Figure 1(b)–(c)). In this setting, stable but non-robust diffusion-based solvers can drive the posterior far from the target one. One way to deal with likelihood misspecifications is to temper the likelihood with heuristic *temperatures* [Zhu et al., 2023, Chung et al., 2023, Song et al., 2023a,b, Wu et al., 2023, Cardoso et al., 2024], thereby globally down-weighting the likelihood term in the reverse diffusion processes. However, this global control is intended to stabilize the reverse sampling dynamics rather than explicitly mitigate the effect of likelihood misspecifications for BIPs.

Contributions. To address this problem, we provide a systematic investigation through theoretical and empirical analysis. We summarize our contributions as follows:

- We establish observations and theoretical results on the *stability of diffusion posterior sampling* for linear Bayesian inverse problems under measurement noise, providing insight into their empirical effectiveness.
- We show that this stability analysis also reveals a vulnerability under outlier contamination. Motivated by this finding, we introduce *robust diffusion posterior sampling* (RDP), a GBI-inspired framework for robust BIPs. RDP serves as a plug-and-play robust guidance module for existing gradient-based diffusion samplers, as illustrated in Figure 1(a), and provides theoretical guaranties of *outlier robustness*.

- We empirically validate our theory across linear and nonlinear scientific inverse problems and image restoration tasks. The results show that RDP mitigates the outlier contamination visualized in Figure 1(b,c), improves performance under likelihood misspecification, and maintains competitive performance in well-specified settings.

2 Background

In this section, we provide the background on BIPs and diffusion-based solvers for these problems.

2.1 Bayesian Inverse Problems

An inverse problem aims to recover the unknown variables $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ with compact $\mathcal{X}$ from noisy observations $\boldsymbol{y} \in \mathbb{R}^{d_y}$ given by:

$$\boldsymbol{y} = F(\boldsymbol{x}) + \boldsymbol{\epsilon}, \tag{1}$$

where $F\colon \mathcal{X} \to \mathbb{R}^{d_y}$ denotes the measurement model, and $\boldsymbol{\epsilon} \in \mathbb{R}^{d_y}$ denotes the additive measurement noise [Sullivan, 2015]. We consider linear inverse problems, i.e. $F \in \mathbb{R}^{d_y \times d_x}$ for theoretical analysis in this work; though we also observe that our method shows promising empirical results for nonlinear BIPs as well (see Section 4). Typically $d_x \gg d_y$, the inverse problem is under-determined and requires regularization from prior knowledge.

Accordingly, the BIP models the unknown $\boldsymbol{x}$ as a random variable by specifying a prior $p(\boldsymbol{x})$ on $\mathcal{X}$ and a likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ induced by (1) [Stuart, 2010, Sullivan, 2015]. In this paper, we focus on the commonly used Gaussian likelihood, which factorizes across measurement components, i.e., $p(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{d_y} p(y_i|\boldsymbol{x})$. Given an observation $\boldsymbol{y}$, Bayes' rule yields the posterior $p(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x})\, p(\boldsymbol{x})$. However, traditional priors, such as the standard Gaussian, are often inadequate for representing the complex distributions of real-world data. Diffusion models, instead, offer a compelling choice for BIPs by learning expressive, data-driven priors [Song et al., 2021c,d].

2.2 Diffusion Models for Inverse Problems

Diffusion models provide a powerful framework for learning complex, high-dimensional distributions. In this work, we adopt the continuous-time Variance Preserving (VP) formulation. The forward process transforms a data sample $\boldsymbol{x}_0$ into Gaussian noise over $t \in [0, T]$, governed by the stochastic differential equation (SDE) $\mathrm{d}\boldsymbol{x}_t = -\frac{1}{2}\beta(t)\boldsymbol{x}_t\,\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\boldsymbol{w}$, where $\beta(t)$ is a predefined noise schedule and $\boldsymbol{w}_t$ is a standard Wiener process. The key insight is that the forward process can be reversed in time using:

$$\mathrm{d}\boldsymbol{x}_t = -\beta(t)[\frac{\boldsymbol{x}_t}{2} + \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)]\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\boldsymbol{w}_t. \tag{2}$$

The critical component required to solve the reverse SDE is the score function, $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$, which is approximated by a time-dependent neural network $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$. This network is trained via denoising score matching, minimizing the expected squared error between the predicted and true conditional scores. Once trained, the score network $\boldsymbol{s}_{\boldsymbol{\theta}}$ is used to generate samples by solving the reverse SDE, starting from $\boldsymbol{x}_T \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

Posterior Sampling with Diffusion Models. Solving a BIP requires sampling from the posterior $p(\boldsymbol{x}_0|\boldsymbol{y})$. This can be achieved by guiding the reverse SDE using the score of the posterior distribution at each time step $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{y})$. Using Bayes' theorem, this posterior score can be decomposed into a prior and a likelihood:

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{y}) = \underbrace{\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)}_{\text{Prior Score}} + \underbrace{\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y}|\boldsymbol{x}_t)}_{\text{Likelihood Score}}.$$

The prior score is readily available from the pre-trained score network, i.e., $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$. The likelihood term is difficult to derive, as the time-dependent likelihood function $p(\boldsymbol{y}|\boldsymbol{x}_t) = \int p(\boldsymbol{y}|\boldsymbol{x}_0)p(\boldsymbol{x}_0|\boldsymbol{x}_t)\,\mathrm{d}\boldsymbol{x}_0 = \mathbb{E}\left[p(\boldsymbol{y}|\boldsymbol{x}_0)\right]$ is not tractable in general.

An efficient strategy is to approximate the intractable likelihood $p(\boldsymbol{y}|\boldsymbol{x}_t)$ with a surrogate $p(\boldsymbol{y}|\boldsymbol{x}_t) \approx \tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$, which is the likelihood evaluated at $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$, the estimate of $\boldsymbol{x}_0$ at step t from the

diffusion model [Stein, 1981, Efron, 2011, Graikos et al., 2022, Chung et al., 2023]. We refer to this surrogate likelihood as *DPS* likelihood. While the DPS approximation has proven effective for well-specified likelihood models [Song et al., 2021a, Ho et al., 2022, Wu et al., 2023, Song et al., 2023b], its robustness to likelihood misspecifications remains a concern. Likelihood misspecifications can arise from the misspecified noise distribution or from the presence of a contaminated subset of observations generated under a different data-generating mechanism. In this work, we investigate the impact of such misspecifications and focus on the following questions:

Is diffusion posterior sampling robust to outlier contamination for linear BIPs? If not, can we propose a solution that guaranties robust recovery without losing effectiveness?

3 Method

This section begins by establishing the stable properties of the standard diffusion sampler, which also shows its vulnerability to outlier contamination. We then formulate a robust strategy with theoretical guaranties and detail its practical implementation for general BIPs with diffusion solvers in Section 3.2. We discuss how our robust approach performs in nominal settings. We consider observations contaminated by outliers: $\boldsymbol{y} = F\boldsymbol{x} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ may contain both nominal measurement noise and sparse, large-magnitude outlier corruption. Hereby, the method achieves the design of a posterior that is robust to the presence of outliers.

3.1 Stability and Robustness of Diffusion Posterior

Following the likelihood approximation in Section 2.2, the likelihood score is evaluated as $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y}|\boldsymbol{x}_t) \approx \nabla_{\boldsymbol{x}_t} \log \tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$. The $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ is the denoised estimate of the clean data given $\boldsymbol{x}_t$, defined by:

$$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = \mathbb{E}(\boldsymbol{x}_0|\boldsymbol{x}_t) \approx \frac{1}{\sqrt{\alpha(t)}}(\boldsymbol{x}_t + (1-\alpha(t))\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)),$$

where $\alpha(t) = \exp\left(-\int_0^t \beta(\tau)\mathrm{d}\tau\right)$. Then, one can perform posterior sampling by solving the reverse SDE in Equation 2, substituting the unconditional score $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ with the approximated posterior score $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{y}) = \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) + \nabla_{\boldsymbol{x}_t} \log \tilde{p}(\boldsymbol{y}|\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$. While Chung et al. [2023] and Song et al. [2023b] analyze the DPS likelihood error, their link to the resulting posterior is indirect. The following Lemma 3.1 bridges this gap by establishing the stability of the DPS solver.

Lemma 3.1. *Consider the posterior $p_\delta(\cdot \mid \boldsymbol{y})$ induced by the DPS solver under assumptions D.1–D.2 in Appendix D with the terminal stopping time δ. Suppose the likelihood is linear Gaussian $\boldsymbol{y} \mid \boldsymbol{x} \sim \mathcal{N}(F\boldsymbol{x}, \sigma_y^2 \boldsymbol{I})$. Then, for any finite noisy measurement $\boldsymbol{y} \in \mathbb{R}^{d_y}$ and any $\boldsymbol{x} \in \mathcal{X}$, let the corresponding noiseless measurement be $\boldsymbol{y}_* = F\boldsymbol{x}$. There exists a finite constant $C_{\mathrm{stable}} > 0$ such that*

$$D_{\mathrm{KL}}\left(p_\delta(\cdot \mid \boldsymbol{y}_*) \,\|\, p_\delta(\cdot \mid \boldsymbol{y})\right) \leq C_{\mathrm{stable}} \|\boldsymbol{y}_* - \boldsymbol{y}\|_2^2. \tag{3}$$

The detailed proof is provided in Appendix D. Lemma 3.1 indicates that the sensitivity is bounded by a term quadratic in the measurement noise magnitude $\|\boldsymbol{y}_* - \boldsymbol{y}\|_2$. This result confirms the stability of the diffusion prior under the DPS likelihood approximation, validating its effectiveness for moderate noise levels that align with the assumed Gaussian likelihood model [Chung et al., 2023]. However, since the bound depends on the measurements, it may grow substantially under outlier contamination, as observed empirically in Figure 2. Hence, the result establishes stability but does not provide a robustness guaranty against outlier contamination, motivating our robust posterior guidance.

Before proceeding, we first provide the definition of robustness by quantifying how the posterior distribution changes in response to measurement perturbations. Following Altamirano et al. [2023], Duran-Martin et al. [2024], the posterior influence function (PIF) is defined as the KL divergence such that $\mathrm{PIF}_{\boldsymbol{y}_*}(\boldsymbol{y}) \coloneqq D_{\mathrm{KL}}\left(p(\boldsymbol{x}|\boldsymbol{y}_*) \| p(\boldsymbol{x}|\boldsymbol{y})\right)$. The posterior is called *outlier-robust* if the PIF is uniformly bounded, i.e., $\sup_{\boldsymbol{y} \in \mathbb{R}^{d_y}} |\mathrm{PIF}_{\boldsymbol{y}_*}(\boldsymbol{y})| < \infty$.

3.2 Robust Diffusion Posterior

To address this issue, we consider generalized Bayes (GB) [Sullivan, 2015] in this work. GB considers alternative differentiable loss functions $\ell_{\boldsymbol{y}} : \mathcal{X} \to \mathbb{R}$ in place of the negative log-likelihood

in Bayes' rule. The subscript $\boldsymbol{y}$ denotes the dependence on the observation $\boldsymbol{y}$. The loss $\ell_{\boldsymbol{y}}$ can be flexibly specified under mild regularity conditions; see Section 6.3 of Sullivan [2015]. This yields a generalized posterior [Matsubara et al., 2022, Wild et al., 2023, Altamirano et al., 2023] in the following form: $p^{(\ell)}(\boldsymbol{x}|\boldsymbol{y}) \propto p(\boldsymbol{x}) \exp\{-\tau\, \ell_{\boldsymbol{y}}(\boldsymbol{x})\}$ where $\tau > 0$ is a temperature parameter that controls the influence of the loss relative to the prior. Many choices of ℓ (e.g., density-power) are well established and show improved performance under likelihood misspecification [Huber, 2011, Song and Huang, 2024].

Such settings assume multiple observations, with robustness defined in terms of the contamination of only part of the dataset. Differently, BIPs focus on the posterior conditioning on a *single* (high-dimensional) measurement, where contamination is structured and often occurs at the level of coordinates or components, e.g., pixel-wise outliers in images; see Figure 1(b)–(c), thereby making standard GB constructions nontrivial to apply in this setting.

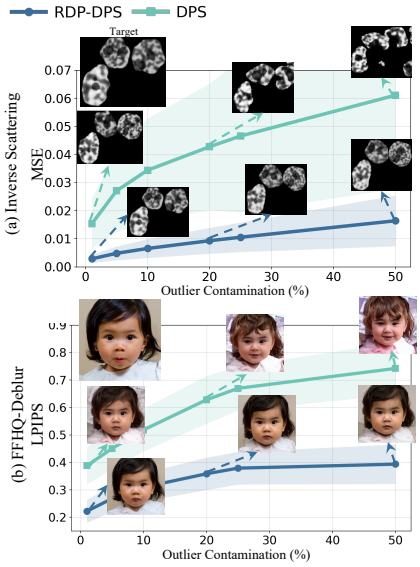


Figure 2: Recovery performances of DPS and RDP-DPS under increasing outlier contamination for inverse scattering and FFHQ-Deblur.

To gain robustness in BIPs, we therefore choose $\ell_{\boldsymbol{y}}$ as a weighted likelihood loss and refer to the resulting method as the *Robust Diffusion Posterior* (RDP). Following the additive likelihood form in Section 2.1, RDP guides the reverse diffusion using a robust weighted likelihood evaluated at the DPS denoised estimate $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$. RDP achieves robustness by transforming the DPS guidance term into a component-wise weighted objective. At diffusion time t, we define the loss function $\ell_{\boldsymbol{y}}(\boldsymbol{x}_t) = -\sum_{i=1}^{d_y} w(r_{t,i}) \log \tilde{p}(y_i|\hat{\boldsymbol{x}}_{0|t})$, where $r_{t,i} = y_i - \hat{y}_{t,i}$ and $\hat{y}_{t,i} := [F\hat{\boldsymbol{x}}_{0|t}]_i$ denote the fitting residuals and the predicted measurements, respectively. The loss function induces the robust diffusion posterior $p_t^{(\ell)}(\boldsymbol{x}_t|\boldsymbol{y}) \propto p(\boldsymbol{x}_t) \exp\{-\tau\, \ell_{\boldsymbol{y}}(\boldsymbol{x}_t)\}$ at time t. The term $w\colon \mathbb{R} \to \mathbb{R}$ is a component-wise adaptive weight function. This weighting mechanism acts as a robust filter that, given small residuals, $w_i \approx 1$, the loss function $\ell_{\boldsymbol{y}}$ recovers standard likelihood; conversely, large residuals are down-weighted to control outliers impact.

