Masked-ResShift: Pixel-Level Residual Shift for Image Inpainting

Rishikesh Bhyri*
rbhyri@buffalo.edu

Yash Rathi*
yrathi2@buffalo.edu

Abstract

The field of image restoration, specifically image inpainting has seen recent remarkable advancements with the development of generative models. Among these, diffusion models have emerged as a powerful class of techniques capable of producing quality image completions by progressively denoising an image noisy low resolution image. This process often involves numerous iterative steps, leading to significant computational overhead and hence slow inference times. Such limitations can hinder the practical application of these models in scenarios requiring rapid image processing.

This project explores and tries to enhance the efficiency of diffusion-based image inpainting. We explore various optimization strategies aimed at accelerating the inference process without substantially compromising the quality of the result images. Our efforts build upon the existing frameworks, such as ResShift known for its efficiency in image restoration tasks, and seek to refine these approaches further. This project implements two distinct strategies, including Masked-Variance Diffusion (MVD), where diffusion noise is primarily applied to unknown image regions, and Exact Masked-Markov Diffusion (EMMD), which defines a true pixel-wise Markov forward process with an exact posterior to guide the reverse diffusion. We examine and implement these techniques, from modifying network architectures to rethinking the diffusion process itself, all with the final goal of achieving faster, yet still effective, image inpainting. Our code and model are available at https://github.com/rishi1134/masked-resshift.

1 Background

Diffusion models are at the core of so many groundbreaking capabilities that we see in image generation and restoration. At their core, these models operate on a simple principle: learning to reverse a noise process. Starting from this noisy state, it learns to carefully denoise the data, step by step, ultimately reconstructing a high-quality, coherent image. This iterative refinement is what allows diffusion models to generate such detailed and realistic outputs, making them particularly adept for complex tasks like image inpainting, where the model must fill missing regions with the surrounding content.

Despite their power, a well-documented challenge with many diffusion models is their computational demand, primarily due to the large number of sampling steps typically required for inference. This can lead to slow generation times, which is a hurdle for real-time or interactive applications. This aim for efficiency, without a significant trade-off in output quality, has given way to research into more optimized diffusion frameworks.

Recent work in this area includes [3], which proposes a framework to optimize generation time by treating image tokens as active or inactive regions based on changes in their content. This approach

*These authors contributed equally to this work.

processes only the active regions for denoising, while caching the inactive ones for faster processing, demonstrating efficient redundancy reduction. Another such work is ResShift [1], designed primarily for image super-resolution (SR). Instead of the conventional approach of diffusing an image to pure noise, ResShift constructs a more direct and efficient Markov chain. This chain facilitates a transition between a high-resolution (HR) image and its low-resolution (LR) counterpart by progressively shifting the residual (the difference, $e_0 = y_0 - x_0$) between them. This elegant modification substantially shortens the diffusion path, enabling ResShift to achieve impressive results in as few as 4 sampling steps.

Figure 1: ResShift Generation Process.

While ResShift marked a significant step forward in efficient image restoration, we recognized that its principles could be further adapted and potentially optimized for the specific challenges of image inpainting. The original ResShift framework, though efficient, treats the entire image (or its latent representation) fairly uniformly throughout its diffusion steps. We hypothesized that for inpainting, where there's a clear distinction between known, fixed regions and unknown regions to be filled, a more targeted approach might yield benefits. Could we reduce redundant processing in known, perfect regions? Furthermore, could the diffusion process be refined for mask awareness, perhaps by adapting noise injection or variance scheduling in these areas, similar to [2]? These considerations paved the way for our exploration into specialized strategies like Masked-Variance Diffusion (MVD) and Exact Masked-Markov Diffusion (EMMD), which we will detail in the subsequent Method section.

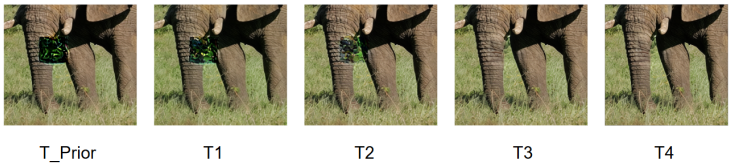

Figure 2: MVD Generation Process.

