GENERATIVE MODELING FROM BLACK-BOX CORRUPTIONS VIA SELF-CONSISTENT STOCHASTIC INTERPOLANTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transport-based methods have emerged as a leading paradigm for building generative models from large, clean datasets. However, in many scientific and engineering domains, clean data are often unavailable: instead, we only observe measurements corrupted through a noisy, ill-conditioned channel. A generative model for the original data thus requires solving an inverse problem at the level of distributions. In this work, we introduce a novel approach to this task based on Stochastic Interpolants: we iteratively update a transport map between corrupted and clean data samples using only access to the corrupted dataset as well as black box access to the corruption channel. Under appropriate conditions, this iterative procedure converges towards a self-consistent transport map that effectively inverts the corruption channel, thus enabling a generative model for the clean data. The resulting method (i) is computationally efficient compared to variational alternatives, (ii) highly flexible, handling arbitrary nonlinear forward models with only black-box access, and (iii) enjoys theoretical guarantees. We demonstrate superior performance on inverse problems in natural image processing and scientific reconstruction, and establish convergence guarantees of the scheme under appropriate assumptions.

1 INTRODUCTION

Generative modeling has become a central tool for learning high-dimensional data distributions. Transport-based methods, including diffusion-based models (Ho et al., 2020; Song et al., 2021) and flow-based models (Albergo & Vanden-Eijnden, 2023; Lipman et al., 2022; Liu et al., 2023) have emerged as leading frameworks for training high-quality generative models, with a wide range of applications from natural image synthesis (Rombach et al., 2022) to molecular design (Watson et al., 2023). These methods rely on access to clean samples $x \sim \pi$ of the target distribution, which are plentiful in many machine learning tasks.

However, in many scientific and engineering applications, such clean data of interest is unavailable. Instead, we only observe corrupted measurements y through a *forward* map $y = \mathcal{F}(x)$ that is typically noisy and ill-conditioned. Examples include medical imaging, where we observe tomographic projections of internal structures, astronomical observations affected by atmospheric distortion, and other measurement processes that introduce noise and information loss (Tarantola, 2005). As a result, the target data x is never observed directly, rendering standard generative modeling inapplicable.

Recent work has begun to tackle this fundamental limitation, aiming to generate clean data x using only corrupted observations y. Most existing approaches, however, require the forward model to be explicitly specified and differentiable, often with additional structural assumptions. For example, Daras et al. (2023); Kawar et al. (2024); Chen et al. (2025); Zhang et al. (2025) train diffusion models with corrupted data under explicit linear forward models and additional rank condition. Akyildiz et al. (2025) learns a generative prior by directly minimizing the sliced-Wasserstein-2 distance between observed data and model outputs. A classical alternative is Empirical Bayes, leading to approaches based on variational inference, such as the EM algorithm in Rozet et al. (2024); Bai et al. (2024), which again depends on a known forward model. In many cases, further restrictions are imposed, such as linearity with Gaussian noise to enable Tweedie's formula for approximate posterior sampling (Daras et al., 2023; Rozet et al., 2024).

In this work, we introduce a framework for this *inverse generative modeling* task using only corrupted observations y and *black-box simulation* of the forward process $\mathcal{F}$. Our approach leverages stochastic interpolants (SI) (Albergo & Vanden-Eijnden, 2023; Albergo et al., 2023) with a self-consistent training procedure: we iteratively transport observed data to clean samples via a learned velocity field, then enforce consistency by requiring that these generated samples, when passed through $\mathcal{F}$, reproduce the original observation distribution. This scheme not only eliminates the need for clean data in generative modeling within the inverse problem settings, but also avoids backpropagation or posterior sampling through $\mathcal{F}$. As a result, our framework applies directly to nonlinear forward models (e.g., motion blur), non-differentiable operators (e.g., JPEG compression), and non-Gaussian noise (e.g., Poisson noise), substantially broadening the applicability of inverse generative modeling. Conceptually, this makes our approach akin to *model-free* reinforcement learning, which optimizes policies through interaction with a simulator, whereas most prior methods resemble *model-based* control, relying on explicit knowledge and differentiability of the underlying physics.

Problem setup We consider a probability distribution of interest $\pi \in \mathcal{P}(\Omega)$, and a forward model $\mathcal{F} : \Omega \to \tilde{\Omega}$ that we allow to be stochastic, i.e., $y = \mathcal{F}(x)$ defines a conditional distribution of y given x on $\tilde{\Omega}$. Some representative examples in $\Omega = \mathbb{R}^d$ include the additive white gaussian noise (AWGN) channel $y = x + \sigma\xi$, with $\xi \sim \gamma_d \equiv \mathcal{N}(0, \mathrm{I}_d)$, or tomography, where $y = M\tilde{x} + \sigma\xi$, $\tilde{x}$ is the Fourier transform of x and M is a certain (possibly random) projection operator along frequency rays.

Since $\mathcal{F}$ is a channel that does not introduce additional information about x, we assume that the observation space $\tilde{\Omega}$ can be embedded back into the data space Ω in a way that preserves all information contained in y. That is, the embedding itself does not introduce any additional information loss beyond what is already incurred by $\mathcal{F}$. With a slight abuse of notation, we therefore redefine $\mathcal{F}$ as a map $\mathcal{F} : \Omega \to \Omega$. We define the kernel $k_{\mathcal{F}}(y, x)$ associated with $\mathcal{F}$ as the conditional distribution of $y = \mathcal{F}(x)$ given x. This channel pushes forward the data distribution π to an observation distribution $\mu \in \mathcal{P}(\Omega)$, given by $\mu = \mathcal{K}_{\mathcal{F}}\pi$, where $\mathcal{K}_{\mathcal{F}}$ is the integral operator with kernel $k_{\mathcal{F}}$, i.e., $\mu(y) = \int k_{\mathcal{F}}(y, x)\,\mathrm{d}\pi(x)$.

The forward model $\mathcal{F}$ is often ill-conditioned, non-deterministic (and therefore non-invertible) as a mapping in Ω, thus justifying the need to regularize the inverse problem of recovering x from the observations $y = \mathcal{F}(x)$. However, the situation is different when viewed at the level of probability measures $\mathcal{P}(\Omega)$: as soon as $\mathcal{K}_{\mathcal{F}}$ is invertible in $\mathcal{P}(\Omega)$, one can hope to recover π from μ by inverting the linear relationship $\mu = \mathcal{K}_{\mathcal{F}}\pi$. To illustrate this point, consider the AWGN channel: while the optimum reconstruction at the level of the samples (in the MSE sense) is given by the posterior mean $\hat{x} = \mathbb{E}[x|\mathcal{F}(x)]$, and generically we have $\mathbb{E}\,\|x - \hat{x}\|^2 > 0$, the associated inverse problem at the level of distributions amounts to a deconvolution, i.e., $\mu = \pi \star \gamma_\sigma$, which is invertible for any noise level σ.

We approach this inverse generative modeling task by first assuming that we have access to μ, either directly, or from a dataset of observations $\{y_i\}_i$, $y_i \sim \mu$ that can be fed into a generative model that produces an estimate $\hat{\mu}$. We also assume only black-box access to a general (potentially nonlinear) forward model $\mathcal{F}$, without requiring its analytical form or gradients.

Additional related works We note growing interest in generative models trained on mixtures of clean and noisy data (Daras et al., 2024; 2025; Lu et al., 2025; Meanti et al., 2025). These approaches assume some clean samples, whereas our setting relies solely on corrupted observations. While our method is tailored for this more constrained regime, it can be applied straightforwardly to settings with partial access to clean data. On the theoretical side, Li et al. (2024; 2025) study inverse problems over measure spaces, analyzing stability, variational structures, and gradient flows. Our work complements these studies by introducing a practical and scalable algorithmic framework while also establishing convergence guarantees under appropriate assumptions.

2 PRELIMINARIES

Standard Stochastic Interpolant (assuming access to clean data) Let π be the clean data distribution we wish to sample from and μ be the distribution of the corrupted data that are available to us, both supported on $\mathbb{R}^d$. Following Albergo et al. (2023), a linear stochastic interpolant I_t between π and μ is defined by

$$I_t = \alpha_t x_0 + \beta_t x_1 + \gamma_t z, \quad t \in [0, 1], \tag{1}$$

where (x_0, x_1) is sampled from a joint distribution (or coupling) $\nu(\mathrm{d}x_0, \mathrm{d}x_1)$ that maintains the marginals $\int_{\mathbb{R}^d} \nu(\cdot, \mathrm{d}x_1) = \pi$, $\int_{\mathbb{R}^d} \nu(\mathrm{d}x_0, \cdot) = \mu$, and $z \sim \gamma_d$ is independent Gaussian noise. The schedules $\alpha_t, \beta_t, \gamma_t$ satisfy boundary conditions $\alpha_0 = \beta_1 = 1$, $\alpha_1 = \beta_0 = 0$, and $\gamma_0 = \gamma_1 = 0$. Define the velocity field

$$b(t, x) \coloneqq \mathbb{E}[\dot{I}_t | I_t = x].$$

The solutions to the probability flow ordinary differential equation (ODE)

$$\dot{X}_t = b(t, X_t) \tag{2}$$

have the property that $X_{t=1} \sim \mu$ if $X_{t=0} \sim \pi$ (forward direction), and $X_{t=0} \sim \pi$ if $X_{t=1} \sim \mu$ (backward direction). The latter enables clean sample generation from the observation distribution by integrating backward using the drift b. The drift b can be learned efficiently in practice by solving the least-squares regression problem

$$\arg\min_{\hat{b}} \int_0^1 \mathbb{E}[|\hat{b}(t, I_t) - \dot{I}_t|^2] \, \mathrm{d}t := \mathcal{E}^b_{\pi,\mu}(\hat{b}) \tag{3}$$

where $\mathbb{E}$ denotes an expectation over the coupling $(x_0, x_1) \sim \nu$ and z.

