

Implicit Image-to-Image Schrödinger Bridge for Image Restoration

Yuang Wang^{a,b}, Siyeop Yoon^b, Pengfei Jin^b, Matthew Tivnan^b, Sifan Song^b,
Zhennong Chen^b, Rui Hu^b, Li Zhang^a, Quanzheng Li^b, Zhiqiang Chen^a, Dufan Wu^{b,*}

^a*The Department of Engineering Physics, Tsinghua University, 30 Shuangqing Road,
Haidian, Beijing, 100084, China*

^b*Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard
Medical School, 55 Fruit Street, Boston, MA 02114, Massachusetts, USA*

Abstract

Diffusion-based models have demonstrated remarkable effectiveness in image restoration tasks; however, their iterative denoising process, which starts from Gaussian noise, often leads to slow inference speeds. The Image-to-Image Schrödinger Bridge ($I^2\text{SB}$) offers a promising alternative by initializing the generative process from corrupted images while leveraging training techniques from score-based diffusion models. In this paper, we introduce the Implicit Image-to-Image Schrödinger Bridge ($I^3\text{SB}$) to further accelerate the generative process of $I^2\text{SB}$. $I^3\text{SB}$ restructures the generative process into a non-Markovian framework by incorporating the initial corrupted image at each generative step, effectively preserving and utilizing its information. To enable direct use of pretrained $I^2\text{SB}$ models without additional training, we ensure consistency in marginal distributions. Extensive experiments across many image corruptions—including noise, low resolution, JPEG compression, and sparse sampling—and multiple image modalities—such as natural, human face, and medical images—demonstrate the acceleration benefits of $I^3\text{SB}$. Compared to $I^2\text{SB}$, $I^3\text{SB}$ achieves the same perceptual quality with fewer generative steps, while maintaining or improving fidelity to the ground truth.

Keywords: Image Restoration; Diffusion Model; Schrödinger Bridge

1. Introduction

Restoring high-quality images from degraded ones is a fundamental yet challenging task in both natural and medical imaging. Image corruptions, can arise from various factors, including noise, low resolution, compression artifacts and sparse sampling. Recently, conditional diffusion models [1, 2] have shown promising performance in addressing this challenge. Rooted in stochastic process theories, diffusion models offer a more stable approach to sampling from complex distributions compared to Generative Adversarial Networks (GANs) [3]. However, the inference speed of diffusion models is

*Corresponding author.

Email address: dwu6@mgh.harvard.edu (Dufan Wu)

often limited by the large number of iterative denoising steps needed to generate clean images starting from pure Gaussian noise.

Instead of starting from Gaussian noise, Schrödinger Bridges establish diffusion bridges between the distributions of clean and corrupted images. By initiating the diffusion process with the corrupted image, which is closer to the clean one than Gaussian noise, Schrödinger Bridges offer a promising approach to generate high-quality conditional samples with fewer diffusion steps. A notable instance of this framework is the Image-to-Image Schrödinger Bridge ($I^2\text{SB}$) [4], which models the transition between paired clean and corrupted images, facilitating efficient training through its connection to score-based diffusion models. The number of neural function evaluations (NFEs) in its generative process controls the trade-off between perceptual quality and fidelity to the ground truth [5]. With a small NFE, $I^2\text{SB}$ tends to produce less distortion but smoother images due to the “regression to the mean effect” [6], whereas with a large NFE, $I^2\text{SB}$ can generate images with high perceptual quality at the expense of some distortion from the ground truth.

In this work, our aim is to accelerate $I^2\text{SB}$ to achieve the same perceptual quality and equal or better fidelity to the ground truth with fewer generative steps. Inspired by the success of incorporating non-Markovian processes into the denoising diffusion probabilistic model (DDPM) [7] to create the denoising diffusion implicit model (DDIM) [8], we proposed the Implicit Image-to-Image Schrödinger Bridge ($I^3\text{SB}$). Our approach employs a non-Markovian process during inference by incorporating the initial corrupted image into each step alongside the current image and the estimated mean, effectively preserving and utilizing the information from corrupted images. By constraining the marginal distribution to match that of $I^2\text{SB}$, the proposed $I^3\text{SB}$ shares the same training loss functions with $I^2\text{SB}$ and can thus reuse pretrained $I^2\text{SB}$ models. We introduced a single parameter to balance the Markovian and non-Markovian components. Additionally, we established a connection between $I^3\text{SB}$ and the probability flow ordinary differential equation (PF-ODE) of the Variance Exploded (VE) endpoint-fixed Schrödinger Bridge [9].

We validate $I^3\text{SB}$ in many image restoration tasks on multiple image modalities, including super-resolution and JPEG restoration for natural and human face images, as well as CT sparse-view reconstruction, super-resolution, and denoising for medical images. Extensive experiments demonstrate that, compared to $I^2\text{SB}$, $I^3\text{SB}$ achieves similar perceptual quality with fewer NFEs while maintaining equal or superior fidelity to the ground truth. Additionally, $I^3\text{SB}$ outperforms several state-of-the-art diffusion-based image restoration models in both quantitative metrics and visual quality.

The main contributions of our work can be summarized as follows:

- We introduce $I^3\text{SB}$, an innovative framework that modifies the generative process of $I^2\text{SB}$ from a Markovian to a non-Markovian approach, incorporating the corrupted image at each generative step, effectively reducing information loss.
- We establish a theoretical connection between $I^3\text{SB}$ and the PF-ODE of the VE endpoint-fixed Schrödinger Bridge.
- Through extensive experiments on image corruptions, including noise, low resolution, JPEG compression, and sparse sampling, we demonstrate the accelerated

performance of I³SB. Compared to I²SB, I³SB achieves similar perceptual quality with fewer NFEs while maintaining or improving fidelity to the ground truth.

2. Related Work

2.1. Image Restoration Methods

Image restoration is a critical task in image processing and pattern recognition, focused on recovering high-quality images from degraded inputs, such as those affected by noise, low resolution, compression artifacts and sparse sampling. Traditional methods typically rely on predefined priors and optimization techniques for image reconstruction. Filtering-based approaches, such as Gaussian smoothing and Non-Local Means Filtering [10], work by reducing noise while preserving edges through averaging nearby or similar pixels. Variational methods [11, 12] assume that images have sparse representations in specific transformation domains (e.g., wavelet or Fourier), framing image restoration as an optimization problem that minimizes a loss function containing both a data consistency term and a regularization term, such as total variation regularization. Dictionary learning approaches model images as sparse combinations of atoms from a learned dictionary, using methods like K-SVD [13] to build dictionaries and restore clean images.

In recent years, deep learning has revolutionized image restoration by replacing handcrafted priors with data-driven representations, enabling models to learn complex image patterns directly from data. Convolutional Neural Networks (CNNs) learn end-to-end mappings from corrupted to clean images. While these methods excel in fidelity, they often produce overly smooth images. Vision Transformers (ViTs), like Restormer [14], leverage self-attention mechanisms to capture long-range dependencies in images, leading to significant improvements in detail restoration. Conditional Variational Autoencoders (CVAEs) [15] learn compact latent representations of clean images, providing a probabilistic framework for generating clean images from Gaussian noise conditioned on the corrupted inputs. Normalizing flow-based models model complex data distributions through invertible transformations, serving as priors to regularize the image restoration process [16]. GAN-based models, such as DGD-cGAN [3], use adversarial training to generate visually realistic textures, although they may suffer from unstable training and issues like mode collapse. Recently, score-based diffusion models have emerged as state-of-the-art methods for image restoration [17], and we will provide an overview of these models in the following subsection.

2.2. Diffusion Based Image Restoration Models

Diffusion based image restoration models can be broadly categorized into task-agnostic and task-specific models. Task-agnostic models train unconditional score functions with denoising score matching [18] as priors for the data distribution, and incorporate data consistency during inference. These models can be used for various types of corruptions with the same trained network, but they require the specific forward operator corresponding to each corruption during inference. For instance, DDNM [19] and DDRM [20, 21] decompose images into the range and null spaces

of the forward operator, applying data consistency to the range space component. DPS [22] and ΠGDM [23] sample from the posterior distribution, using approximations to make the process tractable. Red-Diff [24] formulates image restoration as an optimization problem, incorporating the pretrained score function as regularization.

Task-specific models train conditional score functions using the corrupted image as a condition. These models require separate training for each type of corruption, necessitating pairs of clean and corrupted images during training. However, they often outperform task-agnostic models. Moreover, because task-specific models do not need to incorporate additional data consistency during inference, they are capable of handling blind corruption scenarios where the forward operator is unknown. Notable examples of task-specific models include SR3 [2] and ADM [1].

2.3. Acceleration for Diffusion Models

Accelerating the inference process of diffusion models has become a key focus in the field, leading to the development of two main approaches. The first approach involves generating samples by solving the PF-ODE using high-order solvers instead of sampling from stochastic differential equations (SDEs). These methods modify only the inference process without further training the score function. For instance, DDIM [8] accelerates DDPM by transforming the Markovian process into a non-Markovian process and establishes its connection with the PF-ODE. EDM [25] solve the PF-ODE using the Heun’s 2nd order method, achieving state-of-the-art results in image generation. DPM-solver [26] transforms the ODE discretization into a discretization of the exponentially weighted integral of the score function, applying Taylor expansion for approximation. PNDM [27] introduces a manifold-constrained high-order ODE solver, utilizing linear multi-step and Runge-Kutta methods while maintaining intermediate images on the true noisy manifold.

Diffusion distillation is another approach that accelerates the diffusion inference process by leveraging the PF-ODE, which establishes a deterministic mapping between Gaussian noise and the final generated images. These methods train student models to distill the multi-step outputs of the original diffusion model into a single step. For example, Progressive Distillation [28] introduces binary time distillation, where the student model is trained to predict the two-step output of the teacher model and then serves as the teacher in the next phase. Consistency Model [29] uses the student model, equipped with exponentially moving averaged weights, as a self-teacher to incorporate self-consistency into the student model.

