WEIGHTED CONDITIONAL FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Conditional flow matching (CFM) has emerged as a powerful framework for training continuous normalizing flows due to its computational efficiency and effectiveness. However, standard CFM often produces paths that deviate significantly from straight-line interpolations between prior and target distributions, making generation slower and less accurate due to the need for fine discretization at inference. Recent methods enhance CFM performance by inducing shorter and straighter trajectories but typically rely on computationally expensive mini-batch optimal transport (OT). Drawing insights from entropic optimal transport (EOT), we propose *weighted conditional flow matching* (W-CFM), a novel approach that modifies the classical CFM loss by weighting each training pair (x, y) with a Gibbs kernel. We show that this weighting recovers the entropic OT coupling up to some bias in the marginals, and we provide conditions under which the marginals remain nearly unchanged. Moreover, we establish an equivalence between W-CFM and the minibatch OT method in the large-batch limit, showing how our method overcomes computational and performance bottlenecks linked to batch size. Empirically, we test our method on unconditional generation on various synthetic and real datasets, confirming that W-CFM achieves comparable or superior sample quality, fidelity, and diversity to other alternative baselines while maintaining the computational efficiency of vanilla CFM.

1 INTRODUCTION

Generative modeling seeks to learn a parameterized transformation that maps a simple prior (e.g., a Gaussian) to a complex data distribution. Continuous normalizing flows (CNFs) instead train a time-dependent vector field to solve an ordinary differential equation (ODE) that transports base samples to data samples with exact likelihood computation and invertibility. However, training CNFs by likelihood maximization suffers from training instability and fails to scale efficiently to large or high-dimensional datasets (Chen et al., 2018; Grathwohl et al., 2018; Onken et al., 2021). Flow matching (FM) (Lipman et al., 2023; Albergo et al., 2023; Liu et al., 2023) reframes CNF training as a simple regression problem: a vector field is learned to match the endpoint displacement between a prior sample and its paired data point, yielding near-optimal transport trajectories when the prior is Gaussian. However, the independent pairing of FM cannot ensure that the marginal flow follows an optimal transport geodesic, leading to suboptimal paths in practice. In its most general form, conditional flow matching (CFM) (Lipman et al., 2023; Tong et al., 2024) generalizes FM by learning a vector field that transports samples from an arbitrary transport map, conditioned on paired source and target samples. This method allows for simulation-free training of continuous normalizing flows and can learn conditional generative models from any sampleable source distribution, extending beyond the Gaussian source.

Thanks to its flexibility, CFM has been applied in many areas of science, such as molecule generation (Irwin et al., 2024; Geffner et al., 2025), sequence and time-series modeling (Stark et al., 2024; Zhang et al., 2024; Rohbeck et al., 2025), and text-to-speech translation (Guo et al., 2024). A refinement of CFM is minibatch CFM (OT-CFM) (Pooladian et al., 2023; Tong et al., 2024), which uses an entropic or exact plan as the coupling so that each training pair is drawn according to the optimal transport solution between the minibatch source and target samples. This yields substantially straighter, lower-cost trajectories in the learned flow and improves sample quality with fewer integration steps. However, computing these OT plans—even approximately via Sinkhorn—for every minibatch incurs substantial per-iteration overhead, scaling cubically with batch size (or quadratically under entropic regularization). Moreover, requiring well-balanced class representation in each batch to approximate the global OT plan makes this approach impractical for large, multi-class datasets.

As an alternative that addresses these limitations, we introduce weighted conditional flow matching (W-CFM), which replaces costly batch-level transport computations by simply weighting each independently sampled pair (x, y) with the entropic OT (EOT) Gibbs kernel, $w(x, y) = \exp(-c(x, y)/\varepsilon)$ (Cuturi, 2013). This importance weighting provably recovers the entropic OT (EOT) plan up to a controllable bias in the marginals. As a result, the learned flow follows straight paths without ever explicitly solving an OT problem during training. Moreover, we show that W-CFM matches OT-CFM in the large-batch limit, thereby not incurring any of the batch size-related limitations or any extra costs. In practice, W-CFM delivers straight flows and high-quality samples consistently outperforming CFM and achieving comparable performance to OT-CFM, but with no extra overhead. In short, our contributions are the following:

- We introduce a novel CFM variant that, inspired by EOT, incorporates a Gibbs kernel weight on each sample pair and show that our method serves as a new way to approximate the EOT coupling without any additional cost during training.
- We discuss practical design choices to alleviate the change of the marginals with the new loss and derive sufficient conditions under which this change becomes trivial, so the true marginals are approximately preserved.
- We show that as the batch size grows and assuming that the bias in the marginals remains negligible, W-CFM converges to an entropy-regularized version of OT-CFM, retaining some trajectory straightness without the computational scaling issues associated with the OT plan computations.
- We demonstrate on toy and image-generation benchmarks that W-CFM matches or outperforms existing CFM and OT-CFM methods in sample quality, fidelity, and diversity under a sensible choice of the hyperparameter ε in the Gibbs kernel.

2 BACKGROUND

2.1 ENTROPIC OPTIMAL TRANSPORT

We refer the reader to Nutz (2021) for a comprehensive introduction to the topic and only mention the relevant results for our work. For $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, we recall the definition of the Kullback–Leibler divergence $D_{\mathrm{KL}}(\mu \| \nu) := \int_{\mathbb{R}^d} \log \frac{d\mu}{d\nu}(x) d\mu(x)$ if $\mu \ll \nu$ and $+\infty$ otherwise. Assume that μ, ν have finite first moment. We denote the set of couplings between μ and ν by $\Pi(\mu, \nu) := \{\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) : \pi(\mathbb{R}^d, dy) = \nu(dy), \pi(dx, \mathbb{R}^d) = \mu(dx)\}$. We consider the following entropic optimal transport (EOT) problem with parameter $\varepsilon > 0$:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\pi(x, y) + \varepsilon D_{\mathrm{KL}}(\pi \| \mu \otimes \nu), \tag{1}$$

where $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ denotes a cost function, typically $c(x, y) = \|x - y\|$, which is taken such that equation 1 is finite. When $\varepsilon = 0$, we recover the classical Monge–Kantorovich transportation problem of moving a distribution μ to a distribution ν by minimizing the transport cost as measured by c, whose solution is not necessarily unique; we denote by $\pi^\star \in \Pi(\mu, \nu)$ such a solution. EOT is of significant importance in machine learning and scientific computing (Genevay et al., 2018; Peyré & Cuturi, 2019), as it approximates the original Monge–Kantorovich transport problem and can be solved tractably with Sinkhorn's algorithm (Cuturi, 2013; Altschuler et al., 2017). It is a classical result (see, e.g., Chapter 4 in Peyré & Cuturi (2019)) that solving equation 1 is equivalent to solving the following projection problem $\min_{\pi \in \Pi(\mu,\nu)} D_{\mathrm{KL}}(\pi \| \mathcal{K}_\varepsilon)$, with the Gibbs kernel $\mathcal{K}_\varepsilon(dx, dy) := e^{-c(x,y)/\varepsilon} \mu(dx)\nu(dy)$. A convexity argument can be made to prove that there exists a unique minimizer $\pi_\varepsilon \in \Pi(\mu, \nu)$ of equation 1. More specifically, this projection formulation allows us to write the optimal coupling in terms of $\mathcal{K}_\varepsilon$.

Theorem 1 (Theorem 4.2 in Nutz (2021)). *If $c(x, y) < \infty$ $\mu \otimes \nu$-almost-surely, there exist measurable functions $\phi, \psi : \mathbb{R}^d \to \mathbb{R}$, referred to as Schrödinger potentials, such that the EOT plan for any $\varepsilon > 0$ is given by*

$$\pi_\varepsilon(dx, dy) = \exp\left(\phi_\varepsilon(x) + \psi_\varepsilon(y) - \frac{c(x, y)}{\varepsilon}\right) \mu(dx)\nu(dy). \tag{2}$$

In other words, we have $\pi_\varepsilon(dx, dy) = f_\varepsilon(x)g_\varepsilon(y)\mathcal{K}_\varepsilon(dx, dy)$ for some positive functions $f_\varepsilon(x) = \exp(\phi_\varepsilon(x)), g_\varepsilon(y) = \exp(\psi_\varepsilon(y))$. In particular, the component capturing the dependence in the EOT plan π_ε is exactly given by the Gibbs kernel, which is easy to compute, and one only needs to adjust the marginals independently to obtain π_ε.

