I^{2}SB: Image-to-Image Schrödinger Bridge

Guan-Horng Liu [* 1] **Arash Vahdat** [2] **De-An Huang** [2] **Evangelos A. Theodorou** [1]
Weili Nie [† 2] **Anima Anandkumar** [† 2 3]

Abstract

We propose Image-to-Image Schrödinger Bridge (I^{2}SB), a new class of conditional diffusion models that directly learn the nonlinear diffusion processes between two given distributions. These *diffusion bridges* are particularly useful for image restoration, as the degraded images are structurally informative priors for reconstructing the clean images. I^{2}SB belongs to a tractable class of Schrödinger bridge, the nonlinear extension to score-based models, whose marginal distributions can be computed analytically given boundary pairs. This results in a simulation-free framework for nonlinear diffusions, where the I^{2}SB training becomes scalable by adopting practical techniques used in standard diffusion models. We validate I^{2}SB in solving various image restoration tasks, including inpainting, super-resolution, deblurring, and JPEG restoration on ImageNet 256×256 and show that I^{2}SB surpasses standard conditional diffusion models with more interpretable generative processes. Moreover, I^{2}SB matches the performance of inverse methods that additionally require the knowledge of the corruption operators. Our work opens up new algorithmic opportunities for developing efficient nonlinear diffusion models on a large scale. Project page and codes: https://i2sb.github.io/.
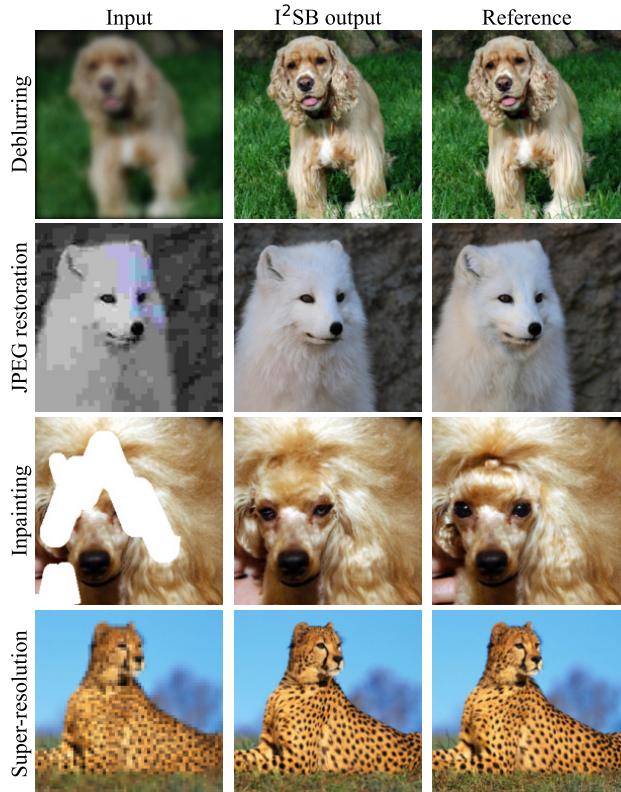
Figure 1. Outputs of our proposed **Image-to-Image Schrödinger Bridge (I^{2}SB)** on ImageNet 256×256 validation set for various image restoration tasks.

1. Introduction

Image restoration is a crucial problem in vision and image processing with applications in optimal filtering (Motwani et al., 2004), data compression (Wallace, 1991), adversarial defense (Nie et al., 2022), and safety-critical systems such as medicine and robotics (Song et al., 2021b; Li et al., 2021). Common image restoration tasks are known to be ill-posed (Banham & Katsaggelos, 1997; Richardson, 1972) and typically solved by modern data-driven approaches with conditional generation (Mirza & Osindero, 2014; Khan et al., 2022), *i.e.*, by learning to sample the underlying (clean) data distribution given the degraded distribution.

Diffusion and score-based generative models (SGMs; Sohl-Dickstein et al. (2015); Song et al. (2020b)) have emerged as powerful conditional generative models with their remarkable successes in synthesizing high-fidelity data (Dhariwal & Nichol, 2021; Rombach et al., 2022; Vahdat et al., 2021). These models rely on progressively diffusing data to noise, and learning the score functions (often parameterized by neural networks) to reverse the processes (Anderson, 1982); the reversed processes enable generation from noise to data.
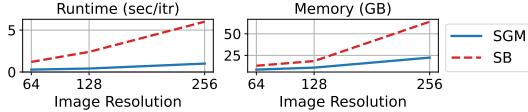
Figure 2. Complexity of SGM and SB (Chen et al., 2021a) On 256×256 resolution, SB is 6× slower and consumes 3× memory.

Figure 3. **I²SB** belongs to a tractable class of SB that shares the same computational framework of SGM and rebases the terminal distribution beyond simple Gaussian priors.

Saharia et al. (2021; 2022) show that these generative processes can be adopted for image restoration by feeding degraded images as extra inputs to the score network so that the processes are biased toward the corresponding intact images. Alternatively, when the mapping between clean and degraded images is known, the tasks can be reformulated as inverse problems that restore the underlying clean signal from the degraded measurement, based on the diffusion priors (Kawar et al., 2022a;b; Wang et al., 2022b).

Notably, all of the aforementioned diffusion models for image restoration begin their generative denoising processes with Gaussian white noise, which has little or no structural information of the clean data distribution. Despite arising naturally from unconditional generation, it remains unclear whether this default setup best suits image-to-image translation problems especially like image restoration, where the degraded images are much more structurally informative compared to random noise.

An alternative that better leverages the problem structure is to directly start the generative processes from degraded images, and build diffusion *bridges* between clean and degraded data distributions. This shares similarity with image-to-image translation GANs (Zhu et al., 2017; Huang et al., 2018). Constructing these diffusion *bridges* often necessitates a new computational framework for reversing general diffusion processes. It has been recently explored in Schrödinger bridge (SB; De Bortoli et al. (2021); Chen et al. (2021a)), a generalized nonlinear score-based model which defines optimal transport between two arbitrary distributions and generalizes beyond Gaussian priors.

Despite the mathematical generalization, computational frameworks for solving SB (Chen et al., 2021b) have been developed independently (hence distinctly) from how diffusion models are typically trained. This makes SB computationally unfavorable compared to its score-based counterpart especially in high-dimensional regimes (see Figure 2), where SB is known to suffer from, *e.g.,* discretization error
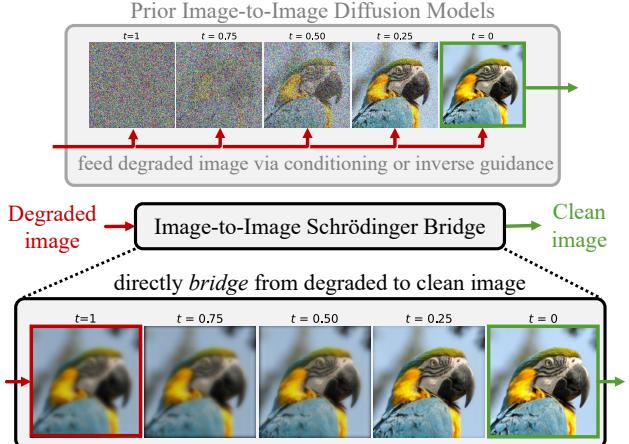

Figure 4. Illustration of **I²SB**. Rather than generating images from random noise as in prior diffusion models, I²SB directly learns the diffusion bridges between degraded and clean distributions, yielding more interpretable generation effective for image restoration.

(De Bortoli et al., 2021), high variance (Chen et al., 2021a), or even divergence (Fernandes et al., 2021). It remains an open question whether SB can be made practical for learning complex nonlinear diffusions on a large scale.

In this work, we propose **Image-to-Image Schrödinger Bridge (I²SB)**, a sub-class of SB with nonlinear diffusion models that share the same computational framework used in standard score-based models. Consequently, practical techniques from diffusion models for learning high-dimensional data distributions (Karras et al., 2022; Song & Ermon, 2020) can be adopted to train nonlinear diffusions. This is achieved by exploiting the linear structure hidden in the nonlinear coupling of SB to construct tractable SBs for transporting between individual clean images and their corresponding degraded distributions, *i.e.,* I²SB. We show that the marginal distributions of I²SB admit analytic solutions given boundary pairs (*i.e.,* clean and degraded image pairs), thereby yielding a simulation-free framework that avoids unfavorable complexity (Chen et al., 2021a). Furthermore, we demonstrate that I²SB can be simulated at test time using DDPM (Ho et al., 2020). Finally, we characterize in how I²SB reduces to an optimal transport ODE (Peyré et al., 2019) when the diffusion vanishes, strengthening the algorithmic connection among dynamic generative models.

We validate our method in many image restoration tasks including super-resolution, deblurring, inpainting, and JPEG restoration on ImageNet 256×256 (Deng et al., 2009); see Figure 1. Through extensive experiments, we show that I²SB surpasses standard conditional diffusion models (Saharia et al., 2022) and matches diffusion-based inverse models (Kawar et al., 2022a;b) *without* exploiting the corruption operators. With these more interpretable generative processes, I²SB enjoys little or no performance drops as the

number of function evaluation (NFE) decreases.

In summary, we present the following contributions.

- We introduce I^2SB, a new class of conditional diffusion models that learns fully nonlinear diffusion bridges between two domain distributions.

- We build I^2SB on a simulation-free computational framework that adopts scalable techniques from standard diffusion models to train nonlinear diffusion processes.

- I^2SB sets new records in many restoration tasks, including super-resolution, deblurring, inpainting, and JPEG restoration. It yields more interpretable generation and enjoys little performance drops as the NFE decreases.

2. Preliminaries

Notation: Let $X_t \in \mathbb{R}^d$ be a d-dimensional stochastic process indexed by $t \in [0, 1]$ and denote the discrete time step by $0 = t_0 < \cdots t_n \cdots < t_N = 1$, we shorthand $X_n \equiv X_{t_n}$. The Wiener process and its reversed counterpart (Anderson, 1982) are denoted by $W_t, \overline{W}_t \in \mathbb{R}^d$. The identity matrix is denoted by $I \in \mathbb{R}^{d \times d}$.

2.1. Score-based Generative Model (SGM)

SGM (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b) is an emerging class of dynamic generative models that, given data X_0 sampled from some domain $p_{\mathcal{A}}$, constructs stochastic differential equations (SDEs),

$$\mathrm{d}X_t = f_t(X_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t, \qquad (1)$$

whose terminal distributions at $t = 1$ approach an approximate Gaussian, *i.e.*, $X_1 \sim \mathcal{N}(0, I)$. This is achieved by properly choosing the diffusion $\beta_t \in \mathbb{R}$ and setting the base drift f_t linear in X_t. It is known that reversing (1) yields another SDE traversing backward in time (Anderson, 1982):

$$\mathrm{d}X_t = [f_t - \beta_t \nabla \log p(X_t, t)]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overline{W}_t, \quad (2)$$

where $p(\cdot, t)$ is the marginal density of (1) at time t and $\nabla \log p$ is its *score function*. The SDE (2) is known as the "reversed process of (1)" in the sense that its path-wise measure equals almost surely to the one induced by (1); thus, the two SDEs also share the same marginal densities.

In practice, given a tuple (X_0, t, X_t) where $X_0 \sim p_{\mathcal{A}}$, $t \sim \mathcal{U}([0, 1])$, and X_t sampled analytically from (1), one can parameterize $\epsilon(X_t, t; \theta)$ with, *e.g.,* U-Net (Ronneberger et al., 2015), and regress its output w.r.t. the rescaled version of denoising score-matching objective (Vincent, 2011),

$$\|\epsilon(X_t, t; \theta) - \sigma_t \nabla \log p(X_t, t | X_0)\|, \qquad (3)$$

where $\nabla \log p(X_t, t | X_0)$ can be computed analytically and σ_t^2 is the variance of $X_t | X_0$, induced by (1), that rescales the regression target to unit variance (Ho et al., 2020).