SI with diffusion The above ODE form for SI can be extended to stochastic differential equation (SDE). Consider another vector-valued function called the *denoiser*

$$g(t, x) \coloneqq \mathbb{E}[z | I_t = x].$$

Similar to the property of the backward ODE, the solutions to the following reverse-time SDE

$$\mathrm{d}X^B_t = b(t, X^B_t)\mathrm{d}t + \epsilon_t \gamma_t^{-1} g(t, X^B_t)\mathrm{d}t + \sqrt{2\epsilon_t}\mathrm{d}W^B_t \tag{4}$$

have the property that if $X^B_{t=1} \sim \mu$ is independent of W^B, then $X^B_{t=0} \sim \pi$. Here $\epsilon_t \in C^0([0, 1])$ with $\epsilon_t \geq 0$ is an arbitrary time-dependent diffusion coefficient, and $W^B_t = -W_{1-t}$. We note (4) is closely related to the SDE in score-based diffusion models: specifically, when $\eta(t) > 0$, we have $s(t, x) := \nabla_x \log \rho(t, x) = -\eta^{-1}(t)\, g(t, x)$, where $\rho(t, x)$ denotes the probability density of I_t. From now on, and for simplicity, we will assume a fixed (i.e., time-independent) diffusion coefficient ϵ. It is straightforward to see when $\epsilon = 0$, the reverse-time SDE becomes ODE. Similar to (3), the denoiser g can be learned efficiently in practice by solving another least-squares regression problem respect to the noise:

$$\arg\min_{\hat{g}} \int_0^1 \mathbb{E}[|\hat{g}(t, I_t) - z|^2] \, \mathrm{d}t := \mathcal{E}^g_{\pi,\mu}(\hat{g}). \tag{5}$$

Notation To simplify notation while covering both ODE and SDE settings, we use Θ to denote the required functions for generative modeling: $\Theta = \{b\}$ in the ODE case, and $\Theta = \{b, g\}$ in the SDE case. Let Φ_Θ denote the backward transport map induced by Θ; that is, $\Phi_\Theta(y) = X_0$ under backward ODE (2) with terminal condition $X_1 = y$ or $\Phi_\Theta(y) = X^B_0$ under reverse-time SDE (4) with terminal condition $X^B_1 = y$. Accordingly, such a transport map induces a pushforward from the observation distribution μ to the clean data distribution, denoted by $\pi_\Theta \coloneqq (\Phi_\Theta)_\# \mu$. Note that in the SDE case, Φ_Θ is a random map due to the Brownian motion, and the pushforward should be interpreted as the expected pushforward, i.e., averaging over the randomness of the Brownian motion. Finally, recall that in (3), we use $\mathcal{E}^b_{\pi,\mu}(\hat{b})$ to denote the loss associated with a candidate drift function $\hat{b}$, defined with respect to the SI between π and μ. Similarly, the objective $\mathcal{E}^g_{\pi,\mu}(\hat{g})$ for the denoiser is defined in (5).

3 Self-Consistent Stochastic Interpolants

In the standard generative modeling setting with direct access to clean data samples $x_0 \sim \pi$ and corrupted samples $x_1 \sim \mu$, one may use, for example, the independent coupling $\nu(\mathrm{d}x_0, \mathrm{d}x_1) = \pi(\mathrm{d}x_0)\mu(\mathrm{d}x_1)$ to construct a Monte Carlo approximation of the expectation in the objective (3)(5). *However, in our inverse problem setting, we only observe corrupted data from μ and lack access to clean samples from π. So it is a priori not obvious how to construct the SI* (1). We now describe how to construct and train a self-consistent SI using only black-box access to the forward map $\mathcal{F}$.

3.1 ITERATIVE SCHEME FOR SELF-CONSISTENCY

Eq (1) provides a natural transport between the observed and clean distribution, but is actionable only when one has sample access to both π and μ. Observe first that if we replace the sample access of μ by oracle access to the forward channel $\mathcal{F}$, we could still build a transport from π to μ by leveraging the fact that $\mu = \mathcal{K}_{\mathcal{F}}\pi$. Indeed,

$$I_t = \alpha_t x + \beta_t \mathcal{F}(x) + \gamma_t z, \quad t \in [0,1], \quad x \sim \pi,\ z \sim \gamma_d,\ x \perp z, \tag{6}$$

defines a valid interpolation between π and μ, and can be directly sampled for training the optimal vector functions Θ^*. For a generic measure ρ replacing π in (6), we denote by b_ρ and g_ρ their associated velocity and denoiser fields.

Observe that the associated backward transport map $\Phi^* := \Phi_{\Theta^*}$ can be used to push observations from μ toward clean samples from π, effectively defining a *local inverse* of the channel, in the sense that $\mathcal{K}_{\mathcal{F}}\Phi^*_{\#}\mu = \mu$[1]. In other words, Φ^* specifies a *self-consistency* condition on the observation probability domain; see Fig. 1.

However, there is a crucial difference in our setup: rather than accessing $\{\pi, \mathcal{F}\}$, we instead have acccess to $\{\mu, \mathcal{F}\}$. The key idea is to turn the self-consistency equation $\mathcal{K}_{\mathcal{F}}\Phi^*_{\#}\mu = \mu$ into a procedure that adjusts Θ to push $\mathcal{K}_{\mathcal{F}}(\Phi_\Theta)_{\#}\mu$ back to μ. We use SIs to connect each of these two distributions to a common 'empirical prior'[2] $\pi_\Theta := (\Phi_\Theta)_{\#}\mu$. Consistency is then enforced by bringing the two SIs close to each other, leading to a natural bi-level fixed-point iteration scheme; see Alg. 1. The outer loop updates $\Theta^{(k)}$ to $\Theta^{(k+1)}$ by constructing, at each step k, the following SI

$$I_t^{(k+1)} = \alpha_t \Phi_{\Theta^{(k)}}(y) + \beta_t \mathcal{F}(\Phi_{\Theta^{(k)}}(y)) + \gamma_t z, \quad t \in [0,1],\ y \sim \mu,\ z \sim \gamma_d,\ y \perp z. \tag{7}$$

This SI is directly sampleable given $\Theta^{(k)}$ and samples from μ, and we train it using standard SI loss (3)(5) via stochastic gradient descent as the inner loop to obtain $\Theta^{(k+1)}$:

$$\Theta^{(k)} \xrightarrow[\text{'E' step}]{\text{via (7)}} I_t^{(k+1)} \xrightarrow[\text{'M' step}]{\text{minimizers in (3)(5) with } I_t^{(k+1)}} \Theta^{(k+1)}. \tag{8}$$

We remark that this bi-level scheme resembles the EM-type algorithm in Rozet et al. (2024); Bai et al. (2024), where the clean data is updated in the outer loop and the score function is retrained in the inner loop. However, their method requires an explicitly linear forward model and relies on uncontrollable approximations to posterior sampling. In contrast, our data-driven backward transport map does not rely on these assumptions and enables learning the SI in Eq. (7) with only black-box access to $\mathcal{F}$.

We easily verify that as soon as the channel is injective, π is the only admissible fixed point of our iterative scheme (see proof in App. A.1), similarly as the consistency guarantees in Daras et al. (2024). In Section 4 we will show that by making additional assumptions beyond injectivity, one can establish unconditional convergence guarantees for our scheme.

Proposition 1 (π is the only admissible fixed point)**.** *Assume that $\mathcal{K}_{\mathcal{F}}$ is injective and that the iterative scheme* (8) *converges to a fixed point Θ^*. Then $\pi_{\Theta^*} = \pi$.*

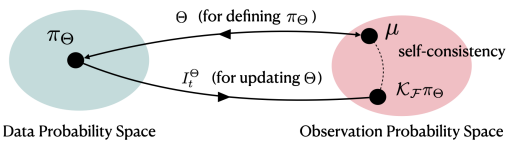


Figure 1: Schematic of the method: the fixed point Θ^* satisfies $\mathcal{K}_{\mathcal{F}}\pi_{\Theta^*} = \mu$, which in turns implies $\pi_{\Theta^*} = \pi$. No samples from π are required—the approach only uses corrupted samples from μ and the map $\mathcal{F}$.

Algorithm 1: Training of Self-Consistent SI

$\Theta \leftarrow \Theta^{(0)}$ `// Initialize`
for k *in* $1 \dots K$ **do**
  for i *in* $1 \dots T_{\text{tr}}$ **do**
    Sample I_t in (7) with $\Theta^{(k)}$
    SGD update of Θ via losses (3)(5)
  $\Theta^{(k)} \leftarrow \Theta$ `// Update transport map`
return $\Theta^{(K)}$

[1] In contrast to a *global inverse*, which would require $\mathcal{K}_{\mathcal{F}}\Phi_{\#}\nu = \nu$ for *all* $\nu \in \mathcal{P}(\Omega)$, a much stronger condition.

[2] In the spirit of Empirical Bayes.

3.2 Truncated Inner-Loop Optimization for Efficiency

For the sake of efficiency, in practice, we do not solve the inner problem to convergence at each step k; instead, we initialize the parameters from $\Theta^{(k)}$ and update them for T_{tr} gradient steps. See Alg. 1 for a description of the resulting algorithm and Alg. 2 in App. B for more details. In the special case $T_{\text{tr}} = 1$, the algorithm is equivalent to treating I_t in (7) as dependent on Θ but applying `stop-gradient` to I_t when computing the gradient of the corresponding loss function. Nevertheless, we retain the two-loop formulation in Alg.1 to emphasize the more general form and better match the bi-level scheme introduced earlier.

4 Theoretical Analysis

In this section we analyze the iterative scheme from Sec. 3.1 and establish convergence in KL divergence to the ground truth distribution focusing in the SDE setting $\epsilon > 0$, by exploiting a contraction property of the scheme. For that purpose, we first introduce two key assumptions that control (i) the stability of the SI drift to initial conditions, and (ii) the condition number of the inverse problem at the distribution level. We focus on the idealised continuous-time limit, and leave time discretization aspects for future work. In the following, we will often use a reference L^2 metric in the space of flows $C([0,1] \times \Omega; \Omega)$, given by $\|b\|^2_{\pi_{[0,1]}} := \mathbb{E}_{t\sim\text{Unif}[0,1]}\mathbb{E}_{x\sim\pi_t}[\|b(t,x)\|^2]$, where π_t is the law of the oracle SI defined in (6). For notational simplicity, we use $\mathcal{K}$ to denote $\mathcal{K}_{\mathcal{F}}$, throughout this section without risk of ambiguity. We also define $\pi^{(k)} := (\Phi_{\Theta^{(k)}})_{\#}\mu$, the estimate of data distribution at (outer loop) iteration k. While we initially introduced the denoiser g, we will henceforth mainly use the score s in analysis for convenience, noting that the two are equivalent.

4.1 Assumptions

Condition Number An important aspect of the problem is that there are two distinct notions of error, whether it is measured on the 'data' side, i.e., $\text{KL}(\pi||\pi^{(k)})$, or on the 'observation' side, i.e., $\text{KL}(\mu||\mu^{(k)}) = \text{KL}(\mathcal{K}\pi||\mathcal{K}\pi^{(k)})$. Since the learner only has access to data from μ, a necessary condition to guarantee that we can recover the original data distribution is *injectivity*, i.e., that $\text{KL}(\mathcal{K}\pi||\mathcal{K}\hat{\pi}) = 0$ implies $\pi = \hat{\pi}$. However, this is not sufficient to provide a quantitative estimate of $\text{KL}(\pi||\hat{\pi})$ in terms of $\text{KL}(\mu||\mathcal{K}\hat{\pi})$. In other words, the inverse problem $\mathcal{K}\pi = \mu$ is generally singular in $\mathcal{P}(\Omega)$, even for the simplest channels, due to the infinite-dimensional nature of the domain.