2.2 Conditional Flow Matching

The flow matching methodology Lipman et al. (2023) is a simulation-free method of training a continuous normalizing flow (Chen et al., 2018), i.e., a smooth vector field $v_\theta(t, x)$, for generating samples from a target distribution $\nu \in \mathcal{P}(\mathbb{R}^d)$ given a source distribution $\mu \in \mathcal{P}(\mathbb{R}^d)$. Assume that there exists a vector field $v_t(x)$ such that the flow given by $dx_t = v_t(x_t)dt$ with initial condition $x_0 \sim \mu$ satisfies $x_1 \sim \nu$. The flow matching loss is given by

$$\mathcal{L}_{\mathrm{FM}}(\theta) \coloneqq \mathbb{E}_{t\sim\mathcal{U}(0,1),X_t\sim p_t}\left[\|v_\theta(t, X_t) - v_t(X_t)\|^2\right], \tag{3}$$

where p_t denotes the distribution of x_t, i.e. a probability path between μ and ν. Under technical assumptions, v_t generates p_t if and only if they satisfy the continuity equation

$$\frac{\partial p_t}{\partial t} + \nabla \cdot (p_t v_t) = 0, \quad p_0 = \mu, \quad p_1 = \nu. \tag{4}$$

Upon finding a minimizer of equation 3, one integrates the ODE $d\tilde{x}_t = v_\theta(t, \tilde{x}_t)dt$ from a source sample $\tilde{x}_0 \sim \mu$, so that $\tilde{x}_1$ is approximately distributed according to ν. In its most general form, conditional flow matching (Lipman et al., 2023) replaces the intractable FM loss equation 3 by an equivalent loss involving a conditional vector field $v_t(x \mid z)$ and conditional probability path $p_t(x \mid z)$ satisfying equation 4 between $\mu(dx \mid z)$ and $\nu(dx \mid z)$

$$\mathcal{L}_{\mathrm{CFM}}(\theta; q) \coloneqq \mathbb{E}_{t\sim\mathcal{U}(0,1),Z\sim q}\mathbb{E}_{X_t\sim p_t(\cdot|Z)}\left[\|v_\theta(t, X_t) - v_t(X_t \mid Z)\|^2\right], \tag{5}$$

where z denotes a latent variable and q the prior distribution, typically choosing $z = x_1$. Tong et al. (2024) have generalized the approach of Lipman et al. (2023) by considering arbitrary latent variables z, showing that minimizing the CFM loss equation 5 is equivalent to minimizing equation 3. Hence, the conditional flow matching method calls for two important design choices:

- The latent variable z and prior q: in this paper we focus on the popular choice of $z = (x_0, x_1)$, and we therefore require q to be a coupling, that is $q \in \Pi(\mu, \nu)$.
- The conditional vector field $v_t(x \mid z)$ and probability path $p_t(x \mid z)$: in this paper we consider the linear interpolation path $X_t = (1-t)X_0 + tX_1$, i.e. $p_t(x \mid x_0, x_1) = \delta_{(1-t)x_0+tx_1}(x)$ together with $v_t(x \mid x_0, x_1) = x_1 - x_0$, this is the implicit choice made for Rectified Flow (Liu et al., 2023).

Choosing $q = \mu \otimes \nu$ in equation 5 leads to the independent conditional flow matching algorithm (I-CFM), which is straightforward to compute but yields irregular trajectories, requiring fine discretizations at inference. On the contrary, choosing $q = \pi^\star$ (i.e., an optimal transport plan between the source and target distributions), leads to the optimal transport conditional flow matching algorithm (OT-CFM), which induces straighter trajectories for the learned model $v_\theta(t, x)$ (Tong et al., 2024). As $\pi^\star$ is *a priori* difficult to sample from, many works have explored approximations by computing short distance couplings at the batch level during training (Tong et al., 2024; Pooladian et al., 2023).

3 Weighted Conditional Flow Matching

By writing $L_\theta(t, X, Y) = \|v_\theta(t, X_t) - v_t(X_t \mid X, Y)\|^2$ where $X_t = (1-t)X + tY$, the I-CFM loss can be written as $\mathcal{L}_{\mathrm{I-CFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),(X,Y)\sim\mu\otimes\nu}\left[L_\theta(t, X, Y)\right]$. In order to enforce straightness of the learned velocity field, one would like to bias this loss towards training sample pairs (x, y) that are close to each other. OT-CFM achieves this by computing the OT plan between discrete sets of points within a batch. We propose to introduce a bias directly inside the expectation by considering modifications of the I-CFM loss function of the form

$$\mathcal{L}_w(\theta) \coloneqq \mathbb{E}_{t\sim\mathcal{U}(0,1),(X,Y)\sim\mu\otimes\nu}\left[w(X, Y)L_\theta(t, X, Y)\right], \tag{6}$$

where $w : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+^*$ denotes a positive weighting function, which should be large whenever X, Y are close and small when X, Y are far apart. This weighting can be understood in terms of a change of measure. In particular, by defining $\pi_w(dx, dy) \propto w(x, y)\mu(dx)\nu(dy)$ we have $\mathcal{L}_w(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,t),(X,Y)\sim\pi_w}[L_\theta(t, X, Y)]$. In other words, weighting the I-CFM loss yields a CFM loss with a new prior distribution $q = \pi_w$. This technique can be thought of as a form of importance sampling, where the baseline distribution $\mu \otimes \nu$ is easy to sample from, and $w(x, y)$ corresponds to the importance sampling weight.

3.1 Approximating EOT with Weighted Conditional Flow Matching

Let $\varepsilon > 0$. We propose a new loss function that can be used as a drop-in replacement within conditional flow matching and call it weighted conditional flow matching (W-CFM). Given the form of the EOT plan equation 2, and a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we consider $\mathcal{L}_w$ equation 6 with the weighting function $w_\varepsilon(x, y) := \exp(-c(x, y)/\varepsilon)$ and introduce the weighted conditional flow matching loss

$$\boxed{\mathcal{L}_{\mathrm{W-CFM}}(\theta; \varepsilon) := \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{(X,Y)\sim\mu\otimes\nu}\left[w_\varepsilon(X, Y)||v_\theta(t, X) - (Y - X)||^2\right].} \tag{7}$$

In particular, Theorem 1 implies that $\mathcal{L}_{\mathrm{W-CFM}}(\theta; \varepsilon) = Z_\varepsilon\mathcal{L}_{\mathrm{CFM}}(\theta; q_\varepsilon)$, where q_ε is the following prior

$$q_\varepsilon(dx, dy) := \frac{\mathcal{K}_\varepsilon(dx, dy)}{Z_\varepsilon} = \pi_\varepsilon(dx, dy)\frac{\exp(-\phi_\varepsilon(x) - \psi_\varepsilon(y))}{Z_\varepsilon}, \tag{8}$$

and $Z_\varepsilon := \int \exp(-c(x, y)/\varepsilon)\mu(dx)\nu(dy)$ is the normalizing constant. Thus, training a CNF model using the W-CFM loss given by equation 7 is equivalent to training a CNF using the EOT plan as the prior distribution, up to a change (a.k.a. tilt) in the marginals given by the Schrödinger potentials $\phi_\varepsilon, \psi_\varepsilon$. Hence, $\mathcal{L}_{\mathrm{W-CFM}}$ can be thought of as an approximation of the following loss function

$$\mathcal{L}_{\mathrm{EOT-CFM}}(\theta; \varepsilon) = \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{(X,Y)\sim\pi_\varepsilon}\left[||v_\theta(t, X_t) - (Y - X)||^2\right], \tag{9}$$

3.2 Marginal Tilting under W-CFM

Using $\mathcal{K}_\varepsilon$ for the prior leads to the following tilted marginals, which are obtained by integrating equation 8 with respect to y and x respectively:

$$\tilde{\mu}_\varepsilon(dx) = \frac{\exp(-\phi_\varepsilon(x))}{Z_\varepsilon^1}\mu(dx), \quad \tilde{\nu}_\varepsilon(dy) = \frac{\exp(-\psi_\varepsilon(y))}{Z_\varepsilon^2}\nu(dy), \tag{10}$$

where $Z_\varepsilon^1, Z_\varepsilon^2$ are normalizing constants. Consequently, training a CNF using the W-CFM loss induces a vector field mapping $\tilde{\mu}_\varepsilon$ to $\tilde{\nu}_\varepsilon$. We formalize this result in the following proposition.