We now provide a sufficient condition for the design of such a weighting function w to guarantee the diffusion posterior robustness, formalized in the following result:

Theorem 3.2. *Consider the robust diffusion posterior with the score* $\nabla_{\boldsymbol{x}_t} \log p_t^{(\ell)}(\boldsymbol{x}_t|\boldsymbol{y}) = \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \ell_{\boldsymbol{y}}(\boldsymbol{x}_t)$. *Let the guidance loss be defined as* $\ell_{\boldsymbol{y}}(\boldsymbol{x}_t) = -\sum_{i=1}^{d_y} w(r_{t,i}) \log p(y_i|\hat{\boldsymbol{x}}_{0|t})$, *where* $r_{t,i} = y_i - [F(\hat{\boldsymbol{x}}_{0|t})]_i$ *denotes the residual and* $w(\cdot)$ *is a weighting function. Suppose assumptions D.1–D.2 hold, if* w *satisfies the following conditions for any* $r \in \mathbb{R}$*:*

$$|r\, w(r)| < \infty \quad \textit{and} \quad |r^2 w'(r)| < \infty, \tag{4}$$

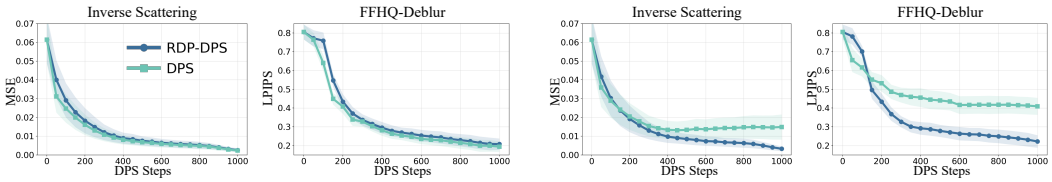
where w' *is the gradient of* w *w.r.t.* r*, the posterior KL divergence is uniformly bounded:*

$$\sup_{\boldsymbol{y} \in \mathbb{R}^{d_y}} D_{\mathrm{KL}}\left(p_\delta^{(\ell)}(\cdot|\boldsymbol{y}_*) \| p_\delta^{(\ell)}(\cdot|\boldsymbol{y})\right) \leq C,$$

where C *is a finite constant independent of the measurement* $\boldsymbol{y}$*. Hence, the posterior* $p_\delta^{(\ell)}$ *is outlier-robust, satisfying* $\sup_{\boldsymbol{y} \in \mathbb{R}^{d_y}} |PIF_{\boldsymbol{y}_*}(\boldsymbol{y})| < \infty$.

We provide the proof in Appendix D.3. This result stands in contrast to the standard DPS solver, whose PIF lacks such a guarantee. The proposed robust conditions control the influence of severe noise misspecification in $\boldsymbol{y}$, mitigating the risk of destabilizing the generation process.

Choice of w. The inverse multi-quadratic (IMQ) weighting function, defined as $w(r) = (1 + r^2/c^2)^{-1/2}$, satisfies the conditions established in Theorem 3.2. Consequently, the resulting diffusion posterior is outlier-robust. In this work, we adopt w in the IMQ form $w(r_i) = \left(1 + r_i^2/c^2\right)^{-1/2}$ for



(a) Well-specification (Gaussian) (b) Contamination of Outliers

Figure 3: Recovery performance of DPS and RDP-DPS (IMQ) across denoising steps under Gaussian noise and outlier contamination, averaged over 100 cases per task.

$i = 1, ..., d_y$, where $c > 0$ is a soft threshold: a larger value of c keeps weights closer to uniform, while a smaller value strongly down-weights large residuals. See Appendix E.2 for additional discussion on alternative weight functions.

Algorithm 1 presents RDP instantiated within the DPS sampler, which provides a realization aligned with the analysis in this paper. More broadly, RDP is a modular plug-in that can be incorporated into a broad class of plug-and-play diffusion-based approaches. We outline a general version in Appendix E.1 Algorithm 2 and provide associated experimental results in Section 4.

Performance given well-specification measurements. So far, we have focused on the robustness of RDP under outlier-contaminated measurements. A natural follow-up question is whether this robustness comes at the cost of degraded performance given well-specified measurements. In this setting, the standard Gaussian likelihood provides the desired guidance. The robust weighted likelihood introduces a bias relative to this Gaussian guidance, and the magnitude of this bias depends on the threshold parameter c. For the IMQ weighting function, $w(r) = (1 + r^2/c^2)^{-1/2}$. When the residual r is small relative to c, $w(r) \approx 1$ that roughly recovers the Gaussian guidance. This is analogous to the discussion in standard Bayesian problems [Knoblauch et al., 2022]. A larger c favors high fidelity for clean data (low bias), while a smaller c provides stronger suppression of large fitting residuals (high robustness).

Algorithm 1: Robust diffusion posterior sampling

Input: measurement $\boldsymbol{y}$, measurement model F, score network $\boldsymbol{s}_{\boldsymbol{\theta}}$, schedules $\{\beta_t, \sigma_t, \tau_t\}_{t=1}^T$, quantile q, scale floor ε_c, noise level σ_y
Initialize $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \ldots, 1$ **do**
  $\alpha_t \leftarrow 1 - \beta_t, \quad \bar{\alpha}_t \leftarrow \prod_{i=1}^t \alpha_i$
  $\hat{\boldsymbol{x}}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}}\big(\boldsymbol{x}_t + (1 - \bar{\alpha}_t)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\big)$
  $\boldsymbol{r}_t \leftarrow \boldsymbol{y} - F(\hat{\boldsymbol{x}}_{0|t})$
  $c_t \leftarrow \max\big\{\text{Quantile}_q(|\boldsymbol{r}_t|), \varepsilon_c\big\}$
  $w(r_{t,i}) \leftarrow \Big(1 + r_{t,i}^2/c_t^2\Big)^{-1/2}, i = 1, \ldots, d_y$
  $\ell_{\boldsymbol{y}}(\boldsymbol{x}_t) = -\sum_{i=1}^{d_y} w(r_{t,i}) \log p(y_i|\hat{\boldsymbol{x}}_{0|t})$
  $\hat{\boldsymbol{s}}_t \leftarrow \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \tau_t \nabla_{\boldsymbol{x}_t} \ell_{\boldsymbol{y}}(\boldsymbol{x}_t)$
  $\boldsymbol{z}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ if $t > 1$, else $\mathbf{0}$
  $\boldsymbol{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t + \beta_t \hat{\boldsymbol{s}}_t\right) + \sigma_t \boldsymbol{z}_t$
end for
Return: $\boldsymbol{x}_0$

Moreover, the denoising process introduces a time-dependent residual scale, which typically decreases as the denoised estimate becomes more accurate. This makes a fixed threshold c less appropriate. In practice, we adopt an adaptive schedule for the parameter c to account for these dynamic changes over time. We dynamically calibrate c_t at each timestep t based on the statistics of the current fitting residuals. Specifically, we set c_t based on the q-th quantile of the absolute residuals: $c_t = \text{Quantile}_q(|\boldsymbol{r}_t|)$, where $q \in (0, 1]$ and $\boldsymbol{r}_t = (r_{t,1}, \ldots, r_{t,d_y})$. By setting c at the q-th quantile (%), the posterior update is primarily driven by the q% of measurements with smaller residual magnitudes, while controlling the impact of the remaining $(1 - q)$% as presented in Figure 4. During gradient evaluation, we hold c_t fixed while retaining gradients through the residual-dependent weights. This adaptive guidance extends the fixed-scale construction of Theorem 3.2 as an empirical design beyond the scope of the current guarantee, with its effectiveness supported by the empirical experiments. With the adaptive design of c, RDP-DPS closely matches DPS under well-specified measurements and provides more stable recovery under outlier contamination. We further include the empirical validation and additional ablations in Section 4.2.

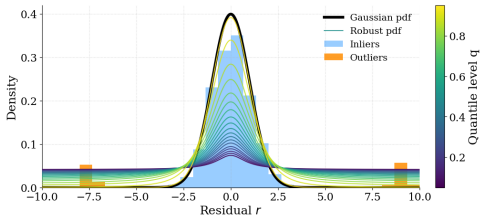


Figure 4: Histogram of residuals and outliers compared against Gaussian (black) and robust pdfs (colored) across quantiles q.

Table 1: Quantitative results on *Inverse Scattering* are reported as PSNR, SSIM, and NMAE (mean $\pm$ std) over 100 test samples. Results on *FFHQ-Inpaint*, *-Deblur*, and *-Phase Retrieval*, reported in LPIPS (↓), VIFL (↓) as mean $\pm$ std, and FID (↓) on 500 test images.

Task	Method	Well-specification (Gaussian)			Noise Misspecification (Student-T)			Contamination of Outliers		
		PSNR(↑)	SSIM(↑)	NMAE(↓)	PSNR(↑)	SSIM(↑)	NMAE(↓)	PSNR(↑)	SSIM(↑)	NMAE(↓)
Inverse Scattering	DPS	26.26 ± 2.83	0.86 ± 0.02	0.16 ± 0.02	21.53 ± 2.93	0.78 ± 0.06	0.25 ± 0.05	19.37 ± 3.44	0.71 ± 0.10	0.33 ± 0.05
	LGD	24.16 ± 2.95	0.76 ± 0.20	0.18 ± 0.09	18.80 ± 2.54	0.62 ± 0.07	0.34 ± 0.10	17.81 ± 2.89	0.53 ± 0.12	0.36 ± 0.05
	ΠGDM	24.23 ± 2.26	0.81 ± 0.13	0.18 ± 0.07	21.17 ± 3.15	0.69 ± 0.05	0.24 ± 0.05	15.14 ± 3.88	0.64 ± 0.15	0.37 ± 0.06
	PnPDM	25.77 ± 2.75	0.87 ± 0.04	$\mathbf{0.14 \pm 0.03}$	20.38 ± 4.14	0.73 ± 0.12	0.27 ± 0.21	15.33 ± 3.23	0.73 ± 0.12	0.30 ± 0.03
	DiffPIR	$\mathbf{27.36 \pm 2.79}$	$\mathbf{0.88 \pm 0.07}$	0.15 ± 0.02	22.91 ± 3.33	0.80 ± 0.10	0.23 ± 0.03	19.08 ± 3.92	0.71 ± 0.13	0.33 ± 0.08
	RDP-DPS	26.01 ± 2.72	0.86 ± 0.02	0.15 ± 0.01	$\mathbf{23.84 \pm 2.59}$	$\mathbf{0.82 \pm 0.03}$	$\mathbf{0.19 \pm 0.04}$	$\mathbf{25.24 \pm 2.74}$	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.15 \pm 0.02}$
	RDP-LGD	24.88 ± 2.17	0.81 ± 0.06	0.18 ± 0.03	21.83 ± 2.25	0.79 ± 0.05	0.23 ± 0.04	25.02 ± 2.23	0.70 ± 0.06	0.21 ± 0.03
	RDP-ΠGDM	23.67 ± 2.15	0.79 ± 0.07	0.17 ± 0.04	22.23 ± 2.18	0.72 ± 0.02	0.19 ± 0.03	22.49 ± 2.36	0.78 ± 0.05	0.21 ± 0.04
	RDP-PnPDM	23.94 ± 2.65	0.86 ± 0.04	0.16 ± 0.02	22.15 ± 3.11	0.79 ± 0.06	0.21 ± 0.02	23.45 ± 2.95	0.81 ± 0.07	0.16 ± 0.03
		LPIPS(↓)	VIFL(↓)	FID(↓)	LPIPS(↓)	VIFL(↓)	FID(↓)	LPIPS(↓)	VIFL(↓)	FID(↓)
Inpaint	DPS	$\mathbf{0.19 \pm 0.04}$	0.62 ± 0.04	27.19	0.25 ± 0.06	0.70 ± 0.04	30.21	0.37 ± 0.06	0.89 ± 0.04	45.27
	LGD	0.21 ± 0.03	0.68 ± 0.04	28.04	0.24 ± 0.04	0.75 ± 0.05	31.79	0.34 ± 0.06	0.82 ± 0.03	37.93
	ΠGDM	0.21 ± 0.05	0.63 ± 0.03	28.72	0.29 ± 0.03	0.77 ± 0.04	40.28	0.38 ± 0.06	0.84 ± 0.04	47.93
	PnPDM	$\mathbf{0.19 \pm 0.03}$	$\mathbf{0.61 \pm 0.04}$	**27.17**	0.25 ± 0.04	0.73 ± 0.04	31.29	0.36 ± 0.04	0.85 ± 0.05	47.28
	DiffPIR	0.24 ± 0.04	0.70 ± 0.06	29.64	0.26 ± 0.07	0.77 ± 0.05	35.83	0.37 ± 0.06	0.87 ± 0.04	46.80
	RDP-DPS	0.20 ± 0.04	0.62 ± 0.05	27.62	$\mathbf{0.21 \pm 0.04}$	$\mathbf{0.66 \pm 0.04}$	**28.03**	$\mathbf{0.24 \pm 0.04}$	$\mathbf{0.70 \pm 0.03}$	**29.95**
	RDP-LGD	0.22 ± 0.04	0.69 ± 0.04	28.60	0.23 ± 0.05	0.72 ± 0.03	29.52	0.25 ± 0.05	0.77 ± 0.04	30.06
	RDP-ΠGDM	0.20 ± 0.04	0.63 ± 0.03	29.15	0.24 ± 0.04	0.70 ± 0.04	28.19	0.25 ± 0.04	0.74 ± 0.05	30.27
	RDP-PnPDM	0.20 ± 0.04	0.65 ± 0.03	27.66	0.23 ± 0.04	0.71 ± 0.03	29.35	$\mathbf{0.24 \pm 0.04}$	0.73 ± 0.04	30.20
Deblur	DPS	0.17 ± 0.04	0.61 ± 0.05	26.98	0.20 ± 0.04	0.69 ± 0.04	29.35	0.32 ± 0.07	0.83 ± 0.07	39.82
	LGD	$\mathbf{0.13 \pm 0.03}$	$\mathbf{0.58 \pm 0.04}$	**22.36**	0.17 ± 0.04	0.67 ± 0.05	22.84	0.33 ± 0.09	0.79 ± 0.08	33.39
	ΠGDM	0.19 ± 0.08	0.62 ± 0.06	30.77	0.29 ± 0.11	0.71 ± 0.14	55.76	0.39 ± 0.15	0.90 ± 0.18	59.32
	PnPDM	0.14 ± 0.04	0.59 ± 0.04	24.11	0.21 ± 0.05	0.70 ± 0.04	35.18	0.35 ± 0.04	0.84 ± 0.04	42.54
	DiffPIR	0.16 ± 0.04	0.61 ± 0.05	25.79	0.22 ± 0.04	0.73 ± 0.06	30.86	0.38 ± 0.06	0.88 ± 0.07	54.78
	RDP-DPS	0.17 ± 0.04	0.62 ± 0.04	26.39	0.18 ± 0.04	0.65 ± 0.03	27.81	0.25 ± 0.06	0.75 ± 0.03	30.60
	RDP-LGD	0.14 ± 0.05	0.61 ± 0.05	22.84	$\mathbf{0.14 \pm 0.03}$	$\mathbf{0.63 \pm 0.04}$	**22.71**	$\mathbf{0.22 \pm 0.05}$	$\mathbf{0.70 \pm 0.04}$	**28.55**
	RDP-ΠGDM	0.20 ± 0.06	0.63 ± 0.05	30.86	0.21 ± 0.04	0.69 ± 0.05	33.94	0.31 ± 0.06	0.83 ± 0.07	38.84
	RDP-PnPDM	0.16 ± 0.04	0.61 ± 0.03	25.73	0.18 ± 0.04	0.65 ± 0.04	27.69	0.25 ± 0.05	0.77 ± 0.04	33.92
Phase Retrieval	DPS	0.45 ± 0.17	0.76 ± 0.03	62.16	0.52 ± 0.18	0.85 ± 0.03	76.01	0.64 ± 0.22	0.96 ± 0.02	118.59
	LGD	0.43 ± 0.19	0.79 ± 0.04	64.86	0.55 ± 0.21	0.89 ± 0.04	95.42	0.60 ± 0.17	0.95 ± 0.05	140.72
	PnPDM	$\mathbf{0.38 \pm 0.13}$	$\mathbf{0.72 \pm 0.04}$	**57.04**	0.49 ± 0.17	0.82 ± 0.03	74.38	0.60 ± 0.20	0.93 ± 0.04	113.82
	RDP-DPS	0.43 ± 0.13	0.76 ± 0.02	61.29	0.49 ± 0.12	0.81 ± 0.02	73.51	$\mathbf{0.51 \pm 0.14}$	$\mathbf{0.82 \pm 0.04}$	**75.07**
	RDP-LGD	0.43 ± 0.16	0.79 ± 0.04	63.37	0.51 ± 0.14	0.83 ± 0.05	76.28	0.54 ± 0.09	0.88 ± 0.03	84.27
	RDP-PnPDM	0.40 ± 0.12	0.74 ± 0.02	58.93	$\mathbf{0.47 \pm 0.13}$	$\mathbf{0.77 \pm 0.04}$	**63.93**	0.52 ± 0.16	$\mathbf{0.82 \pm 0.04}$	75.31

