Stochastic Interpolants with Data-Dependent Couplings

Michael S. Albergo [*1] **Mark Goldstein** [*2] **Nicholas M. Boffi** [2] **Rajesh Ranganath** [2 3] **Eric Vanden-Eijnden** [2]

Abstract

Generative models inspired by dynamical transport of measure – such as flows and diffusions – construct a continuous-time map between two probability densities. Conventionally, one of these is the target density, only accessible through samples, while the other is taken as a simple base density that is data-agnostic. In this work, using the framework of stochastic interpolants, we formalize how to *couple* the base and the target densities, whereby samples from the base are computed conditionally given samples from the target in a way that is different from (but does not preclude) incorporating information about class labels or continuous embeddings. This enables us to construct dynamical transport maps that serve as conditional generative models. We show that these transport maps can be learned by solving a simple square loss regression problem analogous to the standard independent setting. We demonstrate the usefulness of constructing dependent couplings in practice through experiments in super-resolution and in-painting. The code is available at https://github.com/interpolants/couplings.

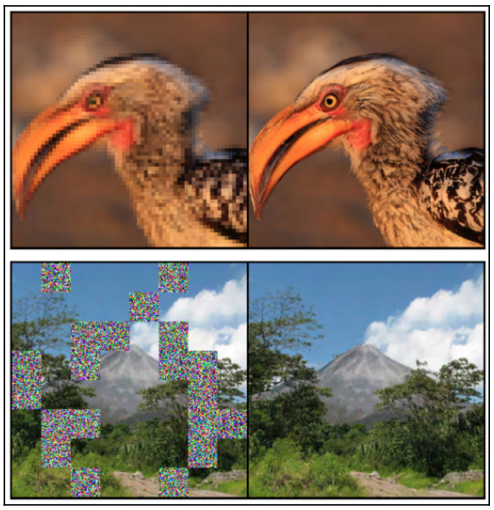

Figure 1: Examples. Super-resolution and in-painting results computed with our formalism.

1. Introduction

Generative models such as normalizing flows and diffusions sample from a target density ρ_1 by continuously transforming samples from a base density ρ_0 into the target. This transport is accomplished by means of an ordinary differential equation (ODE) or stochastic differential equation (SDE), which takes as initial condition a sample from ρ_0 and produces at time $t = 1$ an approximate sample from ρ_1. Typically, the base density is taken to be something simple, analytically tractable, and easy to sample, such as a standard Gaussian. In some formulations, such as score-based diffusion (Sohl-Dickstein et al., 2015; Song & Ermon, 2020; Ho et al., 2020b; Song et al., 2020; Singhal et al., 2023), a Gaussian base density is intrinsically tied to the process achieving the transport. In others, including flow matching (Lipman et al., 2022a; Chen & Lipman, 2023), rectified flow (Liu et al., 2022b; 2023b), and stochastic interpolants (Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023), a Gaussian base is not required, but is often chosen for convenience. In these cases, the choice of Gaussian base represents an absence of prior knowledge about the problem structure, and existing works have yet to fully explore the strength of base densities adapted to the target.

In this work, we introduce a general formulation of stochastic interpolants in which a base density is produced via a *coupling*, whereby samples of this base are computed conditionally given samples from the target. We construct a continuous-time stochastic process that interpolates between the coupled base and target, and we characterize the resulting transport by identification of a continuity equation obeyed by the time-dependent density. We show that the velocity field defining this transport can be estimated by solution of an efficient, simulation-free square loss regression problem analogous to standard, data-agnostic interpolant

[*]Equal contribution [1]Center for Cosmology and Particle Physics, New York University [2]Courant Institute of Mathematical Sciences, New York University [3]Center for Data Science, New York University. Correspondence to: Michael S. Albergo <albergo@nyu.edu>, Mark Goldstein <goldstein@nyu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

and flow matching algorithms.

In our formulation, we also allow for dependence on an external, conditional source of information independent of ρ_1, which we call ξ. This extra source of conditioning is standard, and can be used in the velocity field $b_t(x, \xi)$ to accomplish class-conditional generation, or generation conditioned on a continuous embedding such as a textual representation or problem-specific geometric information. As illustrated in Fig. 2, it is however different from the data-dependent coupling that we propose. Below, we suggest some generic ways to construct coupled, conditional base and target densities, and we consider practical applications to image super-resolution and in-painting, where we find improved performance by incorporating both a data-dependent coupling and the conditioning variable. Together, our **main contributions** can be summarized as:

1. We define a broader way of constructing base and target pairs in generative models based on dynamical transport that adapts the base to the target. In addition, we formalize the use of conditional information – both discrete and continuous – in concert with this new form of *data coupling* in the stochastic interpolant framework. As special cases of our general formulation, we obtain several recent variants of conditional generative models that have appeared in the literature.

2. We provide a characterization of the transport that results from conditional, data-dependent generation, and analyze theoretically how these factors influence the resulting time-dependent density

3. We provide an empirical study on the effect of coupling for stochastic interpolants, which have recently been shown to be a promising, flexible class of generative models. We demonstrate the utility of data-dependent base densities and the use of conditional information in two canonical applications, image inpainting and super-resolution, which highlight the performance gains that can be obtained through the application of the tools developed here .

The rest of the paper is organized as follows. In Section 2, we describe some related work in conditional generative modeling. In Section 3, we introduce our theoretical framework. We characterize the transport that results from the use of data-dependent couplings, and discuss the difference between this approach and conditional generative modeling. In Section 4, we apply the framework to numerical experiments on ImageNet, focusing on image inpainting and image super-resolution. We conclude with some remarks and discussion in Section 5.

2. Related Work

Couplings. Several works have studied the question of how to build couplings, primarily from the viewpoint of optimal transport theory. An initial perspective in this regard comes from (Pooladian et al., 2023; Tong et al., 2023; Klein et al., 2023), who state an unbiased means for building entropically-regularized optimal couplings from minibatches of training samples. This perspective is appealing in that it may give probability flows that are straighter and hence more easily computed using simple ODE solvers. However, it relies on estimating an optimal coupling over minibatches of the entire dataset, which, for large datasets, may become uninformative as to the true coupling. In an orthogonal perspective, (Lee et al., 2023) presented an algorithm to learn a coupling between the base and the target by building dependence on the target into the base. They argue that this can reduce curvature of the underlying transport. While this perspective empirically reduces the curvature of the flow lines, it introduces a potential bias in that they still sample from an independent base, possibly not equal to the marginal of the learned conditional base. Learning a coupling can also be achieved by solving the Schrödinger bridge problem, as investigated e.g. in (De Bortoli et al., 2021; Shi et al., 2023). This leads to iterative algorithms that require solving pairs of SDEs until convergence, which is costly in practice. More closely connected to our work are the approaches proposed in (Liu et al., 2023a; Somnath et al., 2023): by considering generative modeling through the lens of diffusion bridges with *known* coupling, they arrive to a formulation that is operationally similar to, but less general than, ours. Our approach is simpler, and more flexible, as it differentiates between the bridging of the densities and the construction of the generative models. Table 1 summarizes these couplings along with the standard independent pairing.

Generative Modeling and Dynamical Transport. Generative models built upon dynamical transport of measure go back at least to (Tabak & Vanden-Eijnden, 2010; Tabak & Turner, 2013), and were further developed in (Rezende & Mohamed, 2015; Dinh et al., 2017; Huang et al., 2016; Durkan et al., 2019) using compositions of discrete maps, while modern models are typically formulated via a continuous-time transformation. In this context, a major advance was the introduction of score-based diffusion (Song et al., 2021b;a), which relates to denoising diffusion probabilistic models (Ho et al., 2020a), and allows one to generate samples by learning to reverse a stochastic differential equation that maps the data into samples from a Gaussian base density. Methods such as flow matching (Lipman et al., 2022b), rectified flow (Liu, 2022; Liu et al., 2022a), and stochastic interpolants (Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023) expand on the idea of building stochas-

