Reproducing Diffusion Posterior Sampling

Hisham Ahmed **Sigvard Dackevall** **Aqsa Kausar**

Abstract

In this project, we explore the application of diffusion models as generative solvers for noisy inverse problems, building upon the framework proposed by Chung et al. [2022a]. While diffusion models have shown promise in solving linear inverse problems under noiseless conditions, real-world scenarios often involve nonlinear and noisy settings, requiring more robust solutions. Our work re-implements the original methodology and applies it to CIFAR-10 and FFHQ datasets, demonstrating the versatility of this approach in handling noise. By leveraging an approximation of posterior sampling, this method utilizes a blend of diffusion sampling with manifold-constrained gradients, avoiding strict measurement consistency projections. The results of this study show the potential and the limitation of this method with empirical evidence. The code for this study is available on Github: `https://github.com/hishamad/Reproducing-Diffusion-Posterior-Sampling`

1 Introduction

Diffusion models have proven to be powerful tools for modeling complex data distributions, making them highly effective for solving inverse problems—tasks where an unknown signal must be reconstructed from noisy, transformed measurements. Despite their promise, practical challenges emerge when incorporating diffusion models into inverse problem-solving, particularly due to the intractability of the likelihood term in time-dependent diffusion processes. Existing methods often address this through measurement subspace projections, but such approaches struggle in realistic settings involving noise or nonlinear transformations.

Recent advancements have attempted to overcome these challenges by employing spectral-domain techniques like singular value decomposition (SVD) Kawar et al. [2021]. However, while effective for some cases, these methods are computationally intensive and limited to linear inverse problems. Nonlinear inverse problems, which are common in real-world applications, remain largely unexplored in the context of diffusion models.

The Diffusion Posterior Sampling (DPS) framework addresses these limitations by introducing a novel, image-domain approach for solving noisy and nonlinear inverse problems. This method eliminates the computational burden of SVD while accommodating for Gaussian noise and utilizing automatic differentiation for handling complex forward operators. By extending diffusion models to more challenging scenarios, this framework bridges gaps in previous methods and expands their real-world applicability.

Through experiments on CIFAR-10 and FFHQ datasets, we demonstrate the versatility of this approach for both linear tasks, like inpainting and nonlinear tasks, such as non-uniform deblurring. The results underscore the framework's ability to deliver high-quality reconstructions while addressing the key limitations of prior methods in noisy and nonlinear settings.

We successfully reproduced the experiments from the original paper on the FFHQ dataset including applying DPS on three linear inverse problems: box inpainting, super-resolution, and Guassian deblurring and one non-linear inverse problem, non-linear deblurring. As an extension to the original experiments, a colorization experiment was performed on both the FFHQ and the CIFAR-10 dataset which showed some limitations that this method have. Additionally, an investigation on the effect of key hyparameters was conducted.

2 Related Work

Diffusion models have emerged as a cornerstone in generative modeling due to their ability to iteratively transform noise into high-fidelity samples by learning the gradient of the log density (i.e., Stein score, $\nabla_x \log p(x)$) of complex data distributions Song et al. [2020], Ho et al. [2020]. The foundational work on diffusion models introduced denoising diffusion probabilistic models (DDPMs), which established the feasibility of these methods for high-quality image generation and inspired subsequent advancements. This framework was later refined with improvements in sampling efficiency, memory usage, and noise scheduling, enabling broader applications beyond image synthesis.

Building on these foundations, diffusion models have been successfully adapted for solving inverse problems. Early work by Song et al. [2020] demonstrated the use of diffusion models for inverse problems such as inpainting and super-resolution by integrating projection-based approaches to enforce consistency with measurements. Kadkhodaie and Simoncelli [2020] extended these ideas to additional applications like colorization, showcasing the flexibility of diffusion-based frameworks in recovering missing or degraded information from linear measurement processes. However, these methods primarily addressed noiseless scenarios, where the measurement data could be cleanly projected back to the underlying data distribution.

In scenarios involving noisy measurements, methods such as those proposed by Kawar et al. [2021] adapted diffusion models to the spectral domain. By tying noise in the measurement space to the spectral domain through singular value decomposition (SVD), they could solve noisy inverse problems such as deblurring and super-resolution. However, the reliance on SVD introduced computational inefficiencies and limited applicability to specific forward operators, such as separable Gaussian kernels. These constraints underscore the need for more generalizable and computationally efficient solutions.