Other advanced parameterizations that better account for practical training (Karras et al., 2022) have also been explored recently. Importantly, they all produce some ways to predict intact data at $t = 0$ from the network outputs. In other words, the mapping $\epsilon(X_t, t; \theta) \mapsto X_0^\epsilon$ is readily available once ϵ is trained.[1] With this, popular samplers like DDPM (Ho et al., 2020) can be written compactly as recursive posterior sampling:

$$X_n \sim p(X_n | X_0^\epsilon, X_{n+1}), \quad X_N \sim \mathcal{N}(0, I). \qquad (4)$$

2.2. Schrödinger Bridge (SB)

SB (Schrödinger, 1932; Léonard, 2013) is an entropy-regularized optimal transport model that considers the following forward and backward SDEs:

$$\mathrm{d}X_t = [f_t + \beta_t \nabla \log \Psi(X_t, t)]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t, \qquad (5a)$$

$$\mathrm{d}X_t = [f_t - \beta_t \nabla \log \widehat{\Psi}(X_t, t)]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overline{W}_t, \qquad (5b)$$

where $X_0 \sim p_{\mathcal{A}}$ and $X_1 \sim p_{\mathcal{B}}$ are drawn from boundary distributions in two distinct domains. The functions $\Psi, \widehat{\Psi} \in C^{2,1}(\mathbb{R}^d, [0, 1])$ are time-varying energy potentials that solve the following coupled PDEs,

$$\begin{cases} \frac{\partial \Psi(x,t)}{\partial t} = -\nabla \Psi^\top f - \frac{1}{2}\beta \Delta \Psi \\ \frac{\partial \widehat{\Psi}(x,t)}{\partial t} = -\nabla \cdot (\widehat{\Psi} f) + \frac{1}{2}\beta \Delta \widehat{\Psi} \end{cases} \qquad (6a)$$

$$\text{s.t. } \Psi(x,0)\widehat{\Psi}(x,0) = p_{\mathcal{A}}(x), \Psi(x,1)\widehat{\Psi}(x,1) = p_{\mathcal{B}}(x) \quad (6b)$$

In this case, the path measure induced by SDE (5a) equals almost surely to the one induced by SDE (5b), similar to SDEs (1,2). Hence, their marginal densities, denoted by $q(\cdot, t)$ hereafter, are also equivalent.

SGM as a Special Case of SB It is known that SB generalizes SGM to nonlinear structure (Chen et al., 2021a). Indeed, the SDEs between SGM (1,2) and SB (5) differ only by the additional nonlinear forward drift $\nabla \log \Psi$, which allows the processes to transport samples beyond Gaussian priors. In such cases, the backward drift $\nabla \log \widehat{\Psi}$ is no longer the score function of (5a), yet they relate to each other via the Nelson's duality (Nelson, 1967)

$$\Psi(x, t)\widehat{\Psi}(x, t) = q(x, t)$$
$$\Rightarrow \nabla \log \Psi(x, t) - \nabla \log q(x, t) = -\nabla \log \widehat{\Psi}(x, t). \qquad (7)$$

One can verify that reversing (5a) yields

$$\mathrm{d}X_t = [f_t + \beta_t \nabla \log \Psi - \beta_t \nabla \log q]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overline{W}_t, \quad (8)$$

which indeed equals (5b) after substituting (7). Hence, (5b) reverses the nonlinear forward SDE (5a), and vice versa.

---

[1]In all cases, we can write $X_t = a_t X_0 + b_t \epsilon$ for some $a_t, b_t \in \mathbb{R}$ depending on (1) so that the mapping can be defined as $X_0^\epsilon := (X_t - b_t \epsilon_\theta)/a_t$ given any trained ϵ_θ.

3. Image-to-Image Schrödinger Bridge (I²SB)

We propose a tractable class of SB that directly constructs diffusion bridges between two domains, making it suitable for image-to-image translation such as image restoration. All proofs are left to Appendix A due to space constraint.

3.1. Mathematical Framework

Solving SB using SGM Framework Despite the fact that SB generalizes SGM in theory, numerical methods for SB and SGM have been developed independently on distinct computational frameworks. Due to the coupling constraints in (6b), modern SB models often adopt iterative projection methods (Kullback, 1968; Chen et al., 2021b), which have unfavorable complexity as the dimension grows (see Figure 2). It is unclear whether practical techniques in the SGM computational framework can be transferred to efficiently learn nonlinear diffusions.

Let us reexamine the SB theory in detail, but this time through the computational framework of SGM. Notice that

- The nonlinear drifts in (5) resemble the score function in (2) when we view $\Psi(\cdot, t)$ and $\widehat{\Psi}(\cdot, t)$ as the densities.

- Equation (6a) gives the solution to the Fokker-Plank equation (Risken, 1996) that characterizes the marginal density induced by the *linear* SDE in (1).

With these, we can reformulate PDEs (6) in a manner that makes SB more compatible with the SGM framework:

Theorem 3.1 (Reformulating SB drifts as score functions)**.** *When the Schrödinger systems (6) hold, $\nabla \log \widehat{\Psi}(X_t, t)$ and $\nabla \log \Psi(X_t, t)$ are the score functions of the following linear SDEs, respectively:*

$$\mathrm{d}X_t = f_t(X_t)\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}W_t, \quad X_0 \sim \widehat{\Psi}(\cdot, 0), \quad (9a)$$

$$\mathrm{d}X_t = f_t(X_t)\mathrm{d}t + \sqrt{\beta_t}\,\mathrm{d}\overline{W}_t, \quad X_1 \sim \Psi(\cdot, 1). \quad (9b)$$

Theorem 3.1 suggests that the backward drift $\nabla \log \widehat{\Psi}$ in SDE (5b) that transports samples from $p_\mathcal{B}$ to $p_\mathcal{A}$ can also be used to reverse the forward SDE (9a). Crucially, the above linear SDEs (9) have different boundary distributions from nonlinear SDEs (5). Essentially, the nonlinearity of $\nabla \log \widehat{\Psi}$—as the combination of the nonlinear forward drift and its score function (*c.f.* (7))—is absorbed into the initial condition $\widehat{\Psi}(\cdot, 0)$, leaving it compactly as the score function of another linear SDE. Hence, if we can draw samples from $X_0 \sim \widehat{\Psi}(\cdot, 0)$, we can parameterize $\nabla \log \widehat{\Psi}$ with the score network and apply practical techniques from SGM to learn $\nabla \log \widehat{\Psi}$. Similar reasoning applies to $\nabla \log \Psi$.

A Tractable Class of SB Theorem 3.1 is encouraging yet not immediately useful as the boundaries $\widehat{\Psi}(\cdot, 0)$ and $\Psi(\cdot, 1)$ remain intractable due to the couplings in (6b). Below, we present a tractable case that eliminates one of the couplings.
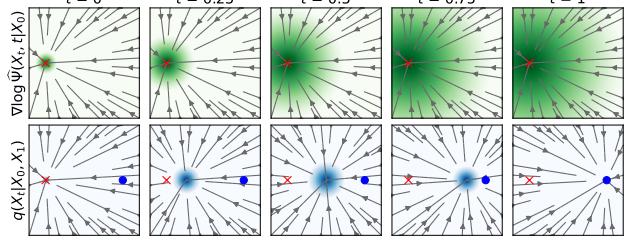

Figure 5. **Top:** The backward drift $\nabla \log \widehat{\Psi}(X_t, t|X_0 = a)$ transporting $p_\mathcal{B}$ toward a (denoted $\times$) corresponds to the score function of a tractable Gaussian (density marked green). **Bottom:** Simulating the backward SDE (5b) with this $\nabla \log \widehat{\Psi}$ yields a diffusion bridge (density marked blue) between $X_0 = a$ and $X_1 \sim p_\mathcal{B}$, whose mean corresponds to the optimal transport path.

Corollary 3.2 (Tractable SB with the Dirac delta boundary)**.** *Let $p_\mathcal{A}(\cdot) := \delta_a(\cdot)$ be the Dirac delta distribution centered at $a \in \mathbb{R}^d$. Then, the initial distributions in (9) are given by*

$$\widehat{\Psi}(\cdot, 0) = \delta_a(\cdot), \quad \Psi(\cdot, 1) = \frac{p_\mathcal{B}}{\widehat{\Psi}(\cdot, 1)}. \quad (10)$$

Comparing (10) to (6b), it is clear that Corollary 3.2 breaks the dependency on Ψ for solving $\widehat{\Psi}(x, 0)$. Intuitively, the optimal[2] backward drift driving the reverse process of (9a) to the Dirac delta $\delta_a(\cdot)$ always flows toward a, regardless of $p_\mathcal{B}$; see Figure 5. The Dirac delta assumption also implicitly appears in the denoising objective (3), which first computes the target $\nabla \log p(X_t, t|X_0 = a)$ for each data point a, as the score between $\delta_a(\cdot)$ and Gaussian, then averages over $X_0 \sim p_\mathcal{A}$. In this vein, Corollary 3.2 adopts the same boundary $\delta_a(\cdot)$ on one side and generalizes the other side from Gaussian to arbitrary $p_\mathcal{B}$. Indeed, we show in Appendix A that when $p_\mathcal{B} = \widehat{\Psi}(\cdot, 1) \approx \mathcal{N}(0, I)$, the forward drift vanishes with $\Psi(\cdot, t) = 1$, reducing the framework to SGM.

Although the singularity of $\delta_a(\cdot)$ may hinder generalization beyond training samples, in practice, the score network generalizes well to unseen samples from the same distributions, for both SGM and our I²SB, partly due to the strong generalization ability of neural networks (Zhang et al., 2021).

To summarize, our theories suggest an efficient pipeline for training $\nabla \log \widehat{\Psi}$ without dealing with the intractability of reversing the nonlinear forward drift. By formulating a tractable SB compatible with the SGM framework, we get both mathematical soundness and computational efficiency.

3.2. Algorithmic Design

In this subsection, we discuss practical designs for applying Corollary 3.2 to image restoration. We will adopt similar setups from prior diffusion models (Saharia et al., 2022) and assume pair information is available during training, *i.e.,* $p(X_0, X_1) = p_\mathcal{A}(X_0)p_\mathcal{B}(X_1|X_0)$. From which, we can

---

[2]The optimality is w.r.t. minimum energy; see Appendix B.

Algorithm 1 Training

1: **Input:** clean $p_{\mathcal{A}}(\cdot)$ and degraded $p_{\mathcal{B}}(\cdot|X_0)$ datasets
2: **repeat**
3:     $t \sim \mathcal{U}([0,1])$, $X_0 \sim p_{\mathcal{A}}(X_0)$, $X_1 \sim p_{\mathcal{B}}(X_1|X_0)$
4:     $X_t \sim q(X_t|X_0, X_1)$ according to (11)
5:     Take gradient descent step on $\epsilon(X_t, t; \theta)$ using (12)
6: **until** converges

Algorithm 2 Generation

1: **Input:** $X_N \sim p_{\mathcal{B}}(X_N)$, trained $\epsilon(\cdot, \cdot; \theta)$
2: **for** $n = N$ **to** 1 **do**
3:     Predict X_0^ϵ using $\epsilon(X_n, t_n; \theta)$
4:     $X_{n-1} \sim p(X_{n-1}|X_0^\epsilon, X_n)$ according to DDPM (4)
5: **end for**
6: **return** X_0

---

construct tractable SBs between individual data points X_0 and their corresponding degraded distributions $p_{\mathcal{B}}(X_1|X_0)$. As rebasing the terminal distribution from Gaussian to $p_{\mathcal{B}}(\cdot|X_0)$ makes f unnecessary, we will drop $f := 0$ and let I^2SB learn the full nonlinear drift by itself.

Sampling Proposal for Training and Generation   Training scalable diffusion models requires efficient computation of X_t. The computation is intractable for I^2SB, if directly from the nonlinear SDE (5a), since its forward drift $\nabla \log \Psi$ is not only generally nonlinear but never explicitly constructed. Computing X_t from the linear SDE (9a) whose score function corresponds to $\nabla \log \widehat{\Psi}$ will not work either. Since the diffusion process in (9a) does *not* converge to the terminal distribution (*i.e.,* $p_{\mathcal{B}}(X_1|X_0)$) of I^2SB, high-probability regions induced by (9a) can be far away from regions where the generative processes actually traverse; see Figure 5. We address the difficulty in the following result.

