Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport

Tong et al.

Presented by Beichen Gu, Siddarth Chilukuri

Problem Setup

- Goal: Generative Modeling
- Learn mapping:
  $p0 \rightarrow p1$
- From simple → complex distributions

- Applications:
- Image generation
- Scientific data modeling

Continuous Normalizing Flows (CNFs)

- Key idea: model transformation as ODE

$$dx = u_t(x)\,dt$$

- ut(x): velocity field
- Defines how samples move over time
- Produces continuous, invertible transformations

From Points to Distributions

- Density evolution follows:

$$\frac{\partial p_t}{\partial t} = -\nabla \cdot (p_t u_t)$$

- Ensures probability is preserved
- Connects: point dynamics → distribution dynamics

CNFs vs Diffusion models

- **CNF:**
- Simulating ODE trajectories
- Computing likelihood via integration
- **Diffusion models:**
- Simple regression training
- No simulation needed
- But Requires many sampling steps→ Slow inference

Flow Matching (Before This Paper)

- Learn velocity field via regression
- Avoid simulation

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,x\sim p_t} \left\| v_\theta(t,x) - u_t(x) \right\|^2$$

- Limitation:
- sampling from pt(x)
- often Gaussian assumptions

Key Idea: Conditional Decomposition

- Break global problem into conditional subproblems
- Each z: defines a simple transport task

$$p_t(x) = \int p_t(x|z)q(z)dz$$

- What is condition z?

$$z = (x_0, x_1)$$

x0: source sample, x1: target sample

True Velocity Field (Intractable)

- Problem:
- Requires pt(x)

$$u_t(x) = \mathbb{E}_{q(z)}\left[u_t(x|z)\frac{p_t(x|z)}{p_t(x)}\right]$$

CFM Objective

- Same gradient as original objective

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,z,x\sim p_t(x|z)} \left\| v_\theta(t,x) - u_t(x|z) \right\|^2$$

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t,x\sim p_t} \left\| v_\theta(t,x) - u_t(x) \right\|^2$$

Design Choice: Coupling

- Independent coupling: simple but inefficient
- Optimal transport coupling: matches points optimally, shorter transport paths

$$z \sim \pi(x_0, x_1)$$

OT-CFM (Key Contribution)

- OT objective:

$$\min \int p_t(x) \|u_t(x)\|^2$$

- Minimizes transport cost
- Efficient trajectories
- Fewer ODE steps

Recap: Flow Matching & CFM

- We want to learn a vector field that transports one distribution → another

- Flow Matching (FM):
  - trains CNFs using a regression objective
  - avoids expensive simulation during training

- Conditional Flow Matching (CFM):
  - generalizes FM to more flexible settings

- Key idea:
  - learn velocity field instead of simulating dynamics

Paper Contributions

- Generalized Conditional Flow Matching (CFM):
  - unified framework for training CNFs without simulation
  - works with arbitrary source distributions

- Introduced OT-CFM:
  - uses optimal transport coupling $\pi(x_0, x_1)$
  - produces straighter, more efficient flows

- Proposed Minibatch OT approximation:
  - scalable alternative to full OT
  - works well in practice despite approximation

