Weighted Conditional Flow Matching

Anonymous Author(s)
Affiliation
Address
email

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 the 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 sample quality, fidelity, and diversity comparable or superior to alternative baselines while maintaining the computational efficiency of vanilla CFM.

1 Introduction

Generative modeling aims to learn a transformation from a simple prior to a complex data distribution. Continuous normalizing flows (CNFs) achieve this via ODE-driven vector fields with exact likelihoods, but likelihood maximization is often unstable and does not scale well [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 training CNFs as regression on endpoint displacements, yielding near-optimal transport when the prior is Gaussian. However, independent pairings can lead to suboptimal paths. Conditional flow matching (CFM) [Lipman et al., 2023, Tong et al., 2024] generalizes FM by conditioning on arbitrary couplings, enabling simulation-free CNF training from any source distribution and supporting applications in molecule design, sequence modeling, and speech synthesis [Irwin et al., 2024, Geffner et al., 2025, Stark et al., 2024, Zhang et al., 2024, Rohbeck et al., 2025, Guo et al., 2024]. A refinement, minibatch optimal transport CFM (OT-CFM) [Pooladian et al., 2023, Tong et al., 2024], couples pairs using an OT plan within each batch, producing straighter trajectories with improved sample quality, but at cubic (or quadratic under entropic regularization) cost per batch and with impractical requirements on balanced class representation 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.

Submitted to Structured Probabilistic Inference & Generative Modeling @ NeurIPS 2025. Do not distribute.

2 Background: Entropic Optimal Transport

We recall only the results relevant to our work (see Nutz [2021] for details). Assume we want to sample from $\nu \in \mathcal{P}(\mathbb{R}^d)$ given samples from $\mu \in \mathcal{P}(\mathbb{R}^d)$. The Kullback–Leibler divergence is $D_{\mathrm{KL}}(\mu\|\nu) = \int \log \frac{d\mu}{d\nu}(x)\, d\mu(x)$ if $\mu \ll \nu$ and $+\infty$ otherwise. Let $\Pi(\mu, \nu)$ be the set of couplings between μ and ν. The entropic optimal transport (EOT) problem with parameter $\varepsilon > 0$ is

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

where $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is typically $c(x, y) = \|x - y\|$. For $\varepsilon = 0$ this reduces to the Monge–Kantorovich problem. EOT is widely used since it approximates OT while being tractable via Sinkhorn iterations [Cuturi, 2013, Altschuler et al., 2017]. It can also be written as

$$\min_{\pi \in \Pi(\mu,\nu)} D_{\mathrm{KL}}(\pi \| \mathcal{K}_\varepsilon), \qquad \mathcal{K}_\varepsilon(dx, dy) = e^{-c(x,y)/\varepsilon}\, \mu(dx)\nu(dy), \tag{2}$$

which has a unique minimizer π_ε.

Theorem 1 (Theorem 4.2 in Nutz [2021])**.** *If $c(x, y) < \infty$ $\mu \otimes \nu$-a.s., there exist measurable Schrödinger potentials $\phi_\varepsilon, \psi_\varepsilon : \mathbb{R}^d \to \mathbb{R}$ such that*

$$\pi_\varepsilon(dx, dy) = \exp\Big(\phi_\varepsilon(x) + \psi_\varepsilon(y) - \tfrac{c(x,y)}{\varepsilon}\Big)\, \mu(dx)\nu(dy). \tag{3}$$

Equivalently, $\pi_\varepsilon(dx, dy) = f_\varepsilon(x) g_\varepsilon(y) \mathcal{K}_\varepsilon(dx, dy)$ with $f_\varepsilon = \exp(\phi_\varepsilon)$ and $g_\varepsilon = \exp(\psi_\varepsilon)$. Thus the Gibbs kernel encodes the dependence, while $f_\varepsilon, g_\varepsilon$ adjust the marginals.