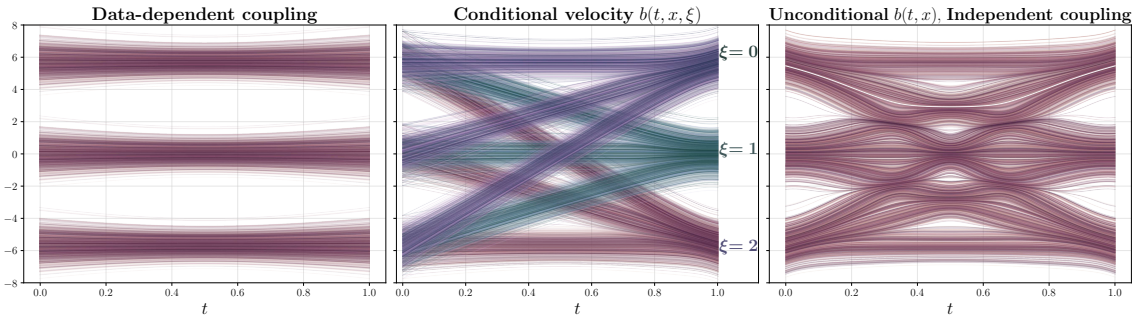

Figure 2: Data-dependent couplings are different than conditioning. Delineating between constructing couplings versus conditioning the velocity field, and their implications for the corresponding probability flow X_t. The transport problem is flowing from a Gaussian Mixture Model (GMM) with 3 modes to another GMM with 3 modes. *Left*: The probability flow X_t arising from the data-dependent coupling $\rho(x_0, x_1) = \rho_1(x_1)\rho_0(x_0|x_1)$. All samples follow simple trajectories. No formation of auxiliary modes form in the intermediate density $\rho(t)$, in juxtaposition to the independent case. *Center*: When the velocity field is conditioned $b_t(x, \xi)$ on each class (mode), it factorizes, resulting in three separate probability flows X_t^ξ with $\xi = 1, 2, 3$. *Right*: The probability flow X_t when taking an unconditional velocity field $b_t(x)$ and an independent coupling $\rho(x_0, x_1) = \rho_0(x_0)\rho_1(x_1)$. Note the complexity of the underlying transport, which motivates us to consider finding correlated base variables directly in the data.

Table 1: Couplings. Standard formulations of flows and diffusions construct generative models built upon an independent coupling (Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023; Lipman et al., 2022a; Liu et al., 2022b). (Lee et al., 2023) learn $q_\phi(x_0|x_1)$ jointly with the velocity to define the coupling during training, but instead sample from $\rho_0 = \mathsf{N}(0, Id)$ for generation. (Tong et al., 2023) and (Pooladian et al., 2023) build couplings by running mini-batch optimal transport algorithms (Cuturi, 2013). Here we focus on couplings enabled by our generic formalism, which bears similarities with (Liu et al., 2023a; Somnath et al., 2023), and can be individualized to each generative task.

Coupling PDF $\rho(x_0, x_1)$	**Base PDF**	**Description**
$\rho_1(x_1)\rho_0(x_0)$	$x_0 \sim \mathsf{N}(0, Id)$	Independent
$\rho(x_0\|x_1)\rho_1(x_1)$	$x_0 \sim q_\phi(x_0\|x_1)$	Learned conditional
mb-OT(x_1, x_0)	$x_0 \sim \mathsf{N}(0, Id)$	Minibatch OT
$\rho_1(x_1)\rho_0(x_0\|x_1)$	$x_0 \sim \rho_0(x_0\|x_1)$	Dependent-coupling **(this work)**

tic processes that connect a base density to the target, but allow for bases that are more general than a Gaussian density. Typically, these constructions assume that the samples from the base and the target are uncorrelated.

Conditional Diffusions and Flows for Images. (Saharia et al., 2022; Ho et al., 2022a) build diffusions for super-resolution, where low-resolution images are given as inputs to a score model, which formally learns a conditional score (Ho & Salimans, 2022). In-painting can be seen as a form of conditioning where the conditioning set determines some coordinates in the target space. In-painting diffusions have been applied to video generation (Ho et al., 2022b) and protein backbone generation (Trippe et al., 2022). In the *replacement method* one directly inputs the clean values of the known coordinates at each step of integration (Ho et al., 2022b); (Schneuing et al., 2022) replace with draws of the diffused state of the known coordinates. (Trippe et al., 2022; Wu et al., 2023) discuss approximation error in this approach and correct with sequential Monte-Carlo. We revisit this problem framing from the velocity modeling perspective in Section 4.1. Recent work has applied flows to high-dimensional conditional modeling (Dao et al., 2023; Hu et al., 2023). A Schrödinger bridge perspective on the conditional generation problem was presented in (Shi et al., 2022).

3. Stochastic interpolants with couplings

Suppose that we are given a dataset $\{x_1^i\}_{i=1}^n$. The aim of a generative model is to draw new samples assuming that the data set comes from a probability density function (PDF) $\rho_1(x_1)$. Following the stochastic interpolant framework (Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023), we

introduce a time-dependent stochastic process that interpolates between samples from a simple base density $\rho_0(x_0)$ at time $t = 0$ and samples from the target $\rho_1(x_1)$ at time $t = 1$:

Definition 3.1 (Stochastic interpolant with coupling)**.** *The stochastic interpolant I_t is the process defined as*[1]

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

where

- *α_t, β_t, and γ_t^2 are differentiable functions of time such that $\alpha_0 = \beta_1 = 1$, $\alpha_1 = \beta_0 = \gamma_0 = \gamma_1 = 0$, and $\alpha_t^2 + \beta_t^2 + \gamma_t^2 > 0$ for all $t \in [0,1]$.*
- *The pair (x_0, x_1) is jointly drawn from a probability density $\rho(x_0, x_1)$ with finite second moments and such that*

$$\int_{\mathbb{R}^d} \rho(x_0, x_1) dx_1 = \rho_0(x_0), \tag{2}$$

$$\int_{\mathbb{R}^d} \rho(x_0, x_1) dx_0 = \rho_1(x_1). \tag{3}$$

- *$z \sim \mathsf{N}(0, Id)$, independent of (x_0, x_1).*

A simple instance of (1) uses $\alpha_t = 1 - t$, $\beta_t = t$, and $\gamma_t = \sqrt{2t(1-t)}$.

The stochastic interpolant framework uses information about the process I_t to derive either an ODE or an SDE whose solutions X_t push the law of x_0 onto the law of I_t for all times $t \in [0,1]$.

As shown in Section 3.1, the drift coefficients in these ODEs/SDEs can be estimated by quadratic regression. They can then be used as generative models, owing to the property that the process x_t specified in Definition 3.1 satisfies $I_{t=0} = x_0 \sim \rho_0(x_0)$ and $I_{t=1} = x_1 \sim \rho_1(x_1)$, and hence samples the desired target density. By drawing samples $x_0 \sim \rho_0(x_0)$ and using them as initial data $X_{t=0} = x_0$ in the ODEs/SDEs, we can then generate samples $X_{t=1} \sim \rho_1(x_1)$ via numerical integration.

In the original stochastic interpolant papers, this construction was made using the choice $\rho(x_0, x_1) = \rho_0(x_0)\rho_1(x_1)$, so that x_0 and x_1 were drawn independently from the base and the target.

Our aim here is to build generative models that are more powerful and versatile by exploring and exploiting dependent couplings between x_0 and x_1 via suitable definition of $\rho(x_0, x_1)$.

[1] More generally, we may set $I_t = I(t, x_0, x_1)$ in (1), where I satisfies some regularity properties in addition to the boundary conditions $I(t = 0, x_0, x_1) = x_0$ and $I(t = 1, x_0, x_1) = x_1$ (Albergo & Vanden-Eijnden, 2022; Albergo et al., 2023). For simplicity, we will stick to the linear choice $I(t, x_0, x_1) = \alpha_t x_0 + \beta_t x_1$.

Remark 3.1 (Incorporating conditioning)**.** *Our formalism allows (but does not require) that each data point $x_1^i \in \mathbb{R}^d$ comes with a label $\xi_i \in D$, such as a discrete class or a continuous embedding like that of a text caption. In this setup, our results can be straightforwardly generalized by making all the quantities (PDF, velocities, etc.) conditional on ξ. This is discussed in Appendix A and used in various forms in our numerical examples.*

3.1. Transport equations and conditional generative models

In this section, we show that the probability distribution of the process I_t defined in (1) has a time-dependent density $\rho_t(x)$ that interpolates between $\rho_0(x)$ and $\rho_1(x)$. We characterize this density as the solution of a transport equation, and we show that both the corresponding velocity field and the score $\nabla \log \rho_t(x)$ are minimizers of simple quadratic objective functions.