1.1 Dataset

To check our explorations and benchmark our methodologies, we utilized the Synthetic CelebA-2k dataset for training and evaluation purposes within this project. This dataset is a derivative of the well-known Large-scale CelebFaces Attributes (CelebA) dataset, which is a prominent public benchmark in the computer vision community. The full CelebA dataset contains over 200,000 images of celebrity faces, each extensively annotated with 40 binary facial attributes, such as hair color, gender, and expression, and includes information like landmark locations across approximately 10,177 unique identities. Given its scale, diversity in pose and background, and rich annotations, CelebA is widely employed for a variety of tasks including face attribute recognition, face detection, landmark localization, and, importantly for our work, face editing and image inpainting.

2 Methodology

2.1 Masked-Variance Diffusion (MVD)

The fundamental concept behind MVD is that we naively injected diffusion noise (and adjusted variance) only in the unknown region. For the image only, the pixels within the designated missing areas are subjected to the iterative noising and denoising characteristic of diffusion models. The

Figure 3: Masked Dataset Images.

known, unmasked pixels are preserved throughout this process. This targeted approach aims to concentrate the model's learning only on the regions that need to be generated.

During the forward (noising) process in MVD, corruption is specifically directed at these unknown areas. The noise schedule itself can be guided by the input mask, ensuring that the diffusion applies exclusively to the regions we intend to inpaint. This allows the model to focus its learning capacity on the intricacies of filling these gaps.

Consequently, during the inference process, noise is added only to these masked patches, and the model works to reconstruct the content within them. The pixels from the original image that were known are carried forward and re-injected at each step, ensuring they remain untouched in the final output. See the supplementary section for a detailed derivation of the forward and reverse processes.

A core assumption under this particular approach was that the mean of the global Gaussian would be similar to the mean of pixel-wise Gaussians. This simplification was made to facilitate the adaptation of standard diffusion posteriors to this masked context. While this method presents a clear path to reducing computational load by not processing the entirety of the image data at each step of the diffusion, its impact on image quality, particularly at the boundaries of the masked regions, was a key aspect of our subsequent evaluation.

2.1.1 Results

For Training MSE and LPIPS (lower is better), our MVD approach generally showed better values compared to the original ResShift baseline over the training steps. This suggested that the texture generated by our model for the inpainted regions was potentially better.

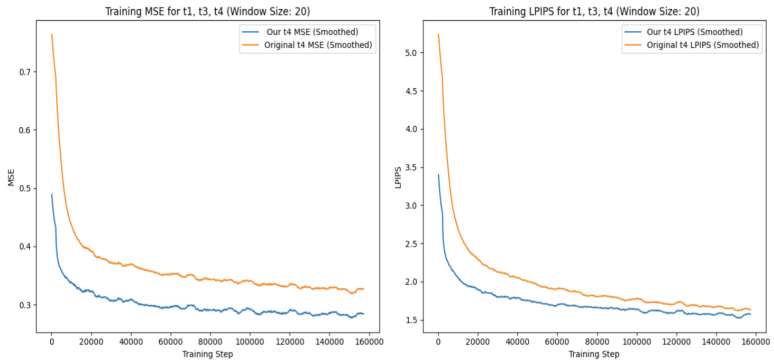

Figure 4: MVD vs Original ResShift Training Graphs

For Validation PSNR (higher is better) and Validation LPIPS (lower is better), the results were more nuanced. While there were fluctuations, our MVD approach demonstrated competitive performance. The texture of our image is better as compared to the ResShift images hence our MSE and LPIPS values are low even though they have seams.