Proposition 1 (Marginal tilting and continuity equation)**.** *Assume $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ have finite second moment. Consider the variational problem*

$$\min_v \ \mathbb{E}_{t\sim\mathcal{U}(0,1)}\,\mathbb{E}_{(X,Y)\sim\mu\otimes\nu}\big[w_\varepsilon(X, Y)\,\|v(t, X_t) - (Y - X)\|^2\big], \quad X_t = (1 - t)\,X + t\,Y. \tag{11}$$

Let ρ_t denote the law of X_t under $(X, Y) \sim q_\varepsilon$. Then, equation 11 admits a minimizer $v_\varepsilon \in L^2([0, 1] \times \mathbb{R}^d; \rho_t(dx)dt)$, which is unique in that space. Moreover (ρ, v_ε) solve the continuity equation in the weak sense

$$\partial_t\rho_t + \nabla\cdot(\rho_t\,v_\varepsilon) = 0, \qquad \rho_0 = \tilde{\mu}_\varepsilon,\ \rho_1 = \tilde{\nu}_\varepsilon.$$

In other words, under mild regularity conditions, the flow generated by v_ε pushes $\tilde{\mu}_\varepsilon$ forward onto $\tilde{\nu}_\varepsilon$. We now present a way to evaluate the marginal tilting. Using equation 8, the density ratios between tilted and original marginals are given by

$$f_\varepsilon(x) = \frac{d\tilde{\mu}_\epsilon}{d\mu}(x) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{c(x, y)}{\varepsilon}\right)\nu(dy), \ \ g_\varepsilon(y) = \frac{d\tilde{\nu}_\epsilon}{d\nu}(y) \propto \int_{\mathbb{R}^d} \exp\left(-\frac{c(x, y)}{\varepsilon}\right)\mu(dx). \tag{12}$$

These integrals can be estimated by Monte Carlo sampling. If $f_\varepsilon(x)$ is constant μ almost-everywhere, then one is guaranteed that the source marginal is preserved, i.e., that $\tilde{\mu}_\varepsilon = \mu$. Similarly, if g_ε is constant ν almost-everywhere, then $\tilde{\nu}_\varepsilon = \nu$. We now give an example which features preservation of the ν marginal.

Proposition 2. *Let $\mathbb{S}_R^{d-1} = \{z \in \mathbb{R}^d : \|z\| = R\}$. Assume $c(x,y) = \|x-y\|$, and that μ is a rotation-invariant measure, for instance $\mu = \mathcal{N}(0, \sigma^2 I)$. Then, $g_\varepsilon(y) \equiv C(R,\varepsilon)$ for all $y \in \mathbb{S}_R^{d-1}$, where $C(R,\varepsilon)$ is a nonnegative constant. In particular, if ν is supported on $\mathbb{S}_R^{d-1}$, then $\tilde{\nu}_\varepsilon = \nu$.*

Proposition 2 implies that using W-CFM with an isotropic distribution as the source, and a target distribution which is supported on a $d-1$-dimensional sphere of fixed radius will induce a coupling that does not tilt the target marginal. In particular, when using a smooth cost and an appropriate ε, we expect that the target distribution will not be tilted significantly provided that its mass is concentrated on a thin annulus, which typically happens when normalizing high-dimensional data. The proofs can be found in Appendix A.

3.2.1 On the Choice of ε

The entropy regularization constant ε controls the trade-off between geometric bias (shorter, straighter flows) and marginal distortion (changing μ, ν into $\tilde{\mu}, \tilde{\nu}$). As suggested in the previous section, equation 12 can be used to build a proxy to identify sensible values for ε. In particular, if the functions f_ε and g_ε are approximately constant over the supports of μ and ν, the reweighted marginals $\tilde{\mu}$ and $\tilde{\nu}$ remain close to the original ones. In this case, the W-CFM loss in equation 7 closely approximates the EOT-CFM loss in equation 9. To quantify this, we estimate f_ε and g_ε by Monte Carlo sampling over a small number of batches and compute their relative variance, defined as $\frac{\mathrm{Var}(f_\varepsilon(X))}{\mathbb{E}[f_\varepsilon(X)]^2}$—and analogously for g_ε. This metric measures how close the functions are to being constant and is invariant to scaling by a constant factor. This invariance ensures the metric is comparable across datasets, allowing consistent evaluation of how much the importance weights distort the marginals. Low relative variance indicates that the induced marginals are close to μ and ν, suggesting that the selected ε yields a good approximation.

As an initial heuristic for choosing ε in high-dimensional settings (e.g., images or language embeddings) with Euclidean cost $c(x,y) = \|x-y\|$, we rely on the observation that normalized high-dimensional data typically concentrates near a thin spherical shell of radius $\sqrt{d}$, where d is the data dimension (see concentration of measures in Vershynin (2018)). Consequently, the typical inter-sample distance (and hence the typical value of the L^2 cost function) is $\mathcal{O}(\sqrt{d})$. Selecting ε on this same scale ensures that the ratio inside the exponential of the weighting function in equation 7 is $\mathcal{O}(1)$[1]. In this way, the kernel varies slowly with respect to typical data variations since most pairwise distances become comparable to ε. This reasoning is similar to the heuristic commonly used in kernel methods (e.g., SVMs), where the Gaussian kernel width is set proportional to the median pairwise distance between points (Christianini et al., 2000).

Following this rationale, in all our experiments we set $\varepsilon = \kappa\sqrt{d}$, with the scalar κ tuned efficiently (within seconds and a few lines of code) using the relative variance proxy described previously. Specifically, we search over a grid of κ values spaced uniformly in log scale and select the smallest value for which the relative variance starts flattening, following an ”elbow rule” heuristic akin to the selection of the number of principal components in PCA (Jolliffe, 2002). Additionally, we experimented with various schedulers for ε (including cosine, exponential, and linear), none of which showed a significant improvement over a fixed constant ε. The exact values of ε we used are reported along with the results in Section 5.

3.3 Equivalence to OT-CFM in the Large Batch Limit

OT–CFM relies on solving a mini-batch optimal transport problem for each batch, which (i) requires batch sizes large enough to represent every mode or class—otherwise the empirical OT plan might be a poor approximation of the true OT plan—and (ii) incurs at least cubic (or quadratic for the entropic approximation) cost in the batch size. By contrast, W-CFM uses a simple and virtually free per-pair Gibbs weight $w_\varepsilon(x,y) = \exp(-c(x,y)/\varepsilon)$, avoiding any coupling step. Under mild regularity (no marginal tilt, bounded support), one can show that as the batch size goes to infinity, the batch-level EOT-CFM loss converges to a limit which is proportional to the W-CFM loss. We

[1] The same logic applies for different cost functions, e.g., when the cost function is the squared Euclidean norm, then our epsilon should be $O(d)$.

formalize this in Proposition 3 below—proof is given in Appendix A. A more detailed discussion can be found in Appendix B.