This result enables us to construct conditional generative models by approximating the velocity (and possibly the score) via minimization over a rich parametric class such as neural networks. We first define the functions:

$$b_t(x) = \mathbb{E}(\dot{I}_t | I_t = x), \quad g_t(x) = \mathbb{E}(z | I_t = x), \tag{4}$$

where the dot denotes time-derivative and $\mathbb{E}(\cdot | I_t = x)$ denotes the expectation over $\rho(x_0, x_1)$ conditional on $I_t = x$. We then have,

Theorem 3.1 (Transport equation with coupling)**.** *The probability distribution of the stochastic interpolant I_t defined in (1) has a density $\rho_t(x)$ that satisfies $\rho_{t=0}(x) = \rho_0(x)$ and $\rho_{t=1}(x) = \rho_1(x)$, and solves the transport equation*

$$\partial_t \rho_t(x) + \nabla \cdot (b_t(x)\rho_t(x)) = 0, \tag{5}$$

where the velocity field $b_t(x)$ is defined in (4). Moreover, for every t such that $\gamma_t \neq 0$, the following identity for the score holds

$$\nabla \log \rho_t(x) = -\gamma_t^{-1} g_t(x). \tag{6}$$

Finally, the functions b and g are the unique minimizers of the objectives

$$\begin{aligned} L_b(\hat{b}) &= \int_0^1 \mathbb{E}\left[|\hat{b}_t(I_t)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t)\right] dt, \\ L_g(\hat{g}) &= \int_0^1 \mathbb{E}\left[|\hat{g}_t(I_t)|^2 - 2z \cdot \hat{g}_t(I_t)\right] dt \end{aligned} \tag{7}$$

where $\mathbb{E}$ denotes an expectation over $(x_0, x_1) \sim \rho(x_0, x_1)$ and $z \sim \mathsf{N}(0, Id)$ with $(x_0, x_1) \perp z$.

A more general version of this result with a conditioning variable is proven in Appendix A. The objectives (7) can

readily be estimated in practice from samples $(x_0, x_1) \sim \rho(x_0, x_1)$ and $z \sim \mathsf{N}(0, 1)$, which will enable us to learn approximations for use in a generative model.

The transport equation (5) can be used to derive generative models, as we now show.

Corollary 3.1 (Probability flow and diffusions with coupling)**.** *The solutions to the probability flow equation*

$$\dot{X}_t = b_t(X_t) \tag{8}$$

enjoy the property that

$$X_{t=1} \sim \rho_1(x_1) \quad \textit{if} \quad X_{t=0} \sim \rho_0(x_0) \tag{9}$$
$$X_{t=0} \sim \rho_0(x_0) \quad \textit{if} \quad X_{t=1} \sim \rho_1(x_1) \tag{10}$$

In addition, for any $\epsilon_t \geq 0$, solutions to the forward SDE

$$dX_t^F = b_t(X_t^F)dt - \epsilon_t \gamma_t^{-1} g_t(X_t^F)dt + \sqrt{2\epsilon_t}dW_t, \tag{11}$$

enjoy the property that

$$X_{t=1}^F \sim \rho_1(x_1) \quad \textit{if} \quad X_{t=0}^F \sim \rho_0(x_0), \tag{12}$$

and solutions to the backward SDE

$$dX_t^R = b_t(X_t^R)dt + \epsilon_t \gamma_t^{-1} g_t(X_t^R)dt + \sqrt{2\epsilon_t}dW_t, \tag{13}$$

enjoy the property that

$$X_{t=0}^R \sim \rho_0(x_0) \quad \textit{if} \quad X_{t=1}^R \sim \rho_1(x_1). \tag{14}$$

A more general version of this result with conditioning is also proven in Appendix A.

Corollary 3.1 shows that the coupling can be incorporated both in deterministic and stochastic generative models derived within the stochastic interpolant framework. In what follows, for simplicity we will focus on the deterministic probability flow ODE (8).

An important observation is that the transport cost of the generative model based on the probability flow ODE (8), which impacts the numerical stability of solving this ODE, is controlled by the time dynamics of the interpolant, as shown by our next result:

Proposition 3.1 (Control of transport cost)**.** *Let $X_t(x_0)$ be the solution to the probability flow ODE (8) for the initial condition $X_{t=0}(x_0) = x_0 \sim \rho_0$. Then*

$$\mathbb{E}_{x_0 \sim \rho_0}\left[|X_{t=1}(x_0) - x_0|^2\right] \leq \int_0^1 \mathbb{E}[|\dot{I}_t|^2]dt < \infty \tag{15}$$

The proof of this proposition is given in Appendix A. Minimizing the left hand-side of (15) would achieve optimal transport in the sense of Benamou-Brenier (Benamou & Brenier, 2000), and the minimum would give the Wasserstein-2 distance between ρ_0 and ρ_1. Various works seek to minimize this distance procedurally either by adapting the coupling (Pooladian et al., 2023; Tong et al., 2023) or by optimizing $\rho_t(x)$ (Albergo & Vanden-Eijnden, 2022), at additional cost. Here we introduce *designed* couplings at no extra cost that can lower the upper bound in (15). This will allow us to show how different couplings enable stricter control of the transport cost in various applications. Let us now discuss a generic instantiation of our formalism involving a specific choice of $\rho(x_0, x_1)$.

3.2. Designing data-dependent couplings

One natural way to allow for a data-dependent coupling between the base and the target is to set

$$\rho(x_0, x_1) = \rho_1(x_1)\rho_0(x_0|x_1) \quad \text{with} \tag{16}$$
$$\int_{\mathbb{R}^d} \rho_0(x_0|x_1)\rho_1(x_1)dx_1 = \rho_0(x_0). \tag{17}$$

There are many ways to construct the conditional $\rho_0(x_0|x_1)$. In the numerical experiments in Section 4.1 & Section 4.2, we consider base densities of a variable x_0 of the generic form

$$x_0 = m(x_1) + \sigma\zeta, \tag{18}$$

where $m(x_1) \in \mathbb{R}^d$ is some function of x_1, possibly random even if conditioned on x_1, $\sigma \in \mathbb{R}^{d\times d}$, and $\zeta \sim \mathsf{N}(0, Id)$ with $\zeta \perp m(x_1)$. In this set-up, the corrupted observation $m(x_1)$ (a noisy, partial, or low-resolution image) is determined by the task at hand and available to us, but we are free to choose the design of the term $\sigma\zeta$ in (18) in ways that can be exploited differently in various applications (and is allowed to depend on any conditional info ξ). Note in particular that, given $m(x_1)$, (18) is easy to generate at sampling time. Note also that, if the corrupted observation $m(x_1)$ is deterministic given x_1, the conditional probability density of (18) is the Gaussian density with mean $m(x_1)$ and covariance $C = \sigma\sigma^\top$:

$$\rho_0(x_0|x_1) = \mathsf{N}(x_0; m(x_1), C), \tag{19}$$

We stress that, even in this case, $\rho(x_0, x_1) = \rho_1(x_1)\rho_0(x_0|x_1)$ and $\rho_0(x_0) = \rho_0(x_0|x_1)$ are non-Gaussian densities in general. In this context, we can use the interpolant from (1) with $\gamma_t = 0$, which reduces to:

$$I_t = \alpha_t(m(x_1) + \sigma\zeta) + \beta_t x_1 \tag{20}$$

Note that the score associated to (20) is still available because of the factor of $\sigma\zeta$, so long as σ is invertible.

3.3. Reducing transport costs via coupling

In the numerical experiments, we will highlight how the construction of a data-dependent coupling enables us to

Algorithm 1 Training

Input: Interpolant coefficients α_t, β_t; velocity model $\hat{b}$; batch size n_b;
repeat
  for $i = 1, \ldots, n_\mathrm{b}$ **do**
    Draw $x_1^i \sim \rho_1(x_1)$, $\zeta_i \sim \mathcal{N}(0, Id)$, $t_i \sim U(0,1)$.
    Compute $x_0^i = m(x_1^i) + \sigma\zeta^i$.
    Compute $I_{t_i} = \alpha_{t_i} x_0^i + \beta_{t_i} x_1^i$.
  end for
  Compute empirical loss
  $\hat{L}_b(\hat{b}) = n_\mathrm{b}^{-1} \sum_{i=1}^{n_\mathrm{b}} [|\hat{b}_{t_i}(I_{t_i})|^2 - 2\dot{I}_{t_i} \cdot \hat{b}_{t_i}(I_{t_i})]$.
  Take gradient step on $\hat{L}_b(\hat{b})$ to update $\hat{b}$.
until converged
Return: Velocity $\hat{b}$.