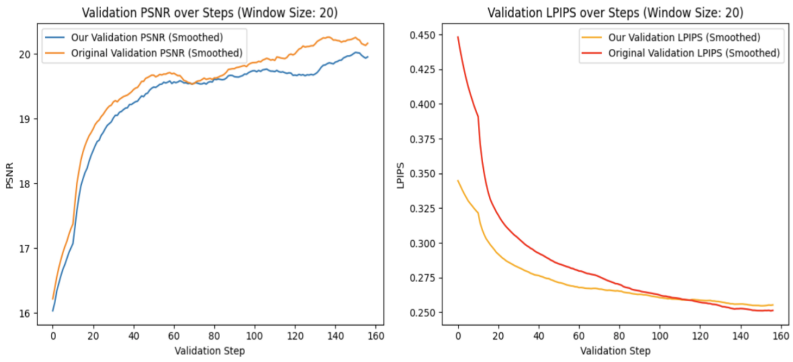

Figure 5: MVD vs Original ResShift Validation Graphs

2.1.2 Drawbacks

Mask-edge artifacts: *Visible seams and noise at the boundary of unknown patches.* By changing the forward noising to only inject noise inside the hole, we broke the Gaussian-Markov structure. The Global Gaussian Mean = Local Pixel Wise Gaussian Mean assumption does not hold well and the closed-form standard posterior it uses is wrong at the boundary. This mismatch shows up as sharp seams where the "no-noise" region meets the "noisy" region.

Figure 6: Comparison of image inpainting results between ResShift (top row) and our MVD approach (bottom row).

2.2 Exact Masked–Markov Diffusion (EMMD)

We define a true pixel-wise Markov forward (identity/delta on known pixels, Gaussian on unknown pixels) and derive its exact posterior mean/variance to restore a valid reverse diffusion. This means that for the known pixels in the image, the forward process is essentially an identity transformation – they remain unchanged, represented by a delta function. For the unknown pixels, a standard pixel-wise Gaussian diffusion process is applied, gradually noising them.

Hence by defining this explicit pixel-wise Markov forward process, we could then derive the mathematically exact posterior mean and variance. This exact posterior is then used in the reverse diffusion process, aiming to provide a more valid and stable mechanism for denoising the unknown regions while perfectly preserving the known ones. The motivation behind EMMD was to create a reverse diffusion process that is precisely consistent with the defined masked forward process, thereby hoping

to mitigate the boundary artifacts and improve the overall coherence of the inpainted image. See the supplementary section for a detailed derivation of the forward and reverse processes.

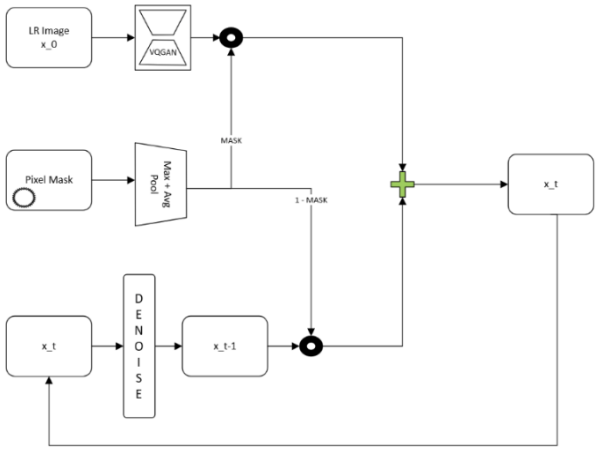

Figure 7: Pixel-Wise Conditional Generation Process.

2.2.1 Results

The quantitative comparison of EMMD, our MVD approach, and the ResShift baseline yielded the following trends:

Training MSE and LPIPS: The graphs for training metrics (Mean Squared Error and LPIPS, where lower is generally better) showed that EMMD often performed worse than MVD. In some cases, EMMD's error metrics were higher than or comparable to the original ResShift, suggesting that the theoretically exact posterior did not straightforwardly translate into improved training dynamics for these metrics.

