Multisample Flow Matching: Straightening Flows with Minibatch Couplings

Aram-Alexandre Pooladian [*1 2] **Heli Ben-Hamu** [*3] **Carles Domingo-Enrich** [*1 4] **Brandon Amos** [1]
Yaron Lipman [1 3] **Ricky T. Q. Chen** [1]

Abstract

Simulation-free methods for training continuous-time generative models construct probability paths that go between noise distributions and individual data samples. Recent works, such as Flow Matching, derived paths that are optimal for each data sample. However, these algorithms rely on independent data and noise samples, and do not exploit underlying structure in the data distribution for constructing probability paths. We propose Multisample Flow Matching, a more general framework that uses non-trivial couplings between data and noise samples while satisfying the correct marginal constraints. At very small overhead costs, this generalization allows us to (i) reduce gradient variance during training, (ii) obtain straighter flows for the learned vector field, which allows us to generate high-quality samples using fewer function evaluations, and (iii) obtain transport maps with lower cost in high dimensions, which has applications beyond generative modeling. Importantly, we do so in a completely simulation-free manner with a simple minimization objective. We show that our proposed methods improve sample consistency on downsampled ImageNet data sets, and lead to better low-cost sample generation.

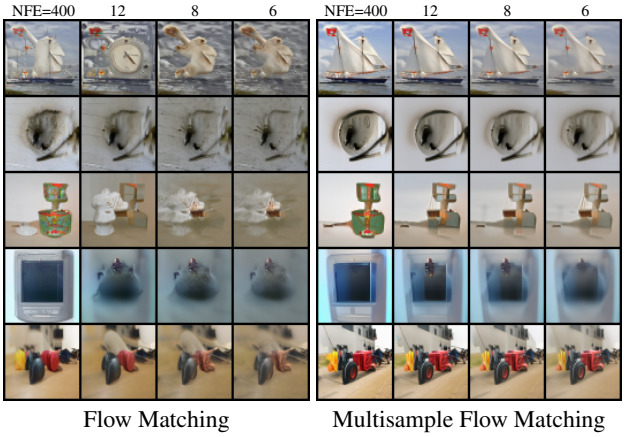


Figure 1. Multisample Flow Matching trained with batch optimal couplings produces more consistent samples across varying NFEs. Note that both flows on each row start from the same noise sample.

1. Introduction

Deep generative models offer an attractive family of paradigms that can approximate a data distribution and produce high quality samples, with impressive results in recent years (Ramesh et al., 2022; Saharia et al., 2022; Gafni et al., 2022). In particular, these works have made use of simulation-free training methods for diffusion models (Ho et al., 2020; Song et al., 2021b). A number of works have also adopted and generalized these simulation-free methods (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023; Liu et al., 2022; Neklyudov et al., 2022) for continuous normalizing flows (CNF; Chen et al. (2018)), a family of continuous-time deep generative models that parameterizes a vector field which flows noise samples into data samples.

Recently, Lipman et al. (2023) proposed *Flow Matching* (FM), a method to train CNFs based on constructing explicit *conditional probability paths* between the noise distribution (at time $t = 0$) and each data sample (at time $t = 1$). Furthermore, they showed that these conditional probability paths can be taken to be the optimal transport path when the noise distribution is a standard Gaussian, a typical assumption in generative modeling. However, this does not imply that the *marginal probability path* (marginalized over the data distribution) is anywhere close to the optimal transport path between the noise and data distributions.

Most existing works, including diffusion models and Flow Matching, have only considered conditional sample paths where the endpoints (a noise sample and a data sample) are sampled independently. However, this results in non-zero gradient variances even at convergence, slow training times, and in particular limits the design of probability paths.

[*]Equal contribution [1]Meta AI (FAIR) [2]Center for Data Science, NYU [3]Weizmann Institute of Science [4]Courant Institute of Mathematical Sciences, NYU. Correspondence to: Ricky T. Q. Chen <rtqichen@meta.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

In turn, it becomes difficult to create paths that are fast to simulate, a desirable property for both likelihood evaluation and sampling.

Contributions: We present a tractable instance of Flow Matching with joint distributions, which we call *Multisample Flow Matching*. Our proposed method generalizes the construction of probability paths by considering non-independent couplings of k-sample empirical distributions.

Among other theoretical results, we show that if an appropriate optimal transport (OT) inspired coupling is chosen, then sample paths become straight as the batch size $k \to \infty$, leading to more efficient simulation. In practice, we observe both improved sample quality on ImageNet using adaptive ODE solvers and using simple Euler discretizations with a low budget number of function evaluations. Empirically, we find that on ImageNet, we can *reduce the required sampling cost by 30% to 60%* for achieving a low Fréchet Inception Distance (FID) compared to a baseline Flow Matching model, while introducing only 4% more training time. This improvement in sample efficiency comes at no degradation in performance, *e.g.* log-likelihood and sample quality.

Within the deep generative modeling paradigm, this allows us to regularize towards the optimal vector field in a *completely simulation-free manner* (unlike *e.g.* Finlay et al. (2020b); Liu et al. (2022)), and avoids adversarial formulations (unlike *e.g.* Makkuva et al. (2020); Albergo & Vanden-Eijnden (2023)). In particular, we are the first work to be able to make use of solutions from optimal solutions on minibatches while preserving the correct marginal distributions, whereas prior works would only fit to the barycentric average (see detailed discussion in Section 5.1). Beyond generative modeling, we also show how our method can be seen as a new way to compute approximately optimal transport maps between arbitrary distributions in settings where the cost function is completely unknown and only minibatch optimal transport solutions are provided.

2. Preliminaries

2.1. Continuous Normalizing Flow

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 path* $p_t : \mathbb{R}^d \to \mathbb{R}_{>0}$, which is a time dependent (for $t \in [0,1]$) probability density function, *i.e.*, $\int p_t(x)dx = 1$, and a *time-dependent vector field*, $u_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$. A vector field u_t constructs a time-dependent diffeomorphic map, called a *flow*, $\psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, defined via the ordinary differential equation (ODE):

$$\frac{d}{dt}\psi_t(x_0) = u_t(\psi_t(x_0)), \quad \psi_0(x_0) = x_0 . \tag{1}$$

To create a deep generative model, Chen et al. (2018) suggested modeling the vector field u_t with a neural network, leading to a deep parametric model of the flow ψ_t, referred to as a *Continuous Normalizing Flow* (CNF). A CNF is often used to transform a density p_0 to a different one, p_1, via the push-forward equation

$$p_t(x) = [\psi_t]_\sharp p_0(x) = p_0(\psi_t^{-1}(x)) \left| \det \left[\frac{\partial \psi_t^{-1}}{\partial x}(x) \right] \right|, \tag{2}$$

where the second equality defines the push-forward (or change of variables) operator $\sharp$. A vector field u_t is said to *generate* a probability path p_t if its flow ψ_t satisfies (2).

2.2. Flow Matching

A simple simulation-free method for training CNFs is the *Flow Matching* algorithm (Lipman et al., 2023), which regresses onto an (implicitly-defined) target vector field that generates the desired probability density path p_t. Given two marginal distributions $q_0(x_0)$ and $q_1(x_1)$ for which we would like to learn a CNF to transport between, Flow Matching seeks to optimize the simple regression objective,

$$\mathbb{E}_{t,p_t(x)} \left\| v_t(x;\theta) - u_t(x) \right\|^2 , \tag{3}$$

where $v_t(x;\theta)$ is the parametric vector field for the CNF, and $u_t(x)$ is a vector field that generates a probability path p_t under the two marginal constraints that $p_{t=0} = q_0$ and $p_{t=1} = q_1$. While Equation (3) is the ideal objective function to optimize, not knowing (p_t, u_t) makes this computationally intractable.

Lipman et al. (2023) proposed a tractable method of optimizing (3), which first defines *conditional* probability paths and vector fields, such that when marginalized over $q_0(x_0)$ and $q_1(x_1)$, provide both $p_t(x)$ and $u_t(x)$. When targeted towards generative modeling, $q_0(x_0)$ is a simple noise distribution and easy to directly enforce, leading to a one-sided construction:

$$p_t(x) = \int p_t(x|x_1) q_1(x_1)\, dx_1 \tag{4}$$

$$u_t(x) = \int u_t(x|x_1) \frac{p_t(x|x_1) q_1(x_1)}{p_t(x)}\, dx_1, \tag{5}$$

where the conditional probability path is chosen such that

$$p_{t=0}(x|x_1) = q_0(x) \quad \text{and} \quad p_{t=1}(x|x_1) = \delta(x - x_1), \tag{6}$$

where $\delta(x-a)$ is a Dirac mass centered at $a \in \mathbb{R}^d$. By construction, $p_t(x|x_1)$ now satisfies both marginal constraints.

Lipman et al. (2023) shows that if $u_t(x|x_1)$ generates $p_t(x|x_1)$, then the marginalized $u_t(x)$ generates $p_t(x)$, and furthermore, one can train using the much simpler objective of *Conditional Flow Matching* (CFM):

$$\mathbb{E}_{t,q_1(x_1),p_t(x|x_1)} \left\| v_t(x;\theta) - u_t(x_t|x_1) \right\|^2 , \tag{7}$$

with $x_t = \psi_t(x_0|x_1)$; see 2.2.1 for more details. Note that this objective has the same gradient with respect to the model parameters θ as Eq. (3) (Lipman et al., 2023, Theorem 2).

2.2.1. Conditional OT (CondOT) path

One particular choice of conditional path $p_t(x|x_1)$ is to use the flow that corresponds to the optimal transport displacement interpolant (McCann, 1997) when $q_0(x_0)$ is the standard Gaussian, a common convention in generative modeling. The vector field that corresponds to this is

$$u_t(x_t|x_1) = \frac{x_1 - x}{1-t}. \tag{8}$$

Using this conditional vector field in (1), this gives the conditional flow

$$x_t = \psi_t(x_0|x_1) = (1-t)x_0 + tx_1\,. \tag{9}$$

Substituting (9) into (8), one can also express the value of this vector field using a simpler expression,

$$u_t(x_t|x_1) = x_1 - x_0\,. \tag{10}$$

It is evident that this results in conditional flows that *(i)* tranports all points x_0 from $t = 0$ to x_1 at exactly $t = 1$ and *(ii)* are straight paths between the samples x_0 and x_1. This particular case of straight paths was also studied by Liu et al. (2022) and Albergo & Vanden-Eijnden (2023), where the conditional flow (9) is referred to as a stochastic interpolant. Lipman et al. (2023) additionally showed that the conditional construction can be applied to a large class of Gaussian conditional probability paths, namely when $p_t(x|x_1) = \mathcal{N}(x|\mu_t(x_1), \sigma_t(x_1)^2 I)$. This family of probability paths encompasses most prior diffusion models where probability paths are induced by simple diffusion processes with linear drift and constant diffusion (*e.g.* Ho et al. (2020); Song et al. (2021b)). However, existing works mostly consider settings where $q_0(x_0)$ and $q_1(x_1)$ are sampled independently when computing training objectives such as (7).

2.3. Optimal Transport: Static & Dynamic

Optimal transport generally considers methodologies that define some notion of distance on the space of probability measures (Villani, 2008; 2003; Santambrogio, 2015). Letting $\mathcal{P}(\mathbb{R}^d)$ be the space of probability measures over $\mathbb{R}^d$, we define the Wasserstein distance with respect to a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ between two measures $q_0, q_1 \in \mathcal{P}(\mathbb{R}^d)$ as (Kantorovitch, 1942)

$$W_c(q_0, q_1) \coloneqq \min_{q \in \Gamma(q_0,q_1)} \mathbb{E}_{q(x_0,x_1)}[c(x_0, x_1)]\,, \tag{11}$$

where $\Gamma(q_0, q_1)$ is the set of joint measures with left marginal equal to q_0 and right marginal equal to q_1, called the set of *couplings*. The minimizer to Equation (11) is called the optimal coupling, which we denote by q_c^*. In the case where $c(x_0, x_1) \coloneqq \|x_0 - x_1\|^2$, the squared-Euclidean distance, Equation (11) amounts to the (squared) 2-Wasserstein distance $W_2^2(q_0, q_1)$, and we simply write the optimal transport plan as q^*.

Considering again the squared-Euclidean cost, in the case where q_0 exhibits a density over $\mathbb{R}^d$ (e.g. if q_0 is the standard normal distribution), Benamou & Brenier (2000) states that $W_2^2(q_0, q_1)$ can be equivalently expressed as a *dynamic* formulation,

$$W_2^2(q_0, q_1) = \min_{p_t, u_t} \int_0^1 \int_{\mathbb{R}^d} \|u_t(x)\|^2 p_t(x) \mathrm{d}x_0 \mathrm{d}t. \tag{12}$$

where u_t generates p_t, and p_t satisfies boundary conditions $p_{t=0} = q_0$ and $p_{t=1} = q_1$. The optimality condition ensures that sample paths x_t are straight lines, i.e. minimize the length of the path, and leads to paths that are much easier to simulate. Some prior approaches have sought to regularize the model using this optimality objective (*e.g.* Tong et al. (2020); Finlay et al. (2020b)). In contrast, instead of directly minimizing (12), we will discuss an approach based on using solutions of the optimal coupling q^* on minibatch problems, while leaving the marginal constraints intact.

3. Flow Matching with Joint Distributions

While Conditional Flow Matching in (7) leads to an unbiased gradient estimator for the Flow Matching objective, it was designed with independently sampled x_0 and x_1 in mind. We generalize the framework from Subsection 2.2 to a construction that uses arbitrary joint distributions of $q(x_0, x_1)$ which satisfy the correct marginal constraints, *i.e.*

$$\int q(x_0, x_1)\mathrm{d}x_1 = q_0(x_0)\,,\ \int q(x_0, x_1)\mathrm{d}x_0 = q_1(x_1). \tag{13}$$

We will show in Subsection 4 that this can potentially lead to lower gradient variance during training and allow us to design more optimal marginal vector fields $u_t(x)$ with desirable properties such as improved sample efficiency.

Building on top of Flow Matching, we propose modifying the conditional probability path construction (6) so that at $t = 0$, we define

$$p_{t=0}(x_0|x_1) = q(x_0|x_1). \tag{14}$$

where $q(x_0|x_1)$ is the conditional distribution $\frac{q(x_0,x_1)}{q_1(x_1)}$. Using this construction, we still satisfy the marginal constraint,

$$p_0(x) = \int p_0(x|x_1)q_1(x_1)dx_1 = \int q(x, x_1)dx_1 = q_0(x)$$

i.e. $p_{t=0}(x) = \int q(x, x_1)dx_1 = q_0(x)$ by the assumption made in (13). Then similar to Chen & Lipman (2023), we

note that the conditional probability path $p_t(x|x_1)$ *need not be explicitly formulated* for training, and that only an appropriate conditional vector field $u_t(x|x_1)$ needs to be chosen such that all points arrive at x_1 at $t = 1$, which ensures $p_{t=1}(x|x_1) = \delta(x - x_1)$. As such, we can make use of the same conditional vector field as prior works, *e.g.* the choice in Equations (8) to (10).