3 Weighted Conditional Flow Matching

Let $L_\theta(t, X, Y) \coloneqq \|v_\theta(t, (1-t)X + tY) - (Y - X)\|^2$. The I-CFM loss with a linearly interpolating conditional path is

$$\mathcal{L}_{\text{I-CFM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),(X,Y)\sim\mu\otimes\nu}[L_\theta(t, X, Y)]. \tag{4}$$

To bias training toward nearby pairs, we introduce

$$\mathcal{L}_w(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,1),(X,Y)\sim\mu\otimes\nu}[w(X, Y)\, L_\theta(t, X, Y)], \tag{5}$$

which is equivalent to (4) with the independent coupling replaced by $\pi_w(dx, dy) \propto w(x, y)\mu(dx)\nu(dy)$. Choosing $w_\varepsilon(x, y) = \exp(-c(x, y)/\varepsilon)$ with cost c and $\varepsilon > 0$ yields

$$\boxed{\mathcal{L}_{\text{W-CFM}}(\theta; \varepsilon) = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \mathbb{E}_{(X,Y)\sim\mu\otimes\nu}[w_\varepsilon(X, Y)\, \|v_\theta(t, X) - (Y - X)\|^2].} \tag{6}$$

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

$$q_\varepsilon(dx, dy) \coloneqq \frac{\mathcal{K}_\varepsilon(dx, dy)}{Z_\varepsilon} = \pi_\varepsilon(dx, dy) \frac{\exp(-\phi_\varepsilon(x) - \psi_\varepsilon(y))}{Z_\varepsilon}, \tag{7}$$

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 (6) 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

$$\mathcal{L}_{\mathrm{EOT-CFM}}(\theta; \varepsilon) = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \mathbb{E}_{(X,Y)\sim\pi_\varepsilon} \left[||v_\theta(t, X_t) - (Y - X)||^2 \right], \tag{8}$$

with the approximation quality depending on the Schrödinger potentials $\phi_\varepsilon, \psi_\varepsilon$.

3.1 Marginal Tilting under W-CFM

Using $\mathcal{K}_\varepsilon$ for the prior leads to the following tilted marginals, which are obtained by integrating (7) with respect to y and x respectively:

$$\tilde{\mu}_\varepsilon(dx) = \frac{\exp(-\phi_\varepsilon(x))}{Z^1_\varepsilon} \mu(dx), \quad \tilde{\nu}_\varepsilon(dy) = \frac{\exp(-\psi_\varepsilon(y))}{Z^2_\varepsilon} \nu(dy), \tag{9}$$

where $Z^1_\varepsilon, Z^2_\varepsilon$ 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*

$$\min_{v} \ \mathbb{E}_{t\sim\mathcal{U}(0,1)} \mathbb{E}_{(X,Y)\sim\mu\otimes\nu}\big[w_\varepsilon(X,Y)\, \|v(t,X_t) - (Y-X)\|^2\big], \quad X_t = (1-t)\,X + t\,Y. \quad (10)$$

Let ρ_t denote the law of X_t under $(X,Y) \sim q_\varepsilon$. Then, (10) *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*

$$\partial_t \rho_t + \nabla\cdot(\rho_t\, v_\varepsilon) = 0, \qquad \rho_0 = \tilde{\mu}_\varepsilon,\ \rho_1 = \tilde{\nu}_\varepsilon. \quad (11)$$

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 (7), the density ratios between tilted and original marginals are given by

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

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 $\tilde{\nu}_\varepsilon = \nu$. Such a situation arises, for instance, when μ and ν are isotropic distributions.

3.1.1 On the Choice of ε

The entropy regularization constant ε trades off geometric bias (straighter flows) against marginal distortion. As shown in (12), if the reweighting functions $f_\varepsilon, g_\varepsilon$ are nearly constant on the supports of μ, ν, then $\tilde{\mu}_\varepsilon, \tilde{\nu}_\varepsilon$ remain close to the μ, ν and the W-CFM loss (6) approximates the EOT-CFM loss (8). We assess this by Monte Carlo estimates of the relative variance $\mathrm{Var}(f_\varepsilon(X))/\mathbb{E}[f_\varepsilon(X)]^2$ (and analogously for g_ε), which is scale-invariant and comparable across datasets: low values indicate minimal marginal distortion whereas large values signal mismatch.

