# FLOW MATCHING FOR GENERATIVE MODELING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We introduce a new paradigm for generative modeling built on Continuous Normalizing Flows (CNFs), allowing us to train CNFs at unprecedented scale. Specifically, we present the notion of Flow Matching (FM), a simulation-free approach for training CNFs based on regressing vector fields of fixed conditional probability paths. Flow Matching is compatible with a general family of Gaussian probability paths for transforming between noise and data samples—which subsumes existing diffusion paths as specific instances. Interestingly, we find that employing FM with diffusion paths results in a more robust and stable alternative for training diffusion models. Furthermore, Flow Matching opens the door to training CNFs with other, non-diffusion probability paths. An instance of particular interest is using Optimal Transport (OT) displacement interpolation to define the conditional probability paths. These paths are more efficient than diffusion paths, provide faster training and sampling, and result in better generalization. Training CNFs using Flow Matching on ImageNet leads to state-of-the-art performance in terms of both likelihood and sample quality, and allows fast and reliable sample generation using off-the-shelf numerical ODE solvers.

## 1 INTRODUCTION

Deep generative models are a class of deep learning algorithms aimed at estimating and sampling from an unknown data distribution. The recent influx of amazing advances in generative modeling, *e.g.*, for image generation Ramesh et al. (2022); Rombach et al. (2022), is mostly facilitated by the scalable and relatively stable training of diffusion-based models Ho et al. (2020); Song et al. (2020b). However, the restriction to simple diffusion processes leads to a rather confined space of sampling probability paths, resulting in very long training times and the need to adopt specialized methods (*e.g.*, Song et al. (2020a); Zhang & Chen (2022)) for efficient sampling.

In this work we consider the general and deterministic framework of Continuous Normalizing Flows (CNFs; Chen et al. (2018)). CNFs are capable of modeling arbitrary probability path and are in particular known to encompass the probability paths modeled by diffusion processes (Song et al., 2021). However, aside from diffusion that can be trained efficiently via, *e.g.*, denoising score matching (Vincent, 2011), no scalable CNF training algorithms are known. Indeed, maximum likelihood training (*e.g.*, Grathwohl et al. (2018)) require expensive numerical ODE simulations, while existing simulation-free methods either involve intractable integrals (Rozen et al., 2021) or biased gradients (Ben-Hamu et al., 2022).

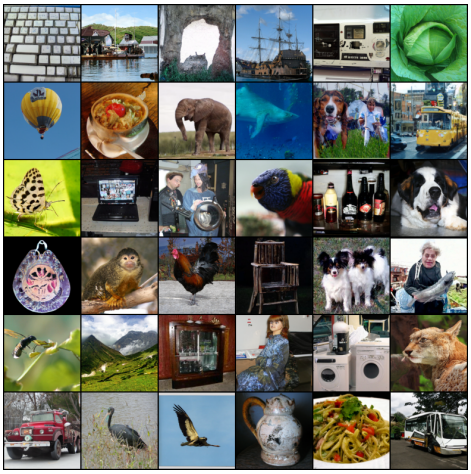


Figure 1: Unconditional ImageNet-128 samples of a CNF trained using Flow Matching with Optimal Transport probability paths.

The goal of this work is to propose Flow Matching (FM), an efficient simulation-free approach to training CNF models, allowing the adoption of general probability paths to supervise CNF training. Importantly, FM breaks the barriers for scalable CNF training beyond diffusion, and sidesteps the need to reason about diffusion processes to directly work with probability paths.

In particular, we propose the Flow Matching objective (Section 3), a simple and intuitive training objective to regress onto a target vector field that generates a desired probability path. We first show that we can construct such target vector fields through per-example (*i.e.*, conditional) formulations. Then, inspired by denoising score matching, we show that a per-example training objective, termed Conditional Flow Matching (CFM), provides equivalent gradients and does not require explicit knowledge of the intractable target vector field. Furthermore, we discuss a general family of per-example probability paths (Section 4) that can be used for Flow Matching, which subsumes existing diffusion paths as special instances. Even on diffusion paths, we find that using FM provides more robust and stable training, and achieves superior performance compared to score matching. Furthermore, this family of probability paths also includes a particularly interesting case: the vector field that corresponds to an Optimal Transport (OT) displacement interpolant (McCann, 1997). We find that conditional OT paths are simpler than diffusion paths, forming straight line trajectories whereas diffusion paths result in curved paths. These properties seem to empirically translate to faster training, faster generation, and better performance.

We empirically validate Flow Matching and the construction via Optimal Transport paths on ImageNet, a large and highly diverse image dataset. We find that we can easily train models to achieve state-of-the-art performance in both likelihood estimation and sample quality amongst competing generative models, through only deterministic simulations. Furthermore, we find that our models produce better trade-offs between computational cost and sample quality compared to prior methods. Figure 1 depicts selected unconditional ImageNet 128×128 samples from our model.

## 2 Preliminaries: Continuous Normalizing Flows

Let $\mathbb{R}^d$ denote the data space with data points $x = (x^1, \ldots, x^d) \in \mathbb{R}^d$. Two important objects we use in this paper are: the *probability density path* $p : [0,1] \times \mathbb{R}^d \to \mathbb{R}_{>0}$, which is a time dependent[1] probability density function, *i.e.*, $\int p_t(x)dx = 1$, and a *time-dependent vector field*, $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$. A vector field $v_t$ can be used to construct a time-dependent diffeomorphic map, called a *flow*, $\phi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, defined via the ordinary differential equation (ODE):

$$\frac{d}{dt}\phi_t(x) = v_t(\phi_t(x)) \tag{1}$$

$$\phi_0(x) = x \tag{2}$$

Previously, Chen et al. (2018) suggested modeling the vector field $v_t$ with a neural network, $v_t(x;\theta)$, where $\theta \in \mathbb{R}^p$ are its learnable parameters, which in turn leads to a deep parametric model of the flow $\phi_t$, called a *Continuous Normalizing Flow* (CNF). A CNF is used to reshape a simple prior density $p_0$ (*e.g.*, pure noise) to a more complicated one, $p_1$, via the push-forward equation

$$p_t = [\phi_t]_* p_0 \tag{3}$$

where the push-forward (or change of variables) operator $*$ is defined by

$$[\phi_t]_* p_0(x) = p_0(\phi_t^{-1}(x)) \det \left[ \frac{\partial \phi_t^{-1}}{\partial x}(x) \right]. \tag{4}$$

A vector field $v_t$ is said to *generate* a probability density path $p_t$ if its flow $\phi_t$ satisfies equation 3. One practical way to test if a vector field generates a probability path is using the continuity equation, which is a key component in our proofs, see Appendix B. We recap more information on CNFs, in particular how to compute the probability $p_1(x)$ at an arbitrary point $x \in \mathbb{R}^d$ in Appendix C.

## 3 Flow Matching

Let $x_1$ denote a random variable distributed according to some unknown data distribution $q(x_1)$. We assume we only have access to data samples from $q(x_1)$ but have no access to the density function itself. Furthermore, we let $p_t$ be a probability path such that $p_0$ is a simple distribution, *e.g.*, the standard normal distribution $p_0(x) = \mathcal{N}(x|0, I)$, and let $p_1$ be approximately equal in distribution to $q$. We will later discuss how to construct such a path. The Flow Matching objective is then designed to match this target probability path, which will allow us to flow from $p_0$ to $p_1$.

[1] We use subscript to denote the time parameter, *e.g.*, $p_t(x)$.

Given a target probability density path $p_t(x)$ and a corresponding vector field $u_t(x)$, which generates $p_t(x)$, we define the Flow Matching (FM) objective as

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,p_t(x)}\|v_t(x) - u_t(x)\|^2, \tag{5}$$

where $\theta$ denotes the learnable parameters of the CNF vector field $v_t$ (as defined in Section 2), $t \sim \mathcal{U}[0,1]$ (uniform distribution), and $x \sim p_t(x)$. Simply put, the FM loss regresses the vector field $u_t$ with a neural network $v_t$. Upon reaching zero loss, the learned CNF model will generate $p_t(x)$.

Flow Matching is a simple and attractive objective, but naïvely on its own, it is intractable to use in practice since we have no prior knowledge for what an appropriate $p_t$ and $u_t$ are. There are many choices of probability paths that can satisfy $p_1(x) \approx q(x)$, and more importantly, we generally don't have access to a closed form $u_t$ that generates the desired $p_t$. In this section, we show that we can construct both $p_t$ and $u_t$ using probability paths and vector fields that are only defined *per sample*, and an appropriate method of aggregation provides the desired $p_t$ and $u_t$. Furthermore, this construction allows us to create a much more tractable objective for Flow Matching.

### 3.1 Constructing $p_t, u_t$ from conditional probability paths and vector fields

A simple way to construct a target probability path is via a mixture of simpler probability paths: Given a particular data sample $x_1$ we denote by $p_t(x|x_1)$ a *conditional probability path* such that it satisfies $p_0(x|x_1) = p_0(x)$ at time $t = 0$, and we design $p_1(x|x_1)$ at $t = 1$ to be a distribution concentrated around $x = x_1$, *e.g.*, $p_1(x|x_1) = \mathcal{N}(x|x_1, \sigma^2 I)$, a normal distribution with $x_1$ mean and a sufficiently small standard deviation $\sigma > 0$. Marginalizing the conditional probability paths over $q(x_1)$ give rise to *the marginal probability path*

$$p_t(x) = \int p_t(x|x_1)q(x_1)dx_1, \tag{6}$$

where in particular at time $t = 1$, the marginal probability $p_1$ is a mixture distribution that closely approximates the data distribution $q$,

$$p_1(x) = \int p_1(x|x_1)q(x_1)dx_1 \approx q(x). \tag{7}$$

Interestingly, we can also define a *marginal vector field*, by "marginalizing" over the conditional vector fields in the following sense (we assume $p_t(x) > 0$ for all $t$ and $x$):

$$u_t(x) = \int u_t(x|x_1)\frac{p_t(x|x_1)q(x_1)}{p_t(x)}dx_1, \tag{8}$$

where $u_t(\cdot|x_1) : \mathbb{R}^d \to \mathbb{R}^d$ is a conditional vector field that generates $p_t(\cdot|x_1)$. It may not seem apparent, but this way of aggregating the conditional vector fields actually results in the correct vector field for modeling the marginal probability path.