Algorithm 2 Sampling (via forward Euler method)

Input: model $\hat{b}$, corrupted sample $m(x_1)$, $N \in \mathbb{N}$.
Draw noise $\zeta \sim \mathcal{N}(0, Id)$
Initialize $\hat{X}_0 = m(x_1) + \sigma\zeta$
for $n = 0, \ldots, N-1$ **do**
  $\hat{X}_{i+1} = \hat{X}_i + N^{-1} \hat{b}_{i/N}(\hat{X}_i)$
end for
Return: clean sample $\hat{X}_N$.

perform various downstream tasks. An additional appeal is that data-dependent couplings facilitate *the design of* more efficient transport than standard generation from a Gaussian, as we now show.

The bound on the transportation cost in (15) may be more tightly controlled by the construction of data-dependent couplings and their associated interpolants. In this case, we seek couplings such that $E[|\dot{I}_t|^2]$ is smaller with coupling than without, i.e. such that

$$\begin{aligned} &\int_{\mathbb{R}^{3d}} |\dot{I}_t|^2 \rho(x_0, x_1) \rho_z(z) dx_0 dx_1 dz \\ &\leq \int_{\mathbb{R}^{3d}} |\dot{I}_t|^2 \rho_0(x_0) \rho_1(x_1) \rho_z(z) dx_0 dx_1 dz, \end{aligned} \tag{21}$$

where $\dot{I}_t = \dot{\alpha}_t x_0 + \dot{\beta}_t x_1 + \dot{\gamma}_t z$ is a function of x_0, x_1 and z. A simple way to design such a coupling is to consider (19) with $m(x_1) = x_1$ and $C = \sigma^2 Id$ for some $\sigma > 0$, which sets the base distribution to be a noisy version of the target. In the case of data-decorruption (which we explore in the numerical experiments), this interpolant directly connects the corrupted conditional density and the uncorrupted density. If we choose $\alpha_t = 1 - t$ and $\beta_t = t$, and set $\gamma_t = 0$, then $\dot{I}_t = x_1 - x_0$, and the left hand-side of (21) reduces to $\mathbb{E}[|\sigma z|^2] = d\sigma^2$, which is less than the right hand-side given by $2\mathbb{E}[|x_1|^2] + d\sigma^2$.

3.4. Learning and Sampling

To learn in this setup, we can evaluate the objective functions (7) over a minibatch of $n_\mathrm{b} < n$ data points x_0^i, x_1^i by using an additional n_b samples $z_i \sim \mathsf{N}(0, Id)$ and $t_i \sim U([0,1])$. This leads to the empirical approximation $\hat{L}_b$ of L_b given by

$$\hat{L}_b(\hat{b}) = \frac{1}{n_\mathrm{b}} \sum_{i=1}^{n_\mathrm{b}} \left[|\hat{b}_{t_i}(I_{t_i})|^2 - 2\dot{I}_{t_i} \cdot \hat{b}_{t_i}(I_{t_i}) \right], \tag{22}$$

with a similar empirical variant for L_z. We approximate the functions $b_t(x)$ and $g_t(x)$ with neural networks and minimize these empirical objectives with stochastic gradient descent. This leads to an approximation of the velocity $b_t(x)$ via (4) and of the score via (6).

Generating data requires sampling an $X_{t=0} \sim \rho_0(x_0)$ as an initial condition to be evolved via the probability flow ODE (8) or the forward SDE (11) to respectively produce a sample $X_{t=1} \sim \rho_1(x_1)$ or $X_{t=1}^F \sim \rho_1(x_1)$. Sampling an x_0 can be performed by picking data point x_1 either from the data set or from some online data acquisition procedure and using it in (18), or using the assumption that one directly observes $x_0 \sim \rho_0(x_0)$ at inference time (e.g. one receives a partial image). The generated samples from either the probability flow ODE or forward SDE will be different from x_1, even with the choices $m(x_1) = x_1$ and $C = \sigma^2 Id$. The probability flow ODE necessarily produces a single sample of x_1 for each x_0, while the SDE produces a collection of samples whose spread can be controlled by the diffusion coefficient ϵ_t. Algorithms 1 and 2 depict these training and sampling procedures, respectively.

4. Numerical experiments

We now explore the interpolants with data-dependent couplings on conditional image generation tasks; we find that the framework is straightforward to scale to high resolution images directly in pixel space.

4.1. In-painting

We consider an in-painting task, whereby $x_1 \in \mathbb{R}^{C \times W \times H}$ denotes an image with C channels, width W, and height H. Given a pre-specified mask, the goal is to fill the pixels in the masked region with new values that are consistent with the entirety of the image. We set the conditioning variable $\xi \in \{0,1\}^{C \times W \times H}$ and additionally provide the model with any potential class labels. For simplicity, the mask takes the same value for all channels in a given spatial location in the image. We define the base density by the relation $x_0 = \xi \circ x_1 + (1 - \xi) \circ \zeta$, where $\circ$ denotes the Hadamard (elementwise) product and $\zeta \in \mathbb{R}^{C \times W \times H}, \zeta \sim \mathsf{N}(0, Id)$ denotes random noise used to initialize the pixels within the masked region (separate noise for each channel). During

training, the mask is drawn randomly by tiling the image into 64 tiles; each tile is selected to enter the mask with probability $p = 0.3$. In our experiments, we set $\rho_1(x_1)$ to correspond to ImageNet (either 256 or 512). This corresponds to using $\rho(x_0, x_1|\xi) = \rho_1(x_1)\rho_0(x_0|x_1, \xi)$. The model sees the mask; we note that we do not need to additionally input the partial image as extra conditioning because it is present, uncorrupted, in x_t for each t because the values are present in x_0 and x_1. In the interpolant (20), we set $\alpha_t = t$ and $\beta_t = 1 - t$. In this setup, the velocity field $b_t(x, \xi)$ is such that $b_t(x, \xi) = 0$ except in the masked regions. This follows because $\xi \circ I_t = \xi \circ x_1$ for every t, i.e., the unmasked pixels in I_t are always those of x_1 for which $\dot{I}_t = 0$. To take this structural information into account, we can build this property into our neural network model, and mask the output of the approximate velocity field to enforce that the unmasked pixels remain fixed. We note that this method does not necessitate any inference time corrections, such as the replacement method or MCMC.

Results. For implementation, we parameterize $b_t(x, \xi)$ using the basic U-Net architecture from (Ho et al., 2020b), where ξ is given to the model as appended channels of the image x. Additional specific experimental details may be found in Appendix B. Samples are shown in Figure 3, as well as Section 1. FIDs are reported in Table 2. As discussed, the missing areas of the image are defined at time zero as independent normal random variables, depicted as colorful static in the images. In each image triple, the left panel is the base distribution sample x_0, the middle is the model sample of $X_{t=1}$ obtained by integrating the probability flow ODE (8), and the right panel is the ground truth. The generated textures, though different from the full sample, correspond to realistic samples from the conditional densities given the observed content. This is an advantage of probabilistic generative models such as ours over models optimized to fit a mean-square error to a ground truth image.

4.2. Super-resolution on Imagenet

We now consider image super-resolution, in which we would like to produce an image with the same content as a given image but at higher resolution. To this end, we let $x_1 \in \mathbb{R}^{C\times W\times H}$ correspond to a high-resolution image, as in Section 4.1. We denote by $\mathcal{D} : \mathbb{R}^{C\times W\times H} \to \mathbb{R}^{C\times W_{\text{low}}\times H_{\text{low}}}$ and $\mathcal{U} : \mathbb{R}^{C\times W_{\text{low}}\times H_{\text{low}}} \to \mathbb{R}^{C\times W\times H}$ image downsampling and upsampling operations, where W_{low} and H_{low} denote the width and height of a low-resolution image. To define the base density, we then set $x_0 = \mathcal{U}(\mathcal{D}(x_1)) + \sigma\zeta$ with $\zeta \in \mathbb{R}^{C\times W\times H}$, $\zeta \sim \mathsf{N}(0, Id)$, and $\sigma > 0$. Defining x_0 in this way frames the transport problem such that each starting pixel is proximal to its intended target. Notice in particular that, with $\sigma = 0$, each x_0 would correspond to a lower-dimensional sample embedded in a higher-dimensional space, and the corresponding distribution would be concentrated on a lower-dimensional manifold. Working with $\sigma > 0$ alleviates the associated singularities by adding a small amount of Gaussian noise to smooth the base density so it is well-defined over the entire higher-dimensional ambient space. In addition, we give the model access to the low-resolution image at all times; this problem setting then corresponds to using $\rho(x_0, x_1|\xi) = \rho_1(x_1)\rho_0(x_0|x_1, \xi)$ with $\xi = \mathcal{U}(\mathcal{D}(x_1))$. In the experiments, we set ρ_1 to correspond to ImageNet (256 or 512), following prior work (Saharia et al., 2022; Ho et al., 2022a).