We then propose the **Joint CFM** objective as

$$\mathcal{L}_{\text{JCFM}} = \mathbb{E}_{t,q(x_0,x_1)} \left\| v_t(x_t;\theta) - u_t(x_t|x_1) \right\|^2, \quad (15)$$

where $x_t = \psi_t(x_0|x_1)$ is the conditional flow. Training only involves sampling from $q(x_0, x_1)$ and does not require explicitly knowing the densities of $q(x_0, x_1)$ or $p_t(x|x_1)$. Note that Equation (15) reduces to the original CFM objective (7) when $q(x_0, x_1) = q_0(x_0)q_1(x_1)$.

A quick sanity check shows that this objective can be used with any choice of joint distribution $q(x_0, x_1)$.

Lemma 3.1. *The optimal vector field $v_t(\cdot;\theta)$ in (15), which is the marginal vector field u_t, maps between the marginal distributions $q_0(x_0)$ and $q_1(x_1)$.*

In the remainder of the section, we highlight some motivations for using joint distributions $q(x_0, x_1)$ that are different from the independent distribution $q_0(x_0)q_1(x_1)$.

Variance reduction Choosing a good joint distribution can be seen as a way to reduce the variance of the gradient estimate, which improves and speeds up training. We develop the gradient covariance at a fixed x and t, and bound its total variance:

Lemma 3.2. *The total variance (i.e. the trace of the covariance) of the gradient at a fixed x and t is bounded as:*

$$\sigma^2_{t,x} = \mathrm{Tr}\Big[\mathrm{Cov}_{p_t(x_1|x)}\Big(\nabla_\theta \left\| v_t(x;\theta) - u_t(x|x_1)\right\|^2\Big)\Big] \quad (16)$$
$$\leq \left\|\nabla_\theta v_t(x;\theta)\right\|^2 \mathbb{E}_{p_t(x_1|x)} \left\| u_t(x) - u_t(x|x_1)\right\|^2$$

Then $\mathbb{E}_{t,p_t(x)}[\sigma^2_{t,x}]$ is bounded above by:

$$\max_{t,x} \left\|\nabla_\theta v_t(x;\theta)\right\|^2 \times \mathcal{L}_{\text{JCFM}} \quad (17)$$

This proves that $\mathbb{E}_{t,p_t(x)}[\sigma^2_{t,x}]$, which is the average gradient variance at fixed x and t, is upper bounded in terms of the Joint CFM objective. That means that minimizing the Joint CFM objective help in decreasing $\mathbb{E}_{t,p_t(x)}[\sigma^2_{t,x}]$. Note also that $\mathbb{E}_{t,p_t(x)}[\sigma^2_{t,x}]$ is not the gradient variance and is always smaller, as it does not account for variability over x and t, but it is a good proxy for it. The proof is in App. D.2.

Sampling x_0 and x_1 independently generally cannot achieve value zero for $\mathbb{E}_{t,p_t(x)}[\sigma^2_{t,x}]$ *even at the optimum*, since there are an infinite number of pairs (x_0, x_1) whose conditional path crosses any particular x at a time t. As shown in (17), having a low optimal value for the Joint CFM objective is a good proxy for low gradient variance and hence a desirable property for choosing a joint distribution $q(x_0, x_1)$. In Section 4, we show that certain joint distributions have optimal Joint CFM values close to zero.

Straight flows Ideally, the flow ψ_t of the marginal vector field u_t (and of the learned v_θ by extension) should be close to a straight line. The reason is that ODEs with straight trajectories can be solved with high accuracy using fewer steps (i.e. function evaluations), which speeds up sample generation. The quantity

$$S = \mathbb{E}_{t,q_0(x_0)}\big[\|u_t(\psi_t(x_0))\|^2 - \|\psi_1(x_0) - x_0\|^2\big], \quad (18)$$

which we call the *straightness* of the flow and was also studied by Liu (2022), measures how straight the trajectories are. Namely, we can rewrite it as

$$S = \mathbb{E}_{t,q_0(x_0)}\left[\|u_t(\psi_t(x_0)) - \mathbb{E}_{t'}\left[u_{t'}(\psi_{t'}(x_0))\right]\|^2\right], \quad (19)$$

which shows that $S \geq 0$ and only zero if $u_t(\psi_t(x_0))$ is constant along t, which is equivalent to $\psi_t(x_0)$ being a straight line.

When x_0 and x_1 are sampled independently, the straightness is in general far from zero. This can be seen in the CondOT plots in Figure 2 (right); if flows were close to straight lines, samples generated with one function evaluation (NFE=1) would be of high quality. In Section 4, we show that for certain joint distributions, the straightness of the flow is close to zero.

Near-optimal transport cost By Lemma 3.1, the flow ψ_t corresponding to the optimal u_t satisfies that $\psi_0(x_0) = x_0 \sim q_0$ and $\psi_1(x_0) \sim q_1$. Hence, $x_0 \mapsto \psi_1(x_0)$ is a transport map between q_0 and q_1 with an associated transport cost

$$\mathbb{E}_{q_0(x_0)}\|\psi_1(x_0) - x_0\|^2. \quad (20)$$

There is no reason to believe that when x_0 and x_1 are sampled independently, the transport cost $\mathbb{E}_{q_0(x_0)}\|\psi_1(x_0) - x_0\|^2$ will be anywhere near the optimal transport cost $W_2^2(p_0, p_1)$. Yet, in Section 4 we show that for well chosen q, the transport cost for ψ_1 does approach its optimal value. Computing optimal (or near-optimal) transport maps in high dimensions is a challenging task (Makkuva et al., 2020; Amos, 2023) that extends beyond generative modeling and into the field of optimal transport, and it has applications in computer vision (Feydy et al., 2017; Solomon et al., 2015; 2016; Liu et al., 2023) and computational biology (Lübeck et al., 2022; Bunne et al., 2021; 2022; Schiebinger et al., 2019), for instance. Hence, Joint CFM may also be viewed as a practical way to obtain approximately optimal transport maps in this context.

4. Multisample Flow Matching

Constructing a joint distribution satisfying the marginal constraints is difficult, especially since at least one of the marginal distributions is based on empirical data. We thus discuss a method to construct the joint distribution $q(x_0, x_1)$ implictly by designing a suitable sampling procedure that leaves the marginal distributions invariant. Note that training with (15) only requires sampling from $q(x_0, x_1)$.

We use a multisample construction for $q(x_0, x_1)$ in the following manner:

1. Sample $\{x_0^{(i)}\}_{i=1}^k \sim q_0(x_0)$ and $\{x_1^{(i)}\}_{i=1}^k \sim q_1(x_1)$.

2. Construct a doubly-stochastic matrix with probabilities $\pi(i, j)$ dependent on the samples $\{x_0^{(i)}\}_{i=1}^k$ and $\{x_1^{(i)}\}_{i=1}^k$.

3. Sample from the discrete distribution, $q^k(x_0, x_1) = \frac{1}{k}\sum_{i,j=1}^k \delta(x_0 - x_0^i)\delta(x_1 - x_1^j)\pi(i, j)$.

Marginalizing $q^k(x_0, x_1)$ over samples from Step 1, we obtain the implicitly defined $q(x_0, x_1)$. By choosing different *couplings* $\pi(i, j)$, we induce different joint distributions. In this work, we focus on couplings that induce joint distributions which approximates, or at least partially satisfies, the optimal transport joint distribution. The following result, proven in App. D.3, guarantees that q has the right marginals.

Lemma 4.1. *The joint distribution $q(x_0, x_1)$ constructed in Steps [1-3] has marginals $q_0(x_0)$ and $q_1(x_1)$.*

That is, the marginal constraints (13) are satisfied and consequently we are allowed to use the framework of Section 3.

4.1. CondOT is Uniform Coupling

The aforementioned multisample construction subsumes the independent joint distribution used by prior works, when the joint coupling is taken to be uniformly distributed, *i.e.* $\pi(i, j) = \frac{1}{k}$. This is precisely the coupling used by (Lipman et al., 2023) under our introduced notion of Multisample Flow Matching, and acts as a natural reference point.

4.2. Batch Optimal Transport (BatchOT) Couplings

The natural connections between optimal transport theory and optimal sampling paths in terms of straight-line interpolations, lead us to the following pseudo-deterministic coupling, which we call Batch Optimal Transport (BatchOT). While it is difficult to solve (11) at the population level, it can efficiently solved on the level of samples. Let $\{x_0^{(i)}\}_{i=1}^k \sim q_0(x_0)$ and $\{x_1^{(i)}\}_{i=1}^k \sim q_1(x_1)$. When defined on batches of samples, the OT problem (11) can be solved exactly and efficiently using standard solvers, as in `POT` (Flamary et al., 2021, Python Optimal Transport). On a batch of k samples, the runtime complexity is well-understood via either the Hungarian algorithm or network simplex algorithm, with an overall complexity of $\mathcal{O}(k^3)$ (Peyré & Cuturi, 2019, Chapter 3). The resulting coupling $\pi^{k,*}$ from the algorithm is a *permutation matrix*, which is a type of doubly-stochastic matrix that we can incorporate into Step 3 of our procedure.

We consider the effect that the sample size k has on the marginal vector field $u_t(x)$. The following theorem shows that in the limit of $k \to \infty$, BatchOT satisfies the three criteria that motivate Joint CFM: variance reduction, straight flows, and near-optimal transport cost.

Theorem 4.2 (Informal)**.** *Suppose that Multisample Flow Matching is run with BatchOT. Then, as $k \to \infty$,*

(i) The value of the Joint CFM objective (Equation (15)) for the optimal u_t converges to 0.
(ii) The straightness S for the optimal marginal vector field u_t (Equation (18)) converges to zero.
(iii) The transport cost $\mathbb{E}_{q_0(x_0)}\|\psi_1(x_0) - x_0\|^2$ (Equation (20)) associated to u_t converges to the optimal transport cost $W_2^2(p_0, p_1)$.

As $k \to \infty$, result *(i)* implies that the gradient variance both during training and at convergence is reduced due to Equation (17); result *(ii)* implies the optimal model will be easier to simulate between t=0 and t=1; result *(iii)* implies that Multisample Flow Matching can be used as a simulation-free algorithm for approximating optimal transport maps.

The full version of Thm. 4.2 can be found in App. D, and it makes use of standard, weak technical assumptions which are common in the optimal transport literature. While Thm. 4.2 only analyzes asymptotic properties, we provide theoretical evidence that the transport cost decreases with k, as summarized by a monotonicity result in Thm. D.8.

4.3. Batch Entropic OT (BatchEOT) Couplings

For k sufficiently large, the cubic complexity of the BatchOT approach is not always desirable, and instead one may consider approximate methods that produce couplings sufficiently close to BatchOT at a lower computational cost. A popular surrogate, pioneered in (Cuturi, 2013), is to incorporate an entropic penalty parameter on the doubly stochastic matrix, pulling it closer to the independent coupling:

$$\min_{q \in \Gamma(q_0, q_1)} \mathbb{E}_{(x_0, x_1) \sim q}\|x_0 - x_1\|^2 + \varepsilon H(q)\,,$$

where $H(q) = -\sum_{i,j} q_{i,j}(\log(q_{i,j}) - 1)$ is the entropy of the doubly stochastic matrix q, and $\varepsilon > 0$ is some finite regularization parameter. The optimality conditions of this strictly convex program leads to Sinkhorn's algorithm, which has a runtime of $\tilde{\mathcal{O}}(k^2/\varepsilon)$ (Altschuler et al., 2017).

The output of performing Sinkhorn's algorithm is a doubly-stochastic matrix. The two limiting regimes of the regular-

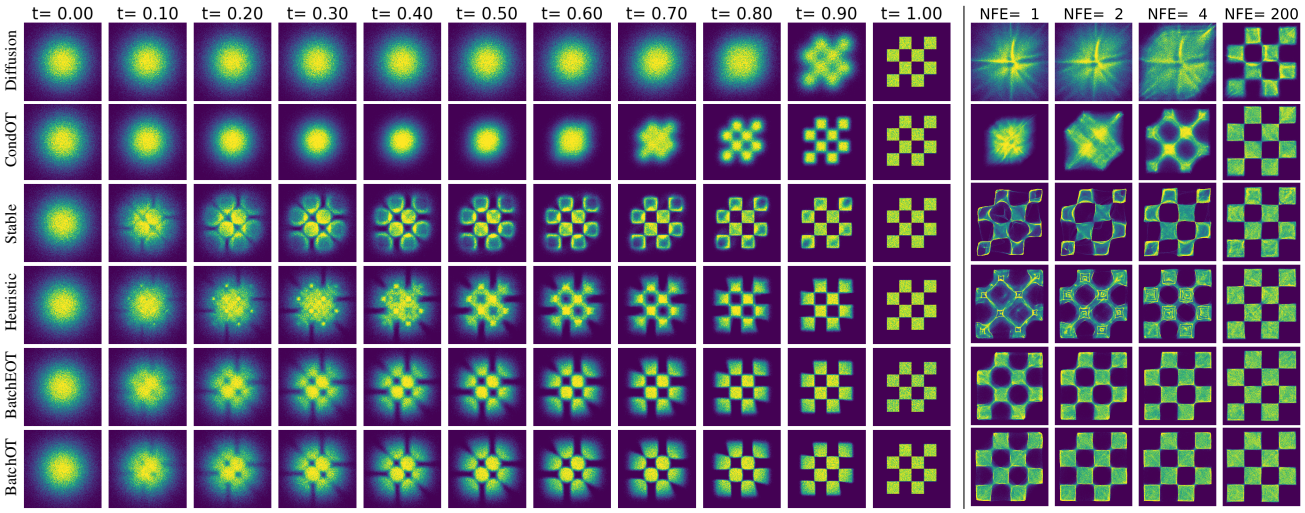


Figure 2. Multisample Flow Matching learn probability paths that are much closer to an optimal transport path than baselines such as Diffusion and CondOT paths. (*Left*) Exact marginal probability paths. (*Right*) Samples from trained models at $t = 1$ for different numbers of function evaluations (NFE), using Euler discretization. Furthermore, the final values of the Joint CFM objective (15)—upper bounds on the variance of u_t at convergence—are: CondOT: 10.72; Stable: 1.60, Heuristic: 1.56; BatchEOT: 0.57, BatchOT: 0.24.

ization parameter are well understood (c.f. Peyré & Cuturi (2019), Proposition 4.1, for instance): as $\varepsilon \to 0$, BatchEOT recovers the BatchOT permutation matrix from Section 4.2; as $\varepsilon \to \infty$, BatchEOT recovers the independent coupling on the indices from Section 4.1.