Our first key observation is this:

*The marginal vector field (equation 8) generates the marginal probability path (equation 6).*

This provides a surprising connection between the conditional VFs (those that generate conditional probability paths) and the marginal VF (those that generate the marginal probability path). This connection allows us to break down the unknown and intractable marginal VF into simpler conditional VFs, which are much simpler to define as these only depend on a single data sample. We formalize this in the following theorem.

**Theorem 1.** *Given vector fields $u_t(x|x_1)$ that generate conditional probability paths $p_t(x|x_1)$, for any distribution $q(x_1)$, the marginal vector field $u_t$ in equation 8 generates the marginal probability path $p_t$ in equation 6, i.e., $u_t$ and $p_t$ satisfy the continuity equation (equation 25).*

The full proofs for our theorems are all provided in Appendix A.

### 3.2 CONDITIONAL FLOW MATCHING

Unfortunately, due to the intractable integrals in the definitions of the marginal probability path and VF (equations 6 and 8), it is still intractable to compute $u_t$, and consequently, intractable to naïvely compute an unbiased estimator of the original Flow Matching objective. Instead, we propose a simpler objective, which surprisingly will result in the same optima as the original objective. Specifically, we consider the *Conditional Flow Matching* (CFM) objective,

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)} \left\| v_t(x) - u_t(x|x_1) \right\|^2, \tag{9}$$

where $t \sim \mathcal{U}[0,1]$, $x_1 \sim q(x_1)$, and now $x \sim p_t(x|x_1)$. Unlike the FM objective, the CFM objective allows us to easily sample unbiased estimates as long as we can efficiently sample from $p_t(x|x_1)$ and compute $u_t(x|x_1)$, both of which can be easily done as they are defined on a per-sample basis.

Our second key observation is therefore:

*The FM (equation 5) and CFM (equation 9) objectives have identical gradients w.r.t. $\theta$.*

That is, optimizing the CFM objective is equivalent (in expectation) to optimizing the FM objective. Consequently, this allows us to train a CNF to generate the marginal probability path $p_t$—which in particular, approximates the unknown data distribution $q$ at $t$=1— without ever needing access to either the marginal probability path or the marginal vector field. We simply need to design suitable *conditional* probability paths and vector fields. We formalize this property in the following theorem.

**Theorem 2.** *Assuming that $p_t(x) > 0$ for all $x \in \mathbb{R}^d$ and $t \in [0,1]$, then, up to a constant independent of $\theta$, $\mathcal{L}_{CFM}$ and $\mathcal{L}_{FM}$ are equal. Hence,*

$$\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{CFM}(\theta). \tag{10}$$

## 4 CONDITIONAL PROBABILITY PATHS AND VECTOR FIELDS

The Conditional Flow Matching objective works with any choice of conditional probability path and conditional vector fields. In this section, we discuss the construction of $p_t(x|x_1)$ and $u_t(x|x_1)$ for a general family of Gaussian conditional probability paths. Namely, we consider conditional probability paths of the form

$$p_t(x|x_1) = \mathcal{N}(x \mid \mu_t(x_1), \sigma_t(x_1)^2 I), \tag{11}$$

where $\mu : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is the time-dependent mean of the Gaussian distribution, while $\sigma : [0,1] \times \mathbb{R} \to \mathbb{R}_{>0}$ describes a time-dependent scalar standard deviation (std). We set $\mu_0(x_1) = 0$ and $\sigma_0(x_1) = 1$, so that all conditional probability paths converge to the same standard Gaussian noise distribution at $t = 0$. We then set $\mu_1(x_1) = x_1$ and $\sigma_1(x_1) = \sigma_{\text{min}}$, which is set sufficiently small so that $p_1(x|x_1)$ is a highly concentrated Gaussian distribution centered at $x_1$.

There is potentially an infinite number of vector fields that generate any particular probability path, but the vast majority of these is due to the presence of components that leave the underlying distribution invariant—for instance, rotational components when the distribution is rotation-invariant—leading to unnecessary extra compute. We decide to use the simplest vector field corresponding to a canonical transformation for Gaussian distributions. Specifically, consider the flow

$$\psi_t(x) = \sigma_t(x_1)x + \mu_t(x_1). \tag{12}$$

When $x$ is distributed as a standard Gaussian, $\psi_t(x)$ is the affine transformation that maps to a normally-distributed random variable with mean $\mu_t(x_1)$ and std $\sigma_t(x_1)$. That is to say, according to equation 4, $\psi_t$ pushes the noise distribution $p_0(x|x_1)$ to $p_t(x|x_1)$, *i.e.*,

$$[\psi_t]_* \, p_0(\cdot|x_1) = p_t(\cdot|x_1). \tag{13}$$

This flow then provides a vector field that generates the conditional probability path:

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)|x_1). \tag{14}$$

Since $\psi_t$ is a simple (invertible) affine map we can use this equation to solve for $u_t$ in a closed form. Let $f'$ denote the derivative with respect to time, *i.e.*, $f' = \frac{d}{dt} f$, for a time-dependent function $f$.

**Theorem 3.** *Let $p_t(x|x_1)$ be a Gaussian probability path as in equation 11, and $\psi_t$ its corresponding flow map as in equation 12. Then, the unique vector field that defines $\psi_t$ has the form:*

$$u_t(x|x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}\left(x - \mu_t(x_1)\right) + \mu_t'(x_1). \tag{15}$$

*Consequently, $u_t(x|x_1)$ generates the Gaussian path $p_t(x|x_1)$.*

### 4.1 Special instances of Gaussian conditional probability paths

Our formulation is fully general for arbitrary (differentiable) functions $\mu_t(x_1)$ and $\sigma_t(x_1)$, and we can set them to any reasonable function. We first discuss the special cases that recover probability paths corresponding to previously-used diffusion processes. Since we directly work with probability paths, we can simply depart from reasoning about diffusion processes altogether. Therefore, in the second example below, we directly formulate a probability path based on the Wasserstein-2 optimal transport solution as an interesting instance.

**Example I: Diffusion conditional VFs** Diffusion models start with data points and gradually add noise until it approximates pure noise. These can be formulated as stochastic processes, which have strict requirements in order to obtain closed form representation at arbitrary times $t$, resulting in Gaussian conditional probability paths $p_t(x|x_1)$ with specific choices of mean $\mu_t(x_1)$ and std $\sigma_t(x_1)$ (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020b). For example, the reversed (noise→data) Variance Exploding (VE) path has the form

$$p_t(x) = \mathcal{N}(x|x_1, \sigma_{1-t}^2 I), \tag{16}$$

where $\sigma_t$ is an increasing function, $\sigma_0 = 0$, and $\sigma_1 \gg 1$. Next, equation 16 provides the choices of $\mu_t(x_1) = x_1$ and $\sigma_t(x_1) = \sigma_{1-t}$. Plugging these into equation 15 of Theorem 3 we get

$$u_t(x|x_1) = -\frac{\sigma_{1-t}'}{\sigma_{1-t}}(x - x_1). \tag{17}$$

The reversed (noise→data) Variance Preserving (VP) diffusion path has the form

$$p_t(x|x_1) = \mathcal{N}(x \mid \alpha_{1-t}x_1, \left(1 - \alpha_{1-t}^2\right) I), \text{where } \alpha_t = e^{-\frac{1}{2}T(t)}, T(t) = \int_0^t \beta(s)ds, \tag{18}$$

and $\beta$ is the noise scale function. Equation 18 provides the choices of $\mu_t(x_1) = \alpha_{1-t}x_1$ and $\sigma_t(x_1) = \sqrt{1 - \alpha_{1-t}^2}$. Plugging these into equation 15 of Theorem 3 we get

$$u_t(x|x_1) = \frac{\alpha_{1-t}'}{1 - \alpha_{1-t}^2}\left(\alpha_{1-t}x - x_1\right) = -\frac{T'(1-t)}{2}\left[\frac{e^{-T(1-t)}x - e^{-\frac{1}{2}T(1-t)}x_1}{1 - e^{-T(1-t)}}\right]. \tag{19}$$

Our construction of $u_t(x|x_1)$ does in fact coincide with the vector field previously used in the deterministic probability flow (Song et al. (2020b), equation 13) for these conditional diffusion processes; see details in Appendix D. Nevertheless, combining our method of marginalizing the conditional vector fields with the Flow Matching objective offers an attractive training alternative—which we find to be more stable and robust in our experiments—to existing score matching approaches.

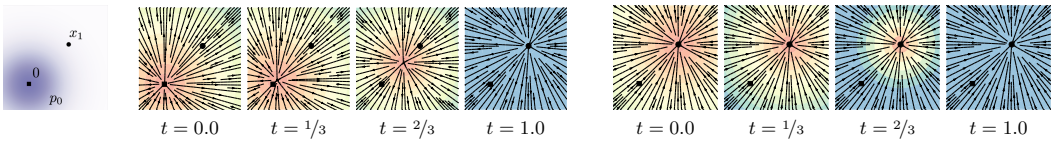
Another important observation is that, as these probability paths were previously derived as solutions of diffusion processes, they do not actually reach a true noise distribution in finite time. In practice, $p_0(x)$ is simply approximated by a suitable Gaussian distribution for sampling and likelihood evaluation. Instead, our construction provides full control over the probability path, and we can just directly set $\mu_t$ and $\sigma_t$, as we will do next.