Proposition 3. *Let $\varepsilon > 0$. Suppose that μ, ν, c are such that equation 1 is finite and μ, ν have bounded support. Let $(t_n, x_n, y_n)_{n\geq 1}$ be iid samples of $\mathcal{U}(0,1) \otimes \mu \otimes \nu$. Assume that $\tilde{\mu}_\varepsilon = \mu$ and $\tilde{\nu}_\varepsilon = \nu$. Let π_ε be the optimal EOT plan between μ and ν. Let π_ε^n be the optimal EOT plan between the empirical distributions $\mathbf{x}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\mathbf{y}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Then, $\pi_\varepsilon^n \to \pi_\varepsilon$ almost surely as $n \to \infty$ in the weak sense. In particular, if $\mathcal{B}_n = \{(t_i, x_i, y_i) : 1 \leq i \leq n\}$ and $v_\theta(t, z)$ is uniformly integrable in $t \in [0,1]$, continuous in $z \in \mathbb{R}^d$, we have, for any θ*

$$\mathbb{E}\left[L_{\mathrm{EOT-CFM}}(\mathcal{B}_n, \theta; \varepsilon)\right] \to l(\theta; \varepsilon) \propto \mathcal{L}_{\mathrm{W-CFM}}(\theta; \varepsilon), \text{ as } n \to \infty,$$

where the expectation is taken over the random batch $\mathcal{B}_n$ and

$$L_{\mathrm{EOT-CFM}}(\mathcal{B}_n, \theta; \varepsilon) = \frac{1}{n}\sum_{i,j=1}^{n} \pi_\varepsilon^n(x_i, y_j)\|v_\theta(t_i, (1-t_i)x_i + t_i y_j) - (y_j - x_i)\|^2.$$

4 Related Work

Straightening Sample Paths with Optimal Transport. We review the existing literature on techniques to learn straighter sample paths for flow-based models. Including OT priors within the maximum likelihood training of CNFs has been considered by Onken et al. (2021). Zhang et al. (2025) propose to learn an acceleration field to capture a random vector field, thereby allowing sample trajectories to cross and leading to straighter trajectories. As highlighted above, our work is closely related to the minibatch optimal transport method proposed by Lipman et al. (2023) and Pooladian et al. (2023) concurrently, which leads to OT-CFM. Within OT-CFM, Klein et al. (2025) have advocated for using the Sinkhorn algorithm (hence computing EOT) to scale to larger batch sizes. Other works have explored an adaptation of OT-CFM to conditional generation through conditional optimal transport (Kerrigan et al., 2024; Cheng & Schwing, 2025). Finally, recent efforts have been made to learn straighter marginal probability paths directly using Wasserstein gradient flows and the JKO scheme (Choi et al., 2024). Compared to OT-CFM, our method has no issue scaling with the batch size, see the discussion in Section 3.3.

Other Approaches for Faster Inference. One important line of research has been to distill existing models (Luhman & Luhman, 2021; Salimans & Ho, 2022), either diffusion or flow-based, and learn the corresponding flow maps so that inference can be done in very few steps. Liu et al. (2023) propose to learn straighter trajectories via ReFlow steps, where one trains a student model using couplings generated by a base teacher model trained by flow matching. The ReFlow paradigm is still of interest and continues to be improved Kim et al. (2025). Consistency models Song et al. (2023) were developed to overcome the cost of discretizing the probability flow ODE in diffusion models, by using the self-consistency property of the underlying flow map. This approach has been translated to flow matching models by Yang et al. (2024). Finally, some great effort has been put into using efficient integrators tailored for these models (Lu et al., 2022; Liu et al., 2022; Zhang & Chen, 2023; Sabour et al., 2024; Williams et al., 2024). Compared to ReFlow, our method does not rely on synthetic data for training and does not require integrating a base neural ODE to generate the training pairs.

Finally, we acknowledge that a potential drawback of our approach is that it effectively learns a mapping between tilted versions of the source and target distributions. Although we conclude that the tilting is negligible in high-dimensional settings with a careful choice of ε in Section 3.2.1, one could benefit from being able to adjust for the tilting at inference, in order to get a genuine mapping between μ and ν. We leave this challenge for future research.

5 Experiments

We evaluate our flow matching framework across three complementary domains: 2D toy transports, unconditional image generation, and a fidelity and diversity analysis of the generated samples. In these experiments, we demonstrate our framework's performance and competitiveness in comparison to well-established baselines.

5.1 Experimental Setup

To visually probe the benefits of our weighted loss, we design similar low-dimensional transport benchmarks as in Tong et al. (2024). First, we focus on mapping a distribution concentrated on an annulus to a configurable Mixture of Gaussians (MoG). The second setup consists of recovering the moons 2D dataset from a MoG source. We compare W-CFM for different choices of ε with the cost $c(x, y) = \|x - y\|$ against both OT-CFM and I-CFM, training a two-layer ELU-MLP with 64 hidden units per layer via Adam with a learning rate of 10^{-3} for 60,000 iterations with a default batch size of 64. We evaluate sample quality, path straightness, and marginal density estimates using KDE contours. When using W-CFM, the training loss is a sample average of equation 7, rescaled by a Monte-Carlo approximation of Z_ε^{-1} computed over a single epoch as a preprocessing step.

To validate our approach in higher-dimensional settings, we evaluate on CIFAR-10, CelebA64, and ImageNet64-10—a 64×64 version of 10 ImageNet classes (Deng et al., 2009). We use a UNet backbone (Ronneberger et al., 2015) adapted to each dataset: for CIFAR-10, a smaller model with two residual blocks, 64 base channels, and 16×16 attention; for the rest of the datasets, a deeper UNet with three residual blocks, 128 base channels, a [1, 2, 2, 4] channel multiplier, and additional attention at 32×32 for ImageNet64-10, Food20, and Intel. All models are trained with Adam, a learning rate of 2×10^{-4}, cosine learning rate scheduling with 5,000 warmup steps, and EMA (decay 0.9999), for 400,000 steps using batch sizes of 128 for CIFAR-10, 64 for CelebA and ImageNet64-10, and 48 for Food20 and Intel. Our goal is not to reach state-of-the-art performance, but to compare flow matching variants under matched computational budgets and architectures.

5.2 Illustrative Comparison on Toy Datasets

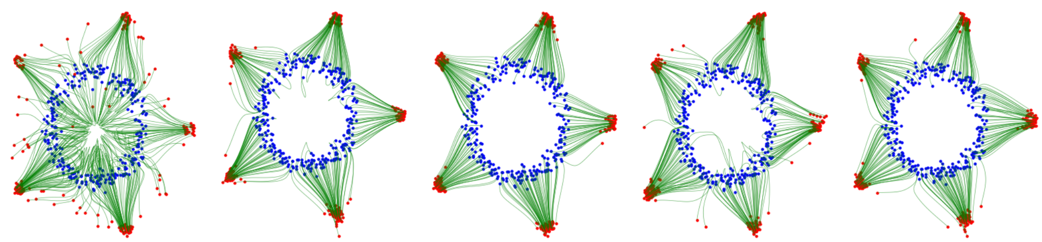

Figure 1: Sample trajectories for circular MoG $\to$ 5 Gaussians. Left to right, the models used are trained with: I-CFM, W-CFM ($\varepsilon = 0.4$), W-CFM ($\varepsilon = 0.2$), OT-CFM (batch size 16), and OT-CFM.

Sample Trajectories on Mixture of Gaussians. We consider the task of mapping a mixture of many low-variance Gaussians, whose support is concentrated on an annulus, to a target distribution consisting of a mixture of five low-variance Gaussians. We train W-CFM with $\varepsilon = 0.2$ (small ε) and $\varepsilon = 0.4$ (large ε). A specific instance where OT-CFM faces challenges is when a typical batch is not representative of the true target distribution (Kornilov et al., 2024; Klein et al., 2025). Hence, on top of training OT-CFM with the default parameters, we train OT-CFM with a smaller batch size to emulate a scenario with a high number of clusters to batch size ratio (we compensate for the smaller batch size by increasing the number of iterations for this version). We plot sample trajectories of the trained models in Figure 1, and report the performance of the models in Table 1. We use the W_2^2 distance between generated samples and the true target for overall sample quality, and compute the normalized path energy $\mathrm{NPE}(\theta) = W_2^{-2}(\mu, \nu) \left| \mathbb{E}\left[\int_0^1 ||v_\theta(t, x_t)||^2 dt\right] - W_2^2(\mu, \nu) \right|$ for straightness of the trajectories (Tong et al., 2024). Overall, W-CFM leads to better sample quality than OT-CFM with trajectories of similar straightness. We also observe a reduction in straightness for some of the paths when training OT-CFM with a smaller batch size. The W-CFM method outperforms OT-CFM and I-CFM in terms of sample quality, and it exhibits straightness on par with OT-CFM. In particular, we observe that the trajectories of W-CFM are straighter than those in OT-CFM with a small batch size. Overall, this suggests that our method can be well-suited for unconditional generation involving a target distribution with many clusters.