Efforts to expand the application of diffusion models to nonlinear inverse problems have been limited. Most works, including those in compressed sensing MRI (CS-MRI) Song et al. [2021] and computed tomography (CT) Chung et al. [2022b], have focused on linear measurement processes. The introduction of manifold-constrained gradients (MCG) by Chung et al. [2022b] represented a step toward addressing noiseless nonlinear inverse problems, using gradient information to refine reconstructions without explicit projections. However, MCG remains restricted to noiseless settings, leaving a gap in addressing the challenges posed by real-world noisy nonlinear problems.

3 Methods

This project builds on the Diffusion Posterior Sampling (DPS) framework introduced by Chung et al. [2022a]. This method addresses the challenges of solving noisy and non-linear inverse problems by leveraging the generative capabilities of diffusion models. Unlike traditional approaches that rely on strict projection-based methods, this framework incorporates a smoother posterior sampling process by blending learned priors with likelihood approximations.

3.1 Diffusion-Based Generative Modeling

Diffusion models define a generative process as the reverse of a forward noising process. The forward process progressively corrupts data by adding Gaussian noise, modeled using a variance-preserving stochastic differential equation (VP-SDE). The reverse process, guided by the learned score function $\nabla_x \log p(x)$, iteratively denoises data back to the original distribution.

For inverse problems, the goal is to recover the clean signal x from noisy measurements y, which are generated via a forward operator $A(x)$ corrupted with noise n. In this framework, the diffusion process is adapted to sample from the posterior distribution $p(x|y)$ by combining the score function with the gradient of the log-likelihood $\nabla_x \log p(y|x)$. However, due to the intractability of $\nabla_x \log p(y|x)$ at intermediate steps, this term is approximated using a tractable surrogate which uses the posterior mean $\hat{x_0}$, leading to $\nabla_{x_t} \log p(y|x) \approx \nabla_{x_t} \log p(y|\hat{x_0})$ making it analytically tractable.

3.2 Algorithm

The proposed diffusion posterior sampling (DPS) algorithm is detailed below step by step:

- 1) Start from a Gaussian prior $x_T \sim \mathcal{N}(0, I)$, where T is the last time step.
- 2) Start the reverse diffusion process loop (3-8):
- 3) Feed the current x_i into the pre-trained diffusion model
- 4) Using the estimated score $\hat{s}$, calculate the posterior mean $\hat{x_0}$.
- 5) Generate Gaussian noise sample z
- 6) Using the posterior mean and the variance, calculate the next sample x_{i-1}.
- 7) Update x_{i-1} using the gradient of the norm of the difference between the measurement y and the approximated measurement $A(\hat{x}_0)$. Back to (3) until we reach time step 0.
- 8) return x_0

3.3 Modifications and Extensions

While we followed the general framework outlined in the paper, several modifications and some incremental changes were made to adapt the method to our experimental setup and datasets:

We experimented with CIFAR-10, a smaller dataset with low-resolution images. This allowed us to evaluate the method's performance on datasets with reduced representational capacity, revealing some limitations in handling smaller resolution. For this experiment, we use the pre-trained DDPM model for CIFAR-10 from Google by Ho et al. [2020], which is available on HuggingFace. The main difference between this pre-trained model and the FFHQ pre-trained model used in the original paper is that the estimated scores are returned as means unlike the FFHQ pre-trained model which return both the mean and the variance. This is an important difference as learned variance has been shown to boost the performance of the models for lower time steps in the reverse diffusion process Dhariwal and Nichol [2021]. To address this issue for CIFAR-10, we used the fixed variance method instead.

4 Data

In this project, two datasets are used: the Flickr Faces HQ (FFHQ) dataset and the popular CIFAR-10 dataset. FFHQ contains high-quality pictures of human faces which were generated using Generative Adversarial Networks (GANs). As working with diffusion models is computationally expensive, only a subset of 100 images of size 256x256 are used in the experiments. The CIFAR-10 dataset contains 32x32 images from 10 different classes airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. For this dataset, only a subsets of 100 images are used in this project to limit the need for computational resources.

Regarding pre-processing, for both datasets, only standard normalization is applied to the images. The normalization is a common pre-processing step when working with images that contain RBG values ranging from 0 to 255 to avoid having large activation values which helps speed up the convergence.

5 Experiments and Results

One of the main goals of this project is to reproduce the experiments from the original paper. The first expirements performed in this study is a reproduction of most of the original experiments on the FFHQ dataset including applying DPS on three linear inverse problems: box inpainting, super-resolution, and Guassian deblurring and one non-linear inverse problem, non-linear deblurring. These methods are implemented the same way as the original paper, where each of the following methods is applied to the images together with Gaussian noise ($\sigma = 0.05$). Here is a description of each method:

- For box inpainting, a square region of a fixed-size 128x128 is masked out from the images in a random place, corresponding to 25% of the image area.
- For super-resolution, the bi-cubic interpolation for down-sampling is used where each pixel in the down-sampled image is a weighted average of the nearest 16 pixels around it representing a 4x4 region.