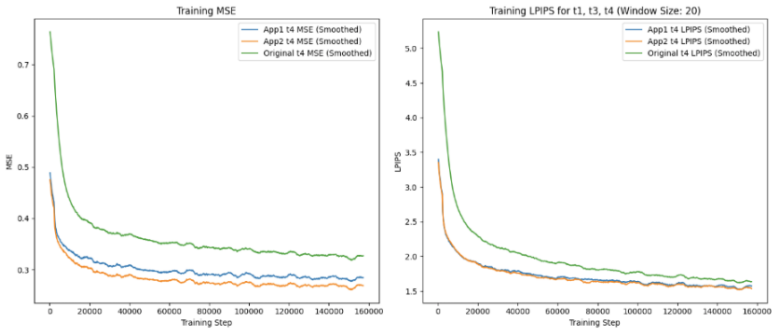

Figure 8: **EMMD Training Graphs**: Comparison between EMMD (App2), MVD(App1) & Original ResShift

Validation PSNR and LPIPS: Similarly, for validation (where PSNR is higher is better, and LPIPS is lower is better), EMMD did not consistently outperform MVD and, in several instances, showed metrics that were less favorable than even the original baseline. The presentation indicates that, overall, the generation quality with EMMD was a concern.

2.2.2 Drawbacks

The artifacts persist, and the quality of generation has deteriorated. This is mainly due to the following reasons:

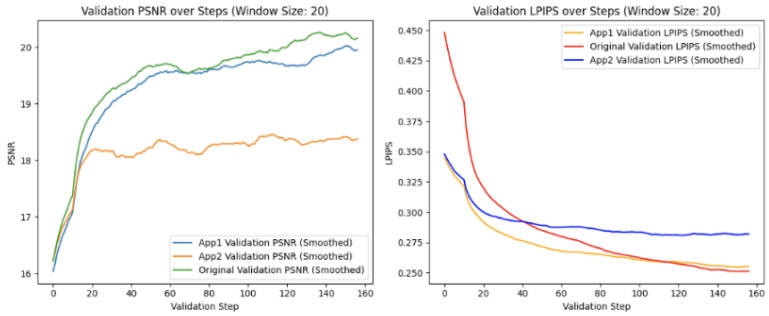

Figure 9: **EMMD Validation Graphs**: Comparison between EMMD (App2), MVD(App1) & Original ResShift

1. Extreme variance discontinuity: We go from "zero variance" outside the mask to "full DDPM variance" inside. That $0 \to \beta_t$ jump is a brutally sharp step-function in the latent-space distributions. Any tiny prediction error at the seam leaves a visible artifact.

2. Split training signal: During training, known regions sees an identity mapping (noising$= 0 \to$ denoising=identity) while the unknown sees a full multi-step diffusion. The U-Net must learn two completely different tasks in parallel, with no shared statistics at the boundary.

Figure 10: Comparison of image inpainting results between ResShift (top row) and our EMMD approach (bottom row).

2.3 Addressing the Drawbacks of EMMD

To tackle the limitations observed with EMMD, such as extreme variance discontinuity and a split training signal, we propose two primary refinements for future work:

Mitigating Variance Discontinuity: We propose smoothing the transition masks between known and unknown regions and potentially retraining with a minimal noise floor in known areas. This aimed to reduce the abrupt variance shift and lessen seam artifacts, with initial explorations into mask smoothing showing some promise.

Simplifying the Learning Task with DiTs: To resolve the split training signal, following a RAS [3] based approach to process only the unknown regions would allow the neural network to focus solely on the denoising task for the inpainted area, rather than simultaneously learning an identity mapping for known regions, thereby simplifying its objective.

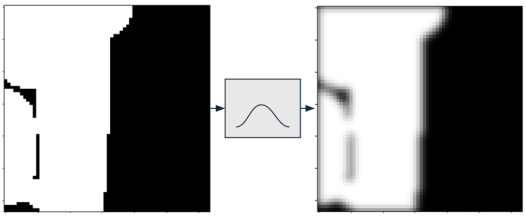

Figure 11: **Mask Smoothening**: Smoothen the masks with a guassian filter such that change between region is not abrupt.

Initial experiments using a Gaussian filter for smooth mask transitions yielded promising results (Figure 11). Specifically, we can further refine mask handling by smoothing transitions between regions and retraining with a minimal noise floor in the known areas to ensure variance remains non-zero.