Table 2: FID for Inpainting Task. FID comparison between under two paradigms: a baseline, where ρ_0 is a Gaussian with independent coupling to ρ_1, and our data-dependent coupling detailed in Section 4.1.

Model	FID-50k
Uncoupled Interpolant (Baseline)	1.35
Dependent Coupling (**Ours**)	**1.13**

Table 3: FID-50k for Super-resolution, 64x64 to 256x256. FIDs for baselines taken from (Saharia et al., 2022; Ho et al., 2022a; Liu et al., 2023a).

Model	Train	Valid
Improved DDPM (Nichol & Dhariwal, 2021)	12.26	–
SR3 (Saharia et al., 2022)	11.30	5.20
ADM (Dhariwal & Nichol, 2021)	7.49	3.10
Cascaded Diffusion (Ho et al., 2022a)	4.88	4.63
I^2SB (Liu et al., 2023a)	–	2.70
Dependent Coupling (**Ours**)	**2.13**	**2.05**

Results. Similarly to the previous experiment, we append the upsampled low-resolution images ξ to the channel dimension of the input x of the velocity model, and likewise include the ImageNet class labels. Samples are displayed in Fig. 4, as well as Section 1. Similar in layout to the previous experiment, the left panel of each triplet is the low-resolution image, the middle panel is the model sample $X_{t=1}$, and the right panel is the high-resolution image. The differences are easiest to see when zoomed-in. While the increased resolution of the model sample is very noticeable for 64 to 256, the differences even in ground truth images between 256 and 512 are more subtle. We also display FIDs for the 64x64 to 256x256 task, which has been studied in other works, in Table 3.

5. Discussion, challenges, and future work

In this work, we introduced a general framework for constructing data-dependent couplings between base and target densities within the stochastic interpolant formalism.

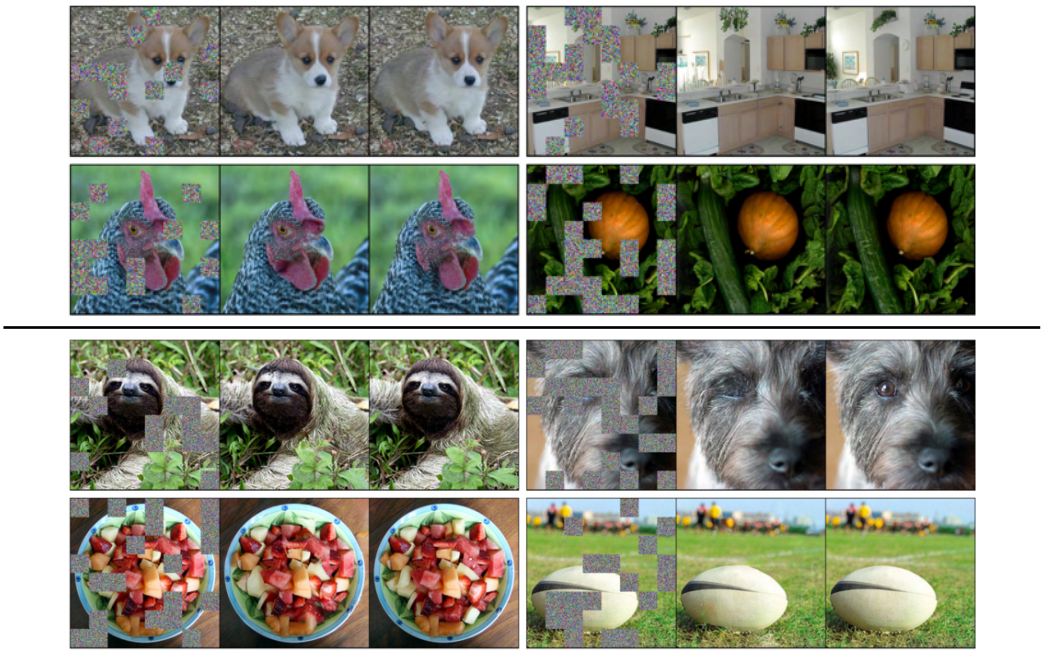

Figure 3: Image inpainting: ImageNet-256×256 and ImageNet-512×512. *Top panels*: Six examples of image in-filling at resolution 256×256, where the left columns display masked images, the center corresponds to in-filled model samples, and the right shows full reference images. The aims are not to recover the precise content of the reference image, but instead, to provide a conditionally valid in-filling. *Bottom panels*: Four examples at resolution 512×512.

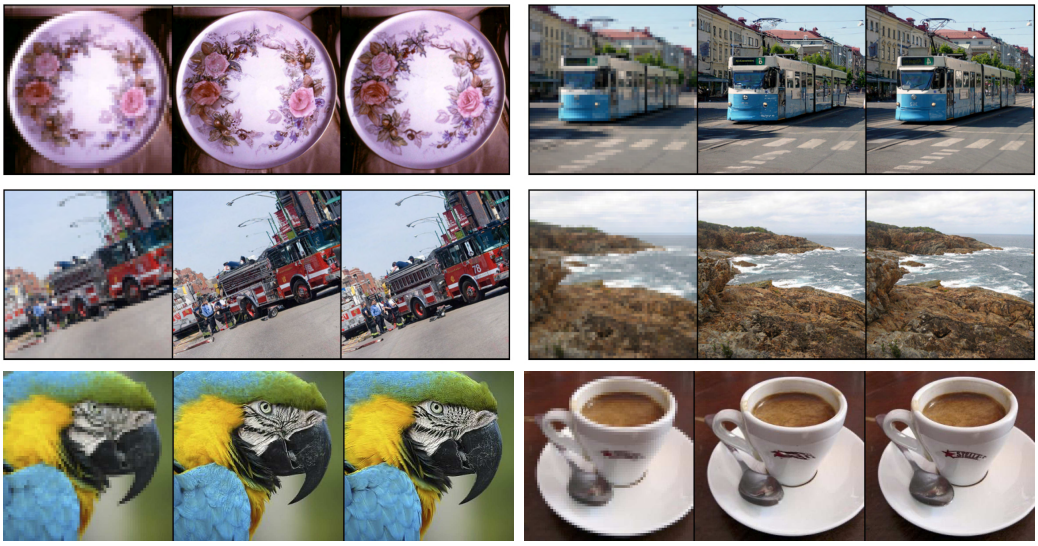

Figure 4: Super-resolution: *Top four rows*: Super-resolved images from resolution $64 \times 64 \mapsto 256 \times 256$, where the left-most image is the lower resolution version, the middle is the model output, and the right is the ground truth. Examples for $256 \times 256 \mapsto 512 \times 512$ are given in Fig. 6.

We provide some suggestions for specific forms of data-dependent coupling, such as choosing for ρ_0 a Gaussian distribution with mean and covariance adapted to samples from the target, and showed how they can be used in practical problem settings such as image inpainting and super-resolution. There are many interesting generative modeling problems that stand to benefit from the incorporation of data-dependent structure. In the sciences, one potential application is in molecule generation, where we can imagine using data-dependent base distributions to fix a chemical

backbone and vary functional groups. The dependency and conditioning structure needed to accomplish a task like this is similar to image inpainting. In machine learning, one potential application is in correcting autoencoding errors produced by an architecture such as a variational autoencoder (Kingma & Welling, 2013), where we could take the target density to be inputs to the autoencoder and the base density to be the output of the autoencoder.

Acknowledgements

We thank Raghav Singhal for insightful discussions. MG and RR are partly supported by the NIH/NHLBI Award R01HL148248, NSF Award 1922658 NRT-HDR: FUTURE Foundations, Translation, and Responsibility for Data Science, NSF CAREER Award 2145542, ONR N00014-23-1-2634, and Apple. MSA and NMB are funded by the ONR project entitled Mathematical Foundation and Scientific Applications of Machine Learning. EVE is supported by the National Science Foundation under Awards DMR-1420073, DMS-2012510, and DMS-2134216, by the Simons Collaboration on Wave Turbulence, Grant No. 617006, and by a Vannevar Bush Faculty Fellowship.

Impact Statement

While this paper presents work whose goal is to advance the field of machine learning, and there are many potential societal consequences of our work, we wish to highlight that generative models, as they are currently used, pose the risk of perpetuating harmful biases and stereotypes.