Proposition 3.3 (Analytic posterior given boundary pair)**.** *The posterior of (5) given some boundary pair* (X_0, X_1), *provided* $f := 0$, *admits an analytic form:*

$$q(X_t|X_0, X_1) = \mathcal{N}(X_t; \mu_t(X_0, X_1), \Sigma_t), \quad (11)$$

$$\mu_t = \frac{\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} X_0 + \frac{\sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} X_1, \quad \Sigma_t = \frac{\sigma_t^2 \bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2} \cdot I,$$

where $\sigma_t^2 := \int_0^t \beta_\tau \mathrm{d}\tau$ *and* $\bar{\sigma}_t^2 := \int_t^1 \beta_\tau \mathrm{d}\tau$ *are variances accumulated from either sides. Further, this posterior marginalizes the recursive posterior sampling in DDPM (4):*

$$q(X_n|X_0, X_N) = \int \Pi_{k=n}^{N-1} p(X_k|X_0, X_{k+1}) \mathrm{d}X_{k+1}.$$

Proposition 3.3 suggests that the analytic posterior of SB given a boundary pair (X_0, X_1) is the marginal density induced by DDPM, $p(X_k|X_0^\epsilon, X_{k+1})$, when $X_0^\epsilon := X_0$ and $X_N \sim p_{\mathcal{B}}$. Practically, this suggests that *(i)* during training when (X_0, X_1) are available from $p_{\mathcal{A}}(X_0)$ and $p_{\mathcal{B}}(X_1|X_0)$, we can sample X_t directly from (11) without solving any nonlinear diffusion as in prior SB models (Vargas et al., 2021), and *(ii)* during generation when only $X_1 \sim p_{\mathcal{B}}$ is given, running standard DDPM starting from X_1 induces the same marginal density of SB paths so long as the predicted X_0^ϵ is close to X_0. Therefore, the proposed sampling proposal in (11) is both tractable and able to cover regions traversed by generative processes.

Parameterization & Objective   Since I^2SB requires no conditioning modules, we adopt the same network parameterization $\epsilon(X_t, t; \theta)$ from SGM (Dhariwal & Nichol, 2021). Similar to the objective (3), we can compute the score function for $\nabla \log \widehat{\Psi}(X_t, t|X_0) \equiv \nabla \log p^{(9a)}(X_t, t|X_0)$, except X_t being drawn from (11). This leads to

$$\left\| \epsilon(X_t, t; \theta) - \frac{X_t - X_0}{\sigma_t} \right\| \quad (12)$$

as we adopt $f := 0$. Algorithms 1 and 2 summarize the training and generation procedures of I^2SB, respectively.

3.3. Connection to Flow-based Optimal Transport (OT)

It is known that the solution to SB, as an entropic optimal transport model, converges weakly to the optimal transport plan (Mikami, 2004) as the diffusion degenerates. The following result characterizes this infinitesimal limit.

Proposition 3.4 (Optimal Transport ODE; OT-ODE)**.** *When* $\beta_t \to 0$, *the SDE between* (X_0, X_1) *reduces to an ODE:*

$$\mathrm{d}X_t = v_t(X_t|X_0)\mathrm{d}t, \ v_t(X_t|X_0) = \frac{\beta_t}{\sigma_t^2}(X_t - X_0), \ (13)$$

whose solution $\mu_t(X_0, X_1)$ *is the posterior mean of (11).*

Note that the OT-ODE (13) is *not* a probability flow ODE, which has the same marginal as the corresponding SDE, in the SGM literature (Chen et al., 2018; Song et al., 2021a). Instead, the OT-ODE (13) simulates an OT plan (Peyré et al., 2019) only when the stochasticity of the SDE vanishes.

Proposition 3.4 suggests that the mean of the posterior q represents the OT-ODE paths. Hence, I^2SB can also be instantiated as a simulation-free OT by replacing the posteriors with their means, *i.e.,* by removing the noise injected into X_t in both training and generation (the lines 4 in Algorithms 1 and 2). The ratio $\frac{\beta_t}{\sigma_t^2}$ characterizes how fast the OT-ODE approaches X_0, in a similar vein to the noise scheduler in SGM (Nichol & Dhariwal, 2021). With this interpretation in mind, we introduce our final result, which complements recent advances in flow-matching (Lipman et al., 2022) except for image-to-image problem setups.

Corollary 3.5. *For sufficiently small* $\beta_t := \beta$ *that remains constant over* t, *we have* $v_t = \frac{X_t - X_0}{t}$ *and* $\mu_t = (1-t)X_0 + tX_1$, *which recover the OT displacement (McCann, 1997).*

Table 1. Comparison of different diffusion models in boundary distributions and tractability of forward and backward drifts. Note that I^2SB requires pair information compared to standard SB.

Model	$p(X_0)$	$p(X_1)$	$\nabla \log \Psi$	$\nabla \log \widehat{\Psi}$
(C)SGM	$p_{\mathcal{A}}$	$\mathcal{N}(0, I)$	0	tractable
I^2SB	$p_{\mathcal{A}}$	$p_{\mathcal{B}}(\cdot|X_0)$	intractable	tractable
SB	$p_{\mathcal{A}}$	$p_{\mathcal{B}}(\cdot)$	intractable	intractable

3.4. Comparison to Standard Conditional Diffusion Model

I^2SB can be thought of as a new class of conditional diffusion models that better leverages the degraded images as the structurally informative priors. It differs from the standard conditional SGM (CSGM; Rombach et al. (2022); Saharia et al. (2022)), which simply constructs a *conditional score function* with the newly available information (in this case, the degraded images) as an additional input. The generative denoising process in CSGM remains the same as the SDE (2) in SGM that starts from a Gaussian prior. Intuitively, it is more efficient to learn the direct mappings between clean and degraded images given that they are already close to each other. We summarize the comparison of I^2SB with other diffusion models in Table 1.

4. Related Work

Conditional SGMs (CSGMs) for image restoration refers to a class of diffusion models that bias the generative processes (Song et al., 2020b) toward the underlying intact image of some degraded measurements. This is typically achieved by conditioning the network with the degraded images via, *e.g.,* concatenation or attention (Rombach et al., 2022). CSGMs have demonstrated impressive results in many restoration tasks such as deblurring (Whang et al., 2022), super-resolution (Saharia et al., 2021), and inpainting (Saharia et al., 2022); yet, all of them start the generative processes from noise, which has little structural information of the clean data distribution. Pandey et al. (2022) explored a new reparametrization of the linear forward SDE to refine a VAE's output. In contrast, our I^2SB is built on a tractable SB framework and is the first to directly bridge clean and degraded image distributions for image restoration.

Diffusion-based inverse model (DIM) combines inverse problem techniques (Song et al., 2021b) with the diffusion priors (Ramesh et al., 2022; Wang et al., 2022a) and aims to restore the underlying clean image signal from the (noisy) measurement given by the degraded image. DIM typically performs projection at each generative step via, *e.g.,* Baye's rule (Chung et al., 2022b; Song et al., 2022) so that the generation best aligns with the observed measurement. This, however, requires knowing the degraded operators, whether linear (Kawar et al., 2022a; Wang et al., 2022b) or nonlinear

(Kawar et al., 2022b; Chung et al., 2022a), in both training and test time. In contrast, our I^2SB, similar to other CSGMs, does not require knowing these operators, making it generally applicable without task-specific manipulations.

5. Experiment

5.1. Experimental Setup

Model We parameterize $\epsilon(X_t, t; \theta)$ with U-Net (Ronneberger et al., 2015) and initialize the network with the unconditional ADM checkpoint (Dhariwal & Nichol, 2021) trained on ImageNet 256×256. Other parameterization, *e.g.,* preconditioning (Karras et al., 2022), is also applicable upon proper adaptation (see Appendix C.3), yet we observed little performance difference. We set $f := 0$ and consider a symmetric scheduling of β_t where the diffusion shrinks at both boundaries; see Figure 6. This is suggested by prior SB models (De Bortoli et al., 2021; Chen et al., 2021a). By default, we use 1000 sampling time steps for all tasks with quadratic discretization (Song et al., 2020a).

Figure 6. Symmetric noise scheduling.

Baselines We compare I^2SB with three classes of diffusion models for image restoration, namely CSGM and DIM discussed in Section 4 and standard SB models. Specifically, we consider Palette (Saharia et al., 2022) and ADM (Dhariwal & Nichol, 2021) for CSGM baselines. For DIM models, we consider DDRM (Kawar et al., 2022a;b), DDNM (Wang et al., 2022b), and ΠGDM (Song et al., 2022), but stress that they require additionally knowing the corruption operators at both training and generation. This is in contrast to CSGM models—including Palette and I^2SB. We report the results of DIM models for completeness. Finally, for the SB baseline, we consider CDSB (Shi et al., 2022) which extends the work of De Bortoli et al. (2021) to conditional generation.

Evaluation We showcase the performance of I^2SB in solving various image restoration problems, including inpainting, JPEG restoration, deblurring, and 4× super-resolution (64×64 to 256×256), on ImageNet 256×256. For each restoration problem, we consider 2-3 tasks by varying, *e.g.,* the quality factors, filtering kernels, and mask types. We keep the implementation and setup of each restoration task the same as the baselines (Kawar et al., 2022a;b; Saharia et al., 2022) for a fair comparison; see Appendix C for details. For quantitative metrics, we choose the Frechet Inception Distance (FID; Heusel et al. (2017)) and Classifier Accuracy (CA) of a pre-trained ResNet50 (He et al., 2016). Similar to the baselines (Saharia et al., 2022; Song et al., 2022), we report super-resolution results on the full ImageNet validation set and report the remaining results on a 10k validation subset.[3]
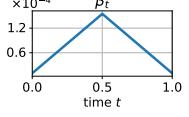
---

[3] https://bit.ly/eval-pix2pix

Table 2. **4× super-resolution** results w.r.t different filters. We report FID and Classifier Accuracy (CA, unit:%) on a pre-trained ResNet50. In all Tables 2 to 5, dark-colored rows denote methods requiring additional information such as corruption operators, as opposed to conditional diffusion models like Palette and our I²SB.

Filter	Method	FID ↓	CA↑
Pool	DDRM (Kawar et al., 2022a)	14.8	64.6
	DDNM (Wang et al., 2022b)	9.9	67.1
	ΠGDM (Song et al., 2022)	3.8	72.3
	ADM (Dhariwal & Nichol, 2021)	3.1	**73.4**
	CDSB (Shi et al., 2022)	13.0	61.3
	I²SB (Ours)	**2.7**	71.0
Bicubic	DDRM (Kawar et al., 2022a)	21.3	63.2
	DDNM (Wang et al., 2022b)	13.6	65.5
	ΠGDM (Song et al., 2022)	3.6	**72.1**
	ADM (Dhariwal & Nichol, 2021)	14.8	66.7
	CDSB (Shi et al., 2022)	13.6	61.0
	I²SB (Ours)	**2.8**	70.7

Table 3. **JPEG restoration** w.r.t different quality factors (QF).

QF	Method	FID-10k ↓	CA↑
5	DDRM (Kawar et al., 2022b)	28.2	53.9
	ΠGDM (Song et al., 2022)	8.6	64.1
	Palette (Saharia et al., 2022)	8.3	64.2
	CDSB (Shi et al., 2022)	38.7	45.7
	I²SB (Ours)	**4.6**	**67.9**
10	DDRM (Kawar et al., 2022b)	16.7	64.7
	ΠGDM (Song et al., 2022)	6.0	71.0
	Palette (Saharia et al., 2022)	5.4	70.7
	CDSB (Shi et al., 2022)	18.6	60.0
	I²SB (Ours)	**3.6**	**72.1**

Table 4. **Inpainting** results w.r.t different masks.

Mask	Method	FID-10k ↓	CA↑
Center 128×128	DDRM (Kawar et al., 2022a)	24.4	62.1
	ΠGDM (Song et al., 2022)	7.3	**72.6**
	DDNM (Wang et al., 2022b)	15.1	55.9
	Palette (Saharia et al., 2022)	6.1	63.0
	CDSB (Shi et al., 2022)	50.5	49.6
	I²SB (Ours)	**4.9**	66.1
Freeform 10%-20%	DDRM (Kawar et al., 2022a)	9.7	67.6
	DDNM (Wang et al., 2022b)	3.2	73.6
	Palette (Saharia et al., 2022)	4.0	73.7
	CDSB (Shi et al., 2022)	8.5	71.2
	I²SB (Ours)	**2.9**	**74.9**
Freeform 20%-30%	DDRM (Kawar et al., 2022a)	8.6	71.9
	ΠGDM (Song et al., 2022)	5.3	**75.3**
	DDNM (Wang et al., 2022b)	4.2	70.8
	Palette (Saharia et al., 2022)	4.1	71.8
	CDSB (Shi et al., 2022)	16.5	64.5
	I²SB (Ours)	**3.2**	73.4

Table 5. **Deblurring** results w.r.t different kernels.