For normalized high-dimensional data with Euclidean cost, typical pairwise distances scale as $\mathcal{O}(\sqrt{d})$ due to concentration of measure on a sphere of radius $\sqrt{d}$ [Vershynin, 2018]. Setting ε on this scale keeps the Gibbs weights well-conditioned, analogous to kernel width selection in SVMs [Christianini et al., 2000]. Accordingly, we parameterize $\varepsilon = \kappa\sqrt{d}$ and tune κ using the relative variance proxy: a log-scale grid search selects the smallest κ where variance flattening occurs (an "elbow rule" akin to PCA [Jolliffe, 2002]). Schedulers for ε (cosine, exponential, linear) showed no clear benefit, so we use a fixed ε, reported per dataset in Section 4.

3.2 Equivalence to OT-CFM in the Large Batch Limit

OT-CFM requires costly minibatch OT plans (cubic/quadratic in batch size and sensitive to mode coverage), whereas W-CFM replaces them with simple Gibbs weights $w_\varepsilon(x,y) = \exp(-c(x,y)/\varepsilon)$, and under mild conditions, its loss coincides with EOT-CFM in the large-batch limit. We formalize this in Proposition 2 below—proof is given in Appendix A. A more detailed discussion can be found in Appendix B.

Proposition 2. *Let $\varepsilon > 0$. Suppose that μ, ν, c are such that* (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*

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

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

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

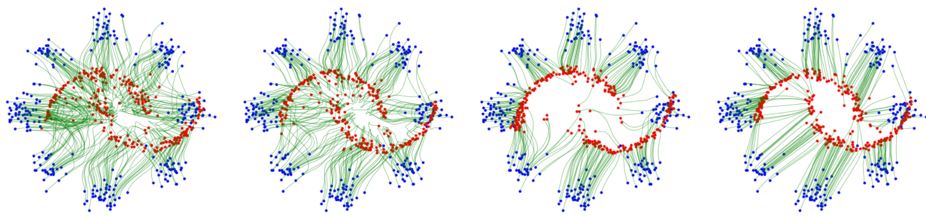

Figure 1: Sample trajectories for moons generation. Blue dots are source samples, red dots are generated samples. From left to right: I-CFM, W-CFM ($\varepsilon = 10$), W-CFM ($\varepsilon = 2$), and OT-CFM.

4 Experiments

Toy datasets. On mapping mixtures of Gaussians to structured targets (Table 1, Figure 1 and 2), W-CFM consistently achieves sample quality comparable to the baselines with careful choice of ε, while significantly improving straightness over I-CFM. Small ε values can distort marginals, but larger ε mitigates this while retaining some path straightness (Figure 5 and 4). OT-CFM with small batches sometimes yields very low NPE, but with mixed sample quality, indicating that straightness alone can be misleading.

Table 1: Performance of CFM variants on 2D datasets (5 seeds). W_2^2 measures sample quality and NPE trajectory straightness (both lower is better).

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 $\pm$ 0.071	1.703 $\pm$ 0.107	0.680 $\pm$ 0.146	1.033 $\pm$ 0.070
OT-CFM	0.029 $\pm$ 0.011	**0.032 $\pm$ 0.019**	**0.232 $\pm$ 0.043**	0.125 $\pm$ 0.011
OT-CFM ($B = 16$)	0.041 $\pm$ 0.014	0.188 $\pm$ 0.041	0.564 $\pm$ 0.125	**0.067 $\pm$ 0.024**
W-CFM (small ε)	**0.018 $\pm$ 0.008**	0.086 $\pm$ 0.021	1.823 $\pm$ 0.166	0.289 $\pm$ 0.008
W-CFM (large ε)	0.029 $\pm$ 0.011	0.097 $\pm$ 0.024	0.843 $\pm$ 0.321	0.463 $\pm$ 0.061

Image datasets. On CIFAR-10, CelebA64, ImageNet64-10, Intel, and Food20, W-CFM matches or surpasses baselines (Table 3), achieving the best FID on all datasets except CelebA64, where OT-CFM benefits from its unimodal structure. Efficiency comparisons (Table 2) show W-CFM reaches competitive or better FIDs with fewer function evaluations.