4 Experiment

In this section, we validate the effectiveness and robustness of the proposed RDP method through extensive experiments and ablation studies. We demonstrate RDP's generality by integrating it with four representative diffusion-based inverse solvers, including DPS [Chung et al., 2023], LGD [Song et al., 2023b], ΠGDM [Song et al., 2023a], and PnPDM [Wu et al., 2024][3]. We also include DiffPIR [Zhu et al., 2023] as another representative diffusion restoration approach. See Algorithm 1 and Algorithm 2 for the pseudo-code of our algorithms. For fair comparisons, all methods are evaluated using the same pretrained score/denoising models and sampling schedule. We refer to the RDP version of method X as RDP-X; for example, RDP-DPS denotes DPS with RDP guidance. See additional experimental results and more ablation studies in Appendix F.4.

4.1 Experiment Settings

Tasks. We conduct experiments on a diverse range of scientific inverse problems and natural image restorations using the open-source benchmark *InverseBench*[4] [Zheng et al., 2025]. The tasks include: (1) inverse scattering (IS), (2) image inpainting, (3) deblurring, and (4) phase retrieval on the FFHQ-256 dataset (FFHQ-D/PR), covering both linear and nonlinear cases. In practice, these problems often suffer from likelihood misspecification; for example, phase retrieval can be affected by unexpected scattering from inhomogeneous media and other unmodeled physical corruptions. Such effects can introduce heavy-tailed noise or outlier contamination that violate Gaussian assumptions, making these tasks well suited for evaluating multi-perspective robust reconstructions.

Experimental Design. A central focus of our evaluation is the performance regarding the mismatched noise distribution. We introduce two different schemes applied in the observation domain: (1) *Heavy-tailed noise*: i.i.d. noises from the Student-t distribution for evaluating the performance under the misspecification of noise; and (2) *Sparse impulsive outlier*: a fraction of components/pixels

[3]PnPDM does not rely on the Tweedie approximation and therefore lies outside the formal analysis of Sec. 3. However, it also employs gradient-based guidance; adding RDP yields improved performance.

[4]`https://github.com/devzhk/InverseBench`

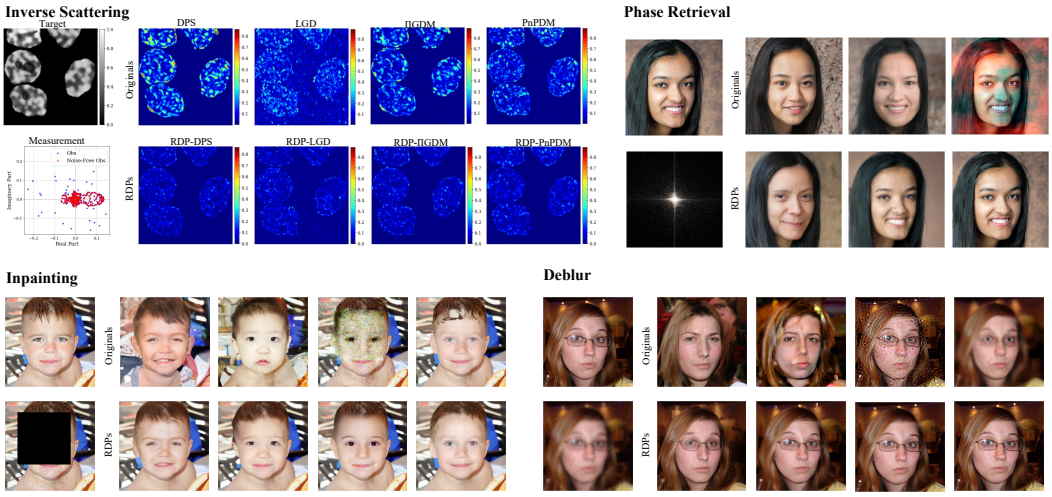


Figure 5: Qualitative results on all outlier-corrupted tasks: (a) Inverse Scattering, (b) Phase Retrieval, (c) Deblurring, and (d) Inpainting. In each panel, the leftmost column shows the ground truth and the measurements. The remaining columns compare the originals (top row) with their robust versions (bottom row); we report reconstruction absolute errors in (a) and reconstructed samples in (b)–(d).

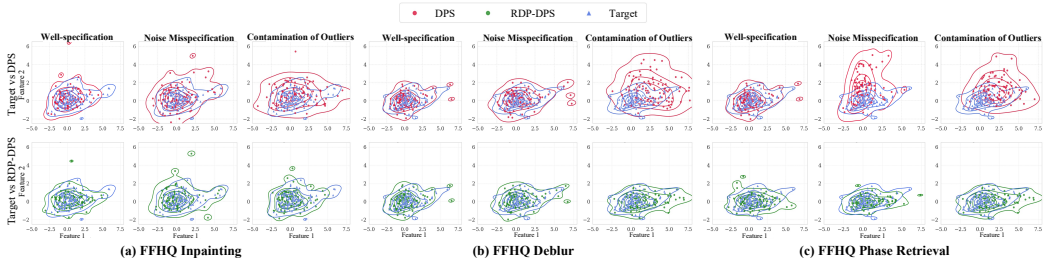


Figure 6: Posterior visualizations in Inception-v3 latent space for (a) Inpainting, (b) Deblur, and (c) Phase Retrieval under well-specified, misspecified, and outlier-contaminated settings. Top: DPS (●); bottom: RDP-DPS (●); target: ▲. The density contours are estimated by kernel density estimation.

corrupted by large-magnitude perturbations for evaluating the performance under outlier contamination. Additionally, to evaluate our approaches under normal measurements, we also report results with the conventional i.i.d. Gaussian noise distribution. Appendix F.2 includes detailed settings.

Evaluation Metric. For the IS problem, we quantify the reconstruction quality using peak signal-to-noise ratio (PSNR), structural similarity (SSIM), and the normalized mean absolute error (NMAE) metrics. In the two natural image tasks, we use standard perceptual quality metrics, including learned perceptual image patch similarity (LPIPS), visual information fidelity loss (VIFL), and Fréchet Inception Distance (FID) using the pre-trained InceptionV3 [Szegedy et al., 2016].

4.2 Result Analysis

In this section, we answer the following questions with key findings from our experiments.

(1) *When measurements match the presumed likelihood model, a diffusion-model prior lead to posterior stability and convergence.*

Quantitatively, Table 1 shows that, when the Gaussian likelihood is well specified, RDP performs competitive with standard methods such as DPS, LGD, ΠGDM, and PnPDM, indicating that the robust likelihood approximation preserves stable convergence. This is also supported by Figure 6 (left columns), where the sample and target density contours largely overlap in the first two latent dimensions. Figure 3 further shows that, with the adaptive choice of c_t, RDP-DPS closely tracks DPS over diffusion time in the well-specified setting, while achieving more robust recovery under outlier

contamination. Additional ablations in Section 4.2 also show that the method is not highly sensitive to the choice of q.

(2) *Diffusion samplers can fail under misspecification, and RDP effectively mitigate this failure.*

(DEGRADATION OF ORIGINAL APPROACHES.) In IS, the baselines in Table 1 degrade by about $10\% - 30\%$ due to the noise-distribution misspecification and $15\% - 50\%$ under outlier contamination across metrics, compared with the well-specified setting. Original approaches typically rely on a presumed Gaussian likelihood and can degrade substantially under likelihood misspecification given non-Gaussian noise and outlier contamination. Similar degradation phenomena are observed in image restoration tasks. Qualitatively in Figure 5, outliers induce visible artifacts, such as bright spots in the deblurring results of ΠGDM and PnPDM , indicating limited robustness.

(RDP IMPROVEMENT.) The integration of RDP substantially mitigates these issues and improves robustness to likelihood misspecification. In IS, RDP improves over baselines by roughly $5\% - 29\%$ under noise-distribution misspecification and $20\% - 50\%$ under outlier contamination across metrics. Similar gains are observed on FFHQ-D and FFHQ-PR in Table 1. Qualitatively, RDP removes bright-spot artifacts in ΠGDM and PnPDM variants, as shown in Figures 5 and 8, consistent with its adaptive down-weighting of extreme residuals. Evaluated in the Inception-v3 latent space (Figure 6, middle/right), RDP-DPS remains closer to the target distribution than DPS under misspecification.

RDP consistently improves performance in both scenarios, but its gains are larger under outlier contamination than under heavy-tailed noise misspecification. We attribute this difference to the fact that large outliers are easier to identify and down-weight, whereas heavy-tailed noise is less separable from normal measurement noise.

(3) *RDP are generalizable in (i) integration into different diffusion samplers and (ii) in validation across different tasks.*

This generality follows from our design: the robust gradient-based guidance term is decoupled from the underlying diffusion solver, making RDP a sampler-agnostic module that can be integrated into different gradient-based posterior samplers. Across image restoration and scientific tasks, Table 1 shows that RDP maintains competitive performance under well-specified Gaussian noise while improving the original samplers under model misspecification. The qualitative results in Figures 5 and 8 further support these findings across tasks.

4.3 Ablation Study

We revisit the IS task to ablate RDP under different conditions and design choices. Detailed settings and additional results are provided in Appendix F.4. We summarize the main findings below.

RDP is more beneficial when measurements are limited. We first examine the performances of RDP under different measurement dimensions d_y. As shown in Table 6, the advantage of RDP becomes more pronounced as the measurement dimension d_y decreases under both outlier and heavy-tailed noise. This suggests that robust guidance is particularly effective when measurements are limited, where each corrupted observation has a larger influence on the posterior update.

RDP is robust to weighting designs. We examine the sensitivity of RDP to the weighting function and the adaptive quantile q. As shown in Table 9, IMQ and Huber weighting achieve comparable performance under Gaussian and heavy-tailed noise, while IMQ gives slightly better results under outlier contamination. This suggests that the key factor is not the specific weighting form, but the adaptive down-weighting of large residuals. Since both IMQ and Huber use quantile-calibrated thresholds, they suppress a similar set of unreliable measurements. Table 10 shows that the performance is stable for $q \geq 0.25$, while an overly small quantile such as $q = 0.10$ degrades performance by aggressively down-weighting useful measurements. These results indicate that RDP is not highly sensitive to the precise weighting choice or quantile.

5 Conclusion

We studied the *stability* and *robustness* of Bayesian inverse problems with diffusion model priors, providing both theoretical insights and empirical evidence. Building on this analysis, we introduced *robust diffusion posterior sampling* (RDP), a unified plug-and-play approach that can be readily

incorporated into existing gradient-based diffusion posterior samplers. RDP preserves the stability of standard diffusion posterior sampling while providing robustness under outlier contamination. Extensive experiments across scientific inverse problems and natural image restoration tasks validate our theory and demonstrate the effectiveness and robustness of the proposed method.

Limitations. This work has two main limitations. First, our theoretical analysis mainly focuses on linear BIPs with a fixed threshold, and extending it to nonlinear settings and adaptive quantile thresholding remains an important direction. Second, our experiments primarily consider isotropic noise, while structured noise, such as correlated or spatially varying noise, is also common in practice.

Acknowledgements

Zhuo Sun is supported by Fundamental Research Funds for the Central Universities 2025110590 of Shanghai University of Finance and Economics. This work is supported by the Shanghai Engineering Research Center of Finance Intelligence (Grant No,19DZ2254600).

References

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and scalable bayesian online changepoint detection. In *International Conference on Machine Learning*, pages 642–663. PMLR, 2023.

Aleksandr Y Aravkin, James V Burke, and Gianluigi Pillonetto. Robust and trend-following student's t kalman smoothers. *SIAM Journal on Control and Optimization*, 52(5):2891–2916, 2014.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. In *International Conference on Machine Learning*, pages 1555–1584. PMLR, 2022a.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2022b.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and Omer Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *Transactions on Machine Learning Research*, 2024.

Gabriel Victorino Cardoso, Yazid Janati El Idrissi, Sylvain Le Corff, and Eric Moulines. Monte carlo guided diffusion for bayesian linear inverse problems. In *International Conference on Learning Representations*, 2024.

Haoxuan Chen, Yinuo Ren, Martin Renqiang Min, Lexing Ying, and Zachary Izzo. Solving inverse problems via diffusion-based priors: An approximation-free ensemble sampling approach. *arXiv preprint arXiv:2506.03979*, 2025.

Xiaoyuan Cheng, Xiaohang Tang, and Yiming Yang. Safe and stable control via lyapunov-guided diffusion models. *Advances in Neural Information Processing Systems*, 38:144946–144985, 2026a.

Xiaoyuan Cheng, Haoyu Wang, Wenxuan Yuan, Ziyan Wang, Zonghao Chen, Li Zeng, and Zhuo Sun. Fisher Decorator: Refining flow policy via a local transport map, 2026b. URL `https://arxiv.org/abs/2604.17919`.