Kernel	Method	FID-10k ↓	CA↑
Uniform	DDRM (Kawar et al., 2022a)	9.9	68.0
	DDNM (Wang et al., 2022b)	**3.0**	**75.5**
	Palette (Saharia et al., 2022)	4.1	74.0
	CDSB (Shi et al., 2022)	15.5	65.1
	I²SB (Ours)	3.9	73.7
Gaussian	DDRM (Kawar et al., 2022a)	6.1	72.5
	DDNM (Wang et al., 2022b)	**2.9**	**75.6**
	Palette (Saharia et al., 2022)	3.1	75.4
	CDSB (Shi et al., 2022)	7.7	71.1
	I²SB (Ours)	3.0	75.0

5.2. Experimental Results

I²SB surpasses standard CSGM on many tasks Tables 2 to 5 summarize the quantitative results on each restoration task. We use the official values reported by each baseline and, if not available, compute them using the official implementations with default hyperparameters, except for Palette on deblurring and inpainting tasks which we implemented by ourselves. I²SB clearly surpasses standard CSGMs such as Palette and ADM on six out of nine tasks, including super-resolution (*Bicubic*), JPEG restoration (for both QFs), and inpainting (for all masks). Despite that ADM and Palette obtain higher CA on super-resolution (*Pool*) and both deblurring tasks, I²SB yields lower, hence better, FID.

I²SB matches DIM without knowing corrupted operators and outperforms standard SB on all tasks Compared to DIM models, I²SB provides a competitive alternative with similar performance yet *without* knowing the corrupted operators during either training or generation. In fact, I²SB achieves state-of-the-art FID on seven out of nine tasks and set new records for CA on JPEG restoration (both

QFs) and inpainting (*Freeform 10-20%*). Finally, I²SB outperforms CDSB on *all* restoration tasks by a large margin. These results highlight I²SB as the *first* nonlinear diffusion model that scales to high-dimensional applications.

I²SB yields interpretable & efficient generation As I²SB directly constructs diffusion bridges between two domains, it generates more interpretable processes that progressively restore the intact images from the degradations; see Figure 7. *More interpretable generation also implies sampling efficiency*. Since the clean and degraded images are typically close to each other, the generation of I²SB starts from a much more structurally informative prior compared to random noise. We validate these concepts in Figures 8 and 9 by tracking how the performance of I²SB and Palette changes as the number of function evaluation (NFE) decreases in sampling. For a fair comparison, we train both models with 1000 discrete steps and sample with DDPM (4) so that they differ mainly in the boundary distributions, *i.e.*, $p_{\mathcal{B}}(\cdot|X_0)$ *vs.* $\mathcal{N}(0, I)$. From Figure 8, we see that across
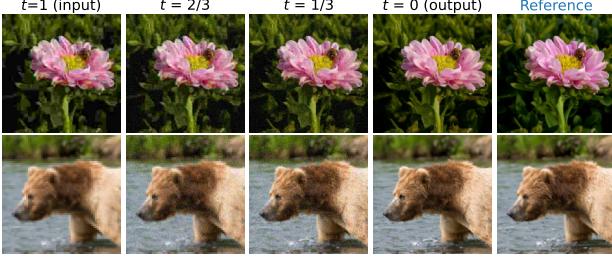
Figure 7. I^{2}SB features more natural and interpretable generative diffusion processes from degraded to clean images. **Top**: JPEG restoration (QF=5). **Bottom**: 4× super-resolution (*Bicubic*).
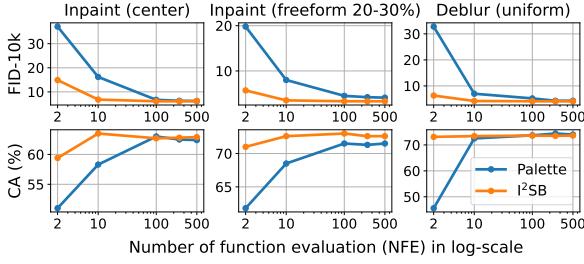

Figure 8. Quantitative comparison between Palette (Saharia et al., 2022) and our I^{2}SB across different NFEs in sampling. I^{2}SB enjoys much smaller performance drops as NFE decreases.

Table 6. How the performance of I^{2}SB improves or degrades with OT-ODE, *i.e.,* by sampling X_t from the mean of $q(X_t|X_0, X_1)$.

| | JPEG restoration | | Deblurring | |
	QF=5	10	*Uniform*	*Gaussian*
FID difference	+5.3	+4.2	-0.3	-0.6
CA difference	-4.7	-3.8	+6.0	+4.1

Figure 9. Qualitative comparison between Palette (Saharia et al., 2022) and our I^{2}SB w.r.t. different NFE on (**top**) inpainting (*Freeform 20%-30%*) and (**bottom**) deblurring (*Uniform*).
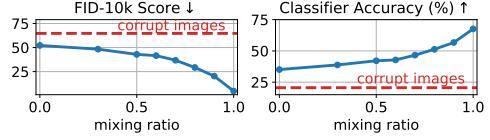

Figure 10. Effect of sampling proposal of X_t on JPEG restoration (QF=5). The x-axis is the mixing ratio between (**left**) the distribution induced by (9a) and (**right**) the posterior $q(X_t|X_0, X_1)$. Both metrics improve as the proposal approaches $q(X_t|X_0, X_1)$.

various tasks, I^{2}SB enjoys much smaller performance drops as NFE decreases. On inpainting (*Freeform 20-30%*), for example, I^{2}SB needs only 2∼10 NFEs while Palette needs at least 100 NFEs to achieve the similar best performance. Qualitatively, Figure 9 also demonstrates that I^{2}SB clearly outperforms Palette in the small NFE regime. Particularly for inpainting, I^{2}SB is able to repaint the masked region with semantic structures with only two NFEs (and further fills in textural details as the NFE increases). On the contrary, Palette tends to generate unnatural images with noisy repainting or contrast shift when the NFE is small.

5.3. Discussions

Sampling proposals I^{2}SB shares much algorithmic similarity with SGM except drawing X_t from an interpolation between clean and degraded images according to $q(X_t|X_0, X_1)$. This posterior differs from the distribution induced by the forward SDE (9a) and, according to Proposition 3.3, better covers regions traversed by the generative processes. To verify this, Figure 10 shows how the perfor-

mance changes when X_t is sampled by mixing these two distributions with different ratios during training. Clearly, both metrics deteriorate as the sampling proposal deviates from $q(X_t|X_0, X_1)$ towards the distribution induced by (9a).

Diffusion *vs.* OT-ODE Table 6 reports the performance difference when we adopt the OT-ODE in Proposition 3.4, *i.e.,* by sampling X_t with the mean of $q(X_t|X_0, X_1)$ in both training and generation. Our result suggests that OT-ODE favors restoration tasks where deterministic mapping is possible (*e.g., deblurring*) yet is biased against those with large uncertainties (*e.g., JPEG restoration*). It reexamines the role of stochasticity in modern dynamic generative models.

General image-to-image translation Since our framework does not impose any assumptions or restrictions on the underlying prior distributions, I^{2}SB can be applied to general image-to-image translation by adopting the same training and sampling procedures (Algorithms 1 and 2), except conditioning the network additionally on the inputs, *i.e.,* $\epsilon(X_t, t, X_1|\theta)$. Aligned with the discussions in Appendix C.3, we found it beneficial when the priors have
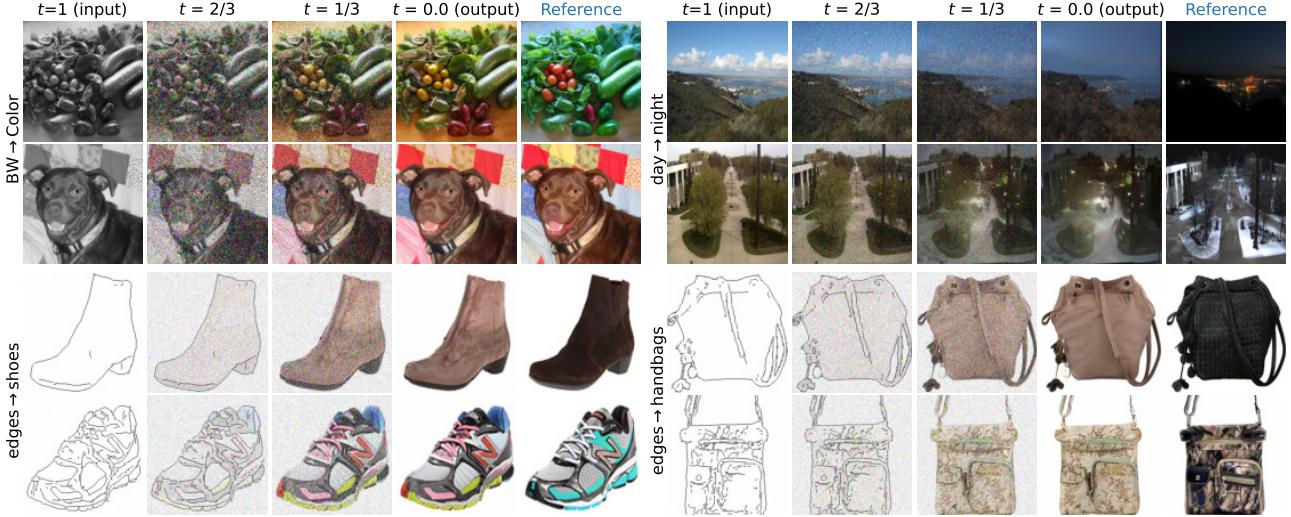
Figure 11. Application of our I^{2}SB to four **general image-to-image translation** tasks from Pix2pix (Isola et al., 2017). We consider ImageNet dataset for the colorization task (BW → Color) and adopt the datasets proposed by Isola et al. (2017) for the remaining three tasks, namely `edges2shoes`, `day2night`, and `edges2handbags`. All images are in 256×256 resolution.

Table 7. Quantitative (FID) results on two general image-to-image translation tasks. Our I^{2}SB matches Pix2pix with only *one* NFE and quickly outperforms it by refining the generation processes.

	Pix2pix	I^{2}SB		
		NFE=1	NFE=5	NFE=1000
edges→shoes	73.9	73.9	54.2	**37.8**
day→night	196.4	196.3	185.8	**153.6**

large information loss. Figure 11 demonstrates the qualitative results, and Table 7 reports the FID w.r.t. the statistics of each validation set. It is clear that our I^{2}SB achieves similar performance to Pix2pix (Isola et al., 2017) with *one* NFE and quickly outperforms it by refining the generation processes. These results highlight the applicability of I^{2}SB to general image-to-image translation tasks.

Comparison to inpainting GANs Table 8 reports the generation quality and efficiency between two inpainting GANs, *i.e.,* DeepFillv2 (Yu et al., 2019) and HiFill (Yi et al., 2020), Palette, and our I^{2}SB. For a fair comparison, we reduce the sampling step of all diffusion models to 1. In other words, "I^{2}SB (NFE=1)" generates images with *one* network call. It is clear that I^{2}SB achieves best generation quality among all models on both tasks. Note that since all models generate images in one network call, the difference in their inference times is mainly due to the network size.

Limitation Despite these encouraging results, the tractability of I^{2}SB requires knowing paired data (*e.g.,* clean and degraded image pairs) during training. While paired data is typically available at nearly no cost, especially for image restoration tasks, it nevertheless limits the application of I^{2}SB to *unpaired* image translation tasks like CycleGAN

Table 8. Comparison between GANs (DFill and HiFill) and our I^{2}SB with *one* NFE on two inpainting tasks. We include Palette for comparison. The inference time is measured on a V100 16G.