Marginal Tilting on Moons Generation. As highlighted above, our method might induce a tilting of the marginal distributions. We investigate this phenomenon when generating moons from a

Table 1: Comparison of CFM training algorithms' performance on 2D datasets generation on 5 random seeds. W_2^2 measures the overall quality of sample generation (lower is better), NPE measures the straightness of trajectories (lower is better), using the true optimal transport cost as a reference.

Dataset $\rightarrow$	Circular MoG $\rightarrow$ 5 Gaussians		8 Gaussians $\rightarrow$ moons	
Algorithm $\downarrow$ Metric $\rightarrow$	W_2^2 ($\downarrow$)	NPE ($\downarrow$)	W_2^2 ($\downarrow$)	NPE ($\downarrow$)
I-CFM	0.091 ± 0.071	1.703 ± 0.107	0.680 ± 0.146	1.033 ± 0.070
OT-CFM	0.029 ± 0.011	$\mathbf{0.032 \pm 0.019}$	$\mathbf{0.232 \pm 0.043}$	0.125 ± 0.011
OT-CFM ($B = 16$)	0.041 ± 0.014	0.188 ± 0.041	0.564 ± 0.125	$\mathbf{0.067 \pm 0.024}$
W-CFM (small ε)	$\mathbf{0.018 \pm 0.008}$	0.086 ± 0.021	1.823 ± 0.166	0.289 ± 0.008
W-CFM (large ε)	0.029 ± 0.011	0.097 ± 0.024	0.843 ± 0.321	0.463 ± 0.061

mixture of 8 Gaussians, and training W-CFM with $\varepsilon = 2$ (small ε) and $\varepsilon = 10$ (large ε). As seen in Figure 2, using W-CFM leads to straighter paths compared to I-CFM as measured by NPE, even for a large value of ε; however, it also causes some bias in the marginals as ε decreases, as shown in Figure 3, which leads to poor sample generation quality, see Table 1. This can be mitigated by choosing a larger value of ε, while maintaining some straightening effect. We present results for $\varepsilon \in \{2, 4, 6, 8, 10\}$ for further validation of this tradeoff in Appendix C. Finally, although OT-CFM with a small batch size seems to outperform the other models in NPE, we want to highlight that this metric measures straightness conditioned on the fact that the model does generate good samples of the target distribution, which is arguable for this instance of the model.

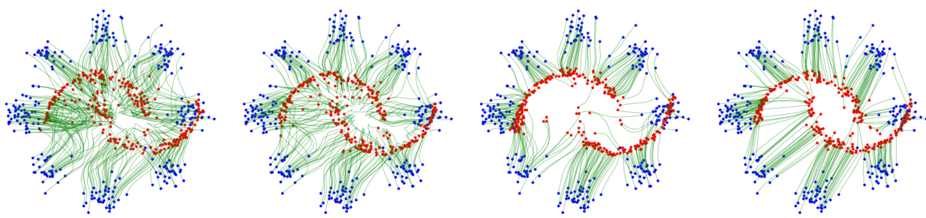

Figure 2: Sample trajectories for moons generation. Source samples are in blue, generated samples are in red. From left to right, we use: I-CFM, W-CFM ($\varepsilon = 10$), W-CFM ($\varepsilon = 2$), and OT-CFM.

Figure 3: Contour plots of learned density for moons (using 50,000 generated samples). The leftmost plot corresponds to the true target distribution. Then, from left to right, the models used are trained with: I-CFM, W-CFM ($\varepsilon = 10$), W-CFM ($\varepsilon = 2$), and OT-CFM. We observe that choosing a small value of ε for W-CFM leads to a "disentanglement" of the two generated moons.

5.3 Comparison on Image Datasets

To evaluate the generation capabilities of W-CFM beyond low-dimensional toy examples, we conducted unconditional image generation experiments across five different benchmarks: CIFAR-10, CelebA64, ImageNet64-10, Intel Image Classification, and Food20 (a subset of Food101 (Bossard et al., 2014)). Table 2 summarizes the Fréchet Inception Distance (FID, Heusel et al. (2017)) scores obtained by our proposed W-CFM method compared to I-CFM and OT-CFM with an adaptive solver. W-CFM consistently matches or outperforms the baselines, achieving the best FID scores on CIFAR-10, ImageNet64-10, Intel, and Food20, while remaining competitive on CelebA64. The slightly better performance of OT-CFM on CelebA64 is likely explained by the relatively unimodal nature of the dataset (Zhang, 2023), which alleviates the mode representation issues intrinsic to minibatch OT methods (as discussed in Section 3.3). Conversely, the multimodal structure of datasets like CIFAR-10, ImageNet64-10, Intel, and Food20 highlights the advantage of W-CFM. Batch sizes were 128 for CIFAR-10, 64 for ImageNet64-10 and CelebA64, and 48 for the remaining datasets. Applying the proxy described in Section 3.2.1 for selecting ε, we used $\varepsilon = 5$ for CIFAR-10, $\varepsilon = 13$ for CelebA64, $\varepsilon = 14$ for ImageNet64-10, $\varepsilon = 14$ for Intel, and $\varepsilon = 15$ for Food20.

Table 2: FID ↓ (lower is better) across datasets for different flow matching models with Dopri5.

Model	CIFAR-10	CelebA64	ImageNet64-10	Intel	Food20
I-CFM	7.44	21.99	13.86	27.54	8.15
OT-CFM	7.60	**20.93**	14.39	25.63	8.23
W-CFM	**7.33**	21.96	**13.56**	**25.22**	**7.93**

We further assessed model efficiency by comparing FID scores at various numbers of neural function evaluations (NFEs) during Euler integration. Table 3 shows that W-CFM consistently achieves lower or comparable FID scores at fewer NFEs compared to both I-CFM and OT-CFM. On CelebA64, OT-CFM outperforms both models, which again reflects the dataset's unimodal structure that mitigates OT-CFM's batch limitations. Appendix C shows example samples for each dataset.

Table 3: FID ↓ at varying number of neural function evaluations (NFE) using Euler discretization. Each column reports FID at 50, 100, and 120 NFE.

	FID @ NFE								
Dataset	I-CFM			OT-CFM			W-CFM		
	50	100	120	50	100	120	50	100	120
CIFAR-10	10.87	9.76	8.68	11.03	9.89	8.53	**10.53**	**9.28**	**8.08**
CelebA64	29.49	25.26	24.50	**27.76**	**23.86**	**22.93**	29.32	25.22	24.37
ImageNet64-10	**14.94**	**13.91**	13.86	15.67	14.78	14.68	15.82	14.17	**13.71**
Intel	26.72	26.40	26.20	25.45	25.98	24.26	**25.01**	**24.47**	**24.08**
Food20	10.10	8.98	8.85	10.16	9.17	8.95	**10.01**	**8.97**	**8.57**

5.4 Fidelity and Diversity of the Generated Samples

A natural concern with the importance weighting in W-CFM is that it could skew the learned flow toward easier-to-transport pairs, potentially under-representing low-probability modes or degrading sample quality in certain regions of the data manifold. To test this, we evaluate sample fidelity and diversity using the precision-recall-density-coverage (PRDC) suite of metrics (Naeem et al., 2020) on 10,000 generated samples and 5,000 held-out real images per dataset. Precision measures the fraction of generated samples near the real data manifold, recall assesses coverage of the real distribution, and density and coverage estimate support concentration and breadth, respectively. F1 summarizes the trade-off via the harmonic mean of precision and recall.

Table 4: Sample quality and diversity metrics on **ImageNet64-10**.

Model	Precision (↑)	Recall (↑)	Density (↑)	Coverage (↑)	F1 (↑)
I-CFM	**0.75**	**0.69**	0.91	0.94	**0.72**
OT-CFM	0.74	0.67	0.91	**0.96**	0.71
W-CFM	**0.75**	0.68	**0.94**	0.92	**0.72**

Table 4 reports the results on ImageNet64-10, showing that W-CFM maintains comparable or better recall, F1, and density than both I-CFM and OT-CFM, without sacrificing coverage or precision. We observe the same qualitative trends across the remaining datasets, with all three methods exhibiting nearly identical PRDC metrics. For completeness, we also include the corresponding tables for CIFAR-10 and CelebA64 in Appendix C. These results confirm that W-CFM does not impair diversity or mode coverage, even when trained with independent pairwise weighting.