**Example II: Optimal Transport conditional VFs** A more natural choice for conditional probability paths is to define the mean and the std to simply change linearly in time, *i.e.*,

$$\mu_t(x) = tx_1, \text{ and } \sigma_t(x) = 1 - (1 - \sigma_{\text{min}})t. \tag{20}$$

According to Theorem 3 this path is generated by the VF

$$u_t(x|x_1) = \frac{x_1 - (1 - \sigma_{\text{min}})x}{1 - (1 - \sigma_{\text{min}})t}, \tag{21}$$




Diffusion path – conditional score function

OT path – conditional vector field

Figure 2: Compared to the diffusion path's conditional score function, the OT path's conditional vector field has constant direction in time and is arguably simpler to fit with a parametric model. Note the blue color denotes larger magnitude while red color denotes smaller magnitude.

which, in contrast to the diffusion conditional VF (equation 19), is defined for all $t \in [0, 1]$. Allowing the mean and std to change linearly not only leads to simple and intuitive paths, but it is actually also optimal in a certain sense. Consider the conditional flow that corresponds to $u_t(x|x_1)$,

$$\psi_t(x) = (1 - (1 - \sigma_{\text{min}})t)x + tx_1. \tag{22}$$

This is in fact the Optimal Transport (OT) *displacement map* between the two Gaussians $p_0(x|x_1)$ and $p_1(x|x_1)$. The OT *interpolant*, which is a probability path, is defined to be (see Definition 1.1 in McCann (1997)):

$$p_t = [(1 - t)\text{id} + t\psi]_\star p_0 \tag{23}$$

where $\psi : \mathbb{R}^d \to \mathbb{R}^d$ is the OT map pushing $p_0$ to $p_1$, id denotes the identity map, *i.e.*, $\text{id}(x) = x$, and $(1 - t)\text{id} + t\psi$ is called the OT displacement map. Example 1.7 in McCann (1997) shows, that in our case of two Gaussians where the first is a standard one, the OT displacement map takes the form of equation 22.

Intuitively, particles under the OT displacement map always move in straight line trajectories and with constant speed. Figure 3 depicts sampling paths for the diffusion and OT conditional VFs. Interestingly, we find that sampling trajectory from diffusion paths can "overshoot" the final sample, resulting in unnecessary backtracking, whilst the OT paths are guaranteed to stay straight.

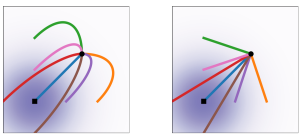



Figure 3: Diffusion and OT trajectories.

Figure 2 compares the diffusion conditional score function (the regression target in a typical diffusion methods), *i.e.*, $\nabla \log p_t(x|x_1)$ with $p_t$ defined as in equation 18, with the OT conditional VF (equation 21). The start ($p_0$) and end ($p_1$) Gaussians are identical in both examples. An interesting observation is that the OT VF has a constant direction in time, which arguably leads to a simpler regression task. This property can also be verified directly from equation 21 as the VF can be written in the form $u_t(x|x_1) = g(t)h(x|x_1)$. Figure 7 in the Appendix shows a visualization of the Diffusion VF. Lastly, we note that although the conditional flow is optimal, this by no means imply that the marginal VF is an optimal transport solution. Nevertheless, we expect the marginal vector field to remain relatively simple.

## 5 Related Work

Continuous Normalizing Flows were introduced in (Chen et al., 2018) as a continuous-time version of Normalizing Flows (see *e.g.*, Kobyzev et al. (2020); Papamakarios et al. (2021) for an overview). Originally, CNFs are trained with the maximum likelihood objective, but this involves expensive ODE simulations for the forward and backward propagation, resulting in high time complexity due to the sequential nature of ODE simulations. Although some works demonstrated the capability of CNF generative models for image synthesis (Grathwohl et al., 2018), scaling up to very high dimensional images is inherently difficult. A number of works attempted to regularize the ODE to be easier to solve, *e.g.*, using augmentation (Dupont et al., 2019), adding regularization terms (Yang & Karniadakis, 2019; Finlay et al., 2020; Onken et al., 2021; Tong et al., 2020; Kelly et al., 2020), or stochastically sampling the integration interval (Du et al., 2022). These works merely aim to regularize the ODE but do not change the fundamental training algorithm.

In order to speed up CNF training, some works have developed simulation-free CNF training frameworks by explicitly designing the target probability path and the dynamics. For instance, Rozen et al. (2021) consider a linear interpolation between the prior and the target density but involves integrals that were difficult to estimate in high dimensions, while Ben-Hamu et al. (2022) consider general

| Model | CIFAR-10 | | ImageNet 32×32 | | ImageNet 64×64 | |
|---|---|---|---|---|---|---|
| | NLL↓ | FID↓ | NLL↓ | FID↓ | NLL↓ | FID↓ |
| *Normalizing Flow* | | | | | | |
| FFJORD (Grathwohl et al., 2018) | 3.40 | | | | | |
| Glow (Kingma & Dhariwal, 2018) | 3.35 | | 4.09 | | 3.81 | |
| Residual Flow (Chen et al., 2019) | 3.28 | | 4.01 | | 3.76 | |
| Flow++ (Ho et al., 2019) | 3.09 | | 3.86 | | 3.69 | |
| *Variational Autoencoder* | | | | | | |
| NVAE (Vahdat & Kautz, 2020) | 2.91 | | 3.92 | | | |
| Very Deep VAE (Child, 2020) | 2.87 | | 3.80 | | 3.52 | |
| *Diffusion Model* | | | | | | |
| DDPM (Ho et al., 2020) | 3.75 | 3.17 | | | | |
| VDM (Kingma et al., 2021) | **2.65** | 7.41 | 3.72 | | 3.40 | |
| Score SDE (Song et al., 2020b) | 2.99 | **2.92** | | | | |
| Soft Truncation (Kim et al., 2022) | 2.88 | 3.45 | 3.85 | 8.42 | | |
| ScoreFlow (Song et al., 2021) | 2.81 | 5.40 | 3.76 | 10.18 | | |
| *Ablation* | | | | | | |
| Score Matching w/ Diffusion path | 3.16 | 21.96 | 3.57 | 22.38 | 3.40 | 19.61 |
| *Ours* | | | | | | |
| Flow Matching w/ Diffusion path | 3.10 | 10.31 | 3.56 | 8.02 | 3.33 | 16.06 |
| Flow Matching w/ OT path | 3.00 | 6.96 | **3.53** | **5.25** | **3.31** | **14.00** |

Table 1: Likelihood and quality of generated samples.

probability paths similar to this work but suffers from biased gradients in the stochastic minibatch regime. In contrast, the Flow Matching framework allows simulation-free training with unbiased gradients and readily scales to very high dimensions.

Another approach to simulation-free training relies on the construction of a diffusion process to indirectly define the target probability path (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song & Ermon, 2019). Song et al. (2020b) shows that diffusion models are trained using denoising score matching (Vincent, 2011), a conditional objective that provides unbiased gradients with respect to the score matching objective. Conditional Flow Matching draws inspiration from this result, but generalizes to matching vector fields. Due to the ease of scalability, diffusion models have received increased attention, producing a variety of improvements such as loss-rescaling (Song et al., 2021), learning the noise schedule (Nichol & Dhariwal, 2021; Kingma et al., 2021), and adding classifier guidance along with architectural improvements (Dhariwal & Nichol, 2021). While existing works make use of a connection between diffusion processes and continuous normalizing flows with the same probability path (Maoutsa et al., 2020b; Song et al., 2020b; 2021), we find that it is possible to completely sidestep the diffusion process construction and reason directly with probability paths. This allows us to generalize beyond the class of probability paths modeled by simple diffusion.

## 6 Experiments

We explore the empirical benefits of using Flow Matching, especially when compared to prior methods for training continuous-time generative models, such as maximum likelihood and score matching. We also ablate the choice of diffusion paths, particularly between the standard variance preserving diffusion path and the optimal transport path. We experiment with 2D (checkerboard pattern) and image datasets such as CIFAR-10 (Krizhevsky et al., 2009) and ImageNet at resolutions 32, 64, and 128 (Chrabaszcz et al., 2017; Deng et al., 2009). Finally, we discuss how sample generation is improved by directly parameterizing the generating vector field and using the Flow Matching objective. Unless otherwise specified, we evaluate likelihood and samples from the model using `dopri5` (Dormand & Prince, 1980) at absolute and relative tolerances of 1e-5. All implementation details can be found in Appendix E.

### 6.1 Density Modeling and Sample Quality on ImageNet

Table 1 summarizes our results alongside prior related works, reporting both negative log-likelihood (NLL) in units of bits per dimension (BPD) and sample quality as measured by the Frechet Inception Distance (FID; Heusel et al. (2017)). Note that these are *unconditional* models. We directly use the U-Net architecture from Dhariwal & Nichol (2021) with minimal changes. Although this

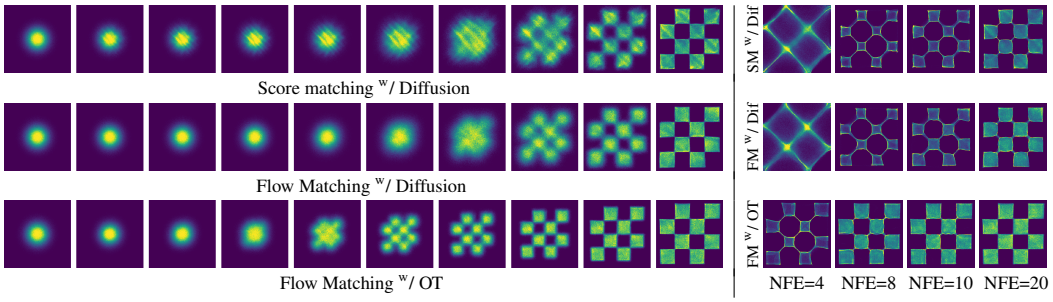



Figure 4: (*left*) Trajectories of CNFs trained with different objectives on 2D checkerboard data. The OT path introduces the checkerboard pattern much earlier, while FM results in more stable training. (*right*) FM with OT results in more efficient sampling, solved using the midpoint scheme.