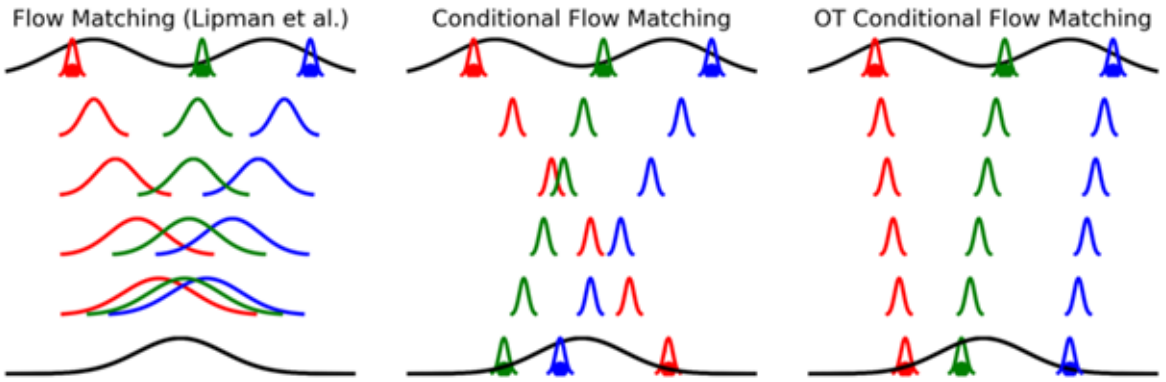
Paper Contributions

- Extensive experiments:
  - low-dimensional OT tasks
  - CIFAR-10 image generation
  - single-cell dynamics
  - image translation + Schrödinger bridge

Problem with CFM

- In CFM, we sample pairs (x_0, x_1)
- These pairs define how mass is transported
- Current approach:
  - pairs are sampled independently
- Problem:
  - transport paths are arbitrary
  - flows become inefficient and noisy
  - harder to train, slower to simulate

Problem with CFM



Source: Tong et al., “Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport”, 2023 (Figure 1)

Optimal Transport (OT)

- Instead of random pairing, use optimal pairing
- Match points that are closest and most efficient to transport
- Objective:
  - minimize total transport cost
- Produces a transport plan $\pi(x_0, x_1)$
- Gives “straight” and efficient paths

OT-CFM

- Keep same framework (CFM)

- Change **ONLY** how we sample pairs (x_0, x_1) :
  - Instead of random pairing,
  - Use optimal transport plan $\pi(x_0, x_1)$

- Leads to:
  - straighter trajectories
  - smoother vector fields

- This reduces:
  - variance in training
  - complexity of learned dynamics

Connection to Dynamic Optimal Transport

- OT-CFM uses optimal transport coupling $\pi(x_0, x_1)$
- Paper shows:
  - learned vector field approximates dynamic OT solution
- As noise → 0:
  - flow minimizes transport energy
- This provides theoretical grounding, not just heuristic improvement

OT Computation Cost

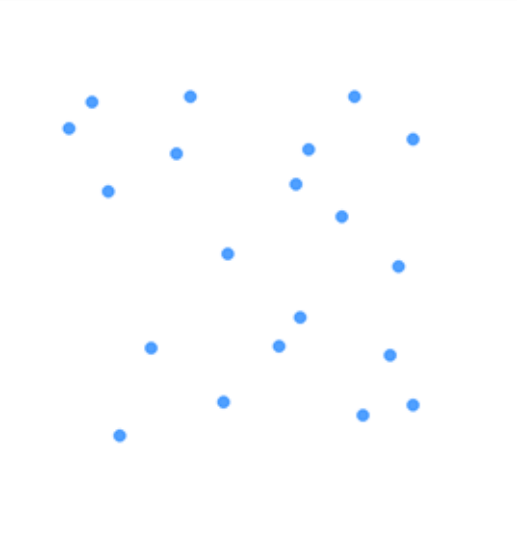
- Exact OT:
  - cubic time complexity
  - quadratic memory

- Not scalable for large datasets

Solution: Minibatch OT

- Compute OT within batches
- Instead of global OT:
  - approximate locally
- For each batch:
  - Sample batch from source and target
  - compute OT matching within batch
  - train model on matched pairs