6 Conclusion

In this work, we introduced weighted conditional flow matching (W-CFM), a novel method that leverages insights from entropic optimal transport (EOT) to efficiently improve path straightness and sample quality in continuous normalizing flows. By weighting each training pair using a Gibbs kernel, our approach approximates the EOT plan without incurring the computational cost of OT plan computations and without being limited by the batch size when solving the optimal transport problem. We derived theoretical conditions under which our approximation preserves the original marginals and proved equivalence with OT-CFM in the large-batch limit, given that the marginals remain unchanged. Empirical evaluations across low-dimensional toy problems, unconditional image generation tasks, and fidelity-diversity analyses confirm that W-CFM achieves performance competitive with OT-CFM while maintaining the computational efficiency of standard CFM.

REFERENCES

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

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, 2017.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

Ho Kei Cheng and Alexander Schwing. The curse of conditions: Analyzing and improving optimal transport for conditional flow-based generation. *arXiv preprint arXiv:2503.10636*, 2025.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. *arXiv preprint arXiv:2402.05443*, 2024.

Nello Christianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press Cambridge, 2000.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018.

Promit Ghosal, Marcel Nutz, and Espen Bernton. Stability of entropic optimal transport and schrödinger bridges. *Journal of Functional Analysis*, 2022.

Will Grathwohl, Ricky TQ Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Yiwei Guo, Chenpeng Du, Ziyang Ma, Xie Chen, and Kai Yu. Voiceflow: Efficient text-to-speech with rectified flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Ross Irwin, Alessandro Tibo, Jon Paul Janet, and Simon Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport. In *International Conference on Machine Learning*, 2024.

I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Dynamic conditional optimal transport through simulation-free flows. In *Advances in Neural Information Processing Systems*, 2024.

Beomsu Kim, Yu-Guan Hsieh, Michal Klein, Marco Cuturi, Jong Chul Ye, Bahjat Kawar, and James Thornton. Simple reflow: Improved techniques for fast flow models. In *International Conference on Learning Representations*, 2025.

Michal Klein, Alireza Mousavi-Hosseini, Stephen Zhang, and Marco Cuturi. On fitting flow models with large sinkhorn couplings. *arXiv preprint arXiv:2506.05526*, 2025.

Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Aleksandr Korotin. Optimal flow matching: Learning straight trajectories in just one step. In *Advances in Neural Information Processing Systems*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022.

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

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems*, 2022.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, 2020.

Marcel Nutz. Introduction to entropic optimal transport. *Lecture notes, Columbia University*, 2021.

Derek Onken, Samy Wu Fung, Xingjian Li, and Lars Ruthotto. Ot-flow: Fast and accurate continuous normalizing flows via optimal transport. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Now Publishers, Inc., 2019.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, 2023.

Martin Rohbeck, Charlotte Bunne, Edward De Brouwer, Jan-Christian Huetter, Anne Biton, Kelvin Y. Chen, Aviv Regev, and Romain Lopez. Modeling complex system dynamics with flow matching across time and conditions. In *International Conference on Learning Representations*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015.

Amirmojtaba Sabour, Sanja Fidler, and Karsten Kreis. Align your steps: Optimizing sampling schedules in diffusion models. In *International Conference on Machine Learning*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Veeravalli S Varadarajan. On the convergence of sample probability distributions. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 1958.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Christopher Williams, Andrew Campbell, Arnaud Doucet, and Saifuddin Syed. Score-optimal diffusion schedules. In *Advances in Neural Information Processing Systems*, 2024.

Ling Yang, Zixiang Zhang, Zhilong Zhang, Xingchao Liu, Minkai Xu, Wentao Zhang, Chenlin Meng, Stefano Ermon, and Bin Cui. Consistency flow matching: Defining straight flows with velocity consistency. *arXiv preprint arXiv:2407.02398*, 2024.

Na Zhang. A study on the impact of face image quality on face recognition in the wild. *arXiv preprint arXiv:2307.02679*, 2023.

Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. In *International Conference on Learning Representations*, 2023.

Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. In *Advances in Neural Information Processing Systems*, 2024.

Yichi Zhang, Yici Yan, Alex Schwing, and Zhizhen Zhao. Towards hierarchical rectified flow. In *International Conference on Learning Representations*, 2025.

A Proofs of Theoretical Results

A.1 Proof of Proposition 1

Recall the prior q_ε induced by the Gibbs kernel $\mathcal{K}_\varepsilon(dx, dy) = \exp(-c(x,y)/\varepsilon)\mu(dx)\nu(dy)$ defined in equation 8. Recall that ρ_t denotes the distribution of $X_t = (1-t)X + tY$ under $(X,Y) \sim q_\varepsilon$. For any $v \in L^2([0,1] \times \mathbb{R}^d; \rho_t(dx)dt)$, we have

$$\begin{aligned}
&\mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{(X,Y)\sim\mu\otimes\nu}\left[w_\varepsilon(X,Y)||v(t,X_t) - (Y-X)||^2\right] \\
&= \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{(X,Y)\sim q_\varepsilon}\left[\frac{d(\mu\otimes\nu)}{dq_\varepsilon}(X,Y)\exp(-c(X,Y)/\varepsilon)||v(t,X_t) - (Y-X)||^2\right] \\
&= Z_\varepsilon \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{(X,Y)\sim q_\varepsilon}\left[||v(t,X_t) - (Y-X)||^2\right],
\end{aligned}$$

where Z_ε denotes the normalizing constant $Z_\varepsilon := \mathbb{E}_{(X,Y)\sim\mu\otimes\nu}[w_\varepsilon(X,Y)] > 0$. Hence the variational problem given by equation 11 is equivalent to

$$\min_v \mathbb{E}_{t\sim\mathcal{U}(0,1)}\mathbb{E}_{(X,Y)\sim q_\varepsilon}\left[||v(t,X_t) - (Y-X)||^2\right]. \tag{13}$$

By the L^2-projection property of conditional expectations, the variational problem of equation 13 is solved by the function $v_\varepsilon : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$v_\varepsilon(t,z) = \mathbb{E}_{(X,Y)\sim q_\varepsilon}[Y - X \mid X_t = z]. \tag{14}$$

Note that this definition is unique in $L^2([0,1]\times\mathbb{R}^d; \rho_t(dx)dt)$. We now check that v_ε generates a valid probability path between $\tilde{\mu}_\varepsilon$ and $\tilde{\nu}_\varepsilon$, i.e., that (ρ, v_ε) satisfy the continuity equation equation 4 in the weak sense. Clearly, $v_\varepsilon(t,\cdot) \in L^1(\mathbb{R}^d, \rho_t)$ and $\int_0^1\int_{\mathbb{R}^d}|v_\varepsilon(t,x)|p(t,x)dxdt < \infty$. By Proposition 4.2 in Santambrogio (2015), it is enough to check that the continuity equation is satisfied in the sense of distributions. Let $\phi \in C_c^1((0,1)\times\mathbb{R}^d)$, then

$$\begin{aligned}
&\int_0^1\int_{\mathbb{R}^d}\partial_t\phi(t,x)\rho_t(dx)dt + \int_0^1\int_{\mathbb{R}^d}\nabla\phi(t,x)\cdot v_\varepsilon(t,x)\rho_t(dx)dt \\
&= \int_0^t \mathbb{E}[\partial_t\phi(t,X_t) + \nabla\phi(t,X_t)\cdot(Y-X)]dt = \mathbb{E}[\phi(1,Y) - \phi(0,X)] = 0.
\end{aligned}$$