Figure 12: Comparison of image inpainting results between ResShift (top row) and our Smoothened EMMD approach (bottom row).

3 Training Details

We maintained the ResShift training configuration, with two key changes: the Swin window size was reduced from 8 (as used in ResShift) to 4, and the training duration was set to 200k iterations.

Figure 13: Training parameters

Configuration	Setting
Optimizer	AdamW
Optimizer betas	{0.9, 0.999}
EMA Rate	0.99
Base learning rate	5e-5
LR Scehdule	Cosine
Warmup steps	5000
Training steps	2e5
Loss Coefficients	MSE: 1, LPIPS: 10
Image size	256×256
Latent size	64×64
Batch size	4
Diffusion steps	4
Noise schedule	Exponential
Kappa	2.0
U-Net Attention Resolutions	[64, 30, 16, 8]

4 Conclusion

We tackled efficient diffusion-based image inpainting, aiming for faster inference without significant quality loss. Our investigation introduced two main approaches: Masked-Variance Diffusion (MVD) and Exact Masked-Markov Diffusion (EMMD), based on refining efficient frameworks like ResShift. MVD, applying noise and variance adjustments only to unknown regions, showed training promise but suffered from mask-edge artifacts due to disrupted Markov structure. EMMD, designed with a theoretically sound Markov forward process, also exhibited artifacts and quality issues, primarily from variance discontinuities and a split training signal. To mitigate these, we proposed future work including smoothed mask transitions, a minimal noise floor in known regions, and using Diffusion Transformers (DiTs) to focus processing on unknown areas.

References

[1] Zongcai Yue & Jian Wang & Chen Change Loy (2024) ResShift: Efficient Diffusion Model for Image Super-resolution by Residual Shifting. *Thirty-eighth Conference on Neural Information Processing Systems*

[2] Thibault Mayet & Peyman Shamsolmoali & Sebastien Bernard & Eric Granger & Romain Hérault & Charles Chatelain (2024) TD-Paint: Faster Diffusion Inpainting Through Time Aware Pixel Conditioning *arXiv: 2410. 09306*

[3] Zheng Liu et al (2025) Region-Adaptive Sampling for Diffusion Transformers *arXiv: 2502. 10389*

Table 1: Author Contributions

Author	**Contribution**
Rishikesh Bhyri	Contributed 50% to the project, including idea development, implementation, experimentation, and writing.
Yash Rathi	Contributed 50% to the project, including idea development, implementation, experimentation, and writing.

Appendix / supplemental material

1. ResShift Forward and Reverse Process

Let,

$$x_0 \in R^{C\times H\times W} \quad \text{the clean latent,}$$
$$y \in R^{C\times H\times W} \quad \text{the degraded (known) latent,}$$
$$\mathbf{M} \in \{0,1\}^{1\times H\times W} \quad \text{the binary mask,}$$

The Forward Process is defined as below where $e_0 = y_0 - x_0$ is the residual

$$q(x_t|x_0, y_0) = \mathcal{N}(x_t; x_0 + \eta_t e_0, \kappa^2\eta_t I), \quad t = 1, 2, \cdots, T. \tag{1}$$

and the closed form Reverse process is defined as,

$$q(x_{t-1}|x_t, x_0, y_0) = \mathcal{N}\left(x_{t-1}; \frac{\eta_{t-1}}{\eta_t}x_t + \frac{\alpha_t\eta_{t-1}}{\eta_t}x_0, \frac{\kappa^2\eta_{t-1}}{\eta_t}\alpha_t I\right) \tag{2}$$

So the tractable approximation using a U-Net is defined as,

$$\mu_\theta(x_t, y_0, t) = \frac{\eta_{t-1}}{\eta_t}x_t + \frac{\alpha_t\eta_{t-1}}{\eta_t}f_\theta(x_t, y_0, t) \tag{3}$$

2. Our Approach

I. Modified Forward Process

We inject both the guidance drift and the Gaussian noise *only* inside the masked region. So modifying the (5) to define the gaussian at pixel level,