Xiaoyuan Cheng, Wenxuan Yuan, Boyang Li, Yuanchao Xu, Yiming Yang, Hao Liang, Bei Peng, Robert Loftin, Zhuo Sun, and Yukun Hu. How does the lagrangian guide safe reinforcement learning through diffusion models? In *Forty-third International Conference on Machine Learning*, 2026c. URL `https://openreview.net/forum?id=mK3xWE7p2T`.

Xiaoyuan Cheng, Wenxuan Yuan, Zhancun Mu, Yuanzhao Zhang, Yiming Yang, Hai Wang, Zhuo Sun, and Che Liu. Scaling world-model reinforcement learning through diffusion policy optimization. *arXiv preprint arXiv:2605.26282*, 2026d.

Hyungjin Chung, Jeongsol Kim, Michael T McCann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 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.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Ernesto De Vito, Lorenzo Rosasco, Andrea Caponnetto, Umberto De Giovannini, Francesca Odone, and Peter Bartlett. Learning from examples as an inverse problem. *Journal of Machine Learning Research*, 6(5), 2005.

Zehao Dou and Yang Song. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Randal Douc and Eric Moulines. Limit theorems for weighted samples with applications to sequential monte carlo methods. *The Annals of Statistics*, pages 2344–2376, 2008.

Arnaud Doucet, Nando De Freitas, Neil James Gordon, et al. *Sequential Monte Carlo methods in practice*, volume 1. Springer, 2001.

Gerardo Duran-Martin, Matias Altamirano, Alexander Y Shestopaloff, Leandro Sánchez-Betancourt, Jeremias Knoblauch, Matt Jones, François-Xavier Briol, and Kevin Murphy. Outlier-robust kalman filtering through generalised bayes. In *International Conference on Machine Learning*, pages 12138–12171, 2024.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *Advances in Neural Information Processing Systems*, 35:14715–14728, 2022.

Shishuo Guo, Xiaoyuan Cheng, and Zhuo Sun. Random-projection tree stein variational gradient descent. In *Advances in Neural Information Processing Systems*, volume 39, 2026.

Jiajun He, José Miguel Hernández-Lobato, Yuanqi Du, and Francisco Vargas. Rne: a plug-and-play framework for diffusion density estimation and inference-time control. *arXiv preprint arXiv:2506.05668*, 2025.

Yi He, Zimo Zhao, Yiming Yang, Xiaoyuan Cheng, Chao He, and Yukun Hu. MMPD-bench: Bridging multimodal fission with multi-polarimetric modalities decomposition. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=EGvoEpzkWG`.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Bamdad Hosseini and Nilima Nigam. Well-posed bayesian inverse problems: Priors with exponential tails. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):436–465, 2017.

Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

Yazid Janati, Badr Moufad, Alain Durmus, Eric Moulines, and Jimmy Olsson. Divide-and-conquer posterior sampling for denoising diffusion priors. *Advances in Neural Information Processing Systems*, 37:97408–97444, 2024.

Yazid Janati, Eric Moulines, Jimmy Olsson, and Alain Oliviero-Durmus. Bridging diffusion posterior sampling and monte carlo methods: a survey. *Philosophical Transactions A*, 383(2299):20240331, 2025.

Tijana Janjić, Niels Bormann, Marc Bocquet, JA Carton, Stephen E Cohn, Sarah L Dance, Svetlana N Losa, Nancy K Nichols, Roland Potthast, Joanne A Waller, et al. On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144(713):1257–1278, 2018.

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.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Bartek T Knapik, Aad W Van Der Vaart, and J Harry van Zanten. Bayesian inverse problems with gaussian priors. *The Annals of Statistics*, pages 2626–2657, 2011.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23 (132):1–109, 2022.

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

Ronan Legin, Alexandre Adam, Yashar Hezaveh, and Laurence Perreault-Levasseur. Beyond gaussian noise: A generalized approach to likelihood analysis with non-gaussian noise. *The Astrophysical Journal Letters*, 949(2):L41, 2023.

Pablo Lemos, Adam Coogan, Yashar Hezaveh, and Laurence Perreault-Levasseur. Sampling-based accuracy testing of posterior estimators for general inference. In *International Conference on Machine Learning*, pages 19256–19273. PMLR, 2023.

Kaiyu Li, Yiming Yang, Xiaoyuan Cheng, Yi He, and Zhuo Sun. Multilevel control functional. In *International Conference on Learning Representations*, volume 2026, pages 146510–146535, 2026.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper_files/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf`.

Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International conference on machine learning*, pages 14429–14460. PMLR, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J Oates. Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022, 2022.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34: 25359–25369, 2021.

Koichi Miyasawa et al. An empirical bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist*, 38(181-188):1–2, 1961.

Badr Moufad, Yazid Janati, Lisa Bedin, Alain Oliviero Durmus, randal douc, Eric Moulines, and Jimmy Olsson. Variational diffusion posterior sampling with midpoint guidance. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6EUtjXAvmj.

Zijing Ou, Mingtian Zhang, Andi Zhang, Tim Z Xiao, Yingzhen Li, and David Barber. Improving probabilistic diffusion models with optimal diagonal covariance matching. In *International Conference on Learning Representations*, 2025.

Marien Renaud, Jiaming Liu, Valentin De Bortoli, Andres Almansa, and Ulugbek Kamilov. Plug-and-play posterior sampling under mismatched measurement and prior models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=66arKkGiFy.

Marien Renaud, Valentin De Bortoli, Arthur Leclaire, and Nicolas Papadakis. From stability of langevin diffusion to convergence of proximal MCMC for non-log-concave sampling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=mcYO3DccbQ.

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.

M. Berk Sahin, Behzad Sharif, and Abolfazl Hashemi. Provable derivative-free inference with score-based generative priors, 2026. URL https://openreview.net/forum?id=EKLhCNjzQR.

Malcolm Sambridge and Klaus Mosegaard. Monte carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3):3–1, 2002.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

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

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pages 32483–32498. PMLR, 2023b.

Lingfei Song and Hua Huang. Robust image restoration with an adaptive huber function based fidelity. *International Journal of Computer Vision*, 132(12):6127–6141, 2024.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

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, 2021b.

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*, 2021c.

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*, 2021d.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta numerica*, 19:451–559, 2010.

Timothy John Sullivan. *Introduction to uncertainty quantification*, volume 63. Springer, 2015.

Yu Sun, Zihui Wu, Yifan Chen, Berthy T Feng, and Katherine L Bouman. Provable probabilistic imaging using score-based generative priors. *IEEE Transactions on Computational Imaging*, 10: 1290–1305, 2024.

Zhuo Sun. *Transfer learning in Monte Carlo Methods and Beyond*. PhD thesis, UCL (University College London), 2023.

Zhuo Sun, Alessandro Barp, and François-Xavier Briol. Vector-valued control variates. In *International Conference on Machine Learning*, pages 32819–32846. PMLR, 2023a.

Zhuo Sun, Chris J Oates, and François-Xavier Briol. Meta-learning control variates: Variance reduction with limited data. In *Uncertainty in Artificial Intelligence*, pages 2047–2057. PMLR, 2023b.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Dilin Wang, Ziyang Tang, Chandrajit Bajaj, and Qiang Liu. Stein variational gradient descent with matrix-valued kernels. *Advances in neural information processing systems*, 32, 2019.

Zheng Wang, Johnathan M Bardsley, Antti Solonen, Tiangang Cui, and Youssef M Marzouk. Bayesian inverse problems with l_1 priors: a randomize-then-optimize approach. *SIAM Journal on Scientific Computing*, 39(5):S140–S166, 2017.

Veit David Wild, Sahra Ghalebikesabi, Dino Sejdinovic, and Jeremias Knoblauch. A rigorous link between deep ensembles and (variational) bayesian methods. *Advances in Neural Information Processing Systems*, 36:39782–39811, 2023.

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.

Yiming Yang, Xiaoyuan Cheng, Daniel Giles, Sibo Cheng, Yi He, Xiao Xue, Boli Chen, and Yukun Hu. Tensor-Var: Efficient four-dimensional variational data assimilation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 70649–70679. PMLR, 2025. URL `https://proceedings.mlr.press/v267/yang25h.html`.

Yiming Yang, Valentin Brekke, James Briant, and Serge Guillas. Diffusion distillation for efficient weather ensembles. *arXiv preprint arXiv:2608.27728*, 2026.

Mingtian Zhang, Alex Hawkins-Hooker, Brooks Paige, and David Barber. Moment matching denoising gibbs sampling. *Advances in Neural Information Processing Systems*, 36:23590–23606, 2023.

Hongkai Zheng, Wenda Chu, Bingliang Zhang, Zihui Wu, Austin Wang, Berthy Feng, Caifeng Zou, Yu Sun, Nikola Borislavov Kovachki, Zachary E Ross, Katherine Bouman, and Yisong Yue. Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. In *International Conference on Learning Representations*, 2025.

Yang Zheng, Jiahua Liu, Tongyao Pang, Wen Li, and Zhaoqiang Liu. Outlier-robust diffusion solvers for inverse problems. In *Conference on Computer Vision and Pattern Recognition 2026*, 2026. URL `https://openreview.net/forum?id=IuH2pXyLT2`.

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*, pages 1219–1229, 2023.

A Notation

Table 2: Summary of Key Notations.

Notation	Meaning
	Inverse Problem
$\boldsymbol{x}$	target variable
$\boldsymbol{y}_*$	noise-free measurement
$\boldsymbol{y}$	observed measurement
$F(\cdot)$	Measurement model
$\boldsymbol{\epsilon}$	Measurement noise ($\boldsymbol{y} = F(\boldsymbol{x}) + \boldsymbol{\epsilon}$)
	Diffusion Model
p_{data}	True data distribution
$\boldsymbol{x}_t$	Noised target variable at time $t \in [0, T]$
$\boldsymbol{x}_0$	Clean data samples
$\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$	Learned score function estimator
$\boldsymbol{x}_0(\boldsymbol{x}_t)$	Ideal denoised estimate using the true score $\nabla \log p_t$; $\boldsymbol{x}_{0\|t} = \mathbb{E}[\boldsymbol{x}_0 \| \boldsymbol{x}_t]$
$\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$	Predicted denoised estimate using the score network $\boldsymbol{s}_{\boldsymbol{\theta}}$; approximation of $\boldsymbol{x}_{0\|t}$
$\beta_t, \bar{\alpha}_t$	Noise schedule and scaling factor
$p(\boldsymbol{x}_t \| \boldsymbol{y})$	intermediate posterior distribution at time t
	Generalized Bayesian Inference
$\ell_{\boldsymbol{y}}(\boldsymbol{x}_t)$	Robust likelihood function
τ_t	temperature (learning rate) at time t
$p^{(\ell)}(\boldsymbol{x}_t \| \boldsymbol{y})$	Generalized posterior distribution at time t
$\hat{\boldsymbol{y}}_t$	Predicted observation; $\hat{\boldsymbol{y}}_t = F(\hat{\boldsymbol{x}}_{0\|t})$
$\boldsymbol{r}_t$	Residual vector; $\boldsymbol{r}_t = \boldsymbol{y} - \hat{\boldsymbol{y}}_t$
ω_t	Learning rate/temperature at time t
$W(\boldsymbol{r}_t)$	Robust weighting matrix
PIF	Posterior Influence Function
	General Notation
$\mathbb{E}[\cdot]$	Expectation
$\mathbb{V}[\cdot]$	Variance (scalar) or Covariance matrix (vector)
$\nabla_{\boldsymbol{z}}$	Gradient with respect to $\boldsymbol{z}$
$\boldsymbol{I}$	Identity matrix
$\|\cdot\|_p$	L_p norm
$J_f(\cdot)$	Jacobian of function f
$H_f(\cdot)$	Hessian of function f

B Review of Tweedie approximation in posterior sampling.

Conditional Score Estimation: The *Plug-and-Play* Approach. The most direct application of Tweedie's approximation in posterior sampling is found in methods that attempt to estimate the conditional score $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y}|\boldsymbol{x}_t)$ and plug it into the reverse SDE. This plug-and-play (PnP) family, commonly referred to as Diffusion Posterior Sampling (DPS) [Graikos et al., 2022, Chung et al., 2023], replaces an intractable conditional expectation with a Tweedie-based point estimator. We focus our analysis on this class. Despite its empirical effectiveness, the approximation is heuristic and can induce systematic bias [Chung et al., 2023]. To mitigate this approximation error, other methods utilize the structure of (linear) measurement model to derive more principled guidance signals. For instance, DDRM [Kawar et al., 2022] approaches linear inverse problems by performing diffusion in the spectral domain of the linear measurement model via SVD, while the ΠGDM [Song et al., 2023a] employs the Moore-Penrose pseudoinverse to project the guidance signal directly onto the solution manifold consistent with the measurements. Moving beyond first-order point estimates, recent works incorporate second-order corrections. Rozet et al. [2024] proposes matching the variance of $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$ by approximating it as a Gaussian rather than a Dirac delta. Similarly, Tweedie Moment Projected Diffusions (TMPD) [Boys et al., 2024] use higher-order Tweedie expansions to estimate the posterior covariance. Though statistically optimal for near-Gaussian distributions, this approach improves stability at the cost of computing the explicit Jacobian of the Tweedie estimator.

Sequential Monte Carlo: Unbiased Estimator. While the conditional score approach introduces systematic bias into the generation process due to Tweedie's approximation, an alternative line of

research utilizes Sequential Monte Carlo (SMC) to correct these errors [Wu et al., 2023, Janati et al., 2025]. SMC methods, such as the Twisted Diffusion Sampler (TDS) [Wu et al., 2023] and Monte Carlo Guided Diffusion (MCGDiff) [Cardoso et al., 2024], provide theoretically well-founded algorithms for posterior sampling. Unlike the DPS, SMC offers strong guaranties: the weighted particle approximation is asymptotically exact as the number of particles goes to infinity [Doucet et al., 2001, Douc and Moulines, 2008]. Recent work on the Radon-Nikodym Estimator (RNE) [He et al., 2025] has further unified these approaches. RNE demonstrates that the heuristic weight functions used in methods like TDS actually correspond to the theoretically optimal weights derived from the ratio of path measures between the forward and backward processes. This theoretical unification confirms that Tweedie's approximation acts as the crucial link that makes these rigorous likelihood-ratio weights computable in practice. For linear inverse problems, Song et al. [2021c] introduces an auxiliary noise measurements sequence $\{\boldsymbol{y}_t\}_t$ that is coupled to the latent diffusion trajectory $\{\boldsymbol{x}_t\}_t$ through the same noise schedule. This coupling renders the intermediate likelihood $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ linear Gaussian, yielding closed-form importance weights. This ensures that standard SMC updates remain tractable. Dou and Song [2024] further formalizes this approach within the Filtering for Posterior Sampling (FPS) framework. The resulting FPS-SMC algorithm provides a global consistency guaranty: as the number of particles approaches infinity, the approximation converges to the true Bayesian posterior. While SMC provides asymptotic guaranties, it suffers from the curse of dimensionality and requires a sufficient number of samples.