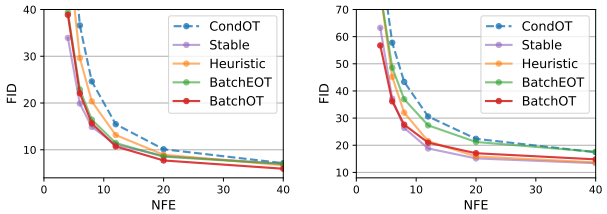
4.4. Stable and Heuristic Couplings

An alternative approach is to consider faster algorithms that satisfy at least some desirable properties of an optimal coupling. In particular, an optimal coupling is *stable*. A permutation coupling is stable if *no pair of* $x_0^{(i)}$ *and* $x_1^{(j)}$ *favor each other over their assigned pairs based on the coupling.* Such a problem can be solved using the Gale-Shapeley algorithm (Gale & Shapley, 1962) which has a compute cost of $\mathcal{O}(k^2)$ given the cross set ranking of all samples. Starting from a random assignment, it is an iterative algorithm that reassigns pairs if they violate the stability property and can terminate very early in practice. Note that in a cost-based ranking, one has to sort the coupling costs of each sample with all samples in the opposing set, resulting in an overall $\mathcal{O}(k^2 \log(k))$ compute cost.

The Gale-Shapeley algorithm is agnostic to any particular costs, however, as stability is only defined in terms of relative rankings of individual samples. We design a modified version of this algorithm based on a heuristic for satisfying the cyclical monotonicity property of optimal transport, namely that should pairs be reassigned, the reassignment should not increase the total cost of already matched pairs. We refer to the output of this modified algorithm as a *heuristic coupling* and discuss the details in Appendix A.2.

5. Related Work

Generative modeling and optimal transport are inherently intertwined topics, both often aiming to learn a transport between two distributions but with very different goals. Optimal transport is widely recognized as a powerful tool for large-scale generative modeling as it can be used to stabilize training (Arjovsky et al., 2017). In the context of continuous-time generative modeling, optimal transport has been used to regularize continuous normalizing flows for easier simulation (Finlay et al., 2020b; Onken et al., 2021), and increase interpretability (Tong et al., 2020). However, the existing methods for encouraging optimality in a generative model generally require either solving a potentially unstable min-max optimization problem (*e.g.* (Arjovsky et al., 2017; Makkuva et al., 2020; Albergo & Vanden-Eijnden, 2023)) or require simulation of the learned vector field as part of training (*e.g.* Finlay et al. (2020b); Liu et al. (2022)). In contrast, the approach of using batch optimal couplings can be used to avoid the min-max optimization problem, but has not been successfully applied to generative modeling as they do not satisfy marginal constraints—we discuss this further in the following Section 5.1. On the other hand, neural optimal transport approaches are mainly centered around the quadratic cost (Makkuva et al., 2020; Amos, 2023; Finlay et al., 2020a) or rely heavily on knowing the exact cost function (Fan et al., 2021; Asadulaev et al., 2022). Being capable of using batch optimal couplings allows us to build generative models to approximate optimal maps under any cost function, and even when the cost function is unknown.



ImageNet 32×32 ImageNet 64×64

Figure 3. Sample quality (FID) vs compute cost (NFE) using Euler discretization. CondOT has significantly higher FID at lower NFE compared to proposed methods.

	ImageNet 32×32 NFE @ FID = 10	**ImageNet 64×64** NFE @ FID = 20
Diffusion	≥40	≥40
FM w/ CondOT	20	29
MultisampleFM w/ Heuristic	18	12
MultisampleFM w/ Stable	**14**	**11**
MultisampleFM w/ BatchOT	**14**	12

Table 1. Derived results shown in Figure 3, we can determine the approximate NFE required to achieve a certain FID across our proposed methods. The baseline diffusion-based methods (e.g. ScoreFlow and DDPM) require more than 40 NFE to achieve these FID values.

NFE	DDPM	ScoreSDE	BatchOT	Stable
Adaptive	5.72	6.84	**4.68**	5.79
40	19.56	16.96	**5.94**	7.02
20	63.08	58.02	**7.71**	8.66
8	232.97	218.66	15.64	**14.89**
6	275.28	266.76	22.08	**19.88**
4	362.37	340.17	38.86	**33.92**

Table 2. FID of model samples on ImageNet 32×32 using varying number of function evaluations (NFE) using Euler discretization.

5.1. Minibatch Couplings for Generative Modeling

Among works that use optimal transport for training generative models are those that make use of batch optimal solutions and their gradients such as Li et al. (2017); Genevay et al. (2018); Fatras et al. (2019); Liu et al. (2019). However, *naïvely using solutions to batches only produces, at best, the barycentric map*, *i.e.* the map that fits to average of the batch couplings (Ferradans et al., 2014; Seguy et al., 2017; Pooladian & Niles-Weed, 2021), and does not correctly match the true marginal distribution. This is a well-known problem and while multiple works (*e.g.* Fatras et al. (2021); Nguyen et al. (2022)) have attempted to circumvent the issue through alternative formulations of optimality, the lack of marginal preservation has been a major downside of using batch couplings for generative modeling as they do not have the ability to match the target distribution for finite batch sizes. This is due to the use of building models within the *static* setting, where the map is parameterized directly with a neural network. In contrast, we have shown in Lemma 4.1 that in our *dynamic* setting, where we parameterize the map as the solution of a neural ODE, it is possible to preserve the marginal distribution exactly. Furthermore, we have shown in Proposition D.7 (App. D.5) that our method produces a map that is no higher cost than the joint distribution induced from BatchOT couplings.

Concurrently, Tong et al. (2023) motivates the use of BatchOT solutions within a similar framework as our Joint CFM, but from the perspective of obtaining accurate solutions to dynamic optimal transport problems. Similarly, Lee et al. (2023) propose to explicitly learn a joint distribution, parameterized with a neural network, with the aim of minimizing trajectory curvature; this is done using through an auxiliary VAE-style objective function. In contrast, we propose a family of couplings that all satisfy the marginal constraints, all of which are easy to implement and have negligible cost during training. Our construction allow us to focus on (i) fixing consistency issues within simulation-free generative models, and (ii) using Joint CFM to obtain more optimal solutions than the original BatchOT solutions.

6. Experiments

We empirically investigate Multisample Flow Matching on a suite of experiments. First, we show how different couplings affect the model on a 2D distribution. We then turn to benchmark, high-dimensional datasets, namely ImageNet (Deng et al., 2009). We use the official *face-blurred* ImageNet data and then downsample to 32×32 and 64×64 using the open source preprocessing scripts from Chrabaszcz et al. (2017). Finally, we explore the setting of unknown cost functions while only batch couplings are provided. Full details on the experimental setting can be found in Appendix E.2.

6.1. Insights from 2D experiments

Figure 2 shows the proposed Multisample Flow Matching algorithm on fitting to a checkboard pattern distribution in 2D. We show the marginal probability paths induced by different coupling algorithms, as well as low-NFE samples of trained models on these probability paths.

The diffusion and CondOT probability paths do not capture intricate details of the data distribution until it is almost at the end of the trajectory, whereas Multisample Flow Matching approaches provide a gradual transition to the target distribution along the flow. We also see that with a fixed step solver, the BatchOT method is able to produce an accurate target distribution in just one Euler step in this low-dimensional setting, while the other coupling approaches also get pretty close. Finally, it is interesting that both Stable and Heuristic exhibit very similar probability paths to

	ImageNet 32×32		ImageNet 64×64	
	CondOT	BatchOT	CondOT	BatchOT
Consistency(m=4)	0.141	**0.101**	0.174	**0.157**
Consistency(m=6)	0.105	**0.071**	0.151	**0.134**
Consistency(m=8)	0.079	**0.052**	0.132	**0.115**
Consistency(m=12)	0.046	**0.030**	0.106	**0.085**

Table 3. BatchOT produces samples with more similar content to its true samples at low NFEs (using midpoint discretization). Visual examples of this consistency are shown in Figure 1.

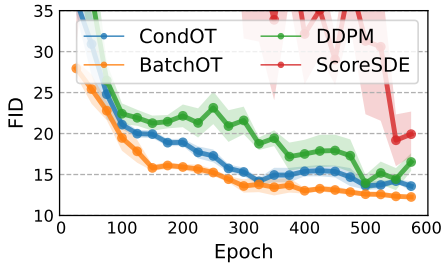


Figure 4. Multisample Flow Matching with BatchOT shows faster convergence due to reduced variance (ImageNet64).

optimal transport despite only satisfying weaker conditions.

6.2. Image Datasets

We find that Multisample Flow Matching retains the performance of Flow Matching while improving on sample quality, compute cost, and variance. In Table 6 of Appendix B.1, we report sample quality using the standard Fréchet Inception Distance (FID), negative log-likelihood values using bits per dimension (BPD), and compute cost using number of function evaluations (NFE); these are all standard metrics throughout the literature. Additionally, we report the variance of $u_t(x|x_0, x_1)$, estimated using the Joint CFM loss (15) which is an upper bound on the variance. We do not observe any performance degradations while simulation efficiency improves significantly, even with small batch sizes.

Additionally, in Appendix B.5, we include runtime comparisons between Flow Matching and Multisample Flow Matching. On ImageNet32, we only observe a 0.8% relative increase in runtime compared to Flow Matching, and a 4% increase on ImageNet64.

Higher sample quality on a compute budget We observe that with a fixed NFE, models trained using Multisample Flow Matching generally achieve better sample quality. For these experiments, we draw $x_0 \sim \mathcal{N}(0, I_d)$ and simulate $v_t(\cdot, \theta)$ up to time $t = 1$ using a fixed step solver with a fixed NFE. Figures 3 show that even on high dimensional data distributions, the sample quality of of multisample methods improves over the naïve CondOT approach as the number of function evaluations drops. We compare to the FID of diffusion baseline methods in Table 2, and provide additional results in Appendix B.4.

Interestingly, we find that the Stable coupling actually performs on par, and some times better than the BatchOT coupling, despite having a smaller asymptotic compute cost and only satisfying a weaker condition within each batch.

As FID is computed over a full set of samples, it does not show how varying NFE affects individual sample paths. We discuss a notion of consistency next, where we analyze the similarity between low-NFE and high-NFE samples.

Consistency of individual samples In Figure 1 we show samples at different NFEs, where it can be qualitatively seen that BatchOT produces samples that are more consistent between high- and low-NFE solutions than CondOT, despite achieving similar FID values.

To evaluate this quantitatively, we define a metric for establishing the *consistency* of a model with respect to an integration scheme: let $x^{(m)}$ be the output of a numerical solver initialized at x using m function evalutions to reach $t = 1$, and let $x^{(*)}$ be a near-exact sample solved using a high-cost solver starting from x_0 as well. We define

$$\text{Consistency}(m) = \tfrac{1}{D}\mathbb{E}_{x\sim q_0}\|\mathcal{F}(x^{(m)}) - \mathcal{F}(x^{(*)})\|^2 \quad (21)$$

where $\mathcal{F}(\cdot)$ outputs the hidden units from a pretrained InceptionNet[1], and D is the number of hidden units. These kinds of perceptual losses have been used before to check the content alignment between two image samples (*e.g.* Gatys et al. (2015); Johnson et al. (2016)). We find that Multisample Flow Matching has better consistency at all values of NFE, shown in Table 3.

Training efficiency Figure 4 shows the convergence of Multisample Flow Matching with BatchOT coupling compared to Flow Matching with CondOT and diffusion-based methods. We see that by choosing better joint distributions, we obtain faster training. This is in line with our variance estimates reported in Table 6 and supports our hypothesis that gradient variance is reduced by using non-trivial joint distributions.

6.3. Improved Batch Optimal Couplings

We further explore the usage of Multisample Flow Matching as an approach to improve upon batch optimal solutions. Here, we experiment with a different setting, where the cost is unknown and only samples from a batch optimal coupling are provided. In the real world, it is often the case that the preferences of each person are not known explicitly, but when given a finite number of choices, people can more easily find their best assignments. This motivates

[1]We take the same layer as used in standard FID computation.

	2-D Cost			2-D KL		32-D Cost			32-D KL		64-D Cost			64-D KL	
Cost Fn. $c(x_0, x_1)$	B	B-ST	B-FM	B-ST	B-FM	B	B-ST	B-FM	B-ST	B-FM	B	B-ST	B-FM	B-ST	B-FM
$\|x_1 - x_0\|_2^2$	0.90	0.60	0.72	0.07	4E-3	41.08	31.58	38.73	151.47	0.06	92.90	65.57	87.97	335.38	0.14
$\|x_1 - x_0\|_1$	1.09	0.86	0.98	0.18	4E-3	27.92	24.51	27.26	254.59	0.08	60.27	50.49	58.38	361.16	0.16
$1 - \frac{\langle x_0, x_1 \rangle}{\|x_0\|\|x_1\|}$	0.03	2E-4	3E-3	5.91	4E-3	0.62	0.53	0.58	179.48	0.06	0.71	0.60	0.68	337.63	0.12
$\|A(x_1 - x_0)\|_2^2$	0.91	0.54	0.65	0.07	4E-3	32.66	24.61	30.13	256.90	0.06	78.70	58.11	78.50	529.09	0.19

Table 4. Matching couplings from an oracle BatchOT solver with unknown costs. Multisample Flow Matching is able to match the marginal distribution correctly while being at least a optimal as the oracle, but static maps fail to preserve the marginal distribution.

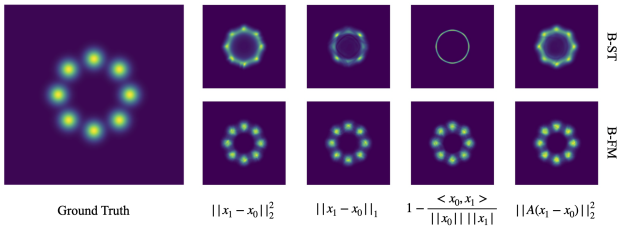


Figure 5. 2D densities on the 8-Gaussians target distribution. (Left) Ground truth density. (Right) Learned densities with static maps in the top row and Multisample Flow Matching dynamic maps in the bottom row. Models within each column were trained using batch optimal couplings with the corresponding cost function.

