# Multi-Step Guided Diffusion for Image Restoration on Edge Devices: Toward Lightweight Perception in Embodied AI

Aditya Chakravarty
Independent Research
San Francisco, CA
chakravarty.aditya28@gmail.com

## 1. Introduction

Diffusion models have emerged as powerful tools for solving inverse problems without task-specific retraining. Methods like Diffusion Posterior Sampling (DPS) [4] and FreeDoM [17] guide the generative process using externally defined objectives, enabling flexible and modular inference. Related multi-call approaches [1, 7, 11] extend these ideas through iterative guidance, but often require retraining, complex scheduling, or domain-specific tuning.

Manifold Preserving Guided Diffusion (MPGD) [6] was recently proposed as a training-free alternative that constrains guidance to the tangent space of a learned image manifold, improving stability and realism during restoration. However, MPGD and similar methods typically apply only a single gradient update per denoising step, leaving the potential of deeper, multi-step conditioning unexplored. Moreover, MPGD was developed and evaluated primarily on face-centric datasets, raising questions about its robustness on more diverse or out-of-distribution content.

In this work, we revisit MPGD through the lens of multi-step optimization: applying several gradient descent updates within each denoising timestep. Inspired by prior observations in RePaint [1] and LGD [11] that repeated updates can improve fidelity, we conduct an empirical study to probe the trade-offs between quality, diversity, and inference cost. We find that increasing the number of guidance steps significantly boosts both perceptual quality (LPIPS) and pixel-level accuracy (PSNR), while also enhancing robustness to degraded or out-of-distribution inputs. Notably, we show that MPGD—despite being trained on face datasets—can effectively restore generic, non-face natural images through multi-step conditioning.

Our experiments focus on two canonical inverse problems: $4\times$ super-resolution and Gaussian deblurring. We evaluate across a range of optimization depths using a Jetson Orin Nano, a compact edge GPU platform aligned with the practical constraints of embodied AI systems. Results on both ImageNet and UAV123 aerial imagery demonstrate that MPGD with multi-step optimization is a viable, lightweight solution for real-time visual restoration in robotics and mobile AI agents operating in unconstrained environments.

## 2. Method and Experimental Setup

We consider $4\times$ super-resolution (bicubic downsampling) and Gaussian deblurring (kernel size 61, intensity 3.0), both with additive noise $\sigma = 0.05$, following [6, 16]. We use the pixel-space MPGD implementation with DDIM sampling [10] and a pretrained FFHQ autoencoder [12]. We generate outputs for 1000 ImageNet [5] test images, sweeping steps $\in \{1, 3, 7, 15, 20\}$ , timesteps $\in \{20, 50, 100\}$, and guidance scales $\in \{4, 7.5, 17.5\}$. For evaluation we consider LPIPS [18], SSIM [13], inference time and PSNR. All experiments run on a single NVIDIA Jetson Orin Nano (8 GB VRAM).

**Application Case: Aerial Inspection from UAV123** To assess real-world viability in an embodied AI context, we evaluate MPGD on degraded aerial footage from the UAV123 dataset [9]. This dataset comprises UAV video sequences over buildings, roads, and industrial sites, which are representative of visual inspection scenarios.

We extract 300 diverse frames spanning multiple scenes. These are processed using the same MPGD configuration as the main ImageNet benchmarks, without further training or adaptation. Outputs are compared across LPIPS, PSNR, SSIM, and inference time.

## 3. Results and Discussion

Across both tasks, we observe that performance improves with more optimization steps and moderate guidance scales, saturating around 15 steps. The reconstructions (Figure 1 ) evolve from generic face-like shapes at 1 step to well-structured outputs at 15 steps, even for out-of-distribution images. Inference latency ranges from 50–100ms per image

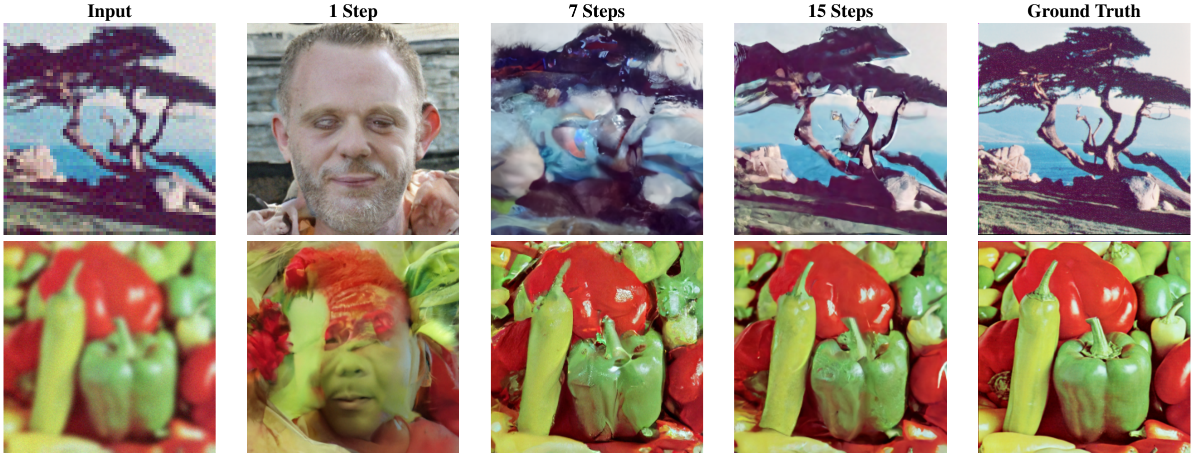


Figure 1. Comparison of SR and Deblur results at 1, 7, and 15 steps.

on Jetson Orin Nano (Table 1), validating MPGD's applicability for real-time embedded perception.