To mitigate this issue, we need to regularize this inverse problem by restricting (or penalizing) the domain of possible velocities and scores arising from the SI objectives (3)(5), so that the resulting constrained optimization returns $\hat{b}_{\pi^{(k)}}$, $\hat{s}_{\pi^{(k)}} \in \mathcal{B}_\lambda$, where $\mathcal{B}_\lambda$ is indexed by a complexity measure λ, e.g., neural networks with $O(\lambda^{-1})$ parameters. In turn, these regularised objectives inject regularity in $\pi^{(k)}$, i.e., for all k we have $\pi^{(k)} \in \mathcal{S}_\lambda$, the class of terminal densities obtained by running a Fokker-Plank equation with drifts in $\mathcal{B}_\lambda$. We can now consider the *condition number* of $\mathcal{K}$ around π:

$$\chi := \sup_{\rho\in\mathcal{S}_\lambda} \frac{\text{KL}(\pi||\rho)}{\text{KL}(\mathcal{K}\pi||\mathcal{K}\rho)} \,. \tag{9}$$

We verify in Appendix A.2 that χ is well-defined. Note that by the data-processing inequality, we always have $\chi \geq 1$. The (regularised) inverse problem becomes non-singular whenever $\chi < \infty$. The purpose of regularisation, in this abstract context, is to restrict the range $\mathcal{S}_\lambda$ as to make χ small, while maintaining a small approximation error.

A particularly simple form of regularisation is to consider a continuous parametric model $\{b_\omega, s_\omega\}$ where $\omega \in \mathcal{D}$ is in a *compact* domain, which encompasses most practical setups. Combined with the injectivity of the channel, this allows us to have $\chi < \infty$. For technical reasons, we consider the misspecified setting:

Proposition 2 (Finite condition number for compact hypothesis class)**.** *Assume that $\mathcal{K}$ is injective, that $\mathcal{D}$ is a compact parameter space, with continuous parametrization of the drift and score models, and that π cannot be exactly represented by the model. Then $\chi < \infty$.*

Unsuprisingly, under such general conditions, we are unable to quantify the condition number. We expand the condition number properties in Appendix A.2.

Lipschitz stability of SI Recall our definition of b_π, s_π, the velocity and score associated with the SI in (6). In order to control the contraction of our iterative scheme, we will assume that the function $\pi \mapsto f_\pi := b_\pi + \epsilon s_\pi$, that maps the candidate data model π to the drift of a Fokker-Plank equation transporting π to $\mathcal{K}\pi$, is Lipschitz with respect to the KL divergence:

$$\forall \pi, \tilde{\pi} \ , \ \|f_\pi - f_{\tilde{\pi}}\|^2_{\pi_{[0,1]}} \leq L \mathrm{KL}(\pi||\tilde{\pi}) \ . \tag{10}$$

We denote by $L = L_{\mathcal{K},\epsilon}$ its Lipscthiz constant. In words, a SI builds a diffusion bridge between π and $\mathcal{K}\pi$, and L measures the sensitivity of its drift to initial conditions. Notice that this Lispschitz constant depends on the design of the SI, and could potentially be lowered by going beyond the usual linear interpolants, and 'preconditioning' to $\mathcal{K}$; this strategy is beyond the scope of this work and is left for future exploration.

4.2 Contraction in KL divergence

Our main result is the following:

Theorem 1 (Contraction in KL)**.** *Assume $\epsilon > 0$ and that SI satisfies (10). Let χ be the condition number of the regularized channel. Let $\delta^{(k)} = \max(\|b^{(k)} - \hat{b}^{(k)}\|_{\pi_{[0,1]}}, \|s^{(k)} - \hat{s}^{(k)}\|_{\pi_{[0,1]}})$ be the error incurred at iteration k, and assume that $\delta^{(k)} \leq \delta$ for all k. Then, if $L < 4\epsilon\chi^{-1}$, we have*

$$\mathrm{KL}(\pi||\pi^{(k)}) \leq 2(1 + \frac{L}{4\epsilon} - \chi^{-1})^k \mathrm{KL}(\pi||\pi^{(0)}) + O(\delta^2) \ . \tag{11}$$

The proof is in Appendix A.4, and exploits explicit KL inequalities of Fokker-Plank equations. Instrumental to the contraction is the ability to relate errors in measurement space back to data space — precisely what is enabled by the condition number. An interesting interpretation of Theorem 1 is that it provides *global convergence* guarantees for a seemingly complex non-convex objective function, given in (14), by replacing the ubiquitous gradient descent strategy with a tailored 'Picard-type' iterative scheme. In that sense, our guarantees go beyond the qualitative results of the self-consistency loss in Daras et al. (2024). The upper bound (11) captures the typical tradeoff between approximation and estimation errors: a 'small' function class has a smaller condition number, which improves the contraction rate, but in turn causes the error δ to increase (the proof provides explicit error dependencies in δ). That said, a quantitative analysis of this tradeoff in specific function classes is beyond the scope of this work, but an interesting question deserving further attention.

Remark 2 (Stability)**.** *Theorem 1 shows that the scheme is stable to estimation errors of the drift and score. However, notice that the error is measured on a path distribution $(\pi_t)_{t\in[0,1]}$ different from the training distribution $(\pi_t^{(k)})_{t\in[0,1]}$, and we rely on a uniform guarantee across all iterations. In that sense, the quantity δ captures an out-of-distribution error which is more stringent than the typical Fisher stability bounds in generative diffusion literature (Chen et al., 2022; Benton et al., 2023). Finally, we remark that if one has access to an estimate $\hat{\mu}$ rather than μ, the scheme pays an additional $O(\mathrm{KL}(\mu||\hat{\mu}))$ additive term, following a standard data-processing argument.*

Contraction in Fokker-Plank Channels An interesting class of channels where the previous result is more explicit is given by *Fokker-Plank Channels* (Wibisono et al., 2017). These are channels where the forward map $\mathcal{F}$ can be expressed as a diffusion process itself; in other words, the law of $\mathcal{F}(x)$ given x agrees with the law of X_1, where X_t solves

$$\mathrm{d}X_t = f(t, X_t)\mathrm{d}t + \sqrt{2\epsilon}\mathrm{d}W_t \ , \quad X_0 = x \ , \tag{12}$$

for some well-posed drift f. In this case, if the Fokker-Plank representation of the channel is known, we can replace the linear SI in (6) by (12). A prominent example of a Fokker-Plank channel is the AWGN channel, where $f \equiv 0$. Now, observe that in this case the drift of the forward process does not depend on initial conditions, thus $L = 0$. We then immediately obtain:

Corollary 1 (Contraction for Fokker-Plank channels)**.** *Under the same hypothesis as in Theorem 1, using the Fokker-Plank interpolant (12) yields a KL exponential contraction with rate $1 - \chi^{-1}$.*

5 Experiments

We apply our method to a variety of forward models across three settings: (i) synthetic low-dimensional datasets, (ii) imaging tasks, and (iii) a scientific application in quasar spectral recovery.

In some tasks, a latent variable M associated with $\mathcal{F}$ is observed, such as the random mask accompanying each observation in the masking task. In such cases, we additionally condition the vector fields on M. This procedure is fully compatible with our framework: it is equivalent to appending M to the observation, redefining the forward map as $\mathcal{F}(x) = (y, M)$, and keeping the corresponding channels of the SI constant with value M. More implementation details are provided in App. C

5.1 Low dimensional synthetic models

We use this setting to compare between the ODE and SDE formalism. We take the two-moon dataset for the true data distribution and consider the AWGN channel, i.e., $\mathcal{F}$ as the corruption with Gaussian noise of fixed variance σ_n. In Fig. 2, we show the results for a high noise ($\sigma_n = 1.0$) and low noise ($\sigma_n = 0.5$) setup. For low or intermediate noise, both formalisms give similar results. However, for the high noise case, ODE restoration collapses into artificially thin arms for the two moons while SDE results remain stable.

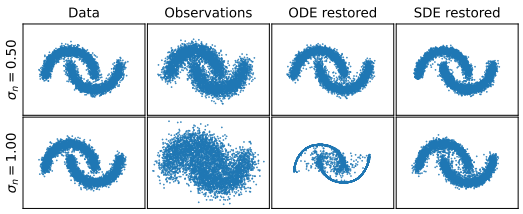


Figure 2: AWGN channel: Comparing ODE and SDE restoration for different noise levels (σ_n).

We use these observations to guide our large scale experiments. While the SDE formalism is more robust for highly corrupting forward models, and a positive diffusion coefficient $\epsilon > 0$ is necessary in our contraction results (see Theorem 1), both approaches work well for moderate corruptions in practice. On the other hand, the SDE approach is computationally more expensive as it requires training two networks and converges more slowly. Thus we investigated ODE formalism on the following high-dimensional examples. We broadly found them to be sufficient and present these results next.

5.2 Imaging Tasks

Setup We use CIFAR-10 dataset (and CelebA dataset for JPEG compression) as the clean data distribution π and generate one observation y per image with the forward model. We model the velocity field b in our SI with the U-net from Dhariwal & Nichol (2021), but using only 64 channels resulting in $\sim$32 million parameters[3]. We train all networks for 50,000 iterations, which required $\sim$54 GPU hours on A100.

DPS Baseline We quantify the quality of the restored samples from our trained SI by evaluating the LPIPS metric (Zhang et al., 2018) and compare it with DPS approach from Chung et al. (2022), a popular and strong inverse solution based on diffusion models. DPS requires a pre-trained diffusion model on the original dataset to solve the inverse problem. Hence we train a large diffusion model with similar U-net architecture but 96 channels instead of 64 ($\sim$ 70 million parameters), which achieved an FID of 5.16. For every inverse problem, we also did a grid search to select the best guidance strength hyperparameter as we found the good values to be very different from the recommendations in the original paper. Hence, this baseline has four advantages over our approach: i) most importantly, it *uses the clean data to train a generative model*, ii) it requires gradients of the forward map unlike our black-box only access, iii) our implementation uses a 2x larger neural network, and iv) benefits from a task-specific hyperparamter search.

i) Random masking Following Daras et al. (2023); Rozet et al. (2024), this map generates an observation y by masking each pixel of an image x independently with probability ρ, and adding isotropic Gaussian noise (σ_n). As in their setting, we assume access to the mask M for each y and use it to condition our SI. We also pre-process the observations by adding independent standard Gaussian noise to masked pixels as it improves the final results. We show the restored images in Fig. 3a and LPIPS metric for different levels of added noise (σ_n) in Table 1. Our restored samples are comparable to DPS in the low noise case but better in the high noise case. We find this to be the case in other examples as well.

[3]Specifically, we use the implementation here.