2.4. Paired Data Schrödinger Bridge

Unlike diffusion models that start from Gaussian noise, Schrödinger Bridges initiate their generative processes from corrupted images, offering a promising alternative for generating high-quality conditional samples with fewer generative steps. Schrödinger Bridges can be broadly categorized into unpaired and paired data Schrödinger Bridges. While unpaired data Schrödinger Bridges [30] often face challenges with inefficient training, paired data Schrödinger Bridges, such as I²SB [4] and InDI [6], can be trained as efficiently as score-based diffusion models. Techniques developed for diffusion models have been adapted to paired data Schrödinger Bridges. For instance, DDBM [9] applies Heun’s 2nd-order method to solve the PF-ODE of paired

data Schrödinger Bridges and demonstrates good performance in image translation tasks. CDBB [5] enhances I²SB by incorporating data consistency with techniques from DDS [31] and DPS. Our proposed I³SB, inspired by DDIM and transforming the generative process from Markovian to non-Markovian, demonstrates acceleration benefits compared to I²SB.

3. Preliminaries

Notation: Let $X_t \in \mathbb{R}^d$ represent a d -dimensional stochastic process indexed by $t \in [0, 1]$, and N denote the number of generative steps. We denote the discrete generative time steps as $0 = t_0 < \dots < t_n \dots < t_N = 1$, and shorthand $X_n \equiv X_{t_n}$.

3.1. Image-to-Image Schrödinger Bridge

The I²SB [4] establishes direct diffusion bridges between paired clean and corrupted images. With X_0 representing clean images and X_1 representing corresponding corrupted images, X_t is designed to follow the Gaussian distribution $q(X_t|X_0, X_1)$:

$$q(X_t|X_0, X_1) = \mathcal{N}\left(X_t; \frac{\bar{\sigma}_t^2}{\sigma_1^2}X_0 + \frac{\sigma_t^2}{\sigma_1^2}X_1, \frac{\sigma_t^2\bar{\sigma}_t^2}{\sigma_1^2}I\right), \quad (1)$$

where $\sigma_t^2 = \int_0^t \beta_\tau d\tau$ and $\bar{\sigma}_t^2 = \int_t^1 \beta_\tau d\tau$ represent variances accumulated from either side, $\sigma_1^2 = \int_0^1 \beta_\tau d\tau$, and β_τ determines the speed of diffusion. Since X_t can be sampled analytically using equation (1), the network ϵ_θ can be efficiently trained to predict the difference between X_t and X_0 by minimizing the loss function:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{X_0, X_1} \mathbb{E}_{t, X_t} \|\epsilon_\theta(X_t, t) - \frac{X_t - X_0}{\sigma_t}\|, \quad (2)$$

where $t \sim \mathcal{U}[0, 1]$ and $X_t \sim q(X_t|X_0, X_1)$.

In the generative process, I²SB begins with the corrupted image X_N and iteratively approaches the clean image X_0 . In the step from X_n to X_{n-1} , $\hat{X}_0^{(n)}$, the expected mean of X_0 at time t_n , is first calculated using the trained network ϵ_{θ^*} and X_n :

$$\hat{X}_0^{(n)} = X_n - \sigma_n \epsilon_{\theta^*}(X_n, t_n), \quad (3)$$

where we use $\sigma_n \equiv \sigma_{t_n}$. Subsequently, X_{n-1} is sampled from the DDPM posterior p described by $\hat{X}_0^{(n)}$ and X_n :

$$X_{n-1} \sim p(X_{n-1}|\hat{X}_0^{(n)}, X_n). \quad (4)$$

Here, the DDPM posterior p is expressed as:

$$p(X_{n-1}|\hat{X}_0^{(n)}, X_n) = \mathcal{N}\left(X_{n-1}; \frac{\alpha_{n-1}^2}{\sigma_n^2}\hat{X}_0^{(n)} + \frac{\sigma_{n-1}^2}{\sigma_n^2}X_n, \frac{\sigma_{n-1}^2\alpha_{n-1}^2}{\sigma_n^2}I\right), \quad (5)$$

where $\alpha_{n-1}^2 = \int_{t_{n-1}}^{t_n} \beta_\tau d\tau$ denotes the accumulated variance between consecutive time steps t_{n-1} and t_n .

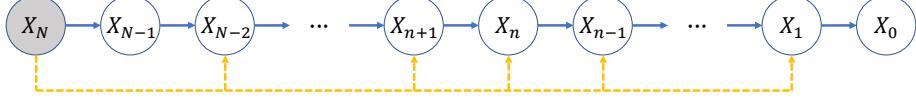


Figure 1: Non-Markovian generative process of $I^3\text{SB}$. Solid arrows denote original dependencies in $I^2\text{SB}$, and dotted arrows signify additional dependencies in $I^3\text{SB}$.

Algorithm 1 Generative Process of $I^3\text{SB}$

Input: $N, \{t_n\}$, corrupted image X_N , trained network ϵ_{θ^*}

```

for  $n = N$  to  $1$  do
    Predict  $\hat{X}_0^{(n)}$  using  $\epsilon_{\theta^*}(X_n, t_n)$  and  $X_n$ 
    if  $n == N$  then
        Sample  $X_{n-1}$  from  $p(X_{n-1}|\hat{X}_0^{(n)}, X_n)$ 
    else if  $1 < n < N$  then
        Sample  $X_{n-1}$  from  $p_G(X_{n-1}|\hat{X}_0^{(n)}, X_n, X_N)$ 
    else
         $X_0 = \hat{X}_0^{(n)}$ 
    end if
end for
return  $X_0$ 

```

4. Method

4.1. Implicit Image-to-Image Schrödinger Bridge

4.1.1. Motivation

We preserve the training process of $I^2\text{SB}$ and aim to accelerate its generative process in our proposed $I^3\text{SB}$. The generative process of $I^2\text{SB}$ is essentially a Markovian chain, where X_{n-1} depends solely on X_n given the trained network ϵ_{θ^*} (with $\hat{X}_0^{(n)}$ also being a function of X_n). In this setup, the information in X_N is only used in the first step and may be gradually lost through the Markovian chain. $I^3\text{SB}$ fully utilizes the information in X_N by incorporating X_N in each generative step, changing the generative process to a non-Markovian chain, as shown in Figure 1.

4.1.2. Algorithm

In the generative step from X_n to X_{n-1} , we first compute $\hat{X}_0^{(n)}$ using equation (3) and then sample X_{n-1} from a distribution p_G :

$$X_{n-1} \sim p_G(X_{n-1}|\hat{X}_0^{(n)}, X_n, X_N), \quad (6)$$

where X_N is included. Following DDIM [8], we design $p_G(X_{n-1}|\hat{X}_0^{(n)}, X_n, X_N)$ as a Gaussian distribution with a linear combination of $\hat{X}_0^{(n)}$, X_n and X_N as its mean and $g_n^2 I$ as its covariance matrix :

$$p_G(X_{n-1}|\hat{X}_0^{(n)}, X_n, X_N) = \mathcal{N}(X_{n-1}; A_n \hat{X}_0^{(n)} + B_n X_n + C_n X_N, g_n^2 I), \quad (7)$$

where A_n , B_n and C_n represent the weights of $\hat{X}_0^{(n)}$, X_n and X_N , respectively, and g_n is a hyperparameter to be discussed later. Since the trained network from I²SB is directly used to predict $\hat{X}_0^{(n)}$ in equation (3), the marginal distribution of I³SB should align with that of I²SB. Specifically, if $\hat{X}_0^{(n)}$ equals to X_0 for any n , X_{n-1} sampled from $p_G(X_{n-1}|X_0, X_n, X_N)$ should follow the distribution of $q(X_{n-1}|X_0, X_N)$. Therefore, the distribution p_G must satisfy the following equation:

$$q(X_{n-1}|X_0, X_N) = \int p_G(X_{n-1}|X_0, X_n, X_N) q(X_n|X_0, X_N) dX_n. \quad (8)$$

Substituting equation (7) into equation (8) with $\hat{X}_0^{(n)}$ equal to X_0 , the weights A_n , B_n and C_n can be analytically expressed in terms of g_n :

$$A_n = \frac{\bar{\sigma}_{n-1}^2}{\sigma_N^2} - \frac{\bar{\sigma}_n^2}{\sigma_N^2} \frac{\sqrt{\sigma_{n-1}^2 \bar{\sigma}_{n-1}^2 - g_n^2 \sigma_N^2}}{\sigma_n \bar{\sigma}_n}, \quad (9a)$$

$$B_n = \frac{\sqrt{\sigma_{n-1}^2 \bar{\sigma}_{n-1}^2 - g_n^2 \sigma_N^2}}{\sigma_n \bar{\sigma}_n}, \quad (9b)$$

$$C_n = \frac{\sigma_{n-1}^2}{\sigma_N^2} - \frac{\sigma_n^2}{\sigma_N^2} \frac{\sqrt{\sigma_{n-1}^2 \bar{\sigma}_{n-1}^2 - g_n^2 \sigma_N^2}}{\sigma_n \bar{\sigma}_n}. \quad (9c)$$

Therefore, X_{n-1} can be efficiently sampled from $p_G(X_{n-1}|\hat{X}_0^{(n)}, X_n, X_N)$ using:

$$X_{n-1} = \left(\frac{\bar{\sigma}_{n-1}^2}{\sigma_N^2} \hat{X}_0^{(n)} + \frac{\sigma_{n-1}^2}{\sigma_N^2} X_N \right) + \sqrt{\frac{\sigma_{n-1}^2 \bar{\sigma}_{n-1}^2 - g_n^2 \hat{\epsilon}^{(n)}}{\sigma_N^2} + g_n \epsilon}, \quad (10)$$

where $\epsilon \sim \mathcal{N}(0, I)$ is Gaussian noise, and $\hat{\epsilon}^{(n)}$ is the normalized estimated Gaussian noise from X_n , defined as:

$$\hat{\epsilon}^{(n)} = \frac{\sigma_N}{\sigma_n \bar{\sigma}_n} \left(X_n - \left(\frac{\bar{\sigma}_n^2}{\sigma_N^2} \hat{X}_0^{(n)} + \frac{\sigma_n^2}{\sigma_N^2} X_N \right) \right). \quad (11)$$

The proofs for equations (9) and (10) are provided in Appendix A.1. The generative process of I³SB is summarized in Algorithm 1.