Mask	Method	FID ↓ (10k)	CA ↑ (%)	Inference time (sec/image)
Freeform 10%-20%	DeepFillv2	6.7	71.6	**0.01**
	HiFill	7.5	70.1	0.03
	I^{2}SB (NFE=1)	**4.1**	**73.4**	0.14
	Palette (NFE=1)	9.6	69.9	0.14
Freeform 20%-30%	DeepFillv2	9.4	68.8	**0.01**
	HiFill	12.4	65.7	0.03
	I^{2}SB (NFE=1)	**6.7**	**69.9**	0.14
	Palette (NFE=1)	19.8	61.8	0.14

(Zhu et al., 2017) or DDIB (Su et al., 2022). Constructing simulation-free diffusion bridges (like our I^{2}SB) under more flexible setups will be an interesting future direction.

6. Conclusion

We developed I^{2}SB, a new conditional diffusion model that transport between clean and degraded image distributions based on a tractable class of Schrödinger bridge. I^{2}SB yields interpretable generation, enjoys sampling efficiency, and sets new records on image restoration. It will be interesting to combine I^{2}SB with inverse problem techniques.

Acknowledgements

References

Anderson, B. D. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Banham, M. R. and Katsaggelos, A. K. Digital image restoration. *IEEE signal processing magazine*, 14(2):24–41, 1997.

Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

Caluya, K. and Halder, A. Wasserstein proximal algorithms for the schrödinger bridge problem: Density control with nonlinear drift. *IEEE Transactions on Automatic Control*, 2021.

Chen, T., Liu, G.-H., and Theodorou, E. A. Likelihood training of schrödinger bridge using forward-backward sdes theory. *arXiv preprint arXiv:2110.11291*, 2021a.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pp. 6572–6583, 2018.

Chen, Y., Georgiou, T. T., and Pavon, M. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrödinger bridge. *SIAM Review*, 63(2):249–313, 2021b.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022a.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *arXiv preprint arXiv:2206.00941*, 2022b.

Cole, J. D. On a quasi-linear parabolic equation occurring in aerodynamics. *Quarterly of applied mathematics*, 9(3):225–236, 1951.

Dai Pra, P. A stochastic control approach to reciprocal diffusion processes. *Applied mathematics and Optimization*, 23(1):313–329, 1991.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. *arXiv preprint arXiv:2106.01357*, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Evans, L. C. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Fernandes, D. L., Vargas, F., Ek, C. H., and Campbell, N. D. Shooting schrödinger's cat. In *Fourth Symposium on Advances in Approximate Bayesian Inference*, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Hopf, E. The partial differential equation ut+ uux= μxx. *Communications on Pure and Applied mathematics*, 3(3):201–230, 1950.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv preprint arXiv:2206.00364*, 2022.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022a.

Kawar, B., Song, J., Ermon, S., and Elad, M. Jpeg artifact correction using denoising diffusion restoration models. *arXiv preprint arXiv:2209.11888*, 2022b.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022.

Kullback, S. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236–1243, 1968.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings*

of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690, 2017.

Léonard, C. From the schrödinger problem to the monge–kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

Léonard, C. A survey of the schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.

Li, G., Yang, Y., Qu, X., Cao, D., and Li, K. A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 213:106617, 2021.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, G.-H., Chen, T., So, O., and Theodorou, E. A. Deep generalized schr\" odinger bridge. *arXiv preprint arXiv:2209.09893*, 2022.

McCann, R. J. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

Menon, S., Damian, A., Hu, S., Ravi, N., and Rudin, C. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 2437–2445, 2020.

Mikami, T. Monge's problem with a quadratic cost by the zero-noise limit of h-path processes. *Probability theory and related fields*, 129(2):245–260, 2004.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Motwani, M. C., Gadiya, M. C., Motwani, R. C., and Harris, F. C. Survey of image denoising techniques. In *Proceedings of GSPX*, volume 27, pp. 27–30, 2004.

Nelson, E. Dynamical theories of brownian motion. 1967.

Nelson, E. *Dynamical theories of Brownian motion*, volume 106. Princeton university press, 2020.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., and Anandkumar, A. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *arXiv preprint arXiv:2201.00308*, 2022.

Pavon, M. and Wakolbinger, A. On free energy, stochastic control, and schrödinger processes. In *Modeling, Estimation and Control of Systems with Uncertainty*, pp. 334–348. Springer, 1991.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Richardson, W. H. Bayesian-based iterative method of image restoration. *JoSA*, 62(1):55–59, 1972.

Risken, H. Fokker-planck equation. In *The Fokker-Planck Equation*, pp. 63–95. Springer, 1996.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *arXiv preprint arXiv:2104.07636*, 2021.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

Särkkä, S. and Solin, A. *Applied stochastic differential equations*, volume 10. Cambridge University Press, 2019.

Schrödinger, E. *Über die umkehrung der naturgesetze*. Verlag der Akademie der Wissenschaften in Kommission bei Walter De Gruyter u . . . , 1931.

Schrödinger, E. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pp. 269–310, 1932.

Shi, Y., De Bortoli, V., Deligiannidis, G., and Doucet, A. Conditional simulation using diffusion schrödinger bridges. In *Uncertainty in Artificial Intelligence*, pp. 1792–1802. PMLR, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2022.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *arXiv preprint arXiv:2006.09011*, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *arXiv e-prints*, pp. arXiv–2101, 2021a.

Song, Y., Shen, L., Xing, L., and Ermon, S. Solving inverse problems in medical imaging with score-based generative models. *arXiv preprint arXiv:2111.08005*, 2021b.

Su, X., Song, J., Meng, C., and Ermon, S. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Vahdat, A., Kreis, K., and Kautz, J. Score-based generative modeling in latent space. *arXiv preprint arXiv:2106.05931*, 2021.

Vargas, F., Thodoroff, P., Lawrence, N. D., and Lamacraft, A. Solving schrödinger bridges via maximum likelihood. *arXiv preprint arXiv:2106.02081*, 2021.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wallace, G. K. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

Wang, T., Zhang, T., Zhang, B., Ouyang, H., Chen, D., Chen, Q., and Wen, F. Pretraining is all you need for image-to-image translation. *arXiv preprint arXiv:2205.12952*, 2022a.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022b.

Whang, J., Delbracio, M., Talebi, H., Saharia, C., Dimakis, A. G., and Milanfar, P. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16293–16303, 2022.

Yi, Z., Tang, Q., Azizi, S., Jang, D., and Xu, Z. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7508–7517, 2020.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4471–4480, 2019.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, Q. and Chen, Y. Path integral sampler: a stochastic control approach for sampling. *arXiv preprint arXiv:2111.15141*, 2021.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A. Proof

Proof of Theorem 3.1. Recall that the density evolution of an Itô process,

$$\mathrm{d}X_t = f_t(X_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t, \quad X_0 \sim p_0 \qquad (14)$$

can be characterized by the Fokker Plank equation (Risken, 1996),

$$\frac{\partial p(x,t)}{\partial t} = -\nabla \cdot (f_t\, p) + \frac{1}{2}\beta_t\Delta p,\, p(x,0) = p_0(x). \quad (15)$$

Comparing (14, 15) to (9a, 6a) readily suggests that the PDE $\frac{\partial \widehat{\Psi}(x,t)}{\partial t}$ in (6a) can be viewed as the Fokker Plank of the SDE in (9a). The equivalence $\widehat{\Psi} \equiv p^{(9a)}$ holds up to some constant which vanishes upon taking the operator "$\nabla \log$" or in the Fokker Plank equation (since all operators are linear). Similar interpretation can be drawn between the PDE $\frac{\partial \Psi(x,t)}{\partial t}$ and the SDE in (9b) by noticing that (6a) can be read *equivalently* from the reversed time coordinate (Chen et al., 2021a; Liu et al., 2022):

$$\begin{cases} \frac{\partial \Psi(x,s)}{\partial s} = \nabla \cdot (\widehat{\Psi} f_s) + \frac{1}{2}\beta_s\Delta\Psi \\ \frac{\partial \widehat{\Psi}(x,s)}{\partial s} = \nabla\Psi^\top f_s - \frac{1}{2}\beta_s\Delta\widehat{\Psi} \end{cases}, \qquad (16)$$

where $s := 1-t$. This suggests that $\Psi(x,s)$ can be seen as the density (up to some constant) of the SDE

$$\mathrm{d}X_s = -f_s(X_s)\mathrm{d}s + \sqrt{\beta_s}\mathrm{d}W_s, \quad X_0 \sim \Psi(\cdot,0),$$

which equals (9b) after substituting back $t = 1-s$. $\qquad\square$

Proof of Corollary 3.2. It suffices to show that the solutions (10) are consistent with the necessary conditions in (6a), *i.e.,* they are the solutions to the two PDEs with the coupled boundary constraints. Notice that the second PDE $\frac{\partial \widehat{\Psi}(x,t)}{\partial t}$ and the constraint $\Psi(\cdot,1)\widehat{\Psi}(\cdot,1) = p_\mathcal{B}(x)$ are satisfied by construction since $\widehat{\Psi}(\cdot,1)$ is the Fokker-Plank solution w.r.t. the initial condition $\widehat{\Psi}(\cdot,0) = \delta_a(\cdot)$. Hence, it remains to be shown that the solution to the following *backward* PDE

$$\frac{\partial \Psi(x,t)}{\partial t} = -\nabla\Psi^\top f_t - \frac{1}{2}\beta_t\Delta\Psi, \quad \Psi(x,1) = \frac{p_\mathcal{B}(x)}{\widehat{\Psi}(x,1)} \quad (17)$$

satisfies the remaining boundary constraint w.r.t. $p_\mathcal{A}$. Precisely, since $p_\mathcal{A}(x) = \widehat{\Psi}(x,0) = \delta_a(x)$, it suffices to show the solution to (17) being $\Psi(a,0) = 1$, which is indeed the case (Zhang & Chen, 2021). For completeness, Zhang & Chen (2021, Theorem 1) identified that the solution to the Hamilton-Jacobi-Bellman (HJB) equation (Evans, 2010), which relates to (17) via exponential transform (Hopf, 1950; Caluya & Halder, 2021), with the terminal cost $\log\frac{p_\mathcal{B}(x)}{\widehat{\Psi}(x,1)}$ is simply 0. Hence, we know that the solution to (17) is $\Psi(a,0) = \exp(0) = 1$, which concludes the proof. $\qquad\square$

How Corollary 3.2 reduces to SGM. When $p_\mathcal{B} := \widehat{\Psi}(\cdot,1)$ and f is chosen such that the terminal distribution of the forward SDE converges to a Gaussian, *i.e.,* $\widehat{\Psi}(\cdot,1) \approx \mathcal{N}(0,I)$, we have $\Psi(\cdot,1) = 1$ from (10). In fact, we will have $\Psi(\cdot,t) = 1$ for all $t \in [0,1]$ since $\frac{\partial\Psi(x,t)}{\partial t} = 0$. In this case, one can verify that the remaining boundary constraint holds, *i.e.,* $p_\mathcal{A}(\cdot) = \Psi(\cdot,0)\widehat{\Psi}(\cdot,0)$, since we set $p_\mathcal{A}(\cdot) = \widehat{\Psi}(\cdot,0) = \delta_a(\cdot)$. $\qquad\square$

Proof of Proposition 3.3. Equation (11) arises naturally by first conditioning the Nelson's duality (Nelson, 2020), *i.e.,* $q(\cdot,t) = \Psi(\cdot,t)\widehat{\Psi}(\cdot,t)$, on a boundary pair (X_0, X_1),

$$q(X_t|X_0, X_1) = \Psi(X_t,t|X_0)\widehat{\Psi}(X_t,t|X_1).$$

Since $\Psi(X_t,t|X_0)$ and $\widehat{\Psi}(X_t,t|X_1)$ are solutions to Fokker-Plank equations (see the proof of Theorem 3.1), we can rewrite the posterior as the product of two Gaussians:

$$\Psi(X_t,t|X_0)\widehat{\Psi}(X_t,t|X_1)$$
$$= \exp\left(-\frac{1}{2}\left(\frac{\|X_t - X_0\|^2}{\sigma_t^2} + \frac{\|X_t - X_1\|^2}{\bar{\sigma}_t^2}\right)\right)$$
$$= \mathcal{N}\left(X_t; \frac{\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}X_0 + \frac{\sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}X_1, \frac{\sigma_t^2\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}\cdot I\right),$$

where $\sigma_t^2 := \int_0^t \beta_\tau\mathrm{d}\tau$ and $\bar{\sigma}_t^2 := \int_t^1 \beta_\tau\mathrm{d}\tau$ are analytic marginal variances (Särkkä & Solin, 2019) of the SDEs (9) when $f := 0$.