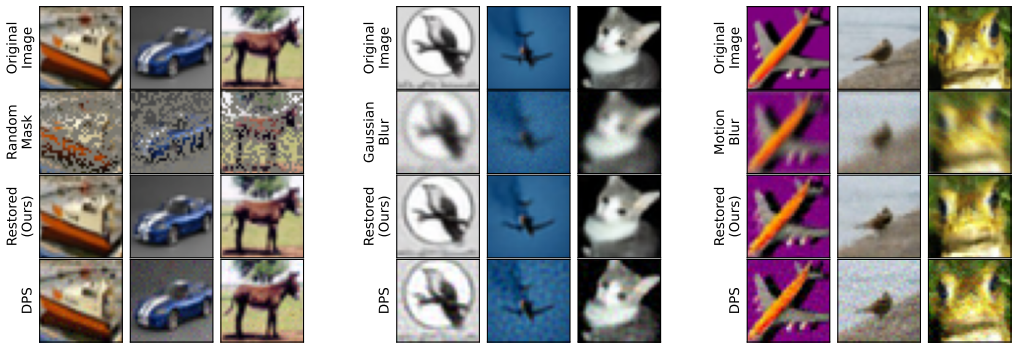
Table 1: LPIPS metric comparing restoration quality of our SI and DPS. Unlike our approach, DPS requires access to clean samples for pre-training and gradients of the forward map for sampling.

Forward Model	Ours	DPS
Random Mask ($\rho = 0.5, \sigma_n = 10^{-6}$)	0.0051	0.0049
Random Mask ($\rho = 0.5, \sigma_n = 0.1$)	0.0064	0.0072
Gaussian Blur ($\sigma_R = 1.0, \sigma_n = 0.1$)	0.005	0.009
Gaussian Blur ($\sigma_R = 1.0, \sigma_n = 0.25$)	0.015	0.025
Motion Blur ($\sigma_R = 1.0, \sigma_n = 10^{-6}$)	0.0072	0.0026
Motion Blur ($\sigma_R = 1.0, \sigma_n = 0.1$)	0.011	0.012

Table 2: FIDs for random masking with different masking probabilities ρ. To account for differing architectures, 'baseline' is FID for our model on the clean CIFAR-10 data.

Method	ρ	FID
Ambient Diffusion	0.20	11.70
	0.40	18.85
EM Posterior	0.25	5.88
	0.50	6.76
Ours (generated)	0.25	**5.38**
	0.50	**6.74**
Baseline	0.00	5.16

To compare with inverse generative models from prior work, we use the trained SI to restore the observations; that is, we transport all observations y to the data space via $\Phi_\Theta(y)$, and use these samples to train a generative diffusion model. We use the same architecture as above, but with 96 channels. Table 2 shows the FID scores for observations with two different masking probabilities and negligible added Gaussian noise ($\Sigma = 10^{-6}$). Our method vastly outperforms Ambient Diffusion (Daras et al., 2023). It is comparable with EM Posterior method (Rozet et al., 2024), but *more computationally efficient*: we required a combined $\sim$ 86 GPU hours (54 and 32 GPU hours to train SI and diffusion model respectively) compared to their 512 GPU hours.



(a) $\mathcal{F}$: Random mask ($\rho = 0.5$, $\sigma_n = 10^{-6}$).

(b) $\mathcal{F}$: Gaussian blur + noise ($\sigma_R = 1, \sigma_n = 0.1$).

(c) $\mathcal{F}$: Motion blur + noise (kernel size $= 5$, $\sigma_n = 0.1$)

Figure 3: Restored samples for different forward maps from our interpolants and DPS.

ii) Gaussian blurring with noise The forward map is blurring with a Gaussian kernel with $\sigma_R = 1$ and adding noise. Here we add Gaussian noises with different levels ($\sigma_n = 0.10, 0.25$) and show the results for Poisson noise case in the appendix. This demonstrates that, unlike previous works, e.g., Daras et al. (2023), our approach can handle non-negligible and non-Gaussian noise.

iii) Motion blurring The previous two examples involve linear forward maps. We now consider a nonlinear one: motion blur. Fig. 3c shows restored samples for observations with a 5-pixel motion kernel and small Gaussian noise ($\Sigma = 10^{-6}$). The blur direction is randomly assigned per image and assumed known for conditioning the SI. While Daras et al. (2023); Rozet et al. (2024) are limited to linear operators, our method handles nonlinear maps with only black-box access.

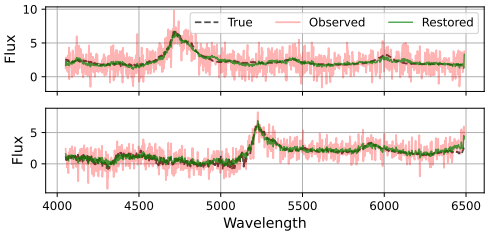
iv) JPEG compression This is another common non-linear corruption operator with real-world applications. The forward map is JPEG compression with quality factor (q) and added Gaussian noise ($\sigma_n = 0.01$). For training, we corrupt every image randomly with a different factor $q \sim \mathcal{U}[0.1, 1]$ (where $q = 1$ implies no compression) and assume this latent parameter q is known for each observation to condition the SI. Fig. 4 shows the restored image with our trained SI for different strengths of compression. The restored image gets closer to the original with lower compressions, and we restore a physically plausible image even for $q = 0.1$. In Appendix D.4, we show that our trained SI is stable to extrapolations of q outside the training regime.

Figure 4: JPEG + noise: results for different compression level (Top: Corrupted; Bottom: Restored).

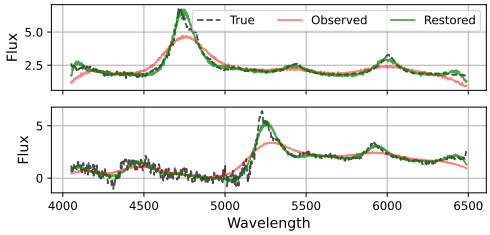
5.3 Quasar Spectra

Scientists observe quasars through telescopes and hence the observed spectra differ from the underlying true spectra due to noise, finite spectral resolution and finite observation time. Recovering the true spectra from these observations is of interest to both study individual objects and to understand the evolution of quasars as a whole.

For the true data distribution, we take the quasar spectra from Sloan Digital Sky Survey data release (Lyke et al., 2020). We isolate 30,000 quasars in redshift $z \in [2.75, 3.25]$ and consider $\lambda \in [400\text{nm}, 650\text{nm}]$ resulting in spectra of length $D = 1024$. We approximate the forward model with a combination of flux calibration error (offset), a Gaussian smoothing, and added Gaussian noise. Unlike imaging examples, for every observation, we randomly vary the size of the smoothing kernel within 5% and add noise with a different magnitude depending on a randomly chosen SNR. We assume we do not have access to these latent parameters of any observation. In Fig. 5, we show the restored spectra for observations in the two extreme regimes that different telescopes operate in: i) observations with high spectral resolution and low SNR (high noise), and ii) those with low spectral resolution and high SNR. In both cases, the restored spectra have much more accurate features (like peak heights and locations), which are used to determine the position and metallic composition of different quasars.



(a) $\mathcal{F}$: High spectral resolution, low SNR
$\Delta \log \lambda = \frac{1}{5000} \pm 5\%$, SNR $\in [0.5, 2]$



(b) $\mathcal{F}$: Low spectral resolution, high SNR
$\Delta \log \lambda = \frac{1}{50} \pm 5\%$, SNR $\in [30, 50]$

Figure 5: Restored quasar spectra for different observational scenarios.

6 Conclusion

We presented a self-consistent SI framework for reconstructing the underlying data distribution using only corrupted observations and a black-box channel. The proposed bi-level iterative scheme is computationally practical and enjoys provable convergence under suitable assumptions. Compared to existing approaches, our method accommodates a much broader class of nonlinear forward models. Experimentally, we demonstrated its effectiveness across a range of inverse problems, achieving competitive performance even against methods that rely on additional access to the forward model (e.g., Ambient Diffusion) or even clean data (e.g., DPS).

Looking ahead, the framework can be naturally combined with large latent-variable models and used to provably model posterior distributions via Föllmer processes Chen et al. (2024), both of which are promising directions for future exploration. On the theoretical side, an important limitation of our current analysis is that it only concerns the SDE setting with $\epsilon > 0$; a natural next step is to explore contraction properties in Wasserstein, which could then capture the ODE variant. Finally, another direction is to quantify the condition number in representative channels for explicit choices of regularization, which would provide further insight and practical guidance for algorithm design.

REFERENCES

O Deniz Akyildiz, Mark Girolami, Andrew M Stuart, and Arnaud Vadeboncoeur. Efficient prior calibration from indirect data. *SIAM Journal on Scientific Computing*, 47(4):C932–C958, 2025.

Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR 2023 Conference*, 2023.

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

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

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. An expectation-maximization algorithm for training clean diffusion models from corrupted observations. *Advances in Neural Information Processing Systems*, 37:19447–19471, 2024.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 2002.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

Nicholas M Boffi and Eric Vanden-Eijnden. Probability flow solution of the fokker–planck equation. *Machine Learning: Science and Technology*, 4(3):035012, 2023.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Tianyu Chen, Yasi Zhang, Zhendong Wang, Ying Nian Wu, Oscar Leong, and Mingyuan Zhou. Denoising score distillation: From noisy diffusion pretraining to one-step high-quality generation. *arXiv preprint arXiv:2503.07578*, 2025.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and f\" ollmer processes. *arXiv preprint arXiv:2403.13724*, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36:288–313, 2023.

Giannis Daras, Alex Dimakis, and Constantinos Costis Daskalakis. Consistent diffusion meets Tweedie: Training exact ambient diffusion models with noisy data. In *Forty-first International Conference on Machine Learning*, 2024.

Giannis Daras, Adrian Rodriguez-Munoz, Adam Klivans, Antonio Torralba, and Constantinos Daskalakis. Ambient diffusion omni: Training good models with bad data. *arXiv preprint arXiv:2506.10038*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Weinan E, Chao Ma, and Lei Wu. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Bahjat Kawar, Noam Elata, Tomer Michaeli, and Michael Elad. GSURE-based diffusion model training with corrupted data. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Qin Li, Maria Oprea, Li Wang, and Yunan Yang. Stochastic inverse problem: stability, regularization and wasserstein gradient flow. *arXiv preprint arXiv:2410.00229*, 2024.

Qin Li, Maria Oprea, Li Wang, and Yunan Yang. Inverse problems over probability measure space. *arXiv preprint arXiv:2504.18999*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

Haoye Lu, Qifan Wu, and Yaoliang Yu. Stochastic forward–backward deconvolution: Training diffusion models with finite noisy datasets. In *Forty-second International Conference on Machine Learning*, 2025.