Table 2: FID $\downarrow$ at varying NFEs (Euler). Lower is better.

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**

Further evaluation in Appendix D. Overall, W-CFM improves sample quality over I-CFM and OT-CFM in multimodal settings, maintains competitive straightness, and achieves superior FID on most image benchmarks with fewer NFEs. A detailed overview of the experimental setup in Appendix C.

5 Conclusion

We propose weighted conditional flow matching (W-CFM), which improves path straightness and sample quality by approximating the entropic OT plan with simple Gibbs weights, avoiding the cost and batch-size limits of explicit OT. With the tuning of a single parameter, one can match the performance of OT-CFM at a fraction of the extra training cost, with minimal impact on the marginals.

References

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

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

R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 2018.

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

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

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

T. Geffner, K. Didi, Z. Zhang, D. Reidenbach, Z. Cao, J. Yim, M. Geiger, C. Dallago, E. Kucukbenli, A. Vahdat, et al. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025.

P. Ghosal, M. Nutz, and E. Bernton. Stability of entropic optimal transport and schrödinger bridges. *Journal of Functional Analysis*, 2022.

W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.

Y. Guo, C. Du, Z. Ma, X. Chen, and K. 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.

R. Irwin, A. Tibo, J. P. Janet, and S. Olsson. Efficient 3d molecular generation with flow matching and scale optimal transport. In *International Conference on Machine Learning*, 2024.

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

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. Technical report.

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

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

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

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

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

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

A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, 2023.

M. Rohbeck, C. Bunne, E. D. Brouwer, J.-C. Huetter, A. Biton, K. Y. Chen, A. Regev, and R. Lopez. Modeling complex system dynamics with flow matching across time and conditions. In *International Conference on Learning Representations*, 2025.

O. Ronneberger, P. Fischer, and T. 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.

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

H. Stark, B. Jing, C. Wang, G. Corso, B. Berger, R. Barzilay, and T. Jaakkola. Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*, 2024.

A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

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

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

X. N. Zhang, Y. Pu, Y. Kawamura, A. Loza, Y. Bengio, D. Shung, and A. Tong. Trajectory flow matching with applications to clinical time series modelling. In *Advances in Neural Information Processing Systems*, 2024.

A Proofs of Theoretical Results

A.1 Proof of Proposition 1

Recall the prior q_ε induced by the Gibbs kernel $\mathcal{K}_\varepsilon(dx, dy) = w_\varepsilon(x, y)\mu(dx)\nu(dy) = \exp(-c(x, y)/\varepsilon)\mu(dx)\nu(dy)$ defined in (7). 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 (10) 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 (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 (11) in the weak sense. Note that $v_\varepsilon(t, \cdot) \in L^1(\mathbb{R}^d, \rho_t)$ 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 $\varepsilon > 0$. 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 between $\mathbf{x}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ 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 (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 (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 (15) [Peyré and 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 (1) with $\varepsilon = 0$.

The sample OT-CFM loss in (16) is a low bias approximation of (17) only when the batch size is large enough. The actual samples for which we compute (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 $\mu \otimes \nu$ are typically 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 (16) approximates $\mathcal{L}_{\mathrm{EOT-CFM}}$ given by (8). 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 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 (6), 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 [Krizhevsky, 2009], CelebA64 [Liu et al., 2015], 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.