4.1.3. Hyperparameter

The term g_n can be freely designed as long as it adheres to the constraint:

$$0 \leq g_n \leq \frac{\sigma_{n-1} \bar{\sigma}_{n-1}}{\sigma_N}, \quad (12)$$

to ensure the meaningfulness of taking the square root in equation (10). The functionality of g_n can be interpreted from multiple aspects: it balances between the Markovian and non-Markovian components, controls the stochasticity of the generative process,

and determines the degree of dependence on the deterministic estimate of noise component. When g_n equals to $\frac{\sigma_{n-1}\alpha_{n-1}}{\sigma_n}$ for any n , p_G becomes equivalent to the DDPM posterior p , and the generative process of I³SB reverts to that of I²SB. We parameterized g_n as:

$$g_n = \eta \frac{\sigma_{n-1}\alpha_{n-1}}{\sigma_n}, \quad (13)$$

where η is left as a hyperparameter.

4.1.4. Relevance to PF-ODE

When g_n is set to 0, the generative process becomes fully deterministic. We provide the ODE for the deterministic generative process in Lemma 1 and establish its equivalence to the PF-ODE for the VE endpoint-fixed Schrödinger Bridge in Theorem 1. Proofs are provided in the Appendix A.2 and Appendix A.3.

Lemma 1. *If g_n is set to 0, then equation (10) can be treated as an Euler discretization of the following ODE:*

$$d\frac{X_t}{\sigma_t \bar{\sigma}_t} = \frac{X_1}{\sigma_1^2} d\frac{\sigma_t}{\bar{\sigma}_t} + \frac{\hat{X}_0^{(t)}(X_t)}{\sigma_1^2} d\frac{\bar{\sigma}_t}{\sigma_t}. \quad (14)$$

Theorem 1. *The ODE (14) is equivalent to the PF-ODE for the VE endpoint fixed Schrödinger Bridge.*

4.2. Implementation Details

We validated our proposed method using three groups of experiments: natural image, human face and medical image experiments.

4.2.1. Natural Image Experiments

For the natural image experiments, we used the pretrained I²SB model and evaluated our proposed I³SB on 10,000 images randomly selected from the validation set of ImageNet 256×256 [32]. We tested I³SB on two image restoration tasks: 4× super-resolution with bicubic interpolation (sr4x-bicubic) and JPEG restoration with a quality factor of 10 (JPEG-10). The original images from the dataset served as clean images, and the corresponding corrupted images were generated by downsampling the clean images by a factor of 4 using bicubic interpolation for the sr4x-bicubic task, and by applying JPEG compression with a quality factor of 10 for the JPEG-10 task. The hyperparameter η was set to 0.6.

4.2.2. Human Face Experiments

For the human face experiments, we conducted two groups of tests. In the first group, we used the CelebA-HQ [33] 512×512 dataset, randomly splitting it into 27,000 images for training and 3,000 images for testing. In the second group, following SR3 [2], we trained our model on the FFHQ [34] 512×512 dataset and tested it on the CelebA-HQ 512×512 dataset to assess its robustness to domain differences between

the training and testing datasets. The entire FFHQ 512×512 dataset was used for training, and 10,000 images were randomly selected from the CelebA-HQ 512×512 dataset for testing. We evaluated our method on two image restoration tasks: sr4x-bicubic and JPEG-10. The original images from the dataset served as clean images, and the corresponding corrupted images were generated using the same procedure as in the natural image experiments. The hyperparameter η was set to 0 in the sr4x-bicubic task in first group, and set to 0.2 in other tasks.

The neural network $\epsilon_\theta(X_n, t_n)$ we trained is a 2D residual U-Net with the same architecture used in DDPM [7]. We concatenated X_N with X_n along the channel dimension to serve as an additional condition for the network. During training, we used 1000 diffusion time steps with quadratic discretization, and adopted a symmetric scheduling of β_t [30]. The model was trained on randomly cropped patches of size 128×128 and tested on the entire 512×512 images. A batch size of 64 was employed during training, using the Adam algorithm with a learning rate of 8×10^{-5} for 200,000 iterations.

4.2.3. Medical Image Experiments

For the medical image experiments, we evaluated our method on CT sparse view reconstruction, 4× super-resolution (sr4x) and denoising tasks. For CT sparse view reconstruction and sr4x tasks, we used the RPLHR-CT-tiny dataset [35], consisting of anonymized chest CT volumes. The original CT images served as clean images, and the corresponding corrupted images were generated using the FBP algorithm with projections from 60 distinct views in a fan beam geometry for the CT sparse view reconstruction task, and by downsampling the clean images by a factor of 4 in the projection domain for the CT sr4x task. We used 40 cases (11,090 slices) for training and 5 cases (1,425 slices) for testing. For the CT denoising task, we utilized the Mayo Grand Challenge dataset [36], which includes anonymized abdominal CT scans from 10 patients (5,936 slices) with matched full-dose (FD) data and simulated quarter-dose (QD) data. The FD data served as clean images, and the corresponding QD data served as corrupted images. We used data from 8 patients for training and 2 patients for testing in our experiment. We trained the neural network $\epsilon_\theta(X_n, t_n)$ using the same architecture and hyperparameters as in the human face experiments, with η set to 0 during inference.

5. Results

5.1. Quantitative Results

5.1.1. Quantitative Results for Natural Image Experiments

In natural image experiments, we compared I³SB with several state-of-the-art methods, which can be grouped into three categories: (1) paired data Schrödinger Bridge, including I²SB [4], (2) conditional diffusion models, such as ADM [1], and (3) diffusion-based task-agnostic models, including DDNM [19], DDRM [20, 21], ΠGDM [23], and DPS [22]. For quantitative evaluation, we used Frechet Inception Distance (FID) [37] and Learned Perceptual Image Patch Similarity (LPIPS) [38] to assess perceptual quality and texture restoration, and Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) to evaluate fidelity to the ground truth. The experimental results for the sr4x-bicubic task are shown in Table 1, and for the JPEG-10 task in Table

Method	Time (s)	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow
ADM [1]	4.91	27.18	0.7767	0.2484	13.906
DDNM [19]	2.77	<u>27.54</u>	<u>0.7826</u>	0.2514	13.997
DDRM [20]	0.56	27.17	0.7740	0.2730	19.700
Π GDM [23]	9.84	25.48	0.7321	0.2130	4.382
DPS [22]	121.65	25.40	0.6940	0.3178	10.251
I^2 SB [4] (NFE=25)	0.67	26.11	0.7279	0.2671	6.633
I^2 SB [4] (NFE=50)	1.38	25.72	0.7130	0.2625	5.060
I^2 SB [4] (NFE=100)	2.79	25.44	0.7009	0.2598	4.128
I^3 SB (ours, NFE=1) [†]	0.03	29.07	0.8240	0.2368	13.299
I^3 SB (ours, NFE=25)	0.67	25.91	0.7150	<u>0.2568</u>	4.115
I^3 SB (ours, NFE=50)	1.38	25.63	0.7043	0.2586	<u>3.766</u>
I^3 SB (ours, NFE=100)	2.79	25.49	0.6992	0.2600	3.648

Table 1: Quantitative results and computation time (per image) for the sr4x-bicubic task in natural image experiments. [†] I^3 SB and I^2 SB yield identical results when NFE=1. **Bold**: best, under: second best.

Method	Time (s)	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	FID \downarrow
DDRM [21]	5.42	<u>28.50</u>	<u>0.8175</u>	0.2955	19.977
Π GDM [23]	11.39	25.82	0.7349	0.2711	6.137
I^2 SB [4] (NFE=25)	0.79	27.09	0.7736	0.2529	5.970
I^2 SB [4] (NFE=50)	1.63	26.81	0.7623	0.2503	4.682
I^2 SB [4] (NFE=100)	3.29	26.62	0.7530	0.2488	3.871
I^3 SB (ours, NFE=1) [†]	0.03	29.39	0.8400	0.2677	16.977
I^3 SB (ours, NFE=25)	0.79	26.86	0.7597	0.2448	3.764
I^3 SB (ours, NFE=50)	1.63	26.67	0.7520	<u>0.2470</u>	<u>3.471</u>
I^3 SB (ours, NFE=100)	3.29	26.58	0.7481	0.2485	3.357

Table 2: Quantitative results and computation time (per image) for the JPEG-10 task in natural image experiments. [†] I^3 SB and I^2 SB yield identical results when NFE=1. **Bold**: best, under: second best.

2. Baseline values were computed using the official implementations of these methods with default hyperparameters. All experiments were conducted on a single A100 GPU, with computation times for each method included in the tables for reference.

