Assessing Image Restoration, Meaningfulness and Diversity in Guided Diffusion Models

M. Parsa Dini [1] **Human Jafari** [1] **Arash H.Nezhad** [2] **Sajjad Amini** [1]
mp.dini@sharif.edu, human.jafari@sharif.edu,
arash.hajian@teh.edu, s_amini@sharif.edu

Abstract

In this paper we explore the diversity and semantic understanding of guided diffusion models. Up until now, most existing approaches in diffusion models restore degraded images by leveraging the posterior distribution of the restored image given the degraded input. We extend this idea by evaluating guided diffusion models across multiple datasets and employing advanced masking techniques. Our goal is to investigate how these models comprehend visual concepts and generate meaningful completions for masked regions.

1. Introduction

Diffusion models have seen significant advancements in recent years,particularly in generative AI fields, where they excel in various restoration tasks in images an beyond, especially in image restoration tasks. However, similar to generative adversarial networks, diffusion models networks tend to adhere closely to the training dataset when reconstructing masked images, leading to limited diversity in generated samples. This limitation arises because the posterior distribution is heavy-tailed, meaning there exist meaningful but low-probability modes. To capture such diverse outputs, multiple samples need to be drawn from the model, which is computationally inefficient.

1.1. Literature/Formulation

Given an input image y that is a degraded version of some high quality image x, our goal is to compose a set of N outputs $\chi = \{x_1, \ldots, x_N\}$ such that each x_i constitutes a plausible reconstruction of x, while χ as a whole reflects the diversity of possible reconstructions in a meaningful manner. By 'meaningful' we mean that rather than adhering to the posterior distribution of x given y (note that $\mathbb{E}(x|y)$ is the optimal MSE estimator), we want χ to cover the perceptual range of plausible reconstructions of x, to the maximal extent possible (depending on N). In practical applications, we would want N to be small (e.g., 5) to avoid the need of tedious scrolling through many restorations.

2. Methodology

2.1. Meaningful Diversity in Image Restoration

Our goal in this section is to mathematically characterize a meaningfully diverse set of solutions for image restoration. Instead of focusing on a specific implementation, we explore fundamental principles that guide the selection of diverse samples. Given a degraded input image y, we generate a large set of **candidate reconstructions** $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{\tilde{N}}\}$ using a diverse image restoration method. We then extract perceptually meaningful features from all samples and use feature-space distances to define different diversity selection strategies.

We consider three approaches:

1. **Cluster Representatives:** We apply K-means clustering to the feature space and select the closest sample to each cluster center, ensuring mode coverage but risking redundancy in high-density regions.

2. **Uniformization:** We reweight samples to increase the likelihood of selecting less probable reconstructions, ensuring broader coverage of the posterior distribution while potentially under-representing dominant modes.

3. **Distant Representatives (Farthest Point Strategy):** We iteratively select the most dissimilar samples in feature space to maximize semantic diversity, ensuring non-redundant and well-separated solutions.

In the vanilla guided diffusion, the guidance term $\eta \log p_\theta(x^i_{t-1}|y)$ adjusts the trajectory of the sample $x^i_{0|t}$ towards more plausible reconstructions conditioned on the observed input y. This ensures that each generated sample reflects the true underlying structure while maintaining diversity across multiple reconstructions.

$$x_{0|t} \leftarrow x_{0|t-1} + \eta \log p_\theta(x_{t-1}|y)$$

Since generating an extensive set $\tilde{X}$ is computationally inefficient, we incorporate these diversity principles directly into the sampling process of diffusion models. Instead of independently sampling multiple reconstructions, we simultaneously generate N diverse outputs, all conditioned on the same input y, but driven by distinct noise realizations. To enforce diversity, we introduce a guidance mechanism that pushes samples apart during denoising. At each timestep t, for each sample $x^i_{0|t}$, we compute its closest neighbor in the batch, $x^{i,NN}_{0|t}$, and adjust it using:

$$x^i_{0|t} \leftarrow x^i_{0|t} + \eta \frac{t}{T}(x^i_{0|t} - x^{i,NN}_{0|t}) I\{\|x^i_{0|t} - x^{i,NN}_{0|t}\| < SD\}$$

where η controls the step size, T is the total number of diffusion steps, and SD sets a minimum threshold for guidance. This approach ensures that generated solutions are semantically diverse while maintaining visual plausibility.

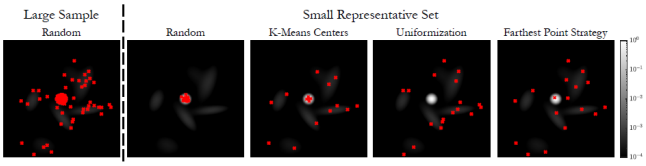

Figure 1. Comparison of three methods for selecting a meaningful subset X of 20 points from an imbalanced set $\tilde{X}$ of 1000 red points drawn from 10 Gaussians. The approaches differ in their ability to cover sparse and dense regions, with the central Gaussian dominating $\tilde{X}$, containing 95% of the probability mass

2.2. Our Approach to Generating Diverse Image Restorations

As mentioned earlier in our proposal, we aimed to achieve two objectives: (1) using masks to evaluate the model's ability to fill in missing regions of images, and (2) assessing whether the model truly comprehends and learns visual concepts.