Brad W. Lyke, Alexandra N. Higley, J. N. McLane, Danielle P. Schurhammer, Adam D. Myers, Ashley J. Ross, Kyle Dawson, Solène Chabanier, Paul Martini, Nicolás G. Busca, Hélion du Mas des Bourboux, Mara Salvato, Alina Streblyanska, Pauline Zarrouk, Etienne Burtin, Scott F. Anderson, Julian Bautista, Dmitry Bizyaev, W. N. Brandt, Jonathan Brinkmann, Joel R. Brownstein, Johan Comparat, Paul Green, Axel de la Macorra, Andrea Muñoz Gutiérrez, Jiamin Hou, Jeffrey A. Newman, Nathalie Palanque-Delabrouille, Isabelle Pâris, Will J. Percival, Patrick Petitjean, James Rich, Graziano Rossi, Donald P. Schneider, Alexander Smith, M. Vivek, and Benjamin Alan Weaver. The Sloan Digital Sky Survey Quasar Catalog: Sixteenth Data Release. *Astrophysical Journal Supplement Series*, 250(1):8, September 2020. doi: 10.3847/1538-4365/aba623.

Giacomo Meanti, Thomas Ryckeboer, Michael Arbel, and Julien Mairal. Unsupervised imaging inverse problems with diffusion distribution matching. *arXiv preprint arXiv:2506.14605*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. *Advances in Neural Information Processing Systems*, 37:87647–87682, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*. SIAM, 2005.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Andre Wibisono, Varun Jog, and Po-Ling Loh. Information and estimation in fokker-planck channels. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2673–2677. IEEE, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yasi Zhang, Tianyu Chen, Zhendong Wang, Ying Nian Wu, Mingyuan Zhou, and Oscar Leong. Restoration score distillation: From corrupted diffusion pretraining to one-step high-quality generation. *arXiv preprint arXiv:2505.13377*, 2025.

A PROOFS

For notational simplicity, we use $\mathcal{K}$ to denote $\mathcal{K}_{\mathcal{F}}$, the forward integral operator that pushes π to μ, and we adopt this shorthand throughout this section without risk of ambiguity.

A.1 PROOF OF PROPOSITION 1

Proof. Recall our iterative scheme as

$$\Theta^{(k)} \xrightarrow[\text{'E' step}]{\text{via (7)}} I_t^{(k+1)} \xrightarrow[\text{'M' step}]{\text{minimizers in (3)(5) with } I_t^{(k+1)}} \Theta^{(k+1)}. \tag{13}$$

If the above iteration converges to a fixed point Θ^*, and the channel is injective at the level of distributions, i.e., $\mathcal{K}\tilde{\pi} = \mathcal{K}\pi$ implies $\tilde{\pi} = \pi$, then the corresponding transport map Φ_{Θ^*} transports corrupted samples from μ into clean samples from π. To see this, consider $\pi_{\Theta^*} \coloneqq (\Phi_{\Theta^*})_{\#}\mu$. We prove only the SDE case, as the ODE case corresponds to the special case when $\epsilon = 0$. By definition of π_{Θ^*} and the property of time-reversal SDE, Θ^* transports π_{Θ^*} to μ under the forward SDE (Anderson, 1982; Song et al., 2021)

$$\mathrm{d}X_t^F = b(t, X_t^F)\mathrm{d}t - \epsilon_t g(t, X_t^F)\mathrm{d}t + \sqrt{2\epsilon_t}\mathrm{d}W_t.$$

On the other hand, since Θ^* is the optimal solution trained from the SI between π_{Θ^*} and $\mathcal{K}\pi_{\Theta^*}$, the above forward SDE also transport samples from π_{Θ^*} to $\mathcal{K}\pi_{\Theta^*}$ (Albergo et al., 2023). As a result we must have $\mathcal{K}_{\mathcal{F}}\pi_{\Theta^*} = \mu$, which means that $\pi_{\Theta^*} = \pi$ thanks to injectivity. ■

Loss function perspective Our iterative scheme can be viewed as a specific procedure to find a fixed point Θ^* satisfying self-consistency. Alternatively, such a fixed point can be characterized as a minimizer of a loss function that penalizes discrepancies between two transport descriptions. Given a generic pair $\Theta = \{b, g\}$ of drift and denoiser models, the corresponding backward transport defines a distribution $\pi_\Theta \coloneqq (\Phi_\Theta)_{\#}\mu$, and then the objectives associated with the SI between π_Θ and $\mathcal{K}\pi_\Theta$ defines minimizers b_{π_Θ} and g_{π_Θ}. We seek to align them with the original pair via the loss

$$\mathcal{L}(b, g) = \|b - b_{\pi_{\{b,g\}}}\|^2 + \|g - g_{\pi_{\{b,g\}}}\|^2 , \tag{14}$$

where $\|\cdot\|$ here denotes an L^2 with respect to an arbitrary base measure. The main challenge when analyzing gradient-based optimization of this loss is the highly non-linear dependencies arising from the transport map.

A.2 ADDITIONAL DETAILS ON CONDITION NUMBER

An important aspect of the problem is that there are two distinct notions of error, whether it is measured on the 'data' side, i.e., $\mathrm{KL}(\pi||\pi^{(k)})$, or on the 'observation' side, i.e., $\mathrm{KL}(\mu||\mu^{(k)}) = \mathrm{KL}(\mathcal{K}\pi||\mathcal{K}\pi^{(k)})$. Since the learner only has access to data from μ, a necessary condition to guarantee that we can recover the original data distribution is *injectivity*, i.e., that $\mathrm{KL}(\mathcal{K}\pi||\mathcal{K}\hat{\pi}) = 0$ implies $\pi = \hat{\pi}$. However, this is not sufficient to provide a quantitative estimate of $\mathrm{KL}(\pi||\hat{\pi})$ in terms of $\mathrm{KL}(\mu||\mathcal{K}\hat{\pi})$. In other words, the inverse problem $\mathcal{K}\pi = \mu$ is generally singular in $\mathcal{P}(\Omega)$, even for the simplest channels, due to the infinite-dimensional nature of the domain. Regularisation is thus necessary.

For that purpose, we modify the SI objectives (3)(5) with a regularised objective:

$$\hat{b}_\pi = \arg\min_{\hat{b}} \mathcal{E}^b_{\pi,\mathcal{K}\pi}(\hat{b}) + \lambda\mathcal{R}(\hat{b}) \ , \ \hat{g}_\pi = \arg\min_{\hat{g}} \mathcal{E}^g_{\pi,\mathcal{K}\pi}(\hat{g}) + \lambda\mathcal{R}(\hat{g}) \ , \ \hat{s}_\pi(t,x) = -\gamma(t)^{-1}\hat{g}_\pi \ . \tag{15}$$

Here, the term $\mathcal{R}$ enforces some type of regularity (e.g., a RKHS norm (Aronszajn, 1950), or a Barron-type norm (Barron, 2002; E et al., 2022)) in the solution within $L^2(\pi_{[0,1]})$. Indeed, if we assume that $\mathcal{F}$ is bounded in L^2, i.e., $\mathbb{E}[\|\mathcal{F}(X)\|^2] \le C_1\mathbb{E}[\|X\|^2] + C_2$, then the minimisers $\hat{b}_\pi$, $\hat{g}_\pi$ in (15) are guaranteed to satisfy $\max(\mathcal{R}(\hat{b}_\pi), \mathcal{R}(\hat{g}_\pi)) \le \lambda^{-1}\max(\mathbb{E}\|\dot{I}_t\|^2, \mathbb{E}[|z|^2])$, and therefore

$$\hat{b}_\pi, \ \hat{s}_\pi \in \mathcal{B}_\lambda = \{f; \ \mathcal{R}(f) \le (\tilde{C}_1\mathbb{E}_\pi[\|X\|^2] + \tilde{C}_2)\lambda^{-1}\} \ . \tag{16}$$

In turn, these regularised objectives inject regularity in $\pi^{(k)}$, in the sense that for all k we have $\pi^{(k)} \in \mathcal{S}_\lambda$, the class of terminal densities obtained by running a Fokker-Plank equation with drifts in $\mathcal{B}_\lambda$. To simplify the technical analysis, and without sacrificing much generality, we will assume that $\pi \notin \mathcal{S}_\lambda$ for any $\lambda > 0$ [4]. We can now quantify the *condition number* of $\mathcal{K}$ 'centered' at π:

$$\chi := \sup_{\rho \in \mathcal{S}_\lambda} \frac{\mathrm{KL}(\pi||\rho)}{\mathrm{KL}(\mathcal{K}\pi||\mathcal{K}\rho)} \,. \tag{17}$$

Note that by the data-processing inequality, we always have $\chi \geq 1$. The (regularised) inverse problem becomes non-singular whenever $\chi < \infty$. The purpose of regularisation, in this context, is to restrict the range $\mathcal{S}_\lambda$ as to make χ small, while maintaining a small approximation error; this tradeoff will be made explicit next. Observe that, if $\rho \in \mathcal{S}_\lambda$, by Girsanov's theorem we have $\mathrm{KL}(\pi||\rho) \leq \epsilon^{-1}\|b^* - \hat{b}_\rho\|^2 + \epsilon\|s^* - \hat{s}_\rho\|^2 < \infty$, which shows that χ is well-defined.

A.3 Proof of Proposition 2

We restate the result for convenience:

Proposition 3 (Finite Condition number for Compact Hypothesis Class)**.** *Assume that $\mathcal{K}$ is injective, that $\mathcal{D}$ is a compact parameter space, with continuous parametrization of the drift and score models, and that π cannot be exactly represented by the model. Then $\chi < \infty$.*

Proof. Let $F : \mathcal{D} \to \mathcal{P}(\Omega)$ be the function that maps a model $\{b_\Theta, s_\Theta\}$ to $F(\Theta) = \pi_1$, where $(\pi_t)_t$ is the marginal law of $(X_t)_t$, which solves the SDE

$$dX_t = (b_\theta(t, X_t) + 2\epsilon s_\theta(t, X_t))dt + \sqrt{2\epsilon}dW_t \,, \tag{18}$$

$$X_0 \sim \mu \,. \tag{19}$$

Define $G(\Theta) := \mathrm{KL}(\mathcal{K}\pi||\mathcal{K}F(\Theta))$. We claim that G is positive for all $\Theta \in \mathcal{D}$ and that G is lower semi-continuous. Indeed, since we are assuming a misspecified model, we have $\mathrm{KL}(\pi||F(\Theta)) > 0$ for all $\Theta \in \mathcal{D}$, which implies $G(\Theta) > 0$ for all $\Theta \in \mathcal{D}$ thanks to the injectivity of $\mathcal{K}$.