We now prove (by induction) that $q(X_t|X_0, X_1)$ is the marginal density of DDPM posterior $p(X_n|X_0, X_{n+1})$. First, notice that when $f := 0$, $p(X_n|X_0, X_{n+1})$ has an analytic Gaussian form

$$p(X_n|X_0, X_{n+1})$$
$$= \mathcal{N}\left(X_n; \frac{\alpha_n^2}{\alpha_n^2 + \sigma_n^2}X_0 + \frac{\sigma_n^2}{\alpha_n^2 + \sigma_n^2}X_{n+1}, \frac{\sigma_n^2\alpha_n^2}{\alpha_n^2 + \sigma_n^2}\cdot I\right),$$

where we denote $\alpha_n^2 := \int_{t_n}^{t_{n+1}} \beta_\tau\mathrm{d}\tau$ as the accumulated variance between two consecutive time steps (t_n, t_{n+1}). It is clear that at the boundary $t_n := t_{N-1}$, we have

$$q(X_{N-1}|X_0, X_N) = p(X_{N-1}|X_0, X_N)$$

since $\alpha_{N-1} = \int_{t_{N-1}}^{t_N} \beta_\tau\mathrm{d}\tau = \bar{\sigma}_{N-1}^2$. Suppose the relation also holds at t_{n+1}, it suffices to show that

$$q(X_n|X_0, X_N) \qquad\qquad (18)$$
$$\stackrel{?}{=} \int p(X_n|X_0, X_{n+1})q(X_{n+1}|X_0, X_N)\mathrm{d}X_{n+1}.$$

Since both p and q are Gaussians, the RHS of (18) is a Gaussian with the mean (Bishop, 2006)

$$\underbrace{\frac{\alpha_n^2}{\alpha_n^2 + \sigma_n^2}}_{\sigma_{n+1}^2}X_0 + \underbrace{\frac{\sigma_n^2}{\alpha_n^2 + \sigma_n^2}}_{\sigma_{n+1}^2}\left(\underbrace{\frac{\bar{\sigma}_{n+1}^2\, X_0}{\bar{\sigma}_{n+1}^2 + \sigma_{n+1}^2}}_{\bar{\sigma}_n^2 + \sigma_n^2} + \underbrace{\frac{\sigma_{n+1}^2\, X_N}{\bar{\sigma}_{n+1}^2 + \sigma_{n+1}^2}}_{\bar{\sigma}_n^2 + \sigma_n^2}\right)$$

$$= \frac{\alpha_n^2(\bar{\sigma}_{n+1}^2 + \sigma_{n+1}^2) + \sigma_n^2 \bar{\sigma}_{n+1}^2}{\sigma_{n+1}^2(\bar{\sigma}_n^2 + \sigma_n^2)} X_0 + \frac{\sigma_n^2}{\bar{\sigma}_n^2 + \sigma_n^2} X_N$$

$$= \frac{\alpha_n^2 \sigma_{n+1}^2 + \bar{\sigma}_{n+1}^2(\alpha_n^2 + \sigma_n^2)}{\sigma_{n+1}^2(\bar{\sigma}_n^2 + \sigma_n^2)} X_0 + \frac{\sigma_n^2}{\bar{\sigma}_n^2 + \sigma_n^2} X_N$$

$$= \frac{\bar{\sigma}_n^2}{\bar{\sigma}_n^2 + \sigma_n^2} X_0 + \frac{\sigma_n^2}{\bar{\sigma}_n^2 + \sigma_n^2} X_N, \tag{19}$$

where we utilize that $\bar{\sigma}_n^2 + \sigma_n^2$ remains constant for all n and that $\alpha_n^2 = \sigma_{n+1}^2 - \sigma_n^2 = \bar{\sigma}_n^2 - \bar{\sigma}_{n+1}^2$ by construction. Similarly, the RHS of (18) has the covariance

$$\frac{\alpha_n^2 \sigma_n^2}{\alpha_n^2 + \sigma_n^2} + \frac{\bar{\sigma}_{n+1}^2 \sigma_{n+1}^2}{\bar{\sigma}_{n+1}^2 + \sigma_{n+1}^2} \left(\frac{\sigma_n^2}{\alpha_n^2 + \sigma_n^2} \right)^2$$

$$= \frac{\alpha_n^2 \sigma_n^2(\bar{\sigma}_n^2 + \sigma_n^2) + \bar{\sigma}_{n+1}^2 \sigma_n^4}{\sigma_{n+1}^2(\bar{\sigma}_n^2 + \sigma_n^2)}$$

$$= \frac{\sigma_n^2 \left[\alpha_n^2(\bar{\sigma}_n^2 + \cancel{\sigma_n^2}) + (\bar{\sigma}_n^2 - \cancel{\sigma_n^2})\sigma_n^2 \right]}{\sigma_{n+1}^2(\bar{\sigma}_n^2 + \sigma_n^2)} = \frac{\sigma_n^2 \bar{\sigma}_n^2}{\bar{\sigma}_n^2 + \sigma_n^2}. \tag{20}$$

Equations (19) and (20) validate the equality in (18), and we conclude the proof by induction. □

Proof of Proposition 3.4. At the infinitesimal limit when $\beta_t \to 0$, the variance of q, *i.e.*, $\frac{\sigma_t^2 \bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}$, vanishes as the numerator converges faster than the denominator toward zero. On the contrary, its mean remains unchanged as both ratios ($\frac{\bar{\sigma}_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}, \frac{\sigma_t^2}{\bar{\sigma}_t^2 + \sigma_t^2}$) preserve. Hence we know the deterministic solution at the infinitesimal limit is simply $X_t = \mu_t(X_0, X_1)$. In this case, the diffusion of the SDE, *i.e.*, "$\sqrt{\beta_t t} \mathrm{d}W_t$", vanishes while its drift approaches a vector field of the form:

$$\beta_t \nabla \log \widehat{\Psi}(X_t|X_0) = \frac{\beta_t}{\sigma_t^2}(X_t - X_0) := v_t(X_t|X_0).$$

Hence, we have the OT-ODE in (13). □

Proof of Corollary 3.5. When $\beta_t := \beta$ is a sufficiently small constant, the ratio $\frac{\beta_t}{\sigma_t^2}$ decays in the order of $\mathcal{O}(1/t)$ since $\sigma_t^2 = \int_0^t \beta_\tau \mathrm{d}\tau = \beta \cdot t$. With this, Proposition 3.4 yields $\mu_t = (1-t)X_0 + tX_1$ and $v_t = \frac{X_t - X_0}{t}$. Intuitively, the OT-ODE trajectories move with a *constant* velocity from X_1 toward X_0. □

B. Introduction to Schrödinger Bridge

The Schrödinger bridge problem was originally introduced quantum mechanics (Schrödinger, 1931; 1932) and later draws broader interests with its connection to optimal transport (Léonard, 2013; Dai Pra, 1991). The *dynamic* Schrödinger bridge (Pavon & Wakolbinger, 1991; Léonard, 2012) is typically defined as

$$\min_{\mathbb{Q} \in \Pi(p_{\mathcal{A}}, p_{\mathcal{B}})} D_{\mathrm{KL}}(\mathbb{Q}||\mathbb{P}),$$

where $\Pi(p_{\mathcal{A}}, p_{\mathcal{B}})$ is a set of path measure with the marginal densities $p_{\mathcal{A}}$ and $p_{\mathcal{B}}$ at the boundaries. Relating the path measures $\mathbb{Q}$ and $\mathbb{P}$ respectively to some controlled and uncontrolled diffusion processes leads to the following stochastic optimal control (SOC) formulation:

$$\min_{u(X_t, t)} \mathbb{E} \left[\int_0^1 \frac{1}{2} \|u(X_t, t)\| \mathrm{d}t \right]$$
$$\text{s.t. } \mathrm{d}X_t = [f_t(X_t) + u(X_t, t)]\mathrm{d}t + \sqrt{\beta_t} \mathrm{d}W_t \tag{21}$$
$$X_0 \sim p_{\mathcal{A}}, \quad X_1 \sim p_{\mathcal{B}}$$

The programming (21) seeks an optimal control process $u(X_t, t)$ such that the energy cost accumulated over the time horizon $[0, 1]$ is minimized while obeying the distributional boundary constraints. The coupled PDEs in (6a) result directly from applying the Hopf-Cole transform (Hopf, 1950; Cole, 1951) to the necessary conditions to (21). This yields $u^\star(X_t, t) = \beta_t \nabla \log \Psi(X_t, t)$ and hence the SDE in (5a). Similar reasoning applies to (5b), where $\beta_t \nabla \log \widehat{\Psi}(X_t, t)$ serves as the optimal control process to a SOC similar to (21) except running backward in time.

C. Experiment Details

Official Pytorch implementation of our I²SB can be found in https://github.com/NVlabs/I2SB.

C.1. Additional Experimental Setup

Deblurring and JPEG restoration We adopt the implementation of blurring kernels from Kawar et al. (2022a) and the implementation of JPEG quality factor from Kawar et al. (2022b). Following the baselines (Saharia et al., 2022; Song et al., 2022), the FID is evaluated over the reconstruction results on the 10k ImageNet validation subset,[4] and compared against the statistics of the entire ImageNet validation set.

4× super-resolution We adopt the same implementation of filters from DDRM (Kawar et al., 2022a). We first generate 64×64 images then upsample them to 256×256 before passing into I²SB, since the model transports between clean and degraded images of the same size. Following the baselines (Saharia et al., 2022; Song et al., 2022), the FID is evaluated over the reconstruction results on the entire ImageNet validation set, and compared against the statistics of the entire ImageNet training set.

Inpainting We use the same freeform masks provided by Palette (Saharia et al., 2022),[4] which contains 10000 masks for both *10%-20%* and *20%-30%* ratios. We randomly select these masks during training and iterate them through the 10k ImageNet validation subst[4] for reproducible evaluation. We follow the same instructions from Palette and

---

[4] https://bit.ly/eval-pix2pix

Table 9. Additional ablation study on the effect of stochasticity on inpainting tasks. OT-ODE exhibits severe degradation with *noiseless* masks but yields slightly better results after injecting additional noise to the masked regions of degraded inputs.

	mask		mask + noise	
	Center	*Ff. 20-30%*	*Center*	*Ff. 20-30%*
FID diff.	+50.9	+13.0	-0.1	-0.1
CA diff.	-14.1	-7.3	0.0	+0.7

Figure 12. The numerical values of the coefficients $c_t^{\text{in}}, c_t^{\text{skip}}, c_t^{\text{out}}$ adopted in (12, 23) for training $\epsilon(X_t, t; \theta)$, where X_t interpolates between clean and corrupted image pair (X_0, X_1).

set up I²SB such that *(i)* the training loss is restricted to only the masked regions, *(ii)* the masked regions are filled with Gaussian noise as inputs (see Figure 13), and *(iii)* the model predicts only the masked regions during generation. The FID is evaluated over the reconstruction results on the 10k ImageNet validation subset and compared against the statistics of the entire ImageNet validation set.

Evaluation We use `cleanfid` package[5] with the option "`legacy_pytorch`" to compute FID values. For the reference statistics, we take the ones provided by ADM (Dhariwal & Nichol, 2021) for the ImageNet training set and compute the ones for the ImageNet validation set by resizing and center-cropping the images to 256×256, similar to ADM. The Classifier Accuracy is based on a pre-trained ResNet50 (He et al., 2016). Following the suggestions from Saharia et al. (2022), we avoid pixel-level metrics like PSNR and SSIM as they tend to prefer blurry regression outputs (Menon et al., 2020; Ledig et al., 2017).