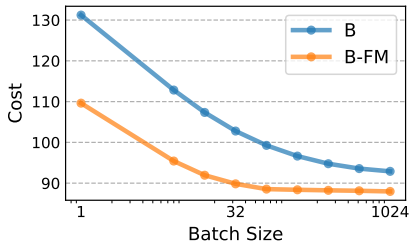


Figure 6. Transport cost vs. batch size (k) for computing couplings on the 64D synthetic dataset. The number of samples used for performing gradient steps during training and the resulting KL divergences were kept the same.

us to consider the case of unknown cost functions, and information regarding the optimal coupling is only given by a weak oracle that acts on finite samples, denoted $q^k_{OT,c}$. We consider two baselines: (i) the BatchOT cost (B) which corresponds to $\mathbb{E}_{q^k_{OT,c}(x_0,x_1)}\left[c(x_0, x_1)\right]$, and (ii) learning a static map that mimics the BatchOT couplings (B-ST) by minimizing the following objective:

$$\mathbb{E}_{q^k_{OT,c}(x_0,x_1)} \left\|x_1 - \psi_\theta(x_0)\right\|^2 . \tag{22}$$

This can be viewed as learning the barycentric projection (Ferradans et al., 2014; Seguy et al., 2017), *i.e.* $\psi^*(x_0) = E_{q^k_{OT,c}(x_1|x_0)}\left[x_1\right]$, a well-studied quantity but is known to not preserve the marginal distribution (Fatras et al., 2019).

We experiment with 4 different cost functions on three synthetic datasets in dimensions $\{2, 32, 64\}$ where both q_0 and q_1 are chosen to be Gaussian mixture models. In Table 4 we report both the transport cost and the KL divergence between q_1 and the distribution induced by the learned map, *i.e.* $[\psi_1]_\sharp q_0$. We observe that while B-ST always results in lower transport costs compared to B-FM, its KL divergence is always very high, meaning that the pushed-forward distribution by the learned static map poorly approximates q_1. Another interesting observation is that B-FM always reduces transport costs compared to B, providing experimental support to the theory (Theorem D.8).

Flow Matching improves optimality Figure 6 shows the cost of the learned model as we vary the batch size for computing couplings, where the models are trained sufficiently to achieve the same KL values as reported in Table 4. We see that our approach decreases the cost compared to the BatchOT oracle for any fixed batch size, and furthermore, converges to the OT solution faster than the batchOT oracle. Thus, since Multisample Flow Matching retains the correct marginal distributions, it can be used to better approximate optimal transport solutions than simply relying on a minibatch solution.

7. Conclusion

We propose Multisample Flow Matching, building on top of recent works on simulation-free training of continuous normalizing flows. While most prior works make use of training algorithms where data and noise samples are sampled independently, Multisample Flow Matching allows the use of more complex joint distribution. This introduces a new approach to designing probability paths. Our framework increases sample efficiency and sample quality when using low-cost solvers. Unlike prior works, our training method does not rely on simulation of the learned vector field during training, and does not introduce any min-max formulations. Finally, we note that our method of fitting to batch optimal couplings is the first to also preserve the marginal distributions, an important property in both generative modeling and solving transport problems.

References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *International Conference on Learning Representations*, 2023.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30*, 2017.

Amos, B. On amortizing convex conjugates for optimal transport. *International Conference on Learning Representations*, 2023.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Asadulaev, A., Korotin, A., Egiazarian, V., and Burnaev, E. Neural optimal transport with general cost functionals. *arXiv preprint arXiv:2205.15403*, 2022.

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.

Bunne, C., Stark, S. G., Gut, G., del Castillo, J. S., Lehmann, K.-V., Pelkmans, L., Krause, A., and Rätsch, G. Learning single-cell perturbation responses using neural optimal transport. *bioRxiv*, 2021.

Bunne, C., Krause, A., and Cuturi, M. Supervised training of conditional Monge maps. *arXiv preprint arXiv:2206.14262*, 2022.

Caffarelli, L. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5:99–104, 1992.

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

Chen, R. T. Q. and Lipman, Y. Riemannian flow matching on general geometries. *International Conference on Machine Learning*, 2023.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A downsampled variant of ImageNet as an alternative to the Cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Cuesta-Albertos, J., Matrán, C., and Tuero-Diaz, A. Optimal transportation plans and convergence in distribution. *Journal of Multivariate Analysis*, 60(1):72–83, 1997.

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

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

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Fan, J., Liu, S., Ma, S., Chen, Y., and Zhou, H. Scalable computation of monge maps with general costs. *arXiv preprint arXiv:2106.03812*, pp. 4, 2021.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch Wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Fatras, K., Zine, Y., Majewski, S., Flamary, R., Gribonval, R., and Courty, N. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

Ferradans, S., Papadakis, N., Peyré, G., and Aujol, J.-F. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

Feydy, J., Charlier, B., Vialard, F.-X., and Peyré, G. Optimal transport for diffeomorphic registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 291–299. Springer, 2017.

Finlay, C., Gerolin, A., Oberman, A. M., and Pooladian, A.-A. Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv preprint arXiv:2006.06033*, 2020a.

Finlay, C., Jacobsen, J.-H., Nurbekyan, L., and Oberman, A. How to train your neural ode: the world of jacobian and kinetic regularization. In *International Conference on Machine Learning*, pp. 3154–3164. PMLR, 2020b.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Gale, D. and Shapley, L. S. College admissions and the stability of marriage. *The American Mathematical Monthly*, 69(1):9–15, 1962.

Gatys, L. A., Ecker, A. S., and Bethge, M. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018.

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

Huang, C.-W., Chen, R. T., Tsirigotis, C., and Courville, A. Convex potential flows: Universal probability distributions with optimal transport and convex optimization. In *International Conference on Learning Representations*, 2020.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.

Kantorovitch, L. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.

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

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *International Conference on Learning Representations*, 2023.

Liu, G.-H., Vahdat, A., Huang, D.-A., Theodorou, E. A., Nie, W., and Anandkumar, A. I^2sb: Image-to-image Schr\" odinger bridge. *International Conference on Machine Learning*, 2023.

Liu, H., Gu, X., and Samaras, D. Wasserstein gan with quadratic transport cost. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4832–4841, 2019.

Liu, Q. Rectified flow: A marginal preserving approach to optimal transport, 2022.

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*, 2022.

Lübeck, F., Bunne, C., Gut, G., del Castillo, J. S., Pelkmans, L., and Alvarez-Melis, D. Neural unbalanced optimal transport via cycle-consistent semi-couplings. *arXiv preprint arXiv:2209.15621*, 2022.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pp. 6672–6681. PMLR, 2020.

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

Neklyudov, K., Severo, D., and Makhzani, A. Action matching: A variational method for learning stochastic dynamics from samples. *arXiv preprint arXiv:2210.06662*, 2022.

Nguyen, K., Nguyen, D., Pham, T., Ho, N., et al. Improving mini-batch optimal transport via partial transportation. In *International Conference on Machine Learning*, pp. 16656–16690. PMLR, 2022.

Nurbekyan, L., Iannantuono, A., and Oberman, A. M. No-collision transportation maps. *Journal of Scientific Computing*, 82(2), 2020.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 9223–9232, 2021.

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

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. *arXiv preprint arXiv:2109.12004*, 2021.

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

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Santambrogio, F. Optimal transport for applied mathematicians. 2015.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

Seguy, V., Damodaran, B. B., Flamary, R., Courty, N., Rolet, A., and Blondel, M. Large-scale optimal transport and mapping estimation. *arXiv preprint arXiv:1711.02283*, 2017.

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

Solomon, J., Peyré, G., Kim, V. G., and Sra, S. Entropic metric alignment for correspondence problems. *ACM Trans. Graph.*, 35(4):72:1–72:13, 2016.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021a.

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

Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Varadarajan, V. S. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 19(1/2):23–26, 1958.

Villani, C. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

Wolansky, G. Semi-discrete optimal transport. *arXiv preprint arXiv:1911.04348*, Sep 2020.

A. Coupling algorithms

Multisample FM makes use of batch coupling algorithms to construct an implicit joint distribution satisfying the marginal constraints. While BatchOT coupling is motivated by approximating the OT map, we consider other lower complexity coupling algorithms which produce coupling that satisfy some desired property of optimal couplings. In Table 5 we summarize the runtime complexities for the different algorithms used in this work. We will now describe in detail the Stable and Heuristic coupling algorithms.

	CondOT	BatchOT	BatchEOT	Stable	Heuristic
Runtime Complexity	$\mathcal{O}(1)$	$\mathcal{O}(k^3)$	$\tilde{\mathcal{O}}(k^2/\varepsilon)$	$\mathcal{O}(k^2\log(k))$	$\mathcal{O}(k^2\log(k))$

Table 5. Runtime complexities of the different coupling algorithms as a function of the batch size k.

A.1. Stable couplings

(Wolansky, 2020) surveys discrete optimal transport from a stable coupling perspective proving that stability is a necessary condition for OT couplings. Although stable couplings are not OT, they are cheaper to compute and are therefore an appealing approach to pursue. For completeness we formulate the Gale Shapely Algorithm in our setting in Algorithm 1. The rankings R_0, R_1 hold the preferences of the samples in $\{x_0^{(i)}\}_{i=1}^k$ and $\{x_1^{(i)}\}_{i=1}^k$ respectively. Where $R_0(i,j)$ is the rank of $x_1^{(j)}$ in $x_0^{(i)}$'s preferences and $R_1(i,j)$ is the rank of $x_0^{(j)}$ in $x_1^{(i)}$'s preferences.

A.2. Heuristic couplings

The stable coupling is agnostic to the cost of pairing samples and only takes into account the ranks. Therefore, reassignments during the Gale Shapely algorithms might increase the total cost although the rankings of assigned samples are improved. We draw inspiration from the cyclic monotonicity of OT couplings (Villani, 2008) and from the marriage with sharing formulation in (Wolansky, 2020) and modify the reassignment condition in the Gale Shapely algorithm (see Algorithm 2). The modified condition encourages "local" monotonicity between the reassigned pairs only, reassigning a pair only if the potentially newly assigned pairs have a lower cost.

Algorithm 1 Stable Coupling (Gale Shapely)

Result: assignment σ
Data: $\{x_0^{(i)}\}_{i=1}^k \sim q_0(x_0)$, $\{x_1^{(i)}\}_{i=1}^k \sim q_1(x_1)$, rankings R_0, R_1
initialization: σ empty assignment
while $\exists\, i \in [k]$ `s.t.` $\sigma(i)$ *is empty* **do**
  $j \leftarrow$ first sample in $R_0(i,\cdot)$ whom $x_0^{(i)}$ has not tried to match with yet
  if $\exists\, i'$ `s.t.` $\sigma(i') = j$ **then**
    if $R_1(j,i) < R_1(j,i')$ **then**
      $\sigma(i') \leftarrow$ empty
      $\sigma(i) \leftarrow j$
    end
  else
    $\sigma(i) \leftarrow j$
  end
end

Algorithm 2 Heuristic Coupling

Result: assignment σ
Data: $\{x_0^{(i)}\}_{i=1}^k \sim q_0(x_0)$, $\{x_1^{(i)}\}_{i=1}^k \sim q_1(x_1)$, rankings R_0, R_1, cost matrix C
initialization: σ empty assignment
while $\exists\, i \in [k]$ `s.t.` $\sigma(i)$ *is empty* **do**
  $j \leftarrow$ first sample in $R_0(i,\cdot)$ whom $x_0^{(i)}$ has not tried to match with yet
  if $\exists\, i'$ `s.t.` $\sigma(i') = j$ **then**
    $j' \leftarrow$ first sample in $R_0(i',\cdot)$ whom $x_0^{(i')}$ has not tried to match with yet
    $l \leftarrow$ second sample in $R_0(i,\cdot)$ whom $x_0^{(i)}$ has not tried to match with yet
    if $C(i,j) + C(i',j') < C(i,l) + C(i',j)$ **then**
      $\sigma(i') \leftarrow$ empty
      $\sigma(i) \leftarrow j$
    end
  else
    $\sigma(i) \leftarrow j$
  end
end

B. Additional tables and figures

B.1. Full results on ImageNet data

Model	ImageNet 32×32				ImageNet 64×64			
	NLL	FID	NFE	Var(u_t)	NLL	FID	NFE	Var(u_t)
Ablations†								
DDPM (Ho et al., 2020)	3.61	5.72	330		3.27	13.80	323	
ScoreSDE (Song et al., 2021b)	3.61	6.84	198		3.30	26.64	365	
ScoreFlow (Song et al., 2021a)	3.61	9.53	189		3.34	32.78	554	
Flow Matching w/ Diffusion (Lipman et al., 2023)	3.60	6.36	165		3.35	15.11	162	
Rectified Flow (Liu et al., 2022)	3.59	5.55	111		3.31	13.02	129	
Flow Matching w/ CondOT (Lipman et al., 2023)	3.58	5.04	139	594	3.27	13.93	131	1880
Ours								
Multisample Flow Matching w/ StableCoupling	3.59	5.79	148	523	3.27	11.82	132	1782
Multisample Flow Matching w/ HeuristicCoupling	3.58	5.29	133	555	3.26	13.37	110	1816
Multisample Flow Matching w/ BatchEOT	3.58	6.14	132	508	3.26	14.92	141	1736
Multisample Flow Matching w/ BatchOT	3.58	4.68	146	507	3.27	12.37	135	1733

Table 6. Multisample Flow Matching improves on sample quality and sample efficiency while not trading off performance at all compared to Flow Matching. †Reproduction using the same training hyperparameters (architecture, optimizer, training iterations) as our methods.