How Minibatch OT Works

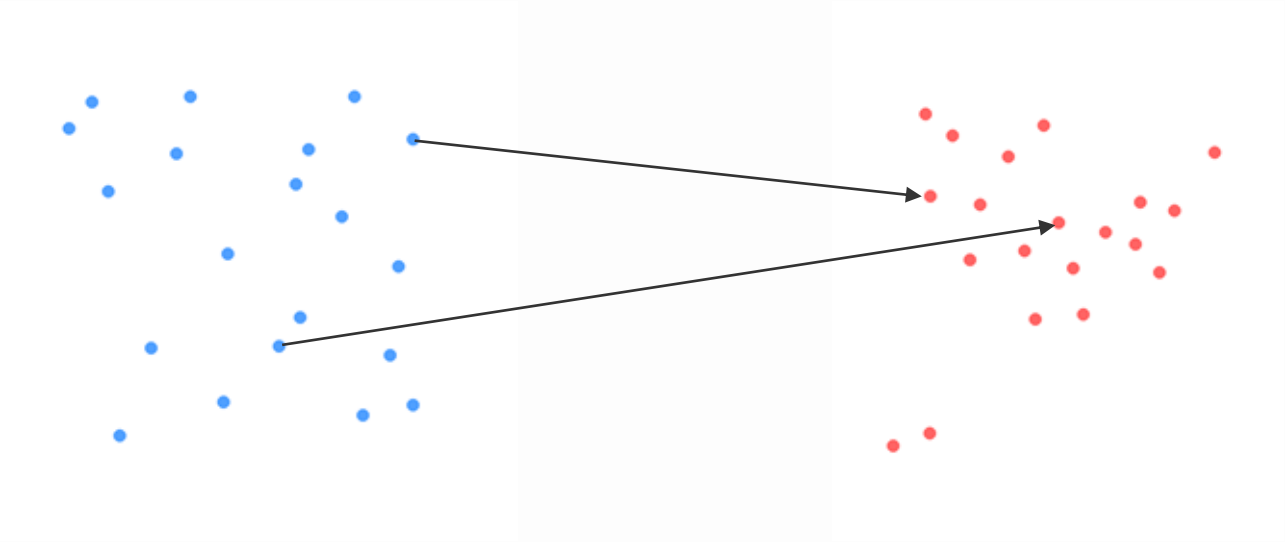


Source distribution
Sample source batch $\boldsymbol{x}_0$

Target distribution
Sample target batch $\boldsymbol{x}_1$

Adapted from Tong et al., 2023

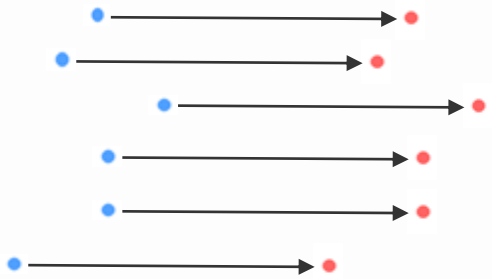
How Minibatch OT Works



OT Matching
Compute OT Matching π_{batch}

Adapted from Tong et al., 2023

How Minibatch OT Works

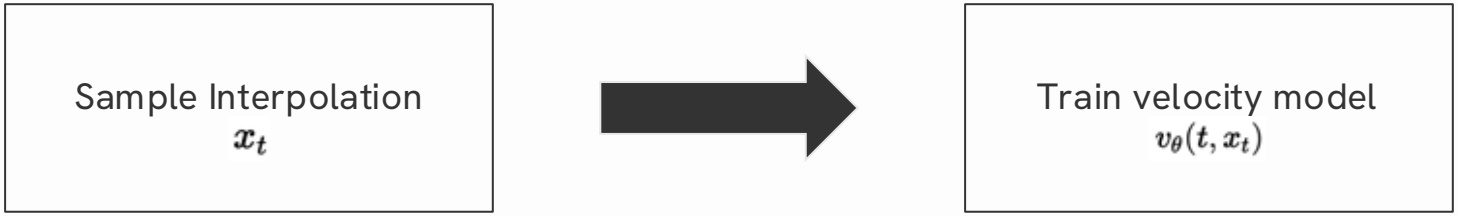


Matched pairs
Form (x_0, x_1)