In the sr4x-bicubic task (Table 1), I^3 SB achieved the highest PSNR and SSIM with NFE=1 and the best FID with NFE=100. When comparing I^3 SB to I^2 SB at the same NFE, I^3 SB consistently attained better FID, with a 38% reduction at NFE=25, a 26% reduction at NFE=50, and a 12% reduction at NFE=100. Notably, compared to 100-step I^2 SB, the 25-step I^3 SB offered a fourfold acceleration in computation time while achieving slightly better performance across all metrics, including a 0.5 improvement in PSNR, a 0.015 increase in SSIM, a 0.003 reduction in LPIPS, and a 0.01 reduction in FID. When compared to ADM, DDNM, and DDRM, I^3 SB with a single generative step outperformed these methods, showing a PSNR increase of 1.5 to 2, a 0.04 to 0.05 improvement in SSIM, a 0.01 to 0.04 decrease in LPIPS, and an FID reduction of 0.6

Method	NFE	Face (CelebA)			Face (FFHQ)			Medical Image		
		Sr4x	JPEG-10	Sr4x	JPEG-10	Sparse	Sr4x	Denoise		
cDDPM [7]	25	8.318	12.999	8.810	15.678	43.958	45.211	44.080		
	50	5.655	11.043	6.371	13.109	36.890	41.325	36.835		
	100	4.263	9.432	4.790	10.376	29.333	36.264	29.584		
cDDIM [8]	25	4.571	10.565	3.925	12.671	42.376	43.053	42.092		
	50	3.410	8.623	3.021	10.293	34.301	39.231	35.357		
	100	<u>3.096</u>	7.115	<u>2.725</u>	8.475	26.203	34.191	28.335		
I ² SB [4]	25	6.361	7.330	5.257	13.245	36.607	38.982	29.570		
	50	4.768	6.126	3.689	11.468	28.888	32.538	22.207		
	100	3.870	5.587	3.111	10.172	<u>22.448</u>	<u>26.551</u>	<u>15.986</u>		
I ³ SB (ours)	25	3.996	5.694	2.842	10.382	33.240	37.682	28.200		
	50	3.237	5.338	2.693	9.129	24.476	30.475	20.387		
	100	3.009	<u>5.475</u>	2.918	<u>8.568</u>	17.673	23.772	14.028		

Table 3: FIDs for all tasks in human face and medical image experiments. "Face (CelebA)" refers to experiments trained and tested on CelebA-HQ, and "Face (FFHQ)" refers to experiments trained on FFHQ and tested on CelebA-HQ. "Sparse" represents the CT sparse view reconstruction task. Lower FID indicates better performance. **Bold**: best, Underlined: second best.

to 6. Against DPS, the 25-step I³SB demonstrated superior performance, with a 0.5 increase in PSNR, a 0.02 improvement in SSIM, a 19% reduction in LPIPS, and a 60% reduction in FID. While PIIGDM achieved the best LPIPS in this task, it required a known forward operator to incorporate data consistency during inference, had a worse FID, and was 14 times slower than the 25-step I³SB.

In the JPEG-10 task (Table 2), I³SB achieved the highest PSNR and SSIM with NFE=1, the best LPIPS with NFE=25, and the best FID with NFE=100. Similar to the sr4x-bicubic task, I³SB consistently outperformed I²SB in FID at the same NFE, with a 37% reduction at NFE=25, a 26% reduction at NFE=50, and a 13% reduction at NFE=100. Notably, compared to the 100-step I²SB, the 25-step I³SB provided a fourfold acceleration in computation time while achieving slightly better performance across all metrics, including a 0.2 increase in PSNR, a 0.007 improvement in SSIM, a 0.004 reduction in LPIPS, and a 0.1 reduction in FID. When compared to DDRM, I³SB with a single generative step delivered superior results, with a 0.9 increase in PSNR, a 0.02 improvement in SSIM, a 0.03 reduction in LPIPS, and a 3-point decrease in FID. Against PIIGDM, the 25-step I³SB demonstrated better performance, achieving a 1-point improvement in PSNR, a 0.025 increase in SSIM, a 10% reduction in LPIPS, and a 40% reduction in FID.

5.1.2. Quantitative Results for Human Face and Medical Image Experiments

In the human face and medical image experiments, we compared I³SB with two groups of state-of-the-art methods: (1) paired data Schrödinger Bridges, such as I²SB [4], and (2) conditional diffusion models, including conditional DDPM (cDDPM) [7] and conditional DDIM (cDDIM) [8]. We implemented cDDPM and cDDIM ourselves, with implementation details provided in Appendix B. For each task, we evaluated the

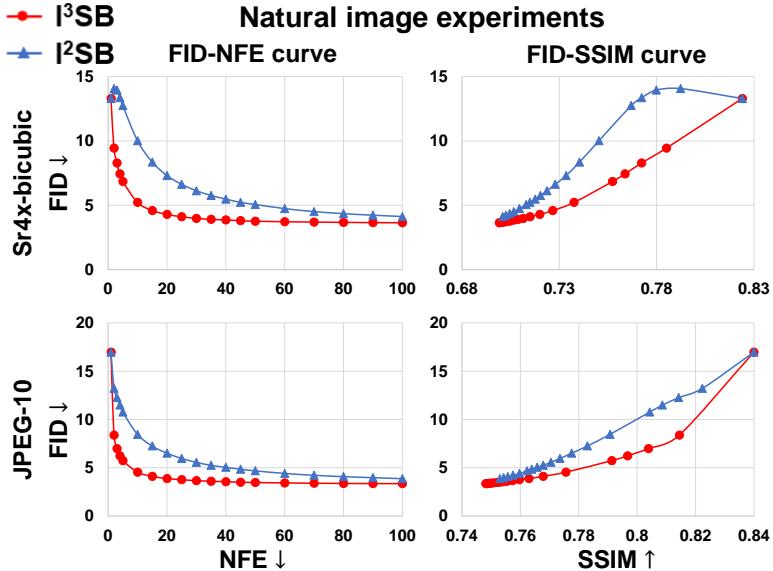


Figure 2: FID-NFE and FID-SSIM curves for the sr4x-bicubic and JPEG-10 tasks in natural image experiments. Each point on the FID-SSIM curves represents the FID and SSIM values at a specific NFE, ranging from 1 to 100. As NFE increases, the FID-SSIM curves shift from the top-right to the bottom-left. Red curves correspond to I^3SB , and blue curves correspond to I^2SB .

perceptual quality of the tested methods using FID at NFE values of 25, 50, and 100. The results are presented in Table 3.

In the human face experiments (Table 3), which included tasks trained on CelebA-HQ and FFHQ datasets with low-resolution and JPEG compression corruptions, I^3SB consistently outperformed I^2SB and cDDPM in terms of FID under the same NFEs. Compared to I^2SB , I^3SB achieved FID reductions of 22%-46% at NFE=25, 13%-32% at NFE=50, and 2%-22% at NFE=100. Notably, the 25-step I^3SB achieved FIDs comparable to the 100-step I^2SB , demonstrating a 4 \times speedup. When compared to cDDPM, I^3SB reduced FID by 34%-68% at NFE=25, 30%-58% at NFE=50, and 17%-42% at NFE=100. While I^3SB and cDDIM achieved similar FIDs at NFE=100, I^3SB delivered superior performance at lower NFEs, with FID reductions of 13%-46% at NFE=25 and 5%-38% at NFE=50.

In the medical image experiments (Table 3), which included CT sparse-view reconstruction, sr4x, and denoising tasks, I^3SB consistently outperformed I^2SB , cD-DPM, and cDDIM in terms of FID under the same NFEs. Compared to I^2SB , I^3SB achieved FID reductions of 3%-9% at NFE=25, 6%-15% at NFE=50, and 12%-21% at NFE=100. When compared to cDDPM and cDDIM, I^3SB delivered FID reductions of 12%-36% at NFE=25, 22%-45% at NFE=50, and 30%-53% at NFE=100.

5.2. FID-NFE and FID-SSIM Curves

To further illustrate the acceleration effect of I^3SB , we plotted the FID-NFE and FID-SSIM curves for both I^2SB and I^3SB across all experiments. Each point on the

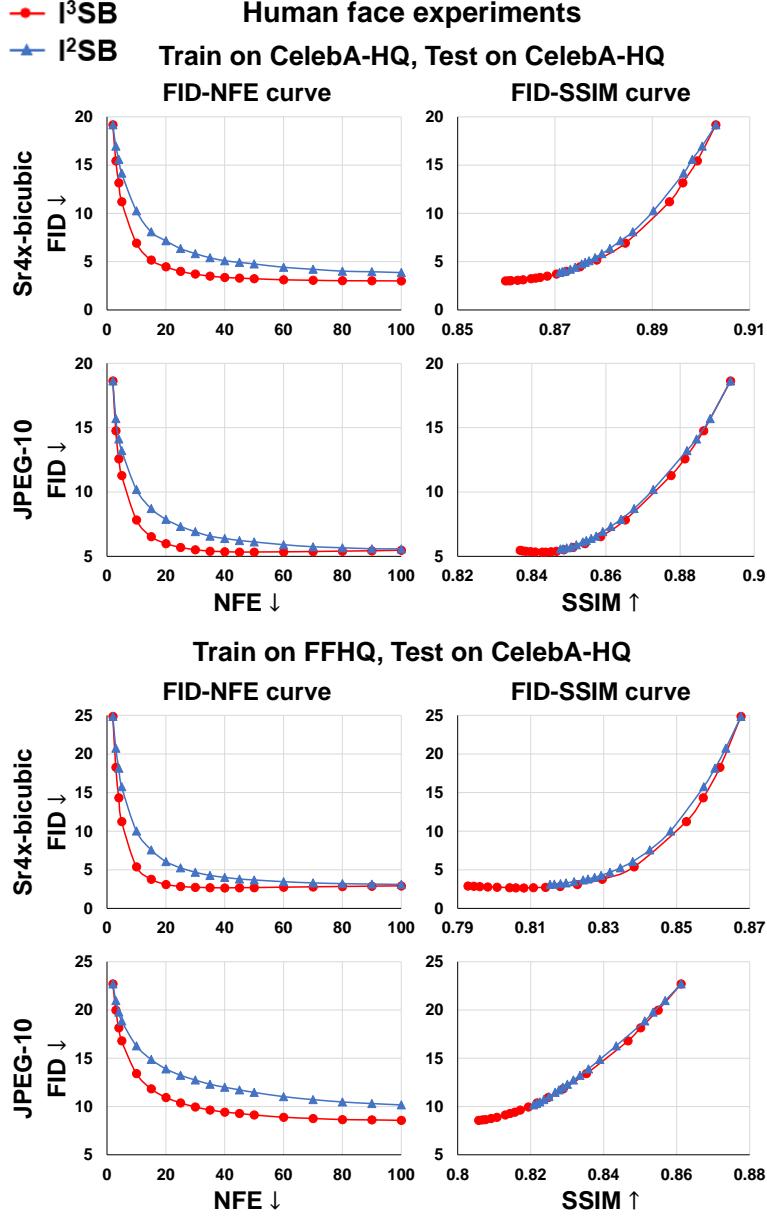


Figure 3: FID-NFE and FID-SSIM curves for sr4x-bicubic and JPEG-10 tasks in human face experiments. Each point on the FID-SSIM curves represents the FID and SSIM values at a specific NFE, ranging from 2 to 100. As NFE increases, the FID-SSIM curves shift from the top-right to the bottom-left. Red curves correspond to I^3SB , and blue curves correspond to I^2SB .