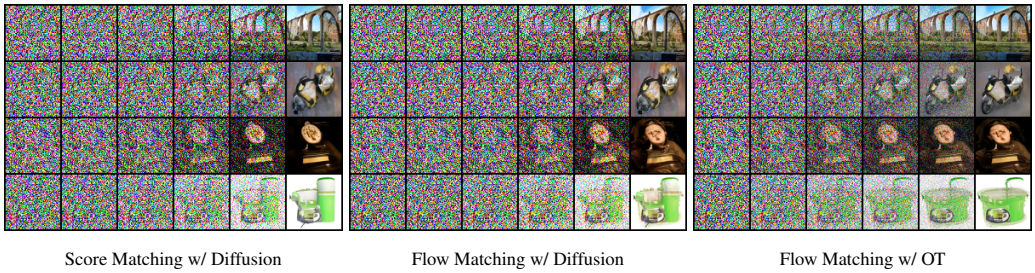
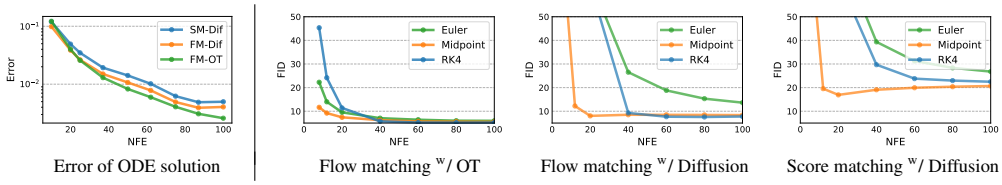



Figure 5: Sample paths from the same initial noise with models trained on ImageNet 64×64. The OT path reduces noise roughly linearly, while diffusion paths visibly remove noise only towards the end of the path. Note also the differences between the generated images.

architecture was not designed with CIFAR-10 in mind, we find that we can achieve very reasonable performance on par with Score SDE (Song et al., 2020b). On ImageNet, we obtain state-of-the-art results in terms of both density modeling and sample quality compared to competing methods. Furthermore, we trained using Flow Matching with the OT path on ImageNet at resolution 128×128, achieving **2.88** BPD and a FID of **21.52**. While NLL is not typically reported at this resolution, we believe our model reaches state-of-the-art performance, and this FID is competitive with strong likelihood-free models such as BigGAN (Lučić et al., 2019) at 25.3 (Casanova et al., 2021). Figures 9, 10, 11 in the Appendix depict non-curated samples from these models.

**Faster convergence** While existing works train diffusion models with a very high number of iterations (*e.g.*, 1.3m and 10m iterations are reported by Score Flow and VDM, respectively), we find that Flow Matching generally converges much faster. For example, Dhariwal & Nichol (2021) train ImageNet-64 for 540k iterations with 2k batch size, while FM trains (with the same architecture) for 157k iterations with equal batch size resulting in 70% less image throughput; for ImageNet-128 Dhariwal & Nichol (2021) train for 4.36m iterations with batch size 256, while FM (with 25% larger model) used 475k iterations with batch size 1.5k, *i.e.*, 35% less image throughput; see Table 3 for exact details. As the energy demand for training large deep learning models is increasing at a rapid pace (Amodei et al., 2018; Thompson et al., 2020), being able to train using less gradient updates / image throughput can lead to significant time and energy savings.

**Ablation experiments** We note that maximum likelihood training for Continuous Normalizing Flows (*e.g.*, as done in FFJORD) is not feasible for larger and more diverse datasets, so we mainly compare to the more competitive simulation-free methods. As our U-Net architecture may be different from prior works, we also perform our own ablation experiments on comparing models trained on a VP diffusion probability path and score matching. All models are trained using the same architecture, hyperparameter values and number of training iterations. We find that there is a noticeable difference when comparing Flow Matching to Score Matching. For instance, on ImageNet 64×64, while the score matching ablation matches the results of VDM, simply switching to Flow Matching immediately outperforms VDM by a large margin.




Error of ODE solution | Flow matching w/ OT | Flow matching w/ Diffusion | Score matching w/ Diffusion

Figure 6: Flow Matching, especially when using OT paths, allows us to use fewer evaluations for sampling while retaining similar numerical error (left) and sample quality (right). Results are shown for models trained on ImageNet 32×32, and numerical errors are for the midpoint scheme.

### 6.2 Sampling Efficiency

For sampling, we first draw a random noise sample $x_0 \sim \mathcal{N}(0, I)$ then compute $\phi_1(x_0)$ by solving equation 1 with the trained VF, $v_t$, on the interval $t \in [0, 1]$ using an ODE solver. While diffusion models can also be sampled through an SDE formulation, this can be highly inefficient and many methods that propose fast samplers (*e.g.*, Song et al. (2020a); Zhang & Chen (2022)) directly make use of the ODE perspective (see Appendix D). In part, this is due to ODE solvers being much more efficient—yielding lower error at similar computational costs (Kloeden et al., 2012)—and the multitude of available ODE solver schemes. When compared to our ablation models, we find that models trained using Flow Matching with the OT path always result in the most efficient sampler, regardless of ODE solver, as demonstrated next.

**Sample paths** We first qualitatively visualize the difference in sampling paths between diffusion and OT. Figure 5 shows samples from ImageNet-64 models using identical random seeds, where we find that the OT path model starts generating images sooner than the diffusion path models, where noise dominates the image until the very last time point. We additionally depict the probability density paths in 2D generation of a checkerboard pattern, Figure 4 (left), noticing a similar trend.

**Computational costs** Table 2 shows the average number of function evaluations (NFE) required for sampling at tolerance of 1e-5, showing that Flow Matching can result in a much lower number of function evaluations, where a model trained using score matching can use 3× as much compute. Furthermore, the NFE required for sampling from a model can drastically change during training for score matching, whereas the sampling cost stays constant when training with Flow Matching (see Figure 8 in Appendix).

| | CIFAR-10 | ImageNet 32×32 | ImageNet 64×64 |
|---|---|---|---|
| SM w/ Dif | 232 | 157 | 436 |
| FM w/ Dif | 179 | 165 | 179 |
| FM w/ OT | 156 | 164 | 162 |

Table 2: Sample costs (NFE) at fixed error tolerance, averaged over 50k samples.

**Low-cost sampling** We next switch to fixed-step solvers and compare low (≤100) NFE samples computed with the ImageNet-32 models from Table 1. In Figure 6 (left), we compare the per-pixel MSE of low NFE solutions compared with 1000 NFE solutions (we use 256 random noise seeds), and notice that the FM with OT model produces the best numerical error, in terms of computational cost, requiring roughly only 60% of the NFEs to reach the same error threshold as diffusion models. Secondly, Figure 6 (right) shows how FID changes as a result of the computational cost, where we find FM with OT is able to achieve decent FID even at very low NFE values, producing better trade-off between sample quality and cost compared to ablated models. Figure 4 (right) shows low-cost sampling effects for the 2D checkerboard experiment.

## 7 Conclusion

We introduced Flow Matching, a new simulation-free framework for training Continuous Normalizing Flow models, relying on conditional constructions to effortlessly scale to very high dimensions. Furthermore, the FM framework provides an alternative view on diffusion models, and suggests forsaking the stochastic/diffusion construction in favor of more directly specifying the probability path, allowing us to, *e.g.*, construct paths that allow faster sampling and/or improve generation. We experimentally showed the ease of training and sampling when using the Flow Matching framework, and in the future, we expect FM to open the door to allowing a multitude of probability paths (*e.g.*, non-isotropic Gaussians or more general kernels altogether). In particular, FM can be generalized to manifold data, where diffusion process are less understood and no longer simulation-free (*e.g.*, De Bortoli et al. (2022)), but probability paths can be easily defined (*e.g.*, Ben-Hamu et al. (2022)).

## REFERENCES


Dario Amodei, Danny Hernandez, Girish SastryJack, Jack Clark, Greg Brockman, and Ilya Sutskever. Ai and compute. https://openai.com/blog/ai-and-compute/, 2018.

Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Aditya Grover, Maximilian Nickel, Ricky T. Q. Chen, and Yaron Lipman. Matching normalizing flows and probability paths on manifolds. *arXiv preprint arXiv:2207.04711*, 2022.

Arantxa Casanova, Marlene Careil, Jakob Verbeek, Michal Drozdzal, and Adriana Romero Soriano. Instance-conditioned gan. *Advances in Neural Information Processing Systems*, 34:27517–27529, 2021.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Ricky T. Q. Chen, Jens Behrmann, David K Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. *Advances in Neural Information Processing Systems*, 32, 2019.

Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modeling. *arXiv preprint arXiv:2202.02763*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt.

John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

Shian Du, Yihong Luo, Wei Chen, Jian Xu, and Delu Zeng. To-flow: Efficient continuous normalizing flows with temporal optimization adjoint with moving speed, 2022. URL https://arxiv.org/abs/2203.10335.

Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural odes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/21be9a4bd4f81549a9d1d241981cec3c-Paper.pdf.

Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M. Oberman. How to train your neural ode: the world of jacobian and kinetic regularization. In *ICML*, pp. 3154–3164, 2020. URL http://proceedings.mlr.press/v119/finlay20a.html.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models, 2018. URL https://arxiv.org/abs/1810.01367.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Xi Chen, Aravind Srinivas, Yan Duan, and Pieter Abbeel. Flow++: Improving flow-based generative models with variational dequantization and architecture design. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2722–2730. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ho19a.html.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jacob Kelly, Jesse Bettencourt, Matthew J Johnson, and David K Duvenaud. Learning differential equations that are easy to solve. *Advances in Neural Information Processing Systems*, 33:4370–4380, 2020.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. In *International Conference on Machine Learning*, pp. 11201–11228. PMLR, 2022.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=2LdBqxc1Yv.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

Peter Eris Kloeden, Eckhard Platen, and Henri Schurz. *Numerical solution of SDE through computer experiments*. Springer Science & Business Media, 2012.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence*, 43 (11):3964–3979, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *International conference on machine learning*, pp. 4183–4192. PMLR, 2019.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker–planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, jul 2020a. doi: 10.3390/e22080802. URL https://doi.org/10.3390%2Fe22080802.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker–planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, 2020b.

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):9223–9232, May 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17113.

George Papamakarios, Eric T Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

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.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=qGvMv3undNJ.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 2020.

Alexander Tong, Jessie Huang, Guy Wolf, David Van Dijk, and Smita Krishnaswamy. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

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

Liu Yang and George E. Karniadakis. Potential flow generator with $l_2$ optimal transport regularity for generative models. *CoRR*, abs/1908.11462, 2019. URL http://arxiv.org/abs/1908.11462.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022.

## A THEOREM PROOFS

**Theorem 1.** *Given vector fields $u_t(x|x_1)$ that generate conditional probability paths $p_t(x|x_1)$, for any distribution $q(x_1)$, the marginal vector field $u_t$ in equation 8 generates the marginal probability path $p_t$ in equation 6, i.e., $u_t$ and $p_t$ satisfy the continuity equation (equation 25).*

*Proof.* To verify this, we check that $p_t$ and $u_t$ satisfy the continuity equation (equation 25):

$$\begin{aligned}\frac{d}{dt}p_t(x) &= \int \Big(\frac{d}{dt}p_t(x|x_1)\Big)q(x_1)dx_1 = -\int \mathrm{div}\Big(u_t(x|x_1)p_t(x|x_1)\Big)q(x_1)dx_1 \\ &= -\mathrm{div}\Big(\int u_t(x|x_1)p_t(x|x_1)q(x_1)dx_1\Big) = -\mathrm{div}\Big(u_t(x)p_t(x)\Big),\end{aligned}$$

where in the second equality we used the fact that $u_t(\cdot|x_1)$ generates $p_t(\cdot|x_1)$, in the last equality we used equation 8. Furthermore, the first and third equalities are justified by assuming the integrands satisfy the regularity conditions of the Leibniz Rule (for exchanging integration and differentiation). □

**Theorem 2.** *Assuming that $p_t(x) > 0$ for all $x \in \mathbb{R}^d$ and $t \in [0,1]$, then, up to a constant independent of $\theta$, $\mathcal{L}_{CFM}$ and $\mathcal{L}_{FM}$ are equal. Hence,*

$$\nabla_\theta \mathcal{L}_{FM}(\theta) = \nabla_\theta \mathcal{L}_{CFM}(\theta). \tag{10}$$

*Proof.* To ensure existence of all integrals and to allow the changing of integration order (by Fubini's Theorem) in the following we assume that $q(x)$ and $p_t(x|x_1)$ are decreasing to zero at a sufficient speed as $\|x\| \to \infty$, and that $u_t, v_t, \nabla_\theta v_t$ are bounded.

First, using the standard bilinearity of the 2-norm we have that

$$\begin{aligned}\|v_t(x) - u_t(x)\|^2 &= \|v_t(x)\|^2 - 2\left\langle v_t(x), u_t(x)\right\rangle + \|u_t(x)\|^2 \\ \|v_t(x) - u_t(x|x_1)\|^2 &= \|v_t(x)\|^2 - 2\left\langle v_t(x), u_t(x|x_1)\right\rangle + \|u_t(x|x_1)\|^2\end{aligned}$$

Next, remember that $u_t$ is independent of $\theta$ and note that

$$\begin{aligned}\mathbb{E}_{p_t(x)}\|v_t(x)\|^2 &= \int \|v_t(x)\|^2 p_t(x)dx = \int \|v_t(x)\|^2 p_t(x|x_1)q(x_1)dx_1dx \\ &= \mathbb{E}_{q(x_1),p_t(x|x_1)}\|v_t(x)\|^2,\end{aligned}$$

where in the second equality we use equation 6, and in the third equality we change the order of integration. Next,

$$\begin{aligned}\mathbb{E}_{p_t(x)}\left\langle v_t(x), u_t(x)\right\rangle &= \int \left\langle v_t(x), \frac{\int u_t(x|x_1)p_t(x|x_1)q(x_1)dx_1}{p_t(x)}\right\rangle p_t(x)dx \\ &= \int \left\langle v_t(x), \int u_t(x|x_1)p_t(x|x_1)q(x_1)dx_1\right\rangle dx \\ &= \int \left\langle v_t(x), u_t(x|x_1)\right\rangle p_t(x|x_1)q(x_1)dx_1dx \\ &= \mathbb{E}_{q(x_1),p_t(x|x_1)}\left\langle v_t(x), u_t(x|x_1)\right\rangle,\end{aligned}$$

where in the last equality we change again the order of integration. □

**Theorem 3.** *Let $p_t(x|x_1)$ be a Gaussian probability path as in equation 11, and $\psi_t$ its corresponding flow map as in equation 12. Then, the unique vector field that defines $\psi_t$ has the form:*

$$u_t(x|x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}\left(x - \mu_t(x_1)\right) + \mu_t'(x_1). \tag{15}$$

*Consequently, $u_t(x|x_1)$ generates the Gaussian path $p_t(x|x_1)$.*

*Proof.* For notational simplicity let $w_t(x) = u_t(x|x_1)$. Now consider equation 1:

$$\frac{d}{dt}\psi_t(x) = w_t(\psi_t(x)).$$

Since $\psi_t$ is invertible (as $\sigma_t(x_1) > 0$) we let $x = \psi^{-1}(y)$ and get

$$\psi_t'(\psi^{-1}(y)) = w_t(y), \tag{24}$$

where we used the apostrophe notation for the derivative to emphasis that $\psi_t'$ is evaluated at $\psi^{-1}(y)$. Now, inverting $\psi_t(x)$ provides

$$\psi_t^{-1}(y) = \frac{y - \mu_t(x_1)}{\sigma_t(x_1)}.$$

Differentiating $\psi_t$ with respect to $t$ gives

$$\psi_t'(x) = \sigma_t'(x_1)x + \mu_t'(x_1).$$

Plugging these last two equations in equation 24 we get

$$w_t(y) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}\left(y - \mu_t(x_1)\right) + \mu_t'(x_1)$$

as required. □

## B THE CONTINUITY EQUATION

One method of testing if a vector field $v_t$ generates a probability path $p_t$ is the continuity equation (Villani, 2009). It is a Partial Differential Equation (PDE) providing a necessary and sufficient condition to ensuring that a vector field $v_t$ generates $p_t$,

$$\frac{d}{dt}p_t(x) + \mathrm{div}(p_t(x)v_t(x)) = 0, \tag{25}$$

where the divergence operator, div, is defined with respect to the spatial variable $x = (x^1, \ldots, x^d)$, *i.e.*, $\mathrm{div} = \sum_{i=1}^{d} \frac{\partial}{\partial x^i}$.

## C COMPUTING PROBABILITIES OF THE CNF MODEL

We are given an arbitrary data point $x_1 \in \mathbb{R}^d$ and need to compute the model probability at that point, *i.e.*, $p_1(x_1)$. Below we recap how this can be done covering the basic relevant ODEs, the scaling of the divergence computation, taking into account data transformations (*e.g.*, centering of data), and Bits-Per-Dimension computation.

**ODE for computing $p_1(x_1)$** The continuity equation with equation 1 lead to the instantaneous change of variable (Chen et al., 2018; Ben-Hamu et al., 2022):

$$\frac{d}{dt}\log p_t(\phi_t(x)) + \mathrm{div}(v_t(\phi_t(x)) = 0.$$

Integrating $t \in [0, 1]$ gives:

$$\log p_1(\phi_1(x)) - \log p_0(\phi_0(x)) = -\int_0^1 \mathrm{div}(v_t(\phi_t(x)))dt \tag{26}$$

Therefore, the log probability can be computed together with the flow trajectory by solving the ODE:

$$\frac{d}{dt}\begin{bmatrix}\phi_t(x)\\ f(t)\end{bmatrix} = \begin{bmatrix}v_t(\phi_t(x))\\ -\mathrm{div}(v_t(\phi_t(x)))\end{bmatrix} \tag{27}$$

Given initial conditions

$$\begin{bmatrix}\phi_0(x)\\ f(0)\end{bmatrix} = \begin{bmatrix}x_0\\ c\end{bmatrix}. \tag{28}$$

the solution $[\phi_t(x), f(t)]^T$ is uniquely defined (up to some mild conditions on the VF $v_t$). Denote $x_1 = \phi_1(x)$, and according to equation 26,

$$f(1) = c + \log p_1(x_1) - \log p_0(x_0). \tag{29}$$

Now, we are given an arbitrary $x_1$ and want to compute $p_1(x_1)$. For this end, we will need to solve equation 27 in reverse. That is,

$$\frac{d}{ds}\begin{bmatrix}\phi_{1-s}(x)\\ f(1-s)\end{bmatrix} = \begin{bmatrix}-v_{1-s}(\phi_{1-s}(x))\\ \mathrm{div}(v_{1-s}(\phi_{1-s}(x)))\end{bmatrix} \tag{30}$$

and we solve this equation for $s \in [0, 1]$ with the initial conditions at $s = 0$:

$$\begin{bmatrix}\phi_1(x)\\ f(1)\end{bmatrix} = \begin{bmatrix}x_1\\ 0\end{bmatrix}. \tag{31}$$

From uniqueness of ODEs, the solution will be identical to the solution of equation 27 with initial conditions in equation 28 where $c = \log p_0(x_0) - \log p_1(x_1)$. This can be seen from equation 29 and setting $f(1) = 0$. Therefore we get that

$$f(0) = \log p_0(x_0) - \log p_1(x_1)$$

and consequently

$$\log p_1(x_1) = \log p_0(x_0) - f(0). \tag{32}$$

To summarize, to compute $p_1(x_1)$ we first solve the ODE in equation 30 with initial conditions in equation 31, and the compute equation 32.