B.2. How batch size affects the marginal probability paths on 2D checkerboard data

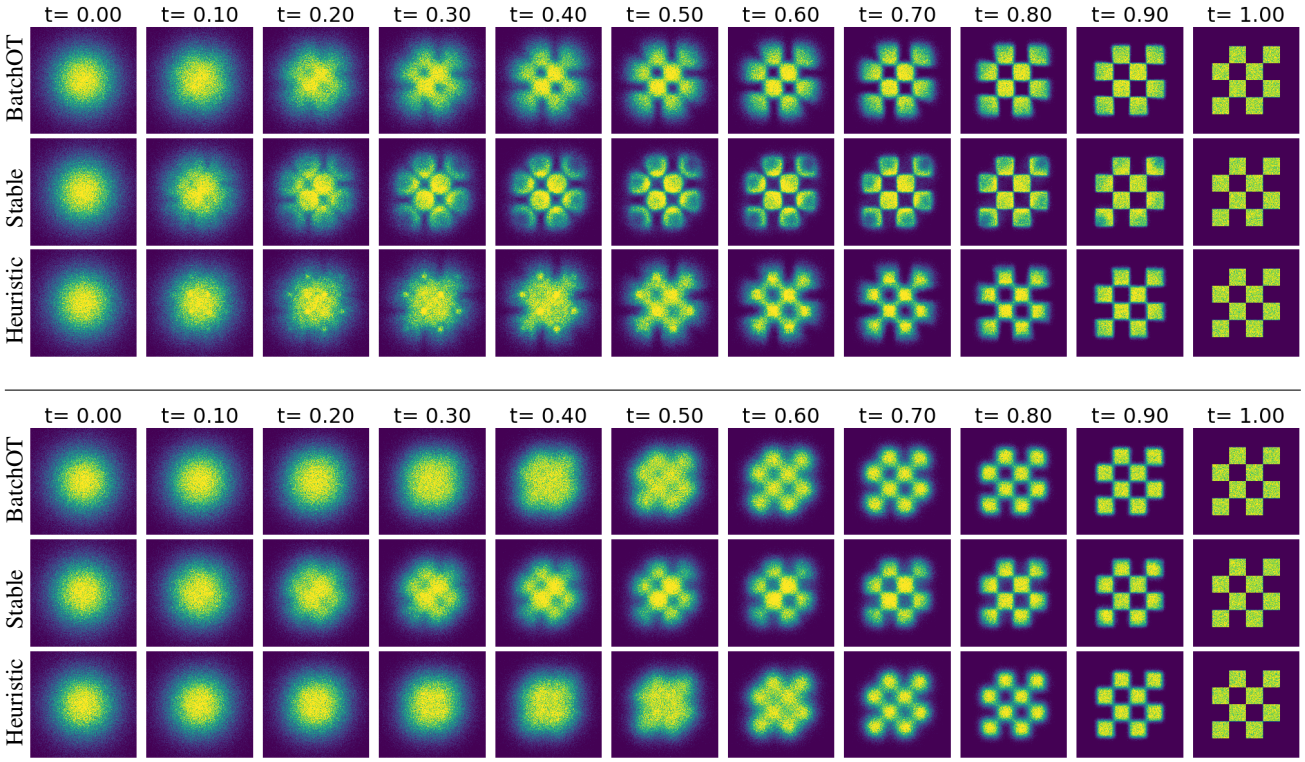
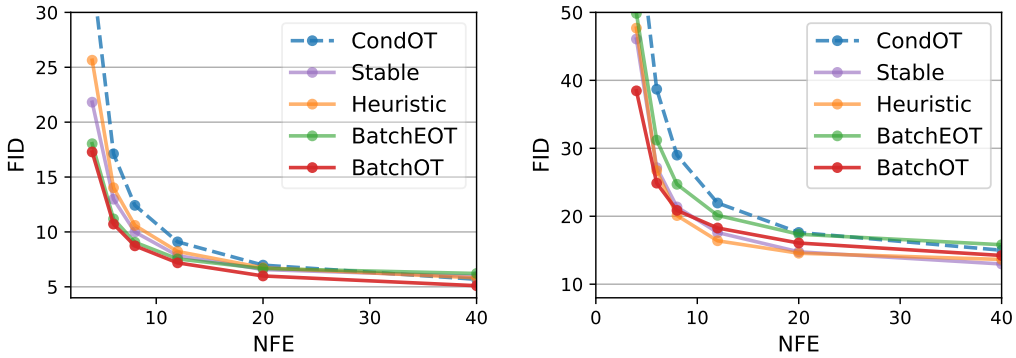


Figure 7. Marginal probability paths. (*Top*) Batch size 64. (*Bottom*) Batch size 8.

B.3. FID vs NFE using midpoint discretization scheme



ImageNet 32×32 ImageNet 64×64

Figure 8. Sample quality (FID) vs compute cost (NFE); midpoint discretization.

B.4. Comparison of FID vs NFE for baseline methods DDPM and ScoreSDE

ImageNet32 FID (Euler)

NFE	DDPM	ScoreSDE	BatchOT	Stable
Adaptive	5.72	6.84	**4.68**	5.79
40	19.56	16.96	**5.94**	7.02
20	63.08	58.02	**7.71**	8.66
12	152.59	140.95	**10.72**	11.10
8	232.97	218.66	15.64	**14.89**
6	275.28	266.76	22.08	**19.88**
4	362.37	340.17	38.86	**33.92**

ImageNet32 FID (Midpoint)

NFE	DDPM	ScoreSDE	BatchOT	Stable
Adaptive	5.72	6.84	**4.68**	5.79
40	6.68	6.48	**5.09**	5.94
20	7.80	8.96	**5.98**	6.57
12	14.87	16.22	**7.18**	7.84
8	56.41	56.73	**8.73**	9.99
6	188.08	168.99	**10.71**	12.98
4	319.41	279.06	**17.28**	21.82

Table 7. Comparing the FID vs. NFE on ImageNet32 for two baselines and two of our methods.

ImageNet64 FID (Euler)

NFE	DDPM	ScoreSDE	BatchOT	Stable
Adaptive	13.80	26.64	12.37	**11.82**
40	25.83	44.16	14.79	**13.39**
20	66.42	82.97	17.06	**15.15**
12	158.46	141.79	20.94	**18.81**
8	258.49	210.29	27.56	**26.38**
6	321.04	262.20	**36.17**	37.14
4	373.08	335.54	**56.75**	63.25

ImageNet64 FID (Midpoint)

NFE	DDPM	ScoreSDE	BatchOT	Stable
Adaptive	13.80	26.64	12.37	**11.82**
40	15.3	26.67	14.22	**12.97**
20	15.05	25.73	16.05	**14.76**
12	18.91	29.99	18.27	**17.60**
8	53.15	67.83	**20.85**	21.36
6	179.79	155.91	**24.87**	27.15
4	330.53	279.00	**38.45**	46.08

Table 8. Comparing the FID vs. NFE on ImageNet64 for two baselines and two of our methods.

B.5. Runtime per iteration is not significantly affected by solving for couplings

	ImageNet 32×32		ImageNet 64×64	
	It./s	Rel. increase	It./s	Rel. increase
CondOT (reference)	1.16	—	1.31	—
BatchOT	1.15	0.8%	1.26	3.9%
Stable	1.15	0.8%	1.26	3.9%

Table 9. Absolute and relative runtime comparisons between CondOT, BatchOT and Stable matching. "It./s" denotes the number of iterations per second, and "Rel. increase" is the relative increase with respect to CondOT. Note that these are on relatively standard batch sizes (refer to Appendix E for exact batch sizes).

B.6. Convergence improves when using larger coupling sizes

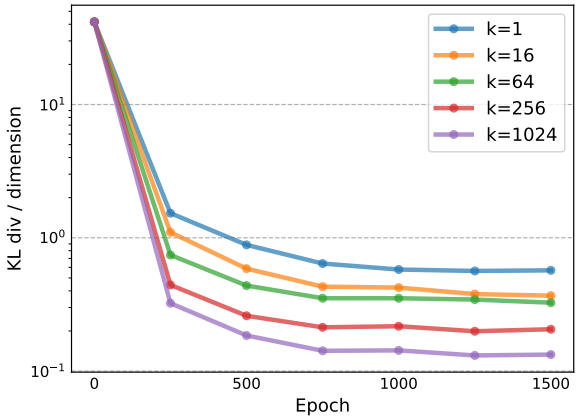


Figure 9. Larger couplings sizes (k) for defining the multisample coupling results in faster and more stable convergence. This is done on the 64-D experiments in Section 6.3. The batch size (number of samples) for training is kept thestr same and only k is varied for solving the couplings.

C. Generated samples

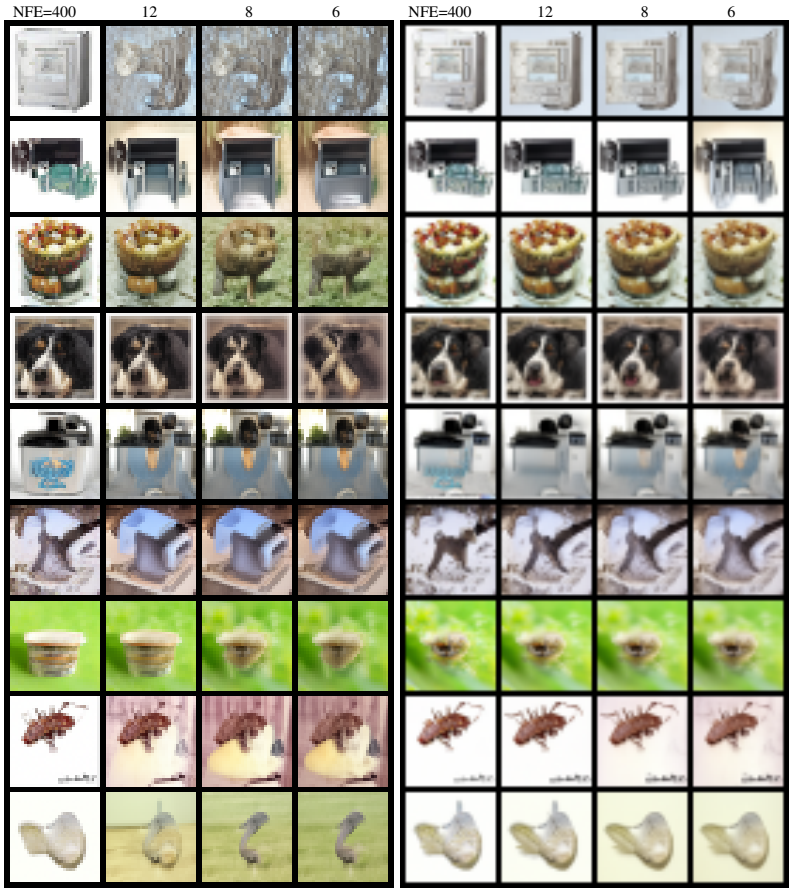


Figure 10. Multisample Flow Matching trained with batch optimal couplings produces more consistent samples across varying NFEs on ImageNet32. From left to right, the NFEs used to generate these samples are 200, 12, 8, and 6 using a midpoint discretization. Note that both flows on each row start from the same noise sample.

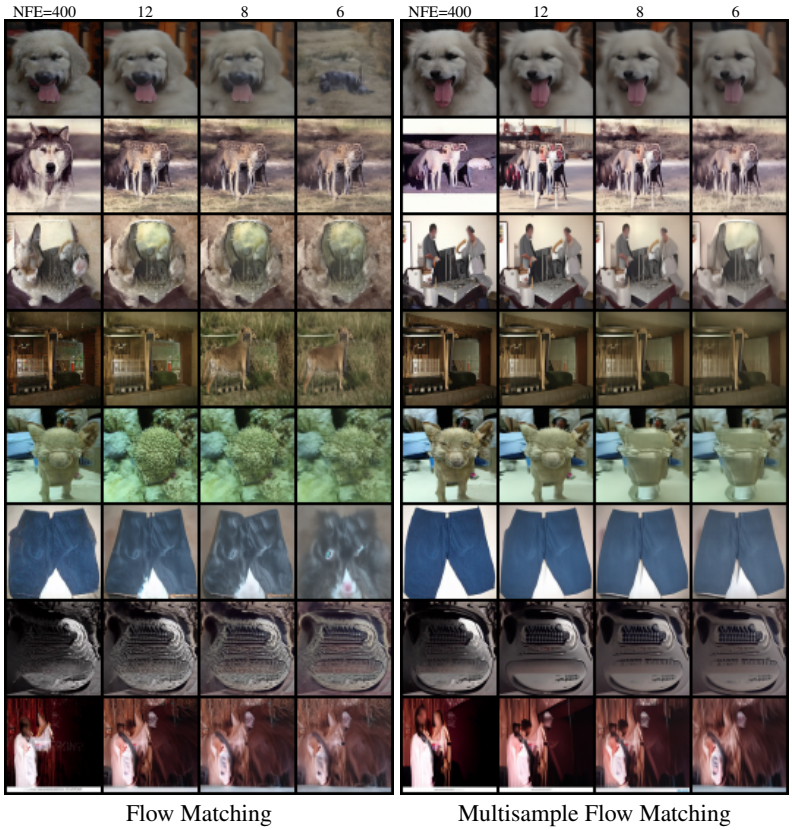


Figure 11. Multisample Flow Matching trained with batch optimal couplings produces more consistent samples across varying NFEs on ImageNet64. From left to right, the NFEs used to generate these samples are 200, 12, 8, and 6 using a midpoint discretization. Note that both flows on each row start from the same noise sample.