To achieve this, we masked the most semantically important parts of an image. For example, in Figure 1, we masked the face of a cat, while in Figure 2, we masked specific facial features such as lips or eyebrows in human faces. Additionally, we experimented with masking half of a bedroom image to analyze how the model reconstructs the missing part and whether the completion exhibits symmetry.

We also tuned the guidance parameters (η, D) to investigate the diversity of the model's outputs. By varying η and monitoring the generated samples as it increased, we observed that, as discussed in the original paper, η is directly related to diversity. When η is too large, the samples become unstable, a phenomenon that holds true across different datasets and masking strategies.

Furthermore, we tested the model's generalization by providing it with images it had never encountered during training. For instance, we fed a statue image into the bedroom model (Figure 3) and observed that the model attempted to reconstruct the missing parts by drawing from similar concepts it had learned previously. This suggests that the model relies on learned patterns rather than simply copying seen examples.

3. Experiments & Results

3.1. (i)

The tables from the paper shows that increasing η enhances diversity in the sampled set, but excessively high values may cause saturation effects. These effects can be mitigated by adjusting D, which enables larger changes through guidance. As D increases, diversity also increases. The impact of setting a minimal distance D is evident in Fig. **??**, where its absence leads to the full influence of η. Setting D helps control the effect on some samples while maintaining a high η to separate similar samples. Both η and D must be balanced, as overly small values reduce diversity. And the tables are depicted down below:

Table 1. Effect of η on results for CelebAMash-HQ in image inpainting. Here D is fixed at 0.0003.

η	LPIPS Div. (↑)	NIQE (↓)
0 (Posterior)	0.090	5.637
0.07	0.105	5.495
0.1	0.109	5.506
0.3	0.113	5.519

Table 2. Effect of D on results for CelebAMash-HQ in image inpainting. Here η is fixed at 0.09.

D	LPIPS Div. (↑)	NIQE (↓)
0	0.090	5.637
0.0002	0.094	5.552
0.0003	0.108	5.472
0.0004	0.126	5.554
∞	0.141	5.450

We also examined our generated metrics for fixing D and sweeping η, and the results are as follows:

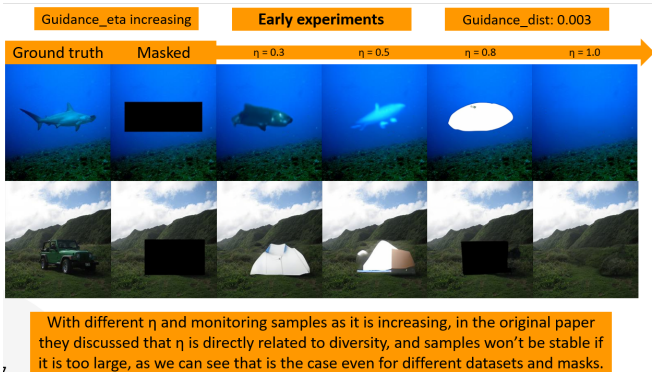

Figure 2. The results of generated images by the Guided Diffusion Model trained on IMAGENET, with $D = 0.003$ and varying η from 0.3 to 1.0. As η approaches zero, the image quality and relevance improve. However, as η increases, the model deviates further from the vanilla guided diffusion, resulting in greater diversity. This highlights the trade-off in setting the guidance parameter.

4. Conclusion

Software and Data

We used the MeaningfulDiversityInIR GitHub repository for their methods of sampling diversely and also the OpenAI guided diffusion models GitHub page for the trained guided diffusion models such as the guided diffusion on IMAGENET, LSUN Cat, LSUN Bedroom, and CelebA-HQ.

Acknowledgements

Many thanks to our mentor, Dr. Amini, for his invaluable guidance and support throughout this project. We also appreciate the insightful feedback and suggestions from our mentor, which significantly contributed to the improvement of this work. We also appreciate OpenAI and the team behind the *MeaningfulDiversityInIR* paper for their contributions to this field.

References

1. Noa Cohen, Hila Manor, Yuval Bahat, Tomer Michaeli, "From Posterior Sampling to Meaningful Diversity in Image Restoration," ICLR 2024, arXiv:2310.16047.

2. Yuyang Hu, Mauricio Delbracio, Peyman Milanfar, Ulugbek S. Kamilov, "A Restoration Network as an Implicit Prior," ICLR 2024, arXiv:2310.01391.

3. Bahjat Kawar, Michael Elad, Stefano Ermon, Jiaming Song, "Denoising Diffusion Restoration Models," NeurIPS 2022, arXiv:2201.11793.

4. Yuanzhi Zhu, "Denoising Diffusion Models for Plug-and-Play Image Restoration," GitHub, 2023, https://github.com/yuanzhi-zhu/DiffPIR.