FID-SSIM curves corresponds to the FID and SSIM values at a specific NFE. As NFE

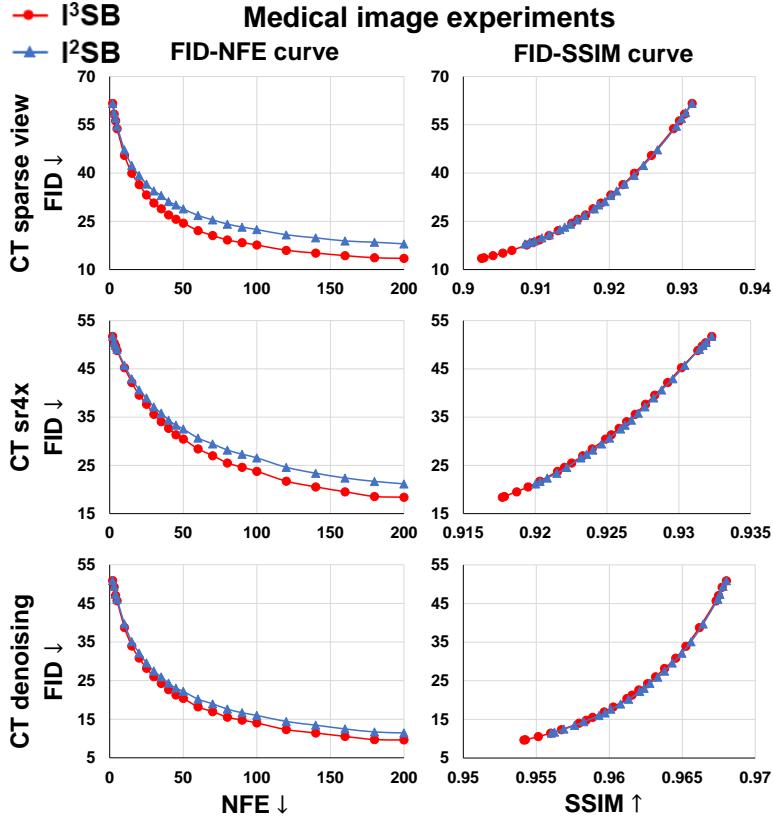


Figure 4: FID-NFE and FID-SSIM curves for CT sparse view reconstruction, sr4x and denoising tasks in medical image experiments. Each point on the FID-SSIM curves represents the FID and SSIM values at a specific NFE, ranging from 2 to 200. As NFE increases, the FID-SSIM curves shift from the top-right to the bottom-left. Red curves correspond to I^3SB , and blue curves correspond to I^2SB .

increases, the FID-SSIM curves shift from the top-right (high FID, high SSIM) to the bottom-left (low FID, low SSIM). These curves are presented in Figure 2 for natural image experiments, Figure 3 for human face experiments, and Figure 4 for medical image experiments.

Consistent trends emerged across all the tasks for all the experiments. As shown in the FID-NFE curves, both I^2SB and I^3SB exhibited decreasing FID as NFE increased, indicating that higher NFEs improved perceptual quality for both models. Notably, FID decreased faster in I^3SB as NFE increased, meaning that I^3SB achieved the same perceptual quality with fewer generative steps.

The FID-SSIM curves highlighted the trade-off between perceptual quality and fidelity to the ground truth, controlled by NFE. As NFE increases, the curves shift toward lower FID (better perceptual quality) but lower SSIM (reduced fidelity), reflecting the inherent compromise between detail generation and distortion [5]. Notably, the FID-SSIM curves for I^3SB either overlapped with or shifted to the lower right of those for

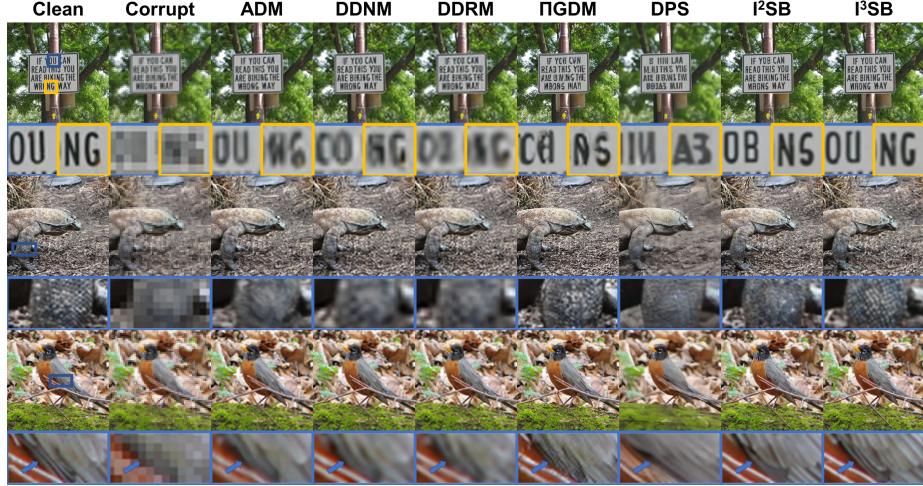


Figure 5: Visualization result for the sr4x-bicubic task in the natural image experiments. Details within blue and yellow boxes are zoomed in for enhanced visual clarity. The NFE for I^2SB is 100, and for I^3SB is 25.

I^2SB , indicating that at equivalent perceptual quality, I^3SB maintained equal or better fidelity. This demonstrated that I^3SB accelerated the generative process of I^2SB , achieving the same perceptual quality with fewer generative steps, without introducing additional distortion from the ground truth.

Additionally, these trends were consistent when training on the FFHQ dataset and testing on the CelebA-HQ dataset in the human face experiments (Figure 3), indicating that the acceleration impact of I^3SB was robust to domain differences.

In natural image experiments, I^3SB achieved comparable performance to the 100-step I^2SB with 25 generative steps, resulting in a $4\times$ speedup. For human face experiments, I^3SB matched the results of the 100-step I^2SB with 20 to 30 steps, providing a $3\times$ to $5\times$ speedup. In medical image experiments, I^3SB showed a $1.4\times$ to $2\times$ acceleration, achieving similar performance as the 200-step I^2SB with 100 steps in the CT sparse view reconstruction task, and 140 steps in both the CT sr4x and denoising tasks.

5.3. Visualization Results

The superior performance of I^3SB is further demonstrated by the visualization results in Figure 5 and Figure 6 for natural image experiments, Figure 7 for human face experiments, and Figure 8 for medical image experiments. Compared to all comparison methods, I^3SB exhibited enhanced detail restoration across all figures. Specifically, it excelled in restoring finer details, such as text characters, lizard scales, and bird features in Figure 5; the bird's eyes and legs, as well as the dog's nose and tongue in Figure 6, wrinkles, eyelashes, teeth, and nose in Figure 7; and pulmonary veins and mammary glands in Figure 8.



Figure 6: Visualization results for the JPEG-10 task in the natural image experiments. Details within blue and yellow boxes are zoomed in for enhanced visual clarity. The NFE for $I^2\text{SB}$ is 100, and for $I^3\text{SB}$ is 25.

6. Conclusion and Discussion

In conclusion, we introduce $I^3\text{SB}$ to accelerate the generative process of $I^2\text{SB}$ for image restoration tasks. By incorporating corrupted images into each generative step, $I^3\text{SB}$ fully leverages their information, reformulating the process into a non-Markovian framework while maintaining the same marginal distribution as $I^2\text{SB}$, enabling direct use of pretrained models. A single hyperparameter balances the Markovian and non-Markovian components. Additionally, we establish the equivalence between $I^3\text{SB}$'s deterministic generative process and the PF-ODE of the VE endpoint-fixed Schrödinger Bridge. Extensive experiments across many image restoration tasks—including super-resolution and JPEG restoration for natural and human face images, as well as sparse-view reconstruction, super-resolution, and denoising for medical images—demonstrate the significant acceleration benefits of $I^3\text{SB}$. Compared to $I^2\text{SB}$, $I^3\text{SB}$ achieves the same perceptual quality with fewer generative steps while maintaining or improving fidelity to the ground truth. Additionally, $I^3\text{SB}$ outperforms several state-of-the-art diffusion-based image restoration models in both quantitative metrics and visual quality.