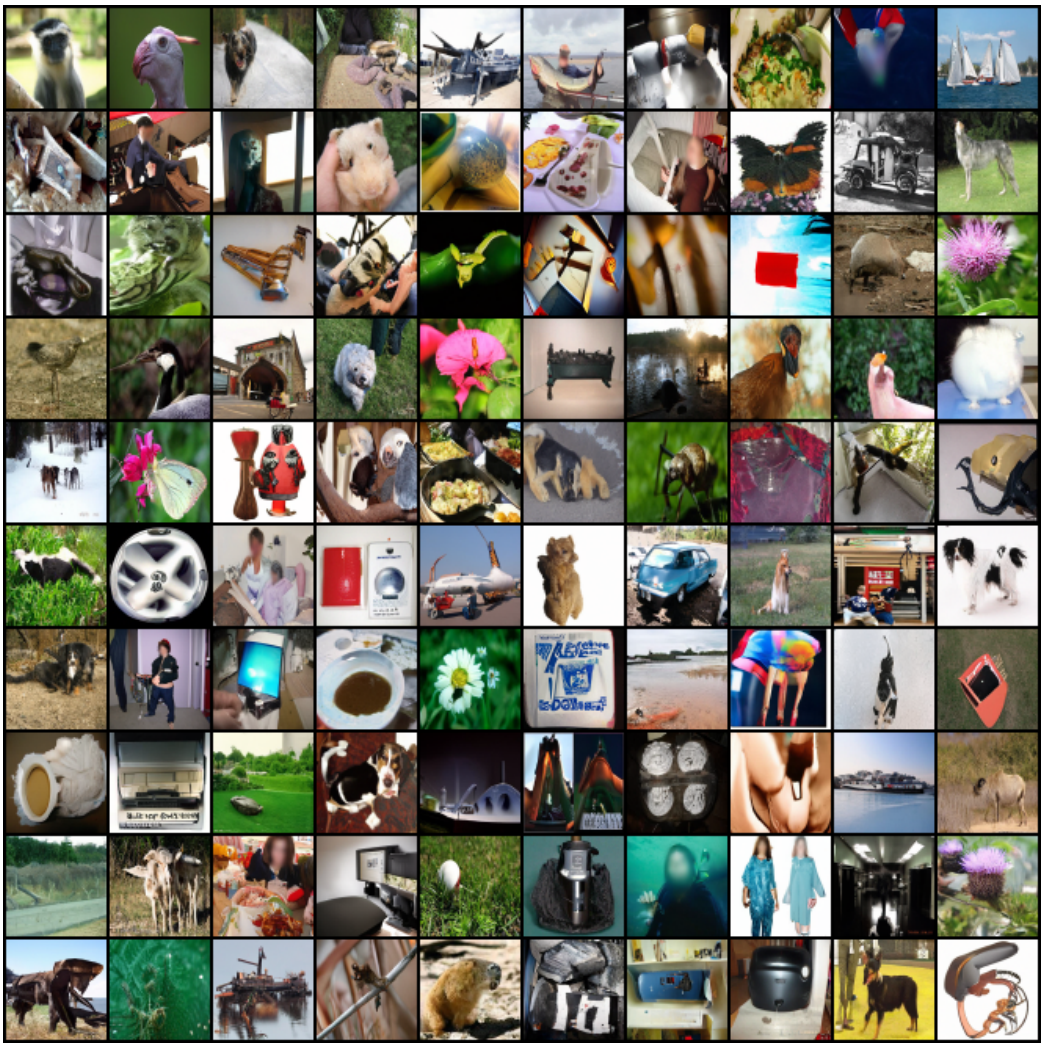


Figure 12. Non-curated generated images for ImageNet64 using Multisample Flow Matching with BatchOT coupling.

D. Theorems and proofs

D.1. Proof of Lemma 3.1

We need only prove that the marginal probability path interpolates between q_0 and q_1.

$$p_0(x) = \int p_0(x|x_1)q_1(x_1)dx_1 = \int q(x|x_1)q_1(x_1)dx_1 = q_0(x). \tag{23}$$

Then since $u_t(x|x_1)$ transports all points $x \in \mathbb{R}^D$ to x_1 at time $t = 1$, we satisfy $p_{t=1}(x|x_1) = \delta(x - x_1)$.

$$p_1(x) = \int p_1(x|x_1)q_1(x_1)dx = \int \delta(x - x_1)q_1(x_1)dx_1 = q_1(x). \tag{24}$$

Theorems 1 and 2 of Lipman et al. (2023) can then be used to prove that (i) the marginal vector field $u_t(x)$ transports between $p_0 = q_0$ and $p_1 = q_1$, and (ii) the Joint CFM objective has the same gradient in expectation as the Flow Matching objective and is uniquely minimized by $v_t(x;\theta) = u_t(x)$.

D.2. Proof of Lemma 3.2

Note that

$$\begin{aligned}&\mathrm{Cov}_{p_t(x_1|x)}\left(\nabla_\theta \left\|v_t(x;\theta) - u_t(x|x_1)\right\|^2\right) = \mathrm{Cov}_{p_t(x_1|x)}\left(\nabla_\theta \left\|v_t(x;\theta)\right\|^2 - \left(\nabla_\theta v_t(x;\theta)\right)^\mathsf{T} u_t(x|x_0,x_1)\right)\\ &= \left(\nabla_\theta v_t(x;\theta)\right)^\mathsf{T} \mathrm{Cov}_{p_t(x_1|x)}\left(u_t(x|x_1)\right)\left(\nabla_\theta v_t(x;\theta)\right),\end{aligned} \tag{25}$$

and that

$$\mathrm{Cov}_{p_t(x_1|x)}\left(u_t(x|x_1)\right) = \mathbb{E}_{p_t(x_1|x)}\left(u_t(x|x_1) - u_t(x)\right)\left(u_t(x|x_1) - u_t(x)\right)^\top. \tag{26}$$

Here, we used that $u_t(x) = \mathbb{E}_{p_t(x_1|x)}\left[u_t(x|x_1)\right]$ by (5). If we take the trace on both sides of (25), we get

$$\begin{aligned}&\mathrm{Tr}\left[\mathrm{Cov}_{p_t(x_1|x)}\left(\nabla_\theta \left\|v_t(x;\theta) - u_t(x|x_1)\right\|^2\right)\right] = \mathrm{Tr}\left[\left(\nabla_\theta v_t(x;\theta)\right)^\mathsf{T} \mathrm{Cov}_{p_t(x_1|x)}\left(u_t(x|x_1)\right)\left(\nabla_\theta v_t(x;\theta)\right)\right]\\ &= \mathrm{Tr}\left[\mathrm{Cov}_{p_t(x_1|x)}\left(u_t(x|x_1)\right)\left(\nabla_\theta v_t(x;\theta)\right)\left(\nabla_\theta v_t(x;\theta)\right)^\mathsf{T}\right] = \left\langle \mathrm{Cov}_{p_t(x_1|x)}\left(u_t(x|x_1)\right), \left(\nabla_\theta v_t(x;\theta)\right)\left(\nabla_\theta v_t(x;\theta)\right)^\mathsf{T}\right\rangle_F\\ &\le \left\|\mathrm{Cov}_{p_t(x_1|x)}\left(u_t(x|x_1)\right)\right\|_F \left\|\left(\nabla_\theta v_t(x;\theta)\right)\left(\nabla_\theta v_t(x;\theta)\right)^\mathsf{T}\right\|_F\\ &\le \mathbb{E}_{p_t(x_1|x)}\left\|\left(u_t(x|x_1) - u_t(x)\right)\left(u_t(x|x_1) - u_t(x)\right)^\top\right\|_F \left\|\left(\nabla_\theta v_t(x;\theta)\right)\left(\nabla_\theta v_t(x;\theta)\right)^\top\right\|_F\\ &= \left\|\nabla_\theta v_t(x;\theta)\right\|^2 \mathbb{E}_{p_t(x_1|x)}\left\|u_t(x|x_1) - u_t(x)\right\|^2.\end{aligned} \tag{27}$$

The second equality holds because $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ when both expressions are well defined, and the third equality holds by the definition of the Frobenius inner product $\langle\cdot,\cdot\rangle_F$. The first inequality holds by the Cauchy-Schwarz inequality. The second inequality holds by equation (26) and by the triangle inequality. In the last equality we used that for any vector v, $\|vv^\top\|_F = (\mathrm{Tr}(vv^\top, vv^\top))^{1/2} = \|v\|^2$. This proves (16).

To prove (17), we write:

$$\begin{aligned}&\mathbb{E}_{t,p_t(x)}[\sigma^2_{t,x}]\\ &\le \mathbb{E}_{t,p_t(x)}[\|\nabla_\theta v_t(x;\theta)\|^2 \mathbb{E}_{p_t(x_1|x)}\|u_t(x|x_1) - u_t(x)\|^2]\\ &\le \max_{x,t}\|\nabla_\theta v_t(x;\theta)\|^2 \times \mathbb{E}_{t,p_t(x)}[\mathbb{E}_{p_t(x_1|x)}\|u_t(x|x_1) - u_t(x)\|^2]\\ &= \max_{x,t}\|\nabla_\theta v_t(x;\theta)\|^2 \times \mathbb{E}_{t,q(x_0,x_1)}[\|u_t(x_t|x_1) - v_t(x_t;\theta)\|^2] \le \max_{t,x}\|\nabla_\theta v_t(x;\theta)\|^2 \times \mathcal{L}_{\mathrm{JCFM}}\end{aligned} \tag{28}$$

Here, the first inequality holds by (16), and the last inequality holds because $u_t(x)$ is the minimizer of $\mathcal{L}_{\mathrm{JCFM}}$.

D.3. Proof of Lemma 4.1

For an arbitrary test function f, by the construction of q we write

$$\mathbb{E}_{q(x_0,x_1)} f(x_0) = \mathbb{E}_{\{x_0^{(i)}\}_{i=1}^k \sim q_0, \{x_1^{(i)}\}_{i=1}^k \sim q_1} \mathbb{E}_{q^k(x_0,x_1)} f(x_0). \tag{29}$$

Since q^k has marginal $\frac{1}{k}\sum_{i=1}^k \delta(x_0 - x_0^{(i)})$ because π is a doubly stochastic matrix, we obtain that $\mathbb{E}_{q^k(x_0,x_1)} f(x_0) = \frac{1}{k}\sum_{i=1}^k f(x_0^{(i)})$ and then the right-hand side is equal to

$$\mathbb{E}_{\{x_0^{(i)}\}_{i=1}^k \sim q_0, \{x_1^{(i)}\}_{i=1}^k \sim q_1} \frac{1}{k}\sum_{i=1}^k f(x_0^{(i)}) = \mathbb{E}_{q_0(x_0)} f(x_0), \tag{30}$$

which proves that the marginal of q for x_0 is q_0. The same argument works for the x_1 marginal.

D.4. Proof of Theorem 4.2

Notation We begin by recalling and introducing some additional notation. Let $\boldsymbol{X}_0 = (x_0^i)_{i=1}^{+\infty}$, $\boldsymbol{X}_1 = (x_1^i)_{i=1}^{+\infty}$ be sequences of i.i.d. samples from the distributions q_0 and q_1, and denote by $\boldsymbol{X}_0^k = (x_0^i)_{i=1}^k$, $\boldsymbol{X}_1^k = (x_1^i)_{i=1}^k$ the finite sequences containing the initial k samples. We denote by q_0^k and q_1^k the empirical distributions corresponding to $\boldsymbol{X}_0^k$ and $\boldsymbol{X}_1^k$, i.e. $q_0^k = \frac{1}{k}\sum_{i=1}^k \delta_{x_0^i}$, $q_1^k = \frac{1}{k}\sum_{i=1}^k \delta_{x_1^i}$. Let q^k be the distribution over $\mathbb{R}^d \times \mathbb{R}^d$ which is output by the matching algorithm; q^k has marginals that are equal to q_0^k and q_1^k. Let q^* be the optimal transport plan between q_0 and q_1, and let $\tilde{q}^k$ be the optimal transport plan between q^k and q under the quadratic cost. Using this additional notation, we rewrite some of the objects that were defined in the main text in a lengthier, more precise way:

(i) The marginal vector field corresponding to sample size k:

$$u_t^k(x) = \mathbb{E}_{\boldsymbol{X}_0^k \overset{\text{iid}}{\sim} q_0, \boldsymbol{X}_1^k \overset{\text{iid}}{\sim} q_1, (x_0,x_1)\sim q^k}[x_1 - x_0 | x = tx_1 + (1-t)x_0], \qquad \forall t \in [0,1]. \tag{31}$$

We made the dependency on k explicit, and we used that $\psi_t(x_0|x_1) = tx_1 + (1-t)x_0$. Note that equivalently, we can write u_t^k as the solution of a simple variational problem.

$$u_t^k = \operatorname*{argmin}_{u_t} \mathbb{E}_{\boldsymbol{X}_0^k \overset{\text{iid}}{\sim} q_0, \boldsymbol{X}_1^k \overset{\text{iid}}{\sim} q_1, (x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t(tx_1 + (1-t)x_0)\|^2, \qquad \forall t \in [0,1]. \tag{32}$$

(ii) The flow $\psi_t^k(x_0)$ corresponding to u_t^k, i.e. the solution of $\frac{dx_t}{dt} = u_t^k(x_t)$ with initial condition x_0. We made the dependency on k explicit.

(iii) The straightness of the flow ψ_t^k:

$$S^k = \mathbb{E}_{t\sim \mathrm{U}(0,1), x_0 \sim q_0}\left[\|u_t^k(\psi_t^k(x_0))\|^2 - \|\psi_1^k(x_0) - x_0\|^2\right]. \tag{33}$$

Assumptions We will use the following three assumptions, which allow us to potentially extend our result beyond BatchOT:

(A1) The distributions q_0 and q_1 over $\mathbb{R}^d$ have bounded supports, i.e. there exists $C > 0$ such that for any $x \in \mathrm{supp}(q_0) \cup \mathrm{supp}(q_1)$, $\|x\| \leq C$.
(A2) q_0 admits a density and the optimal transport map T between q_0 and q_1 under the quadratic cost is continuous.
(A3) We assume that almost surely w.r.t. the draw of $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$, q^k converges weakly to q as $k \to \infty$.

Some comments are in order as to when assumptions **(A2)**, **(A3)** hold, since they are not directly verifiable. By the Caffarelli regularity theorem (see Villani (2008), Ch. 12, originally in Caffarelli (1992)), a sufficient condition for **(A2)** to hold is the following:

(A2') q_0 and q_1 have a common support Ω which is compact and convex, have α-Hölder densities, and they satisfy the lower bound $q_0, q_1 > \gamma$ for some $\gamma > 0$.

Assumption **(A3)** holds when the matching algorithm is BatchOT, that is, when q^k is the optimal transport plan between q_0^k and q_1^k, as shown by the following proposition, which is proven in App. D.4.3.

Proposition D.1. *Let q^k be the optimal transport plan between q_0^k and q_1^k under the quadratic cost (i.e. the result of Steps [1-3] under BatchOT). We have that almost surely w.r.t. the draws of $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$, the sequence $(q_k)_{k\geq 0}$ converges weakly to q^*, i.e. assumption **(A3)** holds.*

Proof structure We split the proof of Theorem 4.2 into two parts: in Subsubsec. D.4.1 we prove that the optimal value of the Joint CFM objective (15) converges to zero as $k \to \infty$. In Subsubsec. D.4.2, we prove that the straightness converges to zero and the transport cost converges to the optimal transport cost as $k \to \infty$.

D.4.1. Convergence of the optimal value of the CFM objective

Theorem D.2. *Suppose that assumptions **(A1), (A2)** and **(A3)** hold. We have that*

$$\lim_{k\to\infty} \mathbb{E}_{t\sim \mathrm{U}(0,1), \boldsymbol{X}_0^k \overset{\mathrm{iid}}{\sim} q_0, \boldsymbol{X}_1^k \overset{\mathrm{iid}}{\sim} q_1, (x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^k(tx_1 + (1-t)x_0)\|^2 = 0, \tag{34}$$

where u_t^k is the marginal vector field as defined in (31).