**Unbiased estimator to $p_1(x_1)$** Solving equation 30 requires computation of div of VFs in $\mathbb{R}^d$ which is costly. Grathwohl et al. (2018) suggest to replace the divergence by the (unbiased) Hutchinson trace estimator,

$$\frac{d}{ds}\begin{bmatrix}\phi_{1-s}(x)\\ \tilde{f}(1-s)\end{bmatrix} = \begin{bmatrix}-v_{1-s}(\phi_{1-s}(x))\\ z^T Dv_{1-s}(\phi_{1-s}(x))z\end{bmatrix}, \tag{33}$$

where $z \in \mathbb{R}^d$ is a sample from a random variable such that $\mathbb{E}zz^T = I$. Solving the ODE in equation 33 exactly (in practice, with a small controlled error) with initial conditions in equation 31 leads to

$$\begin{aligned}
\mathbb{E}_z\left[\log p_0(x_0) - \tilde{f}(0)\right] &= \log p_0(x_0) - \mathbb{E}_z\left[\tilde{f}(0) - \tilde{f}(1)\right]\\
&= \log p_0(x_0) - \mathbb{E}_z\left[\int_0^1 z^T Dv_{1-s}(\phi_{1-s}(x))z\, ds\right]\\
&= \log p_0(x_0) - \int_0^1 \mathbb{E}_z\left[z^T Dv_{1-s}(\phi_{1-s}(x))z\right] ds\\
&= \log p_0(x_0) - \int_0^1 \mathrm{div}(v_{1-s}(\phi_{1-s}(x)))ds\\
&= \log p_0(x_0) - (f(0) - f(1))\\
&= \log p_0(x_0) - (\log p_0(x_0) - \log p_1(x_1))\\
&= \log p_1(x_1),
\end{aligned}$$

where in the third equality we switched order of integration assuming the sufficient condition of Fubini's theorem hold, and in the previous to last equality we used equation 29. Therefore the random variable

$$\log p_0(x_0) - \tilde{f}(0) \tag{34}$$

is an unbiased estimator for $\log p_1(x_1)$. To summarize, for a scalable unbiased estimation of $p_1(x_1)$ we first solve the ODE in equation 33 with initial conditions in equation 31, and then output equation 34.

**Transformed data** Often, before training our generative model we transform the data, *e.g.*, we scale and/or translate the data. Such a transformation is denoted by $\varphi^{-1} : \mathbb{R}^d \to \mathbb{R}^d$ and our generative model becomes a composition

$$\psi(x) = \varphi \circ \phi(x)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is the model we train. Given a prior probability $p_0$ we have that the push forward of this probability under $\psi$ (equation 3 and equation 4) takes the form

$$\begin{aligned} p_1(x) = \psi_* p_0(x) &= p_0(\phi^{-1}(\varphi^{-1}(x))) \det\left[D\phi^{-1}(\varphi^{-1}(x))\right] \det\left[D\varphi^{-1}(x)\right] \\ &= \left(\phi_* p_0(\varphi^{-1}(x))\right) \det\left[D\varphi^{-1}(x)\right] \end{aligned}$$

and therefore

$$\log p_1(x) = \log \phi_* p_0(\varphi^{-1}(x)) + \log \det\left[D\varphi^{-1}(x)\right].$$

For images $d = H \times W \times 3$ and we consider a transform $\phi$ that maps each pixel value from $[-1, 1]$ to $[0, 256]$. Therefore,

$$\varphi(y) = 2^7(y + 1)$$

and

$$\varphi^{-1}(x) = 2^{-7}x - 1$$

For this case we have

$$\log p_1(x) = \log \phi_* p_0(\varphi^{-1}(x)) - 7d \log 2. \tag{35}$$

**Bits-Per-Dimension (BPD) computation** BPD is defined by

$$\text{BPD} = \mathbb{E}_{x_1}\left[-\frac{\log_2 p_1(x_1)}{d}\right] = \mathbb{E}_{x_1}\left[-\frac{\log p_1(x_1)}{d \log 2}\right] \tag{36}$$

Following equation 35 we get

$$\text{BPD} = -\frac{\log \phi_* p_0(\varphi^{-1}(x))}{d \log 2} + 7$$

and $\log \phi_* p_0(\varphi^{-1}(x))$ is approximated using the unbiased estimator in equation 34 over the transformed data $\varphi^{-1}(x_1)$. Averaging the unbiased estimator on a large test test $x_1$ provides a good approximation to the test set BPD.

## D Diffusion conditional vector fields

We derive the vector field governing the Probability Flow ODE (equation 13 in Song et al. (2020b)) for the VE and VP diffusion paths (equation 18) and note that it coincides with the conditional vector fields we derive using Theorem 3, namely the vector fields defined in equations 16 and 19.

We start with a short primer on how to find a conditional vector field for the probability path described by the Fokker-Planck equation, then instantiate it for the VE and VP probability paths.

Since in the diffusion literature the diffusion process runs from data at time $t = 0$ to noise at time $t = 1$, we will need the following lemma to translate the diffusion VFs to our convention of $t = 0$ corresponds to noise and $t = 1$ corresponds to data:

**Lemma 1.** *Consider a flow defined by a vector field $u_t(x)$ generating probability density path $p_t(x)$. Then, the vector field $\tilde{u}_t(x) = -u_{1-t}(x)$ generates the path $\tilde{p}_t(x) = p_{1-t}(x)$ when initiated from $\tilde{p}_0(x) = p_1(x)$.*

*Proof.* We use the continuity equation (equation 25):

$$\begin{aligned} \frac{d}{dt}\tilde{p}_t(x) = \frac{d}{dt}p_{1-t}(x) &= -p'_{1-t}(x) \\ &= \text{div}(p_{1-t}(x)u_{1-t}(x)) \\ &= -\text{div}(\tilde{p}_t(x)(-u_{1-t}(x))) \end{aligned}$$

and therefore $\tilde{u}_t(x) = -u_{1-t}(x)$ generates $\tilde{p}_t(x)$. □

**Conditional VFs for Fokker-Planck probability paths** Consider a Stochastic Differential Equation (SDE) of the standard form

$$dy = f_t dt + g_t dw \tag{37}$$

with time parameter $t$, drift $f_t$, diffusion coefficient $g_t$, and $dw$ is the Wiener process. The solution $y_t$ to the SDE is a stochastic process, *i.e.*, a continuous time-dependent random variable, the probability density of which, $p_t(y_t)$, is characterized by the Fokker-Planck equation:

$$\frac{dp_t}{dt} = -\text{div}(f_t p_t) + \frac{g_t^2}{2}\Delta p_t \tag{38}$$

where $\Delta$ represents the Laplace operator (in $y$), namely $\text{div}\nabla$, where $\nabla$ is the gradient operator (also in $y$). Rewriting this equation in the form of the continuity equation can be done as follows (Maoutsa et al., 2020a):

$$\frac{dp_t}{dt} = -\text{div}\Big(f_t p_t - \frac{g^2}{2}\frac{\nabla p_t}{p_t}p_t\Big) = -\text{div}\Big((f_t - \frac{g_t^2}{2}\nabla \log p_t)p_t\Big) = -\text{div}\Big(w_t p_t\Big)$$

where the vector field

$$w_t = f_t - \frac{g_t^2}{2}\nabla \log p_t \tag{39}$$

satisfies the continuity equation with the probability path $p_t$, and therefore generates $p_t$.

**Variance Exploding (VE) path** The SDE for the VE path is

$$dy = \sqrt{\frac{d}{dt}\sigma_t^2}dw,$$

where $\sigma_0 = 0$ and increasing to infinity as $t \to 1$. The SDE is moving from data, $y_0$, at $t = 0$ to noise, $y_1$, at $t = 1$ with the probability path

$$p_t(y|y_0) = \mathcal{N}(y|y_0, \sigma_t^2 I).$$

The conditional VF according to equation 39 is:

$$w_t(y|y_0) = \frac{\sigma_t'}{\sigma_t}(y - y_0)$$

Using Lemma 1 we get that the probability path

$$\tilde{p}_t(y|y_0) = \mathcal{N}(y|y_0, \sigma_{1-t}^2 I)$$

is generated by

$$\tilde{w}_t(y|y_0) = -\frac{\sigma_{1-t}'}{\sigma_{1-t}}(y - y_0),$$

which coincides with equation 17.

**Variance Preserving (VP) path** The SDE for the VP path is

$$dy = -\frac{T'(t)}{2}y + \sqrt{T'(t)}dw,$$

where $T(t) = \int_0^t \beta(s)ds$, $t \in [0, 1]$. The SDE coefficients are therefore

$$f_s(y) = -\frac{T'(s)}{2}y, \quad g_s = \sqrt{T'(s)}$$

and

$$p_t(y|y_0) = \mathcal{N}(y|e^{-\frac{1}{2}T(t)}y_0, (1 - e^{-T(t)})I).$$

Plugging these choices in equation 39 we get the conditional VF

$$w_t(y|y_0) = \frac{T'(t)}{2}\left(\frac{y - e^{-\frac{1}{2}T(t)}y_0}{1 - e^{-T(t)}} - y\right) \tag{40}$$

Using Lemma 1 to reverse the time we get the conditional VF for the reverse probability path:

$$\begin{aligned}\tilde{w}_t(y|y_0) &= -\frac{T'(1-t)}{2}\left(\frac{y - e^{-\frac{1}{2}T(1-t)}y_0}{1 - e^{-T(1-t)}} - y\right) \\ &= -\frac{T'(1-t)}{2}\left[\frac{e^{-T(1-t)}y - e^{-\frac{1}{2}T(1-t)}y_0}{1 - e^{-T(1-t)}}\right],\end{aligned}$$

which coincides with equation 19.