Despite these encouraging results, there are some limitations to our approach. First, $I^3\text{SB}$ accelerates the generative process of $I^2\text{SB}$ by using the information present in

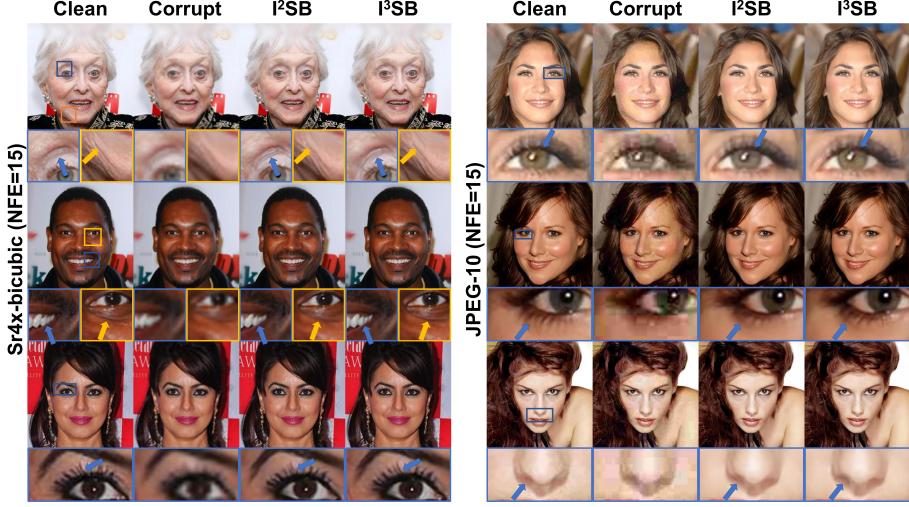


Figure 7: Visualization results for sr4x-bicubic and JPEG-10 tasks in the human face experiments, trained and tested on CelebA-HQ. Details within blue and yellow boxes are zoomed in for enhanced visual clarity.

the corrupted image. Consequently, in cases involving corruptions like occlusion, or extremely severe corruptions (e.g., 64 \times super-resolution), where the corrupted image contains little to no useful information, I³SB may not achieve substantial acceleration. Second, similar to I²SB, I³SB requires paired clean and corrupted images for training, which limits its applicability in unpaired or weakly supervised settings. Third, I³SB implicitly assumes that clean images are noise-free. This assumption may hinder its performance when the clean images contain inherent noise, as is often the case in medical imaging. This is a potential reason why I³SB achieves approximately 4 \times speedup in natural image and human face experiments but is limited to 2 \times speedup in medical image experiments.

In future work, alongside addressing the limitations mentioned above, we plan to extend I³SB in several directions. First, while I³SB is currently based on the I²SB framework, which operates as a Variance Exploding (VE) paired data Schrödinger Bridge, we aim to expand I³SB to the Variance Preserving (VP) paired data Schrödinger Bridges. Second, since I³SB offers an efficient deterministic sampling approach when $\eta = 0$, it could serve as a foundation for distillation in paired data Schrödinger Bridges, potentially further accelerating the generative process. Third, as emphasized by CDDB [5], incorporating data consistency into the generative process—when the forward operator of the corruption is known—can enhance the performance of paired data Schrödinger Bridges. We plan to investigate this by integrating data consistency constraints into I³SB, particularly for tasks like CT sparse view reconstruction, where the forward operator is available. We hope that I³SB will serve as a robust and versatile framework for future research in image restoration, inspiring new advancements in the field.

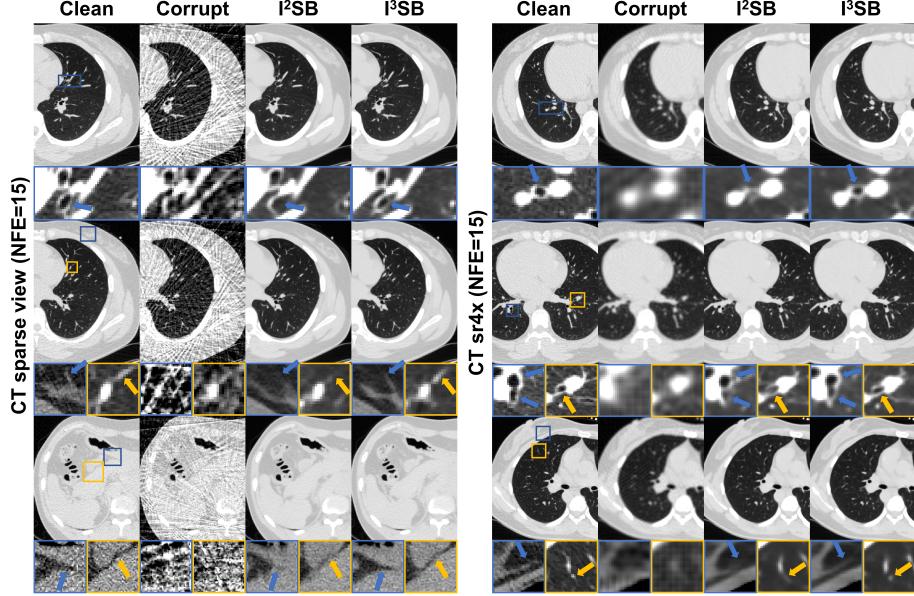


Figure 8: Visualization results for the CT sparse view reconstruction and sr4x tasks in the medical image experiments. The details within the blue and yellow boxes are zoomed in for enhanced visual clarity. The display window for the entire images is set to [-1000HU, 200HU], for the zoomed regions outside the lungs is set to [-160HU, 200HU], and for the zoomed regions inside the lungs is set to [-1000HU, -250HU].

Appendix A. Proofs

Appendix A.1. Proof for Equations (9) and (10)

We define $p(X_{n-1}|X_0, X_N)$ as:

$$p(X_{n-1}|X_0, X_N) = \int p_G(X_{n-1}|X_0, X_n, X_N) q(X_n|X_0, X_N) dX_n. \quad (\text{A.1})$$

According to Bishop (2006) [39], $p(X_{n-1}|X_0, X_N)$ is Gaussian, denoted as $\mathcal{N}(X_{n-1}|\mu_{n-1}, \Sigma_{n-1})$ where

$$\mu_{n-1} = A_n X_0 + C_n X_N + B_n \left(\frac{\bar{\sigma}_n^2}{\sigma_N^2} X_0 + \frac{\sigma_n^2}{\sigma_N^2} X_N \right), \quad (\text{A.2a})$$

$$\Sigma_{n-1} = (g_n^2 + B_n^2 \frac{\sigma_n^2 \bar{\sigma}_n^2}{\sigma_N^2}) I. \quad (\text{A.2b})$$

Given equation (8), $p(X_{n-1}|X_0, X_N)$ equals to $q(X_{n-1}|X_0, X_N)$ for any X_{n-1} , X_0 and X_N . Therefore, the weights for X_0 and X_N , as well as the covariance matrix of $p(X_{n-1}|X_0, X_N)$,

should match those of $q(X_{n-1}|X_0, X_N)$. This leads to the following equations:

$$\frac{\bar{\sigma}_{n-1}^2}{\sigma_N^2} = A_n + B_n \frac{\bar{\sigma}_n^2}{\sigma_N^2}, \quad (\text{A.3a})$$

$$\frac{\sigma_{n-1}^2}{\sigma_N^2} = C_n + B_n \frac{\sigma_n^2}{\sigma_N^2}, \quad (\text{A.3b})$$

$$\frac{\bar{\sigma}_{n-1}^2 \sigma_{n-1}^2}{\sigma_N^2} = g_n^2 + B_n^2 \frac{\sigma_n^2 \bar{\sigma}_n^2}{\sigma_N^2}. \quad (\text{A.3c})$$

By solving the system of equations in (A.3), we obtain the expressions for A_n , B_n and C_n as given in equation (9).

The sample X_{n-1} is drawn from $p_G(X_{n-1}|\hat{X}_0^{(n)}, X_n, X_N)$ as in equation (6):

$$X_{n-1} = A_n \hat{X}_0^{(n)} + B_n X_n + C_n X_N + g_n \epsilon, \quad (\text{A.4})$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise. By substituting the expressions for A_n , B_n and C_n into equation (A.4), we derive the expression for X_n as shown in equation (10).

Appendix A.2. Proof for Lemma 1

When g_n is set to 0, equation (10) simplifies to:

$$X_{n-1} = \left(\frac{\bar{\sigma}_{n-1}^2}{\sigma_N^2} \hat{X}_0^{(n)} + \frac{\sigma_{n-1}^2}{\sigma_N^2} X_N \right) + \frac{\sigma_{n-1} \bar{\sigma}_{n-1}}{\sigma_n \bar{\sigma}_n} \left(X_n - \left(\frac{\bar{\sigma}_n^2}{\sigma_N^2} \hat{X}_0^{(n)} + \frac{\sigma_n^2}{\sigma_N^2} X_N \right) \right). \quad (\text{A.5})$$

This can be rearranged as:

$$\frac{X_{n-1}}{\sigma_{n-1} \bar{\sigma}_{n-1}} - \frac{X_n}{\sigma_n \bar{\sigma}_n} = \left(\frac{\sigma_{n-1}}{\bar{\sigma}_{n-1}} - \frac{\sigma_n}{\bar{\sigma}_n} \right) \frac{X_N}{\sigma_N^2} + \left(\frac{\bar{\sigma}_{n-1}}{\sigma_{n-1}} - \frac{\bar{\sigma}_n}{\sigma_n} \right) \frac{\hat{X}_0^{(n)}}{\sigma_N^2}, \quad (\text{A.6})$$

which can be seen as an Euler discretization of the ODE in equation (14).

Appendix A.3. Proof for Theorem 1

To prove Theorem 1, we begin by deriving the PF-ODE for the VE endpoint-fixed Schrödinger Bridge. Following this, we establish its equivalence with the ODE given in equation (14).