Moreover, the mapping $\nu \mapsto \mathrm{KL}(\mu||\nu)$ is lower semi-continuous in the weak topology. This follows from the Donsker-Varadhan variational representation of the KL divergence:

$$\mathrm{KL}(\mu||\nu) = \sup_{f \in \mathcal{C}_b} \left\{ \langle f, \mu \rangle - \log\langle e^f, \nu \rangle \right\} \,.$$

The map $\nu \mapsto -\log\langle e^f, \nu\rangle$ is weakly continuous for all $f \in \mathcal{C}_b$, and the supremum of continuous functions is lower semicontinuous. Now, consider any sequence $(\Theta_n)_n$ such that $\|\Theta_n - \Theta\| \to 0$ as $n \to \infty$. By Girsanov's theorem, observe that

$$\mathrm{KL}(F(\Theta)||F(\Theta_n)) \leq \epsilon^{-1}\|b_\Theta - b_{\Theta_n}\|^2 + \epsilon\|s_\Theta - s_{\Theta_n}\|^2 \,, \tag{20}$$

which shows that $\mathrm{KL}(F(\Theta)||F(\Theta_n)) \to 0$ as $n \to \infty$ thanks to the continuity of the mappings $\Theta \mapsto \{b, g\}_\Theta$. By Pinsker's inequality, we also have that $\|F(\Theta) - F(\Theta_n)\|_{\mathrm{TV}} \to 0$, which shows that $F(\Theta_n)$ converges weakly to $F(\Theta)$, and therefore

$$\liminf_{n\to\infty} G(\Theta_n) \geq G(\Theta) \,, \tag{21}$$

showing that G is LSC as claimed.

Now, observe that

$$\mathrm{KL}(\pi||F(\Theta)) \leq \epsilon^{-1}\|b_\Theta - b^*\|^2 + \epsilon\|s_\Theta - s^*\|^2 := J(\Theta) \,,$$

and

$$\frac{\mathrm{KL}(\pi||F(\Theta))}{\mathrm{KL}(\mu||\mathcal{K}F(\Theta))} \leq \frac{J(\Theta)}{G(\Theta)} := r(\Theta) \,. \tag{22}$$

The function r is the ratio between a continuous function and a positive, lower semicontinuous function. It follows that r is upper semi-continuous, and therefore

$$\chi \leq \sup_{\Theta \in \mathcal{D}} r(\Theta) < \infty \,,$$

since USC functions attain a maximum over compact sets. ■

[4] That is, we assume we are in the more general misspecified setting; this is to avoid degeneracies in the definition of the condition number where both numerator and denominator can be zero.

A.4 Proof of Theorem 1

Proof. The strategy of the proof is to establish a comparison between $\mathrm{KL}(\pi||\pi^{(k)})$ and $\mathrm{KL}(\pi||\pi^{(k+1)})$ by exploiting the relationship between the diffusion bridges that relate them.

For that purpose, let I_t^* be the *oracle* SI, given by

$$I_t^* = \alpha_t X + \beta_t \mathcal{F}(X) + \gamma_t z \ , \ X \sim \pi^* . \tag{23}$$

Let π_t be the law of I_t^*. It solves the Fokker-Planck equation

$$\begin{aligned} \partial_t \pi_t &= \nabla \cdot ((-b^* - \epsilon s^*)\pi_t) + \epsilon \Delta \pi_t \ , \\ \pi_0 = \pi \ , \ \pi_1 &= \mathcal{K}\pi = \mu \ \ , \end{aligned} \tag{24}$$

where

$$\begin{aligned} b^*(t,x) &:= \mathbb{E}[\dot{I}_t^* \,|\, I_t^* = x] \ , \\ s^*(t,x) &:= -\,\mathbb{E}[\gamma_t^{-1} z \,|\, I_t^* = x] \ , \end{aligned} \tag{25}$$

as well as the reverse Fokker-Planck equation

$$\begin{aligned} \partial_t \pi_t &= \nabla \cdot ((-b^* + \epsilon s^*)\pi_t) - \epsilon \Delta \pi_t \ , \\ \pi_1 = \mathcal{K}\pi = \mu \ \ , \ \ \pi_0 &= \pi \ . \end{aligned} \tag{26}$$

Consider also the SI at iteration k of our algorithm. Given $\pi^{(k)}$, we consider the interpolant

$$I_t^{(k)} = \alpha_t X + \beta_t \mathcal{F}(X) + \gamma_t z \ , \ X \sim \pi^{(k)} \ , \tag{27}$$

its associated (exact) drift and scores

$$\begin{aligned} b^{(k)}(t,x) &:= \mathbb{E}[\dot{I}_t^{(k)} \,|\, I_t^{(k)} = x] \ , \\ s^{(k)}(t,x) &:= -\,\mathbb{E}[\gamma_t^{-1} z \,|\, I_t^{(k)} = x] \ , \end{aligned} \tag{28}$$

as well as the estimated drifts and scores, that we recall are given by

$$\hat{b}^{(k)} = \arg\min_{\hat{b}} \mathcal{E}^b_{\pi^{(k)},\mathcal{K}\pi^{(k)}}(\hat{b}) + \lambda \mathcal{R}(\hat{b}) \ , \ \hat{s}^{(k)} \arg\min_{\hat{s}} \mathcal{E}^s_{\pi^{(k)},\mathcal{K}\pi^{(k)}}(\hat{s}) + \lambda \mathcal{R}(\hat{s}) \quad . \tag{29}$$

They define respectively a forward Fokker-Planck equation

$$\begin{aligned} \partial_t \pi_t &= \nabla \cdot ((-b^{(k)} - \epsilon s^{(k)})\pi_t) + \epsilon \Delta \pi_t \ , \\ \pi_0 = \pi^{(k)} \ , \ \pi_1 &= \mathcal{K}\pi^{(k)} = \mu^{(k)} \ \ , \end{aligned} \tag{30}$$

and a reverse Fokker-Planck equation

$$\begin{aligned} \partial_t \pi_t &= \nabla \cdot ((-\hat{b}^{(k)} + \epsilon \hat{s}^{(k)})\pi_t) - \epsilon \Delta \pi_t \ , \\ \pi_1 = \mu \ \ , \ \ \pi_0 &:= \pi^{(k+1)} \ . \end{aligned} \tag{31}$$

It is also useful to define $f := b + \epsilon s$ to be the total drift of the forward (i.e., from data to measurements) diffusion; with the corresponding oracle f^*, iterate $f^{(k)}$ and estimated $\hat{f}^{(k)}$ versions defined analogously. From (24), (26), (30) and (31) we immediately verify that the reverse drift becomes $-f + 2\epsilon s$.

The following lemma relates the rate of KL along two SDEs. We reproduce the proof later for completeness, but it is a known result, e.g., (Boffi & Vanden-Eijnden, 2023, Proposition 1) or (Albergo et al., 2023, Lemma 2.22):

Lemma 1 (KL divergence along two diffusion processes)**.** *Let $dX_t = b(t, X_t)dt + \sqrt{2\sigma}dW_t$ and $dY_t = a(t, Y_t)dt + \sqrt{2\sigma}dW_t$ be two diffusions, and μ_t, ν_t denote the marginal law of X_t and Y_t respectively. Then*

$$\frac{d}{dt}\mathrm{KL}(\mu_t||\nu_t) = -\sigma \mathrm{I}(\mu_t||\nu_t) + \mathbb{E}_{\mu_t}\left\langle b - a, \nabla \log \mu_t - \nabla \log \nu_t \right\rangle \ , \tag{32}$$

where $\mathrm{I}(\mu||\nu) = \mathbb{E}_\mu[\|\nabla \log \mu - \nabla \log \nu\|^2]$ is the Fisher divergence.

In the particular setting where $b = a$, one obtains a de Bruijn identity:

Lemma 2 (de Bruijn Identity)**.**

$$\frac{d}{dt}\mathrm{KL}(\pi_t||\pi_t^{(k)}) = -\sigma \mathrm{I}(\pi_t||\pi_t^{(k)}) \,. \tag{33}$$

Besides a control of the marginal KL, we will also use Girsanov's theorem to obtain control of the KL divergence between path measures of $(X_t)_t$ and $(Y_t)_t$:

Lemma 3 (Girsanov Theorem)**.** *Let $dX_t = b(t, X_t)dt + \sqrt{2\sigma}dW_t$ and $dY_t = a(t, Y_t)dt + \sqrt{2\sigma}dW_t$ be two diffusions, and let $\mu_{[0,T]}$ and $\nu_{[0,T]}$ be the path measures of X_t and Y_t, respectively. Assume the Novikov integrability condition. Then*

$$\mathrm{KL}(\mu_{[0,T]}||\nu_{[0,T]}) = \mathrm{KL}(\mu_0||\nu_0) + \frac{1}{4\sigma}\,\mathbb{E}_{\mu_{[0,T]}} \int_0^T \|a(t,x) - b(t,x)\|^2 dt \,. \tag{34}$$

By the data processing inequality, a direct consequence of Lemma 3 is

Corollary 2.

$$\mathrm{KL}(\mu_T||\nu_T) \leq \mathrm{KL}(\mu_0||\nu_0) + \frac{1}{4\sigma}\,\mathbb{E}_{\mu_{[0,T]}} \int_0^T \|a(t,x) - b(t,x)\|^2 dt \,. \tag{35}$$

We first apply Corollary 2 from $t = 1$ to $t = 0$ to the two reverse Fokker-Planck equations (26) and (31), respectively sending μ back to π, and the current model sending μ back to $\pi^{(k+1)}$. Since they share the same initial condition, we have

$$\mathrm{KL}(\pi||\pi^{(k+1)}) \leq \frac{1}{4\epsilon}\int_0^1 \mathbb{E}_{\pi_t}\, \|f^*(t,x) - \hat{f}^{(k)}(t,x) - 2\epsilon(s^*(t,x) - \hat{s}^{(k)}(t,x))\|^2 dt \,. \tag{36}$$

We now apply Lemma 1 to the pair of forward Fokker-Planck equations (24) and (30) , to obtain

$$\begin{aligned}
\mathrm{KL}(\pi||\pi^{(k)}) &= \mathrm{KL}(\mu||\mu^{(k)}) + \epsilon\,\mathbb{E}\int_0^1 \|\nabla \log \pi_t - \nabla \log \pi_t^{(k)}\|^2 dt \\
&\quad - \mathbb{E}_\pi \int_0^1 \langle f^* - f^{(k)}, \nabla \log \pi_t - \nabla \log \pi_t^{(k)}\rangle dt \\
&= \mathrm{KL}(\mu||\mu^{(k)}) + \epsilon\,\mathbb{E}\int_0^1 \|s_t^* - s_t^{(k)}\|^2 dt \\
&\quad - \mathbb{E}_\pi \int_0^1 \langle f^* - f^{(k)}, s_t^* - s_t^{(k)}\rangle dt \;\;.
\end{aligned} \tag{37}$$