Proof. The transport plan q^* satisfies the non-crossing paths property, that is, for each $x \in \mathbb{R}^d$ and $t \in [0,1]$, there exists at most one pair (x_0, x_1) such that $x = tx_1 + (1-t)x_0$ (Nurbekyan et al., 2020; Villani, 2003). Consequently, when such a pair (x_0', x_1') exists, we have that the analogue of the vector field in (31) admits a simple expression:

$$u_t^*(x) := \mathbb{E}_{(x_0,x_1)\sim q^*}[x_1 - x_0 | x = tx_1 + (1-t)x_0] = x_1' - x_0' \tag{35}$$

This directly implies that

$$\mathbb{E}_{(x_0,x_1)\sim q^*} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 = 0. \tag{36}$$

Applying this, we can write

$$\begin{aligned} &\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 \\ &= |\mathbb{E}_{(x_0,x_1)\sim q^k}[\mathbb{E}_{t\sim \mathrm{U}(0,1)} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2] \\ &\quad - \mathbb{E}_{(x_0,x_1)\sim q^*}[\mathbb{E}_{t\sim \mathrm{U}(0,1)} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2]|. \end{aligned} \tag{37}$$

Now, define the function $f : \mathrm{supp}(q_0) \times \mathrm{supp}(q_1) \to \mathbb{R}$ as

$$f(x_0, x_1) = \mathbb{E}_{t\sim \mathrm{U}(0,1)} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2. \tag{38}$$

By Lemma D.3, which holds under **(A1)** and **(A2)**, we have that f is bounded and continuous. Assumption **(A3)** states that almost surely w.r.t. the draws of $\boldsymbol{X}_0$ and $\boldsymbol{X}_1$, the measure q_k converges weakly to q^*. We apply the definition of weak convergence of measures, which implies that almost surely,

$$\lim_{k\to\infty} \mathbb{E}_{(x_0,x_1)\sim q^k}[f(x_0,x_1)] = \mathbb{E}_{(x_0,x_1)\sim q}[f(x_0,x_1)]. \tag{39}$$

Equivalently, the right-hand side of (37) converges to zero as k tends to infinity. Hence, $\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 \to 0$ almost surely. Almost sure convergence implies convergence in probability, which means that

$$\Pr(\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 > \epsilon) \xrightarrow{k\to\infty} 0, \quad \forall \epsilon > 0. \tag{40}$$

Here, the randomness comes only from drawing the random variables $\boldsymbol{X}_0^k, \boldsymbol{X}_1^k$. Also, using again that f is bounded, say by the constant $C > 0$, we can write $\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 \le C$, for all $k \ge 0$. A crude bound yields

$$\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 \tag{41}$$

$$\le \epsilon + C\Pr(\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 > \epsilon). \tag{42}$$

In this equation and from now, we write $\boldsymbol{X}_0^k, \boldsymbol{X}_1^k$ instead of $\boldsymbol{X}_0^k \overset{\mathrm{iid}}{\sim} q_0, \boldsymbol{X}_1^k \overset{\mathrm{iid}}{\sim} q_1$ for shortness. We can take ϵ arbitrarily small, and for a given ϵ we can make the second term in the right-hand side arbitrarily small by taking k large enough. Hence, we obtain that

$$\lim_{k\to\infty} \mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1)\sim q^k} \|x_1 - x_0 - u_t^*(tx_1 + (1-t)x_0)\|^2 = 0. \tag{43}$$

To conclude the proof, we use the variational characterization of u_t^k given in (32), which implies that

$$\begin{aligned}&\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1)\sim q^k}\|x_1-x_0-u_t^k(tx_1+(1-t)x_0)\|^2\\&\leq \mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1)\sim q^k}\|x_1-x_0-u_t^*(tx_1+(1-t)x_0)\|^2\to 0.\end{aligned} \tag{44}$$

□

Lemma D.3. *Let f be the function defined in equation (38). Suppose that assumptions **(A1)** and **(A2)** hold. Then, f is bounded and continuous.*

Proof. First, we show that the function u_t^* defined in equation (35) is bounded and continuous wherever it is defined. It is bounded because $u_t^*(x)=x_1'-x_0'$ for some x_0' in $\mathrm{supp}(q_0)$ and x_1' in $\mathrm{supp}(q_1)$, which are both bounded by assumption.

To show that u_t^* is continuous, we use that q_0 is absolutely continuous and that consequently a transport map T exists. Moreover, we have that $x_1'=T(x_0')$. Consider the transport map T_t at time t, defined as $T_t(x)=tT(x)+(1-t)x$. Thus, we can write that $u_t^*(T_t(x_0))=T(x_0)-x_0$. The non-crossing paths property implies that T_t is invertible, which means that an inverse T_t^{-1} exists. We can write

$$u_t^*(x)=T(T_t^{-1}(x))-T_t^{-1}(x). \tag{45}$$

By assumption **(A2)**, the transport map T is continuous, and so is T_t. It is well-known fact that if E,E' are metric spaces, E is compact, and $f:E\to E'$ a continuous bijective function, then $f^{-1}:E'\to E$ is continuous. Thus, T_t^{-1} is also continuous. From equation (45), we conclude that u_t^* is continuous.

The rest of the proof is straightforward: $(x_1,x_0)\mapsto\|x_1-x_0-u_t^*(tx_1+(1-t)x_0)\|^2$ is bounded and continuous on the bounded supports of q_0 and q_1 for all $t\in[0,1]$, and then f is also continuous and bounded since it is an average of continuous bounded functions, applying the dominated convergence theorem. □

D.4.2. Convergence of the straightness and the transport cost

Theorem D.4. *Suppose that assumptions **(A1)** and **(A3)** hold. Then,*

(i) We have that $\lim_{k\to\infty}S^k=0$, where S^k is the straightness defined in (33).

(ii) We also have that

$$\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2\geq \mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0)-x_0\|^2\geq W_2^2(q_0,q_1), \tag{46}$$

$$\lim_{k\to\infty}\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2=\lim_{k\to\infty}\mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0)-x_0\|^2=W_2^2(q_0,q_1). \tag{47}$$

Proof. We begin with the proof of (i). We introduce some additional notation. We define the quantity S^* in analogy with S^k:

$$S^*=\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\left[\|u_t^*(\psi_t^*(x_0))\|^2-\|\psi_1^*(x_0)-x_0\|^2\right], \tag{48}$$

and $\psi_t^*(x_0)$ as the solution of the ODE $\frac{dx_t}{dt}=u_t^*(x_t)$. Since the trajectories for the optimal transport vector field are straight lines, we deduce from the alternative expression of the straightness (equation (19)) that $S^*=0$. An alternative way to see this is by the Benamou-Brenier theorem (Benamou & Brenier, 2000), which states that the dynamic optimal transport cost $\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2$ is equal to the static optimal transport cost $\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|\psi_1^*(x_0)-x_0\|^2$.

We will first prove that $\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2$ converges to $\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2$ and then that $\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|\psi_1^k(x_0)-x_0\|^2$ converges to $\mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|\psi_1^*(x_0)-x_0\|^2$.

For given instances of $\boldsymbol{X}_0^k$ and $\boldsymbol{X}_1^k$, let $\tilde{q}_k$ be the optimal transport plan between the optimal transport plans q and q^k. In other words, $\tilde{q}_k$ is a measure over the variables x_0,x_1,x_0',x_1', and is such that its marginal w.r.t. x_0,x_1 is q, while its marginal w.r.t. x_0',x_1' is q^k.

That is, we will use that for all $t\in[0,1]$, the random variable $tx_1+(1-t)x_0$, with $(x_0,x_1)\sim q^k$, and q^k built randomly from $\boldsymbol{X}_0^k\overset{\text{iid}}{\sim}q_0,\boldsymbol{X}_1^k\overset{\text{iid}}{\sim}q_1$, has the same distribution as the random variable $\psi_t^k(x_0)$, with $x_0\sim q_0$. This is a direct

consequence of Lemma 3.1, i.e. the marginal vector field u_t generates the marginal probability path p_t. An analogous statement holds for q, i.e. the random variable $tx_1 + (1-t)x_0$, with $(x_0, x_1) \sim q$, has the same distribution as the random variable $\psi_t^*(x_0)$, with $x_0 \sim q_0$. However, in this case it can be obtained immediately by the non-crossing paths property of the optimal transport plan. Hence,

$$\begin{aligned}\mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2 &= \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\|u_t^*(tx_1+(1-t)x_0)\|^2,\\ \mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2 &= \mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1)\sim q^k}\|u_t^k(tx_1+(1-t)x_0)\|^2.\end{aligned} \tag{49}$$

Using this and the definition of $\tilde{q}^k$, and applying Jensen's inequality, the Cauchy-Schwarz inequality and the triangle inequality, we can write

$$\begin{aligned}&\left|\mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2 - \mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2\right|\\
&= \left|\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\|u_t^*(tx_1+(1-t)x_0)\|^2 - \mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0',x_1')\sim q^k}\|u_t^k(tx_1'+(1-t)x_0')\|^2\right|\\
&= \left|\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1,x_0',x_1')\sim \tilde{q}^k}\left[\|u_t^*(tx_1+(1-t)x_0)\|^2 - \|u_t^k(tx_1'+(1-t)x_0')\|^2\right]\right|\\
&= \left|\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1,x_0',x_1')\sim \tilde{q}^k}\left[(\|u_t^*(tx_1+(1-t)x_0)\| - \|u_t^k(tx_1'+(1-t)x_0')\|)\right.\right.\\
&\qquad\qquad \left.\left.\times(\|u_t^*(tx_1+(1-t)x_0)\| + \|u_t^k(tx_1'+(1-t)x_0')\|)\right]\right|\\
&\le \left(\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1,x_0',x_1')\sim \tilde{q}^k}\left[(\|u_t^*(tx_1+(1-t)x_0)\| - \|u_t^k(tx_1'+(1-t)x_0')\|)^2\right]\right)^{1/2}\\
&\qquad \times\left(\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1,x_0',x_1')\sim \tilde{q}^k}\left[(\|u_t^*(tx_1+(1-t)x_0)\| + \|u_t^k(tx_1'+(1-t)x_0')\|)^2\right]\right)^{1/2}\\
&\le \left(\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1,x_0',x_1')\sim \tilde{q}^k}\|u_t^*(tx_1+(1-t)x_0) - u_t^k(tx_1'+(1-t)x_0')\|^2\right)^{1/2}\\
&\qquad \times\left(\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^k,\boldsymbol{X}_1^k,(x_0,x_1,x_0',x_1')\sim \tilde{q}^k}\left[(\|u_t^*(tx_1+(1-t)x_0)\| + \|u_t^k(tx_1'+(1-t)x_0')\|)^2\right]\right)^{1/2}.\end{aligned} \tag{50}$$

Remark that the second factor in the right-hand side is bounded because u_t^* and u_t^k are bounded. Using Lemma D.5, we obtain that the first factor in the right-hand side tends to zero as k grows. Thus,

$$\left|\mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2 - \mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0^k}\|u_t^k(\psi_t^k(x_0))\|^2\right| \xrightarrow{k\to\infty} 0. \tag{51}$$

Now, since $\mathbb{E}_{x_0\sim q_0}\|\psi_1^*(x_0) - x_0\|^2$ is the optimal cost and $S^* = 0$, we write

$$\left|\mathbb{E}_{x_0\sim q_0}\|\psi_1^*(x_0) - x_0\|^2 - \mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0) - x_0\|^2\right| = \mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0) - x_0\|^2 - \mathbb{E}_{x_0\sim q_0}\|\psi_1^*(x_0) - x_0\|^2 \tag{52}$$

$$= \mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0) - x_0\|^2 - \mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2. \tag{53}$$

Since ψ_t^k is the flow of u_t^k and by Jensen's inequality, we have that

$$\begin{aligned}\mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0) - x_0\|^2 &= \mathbb{E}_{x_0\sim q_0}\left\|\int_0^1 u_s^k(\psi_s^k(x_0'))\,ds\right\|^2\\ &\le \mathbb{E}_{x_0\sim q_0}\int_0^1 \|u_s^k(\psi_s^k(x_0))\|^2\,ds = \mathbb{E}_{t\sim U(0,1), x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2.\end{aligned}$$

Plugging this into (52), we get that

$$\left|\mathbb{E}_{x_0\sim q_0}\|\psi_1^*(x_0) - x_0\|^2 - \mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0) - x_0\|^2\right| \tag{54}$$

$$\le \mathbb{E}_{t\sim U(0,1), x_0\sim q_0}\|u_t^k(\psi_t^k(x_0))\|^2 - \mathbb{E}_{t\sim \mathrm{U}(0,1), x_0\sim q_0}\|u_t^*(\psi_t^*(x_0))\|^2 \xrightarrow{k\to\infty} 0, \tag{55}$$

where the limit holds by . Putting together (51) and (D.4.2), we end up with $S^k = |S^* - S^k| \xrightarrow{k\to\infty} 0$, which proves (i).