□

A.2 Proof of Proposition 2

Let $y_2, y_2 \in \mathbb{S}_R^{d-1}$. We can get hold of $\varphi \in O(d)$ (i.e. a distance-preserving transformation in $\mathbb{R}^d$) such that $\varphi(y_1) = y_2$. Since $||\varphi(y_1) - \varphi(x)|| = ||y_1 - x||$ for all $x \in \mathbb{R}^d$, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} \exp\left(-\frac{\|y_2 - x\|}{\varepsilon}\right))\mu(dx) &= \int_{\mathbb{R}^d} \exp\left(-\frac{\|\varphi(y_1) - x\|}{\varepsilon}\right))\mu(dx) \\
&= \int_{\varphi^{-1}(\mathbb{R}^d)} \exp\left(-\frac{\|\varphi(y_1) - \varphi(x)\|}{\varepsilon}\right)(\varphi_{\#}^{-1}\mu)(dx) \\
&= \int_{\varphi^{-1}(\mathbb{R}^d)} \exp\left(-\frac{\|y_1 - x\|}{\varepsilon}\right)(\varphi_{\#}^{-1}\mu)(dx) \\
&= \int_{\mathbb{R}^d} \exp\left(-\frac{\|y_1 - x\|}{\varepsilon}\right)\mu(dx),
\end{aligned}$$

which directly implies that $g_\varepsilon(y_2) = g_\varepsilon(y_1)$ □

A.3 Proof of Proposition 3

Let $\varepsilon > 0$. Recall that $(t_n, x_n, y_n)_{n\geq 1}$ are iid samples of $\mathcal{U}(0,1) \otimes \mu \otimes \nu$, and that we assume $\tilde{\mu}_\varepsilon = \mu$ and $\tilde{\nu}_\varepsilon = \nu$. Let π_ε be the optimal EOT plan between μ and ν. Let π_ε^n be the optimal EOT plan betwen $\mathbf{x}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\mathbf{y}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. In this proof, the convergence of probability measures is understood in the weak sense.

First, the almost-sure convergences $\mathbf{x}_n \to \mu$ and $\mathbf{y}_n \to \nu$ come from a classical result in probability theory on the convergence of empirical distributions to the true distribution, see Varadarajan (1958).

Since the minimization problem of equation 1 is non-trivial, an application of Theorem 1.4 in Ghosal et al. (2022) shows that the empirical EOT plan satisfies $\pi_\varepsilon^n \to \pi_\varepsilon$ almost surely. Now, for any $n \geq 1$, we have

$$\begin{aligned}
\mathbb{E}\left[L_{\mathrm{EOT-CFM}}(\mathcal{B}_n, \theta; \varepsilon)\right] &= \mathbb{E}\left[\int_{\mathbb{R}^d\times\mathbb{R}^d}\int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt\pi_\varepsilon^n(dx,dy)\right] \\
&= \mathbb{E}\left[\int_{\mathrm{s}(\mu)\times\mathrm{s}(\nu)}\int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt\pi_\varepsilon^n(dx,dy)\right],
\end{aligned}$$

where $\mathrm{s}(\mu), \mathrm{s}(\nu)$ denote the support of μ and ν respectively, which are assumed to be bounded. Since $(x,y) \mapsto \int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt$ is continuous and bounded on $\mathrm{s}(\mu)\times\mathrm{s}(\nu)$ by our assumption on v_θ, we have

$$\begin{aligned}
&\int_{\mathrm{s}(\mu)\times\mathrm{s}(\nu)}\left(\int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt\right)\pi_\varepsilon^n(dx,dy) \\
&\to \int_{\mathrm{s}(\mu)\times\mathrm{s}(\nu)}\left(\int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt\right)\pi_\varepsilon(dx,dy)
\end{aligned}$$

almost surely as $n \to \infty$. Now, the dominated convergence theorem ensures that this convergence also holds in expectation, i.e.

$$\mathbb{E}\left[L_{\mathrm{EOT-CFM}}(\mathcal{B}_n, \theta; \varepsilon)\right] \to \int_{\mathrm{s}(\mu)\times\mathrm{s}(\nu)}\left(\int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt\right)\pi_\varepsilon(dx,dy).$$

Finally, we want to prove that this integral is proportional to $\mathcal{L}_{\mathrm{W-CFM}}(\theta;\varepsilon)$. Since we assume no tilting of the marginals, i.e. $q_\varepsilon = \pi_\varepsilon$, we have

$$\begin{aligned}
\mathcal{L}_{\mathrm{W-CFM}}(\theta;\varepsilon) &= Z_\varepsilon \mathbb{E}_{t\sim\mathcal{U}(0,1),(X,Y)\sim\pi_\varepsilon}\left[||v_\theta(t,X_t) - (Y-X)||^2\right] \\
&= Z_\varepsilon \int_{\mathrm{s}(\mu)\times\mathrm{s}(\nu)}\left(\int_0^1 \|v_\theta(t,(1-t)x+ty) - (y-x)\|^2\, dt\right)\pi_\varepsilon(dx,dy).
\end{aligned}$$

by using the same change of measure argument as in the proof of Proposition 1. □

B Details on the Equivalence to OT-CFM in the Large Batch Limit

We recall the mini-batch optimal transport technique that is central in the OT-CFM algorithm of Tong et al. (2024). Given a batch of i.i.d. samples $\mathcal{B} = \{(t_i, x_i, y_i) : i = 1, \ldots, B\}$, where t_i are i.i.d. according to $\mathcal{U}(0,1)$, x_i are i.i.d. according to μ, y_i are i.i.d. according to ν, and t_i, x_i, y_i are drawn independently, one can compute the optimal transport plan between the two corresponding discrete distribution, i.e. one computes

$$\pi_{\mathcal{B}} \in \arg\min_{\pi \in \Pi_{\mathcal{B}}} \sum_{i=1}^{B} \sum_{j=1}^{B} c(x_i, y_j) \pi(x_i, y_j), \tag{15}$$

where $\Pi_{\mathcal{B}}$ is the set of couplings between the empirical measures

$$\mathbf{x}_{\mathcal{B}} = \frac{1}{B} \sum_{i=1}^{B} \delta_{x_i}, \quad \mathbf{y}_{\mathcal{B}} = \frac{1}{B} \sum_{i=1}^{B} \delta_{y_i}.$$

In particular, any $\pi \in \Pi_{\mathcal{B}}$ must satisfy $\sum_j \pi(x_i, y_j) = \sum_i \pi(x_i, y_j) = \frac{1}{B}$. Then, given an optimal $\pi_{\mathcal{B}}$, one computes the following

$$L_{\mathrm{OT-CFM}}(\mathcal{B}, \theta) = \frac{1}{B} \sum_{i=1}^{B} (v_\theta(t_i, (1-t_i)x_i + t_i y_{\sigma(i)}) - (y_{\sigma(i)} - x_i))^2, \tag{16}$$

where σ is a permutation corresponding to a Monge map for the problem equation 15, i.e., for some $T : \{x_i : i = 1, \ldots, B\} \to \{y_i : i = 1, \ldots, B\}$ such that $T(x_i) = y_{\sigma(i)}$ and $\pi_{\mathcal{B}} \coloneqq (\mathrm{Id}, T)_\# \mathbf{x}_{\mathcal{B}}$ is a solution to equation 15 (Peyré & Cuturi, 2019). This sample loss is used as an approximation of the following OT-CFM loss

$$\mathcal{L}_{\mathrm{OT-CFM}}(\theta) \coloneqq \mathbb{E}_{t \sim \mathcal{U}(0,1)} \mathbb{E}_{(X,Y) \sim \pi^\star} \left[||v_\theta(t, X_t) - (Y - X)||^2 \right], \tag{17}$$

where $\pi^\star$ solves the unregularized optimal transport problem, that is equation 1 with $\varepsilon = 0$.

The sample OT-CFM loss in equation 16 is a low bias approximation of equation 17 only when the batch size is large enough. The actual samples for which we compute equation 16 are not exactly distributed according to a genuine OT plan between μ and ν, since the OT plan $\pi^\star$ and the product measures $\mu \otimes \nu$ might be mutually singular. Additionally, computing the exact batch OT plan becomes prohibitively expensive as the batch size grows. A solution is to compute an approximate OT plan, by using the Sinkhorn algorithm (Cuturi, 2013), which is an efficient way of computing the entropic OT plan between two discrete sets of measures. In that case, as the batch size increases, the sample OT-CFM loss equation 16 approximates $\mathcal{L}_{\mathrm{EOT-CFM}}$ given by equation 9. Nevertheless, approximating the OT at the batch level is particularly challenging in datasets with multiple modes, as it becomes unrealistic to faithfully approximate the global OT if not all modes are adequately represented within each (or the average) batch. Consequently, the batch size must scale with the number of models or classes present in the dataset.