D Additional Results

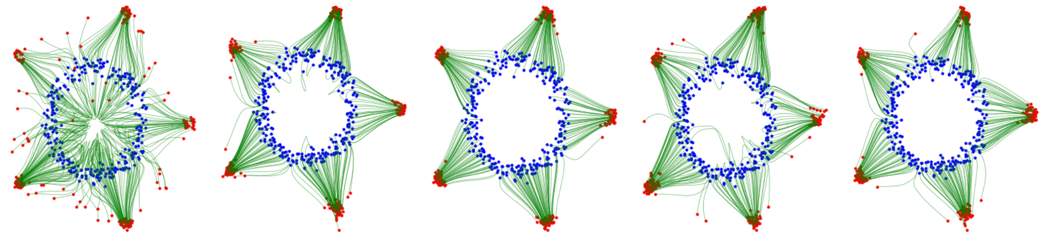

Figure 2: Sample trajectories for Circular MoG → 5 Gaussians. From left to right, the models used are trained with: I-CFM, W-CFM ($\varepsilon = 0.4$), W-CFM ($\varepsilon = 0.2$), OT-CFM (batch size 16), and OT-CFM.

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.

Table 3: FID ↓ across datasets (Dopri5 solver). Lower is better.

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**

Table 4: Comparison of CFM training algorithms' performance on **8 Gaussians → 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	$\mathbf{0.232 \pm 0.043}$	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	$\mathbf{0.033 \pm 0.023}$
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

Table 5: 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**

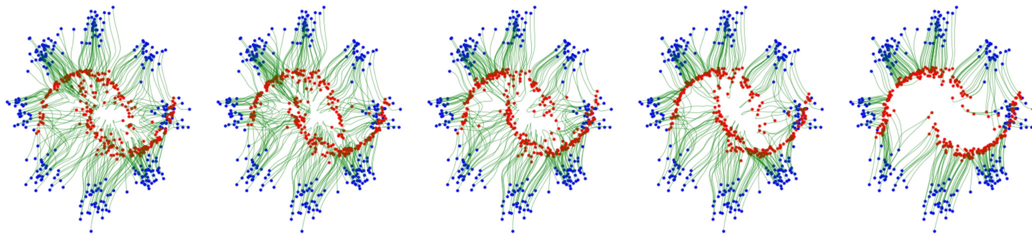

Figure 4: Sample trajectories on **8 Gaussians** $\rightarrow$ **moons** with W-CFM. From left to right, the models used are trained with the following values of ε: 10,8,6,4,2.

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: 2,4,6,8,10.

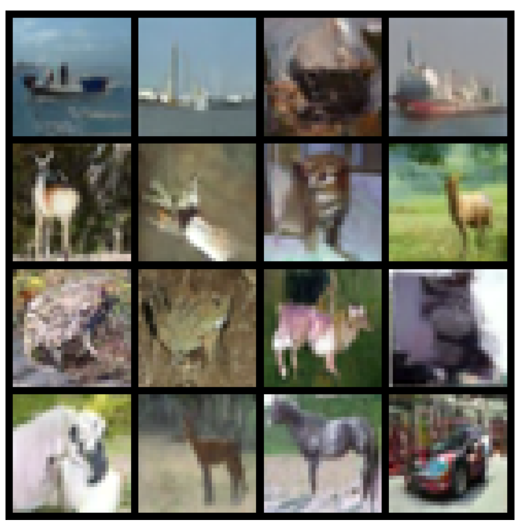

Figure 6: Generated samples from W-CFM trained on CIFAR-10.

Table 6: 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 7: Generated samples from W-CFM trained on CelebA64.