The Schrödinger Bridge constructs diffusion bridges between two arbitrary distributions, p_A and p_B , with the following forward and backward SDEs:

$$dX_t = [f_t + \beta_t \nabla_{X_t} \log \Psi(X_t, t)] dt + \sqrt{\beta_t} dw_t, \quad (\text{A.7a})$$

$$dX_t = [f_t - \beta_t \nabla_{X_t} \log \hat{\Psi}(X_t, t)] dt + \sqrt{\beta_t} d\bar{w}_t. \quad (\text{A.7b})$$

Here, X_0 is sampled from p_A , X_1 is sampled from p_B , and w_t and \bar{w}_t denote the Wiener process and its time-reversed counterpart. To ensure that the path measure induced by the forward SDE (A.7a) is almost surely equal to the one induced by the reverse

SDE (A.7b), the time-varying energy potentials Ψ and $\hat{\Psi}$ should satisfy the following coupled partial differential equations (PDEs):

$$\frac{\partial \Psi}{\partial t} = -\nabla \Psi^\top f - \frac{1}{2}\beta \Delta \Psi, \quad (\text{A.8a})$$

$$\frac{\partial \hat{\Psi}}{\partial t} = -\nabla \cdot (\hat{\Psi} f) + \frac{1}{2}\beta \Delta \hat{\Psi}, \quad (\text{A.8b})$$

with the marginal conditions:

$$\Psi(X_0, 0) \hat{\Psi}(X_0, 0) = p_A(X_0), \quad (\text{A.9a})$$

$$\Psi(X_1, 1) \hat{\Psi}(X_1, 1) = p_B(X_1). \quad (\text{A.9b})$$

As noted by Chen 2021 [30], the ODE

$$dX_t = \left(f_t + \frac{1}{2}\beta_t \nabla_{X_t} \log \frac{\Psi(X_t, t)}{\hat{\Psi}(X_t, t)} \right) dt \quad (\text{A.10})$$

characterizes the probability flow of the forward and reverse processes of the Schrödinger Bridge defined in equation (A.7).

VE endpoint-fixed Schrödinger Bridge defines f_t , p_A and p_B as follows:

$$f_t = 0, \quad (\text{A.11a})$$

$$p_A(X_0) = q_{\text{clean}}(X_0 | X_{\text{corrupt}}), \quad (\text{A.11b})$$

$$p_B(X_1) = \delta(X_1 - X_{\text{corrupt}}), \quad (\text{A.11c})$$

where X_{corrupt} represents a given corrupted image, and $q_{\text{clean}}(\cdot | X_{\text{corrupt}})$ represents the clean image distribution conditioned on X_{corrupt} . The Dirac function δ indicates that the endpoint fixed Schrödinger Bridge is constructed for each specific corrupted image, rather than for the entire corrupted image distribution.

Lemma 2. *If p_A , p_B and f_t are defined as in equations (A.11), then the PDEs (A.8) with the marginal conditions (A.9) admit the following analytical solutions:*

$$\Psi(X_t, t) = \mathcal{N}(X_t | X_{\text{corrupt}}, \bar{\sigma}_t^2 I), \quad (\text{A.12})$$

$$\hat{\Psi}(X_t, t) = \int \hat{\Psi}_{X_0}(X_t, t) q_{\text{clean}}(X_0 | X_{\text{corrupt}}) dX_0, \quad (\text{A.13})$$

where

$$\hat{\Psi}_{X_0}(X_t, t) = C_{X_0} \mathcal{N}(X_t | X_0, \sigma_t^2 I), \quad (\text{A.14})$$

and

$$C_{X_0} = \left(\sqrt{2\pi} \sigma_1 \right)^d \exp \left(\frac{\left(X_0 - X_{\text{corrupt}} \right)^\top \left(X_0 - X_{\text{corrupt}} \right)}{2\sigma_1^2} \right). \quad (\text{A.15})$$

Furthermore,

$$\nabla_{X_t} \log \Psi = -\frac{1}{\bar{\sigma}_t^2} (X_t - X_{\text{corrupt}}), \quad (\text{A.16})$$

and

$$\nabla_{X_t} \log \hat{\Psi} = -\frac{1}{\sigma_t^2} (X_t - \hat{X}_0^{(t)}), \quad (\text{A.17})$$

where the expected mean $\hat{X}_0^{(t)}$ is defined as:

$$\hat{X}_0^{(t)} = \int X_0 q_{\text{clean}}(X_0 | X_t, X_{\text{corrupt}}) dX_0. \quad (\text{A.18})$$

With Lemma 2, we provide proofs for the equivalence between the ODEs (14) and (A.10). Applying the chain rule, we obtain:

$$d\frac{\sigma_t}{\bar{\sigma}_t} = \frac{\sigma_1^2}{2\sigma_t \bar{\sigma}_t^3} \beta_t dt, \quad (\text{A.19a})$$

$$d\frac{\bar{\sigma}_t}{\sigma_t} = -\frac{\sigma_1^2}{2\sigma_t^3 \bar{\sigma}_t} \beta_t dt, \quad (\text{A.19b})$$

and

$$\begin{aligned} d\frac{X_t}{\sigma_t \bar{\sigma}_t} &= \frac{1}{\sigma_t \bar{\sigma}_t} dX_t + X_t d\frac{1}{\sigma_t \bar{\sigma}_t}, \\ &= \frac{1}{\sigma_t \bar{\sigma}_t} dX_t + X_t \frac{\sigma_t^2 - \bar{\sigma}_t^2}{2\sigma_t^3 \bar{\sigma}_t^3} \beta_t dt. \end{aligned} \quad (\text{A.20})$$

Substituting equations (A.19) and (A.20) into equation (14), we find that the ODE (14) is equivalent to:

$$dX_t = \frac{1}{2} \beta_t \left(\frac{1}{\bar{\sigma}_t^2} (X_1 - X_t) + \frac{1}{\sigma_t^2} (X_t - \hat{X}_0^{(t)}) \right) dt. \quad (\text{A.21})$$

Using $X_1 = X_{\text{corrupt}}$ along with equations (A.11a), (A.16) and (A.17), we conclude that the ODE (14) is equivalent to ODE (A.10), which is the PF- ODE for the VE endpoint fixed Schrödinger Bridge.

Appendix A.4. Proof for Lemma 2

To prove Lemma 2, we first demonstrate that Ψ , as defined in equation (A.12), satisfies the PDE (A.8a). This holds because:

$$\begin{aligned} \frac{\partial \Psi}{\partial t} &= \frac{\partial \Psi}{\partial \bar{\sigma}_t^2} \frac{\partial \bar{\sigma}_t^2}{\partial t}, \\ &= -\frac{1}{2} \beta \Psi \left(\frac{(X_t - X_{\text{corrupt}})^T (X_t - X_{\text{corrupt}})}{\bar{\sigma}_t^4} - \frac{d}{\sigma_t^2} \right), \\ &= -\frac{1}{2} \beta \Delta \Psi. \end{aligned} \quad (\text{A.22})$$

Similarly, $\hat{\Psi}_{X_0}(X_t, t)$ satisfies: $\frac{\partial \hat{\Psi}_{X_0}}{\partial t} = \frac{1}{2}\beta \Delta \hat{\Psi}_{X_0}$. Consequently, $\hat{\Psi}$ expressed in equation (A.13) satisfies:

$$\begin{aligned}\frac{\partial \hat{\Psi}}{\partial t} &= \int \frac{\partial \hat{\Psi}_{X_0}}{\partial t} q_{\text{clean}}(X_0|X_{\text{corrupt}}, y) dX_0, \\ &= \frac{1}{2}\beta \int \Delta \hat{\Psi}_{X_0} q_{\text{clean}}(X_0|X_{\text{corrupt}}, y) dX_0, \\ &= \frac{1}{2}\beta \Delta \hat{\Psi}.\end{aligned}\quad (\text{A.23})$$

Therefore, Ψ and $\hat{\Psi}$ satisfy the PDEs (A.8).

Next, we proof that Ψ and $\hat{\Psi}$ satisfy the equations (A.9). At $t = 0$, we have

$$\Psi(X_0, 0) = 1/C_{X_0}, \quad (\text{A.24})$$

where C_{X_0} is defined in equation (A.15). Additionally,

$$\begin{aligned}\hat{\Psi}(X_0, 0) &= \int C_X \delta(X_0 - X) q_{\text{clean}}(X|X_{\text{corrupt}}) dX, \\ &= C_{X_0} q_{\text{clean}}(X_0|X_{\text{corrupt}}).\end{aligned}\quad (\text{A.25})$$

Therefore, the marginal condition (A.9a) holds. At $t = 1$, we have

$$\Psi(X_1, 1) = \delta(X_1 - X_{\text{corrupt}}), \quad (\text{A.26})$$

and since $\hat{\Psi}_{X_0}(X_1 = X_{\text{corrupt}}, 1)$ equals 1, we have:

$$\begin{aligned}\Psi(X_1, 1) \hat{\Psi}(X_1, 1) &= \delta(X_1 - X_{\text{corrupt}}) \int \hat{\Psi}_{X_0}(X_1, 1) q_{\text{clean}}(X_0|X_{\text{corrupt}}) dX_0, \\ &= \delta(X_1 - X_{\text{corrupt}}) \int q_{\text{clean}}(X_0|X_{\text{corrupt}}) dX_0, \\ &= \delta(X_1 - X_{\text{corrupt}}), \\ &= p_B(X_1).\end{aligned}\quad (\text{A.27})$$

Thus, Ψ and $\hat{\Psi}$ satisfy the marginal conditions (A.9).