- For Guassian blurring, the images are convolved using a Gaussian kernel of size 61x61 and a standard deviation of 3.0 as in the original paper.
- For non-linear blurring, similar to the original paper, a pre-trained neural network developed by Tran et al. [2021] is used.

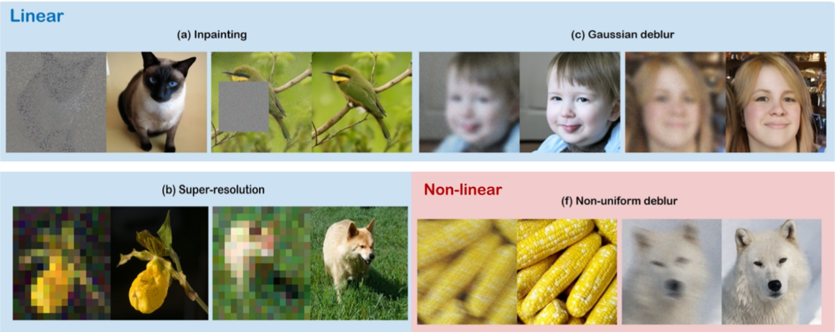

Figure 1: Solving noisy linear, and nonlinear inverse problems with diffusion models from Chung et al. [2022a]

Two main evaluation metrics are used, which are the Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS). Both of these metrics measure the perceptual similarity between the original images and the reconstructed images. These metrics are also used in the original paper which makes it easier to compare our results with theirs. FID compares the distributions of the reconstructed images with the original images. The FID scores ranges from 0 to infinity where 0 indicates a perfect match between the distributions. However, LPIPS compare two individual images at a time by passing them into a pre-trained network and the similarity between the activations of the images is returned. LPIPS scores ranges from 0 to 1 where 0 indicates a perfect match between the images. In this study, the average LPIPS score of the 100 images is taken as the final evaluation and the pre-trained network used is a VGG-net.

The results from this experiment are presented in table 1. For LPIPS, very similar results are achieved to the original paper. For some methods, such as the super-resolution method, we even performed better than the original results, indicating a successful implementation of the original paper method. However, there are some deviations, mainly with the FID score, where we are getting a higher FID score (worse results) across all methods, specially for the non-linear de-blur experiment. However, the scores differences between the methods are somewhat similar to the original results, where DPS performs best on box inpainting.

One reason for these deviations could be because of how the FID score works and the small subset (100 images) we used for evaluation. Unlike LPIPS, where individual images are compared, FID approximates the distributions of the original and reconstructed images, which means that to get more accurate approximates, more data is needed. In this case, only a subset of 100 images is used compared to 1K images used in the original paper, which means that this size may not be enough to approximate the FID score accurately.

Another reason for these deviations can be because of differences in configurations related to the method as these were not disclosed clearly in the original paper. For example, the authors did not disclose what method they used to calculate the variance and the mean in the inference steps. They provided different methods to calculate them in their repository, however, it was not mentioned which specific one was used to run the experiments on the FFHQ dataset.

There are also different step size values used in the original paper and in their code repository on GitHub. The step size refer to the scale used when adjusting the values of the reconstructed images with the new gradients at a certain time-step. This hyperparameter seems to be an important factor in determining the performance of the model which is why the experiments were performed on different time steps as seen in table 1. Higher step size seems to give better results on both FID and LPIPS for the FFHQ dataset, which led us to explore the effect of the step size more closely in the next experiment.

Table 1: results of diffusion posterior sampling using different operators with different step sizes on the FFHQ dataset (100 images)

	Box Inpaint		Super-resolution		Gaussian Deblur		Non-linear Deblur	
Step Size	FID	LPIPS	FID	LPIPS	FID	LPIPS	FID	LPIPS
0.5	62.77	0.207	73.31	0.2609	65.75	0.2399	92.57	0.3149
1.0	46.07	0.169	57.66	0.2097	57.72	0.2197	103.2	0.3240
Original	33.12	0.168	39.35	0.214	44.05	0.257	41.86	0.278

Table 2: Investigating the effect of step size and the standard deviation (sigma) on super-resolution using DPS

(a) Step size FFHQ

	Super-resolution	
Step size	FID	LPIPS
0.1	96.56	0.3928
0.3	85.39	0.3031
0.5	73.31	0.2609
1.0	57.66	0.2097

(b) Sigma

	Super-resolution	
Sigma	FID	LPIPS
0.01	80.00	0.2954
0.05	85.39	0.3031
0.25	90.65	0.3494
1.25	100.2	0.4558

The first experiment on C-level aimed to analyze the impact of step size on model performance using the super-resolution (SR) task. The results for this step size experiment are presented in Table 2a. The hyperparameters were configured to match the original paper, with sigma set to 0.05, a scaling factor of 0.25, with bicubic interpolation for downsampling. The evaluation metrics used were FID and LPIPS, measuring the similarity between the generated images and the ground truth.