Our method does not have these scaling issues with the batch size, since it only involves computing a simple weighting factor $w_\varepsilon(x_i, y_i) = \exp(-c(x_i, y_i)/\varepsilon)$ for every training sample pair (x_i, y_i) in a batch. In other words, if one assumes that the weight does not tilt the marginals, the weighted CFM method corresponds to a large batch limit of OT-CFM (where batch EOT is used).

C Additional Results

Table 5: Comparison of CFM training algorithms' performance on **8 Gaussians** $\rightarrow$ **moons** on 5 random seeds. W_2^2 measures the overall quality of sample generation (lower is better), NPE measures the straightness of trajectories (lower is better), using the true optimal transport cost as a reference. We emphasize on the tradeoff incurred by the choice of ε.

Model	W_2^2 (↓)	**NPE** (↓)
I-CFM	0.680 ± 0.146	1.033 ± 0.070
OT-CFM	$\mathbf{0.232} \pm \mathbf{0.043}$	0.125 ± 0.011
OT-CFM ($B = 16$)	0.564 ± 0.125	0.067 ± 0.024
W-CFM ($\varepsilon = 2$)	1.823 ± 0.166	0.289 ± 0.008
W-CFM ($\varepsilon = 4$)	1.476 ± 0.167	$\mathbf{0.033} \pm \mathbf{0.023}$
W-CFM ($\varepsilon = 6$)	0.960 ± 0.186	0.220 ± 0.050
W-CFM ($\varepsilon = 8$)	0.888 ± 0.217	0.365 ± 0.076
W-CFM ($\varepsilon = 10$)	0.843 ± 0.321	0.463 ± 0.061

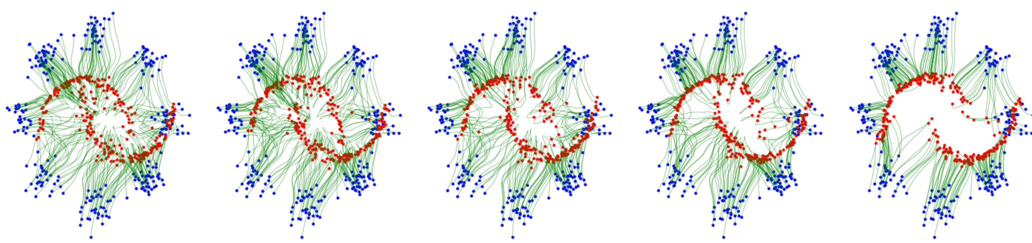

Figure 4: Sample trajectories on **8 Gaussians** $\rightarrow$ **moons** with W-CFM. From left to right, the models used are trained with the following values of ε: 10,8,6,4,2.

Table 6: Sample quality and diversity metrics on **CIFAR-10**.

Model	**Precision** (↑)	**Recall** (↑)	**Density** (↑)	**Coverage** (↑)	**F1** (↑)
I-CFM	**0.83**	0.75	0.98	0.91	**0.78**
OT-CFM	0.80	0.75	**1.00**	**0.92**	0.77
W-CFM	0.81	**0.76**	0.94	0.91	**0.78**

Table 7: Sample quality and diversity metrics on **CelebA64**.

Model	**Precision** (↑)	**Recall** (↑)	**Density** (↑)	**Coverage** (↑)	**F1** (↑)
I-CFM	**0.86**	**0.66**	**1.26**	**0.98**	**0.74**
OT-CFM	0.84	0.65	1.23	0.96	0.73
W-CFM	0.83	**0.66**	1.19	**0.98**	**0.74**

Figure 5: Contour plots of learned target density for **8 Gaussians** $\rightarrow$ **moons**. The leftmost plot corresponds to the true target distribution. Then, from left to right, the models used are trained with the following values of ε: 10,8,6,4,2.

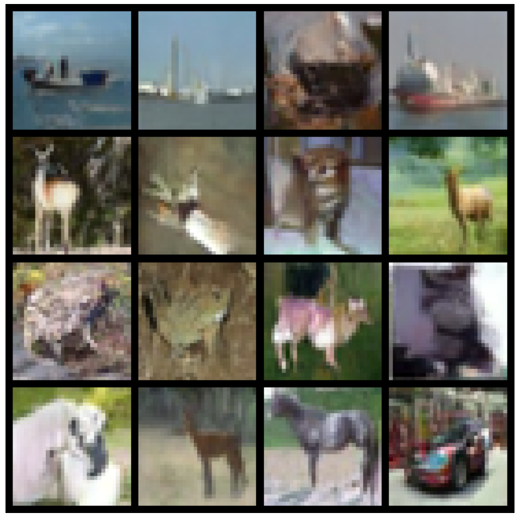

Figure 6: Generated samples from W-CFM trained on CIFAR-10.

Figure 7: Generated samples from W-CFM trained on CelebA64.

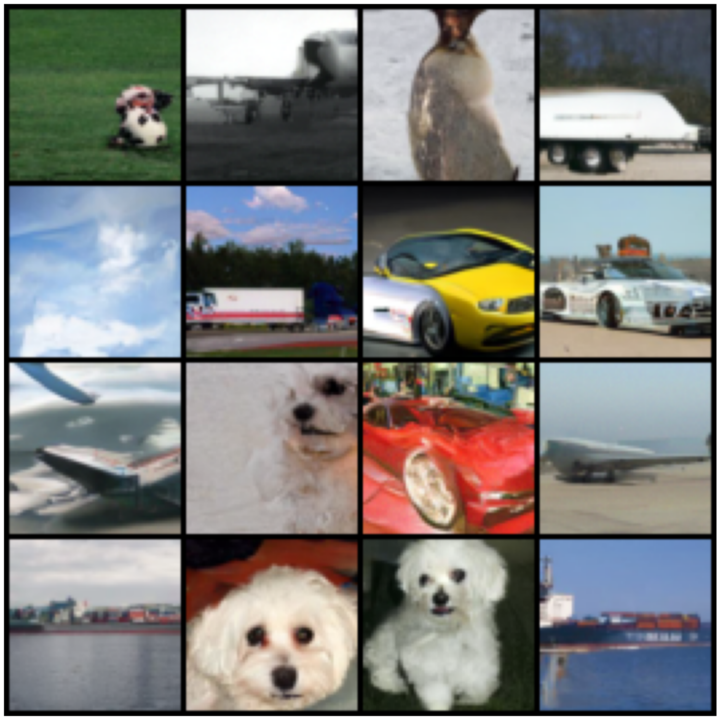

Figure 8: Generated samples from W-CFM trained on ImageNet-10.

Figure 9: Generated samples from W-CFM trained on Food-101.

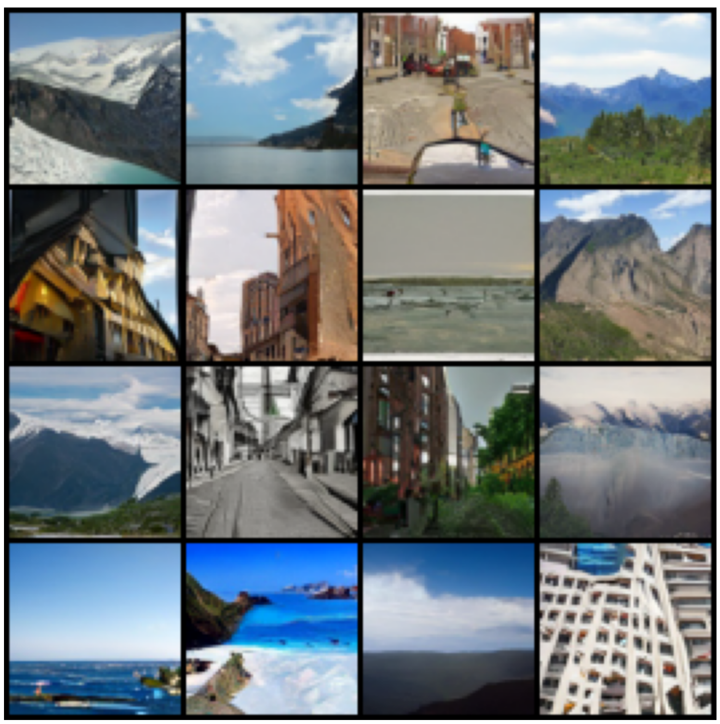

Figure 10: Generated samples from W-CFM trained on Intel Image Classification.