Sample Interpolation
x_t

Adapted from Tong et al., 2023

How Minibatch OT Works



Adapted from Tong et al., 2023

Why Minibatch OT is enough

- Exact OT not needed
- Even with approximate matching, improves training significantly
- Works well empirically

Summary

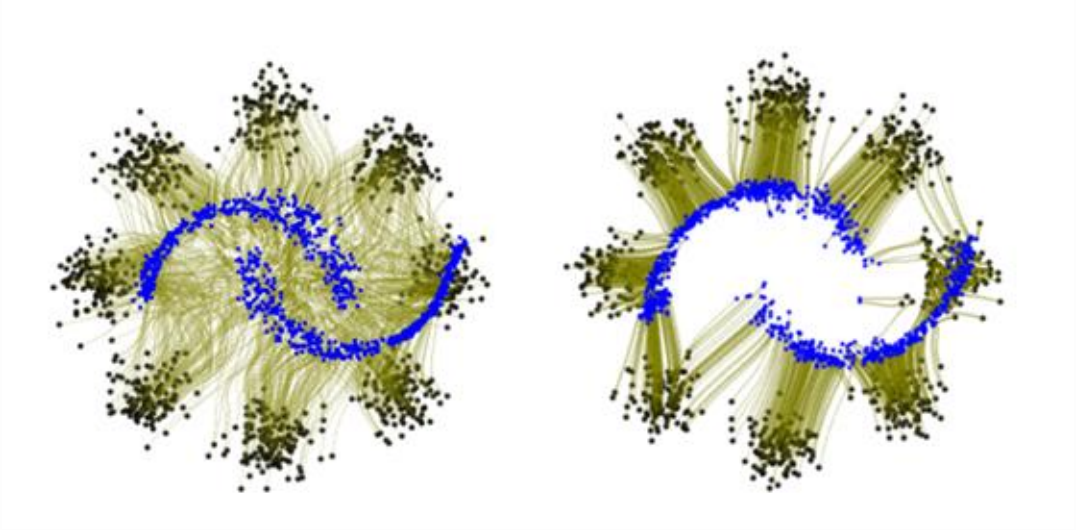
- CFM → regression-based training
- OT-CFM → better pairing using OT
- Minibatch OT → scalable version
- Outcome:
  - better flows
  - faster training
  - faster inference

Experimental Setup

- Evaluated on:
  - low-dimensional datasets (Gaussian, moons, etc.)
  - Schrödinger bridge tasks
  - single-cell data
  - CIFAR-10 image generation
  - CelebA translation

- Metrics:
  - Wasserstein distance
  - path energy
  - FID (images)
  - MMD (translation)

Low-Dimensional Results



OT-CFM achieves lowest normalized path energy, closest to true optimal transport solution

Source: Tong et al., “Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport”, 2023 (Figure 1)