References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

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

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000. doi: 10.1007/s002110050002. URL https://doi.org/10.1007/s002110050002.

Chen, R. T. and Lipman, Y. Riemannian flow matching on general geometries. *arXiv preprint arXiv:2302.03660*, 2023.

Chen, R. T. Q. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Dao, Q., Phung, H., Nguyen, B., and Tran, A. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based generative modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 17695–17709. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/940392f5f32a7ade1cc201767cf83e31-Paper.pdf.

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

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density Estimation Using Real NVP. In *International Conference on Learning Representations*, pp. 32, 2017.

Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. Neural spline flows. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/7ac71d433f282034e088473244df8c02-Paper.pdf.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020b.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022a.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. *arXiv:2204.03458*, 2022b.

Hu, V. T., Zhang, D. W., Tang, M., Mettes, P., Zhao, D., and Snoek, C. G. Latent space editing in transformer-based flow matching. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Deep Networks with Stochastic Depth. *arXiv:1603.09382 [cs]*, July 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *arXiv [Preprint]*, 0, 2013. URL https://arxiv.org/1312.6114v10.

Klein, L., Krämer, A., and Noé, F. Equivariant flow matching, 2023.

Lee, S., Kim, B., and Ye, J. C. Minimizing trajectory curvature of ode-based generative models. *arXiv preprint arXiv:2301.12003*, 2023.

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

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling, 2022b. URL https://arxiv.org/abs/2210.02747.

Liu, G.-H., Vahdat, A., Huang, D.-A., Theodorou, E. A., Nie, W., and Anandkumar, A. I^2sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023a.

Liu, Q. Rectified flow: A marginal preserving approach to optimal transport, 2022. URL https://arxiv.org/abs/2209.14577.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022a. URL https://arxiv.org/abs/2209.03003.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022b.

Liu, X., Zhang, X., Ma, J., Peng, J., and Liu, Q. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2309.06380*, 2023b.

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

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. Multisample flow matching: Straightening flows with minibatch couplings. *arXiv preprint arXiv:2304.14772*, 2023.

Rezende, D. and Mohamed, S. Variational Inference with Normalizing Flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, June 2015.

Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2022.

Schneuing, A., Du, Y., Harris, C., Jamasb, A., Igashov, I., Du, W., Blundell, T., Lió, P., Gomes, C., Welling, M., et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

Shi, Y., Bortoli, V. D., Deligiannidis, G., and Doucet, A. Conditional simulation using diffusion schrödinger bridges. In *The 38th Conference on Uncertainty in Artificial Intelligence*, 2022. URL https://openreview.net/forum?id=H9Lu6P8sqec.

Shi, Y., Bortoli, V. D., Campbell, A., and Doucet, A. Diffusion schrödinger bridge matching, 2023.

Singhal, R., Goldstein, M., and Ranganath, R. Where to diffuse, how to diffuse, and how to get back: Automated learning for multivariate diffusions. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

Somnath, V. R., Pariset, M., Hsieh, Y.-P., Martinez, M. R., Krause, A., and Bunne, C. Aligned diffusion schrödinger bridges. In *The 39th Conference on Uncertainty in Artificial Intelligence*, 2023. URL https://openreview.net/forum?id=BkWFJN7_bQ.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 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*, 2020.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.),

Advances in Neural Information Processing Systems, volume 34, pp. 1415–1428. Curran Associates, Inc., 2021a. URL https://proceedings.neurips.cc/paper/2021/file/0a9fdbb17feb6ccb7ec405cfb85222c4-Paper.pdf.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

Tabak, E. G. and Turner, C. V. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2): 145–164, 2013. doi: https://doi.org/10.1002/cpa.21423. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.21423.

Tabak, E. G. and Vanden-Eijnden, E. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8(1):217–233, 2010. ISSN 15396746, 19450796. doi: 10.4310/CMS.2010.v8.n1.a11.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

Trippe, B. L., Yim, J., Tischer, D., Baker, D., Broderick, T., Barzilay, R., and Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Wu, L., Trippe, B. L., Naesseth, C. A., Blei, D. M., and Cunningham, J. P. Practical and asymptotically exact conditional sampling in diffusion models. *arXiv preprint arXiv:2306.17775*, 2023.

A. Omitted proofs with conditioning variables incorporated

In this Appendix we give the proofs of Theorem 3.1 and Corollary 3.1 in a more general setup in which we incorporate conditioning variables in the definition of the stochastic interpolant.

To this end, suppose that each data point $x_1^i \in \mathbb{R}^d$ in the data set comes with a label $\xi_i \in D$, such as a discrete class or a continuous embedding like a text caption, and let us assume that this data set comes from a PDF decomposed as $\rho_1(x_1|\xi)\eta(\xi)$, where $\rho_1(x_1|\xi)$ is the density of the data x_1 conditioned on their label ξ, and $\eta(\xi)$ is the density of the label. In the following, we will somewhat abuse notation and use $\eta(\xi)$ even when ξ is discrete (in which case, $\eta(\xi)$ is a sum of Dirac measures); we will however assume that $\rho_1(x_1|\xi)$ is a proper density. In this setup we can generalize Definition 3.1 as

Definition A.1 (Stochastic interpolant with coupling and conditioning)**.** *The stochastic interpolant I_t is the stochastic process defined as*

$$I_t = \alpha_t x_0 + \beta_t x_1 + \gamma_t z \qquad t \in [0,1], \tag{23}$$

where

- *α_t, β_t, and γ_t^2 are differentiable functions of time such that $\alpha_0 = \beta_1 = 1$, $\alpha_1 = \beta_0 = \gamma_0 = \gamma_1 = 0$, and $\alpha_t^2 + \beta_t^2 + \gamma_t^2 > 0$ for all $t \in [0,1]$.*
- *The pair (x_0, x_1) are jointly drawn from a conditional probability density $\rho(x_0, x_1|\xi)$ such that*

$$\int_{\mathbb{R}^d} \rho(x_0, x_1|\xi) dx_1 = \rho_0(x_0|\xi), \tag{24}$$

$$\int_{\mathbb{R}^d} \rho(x_0, x_1|\xi) dx_0 = \rho_1(x_1|\xi). \tag{25}$$

- *$z \sim \mathsf{N}(0, Id)$, independent of (x_0, x_1, ξ).*

Similarly, the functions (4) become

$$b_t(x,\xi) = \mathbb{E}(\dot{I}_t | I_t = x, \xi), \quad g_t(x,\xi) = \mathbb{E}(z | I_t = x, \xi) \tag{26}$$

where $\mathbb{E}(\cdot|I_t = x)$ denotes the expectation over $\rho(x_0, x_1|\xi)$ conditional on $I_t = x$, and Theorem 3.1 becomes:

Theorem A.1 (Transport equation with coupling and conditioning)**.** *The probability distribution of the stochastic interpolant I_t specified by Definition A.1 has a density $\rho_t(x|\xi)$ that satisfies $\rho_{t=0}(x|\xi) = \rho_0(x|\xi)$ and $\rho_{t=1}(x|\xi) = \rho_1(x|\xi)$, and solves the transport equation*

$$\partial_t \rho_t(x|\xi) + \nabla \cdot (b_t(x,\xi)\rho_t(x|\xi)) = 0, \tag{27}$$

where the velocity field is given in (26). Moreover, for every t such that $\gamma_t \neq 0$, the following identity for the score holds

$$\nabla \log \rho_t(x|\xi) = -\gamma_t^{-1} g_t(x,\xi). \tag{28}$$

The functions b and g are the unique minimizers of the objective

$$\begin{aligned} L_b(\hat{b}) &= \int_0^1 \mathbb{E}\left[|\hat{b}_t(I_t,\xi)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t,\xi)\right] dt, \\ L_g(\hat{g}) &= \int_0^1 \mathbb{E}\left[|\hat{g}_t(I_t,\xi)|^2 - 2z \cdot \hat{g}_t(I_t,\xi)\right] dt, \end{aligned} \tag{29}$$

where $\mathbb{E}$ denotes an expectation over $(x_0, x_1) \sim \rho(x_0, x_1|\xi)$, $\xi \sim \eta(\xi)$, and $z \sim \mathsf{N}(0, Id)$.

Note that the objectives (29) can readily be estimated in practice from samples $(x_0, x_1) \sim \rho(x_0, x_1|\xi)$, $z \sim \mathsf{N}(0,1)$, and $\xi \sim \eta(\xi)$, which will enable us to learn approximations for use in a generative model.