Palette implementation We implement our own Palette for the results in Tables 4, 5 and 8 and Figures 8 and 9. For all the other tasks, we use the official values reported in their paper. For a fair comparison, we initialize its network of Palette with the same checkpoint from unconditional ADM (Dhariwal & Nichol, 2021) on ImageNet 256×256 and concatenate the first layer with conditional inputs, following Rombach et al. (2022). The SDE uses the same 1000 time steps with quadratic discretization similar to I²SB.

C.2. Additional Qualitative Results

Figures 13 to 16 provide additional qualitative results on each restoration tasks, and Figures 17 to 19 provide additional examples comparing between Palette and I²SB w.r.t. various NFE sampling. Finally, Figure 20 demonstrates that I²SB is able to generate diverse samples.

C.3. Additional Discussions

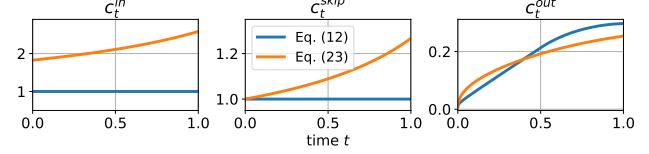**More Ablation Study on OT-ODE** Table 6 shows how OT-ODE seems to disfavor restoration tasks with large un-

---

certainties. We conjecture that it is due to the severe information lost in degraded inputs that hinders the reconstruction of deterministic mapping. This is validated in Table 9, where we compare the performance difference on inpainting tasks with or without injecting Gaussian noise to the masked regions. OT-ODE exhibits severe degradation without any stochasticity but yields comparable results after injecting additional noise to the masked regions of degraded inputs.

Other Parameterization In additional to the standard rescaled score function in (12), we may follow Karras et al. (2022) by considering

$$\|\epsilon(c_t^{\text{in}} \cdot X_t, t; \theta) - \frac{c_t^{\text{skip}} \cdot X_t - X_0}{c_t^{\text{out}}}\|, \qquad (22)$$

where $c_t^{\text{in}}, c_t^{\text{skip}}, c_t^{\text{out}} \in \mathbb{R}$ are time-varying coefficients such that *(i)* the inputs and outputs of ϵ have unit variance and *(ii)* the approximation error induced from ϵ are minimized. In our cases, since X_t now interpolates between clean and corrupted image pairs (rather than images with i.i.d. noises), we re-derive these coefficients in a more general form given estimated $\text{Var}[X_t]$ and $\text{Cov}[X_0, X_t]$:

$$c_t^{\text{in}} = \frac{1}{\sqrt{\text{Var}[X_t]}}, \qquad c_t^{\text{skip}} = \frac{\text{Cov}[X_0, X_t]}{\text{Var}[X_t]},$$
$$c_t^{\text{out}} = \sqrt{\text{Var}[X_0] - \frac{\text{Cov}[X_0, X_t]^2}{\text{Var}[X_t]}}. \qquad (23)$$

These coefficients can be obtained by $\text{Var}[c_{\text{in}} X_t] = 1$ and

$$\text{Var}\left[\frac{c_{\text{skip}} X_t - X_0}{c_{\text{out}}}\right] = 1$$
$$\Rightarrow c_{\text{skip}}^2 \text{Var}[X_t] + \text{Var}[X_0] - 2c_{\text{skip}} \text{Cov}[X_t, X_0] = c_{\text{out}}^2.$$

Choosing c_{skip} such that c_{out}^2 is minimized yields (23). Figure 12 summarizes the difference between (12) and (23). In practice, we find their empirical differences negligible.

Remark C.1 (How (23) recovers Karras et al. (2022)). In the specific case when $X_t := X_0 + \epsilon$, X_0 has variance σ_{data}^2, and ϵ is i.i.d. noise with variance σ^2, we have

$$\text{Var}[X_t] = \sigma_{\text{data}}^2 + \sigma^2$$
$$\text{Cov}[X_0, X_t] = \text{Var}[X_0] + \text{Cov}[X_0, \epsilon] = \sigma_{\text{data}}^2. \qquad (24)$$

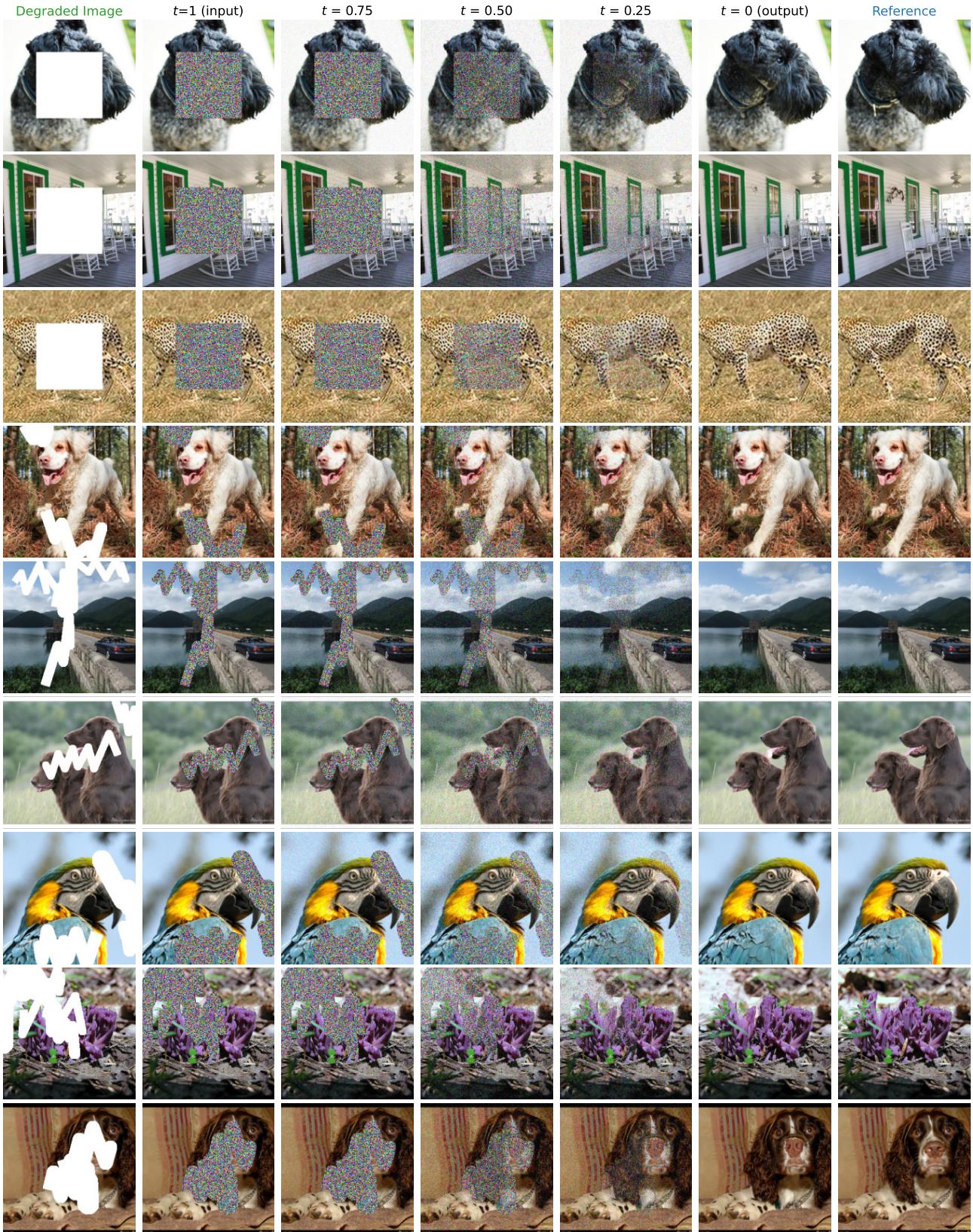Substituting (24) into (23) yields the coefficients suggested in Karras et al. (2022).

Figure 13. Generative processes of I²SB on inpainting tasks. **Top 3 rows**: *Center 128×128* mask. **Middle 3 rows**: *Freeform 10%-20%* mask. **Bottom 3 rows**: *Freeform 20%-30%* mask.

Figure 14. Generative processes of I^{2}SB on JPEG restoration tasks. **Top 5 rows**: QF=5. **Bottom 3 rows**: QF=10.

Figure 15. Generative processes of I²SB on deblurring tasks. **Top 5 rows**: *Uniform* kernel. **Bottom 3 rows**: *Gaussian* kernel.
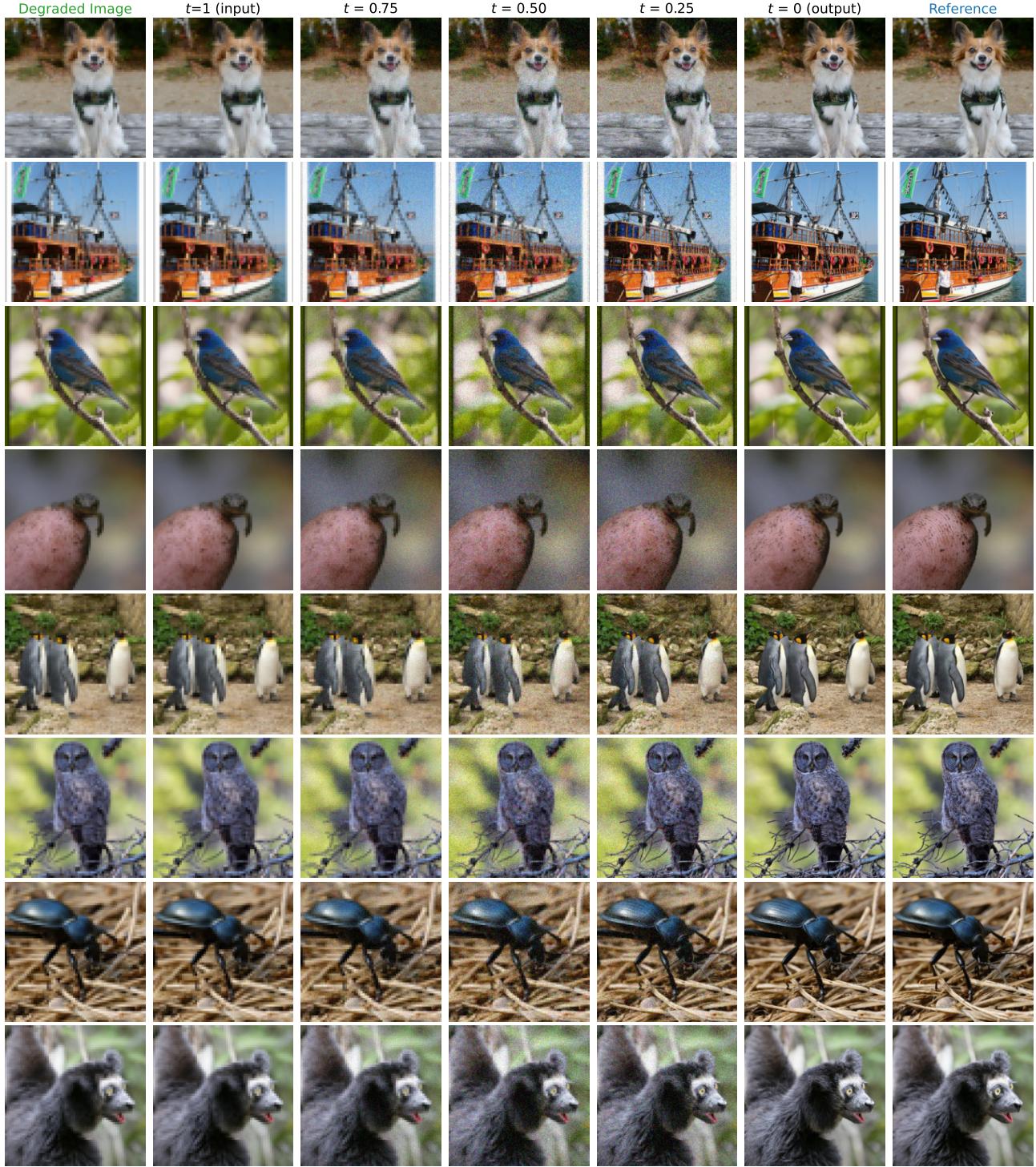
| Degraded Image | $t=1$ (input) | $t = 0.75$ | $t = 0.50$ | $t = 0.25$ | $t = 0$ (output) | Reference |