Diffusion Priors with MCMC Posterior Sampling. Another related line of work formulates posterior sampling with diffusion or score-based priors as Markov chain Monte Carlo (MCMC) targeting an explicit posterior distribution or a controlled approximation. Within this class, the PnP Unadjusted Langevin Algorithm (PnP-ULA)[Laumont et al., 2022] and PnP Monte Carlo (PMC) [Sun et al., 2024] follow a closely related principle: both construct training-free MCMC samplers by combining an explicit measurement likelihood with an implicit denoising or score-based prior. PnP-ULA uses Tweedie's formula to interpret a pretrained score network as an approximate prior score and applies unadjusted Langevin dynamics for Bayesian imaging. PMC and its annealed variant extend this idea more directly to score-based generative priors, providing non-asymptotic stationarity guaranties. A recent extension of PMC, zeroth-order annealed PnP Monte Carlo (ZO-APMC) [Sahin et al., 2026], develops a derivative-free annealed PMC framework that only requires forward-model evaluations and a pretrained score-based prior, making MCMC posterior sampling applicable when likelihood gradients or pseudo-inverses are unavailable. Subsequent work studies the sensitivity of this sampler to mismatched measurement and prior models, providing a posterior discrepancy metric to quantify the deviation of the resulting sampling distribution from the target posterior [Renaud et al., 2024]. Recent proximal MCMC theory further extends these guaranties to broader settings [Renaud et al., 2026]. In a different direction, the PnP Split Gibbs Sampler (PnP-SGS) alternates between a data-consistency update and a denoising-prior update, providing a practical Gibbs-style framework for Bayesian imaging with implicit deep generative priors [Coeurdoux et al., 2024]. Stein variational gradient descent (SVGD) offers a complementary approach to posterior approximation through deterministic particle updates driven by the target score and kernel-based repulsion [Liu and Wang, 2016, Wang et al., 2019, Guo et al., 2026]. These score-based formulations also connect naturally to Stein control variates, which exploit zero-mean identities involving the target score to reduce the variance of Monte Carlo estimators of posterior expectations [Sun et al., 2023a,b, Sun, 2023, Li et al., 2026]. Likelihood modeling provides a complementary direction for addressing misspecification [Legin et al., 2023]. Beyond reconstruction quality and robustness, assessing posterior accuracy is also important [Lemos et al., 2023]. These perspectives motivate future work on broader noise models and the statistical reliability of robust posterior sampling. Related advances of diffusion posterior sampling in reinforcement learning [Cheng et al., 2026a,c,b,d], weather forecasting [Yang et al., 2025, 2026], and computational optics [He et al., 2026] suggest potential opportunities for extending robust diffusion posterior sampling to broader scientific and decision-making applications.

C Preliminaries of Tweedie's Formula in Diffusion Models

Proposition C.1 (First-order Tweedie Approximation [Chung et al., 2023])**.** *For the case of VP-SDE sampling, $p(\boldsymbol{x}_0|\boldsymbol{x}_t)$ has the analytical posterior expectation, denoted as $\boldsymbol{x}_{0|t}(\boldsymbol{x}_t) \coloneqq \mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$, is given by [Chung et al., 2023]:*

$$\boldsymbol{x}_0(\boldsymbol{x}_t) = \frac{1}{\sqrt{\alpha(t)}}\left(\boldsymbol{x}_t + (1-\alpha(t))\nabla \boldsymbol{x}_t \log p_t(\boldsymbol{x}_t)\right), \tag{5}$$

where $\alpha(t) = \exp(-\int_0^t \beta(s)\mathrm{d}s)$ is the accumulated noise variance.

By replacing the $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ with the score network $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$, the posterior expectation can be approximated as $\hat{\boldsymbol{x}}_0 \approx \frac{1}{\sqrt{\alpha(t)}}\big(\boldsymbol{x}_t + (1-\alpha(t))\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\big)$.

Proposition C.2 (Second-order Tweedie Approximation [Meng et al., 2021, Boys et al., 2024])**.** *For the case of VP-SDE, the conditional posterior $p(\boldsymbol{x}_0 \mid \boldsymbol{x}_t)$ has an analytical posterior covariance given by:*

$$\mathbb{V}(\boldsymbol{x}_0 \mid \boldsymbol{x}_t) = \frac{1-\alpha(t)}{\alpha(t)}\left(\boldsymbol{I} + (1-\alpha(t))\,\nabla^2_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right).$$

This follows from applying Tweedie's formula to second-order moments. It introduces a curvature term to the first-order estimate, the Hessian of $\log p(\boldsymbol{x}_t)$—to capture posterior uncertainty. While the first-order formula gives the posterior mean $\mathbb{E}(\boldsymbol{x}_0|\boldsymbol{x}_t)$, the second-order formula gives the corresponding covariance $\mathbb{V}(\boldsymbol{x}_0|\boldsymbol{x}_t)$. Direct evaluation of $\nabla^2_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$ is often impractical because it requires second-order score information. In practice, this Hessian is approximated using the score network, either through automatic differentiation [Meng et al., 2021, Zhang et al., 2023, Boys et al., 2024] or by training an auxiliary network to estimate the Hessian term [Bao et al., 2022a,b, Ou et al., 2025].

Remark C.3. There is a straightforward algebraic relation between Proposition C.1 and C.2 such that

$$\mathbb{V}(\boldsymbol{x}_0|\boldsymbol{x}_t) = \frac{1-\alpha(t)}{\sqrt{\alpha(t)}} J_{\boldsymbol{x}_0}(\boldsymbol{x}_t), \tag{6}$$

where $J_{\boldsymbol{x}_0}$ denotes the Jacobian of the posterior mean map. Note that this identity is not valid for an arbitrary Jacobian: the right-hand side need not be symmetric positive semidefinite and thus may fail to define a covariance matrix. It holds when J_T is the Jacobian of the *true* posterior-mean map induced by the associated density $p(\boldsymbol{x}_t)$.

Proposition C.4 (Conditional Tweedie's Formula)**.** *Assuming a transition kernel $p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha(t)}\boldsymbol{x}_0, (1-\alpha(t))\boldsymbol{I})$ from VP-SDE, the score of the posterior marginal distribution at time t is exactly related to the conditional posterior expectation of $\boldsymbol{x}_0$ by:*

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{y}) = -\frac{\boldsymbol{x}_t - \sqrt{\alpha(t)}\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t, \boldsymbol{y}]}{1-\alpha(t)}. \tag{7}$$

D Proof of main results

Notation and setup.

- **Diffusion process**. As stated in Section 2.2, we consider a VP-SDE with a noise schedule $\beta(t)$ satisfying $0 \leq \beta_{\min} \leq \beta(t) \leq \beta_{\max} < \infty$. Defining $\alpha(t) \coloneqq \exp\left(-\int_0^t \beta(s)\mathrm{d}s\right)$, the transition kernel is $p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha(t)}\boldsymbol{x}_0, (1-\alpha(t))\boldsymbol{I})$. Note that $\alpha(t) \in [e^{-T\beta_{\max}}, 1]$.
- **Endpoint Truncation**. In practice, the reverse process is evaluated from T up to a small positive terminal time $\delta > 0$. We adopt the same truncated-time setting in our theoretical analysis so that all score-based and denoising quantities are considered on $t \in [\delta, T]$.
- **Norms**. $\|\cdot\|$: Unless otherwise stated, $\|\boldsymbol{x}\|$ denotes the Euclidean norm for vectors, and $\|F\| \coloneqq \sup_{\|\boldsymbol{x}\|=1} \|F\boldsymbol{x}\|$ denotes the operator norm for matrices or operators.
- **Denoised estimates**. In the following, we let $\boldsymbol{x}_0(\boldsymbol{x}_t)$ denote the ideal denoised estimate computed using the true score function. We reserve $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ for the approximate estimate derived from the learned score network $\boldsymbol{s}_{\boldsymbol{\theta}}$.

D.1 Assumptions

We introduce the following assumptions, commonly adopted in the theoretical analysis of DMs and BIPs, which will be used throughout the paper:

Assumption D.1 (Score Network Regularity)**.** The score network $\boldsymbol{s}_{\boldsymbol{\theta}}(\cdot, t)$ is continuously differentiable with respect to $\boldsymbol{x}$ for time $t \in [\delta, T]$, and there exists a constant $L_s > 0$ such that $\|\nabla_{\boldsymbol{x}} \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\| \leq L_s$.

Assumption D.2 (Score Network Accuracy)**.** The expected error of the score network is bounded for $t \in [\delta, T]$: $\mathbb{E}\big[\|\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) - \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}, t)\|_2^2\big] \leq \epsilon_s(t)^2$, where $\int_\delta^T \epsilon_s(t)^2 < \infty$.

D.2 Proof of Lemma 3.1

Lemma D.3. *Suppose the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$ is a Gaussian distribution such that $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; F\boldsymbol{x}, \sigma_y^2\boldsymbol{I})$. The likelihood score $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x})$ is uniformly Lipschitz continuous with respect to the measurement $\boldsymbol{y}$. Specifically, for any $\boldsymbol{x} \in \mathbb{R}^{d_x}$, we have:*

$$\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}'|\boldsymbol{x})\| \leq L_{lld}\|\boldsymbol{y} - \boldsymbol{y}'\|, \quad \forall \boldsymbol{y}, \boldsymbol{y}' \in \mathbb{R}^{d_y},$$

where $L_{lld} = \|F\|/\sigma_y^2$.

Proof. The score of the Gaussian measurement model is given as $\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}) = \nabla F(\boldsymbol{x})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - F(\boldsymbol{x}))$, where $\nabla F(\boldsymbol{x})$ is the Jacobian of the measurement model. We have

$$\begin{aligned}\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}|\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p(\boldsymbol{y}'|\boldsymbol{x})\| &= \frac{1}{\sigma_y^2}\|F^\top(\boldsymbol{y} - \cancel{F\boldsymbol{x}}) - F^\top(\boldsymbol{y}' - \cancel{F\boldsymbol{x}})\| \\ &= \frac{1}{\sigma_y^2}\|F^\top(\boldsymbol{y} - \boldsymbol{y}')\| \leq \underbrace{\frac{1}{\sigma_y^2}\|F\|}_{L_{\text{lld}}} \cdot \|\boldsymbol{y} - \boldsymbol{y}'\|.\end{aligned}$$

□

Theorem D.4 (Girsanov Theorem)**.** *For any pair of well-defined diffusion processes $\{\boldsymbol{x}_t\}_{t=\delta}^T$ and $\{\boldsymbol{x}'_t\}_{t=\delta}^T$ on $\mathbb{R}^{d_x}$ defined as follows:*

$$\begin{aligned}\mathrm{d}\boldsymbol{x} &= f(\boldsymbol{x}_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t \\ \mathrm{d}\boldsymbol{x}' &= f'(\boldsymbol{x}'_t, t)\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}'_t\end{aligned}$$

where $f, f' \colon \mathbb{R}^{d_x} \times [\delta, T] \to \mathbb{R}^{d_x}$ are the two drift functions and $g \colon [\delta, T] \to \mathbb{R}_{>0}$is the diffusion function. Assume that: (1) $g(t) > 0$ for all $t \in [\delta, T]$, (2) $p_T \ll p'_T$, and (3) Novikov's condition

holds $\mathbb{E}\left[\exp\left(\frac{1}{2}\int_\delta^T \left\|\frac{f(\boldsymbol{x}_t,t)-f'(\boldsymbol{x}_t,t)}{g(t)}\right\|^2 \mathrm{d}t\right)\right] < \infty$. *Let* p_t *and* p'_t *denote the distributions of* $\boldsymbol{x}_t$ *and* $\boldsymbol{x}'_t$ *for any* $t \in [\delta, T]$*; then we have*

$$D_{KL}(p_\delta \| p'_\delta) \leq D_{KL}(p_T \| p'_T) + \int_\delta^T \frac{1}{2g(t)^2}\mathbb{E}\left[\|f(\boldsymbol{x}_t,t) - f'(\boldsymbol{x}_t,t)\|^2\right] \mathrm{d}t. \tag{8}$$

Proof. The bound is a direct consequence of the Girsanov theorem, which relates the probability measures of diffusion processes under a change of drift. The result has been well established as standard in the analysis of generative diffusion models and is used to upper-bound the KL divergence between the true data distribution and the generated distribution by the time-accumulated mismatch between their respective score functions. For detailed derivations, we refer the reader to prior work, such as Theorem 1 in [Song et al., 2021b], Lemma 2.22 in [Wu et al., 2024], and Lemma C.1 in [Chen et al., 2025]. A version adapted for the probability flow ODE formulation can be found in Theorem 3.1 of [Lu et al., 2022].

In the VP-SDE, the reverse-time drift functions corresponding to the posterior $p_t(\boldsymbol{x}_t|\boldsymbol{y})$ given the measurement $\boldsymbol{y}$ are defined as:

$$f(\boldsymbol{x}_t,t) = -\beta(t)\left[\frac{\boldsymbol{x}_t}{2} + \nabla \log p_t(\boldsymbol{x_t}|\boldsymbol{y})\right].$$

Substituting the diffusion coefficient $g(t) = \sqrt{\beta(t)}$, the squared difference appearing in the Girsanov Theorem (Equation (8)) simplifies to:

$$\frac{1}{2g(t)^2}\mathbb{E}\left[\|f(\boldsymbol{x}_t,t) - f'(\boldsymbol{x}_t,t)\|^2\right] = \frac{\beta(t)}{2}\mathbb{E}\left[\|\nabla \log p_t(\boldsymbol{x}_t|\boldsymbol{y}) - \nabla \log p'_t(\boldsymbol{x}_t|\boldsymbol{y})\|^2\right].$$

Lemma D.5. *Assume the forward process (VP) is defined by the SDE* $\mathrm{d}\boldsymbol{x}_t = -\frac{1}{2}\beta(t)\boldsymbol{x}_t\,\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\boldsymbol{w}_t$ *as in Section 2.2. The Gaussian kernel is given by* $\boldsymbol{x}_t|\boldsymbol{x}_0 \sim \mathcal{N}(\sqrt{\alpha(t)}\boldsymbol{x}_0, (1-\alpha(t))\,\boldsymbol{I})$*, where* $\alpha(t) = \exp\left(-\int_0^t \beta(s)\,\mathrm{d}s\right)$*. Let* $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ *be the Tweedie estimator approximating the posterior mean* $\mathbb{E}[\boldsymbol{x}_0|\boldsymbol{x}_t]$ *with the score network* $\boldsymbol{s_\theta}$*. Under Assumption D.1, the Jacobian of this estimator is bounded:* $\|J_{\hat{\boldsymbol{x}}_0}(\boldsymbol{x}_t)\| \leq \exp(T\beta_{\max}/2)(1+L_s) \coloneqq L_{\hat{\boldsymbol{x}}_0}$ *for any time* $t \in [\delta, T]$.