Proof. By definition of the stochastic interpolant given in (23), its characteristic function is given by

$$\mathbb{E}[e^{ik\cdot I_t}] = \int_{\mathbb{R}^d\times\mathbb{R}^d} e^{ik\cdot(\alpha_t x_0+\beta_t x_1)}\rho(x_0,x_1|\xi)dx_0dx_1e^{-\frac{1}{2}\gamma_t^2|k|^2}, \tag{30}$$

where we used $z \perp (x_0, x_1)$ and $z \sim \mathsf{N}(0, Id)$. The smoothness in k of (30) guarantees that the distribution of I_t has a density $\rho_t(x|\xi) > 0$ globally. By definition of I_t, this density $\rho_t(x|\xi)$ satisfies, for any suitable test function $\phi : \mathbb{R}^d \to \mathbb{R}$,

$$\int_{\mathbb{R}^d}\phi(x)\rho_t(x|\xi)dx = \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}\phi\left(I_t\right)\rho(x_0,x_1|\xi)(2\pi)^{-d/2}e^{-\frac{1}{2}|z|^2}dx_0dx_1dz. \tag{31}$$

Above, $I_t = \alpha_t x_0 + \beta_t x_1 + \gamma_t z$. Taking the time derivative of both sides

$$\begin{aligned}
&\int_{\mathbb{R}^d}\phi(x)\partial_t\rho_t(x|\xi)dx\\
&=\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d}\left(\dot\alpha_t x_0+\dot\beta_t x_1+\dot\gamma_t z\right)\cdot\nabla\phi\left(I_t\right)\rho(x_0,x_1|\xi)(2\pi)^{-d/2}e^{-\frac{1}{2}|z|^2}dx_0dx_1dz\\
&=\int_{\mathbb{R}^d}\mathbb{E}\left[\left(\dot\alpha_t x_0+\dot\beta_t x_1+\dot\gamma_t z\right)\cdot\nabla\phi(I_t)\right]\big|I_t=x\big]\rho_t(x|\xi)dx\\
&=\int_{\mathbb{R}^d}\mathbb{E}\left[\dot\alpha_t x_0+\dot\beta_t x_1+\dot\gamma_t z\big|I_t=x\right]\cdot\nabla\phi(x)\rho_t(x|\xi)dx
\end{aligned} \tag{32}$$

where we used the chain rule to get the first equality, the definition of the conditional expectation to get the second, and the tower property $\phi(I_t) = \phi(x)$ conditioned on $I_t = x$ to get the third. Since

$$\mathbb{E}\left[\dot\alpha_t x_0+\dot\beta_t x_1+\dot\gamma_t z\big|I_t=x\right] = b_t(x) \tag{33}$$

by the definition of b in (26), we can therefore write (32) as

$$\int_{\mathbb{R}^d}\phi(x)\partial_t\rho_t(x|\xi)dx = \int_{\mathbb{R}^d} b_t(x,\xi)\cdot\nabla\phi(x)\rho_t(x|\xi)dx. \tag{34}$$

This equation is (27) written in weak form.

To establish (28), note that if $\gamma_t > 0$, we have

$$\begin{aligned}
\mathbb{E}\left[ze^{i\gamma_t k\cdot z}\right] &= -\gamma_t^{-1}(i\partial_k)\mathbb{E}\left[e^{i\gamma_t k\cdot z}\right],\\
&= -\gamma_t^{-1}(i\partial_k)e^{-\frac{1}{2}\gamma_t^2|k|^2},\\
&= i\gamma_t k e^{-\frac{1}{2}\gamma_t^2|k|^2}.
\end{aligned} \tag{35}$$

As a result, using $z \perp (x_0, x_1)$, we have

$$\mathbb{E}\left[ze^{ik\cdot I_t}\right] = i\gamma_t k\mathbb{E}\left[e^{ik\cdot I_t}\right]. \tag{36}$$

Using the properties of the conditional expectation, the left-hand side of this equation can be written

$$\begin{aligned}
\mathbb{E}\left[ze^{ik\cdot I_t}\right] &= \int_{\mathbb{R}^d}\mathbb{E}\left[ze^{ik\cdot I_t}\big|I_t=x\right]\rho_t(x|\xi)dx,\\
&= \int_{\mathbb{R}^d}\mathbb{E}[z|I_t=x]e^{ik\cdot x}\rho_t(x,\xi)dx,\\
&= \int_{\mathbb{R}^d} g_t(x,\xi)e^{ik\cdot x}\rho_t(x,\xi)dx,
\end{aligned} \tag{37}$$

where we used the definition of g in (26) to get the last equality. Since the right-hand side of (36) is the Fourier transform of $-\gamma_t\nabla\rho_t(x|\xi)$, we deduce that

$$g_t(x,\xi)\rho_t(x|\xi) = -\gamma_t\nabla\rho_t(x|\xi) = -\gamma_t\nabla\log\rho_t(x|\xi)\,\rho_t(x|\xi). \tag{38}$$

Since $\rho_t(x|\xi) > 0$, this implies (28) when $\gamma_t > 0$.

Finally, to derive (29), notice that we can write

$$\begin{aligned} L_b(\hat{b}) &= \int_0^1 \mathbb{E}\left[|\hat{b}_t(I_t,\xi)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t,\xi)\right] dt, \\ &= \int_0^1 \int_{\mathbb{R}^d} \mathbb{E}\left[|\hat{b}_t(I_t,\xi)|^2 - 2\dot{I}_t \cdot \hat{b}_t(I_t,\xi)|I_t = x\right] \rho_t(x|\xi) dx dt \\ &= \int_0^1 \int_{\mathbb{R}^d} \left[|\hat{b}_t(x,\xi)|^2 - 2\mathbb{E}[\dot{I}_t|I_t = x] \cdot \hat{b}_t(x,\xi)\right] \rho_t(x|\xi) dx dt \\ &= \int_0^1 \int_{\mathbb{R}^d} \left[|\hat{b}_t(x,\xi)|^2 - 2b_t(x,\xi) \cdot \hat{b}_t(x,\xi)\right] \rho_t(x|\xi) dx dt \end{aligned} \tag{39}$$

where we used the definition of b in (26). The unique minimizer of this objective function is $\hat{b}_t(x,\xi) = b_t(x,\xi)$, and we can proceed similarly to show that the unique minimizers of $L_g(\hat{g})$ is $\hat{g}_t(x,\xi) = g_t(x,\xi)$, respectively. □

Theorem A.1 implies the following generalization of Corollary 3.1:

Corollary A.1 (Probability flow and diffusions with coupling and conditioning)**.** *The solutions to the probability flow equation*

$$\dot{X}_t = b_t(X_t,\xi) \tag{40}$$

enjoy the property that

$$X_{t=1} \sim \rho_1(x_1|\xi) \quad \textit{if} \quad X_{t=0} \sim \rho_0(x_0|\xi) \tag{41}$$

$$X_{t=0} \sim \rho_0(x_0|\xi) \quad \textit{if} \quad X_{t=1} \sim \rho_1(x_1|\xi) \tag{42}$$

In addition, for any $\epsilon_t \geq 0$, solutions to the forward SDE

$$dX_t^F = b_t(X_t^F,\xi)dt - \epsilon_t\gamma_t^{-1}g_t(X_t^F,\xi)dt + \sqrt{2\epsilon_t}dW_t, \tag{43}$$

enjoy the property that

$$X_{t=1}^F \sim \rho_1(x_1|\xi) \quad \textit{if} \quad X_{t=0}^F \sim \rho_0(x_0|\xi), \tag{44}$$

and solutions to the backward SDE

$$dX_t^R = b_t(X_t^R,\xi)dt + \epsilon_t\gamma_t^{-1}g_t(X_t^R,\xi)dt + \sqrt{2\epsilon_t}dW_t, \tag{45}$$

enjoy the property that

$$X_{t=0}^R \sim \rho_0(x_0|\xi) \quad \textit{if} \quad X_{t=1}^R \sim \rho_1(x_1|\xi). \tag{46}$$

Note that if we additionally draw ξ marginally from $\eta(\xi)$ when we generate the solution to these equations, we can also generate samples from the unconditional $\rho_0(x_0) = \int_D \rho_0(x_0|\xi)\eta(\xi)d\xi$ and $\rho_1(x_1) = \int_D \rho_1(x_1|\xi)\eta(\xi)d\xi$.