From (36) and (37) we thus have

$$\mathrm{KL}(\pi||\pi^{(k+1)}) \leq \frac{1}{4\epsilon}\|f^* - \hat{f}^{(k)}\|_\pi^2 + \epsilon\|s^* - \hat{s}^{(k)}\|_\pi^2 - \langle f^* - \hat{f}^{(k)}, s^* - \hat{s}^{(k)}\rangle_\pi \,. \tag{38}$$

Assuming a drift and score approximation error uniformly bounded by δ, we have

$$\begin{aligned}
\mathrm{KL}(\pi||\pi^{(k+1)}) &\leq \frac{1}{4\epsilon}\mathbb{E}\|f^* - f^{(k)}\|^2 + \epsilon\|s^* - s^{(k)}\|^2 - \mathbb{E}\langle f^* - f^{(k)}, s^* - s^{(k)}\rangle && (39)\\
&\quad + \delta^2\left(\frac{1}{4\epsilon} + \epsilon + 1\right) + \delta\left(\frac{1+2\epsilon}{2\epsilon}\|f^* - f^{(k)}\| + (1+\epsilon)\|s^* - s^{(k)}\|\right) && (40)\\
&\leq \mathrm{KL}(\pi||\pi^{(k)}) - \mathrm{KL}(\mu||\mu^{(k)}) && (41)\\
&\quad + \frac{1}{4\epsilon}\mathbb{E}\|f^* - f^{(k)}\|^2 + C_1(\epsilon)\delta^2 + \delta(C_2(\epsilon)\|b^* - b^{(k)}\| + C_3(\epsilon)\|s^* - s^{(k)}\|) \,. && (42)
\end{aligned}$$

Now, using the condition number and SI Lipschitz assumptions, denoting $\eta = 1 + \frac{L}{4\epsilon} - \chi^{-1}$, and redefining $\tilde{\delta} = \sqrt{C_1}\delta$, we obtain

$$\mathrm{KL}(\pi||\pi^{(k+1)}) \leq \eta \mathrm{KL}(\pi||\pi^{(k)}) + \tilde{\delta}^2 + 2\tilde{\delta}\tilde{C}(\|b^* - b^{(k)}\| + \|s^* - s^{(k)}\|) \,. \tag{43}$$

Observe that from (37) and using Cauchy-Schwartz, we have

$$\|s^* - s^{(k)}\|^2 \leq (1-\chi^{-1})\mathrm{KL}(\pi||\pi^{(k)}) + |\langle f^* - f^{(k)}, s^* - s^{(k)}\rangle| \tag{44}$$

$$\leq (1-\chi^{-1})\mathrm{KL}(\pi||\pi^{(k)}) + \sqrt{L\mathrm{KL}(\pi||\pi^{(k)})}\|s^* - s^{(k)}\| \tag{45}$$

$$\leq \eta\mathrm{KL}(\pi||\pi^{(k)}) + \sqrt{\eta\mathrm{KL}(\pi||\pi^{(k)})}\|s^* - s^{(k)}\| \,, \tag{46}$$

which implies $\|s^* - s^{(k)}\| \leq 2\sqrt{\eta\mathrm{KL}(\pi||\pi^{(k)})}$, and therefore

$$\|b^* - b^{(k)}\| + \|s^* - s^{(k)}\| \leq C\sqrt{\eta\mathrm{KL}(\pi||\pi^{(k)})} \,. \tag{47}$$

Thus, by redefining $\tilde{\delta} = \bar{C}_\epsilon\delta$ for some appropriate constant $\bar{C}_\epsilon$ we obtain

$$\mathrm{KL}(\pi||\pi^{(k+1)}) \leq \eta\mathrm{KL}(\pi||\pi^{(k)}) + \tilde{\delta}^2 + 2\tilde{\delta}\sqrt{\eta\mathrm{KL}(\pi||\pi^{(k)})} \tag{48}$$

$$= \left(\sqrt{\eta\mathrm{KL}(\pi||\pi^{(k)})} + \tilde{\delta}\right)^2 \,. \tag{49}$$

Setting $\alpha_k = \mathrm{KL}(\pi||\pi^{(k)})^{1/2}$, we arrive at the linear recurrence

$$\alpha_{k+1} \leq \sqrt{\eta}\alpha_k + \tilde{\delta} \,. \tag{50}$$

Solving this linear recurrence yields

$$\alpha_k \leq \eta^{k/2}\alpha_0 + \frac{\tilde{\delta}}{1-\sqrt{\eta}} \,, \tag{51}$$

hence

$$\mathrm{KL}(\pi||\pi^{(k)}) \leq \left(\eta^{k/2}\alpha_0 + \frac{\tilde{\delta}}{1-\sqrt{\eta}}\right)^2 \tag{52}$$

$$\leq 2\eta^k\mathrm{KL}(\pi||\pi^{(0)}) + \frac{2\tilde{\delta}^2}{(1-\sqrt{\eta})^2} \,, \tag{53}$$

as claimed. ■

Proof of Lemma 1. Let $K_t = \mathrm{KL}(\mu_t||\nu_t) = \int \mu_t(x)\log\left(\frac{\mu_t(x)}{\nu_t(x)}\right)dx$. By definition, the laws μ_t and ν_t solve the Fokker-Planck equations

$$\partial_t\mu_t = \nabla\cdot((-b + \sigma\nabla\log\mu_t)\mu_t) \,, \tag{54}$$

$$\partial_t\nu_t = \nabla\cdot((-a + \sigma\nabla\log\nu_t)\nu_t) \,. \tag{55}$$

We compute

$$\frac{d}{dt}K_t = -\int\frac{\mu_t(x)}{\nu_t(x)}\partial_t\nu_t(x)dx + \int\log\left(\frac{\mu_t(x)}{\nu_t(x)}\right)\partial_t\mu_t(x)dx \tag{56}$$

$$= -\int\frac{\mu_t}{\nu_t}\nabla\cdot((-a+\sigma\nabla\log\nu_t)\nu_t)dx + \int\log\left(\frac{\mu_t}{\nu_t}\right)\nabla\cdot((-b+\sigma\nabla\log\mu_t)\mu_t)dx \tag{57}$$

$$= \int\langle\nabla\left(\frac{\mu_t}{\nu_t}\right), -a+\sigma\nabla\log\nu_t\rangle\nu_t dx - \int\left\langle\nabla\log\left(\frac{\mu_t}{\nu_t}\right), (-b+\sigma\nabla\log\mu_t)\right\rangle\mu_t dx \tag{58}$$

$$= \int\left\langle\nabla\log\left(\frac{\mu_t}{\nu_t}\right), -a+b-\sigma\nabla\log\left(\frac{\mu_t}{\nu_t}\right)\right\rangle\mu_t \tag{59}$$

$$= -\sigma\mathrm{I}(\mu_t||\nu_t) + \mathbb{E}_{\mu_t}\left\langle b-a, \nabla\log\mu_t - \nabla\log\nu_t\right\rangle \,. \tag{60}$$

■

B Detailed Algorithm Pseudocode

Algorithm 2: Training of Self-Consistent Stochastic Interpolant

Input : Observation distribution μ, Forward mapping $\mathcal{F}$, Interpolant schedule (α, β, γ), Initialization of drift and denoiser $\Theta^0 = \{b^{(0)}, g^{(0)}\}$, Total number of iterations K, Number of transport steps T_{tr}

Output : Optimized networks $\Theta^{(K)} = \{b^{(K)}, g^{(K)}\}$

$\Theta \leftarrow \Theta^{(0)}$ `// Initialize transport map`

for k *in* $1 \dots K$ **do**

  for i *in* $1 \dots T_{\text{tr}}$ **do**

    $y \sim \mu$

    $x = \Phi_{\Theta^{(k-1)}}(y)$ `// Backward transport to get a data sample`

    $\tilde{y} = \mathcal{F}(x)$ `// Map back to observations`

    $z \sim \mathcal{N}(0,1);\ t \sim \mathcal{U}(0,1)$

    $I_t = \alpha_t x + \beta_t \tilde{y} + \gamma_t z$

    SGD update of Θ via losses (3)(5)

  $\Theta^{(k)} \leftarrow \Theta$ `// Update transport map`

return $\Theta^{(K)}$

C Implementation Details

Architecture of models We give the architecture details of our SI and diffusion model here. Both architectures are the U-net from Dhariwal & Nichol (2021), specifically following the implementation here. The main difference is that we reduce the number of model channels in the first layer from default 192 to 96 for the diffusion model and 64 for the stochastic interpolant. This is primarily done for computational reasons. As a result, the small model (64 channels) has ∼32 million parameters while the large model has ∼70 million parameters. Maximum positional embedding for the diffusion model and SI is taken to be 10,000 and 2 respectively.

For 2-D latent parameters as used in random masking, we process them with a small U-net consisting of 2 convolution blocks sandwiched between two mode convolution layers and the number of channels given by channel multiplier. We concatenate this with the image along channel dimension. For 1-D latents as used in motion blur and JPEG compression, we process them with a three layer perceptron and then add them to the time embedding.

Table 3: Model configuration parameters

Parameter	Value
Model channels	96 (64)
Channel multiplier	[1, 2, 3, 4]
Channel multiplier for embeddings	4
Number of blocks	3
Attention on resolutions	[32, 16, 8]
Dropout Fraction	0.10
Max positional embedding	10000 (2)
Number of channels in latent U-Net	8

C.1 Training parameters

We use the same hyperparameters for all the experiments. The backward transport map via ODE or SDE is performed in 64 steps. We experimented with different choices of T_{tr}, the number of backward transport steps in Alg. 1, and observed only minor differences across values, with $T_{\text{tr}} = 1$ already sufficient for all current experiments. For simplicity, we therefore set $T_{\text{tr}} = 1$ throughout this

work. For experiments with ODE, we choose the schedule of SI as $\alpha_t = 1 - t, \beta_t = t, \gamma_t = 0$. For experiments with SDE, we keep the same schedule for α_t, β_t, set $\gamma_t = t(1-t)$, and use $\epsilon = 0.1$.

When $\Theta^{(k)}$ is far from the optimal at the early stages of the outer iteration, the distribution of self-generated observations $\mathcal{K}_{\mathcal{F}}(\Phi_{\Theta^{(k)}})_{\#}\mu$ may differ significantly from μ, and consequently slow down the convergence in practice. To mitigate this effect, we modify the interpolant (7) by replacing $\mathcal{F}(\Phi_{\Theta^{(k)}}(y))$ with a mixture: with probability p (set to 0.9 in our experiments), we use the generated observation, and with probability $1 - p$, we use the original y. As long as $p > 0$, following the same argument in Prop. 1, we know the fixed point still gives us the desired optimal parameters Θ^*.