**Performance on UAV123 dataset** On UAV123 images, MPGD achieves strong perceptual and pixel-level performance, despite scene variability and noise. Table 2 shows MPGD (15 steps) surpasses NAFNet[2] and Uformer[14] in LPIPS and PSNR while maintaining real-time throughput on Jetson Orin Nano. This suggests MPGD's utility as a lightweight plug-in module for real-time image enhancement in aerial infrastructure inspection.

| Method | LPIPS ↓ | SSIM ↑ | PSNR ↑ | Time (ms) ↓ |
|---|---|---|---|---|
| MPGD (15 steps) | 0.32 | 0.90 | 20.91 | 80 |
| NAFNet [2] | 0.36 | 0.86 | 20.13 | 35 |
| Uformer [14] | 0.34 | 0.87 | 19.65 | 58 |

Table 1. Comparison of MPGD with baseline models (ImageNet).

| Method | LPIPS ↓ | SSIM ↑ | PSNR ↑ | Time (ms) ↓ |
|---|---|---|---|---|
| MPGD (15 steps) | 0.35 | 0.88 | 21.20 | 90 |
| NAFNet | 0.38 | 0.84 | 20.10 | 42 |
| Uformer | 0.37 | 0.85 | 19.90 | 65 |

Table 2. Performance on degraded UAV123 frames (300 samples).

Across both tasks, we observe that performance improves with more optimization steps and moderate guidance scales, saturating around 15 steps. Qualitatively, reconstructions evolve from generic face-like shapes at 1 step to well-structured outputs at 15 steps, even for out-of-distribution images. Inference latency ranges from 50–100ms per image, validating MPGD's applicability for real-time embedded perception.

## Conclusions and Ongoing/Future Work

We presented a multi-step Manifold Preserving Guided Diffusion (MPGD) approach for training-free image restoration, targeting deployment on edge devices like the Jetson Orin Nano. Our experiments on both standard benchmarks and the UAV123 dataset demonstrate MPGD's utility as a lightweight, retraining-free vision module for embodied AI agents, such as drones and mobile robots, which operate under stringent power and compute constraints. Crucially, we show that multi-step optimization allows a model trained on face-centric data to generalize surprisingly well to natural, non-face images—suggesting that task-driven optimization can compensate for mismatched training domains in practice.

This work is part of an ongoing effort to extend MPGD to a broader class of embodied perception challenges, including low-light navigation, infrastructure inspection, and visual localization under domain shift. Recent advances in diffusion-based decision-making [3, 8] suggest that generative priors can support robust visual modules across diverse environments. Building on this, we plan to explore adaptive optimization depth during inference [11], as well as lightweight test-time adaptation strategies tailored to resource-constrained edge deployment [15]. We are also investigating extensions of MPGD for multi-modal and nonlinear inverse problems, including conditioning on spatial prompts or language cues. These directions aim to position MPGD as a deployable and adaptable backbone for real-time perception in embodied AI systems.

## References

[1] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Liangyu Chen, Xiaojie Chu, and Xiangyu Zhang. Simple baselines for image restoration. In *European Conference on Computer Vision (ECCV)*, pages 186–202, 2022.

[3] Linxi Chi, Zichen Huang, Tao Yu, Ziyu Ma, Ankur

Handa, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.

[4] Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L.Klasky, and JongChul Ye. Diffusion posterior sampling for general noisy inverse problems. In *Proc. International Conference on Learning Representations (ICLR)*, 2023.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[6] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *Proc. International Conference on Learning Representations (ICLR)*, 2024. poster.

[7] Jiachun Pan, Hanshu Yan, Jun Hao Liew, Jiashi Feng, and Vincent Y.F. Tan. Towards accurate guided diffusion sampling through symplectic adjoint method. *arXiv preprint arXiv:2312.12030*, 2023.

[8] Jacky Liang, Zirui Yuan, Qiyang Wu, et al. Generative pretraining for decision making and control. *arXiv preprint arXiv:2210.03029*, 2023.

[9] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016.

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Yang Song, Chenlin Meng, and Stefano Ermon. Loss-guided diffusion: Learning to denoise images conditioned on a loss function. In *International Conference on Machine Learning (ICML)*, 2023.

[12] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17968–17977, 2021.

[13] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[14] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Guihua Chen, Jin Gu, Jianzhuang Liu, and Chen Dong. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17683–17693, 2021.

[15] Zhen Wang, Qiang Wang, Kai Zhang, Mengyuan Xu, and Yiran Wang. Tta-lite: Memory-efficient test-time adaptation for deep models on the edge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[16] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

[17] Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[18] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.