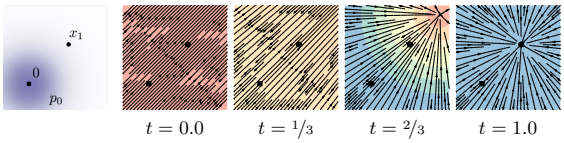



Figure 7: VP Diffusion path's conditional vector field. Compare to Figure 2.

## E Implementation details

For the 2D example we used an MLP with 5-layers of 512 neurons each, while for images we used the UNet architecture from Dhariwal & Nichol (2021). For images, we center crop images and resize to the appropriate dimension, whereas for the 32×32 and 64×64 resolutions we use the same pre-processing as (Chrabaszcz et al., 2017). The three methods (FM-OT, FM-Diffusion, and SM-Diffusion) are always trained on the same architecture, same hyper-parameters, and for the same number of epochs.

### E.1 Diffusion baseline

For the VP diffusion VF (equation 19) we use, as suggested in Song et al. (2020b), $\beta(s) = \beta_{\min} + s(\beta_{\max} - \beta_{\min})$ and consequently

$$T(s) = \int_0^s \beta(r)dr = s\beta_{\min} + \frac{1}{2}s^2(\beta_{\max} - \beta_{\min}),$$

where $\beta_{\min} = 0.1$, $\beta_{\max} = 20$ and time is sampled in $[0, 1-\epsilon]$, $\epsilon = 10^{-5}$ for training and likelihood and $\epsilon = 10^{-5}$ for sampling.

For the Score Matching (SM) loss we use:

$$\mathcal{L}_{\text{SM}}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)}\lambda(t)\left\|s_t(x) - \nabla \log p_t(x|x_1)\right\|^2, \tag{41}$$

where $s_t$ is the learnable score function, and the VP Diffusion path that is given by

$$p_t(y|y_0) = \mathcal{N}(y|e^{-\frac{1}{2}T(t)}y_0, (1 - e^{-T(t)})I)$$

and $\lambda(t) \propto 1/\mathbb{E}\left\|\nabla \log p_t(x|x_1)\right\|^2 \propto (1 - e^{-T(t)})$ for the VP case. Once the score has been trained samples are produced according to the ODE (equation 1) with the vector field

$$u_t(x) = -\frac{T'(1-t)}{2}\left[s_t(x) - x\right]. \tag{42}$$

### E.2 Training & evaluation details

We report the hyper-parameters used in Table 3. We use full 32 bit-precision for training CIFAR10 and ImageNet-32 and 16-bit mixed precision for training ImageNet-64/128. All models are trained using the Adam optimizer with the following parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.0, and $\epsilon = 1e{-}8$. All methods we trained (*i.e.*, FM-OT, FM-Diffusion, SM-Diffusion) using identical architectures, with the same parameters for the the same number of Epochs (see Table 3 for details). We use either a constant learning rate schedule or a polynomial decay schedule (see Table 3). The polynomial decay learning rate schedule includes a warm-up phase for a specified number of training steps. In the warm-up phase, the learning rate is linearly increased from $1e{-}8$ to the peak learning rate (specified in Table 3). Once the peak learning rate is achieved, it linearly decays the learning rate down to $1e{-}8$ until the final training step.

When reporting negative log-likelihood, we dequantize using the standard uniform dequantization. We report an importance-weighted estimate using

$$\log \frac{1}{K}\sum_{k=1}^{K} p_t(x + u_k), \text{ where } u_k \sim \mathcal{U}(0, 1), \tag{43}$$

| | CIFAR10 | ImageNet-32 | ImageNet-64 | ImageNet-128 |
|---|---|---|---|---|
| Channels | 256 | 256 | 192 | 256 |
| Depth | 2 | 3 | 3 | 3 |
| Channels multiple | 1,2,2,2 | 1,2,2,2 | 1,2,3,4 | 1,1,2,3,4 |
| Heads | 4 | 4 | 4 | 4 |
| Heads Channels | 64 | 64 | 64 | 64 |
| Attention resolution | 16 | 16,8 | 32,16,8 | 32,16,8 |
| Dropout | 0.0 | 0.0 | 0.0 | 0.0 |
| Effective Batch size | 256 | 1024 | 2048 | 1536 |
| GPUs | 2 | 4 | 16 | 32 |
| Epochs | 1000 | 200 | 250 | 571 |
| Iterations | 391k | 250k | 157k | 475k |
| Learning Rate | 5e-4 | 1e-4 | 1e-4 | 1e-4 |
| Learning Rate Scheduler | Polynomial Decay | Polynomial Decay | Constant | Polynomial Decay |
| Warmup Steps | 45k | 20k | - | 20k |

Table 3: Hyper-parameters used for training each model

| | **CIFAR-10** | | | **ImageNet 32×32** | | | **ImageNet 64×64** | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | $K$=1 | $K$=20 | $K$=50 | $K$=1 | $K$=5 | $K$=15 | $K$=1 | $K$=5 | $K$=10 |
| *Ablation* | | | | | | | | | |
| SM w/ Diffusion | 3.28 | 3.18 | 3.16 | 3.65 | 3.59 | 3.57 | 3.43 | 3.41 | 3.40 |
| *Ours* | | | | | | | | | |
| FM w/ Diffusion | 3.23 | 3.13 | 3.10 | 3.64 | 3.58 | 3.56 | 3.37 | 3.34 | 3.33 |
| FM w/ OT | 3.12 | 3.02 | 3.00 | 3.62 | 3.56 | 3.53 | 3.35 | 3.33 | 3.31 |

Table 4: Negative log-likelihood (in bits per dimension) on the test set with different values of $K$ using uniform dequantization.

with $x$ is in $\{0, \ldots, 255\}$. Estimated values for different values of $K$ are in Table 4.

When computing FID/Inception scores for CIFAR10, ImageNet-32/64 we use the TensorFlow GAN library [2]. To remain comparable to Dhariwal & Nichol (2021) for ImageNet-128 we use the evaluation script they include in their publicly available code repository [3].

## F ADDITIONAL TABLES AND FIGURES

[2] https://github.com/tensorflow/gan
[3] https://github.com/openai/guided-diffusion

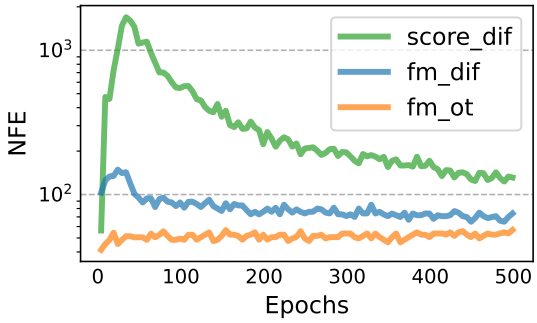



Figure 8: Function evaluations for sampling during training, for models trained on CIFAR-10 using `dopri5` solver with tolerance $1e^{-5}$.

Figure 9: Non-curated unconditional ImageNet-32 generated images of a CNF trained with FM-OT, 200 Epochs, 250k iterations, 3.53 BPD, 5.25 FID, see Table 1.