Furthermore, to enhance computational efficiency, for every data mapped back with ODE integration, we (re)-sample the observations twice to generate two interpolated points. This amortizes the cost of ODE integration, which is the most expensive step in the training process. We fix the learning rate to be 0.0005 and use cosine schedule with warmup. Random masking, motion blur and JPEG experiments are trained for 50,000 iterations while other experiments are trained for 20,000 iterations.

D Additional Results

D.1 Diffusion Model

We train a big and a small diffusion model on clean CIFAR-10 data. For sampling, we use 256 steps. The FID for these models is 5.16 and 6.64 respectively. In Fig. 6 and 7, we show some randomly drawn samples from these models.

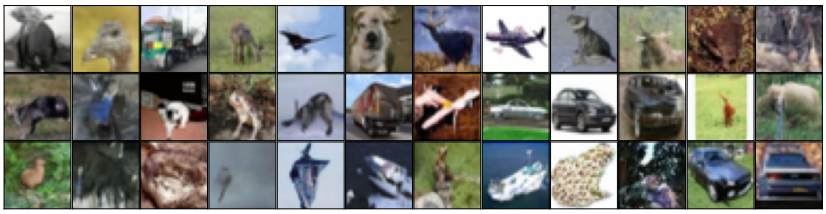


Figure 6: Randomly drawn images from the large diffusion model trained on cleaned images.

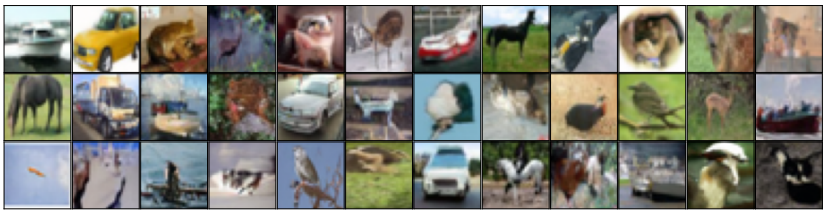
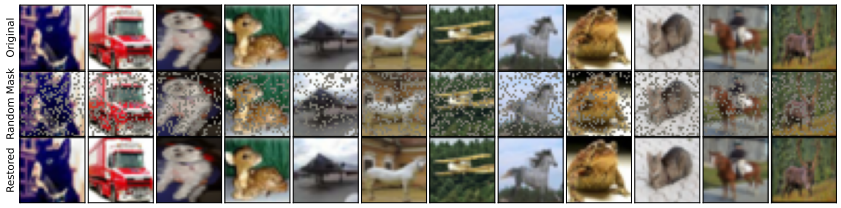


Figure 7: Randomly drawn images from the smaller diffusion model trained on cleaned images.

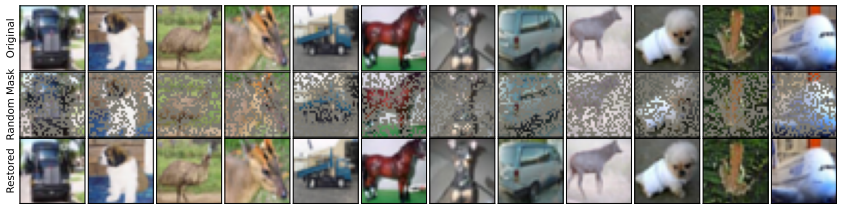
D.2 Random Masking

In Fig. 8, we show additional results for random masking experiment with 25%, 50% and 75% pixels randomly masked. The quality of restored images declines with increasing corruption, but the restored samples are close to original image even for 75% corruption.

For generative modeling, we train a new diffusion model on the samples restored with SI. Fig 9 shows samples from the model trained on the restored samples of the random masking experiment with 50% corruption and negligible noise. As reported in the main text, FID of this model is 6.74.



(a) Random masking with 25% pixels masked.



(b) Random masking with 50% pixels masked.

(c) Random masking with 75% pixels masked.

Figure 8: Restoring images with SI for varying fractions of masked pixels (levels of corruptions).

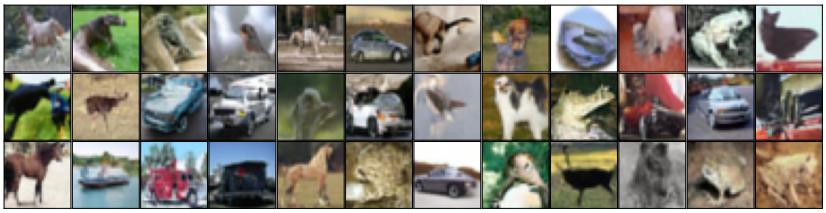


Figure 9: Samples from the diffusion model trained on the restored samples of random masking experiment with 50% corruption.

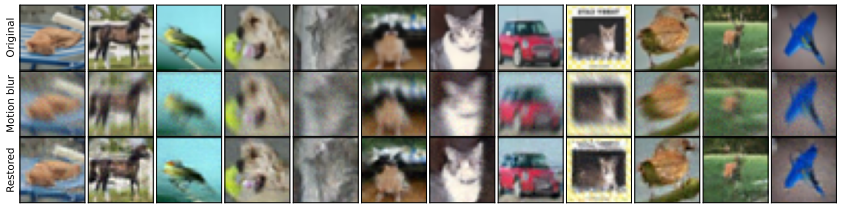
D.3 Motion Blur

In Fig. 10, we show additional results for the motion blur experiment with increasing size of the motion blur kernel from 5 to 9 pixels.

D.4 JPEG Compression

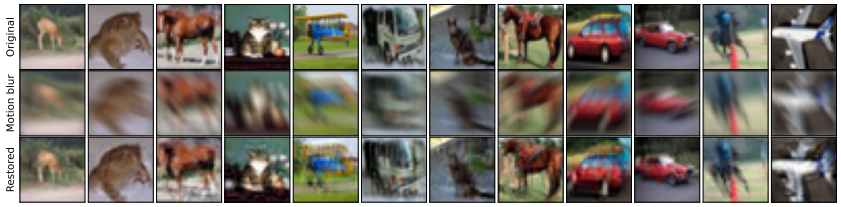
In Fig. 11, we show restorations for JPEG corruption for additional images that have been compressed with randomly chosen ratios. The SI is able to restore samples across a broad range of corruptions.

In addition, we consider another setting where we have training samples that are corrupted with $q \sim \mathcal{U}[0.1, 0.5]$, i.e., we never see high quality samples. The results for the trained SI in this setting are shown in Fig. 12 and 13. The restoration for low-quality samples is poorer than when SI was trained on some samples with compression ratio of more than 0.5. However, note that the SI remains



(a) Motion blur with blur kernel of 5 pixels.

(b) Motion blur with blur kernel of 7 pixels.



(c) Motion blur with blur kernel of 9 pixels.

Figure 10: Restoring images with SI for varying size of motion blur kernel (levels of corruptions).

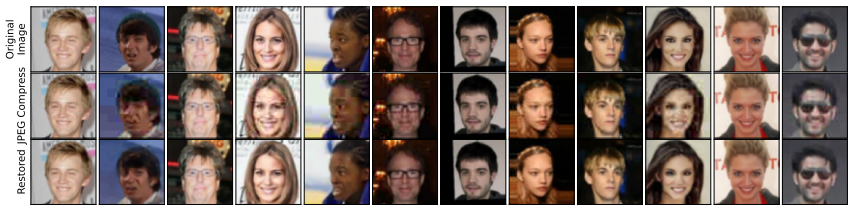


Figure 11: Additional images for JPEG restoration for the model trained on samples with $q \sim \mathcal{U}[0.1, 1.]$.

stable in the extrapolation range, i.e., when restoring sample of $q > 0.5$, the interpolant does indeed improve the restored image even though it has never seen samples in this regime.

D.5 Gaussian Blurring with Poisson noise

In this section, we present additional results for when the forward map is blurring with a Gaussian kernel followed by adding Poisson noise. We add noise with two different levels, $\lambda_n = 0.1$ and 0.5. The restored images here demonstrate that our approach also works in the non-Gaussian noise setting.

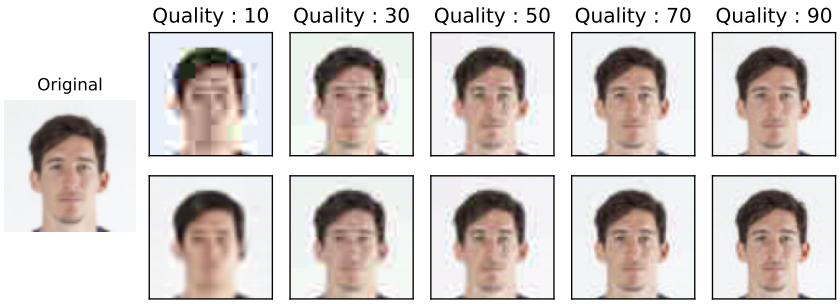


Figure 12: $\mathcal{F}$: JPEG compression + noise ($\sigma_n = 0.01$): results for different compression levels (Top: Corrupted; Bottom: Restored). Model is trained only on samples with $q \sim \mathcal{U}[0.1, 0.5]$. Results for higher qualities are in extrapolation regime.

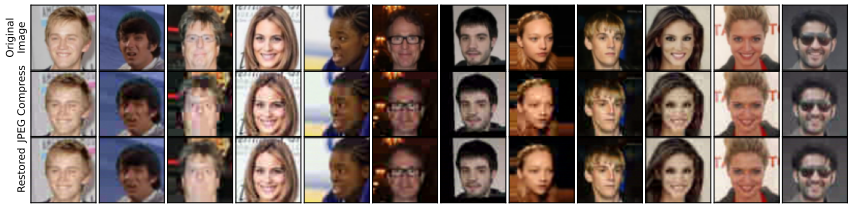
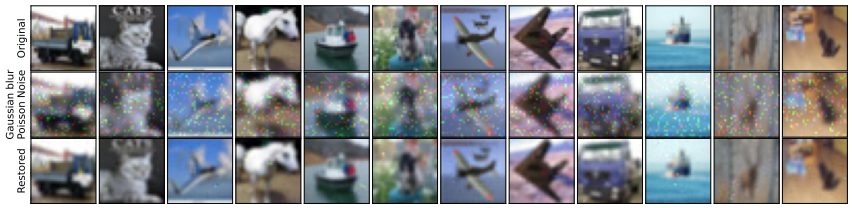


Figure 13: Additional images for JPEG restoration for the model trained on samples with $q \sim \mathcal{U}[0.1, 0.5]$ only.



(a) Gaussian blur ($\sigma_R = 1$) with Poisson noise $\lambda_n = 0.1$.

(b) Gaussian blur ($\sigma_R = 1$) with Poisson noise $\lambda_n = 0.25$.

Figure 14: Restoring images with SI for Gaussian blurring with Poisson noise for different noise levels.