Proof. The proof relies on Tweedie's formula, which relates the score of the marginal distribution to the posterior expectation. For the Gaussian perturbation kernel defined above and Proposition C.1, the Tweedie estimator is given as $\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t) = \frac{1}{\sqrt{\alpha(t)}}\left(\boldsymbol{x}_t + ((1-\alpha(t))\,\boldsymbol{s_\theta}(\boldsymbol{x}_t,t)\right)$. We compute the Jacobian $J_{\hat{\boldsymbol{x}}_0}(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t}\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t)$ by differentiating with respect to $\boldsymbol{x}_t$; then, for any time $t \in [\delta, T]$,

$$\begin{aligned}
\|J_{\hat{\boldsymbol{x}}_0}(\boldsymbol{x}_t)\| &= \|\frac{1}{\sqrt{\alpha(t)}}\Big(\nabla_{\boldsymbol{x}_t}\boldsymbol{x}_t + (1-\alpha(t))\,\nabla_{\boldsymbol{x}_t}\boldsymbol{s_\theta}(\boldsymbol{x}_t,t)\Big)\| \\
&= \frac{1}{\sqrt{\alpha(t)}}\|I + (1-\alpha(t))\,\nabla_{\boldsymbol{x}_t}\boldsymbol{s_\theta}(\boldsymbol{x}_t,t)\| \\
&\leq \frac{1}{\sqrt{\alpha(t)}}\Big(\|I\| + (1-\alpha(t))\,\|\nabla_{\boldsymbol{x}_t}\boldsymbol{s_\theta}(\boldsymbol{x}_t,t)\|\Big) \\
&\leq \frac{1}{\sqrt{\alpha(t)}}\big(1 + (1-\alpha(t))\,L_s\big) \\
&\leq \alpha(T)^{-1/2}(1+L_s) = \exp(T\beta_{\max}/2)(1+L_s)
\end{aligned}$$

Moreover, $\int_\delta^T L_{\hat{\boldsymbol{x}}_0}\mathrm{d}t \leq (T-\delta)\exp(T\beta_{\max}/2)(1+L_s) < \infty$. □

***Proof* of Lemma 3.1.** Given the measurement $\boldsymbol{y} \in \mathbb{R}^{d_y}$, the difference between the two drift terms in the reverse SDEs at time t is

$$\begin{aligned}
\nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{y}_*) - \nabla_{\boldsymbol{x}_t}\log p_t(\boldsymbol{x}_t|\boldsymbol{y}) &= \cancel{\boldsymbol{s_\theta}(\boldsymbol{x}_t,t)} + \nabla_{\boldsymbol{x}_t}\log\tilde{p}_t(\boldsymbol{y}_*|\boldsymbol{x}_t) - \cancel{\boldsymbol{s_\theta}(\boldsymbol{x}_t,t)} - \nabla_{\boldsymbol{x}_t}\log\tilde{p}_t(\boldsymbol{y}|\boldsymbol{x}_t) \\
&= \nabla_{\boldsymbol{x}_t}\log\tilde{p}_t(\boldsymbol{y}_*|\boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t}\log\tilde{p}_t(\boldsymbol{y}|\boldsymbol{x}_t).
\end{aligned}$$

By the chain rule for the DPS likelihood approximation,

$$\nabla_{\boldsymbol{x}_t} \log \tilde{p}_t(\boldsymbol{y} \mid \boldsymbol{x}_t) = J_{\hat{\boldsymbol{x}}_0}(\boldsymbol{x}_t)^\top \nabla_{\hat{\boldsymbol{x}}_{0|t}} \log p(\boldsymbol{y} \mid \hat{\boldsymbol{x}}_{0|t}).$$

The difference between the two drift terms in the reverse SDEs can be bounded as

$$\|f(\boldsymbol{x}_t, t) - f'(\boldsymbol{x}_t, t)\| = \|-\beta(t)\left[\nabla_{\boldsymbol{x}_t} \log \tilde{p}_t(\boldsymbol{y}_* | \boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} \log \tilde{p}_t(\boldsymbol{y} | \boldsymbol{x}_t)\right]\| \le \beta(t) L_{\hat{\boldsymbol{x}}_0} L_{\mathrm{lld}} \|\boldsymbol{y}_* - \boldsymbol{y}\|.$$

Since $g(t) = \sqrt{\beta(t)}$ for the VP-SDE and $\beta(t) \le \beta_{\max}$, the above bound also provides square integrability: $\|\boldsymbol{y}_* - \boldsymbol{y}\|^2 \int_\delta^T \frac{\beta(t)^2 L_{\hat{\boldsymbol{x}}_0}^2}{g(t)^2}\,\mathrm{d}t = \|\boldsymbol{y}_* - \boldsymbol{y}\|^2 L_{\hat{\boldsymbol{x}}_0}^2 L_{\mathrm{lld}}^2 \int_\delta^T \beta(t)\,\mathrm{d}t < \infty$ by lemmas D.3 and D.5. Hence, the Girsanov Theorem D.4 can be applied

$$D_{KL}\big(p_\delta(\boldsymbol{x}|\boldsymbol{y}_*) \| p_\delta(\boldsymbol{x}|\boldsymbol{y})\big) \le D_T + \int_\delta^T \frac{\beta(t)^2}{2g(t)^2} \mathbb{E}\Big[\|\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{y}_*) - \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{y})\|^2\Big]\mathrm{d}t.$$

Since both initial distributions are standard Gaussian, the initial term D_T vanishes, yielding

$$D_{\mathrm{KL}}\big(p_\delta(\cdot \mid \boldsymbol{y}_*) \,\|\, p_\delta(\cdot \mid \boldsymbol{y})\big) \le C_{\mathrm{stable}} \|\boldsymbol{y}_* - \boldsymbol{y}\|^2,\ C_{\mathrm{stable}} \coloneqq \frac{1}{2} L_{\mathrm{lld}}^2 L_{\hat{\boldsymbol{x}}_0}^2 \int_\delta^T \beta(t)\mathrm{d}t < \infty.$$

D.3 Proof of Theorem 3.2

Proof of Theorem 3.2. The proof follows the same Girsanov argument as above. The only change is the likelihood-score sensitivity bound. For the robust likelihood $\tilde{p}^{(\ell)}$, the approximate likelihood score is given by chain rule as

$$\nabla_{\boldsymbol{x}_t} \log \tilde{p}_t^{(\ell)}(\boldsymbol{y} \mid \boldsymbol{x}_t) = -J_{\hat{\boldsymbol{x}}_0}(\boldsymbol{x}_t, t)^\top \nabla_{\hat{\boldsymbol{x}}_{0|t}} \ell_{\boldsymbol{y}}(\hat{\boldsymbol{x}}_{0|t}).$$

where $\ell_{\boldsymbol{y}}(\hat{\boldsymbol{x}}_{0|t}) = \sum_{i=1}^{d_y} \rho(r_i)$, and $r_i = y_i - [F\hat{\boldsymbol{x}}_{0|t}]_i$. For the robust loss $\rho(r) = \frac{1}{2\sigma_y^2} w(r) r^2$, we have $\psi(r) := \rho'(r) = \frac{1}{2\sigma_y^2}\left(2rw(r) + r^2 w'(r)\right)$. Under the robust weight condition on w given in Equation 4, both $|rw(r)|$ and $|r^2 w'(r)|$ are uniformly bounded. Hence there exists $K < \infty$ such that $\sup_{r\in\mathbb{R}} |\psi(r)| \le K$. Therefore,

$$\left\|\nabla_{\hat{\boldsymbol{x}}_{0|t}} \ell_{\boldsymbol{y}}^{(\rho)}(\hat{\boldsymbol{x}}_{0|t})\right\| = \left\|\sum_{i=1}^{d_y} \psi(r_i) \nabla_{\hat{\boldsymbol{x}}_{0|t}} r_i\right\| \le K\|F\|\sqrt{d_y},$$

uniformly over $\boldsymbol{y}$. Thus, for any $\boldsymbol{y}$, $\left\|\nabla_{\hat{\boldsymbol{x}}_{0|t}} \ell_{\boldsymbol{y}_*} - \nabla_{\hat{\boldsymbol{x}}_{0|t}} \ell_{\boldsymbol{y}}\right\| \le 2K\|F\|\sqrt{d_y}$. With Lemma D.5 we obtain

$$\left\|\nabla_{\boldsymbol{x}_t} \log \tilde{p}_t^{(\rho)}(\boldsymbol{y}_* \mid \boldsymbol{x}_t) - \nabla_{\boldsymbol{x}_t} \log \tilde{p}_t^{(\rho)}(\boldsymbol{y} \mid \boldsymbol{x}_t)\right\|^2 \le 4L_{\hat{\boldsymbol{x}}_0}^2 K^2 L_F^2 d_y.$$

Applying Theorem D.4 and using $g(t)^2 = \beta(t)$ for the VP-SDE gives

$$\begin{aligned} D_{\mathrm{KL}}\left(p_\delta^{(\ell)}(\cdot \mid \boldsymbol{y}_*) \,\|\, p_\delta^{(\ell)}(\cdot \mid \boldsymbol{y})\right) &\le \frac{1}{2}\int_\delta^T \beta(t)\, 4L_{\hat{\boldsymbol{x}}_0}^2 K^2 L_F^2 d_y\,\mathrm{d}t \\ &\le 2(T-\delta)\beta_{\max} L_{\hat{\boldsymbol{x}}_0}^2 K^2 L_F^2 d_y < \infty. \end{aligned}$$

Therefore, the KL divergence under the robust likelihood is uniformly bounded with respect to the measurement discrepancy, unlike the Gaussian likelihood case whose bound grows with $\|\boldsymbol{y}_* - \boldsymbol{y}\|^2$.

□

E Supplementary about RDP

E.1 General Pseudo Algorithm

We include the algorithm design of RDP for generic diffusion posterior sampling in this section.

Algorithm 2 RDP Plug in for Generic Diffusion Posterior Sampling

Require: observation $\boldsymbol{y}$, score network $\boldsymbol{s}_{\boldsymbol{\theta}}$, noise schedule $\{\beta_t\}_{t=0}^{T}$, weight function $w(\cdot)$, temperature $\{\tau_t\}_{t=0}^{T}$

Initialize $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, ..., 1$ **do**
  $\alpha_t = 1 - \beta_t, \quad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$
  $\hat{\boldsymbol{x}}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(\boldsymbol{x}_t + (1 - \bar{\alpha}_t)\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t))$
  $\boldsymbol{r}_t \leftarrow \boldsymbol{y} - F(\hat{\boldsymbol{x}}_{0|t})$
  $\ell_{\boldsymbol{y}}(\boldsymbol{x}_t) \leftarrow -\sum_{i=1}^{d_y} w(r_{t,i}) \log p(y_i \mid \hat{\boldsymbol{x}}_{0|t})$
  $\boldsymbol{x}_{t-1}$= UpdateStep($\boldsymbol{x}_t, \boldsymbol{s}_{\boldsymbol{\theta}}, \nabla \ell_{\boldsymbol{y}}(\boldsymbol{x}_t), \alpha_t, \beta_t$)
end for

- The likelihood term $p(\boldsymbol{y} \mid \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))$ enforces data consistency; however, its specific formulation depends on the noise and operator assumptions. Our method can be used to *robustify* a broad family of algorithms. In this paper, we focus on the DPS algorithm, which approximates the likelihood via the surrogate mean $\hat{\boldsymbol{x}}_{0|t}$ as a point estimate. For ΠGDM [Song et al., 2023a] (which we use in our experiments), the method utilizes SVD to define the likelihood in the spectral domain, modeled as $p(\boldsymbol{y}|\boldsymbol{x}_t) \approx \mathcal{N}(\boldsymbol{y} \mid F(\hat{\boldsymbol{x}}_{0|t}), r_t \boldsymbol{I})$. We integrate RDP by reweighting the negative log-likelihood objective: $\ell_{\text{RDP}} \approx -\sum_{i=1}^{d_y} w(r_{t,i}) \log \mathcal{N}(y_i \mid [F(\hat{\boldsymbol{x}}_{0|t})]_i, r_t)$, while additionally incorporating the annealing variance r_t. Furthermore, Boys et al. [2024] proposes a method for matching the variance using the second-order Tweedie's formula. This higher-order approximation of variance can similarly be incorporated with RDP.
- The UPDATESTEP is solver-agnostic, allowing robust guidance to be integrated into various reverse-time discretizations. This includes standard SDE and ODE probability flow sampling, Predictor-Corrector (PC) schemes where Langevin dynamics refine the posterior estimate via $\nabla \ell_{\boldsymbol{y}}$, as well as recent progress in recent midpoint estimation approaches such as Janati et al. [2024], Moufad et al. [2025]. Furthermore, the guidance term seamlessly extends to high-order solvers—such as DPM-Solver++ [Lu et al., 2025] or Heun's method [Karras et al., 2022], enabling efficient sample generation with fewer steps.

E.2 Other Weight functions

- **Mahalanobis-based distance.** When the measurement noise exhibits a nontrivial heterogeneous structure (e.g., $\sigma_i \neq \sigma_j$ for $i \neq j$ in the Gaussian case), it is natural to measure residuals using Mahalanobis-based (MB) distance with a diagonal covariance $\Sigma = (\sigma_1, ..., \sigma_{d_y})$, leading to weights $w_i(y_i, \hat{y}_{t,i}) = (1 + \frac{(y_i - \hat{y}_{t,i})^2}{\sigma_i c^2})^{-\frac{1}{2}}$.
- **Global scale.** Many diffusion samplers choose the step size based on the residual norm. For example, Chung et al. [2023] suggests choosing this step size to be inversely proportional to the measurement error norm $\|\boldsymbol{y} - F(\hat{\boldsymbol{x}}_0)\|_2$ [Daras et al., 2024]. Within the GB framework, this corresponds to taking a global constant $w = \frac{c}{\varepsilon + \|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2}$ for some constant $c \in (0, 1]$ and a small $\varepsilon > 0$. This stabilizes the reverse process by preventing excessively large gradient steps. However, because it uses global scaling, it down-weights clean observations along with outliers and can thus be viewed as a special case of RDP.
- **Huber (L_1 Loss).** Another well known robust loss function is Huberized Loss [Huber, 2011], which combines the sensitivity of squared error (L_2) for small residuals with the robustness of absolute error (L_1) for large residuals. For a residual $r_i = y_i - \hat{y}_{t,i}$ and a

specified transition threshold c, i.e.,

$$\ell_c(r_i) = \begin{cases} \frac{1}{2} r_i^2, & |r_i| \leq c, \\ c\left(|r_i| - \frac{1}{2}c\right), & \text{otherwise.} \end{cases}$$

This loss preserves the quadratic Gaussian penalty for small residuals, while clipping the gradient magnitude to c for large residuals. As a result, large corrupted measurements are prevented from dominating the guidance.