Proof. The probability flow ODE is the characteristic equation of the transport equation (27), which proves the statement about its solutions X_t. To establish the statement about the solution of the forward SDE (43), use expression (28) for $\nabla \log \rho_t(x,\xi)$ together with the identity $\Delta\rho_t(x,\xi) = \nabla\cdot(\nabla \log \rho_t(x,\xi)\, \rho_t(x,\xi))$ to write (27) as the forward Fokker-Planck equation

$$\partial_t\rho_t(x|\xi) + \nabla \cdot \left((b_t(x,\xi) - \epsilon_t\gamma_t^{-1}g_t(x,\xi))\rho_t(x|\xi)\right) = \epsilon_t\Delta\rho_t(x|\xi) \tag{47}$$

to be solved forward in time since $\epsilon_t > 0$. To establish the statement about the solution of the reversed SDE (45), proceed similarly to write (27) as the backward Fokker-Planck equation

$$\partial_t\rho_t(x|\xi) + \nabla \cdot \left((b_t(x,\xi) + \epsilon_t\gamma_t^{-1}g_t(x,\xi))\rho_t(x|\xi)\right) = -\epsilon_t\Delta\rho_t(x|\xi) \tag{48}$$

to be solved backward in time since $\epsilon_t > 0$. □

The generative model arising from Corollary 3.1 has an associated transport cost which is the subject of Corollary 3.1:

Proposition 3.1 (Control of transport cost)**.** *Let $X_t(x_0)$ be the solution to the probability flow ODE (8) for the initial condition $X_{t=0}(x_0) = x_0 \sim \rho_0$. Then*

$$\mathbb{E}_{x_0\sim\rho_0}\left[|X_{t=1}(x_0) - x_0|^2\right] \leq \int_0^1 \mathbb{E}[|\dot{I}_t|^2]dt < \infty \tag{15}$$

Proof. We have

$$\begin{aligned}\mathbb{E}_{x_0\sim\rho_0}\left[|X_{t=1}(x_0) - x_0|^2\right] &= \mathbb{E}_{x_0\sim\rho_0}\left[\Big|\int_0^1 b_t(X_t(x_0))dt\Big|^2\right] \\ &\leq \int_0^1 \mathbb{E}_{x_0\sim\rho_0}\left[|b_t(X_t(x_0))|^2\right]dt \\ &= \mathbb{E}\left[|b_t(I_t)|^2\right]\end{aligned} \tag{49}$$

where we used the probability flow equation (8) for X_t and the property that the law of $X_t(x_0)$ with $x_0 \sim \rho_0$ and I_t coincide. Using the definition of $b_t(x)$ in (4) and Jensen's inequality we have that

$$\mathbb{E}\left[|b_t(I_t)|^2\right] = \mathbb{E}\left[\left|\mathbb{E}[\dot{I}_t|I_t]\right|^2\right] \leq \mathbb{E}\left[\mathbb{E}\left[|\dot{I}_t|^2\middle|I_t\right]\right] = \mathbb{E}[|\dot{I}_t|^2] \tag{50}$$

where the last line is true by the tower property of the conditional expectation. Combining (49) and (50) establishes the bound in (15). □

B. Further experimental details

Architecture For the velocity model we use the U-net from (Ho et al., 2020b) as implemented in lucidrain's denoising-diffusion-pytorch repository; this variant of the architecture includes embeddings to condition on class labels. We use the following hyperparameters:

- Dim Mults: (1,1,2,3,4)
- Dim (channels): 256
- Resnet block groups: 8
- Leanred Sinusoidal Cond: True
- Learned Sinusoidal Dim: 32
- Attention Dim Head: 64
- Attention Heads: 4
- Random Fourier Features: False

Image-shaped conditioning in the Unet. For image-shaped conditioning, we follow (Ho et al., 2022a) and append upsampled low-resolution images to the input x_t at each time step to the velocity model. We also condition on the missingness masks for in-painting by appending them to x_t.

Optimization. We use Adam optimizer (Kingma & Ba, 2014), starting at learning rate 2e-4 with the StepLR scheduler which scales the learning rate by $\gamma = .99$ every $N = 1000$ steps. We use no weight decay. We clip gradient norms at $10,000$ (this is the norm of the entire set of parameters taken as a vector, the default type of norm clipping in PyTorch library).

Integration for sampling We use the Dopri solver from the torchdiffeq library (Chen, 2018).

Miscellaneous We use Pytorch library along with Lightning Fabric to handle parallelism.

Below we include additional experimental illustrations in the flavor of the figures in the main text.

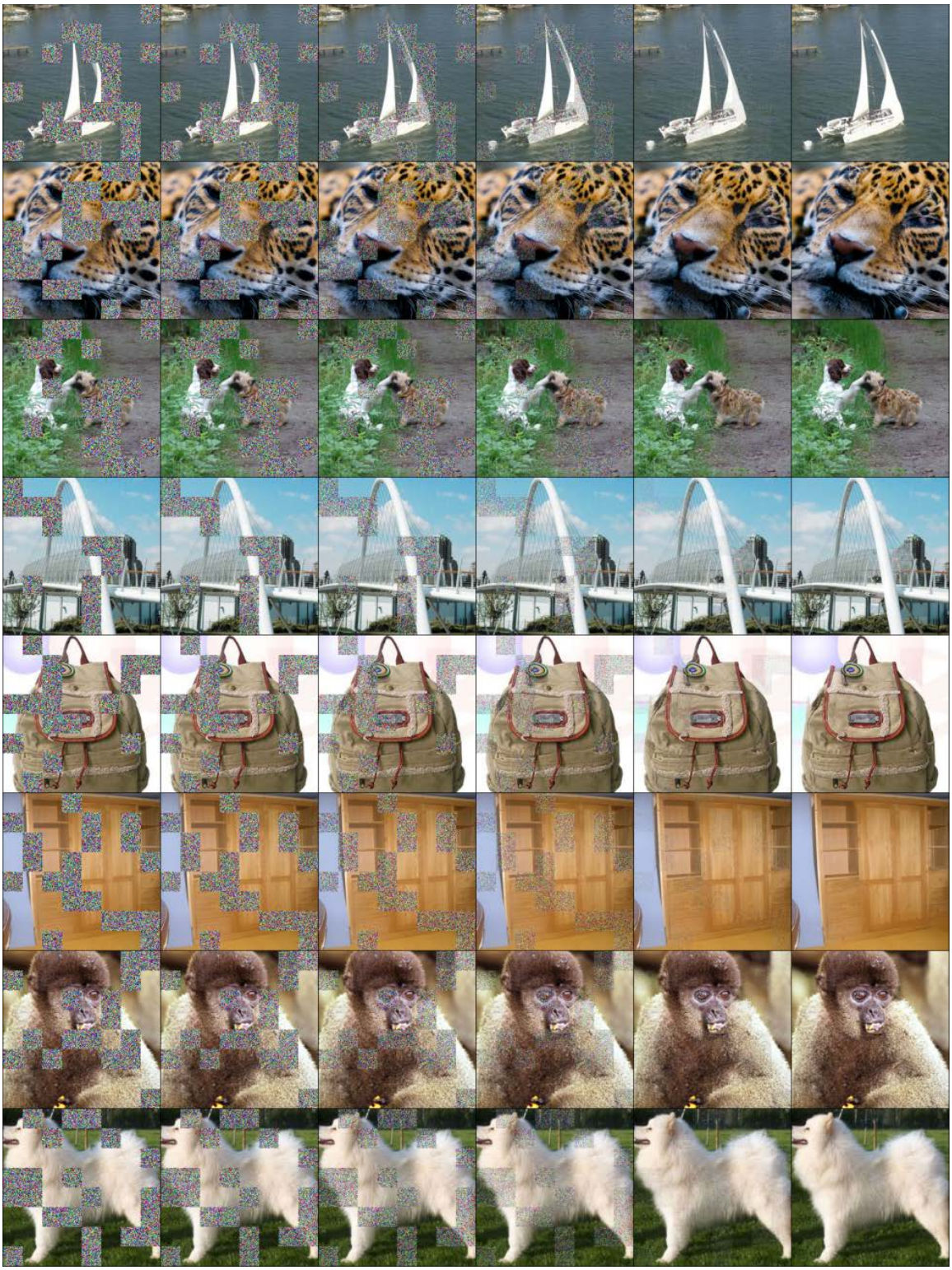

Figure 5: Additional examples of in-filling on the 256×256 resolution images, with temporal slices of the probability flow.

Figure 6: Super-resolution: *Top four rows*: Super-resolved images from resolution $256 \times 256 \mapsto 512 \times 512$, where the left-most image is the lower resolution version, the middle is the model output, and the right is the ground truth.