Our results show that increasing the step size improves the performance of the model, as evidenced by the decreasing FID and LPIPS values. This indicates better perceptual similarity and a closer distributional match to the original image. Specifically, the results highlight that a step size of 1.0 produces the best performance, achieving an FID of 57.66 and an LPIPS of 0.2097, notably a slight improvement in relation to the original papers LPIPS score. Furthermore, our results validate the original paper's conclusions that between the range 0 - 1.0 is an optimal step size for the FFHQ model.

The second experiment on C-level focused on the effect of standard deviation (sigma), which controls the level of Gaussian noise added to the measurements during inference. The results, shown in Table 2b, were obtained by running SR experiments with four different sigma values: 0.01, 0.05, 0.25, and 1.25, while keeping the step-size at 0.3. As in the step size experiment, the metrics used were FID and LPIPS.

The results demonstrate that increasing sigma leads to worse performance in both FID and LPIPS. However, the slight decrease in performance, even when using a large sigma such as 1.25 speaks to that the model is quite robust to noise. This could be due to the fact that diffusion models inherently uses Gaussian noise in the training process and therefore it is well adapted to handle noise in the inference.

Another objective of this project was to investigate how the model performs on the CIFAR-10 dataset, expanding beyond the original experiments on the FFHQ dataset. To achieve this, we reproduced the experiments from earlier tasks, applying DPS on the CIFAR-10 dataset to evaluate four methods: box in-painting, super-resolution, Gaussian de-blurring, and the newly introduced colorization task. Similar to the FFHQ experiments, Gaussian noise with a standard deviation of sigma=0.05 was added to the input data during inference.

The following is a brief description of the colorization method applied to CIFAR-10: The images were converted to grayscale using the standard formula grayscale, Gray = 0.299R + 0.587G + 0.114B. The model then predicted the original colors of the image. For this task, we evaluated the results

Table 3: Results of diffusion posterior sampling using different operators on the CIFAR-10 dataset (100 images).

	Box Inpaint		Super-resolution		Gaussian Deblur		Colorization	
Step size	FID	LPIPS	FID	LPIPS	FID	LPIPS	SSIM	LPIPS
0.5	125.5	0.1758	159.2	0.3012	153.8	0.2943	0.9806	0.4104
1.0	137.5	0.1664	177.74	0.3285	166.5	0.3165	0.9816	0.5287

Table 4: Results of the colorization experiment on the FFHQ dataset.

Step Size/Method	Colorization	
	LPIPS	SSIM
0.1	0.6051	0.7417
0.5	0.5472	0.2434
1.0	0.5219	0.9491

using SSIM, a metric that accounts for color fidelity, alongside LPIPS for perceptual similarity. SSIM score is graded oppositely to FID and LPIPS, going from 0 to 1, where 1 represents a perfect match.

The results from the CIFAR-10 experiments are presented in Table 3. Compared to the FFHQ results, there are some deviations, particularly in the FID and LPIPS scores. For example, box inpainting and super-resolution on CIFAR-10 showed higher FID scores across all step sizes (ie. worse performance), suggesting that the model struggles more with these tasks on lower-resolution datasets like CIFAR-10. Another reason for the poorer results of the CIFAR-10 model, as mentioned in the method section, could be that we used a fixed variance in inference unlike in the FFHQ experiment where the variance was learned.

One of the core differences observed between the results for FFHQ and CIFAR-10 was that a step size of 0.5 outperformed a step size of 1.0 in the CIFAR-10 experiments. We hypothesize that this could be due to the lower data quality in CIFAR-10 dataset. Which we believe could have caused the model to benefit more from smaller gradient step sizes. This is likely because the noisier and less detailed data in CIFAR-10 requires finer adjustments to avoid overshooting during inference. In contrast, the higher-quality FFHQ dataset contains more structured information, allowing the model to gain additional efficiency with larger step sizes.

Another notable difference was the significantly higher FID scores for CIFAR-10 compared to FFHQ, while the LPIPS scores were much closer between the two datasets. We believe this discrepancy could be due from the way these metrics are calculated. FID essentially measures the divergence between the distributions of the set of output images and the set of real images, making it more sensitive to variations caused by lower data quality and limited sample sizes. In contrast, LPIPS (as well as SSIM) is computed pairwise between individual images, making it less susceptible to these distributional issues. Given our smaller sample size of 100 output images and the inherently lower quality of CIFAR-10, FID may have become less reliable as an evaluation metric for this dataset.