F Additional Experiment Details

This appendix provides additional details on the experimental setup, including task definitions, measurement models, noise configurations, hyperparameters, and implementation details. All experiments were run on two NVIDIA `A100` GPUs with `40GiB` memory.

F.1 Tasks

We evaluate our proposed method across a diverse set of inverse problems, ranging from standard linear image restoration tasks to complex nonlinear scientific imaging. The measurement process is generally modeled as $\boldsymbol{y} = F(\boldsymbol{x}) + \boldsymbol{n}$, where $F(\cdot)$ represents the measurement model and $\boldsymbol{\epsilon}$ denotes the measurement noise.

- (Linear) Inverse Scattering: This task represents a linearized version of the general nonlinear inverse scattering problem. The measurement model is given by: $\boldsymbol{y} = \mathbf{H}(\boldsymbol{u}_{\text{in}} \odot \boldsymbol{x}) + \boldsymbol{\epsilon}$, where $\boldsymbol{u}_{\text{in}}$ is the known incident light field, $\mathbf{H}$ is the Green's function propagator, and $\odot$ denotes the element-wise product. See more details in Section 3 of Zheng et al. [2025]
- Image Inpainting: The measurement model F is a diagonal masking matrix $\mathbf{M} \in \{0, 1\}^d$, which effectively removes a random fraction of the pixels. The measurement is given by $\boldsymbol{y} = \mathbf{M} \odot \boldsymbol{x} + \boldsymbol{\epsilon}$. The mask $\mathbf{M}$ consists of a 32×32 box placed at a random location within the image.
- (Gaussian) Deblurring: The measurement model F corresponds to a convolution with an isotropic Gaussian kernel. The kernel is chosen with a standard deviation $\sigma_{\text{blur}} = 3$ and a kernel size 61×61.
- Phase Retrieval: A critical problem in coherent diffractive imaging (e.g., X-ray crystallography), where detectors capture only the intensity of the diffracted wave. The measurement model is defined as $\mathbf{y} = |\mathcal{T}(\mathbf{x})| + \mathbf{n}$, where $\mathcal{T}$ denotes the 2D discrete Fourier transform and $|\cdot|$ is the element-wise magnitude operator. This problem is severely ill-posed due to the loss of phase information.

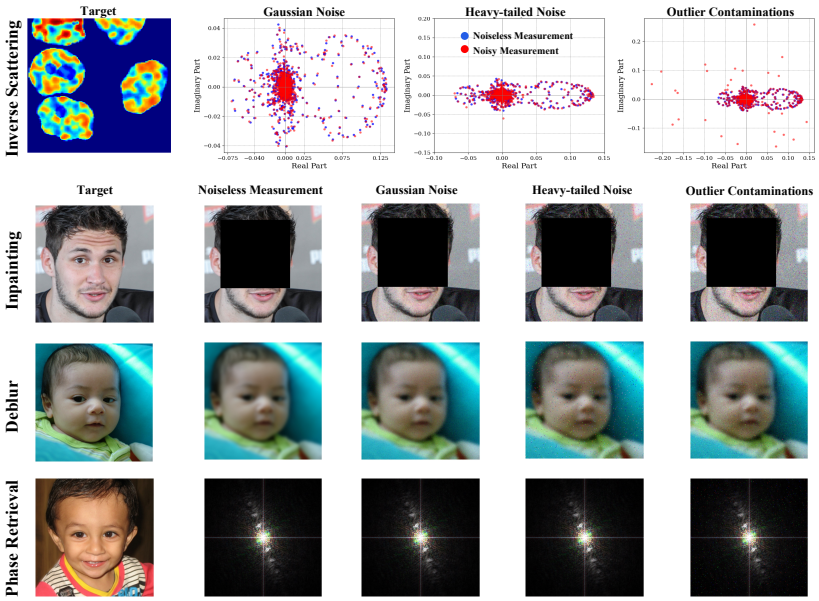


Figure 7: Visualization of measurements across three noise schemes (left→right: Gaussian, heavy-tailed, outlier).

F.2 Experimental Design

In our numerical experiments in Section 4, we assess the robustness of the proposed modular approaches under three representative settings that commonly arise in scientific inverse problems:

- Well-specified Likelihood (Gaussian Noise): We utilize standard additive white Gaussian noise as the baseline control. This represents the ideal setting where the likelihood assumed by standard solvers matches the true noise distribution.

- Distributional Misspecification (Heavy-tailed Noise): To test robustness against violations of the Gaussian assumption, we model noise using a Student-t distribution. The scale parameter is calibrated to match the standard deviation of the corresponding Gaussian baseline, while the degrees of freedom ν control the tail heaviness. Heavy-tailed errors are extensively studied in robust estimation problems such as in Aravkin et al. [2014] and Janjić et al. [2018].

- Outlier Contamination (Sparse Impulsive Noise): We model outlier contamination by adding large-magnitude noise to a randomly selected fraction of the measurement components. The severity is controlled by two parameters: the corruption ratio p (the percentage of contaminated measurements) and the amplification factor m (the magnitude of the outlier values). This setting challenges the solver to identify and reject localized corruptions, simulating practical issues such as heat noise spikes or data transmission scattering errors. To simulate outlier corruptions, we first sample random noise from the base Gaussian distribution and then rescale the noise magnitude of a randomly chosen $p\%$ of the dimensions by a factor of m.

Table 3: Noise Schemes Summary. Well-specified lists the standard deviation σ_y used for Gaussian baselines. Misspecification details the Student-t parameters (ν and calibrated scale). Contamination specifies the outlier percentage (p) and the magnitude factor (m), where m denotes the amplification factor applied to the nominal noise standard deviation σ_y.

	Well-Specified ($\mathcal{N}(0, \sigma_y^2)$)	Misspecification (Student-t)	Contamination (Outliers p, Mag m)
IS	$\sigma_y = 10^{-3}$	$\nu = 2.2$	$p = 1\%, m = 30\times$
FFHQ-I	$\sigma_y = 0.05$	$\nu = 2.5$	$p = 5\%, m = 30\times$
FFHQ-D	$\sigma_y = 0.05$	$\nu = 2.5$	$p = 5\%, m = 30\times$
FFHQ-PR	$\sigma_y = 10^{-3}$	$\nu = 2.5$	$p = 1\%, m = 10\times$

F.3 Pretrained Diffusion Model Details

We use pretrained diffusion models with checkpoints provided by *InverseBench* [Zheng et al., 2025], which employ U-Net backbones, as in Song et al. [2021d]. Detailed configurations are listed in Table 4.

Table 4: Model network configurations for pre-trained diffusion models.

	Inverse Scattering	FFHQ
Input resolution	128×128	$3 \times 256 \times 256$
# Attention blocks in encoder/decoder	5	6
# Residual blocks per resolution	1	3
Attention resolutions	16	16
# Parameters	26.8M	93.6M

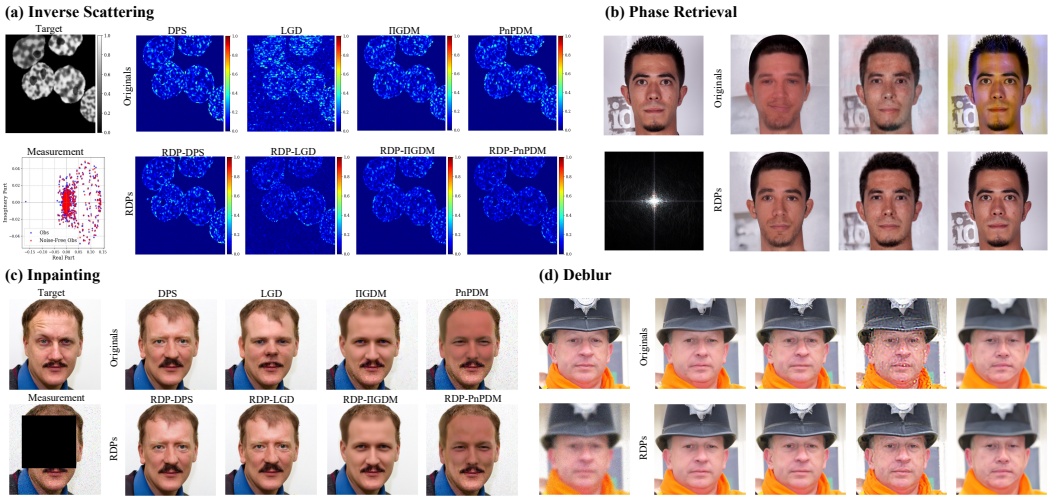


Figure 8: Qualitative results on all noise misspecification tasks: (a) Inverse Scattering, (b) Phase Retrieval, (c) Deblurring, and (d) Inpainting. In each panel, the leftmost column shows the ground truth and the contaminated observation. The remaining columns compare the original methods (top row) with their robustified versions using our sampler (bottom row); we report reconstruction absolute errors in (a) and reconstructed samples in (b)–(d).

Table 5: Hyperparameter tuning for different sampling methods and tasks. For other hyperparameters in PnPDM, we found it does not have significant effects to the results and we kept the configurations with Zheng et al. [2025]

	Inverse Scattering		FFHQ-Inpaint		FFHQ-Deblur		FFHQ-PR	
Method	Used	Range	Used	Range	Used	Range	Used	Range
DPS								
Guidance Scale	350	$[100, 500]$	1	$[0.1, 50]$	5	$[0.1, 50]$	0.5	$[0.1, 50]$
LGD								
Guidance Scale	2500	$[500, 5000]$	1	$[0.1, 50]$	5	$[0.1, 100]$	1	$[0.1, 100]$
ΠGDM								
Stochasticity η	0.3	$(0, 1]$	0.95	$(0, 1]$	0.85	$(0, 1]$	-	-
PnPDM								
Annealing Max ρ_{max}	20	$[1, 100]$	20	$[1, 100]$	10	$[1, 100]$	10	$[1, 100]$
Annealing decay rate ρ	0.9	$[0.5, 1)$	0.9	$[0.5, 1)$	0.9	$[0.5, 1)$	0.9	$[0.5, 1)$
Langevin step size τ	5×10^{-4}	$(0, 1]$	5×10^{-4}	$(0, 1]$	5×10^{-4}	$(0, 1]$	1×10^{-4}	$(0, 1]$
DiffPIR								
Regularization λ	4×10^{-4}	$[10^{-5}, 10^{-2}]$	1	$[10^{-1}, 10^{4}]$	8	$[10^{-1}, 10^{4}]$	8	$[10^{-1}, 10^{4}]$
Stochasticity η	1	$[0, 1]$	1	$[0, 1]$	0.5	$[0, 1]$	0.5	$[0, 1]$



Figure 9: Multiple posterior samples from DPS and LGD with and without the RDPs on FFHQ, given heavy-tailed noises.

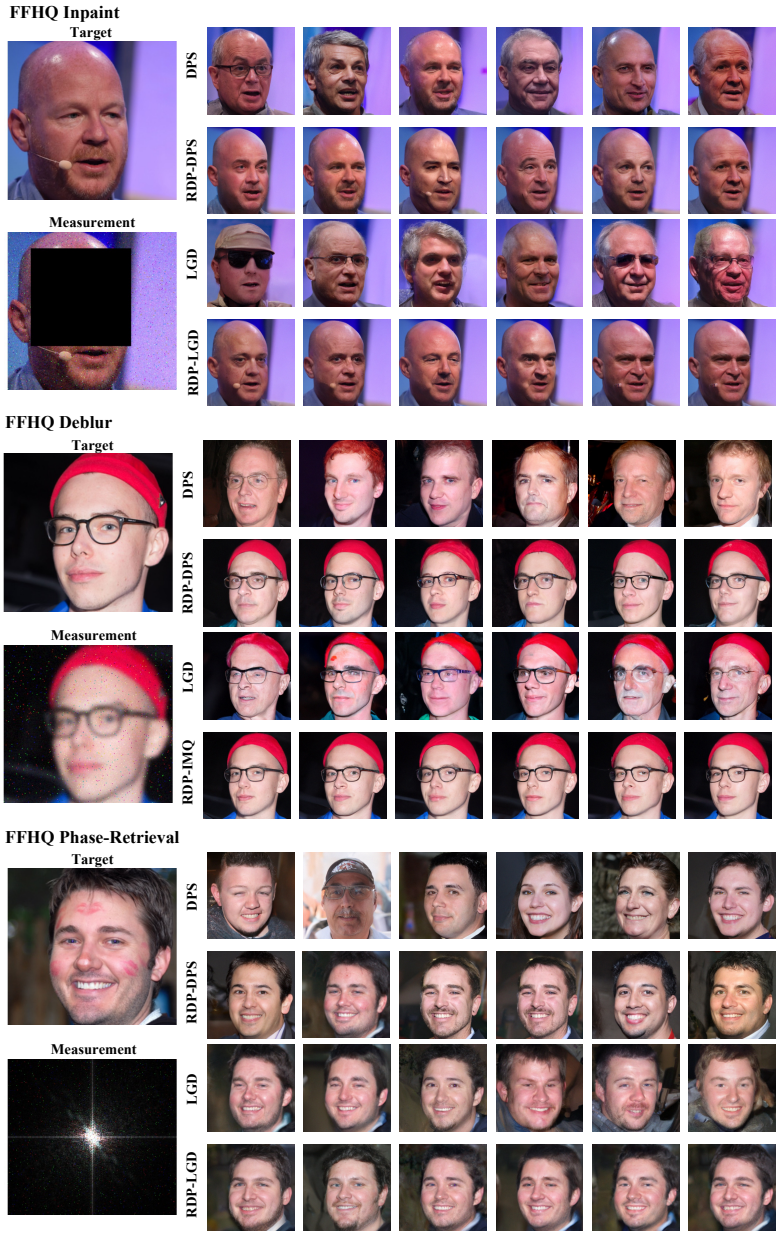


Figure 10: Multiple posterior samples from DPS and LGD with and without the RDPs on FFHQ, given observations corrupted by outliers.

F.4 Additional Experimental Results

Performance under varying measurement dimension. In practice, the measurement dimension d_y, which determines how much information is available from observations, is often more influential for inverse problems than the target dimension d_x. We conduct an ablation on inverse scattering (IS) task, where the number of receivers recording the scattered wavefield controls the amount of observed data and hence changes d_y. Specifically, we evaluate $d_y \in \{360, 180, 90\}$.