Low-Dimensional Results

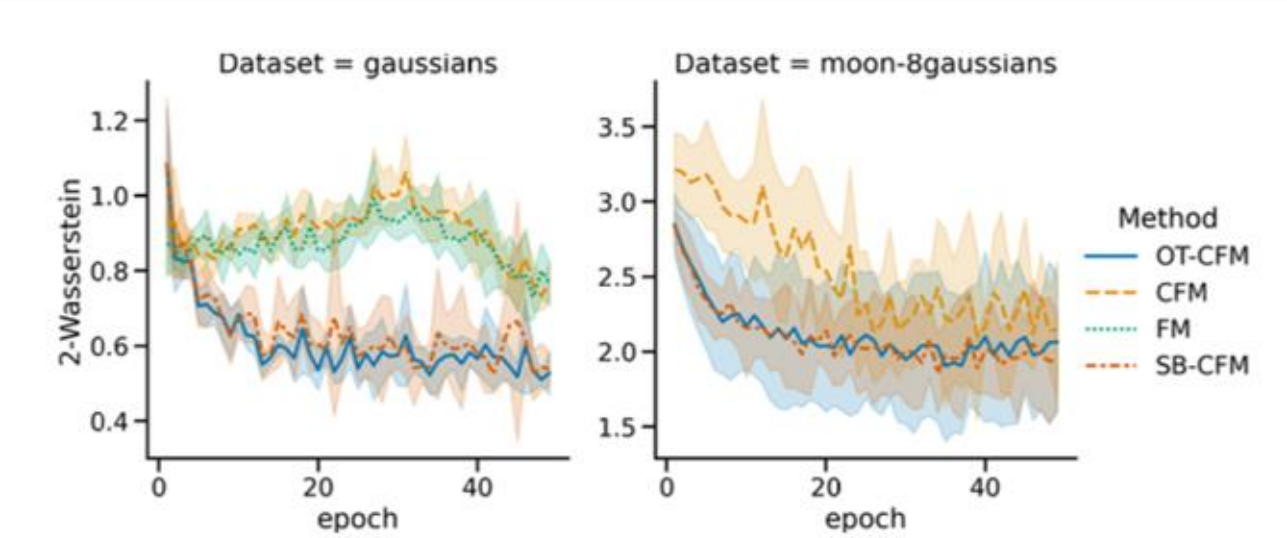
Table 2: Comparison of neural optimal transport methods over four distribution pairs ($\mu \pm \sigma$ over five seeds) in terms of fit (2-Wasserstein), optimal transport performance (normalized path energy), and runtime. '—' indicates a method that requires a Gaussian source. Best in **bold**. CFM and RF models are trained on a single CPU core, other baselines are trained with a GPU and two CPUs.

Dataset →	$\mathcal{N}$→8gaussians		moons→8gaussians		$\mathcal{N}$→moons		$\mathcal{N}$→scurve		Avg. train time
Algorithm ↓ Metric →	W_2^2	NPE	W_2^2	NPE	W_2^2	NPE	W_2^2	NPE	($\times 10^3$ s)
OT-CFM	1.262±0.348	**0.018**±0.014	**1.923**±0.391	**0.053**±0.035	**0.239**±0.048	0.087±0.061	**0.264**±0.093	**0.027**±0.026	1.129±0.335
I-CFM	1.284±0.384	0.222±0.032	1.977±0.266	2.738±0.181	0.338±0.109	0.841±0.148	0.333±0.060	0.867±0.117	**0.630**±0.365
2-RF (Liu, 2022)	1.436±0.344	0.069±0.027	2.211±0.423	0.149±0.101	0.278±0.026	**0.076**±0.067	0.395±0.111	0.112±0.085	0.862±0.166
3-RF (Liu, 2022)	1.337±0.367	0.055±0.043	2.700±0.587	0.123±0.112	0.305±0.026	0.084±0.051	0.395±0.082	0.129±0.075	0.954±0.116
FM (Lipman et al., 2023)	1.062±0.196	0.174±0.030	—	—	0.246±0.077	0.778±0.144	0.377±0.099	0.772±0.081	0.708±0.370
Reg. CNF (Finlay et al., 2020)	1.144±0.075	0.274±0.060	—	—	0.376±0.040	0.620±0.088	0.581±0.195	0.586±0.503	8.021±3.288
CNF (Chen et al., 2018)	**1.055**±0.059	0.151±0.064	—	—	0.387±0.065	2.937±1.973	0.645±0.343	10.548±8.100	18.810±12.677
ICNN (Makkuva et al., 2020)	1.771±0.398	0.747±0.029	2.193±0.136	0.832±0.004	0.532±0.046	0.267±0.010	0.753±0.068	0.344±0.045	2.912±0.626

- OT-CFM outperforms I-CFM and other baselines
- Especially strong on moons ↔ 8 Gaussians (hard case)

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Table 2)