The results from the FFHQ colorization experiments are presented in Table 4. Notably, colorization, an inverse problem not considered in the original paper—emerges as the most challenging task for our model on both CIFAR-10 and FFHQ. In fact, our best LPIPS scores for colorization are approximately twice as large as those achieved on tasks like inpainting or super-resolution, indicating a substantial drop in perceptual fidelity. A key reason for this difficulty may be that, unlike inpainting or super-resolution where structural constraints guide the restoration, colorization requires the model to infer suitable color distributions solely from grayscale intensity patterns. Such ambiguity could make it harder for the diffusion model's learned priors to align with the correct color space.

6 Challenges and Limitations

Reimplementing the methods and conducting experiments posed several challenges that required adaptation. While we initially used the ImageNet dataset and the pre-trained ImageNet diffusion model, the large size of the pre-trained model (2.1GB) made training and inference impractical due to computational limitations, leading us to switch to the FFHQ pre-trained model (365MB) and the FFHQ dataset for manageable experimentation. Limited GPU resources on Google Colab further constrained scalability and batch sizes, impacting efficiency. In terms of presentation, the original paper assumed a high level of familiarity with diffusion-based methods, which conversely none in the group had. This meant that a big part of the project was spent on building the knowledge to be able understand the foundation of what they had done in the original paper what we needed to do to achieve the goals we set out.

Regarding technical implementation, there are many details that were not mentioned in the paper. For example, in the inference algorithm, they provide an equation for calculating the posterior mean, however, when inspecting their repository, they had many methods to calculate the mean and variance which was not explained in the paper. The experimental details in the paper also did not match the configurations used in the code which also made it harder for us to decide for instance what step size to use and so on.

To evaluate the method on lower-resolution images, we experimented with CIFAR-10, but its small resolution posed challenges for generation, resulting in noticeably worse performance compared to FFHQ. We were also unable to perform non-linear blurring experimentation on CIFAR-10 because the pretrained model, GOPRO wVAE, used a kernel size in the architecture, that is not compatible for CIFAR-10 image sizes.

7 Self-Assesment

We firmly believe that we have met and, in some cases, exceeded the scope of our original project proposal, in which we set the ambition to achieve a grade A. At a fundamental level, we successfully reproduced several key results of the original paper, demonstrating our understanding of the methodology and its empirical outcomes. In addition to replication, we conducted a detailed discussion analyzing discrepancies between our results and those reported in the original paper.

For the intermediate grading level (D-C-B), we examined the model's sensitivity to hyperparameters by exploring two hyperparameters which resulted in a total of 8 experiements. While our original plan aimed to investigate noise scheduling and timestep sensitivity, these goals proved unattainable within the scope of the project. Achieving them would have required training the model from scratch, which was infeasible given our time and resource constraints. Instead, we selected two other parameters that we hypothesized would significantly impact the model's performance during inference. This adjustment demonstrates both adaptability and critical thinking, as we prioritized meaningful and achievable experiments to ensure a thorough evaluation.

To align with the requirements for an excellent project (B-A), we extended the scope of the paper's original experiments in several meaningful and original ways. First, we evaluated the method using a CIFAR-10 diffusion model, allowing us to discuss differences in data quality and performance between the CIFAR-10 dataset and the FFHQ dataset used in the original work. Secondly, we conducted a novel experiment: image colorization. We applied this task to both the FFHQ diffusion model and the CIFAR-10 model, producing a completely new line of investigation not considered in the original paper. Additionally we also looked at the effect of varying the step size for the colorization experiment, essentially going beyond the scope of our proposal. These extensions demonstrate our ability to innovate and explore the method's applicability beyond its original scope.

By successfully reproducing key results, performing a critical analysis, testing alternative hyperparameters, and conducting novel experiments, we believe our work fulfills the requirements for an excellent project (A-grade). Our efforts showcase a strong understanding of the method, experimentation, and important extensions that contribute to a broader understanding of the method's capabilities.

References

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022a.

Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022b.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

Zahra Kadkhodaie and Eero P Simoncelli. Solving linear inverse problems using the prior implicit in a denoiser. *arXiv preprint arXiv:2007.13640*, 2020.

Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021.

Phong Tran, Anh Tran, Quynh Phung, and Minh Hoai. Explore image deblurring via blur kernel space. *arXiv preprint arXiv:2104.00317*, 2021.