For known region,

$$q(x_t^{kn}|x_0, y_0) = \delta(x_t^{kn} - x_0^{kn}) \tag{4}$$

and for unknown region,

$$q(x_t^{unk}|x_0, y_0) = \mathcal{N}(x_t^{unk}; x_0^{unk} + \eta_t e_0^{unk}, \kappa^2\eta_t I), \quad t = 1, 2, \cdots, T. \tag{5}$$

Equivalently, writing

$$\boxed{x_t = x_0 \;+\; \mathbf{M} \;\odot\; \Big[\eta_t\, e_0 \;+\; \underbrace{\sqrt{\eta_t}\,\kappa}_{\sigma_t}\;\epsilon\Big], \quad \epsilon \sim \mathcal{N}(0, I)\,.}$$

II. Modified Reverse Process

a. Posterior in the Known Region

For each pixel i with $\mathbf{M}(i) = 0$, the forward step is

$$x_t(i) = x_{t-1}(i) \quad \text{(deterministic),}$$

so the exact reverse is the Dirac delta

$$q\big(x_{t-1}(i) \mid x_t(i)\big) = \delta\big(x_{t-1}(i) - x_t(i)\big),$$

i.e. in the known region

$$\mu_t^{\text{known}}(i) = x_t(i), \quad \sigma_t^{2,\text{known}} = 0.$$

b. Posterior in the Unknown Region

For each pixel i with $\mathbf{M}(i) = 1$, the closed-form DDPM posterior is

$$q\big(x_{t-1}(i) \mid x_t(i), x_0(i)\big) = \mathcal{N}\big(x_{t-1}(i);\ \mu_t^{\text{unk}}(i),\ \sigma_t^{2,\text{unk}}\big),$$

where

$$\mu_t^{\text{unk}}(i) = \frac{\eta_{t-1}}{\eta_t}\, x_t(i)\ +\ \frac{\alpha_t \eta_{t-1}}{\eta_t}\, x_0(i), \quad \sigma_t^{2,\text{unk}} = \frac{\kappa^2 \eta_{t-1}}{\eta_t} \alpha_t I.$$

c. Combined Custom Reverse Step

We now blend the two cases element-wise via the mask $\mathbf{M}$, broadcasting $\mathbf{M}$ to all channels:

$$\mu_t = \mathbf{M} \odot \mu_t^{\text{unk}}\ +\ (1 - \mathbf{M}) \odot x_t, \qquad \sigma_t^2 = \mathbf{M} \odot \sigma_t^{2,\text{unk}}.$$

Hence the final sampling equation is

$$\boxed{x_{t-1} = \mu_t\ +\ \sqrt{\sigma_t^2}\, \epsilon', \quad \epsilon' \sim \mathcal{N}(0, I).}$$

d. Final Re-injection of Clean Known Region

To guarantee the known region stays perfect, you optionally do

$$x_{t-1}\ \leftarrow\ \mathbf{M} \odot x_{t-1}\ +\ (1 - \mathbf{M}) \odot x_0.$$

Forward (partial-noise) step

$$x_t = \big[x_0\ +\ \eta_t\, e_0\big]\ +\ \mathbf{M} \odot \big[\kappa \sqrt{\eta_t}\, \epsilon\big], \quad \epsilon \sim \mathcal{N}(0, I).$$

Reverse (standard DDPM posterior + masked sampling + re-injection)

$$\mu_t = c_{1,t}\, x_t + c_{2,t}\, x_0, \quad \sigma_t^2 = \text{posterior\_variance}_t,$$

$$x'_{t-1} = \mu_t + \sqrt{\sigma_t^2}\, \epsilon', \quad \epsilon' \sim \mathcal{N}(0, I),$$

$$x_{t-1} = \mathbf{M} \odot x'_{t-1} + (1 - \mathbf{M}) \odot y.$$

$$c_{1,t} = \frac{\sqrt{\alpha_t}\,(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}, \quad c_{2,t} = \frac{\sqrt{\bar{\alpha}_{t-1}}\,\beta_t}{1 - \bar{\alpha}_t}.$$