Figure 16. Generative processes of I^{2}SB on $4\times$ super-resolution tasks. **Top 5 rows**: *Pool* filter. **Bottom 3 rows**: *Bicubic* filter.
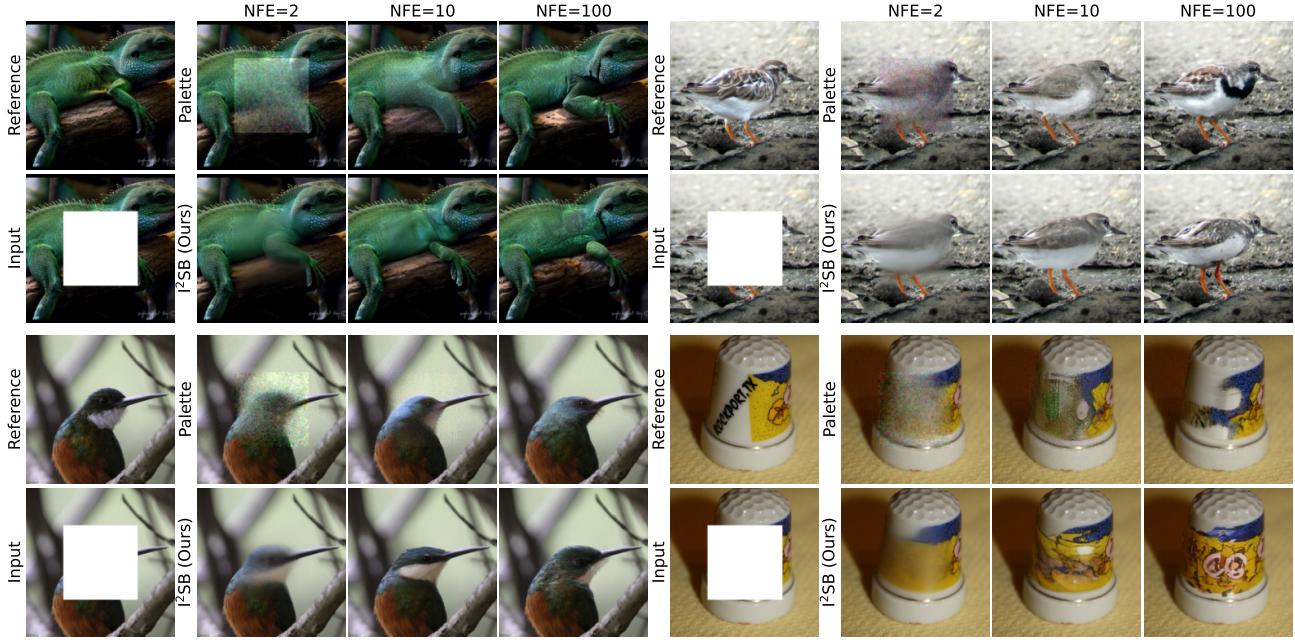
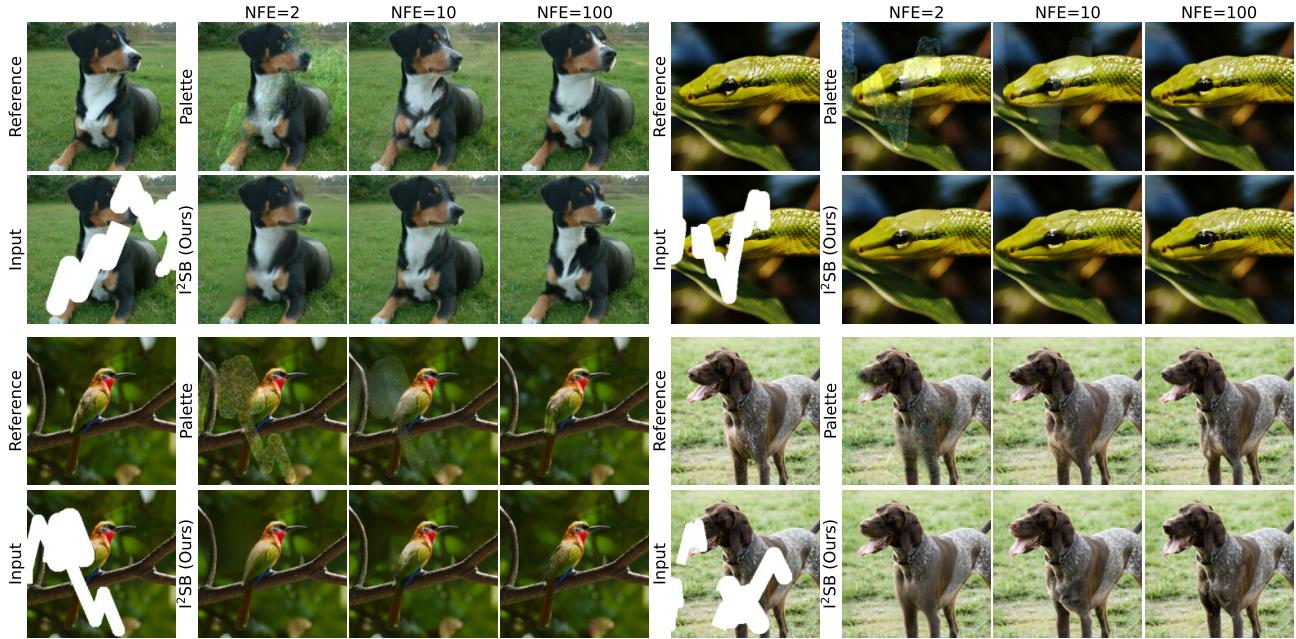*Figure 17.* Additional qualitative comparison between I²SB and Palette on inpainting (*Center 128×128*).

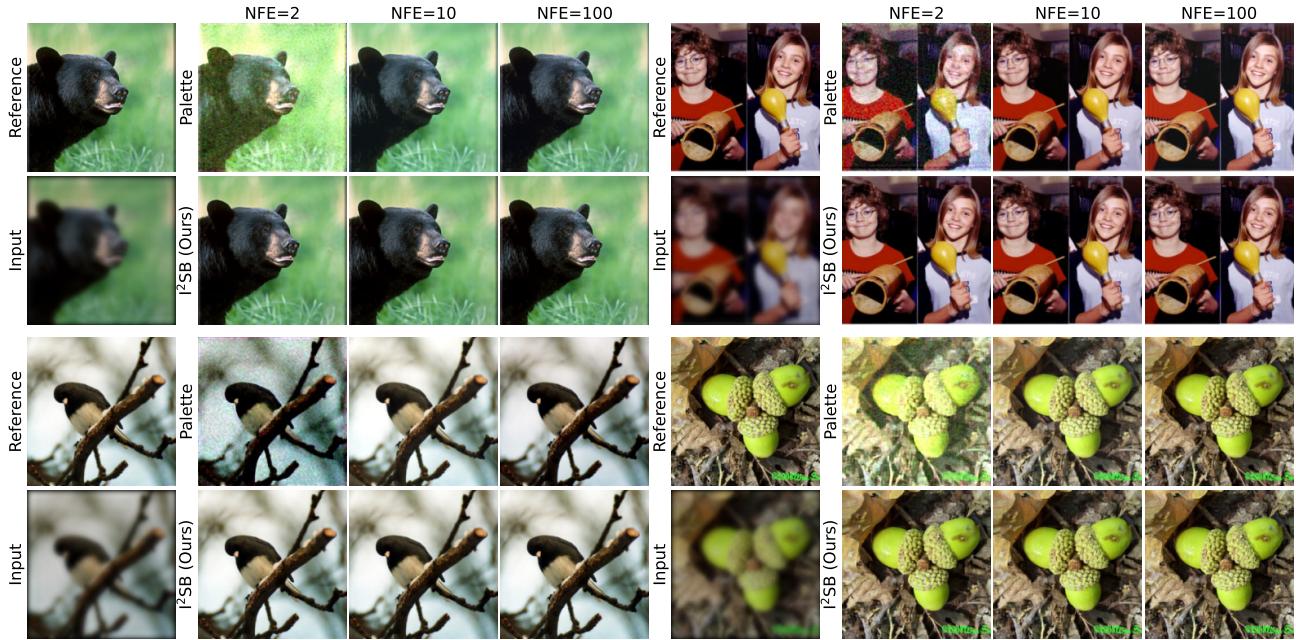*Figure 18.* Additional qualitative comparison between I²SB and Palette on inpainting (*Freeform 20%-30%*).

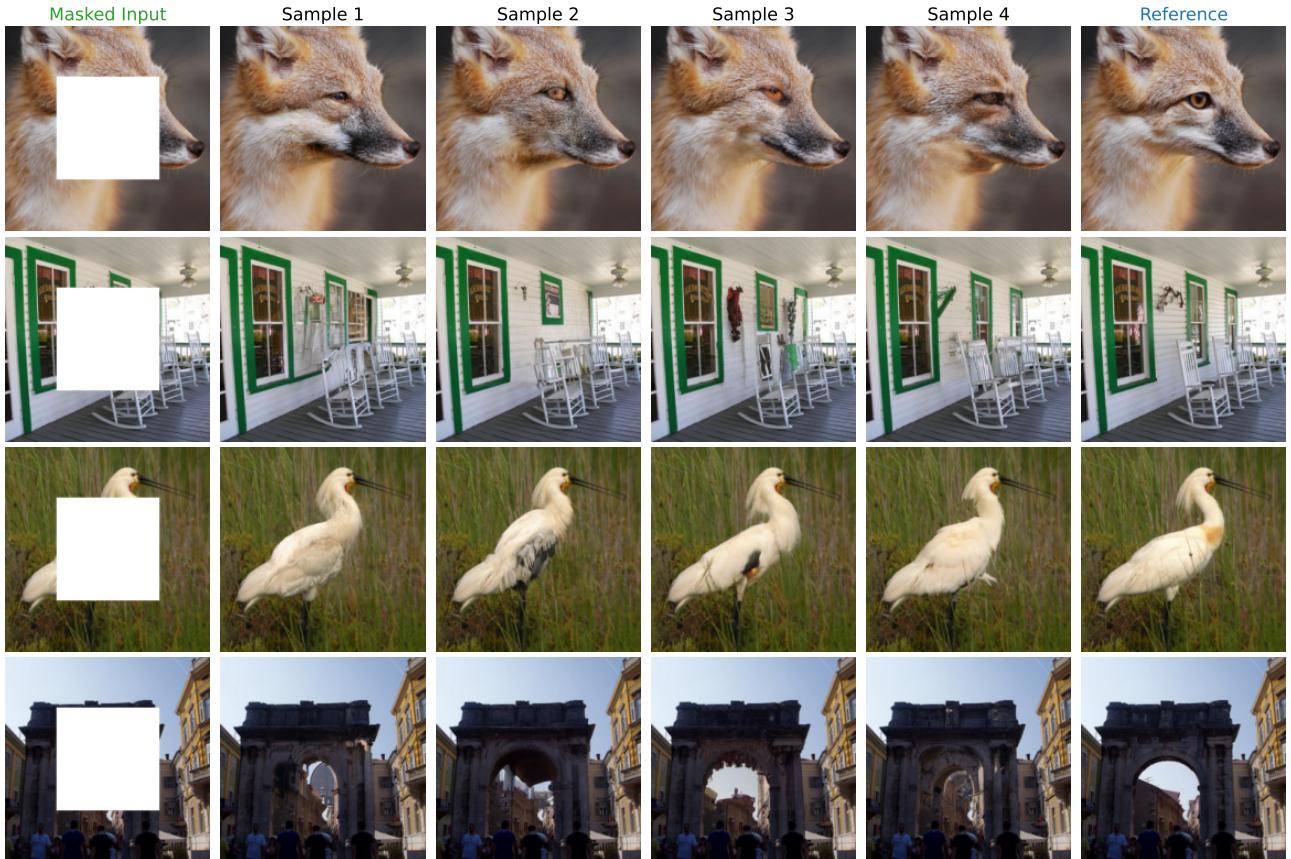*Figure 19.* Additional qualitative comparison between I^{2}SB and Palette on deblurring (*Uniform*).

Figure 20. Diversity of I^{2}SB outputs on inpainting tasks.