Faster Training



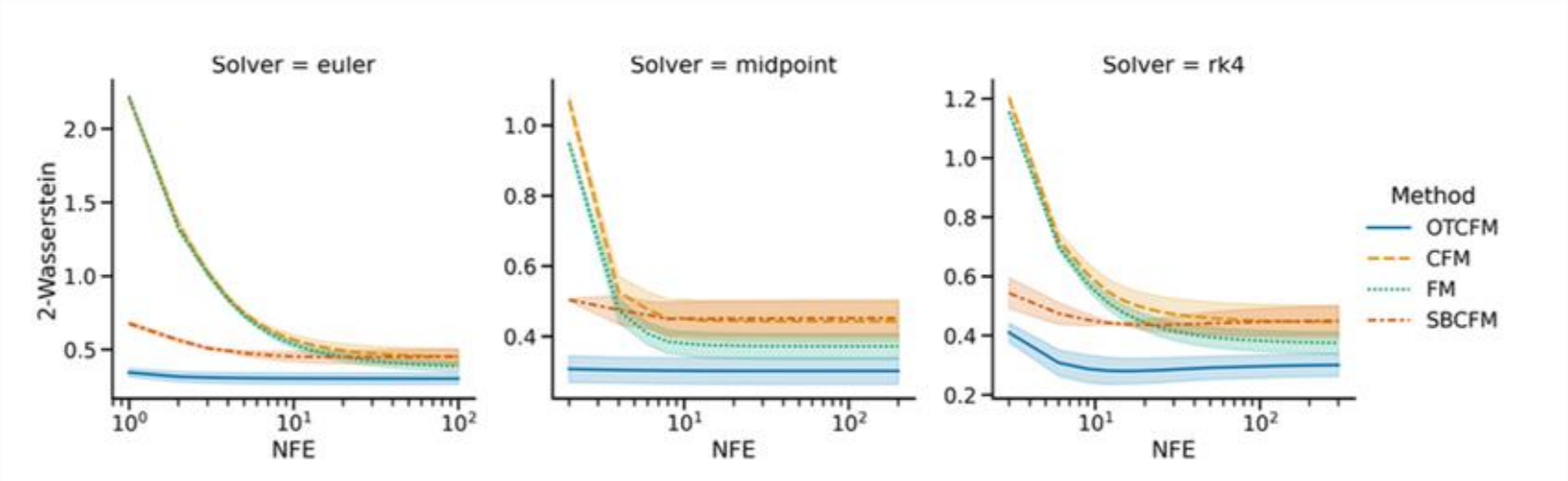
- OT-CFM converges faster (Tong et al., 2023)
- Lower validation error earlier

Because:

- OT pairing reduces variance in training targets
- easier optimization

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Figure 2)

Faster Inference



- OT-CFM needs fewer function evaluations (Tong et al., 2023)
- Same quality with fewer steps

Reason:

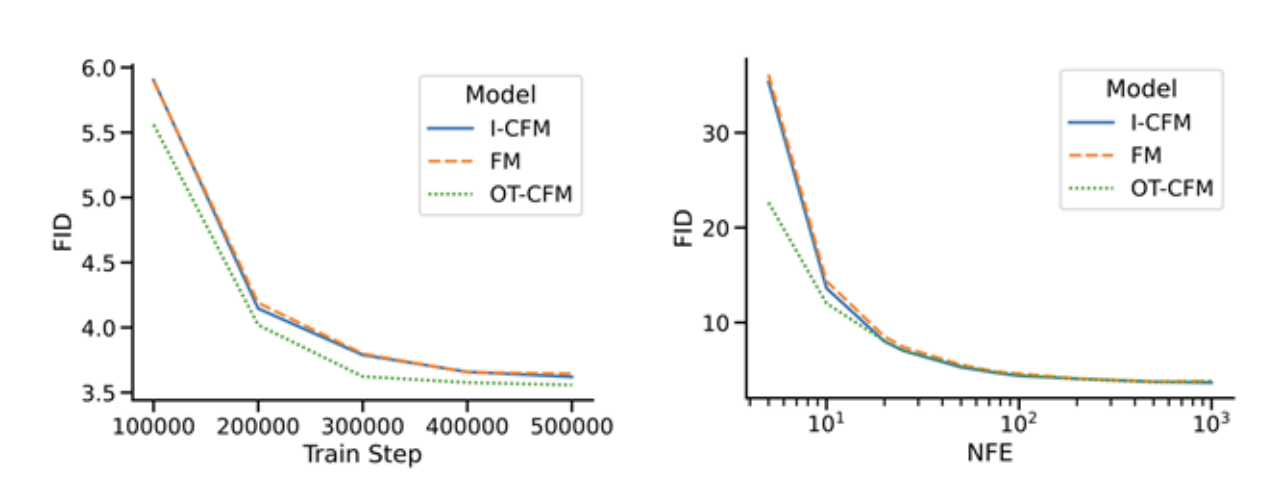
- straighter flows = easier ODE integration