We prove (ii). The first inequality in (46) holds because $S^k \ge 0$ since it can be written in a form analogous to (19). The second inequality in (46) holds because $\mathbb{E}_{x_0\sim q_0}\|\psi_1^k(x_0) - x_0\|^2$ is the squared transport cost for the map $x \mapsto \psi_1^k(x)$, which must be at least as large as the optimal cost. The first equality in (46) is a direct consequence of (i). To prove the second equality in (46), we remark that $W_2^2(q_0, q_1) = \mathbb{E}_{x_0\sim q_0}\|\psi_1^*(x_0) - x_0\|^2$. Then, equation (D.4.2) readily implies that $|\mathbb{E}_{x_0\sim q_0^k}\|\psi_1^k(x_0) - x_0\|^2 - W_2^2(q_0, q_1)| \xrightarrow{k\to\infty} 0$. □

Lemma D.5. *Suppose that assumptions **(A1)** and **(A3)** hold. Let $\tilde{q}^k$ be the optimal transport plan between the optimal transport plans q and q^k. We have that*

$$\lim_{k\to\infty} \mathbb{E}_{t\sim \mathrm{U}(0,1), \boldsymbol{X}_0^k \overset{\mathrm{iid}}{\sim} q_0, \boldsymbol{X}_1^k \overset{\mathrm{iid}}{\sim} q_1, (x_0,x_1,x_0',x_1')\sim \tilde{q}^k} \left[\|u_t^*(tx_1+(1-t)x_0) - u_t^k(tx_1'+(1-t)x_0')\|^2\right] = 0 \tag{56}$$

Proof. For given instances of $\boldsymbol{X}_0^k$ and $\boldsymbol{X}_1^k$, we can write

$$\begin{aligned}
&\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1,x_0',x_1')\sim\tilde{q}^k}\left[\|u_t^*(tx_1+(1-t)x_0) - u_t^k(tx_1'+(1-t)x_0')\|^2\right] \\
&= \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1,x_0',x_1')\sim\tilde{q}^k}\Big[\|\mathbb{E}_{\tilde{x}_0,\tilde{x}_1\sim q}[\tilde{x}_1-\tilde{x}_0 | tx_1+(1-t)x_0 = t\tilde{x}_1+(1-t)\tilde{x}_0] \\
&\qquad\qquad - \mathbb{E}_{\tilde{x}_0',\tilde{x}_1'\sim q^k}[\tilde{x}_1'-\tilde{x}_0' | tx_1'+(1-t)x_0' = t\tilde{x}_1'+(1-t)\tilde{x}_0']\|^2\Big] \\
&\le \mathbb{E}_{(x_0,x_1,x_0',x_1')\sim\tilde{q}^k}\left[\left\|x_1-x_0-(x_1'-x_0')\right\|^2\right] \le 2\mathbb{E}_{(x_0,x_1,x_0',x_1')\sim\tilde{q}^k}\left[\left\|x_1-x_1'\right\|^2 + \left\|x_0-x_0'\right\|^2\right] \\
&= 2\mathbb{E}_{(x_0,x_1,x_0',x_1')\sim\tilde{q}^k}\left[\left\|(x_0,x_1)-(x_0',x_1')\right\|^2\right] = 2W_2^2(q,q^k)
\end{aligned} \tag{57}$$

Assumption **(A3)** implies that almost surely, q^k converges to q weakly. For distributions on a bounded domain, weak convergence is equivalent to convergence in the Wasserstein distance (Villani, 2008, Thm. 6.8), and this means that $W_2^2(q,q^k) \xrightarrow{k\to\infty} 0$ almost surely. Almost sure convergence implies convergence in probability, which means that

$$\Pr(W_2^2(q,q^k) > \epsilon) \xrightarrow{k\to\infty} 0, \quad \forall \epsilon > 0. \tag{58}$$

Note that $W_2^2(q,q^k)$ is a bounded random variable because q and q^k have bounded support as q_0, q_1, q_0^k and q_1^k have bounded support. Suppose that $W_2^2(q,q^k)$ is bounded by the constant C. Hence, we can write

$$\mathbb{E}_{\boldsymbol{X}_0^k,\boldsymbol{X}_1^k}\mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1,x_0',x_1')\sim\tilde{q}^k}\left[\|u_t^*(tx_1+(1-t)x_0) - u_t^k(tx_1'+(1-t)x_0')\|^2\right] \tag{59}$$

$$\le 2\mathbb{E}_{\boldsymbol{X}_0^k,\boldsymbol{X}_1^k} W_2^2(q,q^k) \le 2\big(\epsilon + C\Pr(W_2^2(q,q^k)>\epsilon)\big). \tag{60}$$

We can take ϵ arbitrarily small, and for a given ϵ we can make the second term in the right-hand side arbitrarily small by taking k large enough. The final result follows. □

D.4.3. Proof of Proposition D.1

We have that almost surely, the empirical distributions q_0^k, resp. q_1^k, converge weakly to q_0, resp. q_1 (Varadarajan, 1958). Hence, we can apply Theorem D.6. Since convergence in distribution of random variables is equivalent to weak convergence of their laws, and the law of an optimal coupling is the optimal transport plan, we conclude that $(q_k)_{k\ge 0}$ converges weakly to q^*.

Theorem D.6 ((Cuesta-Albertos et al., 1997), Theorem 3.2)**.** *Let $(P_n)_n$, $(Q_n)_n$, P, Q be probability measures in $\mathcal{P}_2$ (the space of Borel probability measures with bounded second order moment) such that $P \ll \lambda_p$ (P is absolutely continuous with respect to the Lebesgue measure) and $P_n \xrightarrow{w} P$, $Q_n \xrightarrow{w} Q$, where $\xrightarrow{w}$ denotes weak convergence of probability measures. Let (X_n, Y_n) be an optimal coupling between P_n and Q_n, $n\in\mathbb{N}$, and (X,Y) an optimal coupling between P and Q. Then, $(X_n,Y_n) \xrightarrow{\mathcal{L}} (X,Y)$, where $\xrightarrow{\mathcal{L}}$ denotes convergence of random variables in distribution.*

D.5. Bounds on the transport cost and monotone convergence results

The following result shows that for an arbitrary joint distribution $q(x_0,x_1)$, we can upper-bound the transport cost associated to the marginal vector field u_t to a quantity that depends only $q(x_0,x_1)$.

Proposition D.7. *For an arbitrary joint distribution $q(x_0,x_1)$ with marginals $q_0(x_0)$ and $q_1(x_1)$, let ψ_t be the flow corresponding to the marginal vector field u_t. We have that*

$$\mathbb{E}_{q_0(x_0)}\|\psi_1(x_0)-x_0\|^2 \le \mathbb{E}_{q(x_0,x_1)}\|x_1-x_0\|^2, \tag{61}$$

Proof. We make use of the notation introduced in App. D.4. We will rely on the fact that for all $t\in[0,1]$, the random variable $tx_1+(1-t)x_0$, with $(x_0,x_1)\sim q$ has the same distribution as the random variable $\psi_t(x_0)$, with $x_0\sim q_0$. This is

a direct consequence of Lemma 3.1. Using that ψ_t is the flow for u_t and Jensen's inequality twice, we have that

$$\begin{aligned}
&\mathbb{E}_{x_0\sim q_0}\|\psi_1(x_0)-x_0\|^2 \\
&= \mathbb{E}_{x_0\sim q_0}\left\|\int_0^1 u_s(\psi_s(x_0))\,ds\right\|^2 \leq \mathbb{E}_{t\sim \mathrm{U}(0,1),x_0\sim q_0}\|u_t(\psi_t(x_0))\|^2 \\
&= \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\|u_t(tx_1+(1-t)x_0)\|^2 \\
&= \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\|\mathbb{E}_{(x_0',x_1')\sim q}\left[u_t(tx_1+(1-t)x_0|x_0',x_1')|tx_1+(1-t)x_0=tx_1'+(1-t)x_0'\right]\|^2 \\
&\leq \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\mathbb{E}_{(x_0',x_1')\sim q}\left[\|u_t(tx_1+(1-t)x_0|x_0',x_1')\|^2|tx_1+(1-t)x_0=tx_1'+(1-t)x_0'\right] \\
&= \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\mathbb{E}_{(x_0',x_1')\sim q}\left[\|x_1'-x_0'\|^2|tx_1+(1-t)x_0=tx_1'+(1-t)x_0'\right] \\
&= \mathbb{E}_{t\sim \mathrm{U}(0,1),(x_0,x_1)\sim q}\|x_1-x_0\|^2
\end{aligned} \tag{62}$$

as needed. □

Note that that the statement and proof of this proposition is equivalent to Theorem 3.5 of (Liu et al., 2022), although the language and notation that we use is different, which is why we though convenient to include it.

For the case of BatchOT, the following theorem shows that the quantity in the upper bound of (61) is monotonically decreasing in k. The combination of Proposition D.7 and Theorem D.8 provides a weak guarantee that for BatchOT, the transport cost should not get much higher when k increases.

Theorem D.8. *Suppose that Multisample Flow Matching is run with BatchOT. For clarity, we make the dependency on the sample size k explicit and let*[2] $q^{(k)}(x_0,x_1) := q(x_0,x_1)$*, and* $\psi_t^k(x_0) := \psi_t(x_0)$*. Then, for any $k \geq 1$, we have that*

$$\begin{aligned}
&\mathbb{E}_{q_0(x_0)}\|\psi_1^k(x_0)-x_0\|^2 \leq \mathbb{E}_{q^{(k)}(x_0,x_1)}\|x_1-x_0\|^2, \\
&\mathbb{E}_{q^{(k+1)}(x_0,x_1)}\|x_1-x_0\|^2 \leq \mathbb{E}_{q^{(k)}(x_0,x_1)}\|x_1-x_0\|^2.
\end{aligned} \tag{63}$$

Proof. We write

$$\begin{aligned}
&\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^{k+1},\boldsymbol{X}_1^{k+1}}\mathbb{E}_{(x_0,x_1)\sim q^{k+1}}\|x_1-x_0\|^2 = \frac{1}{k}\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^{k+1},\boldsymbol{X}_1^{k+1}}\left[\sum_{i=1}^k \|x_1^{(i)}-x_0^{(\sigma_{k+1}(i))}\|^2\right] \\
&= \frac{1}{k}\frac{1}{k+1}\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^{k+1},\boldsymbol{X}_1^{k+1}}\left[\sum_{j=1}^{k+1}\sum_{i\in[k+1]\setminus\{j\}} \|x_1^{(i)}-x_0^{(\sigma_{k+1}(i))}\|^2\right] \\
&\leq \frac{1}{k}\frac{1}{k+1}\mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^{k+1},\boldsymbol{X}_1^{k+1}}\left[\sum_{j=1}^{k+1}\sum_{i\in[k+1]\setminus\{j\}} \|x_1^{(i)}-x_0^{(\sigma_k^{-j}(i))}\|^2\right] \\
&= \mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^{k},\boldsymbol{X}_1^{k}}\left[\frac{1}{k}\sum_{j=1}^{k} \|x_1^{(i)}-x_0^{(\sigma_k(i))}\|^2\right] = \mathbb{E}_{t\sim \mathrm{U}(0,1),\boldsymbol{X}_0^{k},\boldsymbol{X}_1^{k}}\mathbb{E}_{(x_0,x_1)\sim q^k}\|x_1-x_0\|^2.
\end{aligned} \tag{64}$$

In the first equality, we used that the optimal transport map between the empirical distributions q_0^k and q_1^k can be encoded as a permutation, which we denote by σ_{k+1}. In the inequality, we introduced the notation σ_k^{-j} to denote the optimal permutation within $\{x_0^{(i)}\}_{i\in[k+1]\setminus\{j\}}$. The inequality holds because using the optimality of σ_{k+1}:

$$\begin{aligned}
&\sum_{j=1}^{k+1}\sum_{i\in[k+1]\setminus\{j\}} \|x_1^{(i)}-x_0^{(\sigma_{k+1}(i))}\|^2 \leq \sum_{j=1}^{k+1}\sum_{i\in[k+1]} \|x_1^{(i)}-x_0^{(\sigma_{k+1}(i))}\|^2 \\
&\leq \sum_{j=1}^{k+1}\left(\sum_{i\in[k+1]\setminus\{j\}} \|x_1^{(i)}-x_0^{(\sigma_k^{-j}(i))}\|^2 + \|x_1^{(j)}-x_0^{(j)}\|^2\right) \leq \sum_{j=1}^{k+1}\sum_{i\in[k+1]\setminus\{j\}} \|x_1^{(i)}-x_0^{(\sigma_k^{-j}(i))}\|^2.
\end{aligned} \tag{65}$$

□

[2]Note that here $q^{(k)} := q$ is a marginalized distribution and is different from q^k defined in Step 3.

E. Experimental & evaluation details

	ImageNet-32	ImageNet-64
Channels	256	192
Depth	3	3
Channels multiple	1,2,2,2	1,2,3,4
Heads	4	4
Heads Channels	64	64
Attention resolution	4	8
Dropout	0.0	0.1
Batch size / GPU	256	50
GPUs	4	16
Effective Batch size	1024	800
Epochs	350	575
Effective Iterations	438k	957k
Learning Rate	1e-4	1e-4
Learning Rate Scheduler	Polynomial Decay	Constant
Warmup Steps	20k	-

Table 10. Hyper-parameters used for training each model.

E.1. Image datasets

We report the hyper-parameters used in Table 10. We use the architecture from Dhariwal & Nichol (2021) but with much lower attention resolution. We use full 32 bit-precision for training ImageNet-32 and 16-bit mixed precision for training ImageNet-64. 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 using identical architectures, with the same parameters for the the same number of epochs (see Table 10 for details), with the exception of Rectified Flow, which we trained for much longer starting from the fully trained CondOT model. We use either a constant learning rate schedule or a polynomial decay schedule (see Table 10). 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 10). 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 (Dinh et al., 2016). We report an importance-weighted estimate using

$$\text{BPD}(K) = -\frac{1}{D}\log_2 \frac{1}{K}\sum_{k=1}^{K} p_t(x+u_k), \text{ where } u_k \sim [U(0,1)]^D, \tag{66}$$

with x is in $\{0,\ldots,255\}^D$. We solve for p_t at exactly $t = 1$ with an adaptive step size solver `dopri5` with `atol=rtol=1e-5` using the `torchdiffeq` (Chen, 2018) library. We used K=15 for ImageNet32 and K=10 for ImageNet64.

When computing FID, we use the TensorFlow-GAN library `https://github.com/tensorflow/gan`.

We run coupling algorithms only within each GPU. We also ran coupling algorithms across all GPUs (using the "Effective Batch Size") in preliminary experiments, but did not see noticeable gains in sample efficiency while obtaining slightly worse performance and sample quality, so we stuck to the smaller batch sizes for running our coupling algorithms.

For Rectified Flow, we use the finalized FM-CondOT model, generate 50000 noise and sample pairs, then train using the same FM-CondOT algorithm and hyperparameters on these sampled pairs. This is equivalent to their 2-Rectified Flow approach (Liu et al., 2022). For the rectification process, we train for 300 epochs.

E.2. Improved batch optimal couplings

Datasets. We experimented with 3 datasets in dimensions $\{2, 32, 64\}$ consisting of $50K$ samples. Both q_0 and q_1 were Gaussian mixtures with number of centers described in Table 11.

Neural Networks Architectures. For B-ST we used stacked blocks of Convex Potential Flows (Huang et al., 2020) as an invertible neural network parametrizing the map, which also allowed us to estimate KL divergence:

$$\mathrm{KL}(q_1 || (\psi_1)_\sharp\, q_0) = \mathbb{E}_{x\sim q_1}\left[\log q_1(x) - \log\left((\psi_1)_\sharp q_0\right)(x)\right]. \tag{67}$$

For B-FM we used a simple MLP with Swish activation. For each dataset we built architectures with roughly the same number of parameters.

Hyperparameter Search. For each dataset and each cost we swept over learning rates $\{0.005, 0.001, 0.0005\}$ and chose the best setting.

	2-D	32-D	64-D
q_0 #centers	1	50	100
q_1 #centers	8	50	100
#params	50K	800K	800K
batch size	128	1024	1024
epochs	100	1000	1000

Table 11. Hyperparameters for experiments on synthetic datasets.