Finally, we provide proofs for equations (A.16) and (A.17). For equation (A.16), $\nabla_{X_t} \log \Psi$ can be straightforwardly obtained by direct computation. For equation (A.17), we proceed as follows:

$$\begin{aligned}\nabla \log \hat{\Psi} &= \frac{\nabla \hat{\Psi}}{\hat{\Psi}} \\ &= \frac{1}{\hat{\Psi}} \int \nabla \hat{\Psi}_{X_0}(X_t, t) q_{\text{clean}}(X_0|X_{\text{corrupt}}) dX_0 \\ &= \frac{1}{\hat{\Psi}} \int \left(\frac{X_0 - X_t}{\sigma_t^2} \right) \hat{\Psi}_{X_0} q_{\text{clean}}(X_0|X_{\text{corrupt}}) dX_0 \\ &= -\frac{1}{\sigma_t^2} \left(X_t - \frac{1}{\hat{\Psi}} \int X_0 \hat{\Psi}_{X_0} q_{\text{clean}}(X_0|X_{\text{corrupt}}) dX_0 \right)\end{aligned}\quad (\text{A.28})$$

Using the definition of $\hat{\Psi}_{X_0}$ in equation (A.14), we have:

$$\hat{\Psi}_{X_0}(X_t, t) = k_{X_t} q(X_t | X_0, X_1 = X_{\text{corrupt}}), \quad (\text{A.29})$$

where $q(X_t | X_0, X_1)$ is defined in equation (1), and k_{X_t} is independent of X_0 . Since X_1 is sampled from a Dirac distribution centered at X_{corrupt} , we obtain:

$$q(X_t | X_0, X_1 = X_{\text{corrupt}}) = q(X_t | X_0, X_{\text{corrupt}}). \quad (\text{A.30})$$

Therefore:

$$\nabla \log \hat{\Psi} = -\frac{1}{\sigma_t^2} \left(X_t - \frac{\int X_0 q(X_t | X_0, X_{\text{corrupt}}) q_{\text{clean}}(X_0 | X_{\text{corrupt}}) dX_0}{\int q(X_t | X_0, X_{\text{corrupt}}) q_{\text{clean}}(X_0 | X_{\text{corrupt}}) dX_0} \right) \quad (\text{A.31})$$

Using Bayes' theorem, we know that:

$$q(X_t | X_0, X_{\text{corrupt}}) = \frac{q_{\text{clean}}(X_0 | X_t, X_{\text{corrupt}}) q(X_t | X_{\text{corrupt}})}{q_{\text{clean}}(X_0 | X_{\text{corrupt}})}. \quad (\text{A.32})$$

Thus, we have

$$\begin{aligned} \nabla \log \hat{\Psi} &= -\frac{1}{\sigma_t^2} \left(X_t - \frac{\int X_0 q_{\text{clean}}(X_0 | X_t, X_{\text{corrupt}}) dX_0}{\int q_{\text{clean}}(X_0 | X_t, X_{\text{corrupt}}) dX_0} \right) \\ &= -\frac{1}{\sigma_t^2} (X_t - \hat{X}_0^{(t)}). \end{aligned} \quad (\text{A.33})$$

That completes the proof.

Appendix B. Implementation Details for cDDPM and cDDIM

We implemented cDDPM and cDDIM as comparison methods for human face and medical image experiments. We adopted the sigmoid schedule [40] for β_t and used the same network architecture and training settings as the l²SB model trained for these tasks. The network was trained to predict the ground truth in human face experiments and to predict the noise in medical image experiments. The hyperparameter η in cDDIM was set to 0 for human face experiments and 0.6 for medical image experiments.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgements

The authors acknowledge financial support provided by National Institute of Biomedical Imaging and Bioengineering and Samsung Electronics Co., Ltd.

References

- [1] P. Dhariwal, A. Nichol, Diffusion models beat gans on image synthesis, *Advances in neural information processing systems* 34 (2021) 8780–8794.
- [2] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, M. Norouzi, Image super-resolution via iterative refinement, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (2022) 4713–4726.
- [3] S. Gonzalez-Sabbagh, A. Robles-Kelly, S. Gao, Dgd-cgan: A dual generator for image dewatering and restoration, *Pattern Recognition* 148 (2024) 110159.
- [4] G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, A. Anand-kumar, I2sb: Image-to-image schrödinger bridge, in: International Conference on Machine Learning, 2023. URL: <https://api.semanticscholar.org/CorpusID:257022338>.
- [5] H. Chung, J. Kim, J. C. Ye, Direct diffusion bridge using data consistency for inverse problems, *Advances in Neural Information Processing Systems* 36 (2024).
- [6] M. Delbracio, P. Milanfar, Inversion by direct iteration: An alternative to denoising diffusion for image restoration, *Transactions on Machine Learning Research* (2023). URL: <https://openreview.net/forum?id=VmyFF51L3F>, featured Certification.
- [7] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems* 33 (2020) 6840–6851.
- [8] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations, 2021. URL: <https://openreview.net/forum?id=St1giarCHLP>.
- [9] L. Zhou, A. Lou, S. Khanna, S. Ermon, Denoising diffusion bridge models, in: The Twelfth International Conference on Learning Representations, 2024. URL: <https://openreview.net/forum?id=FKksTayvGo>.
- [10] A. Buades, B. Coll, J.-M. Morel, Non-local means denoising, *Image Processing On Line* 1 (2011) 208–212.
- [11] A. El Hamidi, M. Menard, M. Lugiez, C. Ghannam, Weighted and extended total variation for image restoration and decomposition, *Pattern Recognition* 43 (2010) 1564–1576.
- [12] P. Li, J. Liang, M. Zhang, W. Fan, G. Yu, Joint image denoising with gradient direction and edge-preserving regularization, *Pattern Recognition* 125 (2022) 108506.
- [13] M. Aharon, M. Elad, A. Bruckstein, K-svd: An algorithm for designing over-complete dictionaries for sparse representation, *IEEE Transactions on signal processing* 54 (2006) 4311–4322.

- [14] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, Restormer: Efficient transformer for high-resolution image restoration, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5728–5739.
- [15] W. Harvey, S. Naderiparizi, F. Wood, Conditional image generation by conditioning variational auto-encoders, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=7MV6uLz0ChW>.
- [16] L. Helminger, M. Bernasconi, A. Djelouah, M. Gross, C. Schroers, Generic image restoration with flow based priors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 334–343.
- [17] Y. Liu, J. He, Y. Liu, X. Lin, F. Yu, J. Hu, Y. Qiao, C. Dong, Adaptbir: Adaptive blind image restoration with latent diffusion prior for higher fidelity, Pattern Recognition (2024) 110659.
- [18] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, B. Poole, Score-based generative modeling through stochastic differential equations, in: International Conference on Learning Representations, 2021. URL: <https://openreview.net/forum?id=PxTIG12RRHS>.
- [19] Y. Wang, J. Yu, J. Zhang, Zero-shot image restoration using denoising diffusion null-space model, in: The Eleventh International Conference on Learning Representations, 2023. URL: <https://openreview.net/forum?id=mRieQgMtNTQ>.
- [20] B. Kawar, M. Elad, S. Ermon, J. Song, Denoising diffusion restoration models, Advances in Neural Information Processing Systems 35 (2022) 23593–23606.
- [21] B. Kawar, J. Song, S. Ermon, M. Elad, Jpeg artifact correction using denoising diffusion restoration models, in: Neural Information Processing Systems (NeurIPS) Workshop on Score-Based Methods, 2022.
- [22] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, J. C. Ye, Diffusion posterior sampling for general noisy inverse problems, in: The Eleventh International Conference on Learning Representations, 2023. URL: <https://openreview.net/forum?id=0nD9zGAGTOk>.
- [23] J. Song, A. Vahdat, M. Mardani, J. Kautz, Pseudoinverse-guided diffusion models for inverse problems, in: International Conference on Learning Representations, 2023.
- [24] M. Mardani, J. Song, J. Kautz, A. Vahdat, A variational perspective on solving inverse problems with diffusion models, in: The Twelfth International Conference on Learning Representations, 2024. URL: <https://openreview.net/forum?id=1Y04EE3SPB>.
- [25] T. Karras, M. Aittala, T. Aila, S. Laine, Elucidating the design space of diffusion-based generative models, Advances in neural information processing systems 35 (2022) 26565–26577.

- [26] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, J. Zhu, Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, *Advances in Neural Information Processing Systems* 35 (2022) 5775–5787.
- [27] L. Liu, Y. Ren, Z. Lin, Z. Zhao, Pseudo numerical methods for diffusion models on manifolds, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=P1KWVd2yBkY>.
- [28] T. Salimans, J. Ho, Progressive distillation for fast sampling of diffusion models, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=TIIdIXIpzhoI>.
- [29] Y. Song, P. Dhariwal, M. Chen, I. Sutskever, Consistency models, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.
- [30] T. Chen, G.-H. Liu, E. Theodorou, Likelihood training of schrödinger bridge using forward-backward SDEs theory, in: International Conference on Learning Representations, 2022. URL: <https://openreview.net/forum?id=nioAdKCedXB>.
- [31] H. Chung, S. Lee, J. C. Ye, Decomposed diffusion sampler for accelerating large-scale inverse problems, in: The Twelfth International Conference on Learning Representations, 2024. URL: <https://openreview.net/forum?id=DsEhqQtfAG>.
- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
- [33] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018. URL: <https://openreview.net/forum?id=Hk99zCeAb>.
- [34] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.
- [35] P. Yu, H. Zhang, H. Kang, W. Tang, C. W. Arnold, R. Zhang, Rplhr-ct dataset and transformer baseline for volumetric super-resolution from ct scans, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2022, pp. 344–353.
- [36] AAPM, Low dose ct grand challenge, [Online], 2017. Available: <http://www.aapm.org/GrandChallenge/LowDoseCT/>.
- [37] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, *Advances in neural information processing systems* 30 (2017).

- [38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: CVPR, 2018.
- [39] C. M. Bishop, Pattern recognition and machine learning, Springer google schola 2 (2006) 1122–1128.
- [40] A. Jabri, D. J. Fleet, T. Chen, Scalable adaptive computation for iterative generation, in: Proceedings of the 40th International Conference on Machine Learning, 2023, pp. 14569–14589.