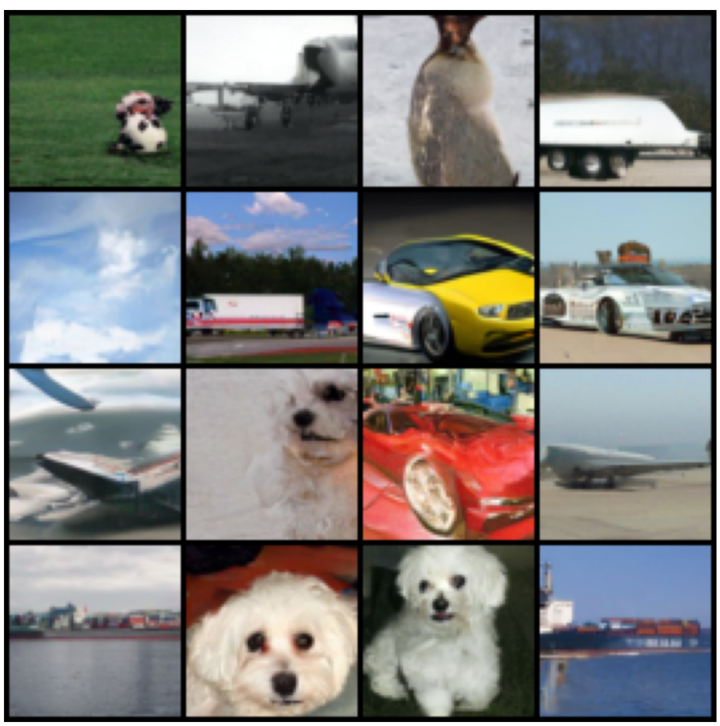

Figure 8: Generated samples from W-CFM trained on ImageNet-10.

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

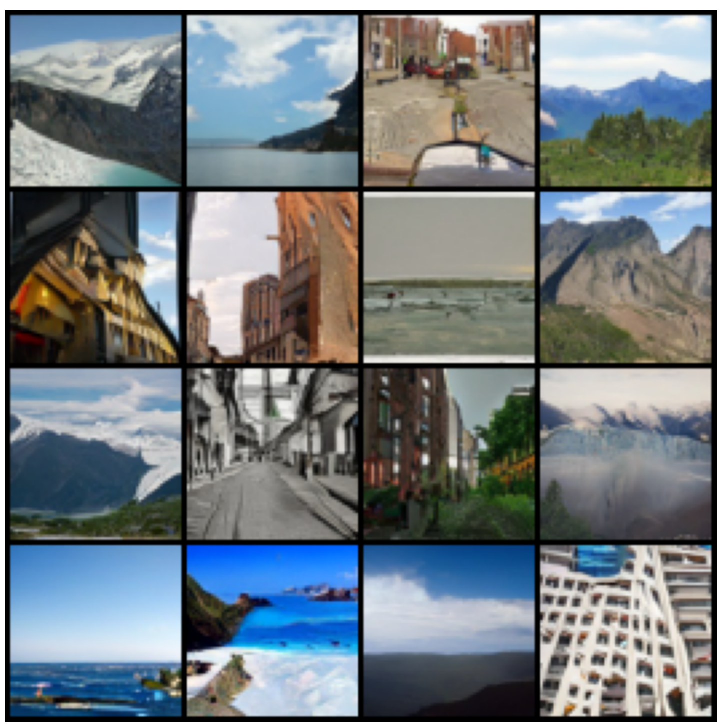

Figure 10: Generated samples from W-CFM trained on Intel Image Classification.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made are detailed in the theoretical justification of our approach, which is detailed in Section 3, and experiments to support our claims are presented in Section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main limitation of our method is the induced tilting of the marginals, which is highlighted both in the presentation of our method (Section 3) and in the experiments on moons generation (Section 4).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions are clearly stated for each result. All complete proofs can be found in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The datasets, model architectures and training algorithms (including hyperparameters) are explicitly discussed in Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our new training loss can easily be implemented by adding a few lines of code to existing CFM training routines. We will release code upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: See Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide frequentist confidence intervals when relevant throughout our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This information is not relevant for the reproducibility and use of our results and method, as the experiments can be performed on a single modern GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We propose a method to make training of an existing class of generative models more efficient, without affecting the overall landscape of what these models can achieve.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our work does not enable misuse beyond what continuous normalizing flows trained with condition flow matching enables.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We reference any other related work and baselines we use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assets are released.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve such risks.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs were only used for the purpose of enhancing the presentation.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.