Under Gaussian noise, the gap between DPS and RDP-DPS remains small, consistent with the main results. In the other two settings, however, the advantage of RDP becomes more pronounced as d_y decreases. This suggests that robust guidance is more important when measurements are limited, since each corrupted measurement has a larger influence on the posterior update.

Table 6: Performances under different noise settings and receiver configurations

Method	Num of Receivers/ d_y	Well-specification (Gaussian)			Noise Misspecification (Student-T)			Contamination of Outliers		
		PSNR	SSIM	NMAE	PSNR	SSIM	NMAE	PSNR	SSIM	NMAE
DPS	360/7200	26.26±2.83	0.86±0.02	0.16±0.02	21.53±2.93	0.78±0.06	0.25±0.05	19.37±3.44	0.71±0.10	0.33±0.05
RDP-DPS	360/7200	26.01±2.72	0.86±0.02	0.15±0.01	23.84±2.59	0.82±0.03	0.19±0.04	25.24±2.74	0.83±0.02	0.15±0.02
DPS	180/3600	24.50±2.90	0.84±0.03	0.19±0.02	19.68±3.15	0.73±0.07	0.31±0.06	17.28±3.68	0.63±0.12	0.41±0.06
RDP-DPS	180/3600	24.20±2.78	0.83±0.03	0.18±0.01	22.50±2.65	0.79±0.04	0.22±0.04	23.58±2.82	0.80±0.03	0.17±0.02
DPS	90/1800	22.76±3.05	0.81±0.04	0.23±0.03	17.56±3.42	0.68±0.08	0.37±0.07	15.76±3.95	0.56±0.13	0.47±0.07
RDP-DPS	90/1800	22.40±2.86	0.80±0.04	0.22±0.02	20.98±2.78	0.75±0.05	0.26±0.05	22.54±2.95	0.77±0.04	0.20±0.03

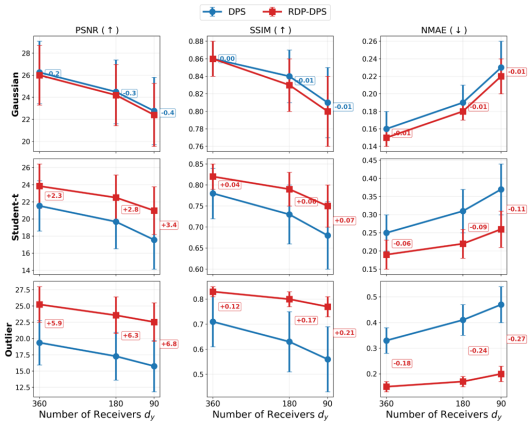


Figure 11: Recovery performance of DPS and RDP-DPS as the measurement dimension d_y varies under three noise settings on the IS task. RDP-DPS shows larger gains when measurements are limited and corrupted.

Performances under noise mismatch levels. To evaluate a more practical form of model misspecification, we further consider moderate noise-level mismatch on the IS task. Specifically, the sampler assumes a Gaussian likelihood with noise standard deviation σ, while the true measurement noise has standard deviation 2σ, 4σ, or 8σ. Results are reported in Table 7.

The advantage of RDP-DPS increases as the mismatch level becomes larger. For example, RDP-DPS achieves comparable performance under $2\times$ mismatch and provides clearer improvements under $4\times$ and $8\times$ mismatch, especially in NMAE. The gains are smaller than those observed under outlier contamination or Student-t noise, which is expected because noise-level mismatch affects all measurements uniformly rather than producing clearly distinguishable corrupted components. In this case, component-wise reweighting is less directly targeted, but it still mitigates the effect of an overconfident likelihood and provides consistent improvements.

Table 7: Performances under mismatch levels.

Method	σ_y mismatch	PSNR	SSIM	NMAE
DPS	$2\times$	23.64±2.27	0.81±0.03	0.23±0.01
RDP-DPS	$2\times$	23.72±2.30	0.82±0.02	0.23±0.01
DPS	$4\times$	20.58±2.19	0.75±0.05	0.33±0.01
RDP-DPS	$4\times$	20.88±2.16	0.76±0.04	0.30±0.02
DPS	$8\times$	17.94±2.30	0.69±0.07	0.40±0.03
RDP-DPS	$8\times$	18.18±2.22	0.71±0.07	0.37±0.03

Wall-clock time. We further evaluate the computational overhead introduced by RDP on the IS task. Since RDP only adds a component-wise residual weighting to the guidance term, it incurs negligible additional cost per sampling step. As shown in Table 8, RDP-DPS has nearly identical wall-clock time to DPS.

Table 8: Wall-clock time per sample on the IS task, averaged over 100 samples.

Method	Samples	Mean	Std
DPS	100	30.2s	1.4s
RDP-DPS	100	30.4s	1.3s

Ablation study on weighting functions and quantiles. The choice of a robust weighting function can affect recovery quality. To examine whether the performance of RDP depends critically on the specific IMQ weighting, we conduct ablations on the IS task over two design choices: the weighting function and the adaptive quantile q.

In addition to IMQ, we evaluate a weighting function that yields the standard Huber loss through:

$$w(r) = \begin{cases} 1, & |r| \le c, \\ 2c/|r| - c^2/r^2, & |r| > c. \end{cases}$$

where the transition threshold c is set adaptively as the 75th percentile of the residual magnitudes $|r|$, following the same calibration strategy as IMQ.

The results show that IMQ and Huber weighting perform comparably under Gaussian and Student-t noise, while IMQ gives a slight advantage under outlier contamination. Since both weights use quantile-calibrated thresholds, they suppress a similar set of large-residual measurements, although IMQ provides a smoother transition.

Table 9: Ablation on weighting function on IS.

Weighting	Noise	PSNR	SSIM	NMAE
IMQ	Gaussian	26.01±2.72	0.86±0.02	0.15±0.01
Huber	Gaussian	26.00±2.45	0.86±0.01	0.15±0.01
IMQ	Student-T	23.84±2.59	0.82±0.03	0.19±0.04
Huber	Student-T	23.82±2.41	0.81±0.02	0.20±0.01
IMQ	Outliers	25.24±2.74	0.83±0.02	0.15±0.02
Huber	Outliers	24.73±2.51	0.83±0.02	0.17±0.01

We further evaluate the IMQ quantile $q \in \{0.10, 0.25, 0.50, 0.75, 0.95\}$. Performance is stable for $q \ge 0.25$, with a marginal drop at $q = 0.25$, whereas $q = 0.10$ is overly aggressive and degrades performance under outlier contamination by down-weighting useful clean measurements. These results indicate that RDP is not highly sensitive to the weighting design and support IMQ with $q = 0.75$ as a reasonable choice.

Table 10: Ablation on quantile q of IMQ on IS.

q	Student-T			Outliers		
	PSNR	SSIM	NMAE	PSNR	SSIM	NMAE
0.10	23.47±2.64	0.81±0.04	0.21±0.04	22.63±2.81	0.80±0.03	0.19±0.02
0.25	23.59±2.60	0.82±0.03	0.20±0.03	24.96±2.76	0.82±0.02	0.16±0.02
0.50	23.81±2.58	0.82±0.03	0.19±0.04	25.24±2.76	0.83±0.02	0.15±0.02
0.75	23.85±2.59	0.82±0.03	0.19±0.04	25.24±2.74	0.83±0.02	0.15±0.02
0.95	23.83±2.60	0.82±0.04	0.19±0.04	25.25±2.73	0.83±0.02	0.15±0.02

Performance under varying outlier magnitudes. To examine the behavior of RDP under varying levels of outlier corruption, we extend our evaluation on the FFHQ-Deblur and FFHQ-Inpainting tasks, comparing DPS and RDP-DPS over a substantially broader range of outlier magnitudes.

Specifically, we fix the clean-measurement noise level and consider magnification factors $m \in \{2, 5, 10, 1000\}$. This sweep covers both the ambiguous low-magnification regime, in which the

Table 11: Performance of DPS and RDP-DPS across outlier magnification factors on FFHQ deblurring and inpainting. Results are reported as mean ± standard deviation over 100 images per task.

Task	Magnification	DPS		RDP-DPS	
		VIFL ↓	LPIPS ↓	VIFL ↓	LPIPS ↓
Deblur	2×	0.63 ± 0.05	0.17 ± 0.04	0.64 ± 0.05	0.15 ± 0.04
	5×	0.65 ± 0.05	0.18 ± 0.04	0.64 ± 0.05	0.13 ± 0.04
	10×	0.73 ± 0.05	0.23 ± 0.05	0.68 ± 0.05	0.18 ± 0.04
	1000×	0.93 ± 0.03	0.35 ± 0.11	0.74 ± 0.05	0.22 ± 0.07
Inpaint	2×	0.62 ± 0.07	0.13 ± 0.04	0.63 ± 0.07	0.15 ± 0.05
	5×	0.67 ± 0.07	0.20 ± 0.05	0.65 ± 0.06	0.17 ± 0.06
	10×	0.79 ± 0.06	0.26 ± 0.05	0.72 ± 0.06	0.14 ± 0.06
	1000×	0.94 ± 0.03	0.41 ± 0.17	0.77 ± 0.04	0.26 ± 0.07

Table 12: Performance of DPS and RDP-DPS on inverse scattering (IS) under different numbers of denoising steps. Results are reported as mean ± standard deviation over 100 test cases.

Method	# Denoising Steps	PSNR ↑	SSIM ↑	NMAE ↓
DPS	100	15.91 ± 2.60	0.32 ± 0.05	0.53 ± 0.03
RDP-DPS	100	19.32 ± 2.67	0.41 ± 0.02	0.38 ± 0.03
DPS	200	16.18 ± 2.99	0.50 ± 0.07	0.45 ± 0.04
RDP-DPS	200	21.68 ± 2.98	0.61 ± 0.03	0.27 ± 0.02
DPS	500	18.41 ± 3.33	0.64 ± 0.09	0.37 ± 0.05
RDP-DPS	500	24.06 ± 3.00	0.80 ± 0.02	0.18 ± 0.02
DPS	1000	19.37 ± 3.44	0.71 ± 0.10	0.33 ± 0.05
RDP-DPS	1000	25.24 ± 2.74	0.83 ± 0.02	0.15 ± 0.02

clean and corrupted measurement distributions overlap substantially, and an extremely severe outlier regime with $\sigma_{\text{outlier}} = 50$. The latter case also provides empirical support for the outlier-robustness guarantee of RDP-DPS established in Theorem 3.2. All other experimental settings, including the outlier proportion, are kept unchanged.

Table 11 summarizes the results across the outlier magnitudes. Overall, RDP remains as effective as the base sampler in well-specified or weakly contaminated settings, while providing clear robustness gains as the outliers become severe. This is practically useful in BIPs, where the exact noise condition is often unknown in advance.

Sensitivity to the number of function evaluations. We study the trade-off between efficiency and robustness under outlier-contaminated measurements by varying the number of function evaluations (NFEs). Specifically, we evaluate DPS and RDP-DPS on inverse scattering (IS) and FFHQ-Deblur with $\{1000, 500, 200, 100\}$ denoising steps. Results are reported in Table 12 and Table 13, respectively.

The NFE sensitivity study yields three observations, which we discuss below.

- *Effectiveness under reduced NFEs.* Across both IS and FFHQ-Deblur, RDP-DPS consistently outperforms DPS at every tested NFE, including the low-NFE settings. This suggests that RDP can remain effective even when the sampling budget is reduced.
- *No additional NFEs requirement.* Both DPS and RDP-DPS improve as the number of denoising steps increases, suggesting that the need for sufficient NFEs is primarily inherited from the base diffusion sampler rather than being a limitation specific to RDP.
- *Adaptive thresholding mitigates early-stage over-suppression.* RDP uses an adaptive quantile threshold rather than a fixed residual threshold. When the residual scale is globally large, the threshold also increases, so the method does not downweight residuals solely because their absolute magnitude is large.

Table 13: Performance of DPS and RDP-DPS on FFHQ-Deblur under different numbers of denoising steps. Results are reported as mean $\pm$ standard deviation.

Method	# Denoising Steps	LPIPS $\downarrow$	VIFL $\downarrow$
DPS	100	0.53 ± 0.06	0.95 ± 0.03
RDP-DPS	100	0.41 ± 0.06	0.924 ± 0.05
DPS	200	0.46 ± 0.06	0.93 ± 0.04
RDP-DPS	200	0.35 ± 0.06	0.87 ± 0.04
DPS	500	0.39 ± 0.07	0.91 ± 0.06
RDP-DPS	500	0.28 ± 0.05	0.83 ± 0.05
DPS	1000	0.32 ± 0.07	0.83 ± 0.07
RDP-DPS	1000	0.25 ± 0.06	0.75 ± 0.03

We further note that a smaller NFE does not correspond to using only early denoising stages: the NFE controls the discretization budget over the entire reverse trajectory, whereas early and late refer to diffusion time. The lower-NFE behavior in Figure 3 therefore reflects the coarser discretization of the full trajectory, rather than RDP suppressing useful guidance in early denoising steps. In general, RDP introduces no additional NFE requirement: it inherits the sampling budget of the base diffusion sampler while providing robustness against outlier contamination.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. The main claims in the abstract accurately reflect the paper's contributions and scope. The paper claims to study the connection between likelihood modeling and recovery quality in diffusion-based solvers for BIPs, provides stability analysis for linear BIPs, identifies robustness limitations under outlier contamination, and proposes a robust diffusion posterior sampling method. The scope is also appropriately stated: the theoretical guarantees are mainly for linear BIPs, while the empirical evaluation covers both linear and nonlinear inverse problems as well as image restoration tasks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a discussion of its main limitation in conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide assumptions and complete proof in Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All code and data are provided in supplementary materials.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: See supplementary materials.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the necessary experimental details to understand and reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports statistical variability for key experimental results using error bars that reflect standard deviation over multiple runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides details of the compute resources used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a discussion of broader impacts. On the positive side, the proposed framework has the potential to improve the robustness of methods for Bayesian inverse problems which could benefit applications such as medical imaging and other real-world systems where diffusion posterior sampling is critical. On the negative side, like other advanced machine learning techniques, there is a risk that the method could be misapplied in high-stakes systems without sufficient verification, potentially leading to unintended behavior. While this work is primarily theoretical and demonstrated in simulation and several benchmark datasets, awareness of deployment risks is important. We believe

transparency and further research into verification and robustness can help mitigate these concerns.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The models and data used in this work do not pose a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets used in this work—such as code and datasets—are properly cited in the paper. The licenses and terms of use are respected and, where applicable, included in the references or appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces new code assets implementing the proposed robust diffusion posterior sampling framework. Documents are provided in supplementary materials.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve any crowdsourcing or research with human subjects. All experiments are conducted in simulated environments without human data or interaction.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: This work does not involve human subjects, user studies, or the collection of human data. Therefore, IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were not used as part of the core methodology, experiments, or technical contributions of this work. Any use of LLMs was solely for writing refinement and editing purposes.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.