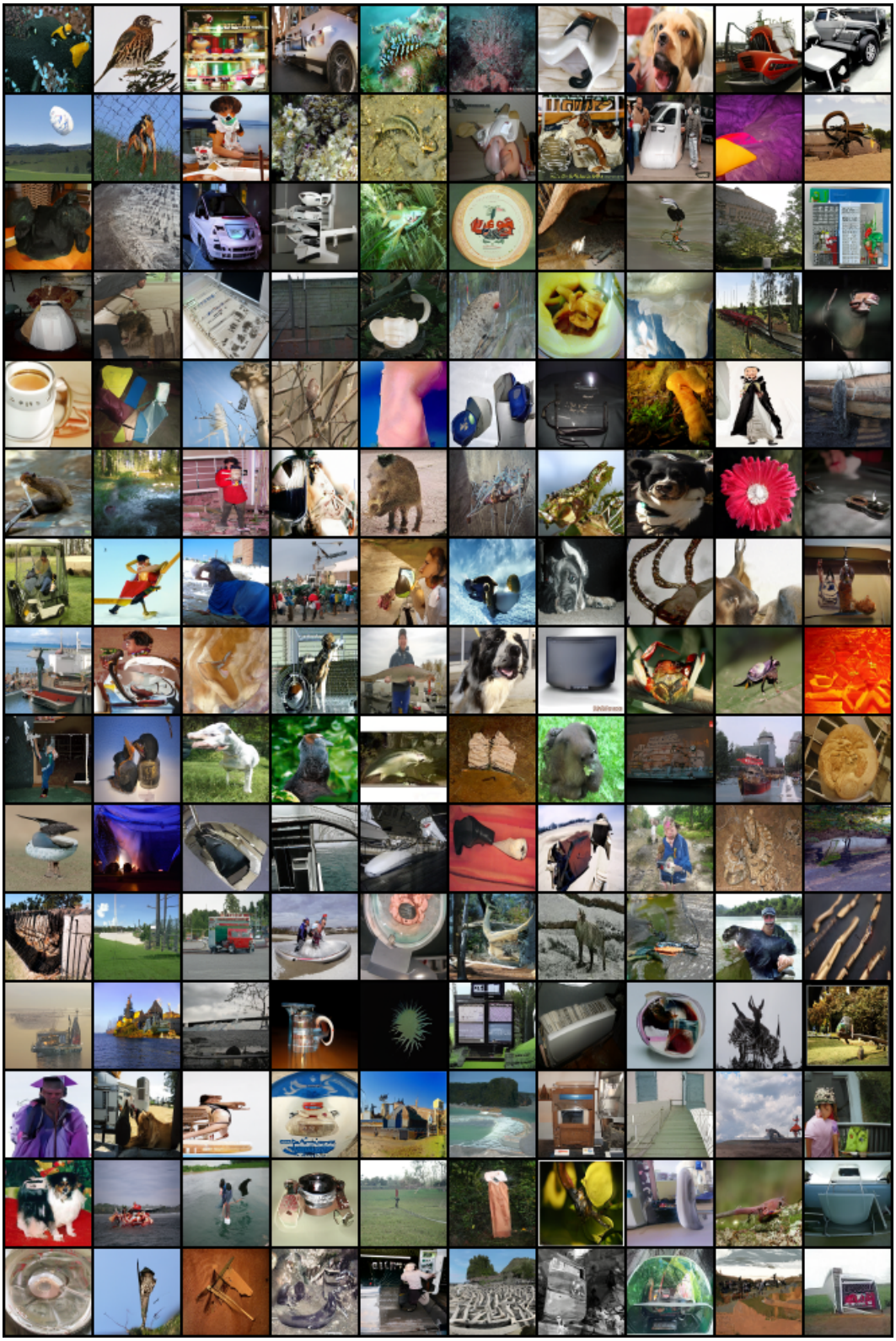


Figure 10: Non-curated unconditional ImageNet-64 generated images of a CNF trained with FM-OT, 250 Epochs, 157k iterations, 3.31 BPD, 14.00 FID, see Table 1.

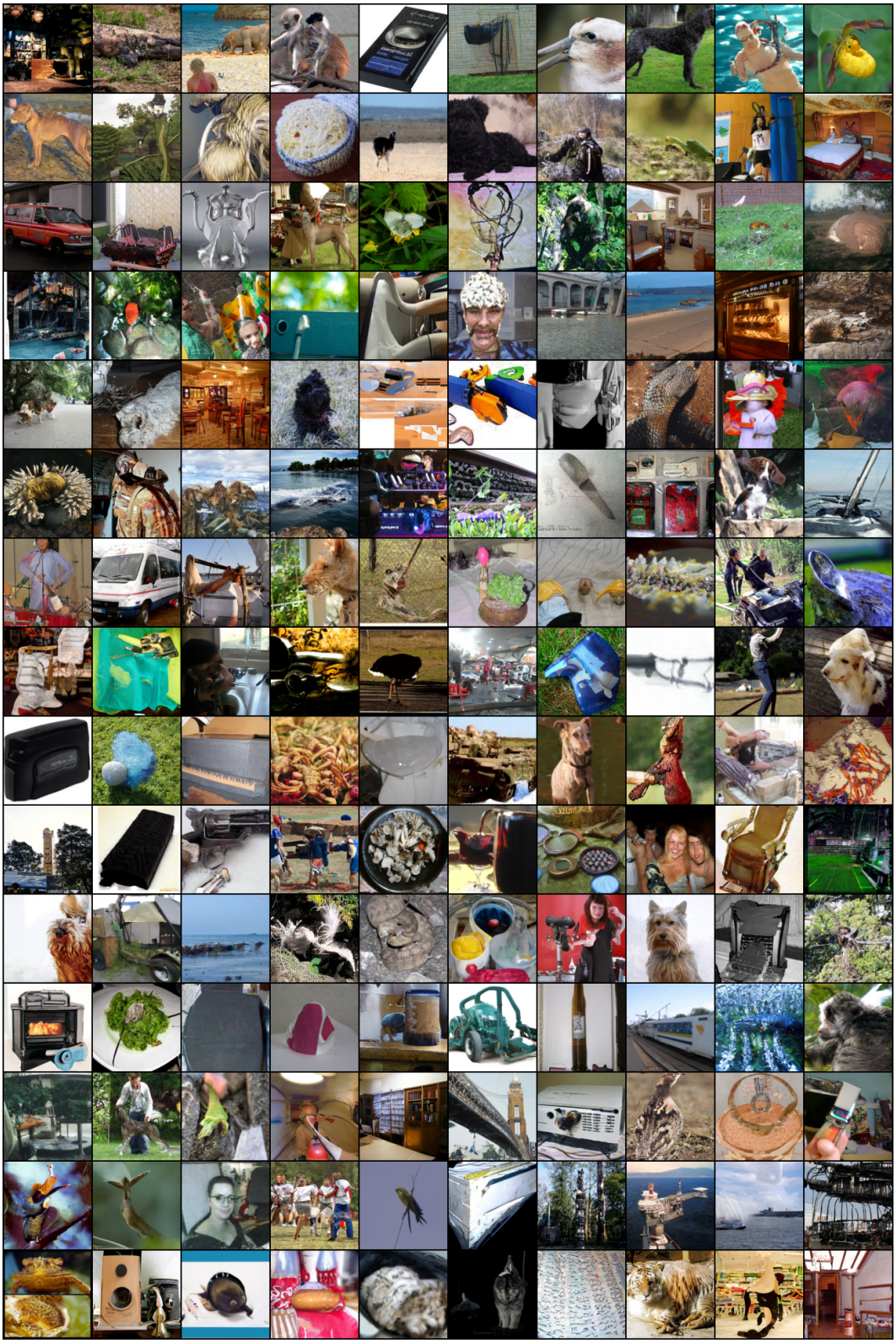
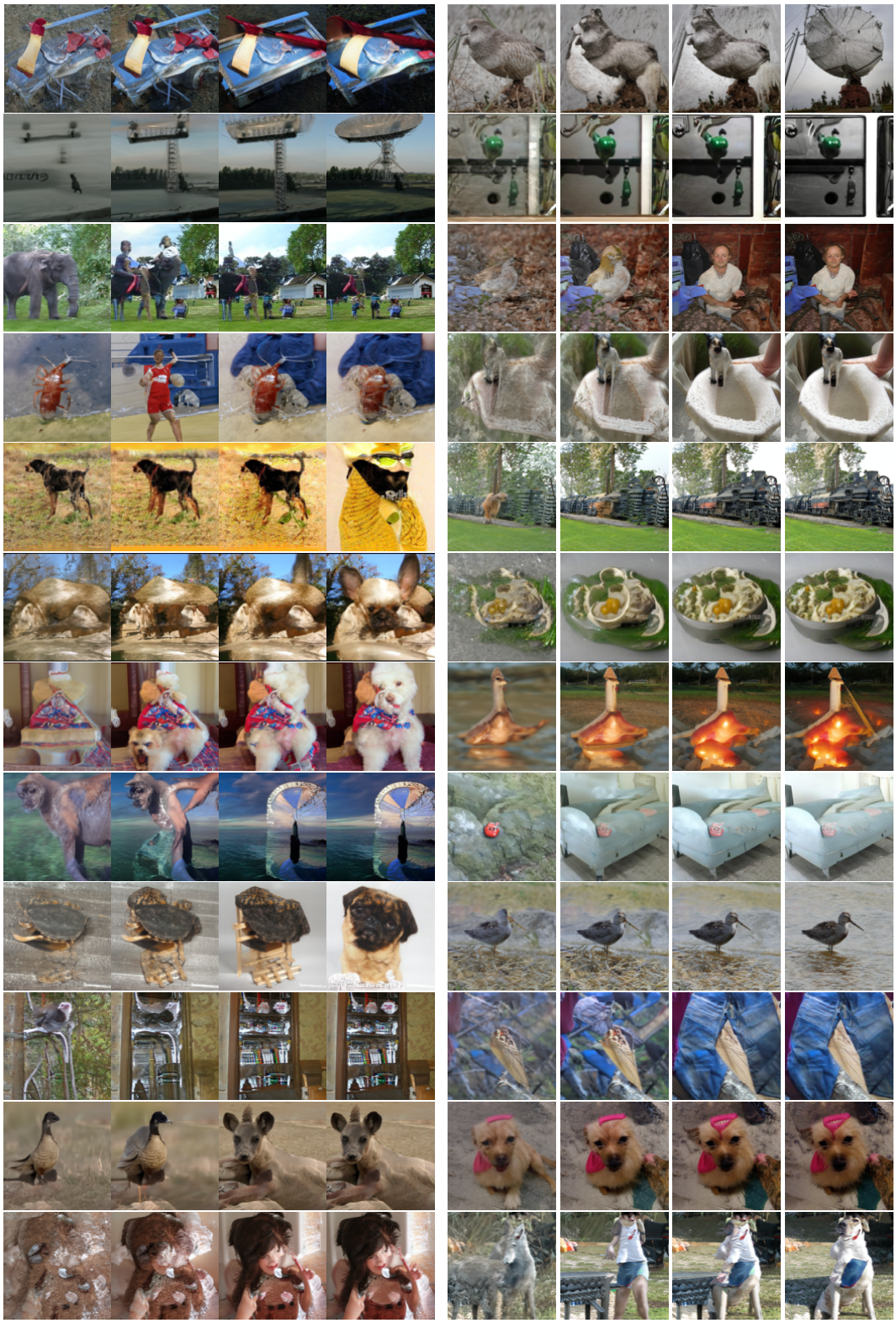


Figure 11: Non-curated unconditional ImageNet-128 generated images of a CNF trained with FM-OT, 527 Epochs, 475k iterations, 2.88 BPD, 21.52 FID.



NFE=10 NFE=20 NFE=40 NFE=100 NFE=10 NFE=20 NFE=40 NFE=100

Figure 12: Generated samples from the same initial noise, but with varying number of function evaluations (NFE). Flow matching with OT path trained on ImageNet-128.