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Figure 2)

High-Dimensional Results (CIFAR-10)

- Evaluated on CIFAR-10 image generation
- Model setup:
  - U-Net architecture (similar to diffusion models)
- Metrics:
  - FID (Fréchet Inception Distance)
  - NFE (Number of Function Evaluations)
- Key results:
  - OT-CFM achieves competitive or better FID scores
  - Requires fewer function evaluations (NFE) for similar quality
  - Performs especially well under limited compute budgets

High-Dimensional Results (CIFAR-10)



- OT-based flows are straighter → easier to integrate
- Leads to faster and more efficient generation

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Figure 3)

Applications of OT-CFM

- Single-cell dynamics
  - Task: interpolate cell distributions over time
  - Metric: Earth mover's distance (EMD)

- OT-CFM achieves lowest error across datasets (Tong et al., 2023)

Table 4: Single-cell comparison over three datasets averaged over leaving out intermediate time-points measuring EMD to left out distribution following Tong et al. (2020). *Indicates values taken from aforementioned work.

Algorithm ↓ Dataset →	Cite	EB	Multi
TrajectoryNet (Tong et al., 2020)*	—	0.848 ± —	—
Reg. CNF (Finlay et al., 2020)*	—	0.825 ± —	—
DSB (De Bortoli et al., 2021)	0.953 ± 0.140	0.862 ± 0.023	1.079 ± 0.117
I-CFM	0.965 ± 0.111	0.872 ± 0.087	1.085 ± 0.099
SB-CFM	1.067 ± 0.107	1.221 ± 0.380	1.129 ± 0.363
OT-CFM	**0.882 ± 0.058**	**0.790 ± 0.068**	**0.937 ± 0.054**

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Table 4)

Applications of OT-CFM

- Image generation
  - Dataset: CIFAR-10
  - Metrics: FID + number of function evaluations (NFE)

- OT-CFM achieves competitive FID.

- Requires fewer evaluations -> faster generation (Tong et al., 2023)

Applications of OT-CFM

- Unsupervised image translation
  - Task: map between attributes (CelebA latent space)
  - Metric: Maximum Mean Discrepancy (MMD)

- OT-CFM achieves better alignment between distributions (Tong et al., 2023)

Table 6: MMD (in units of 10^{-3}) between target and transformed source samples of CelebA latent vectors. Mean and standard deviation over 40 attributes and both translation directions ($- \leftrightarrow +$) for each attribute. 'Identity' refers to performing no translation and treating source samples as approximate samples from the target.

Algorithm ↓	$\sigma = 0.1$	$\sigma = 0.3$	$\sigma = 1$
Identity	9.17 ± 5.68	9.17 ± 5.68	9.17 ± 5.68
I-CFM	4.85 ± 5.09	3.44 ± 2.03	1.59 ± 0.83
OT-CFM	2.81 ± 2.62	1.91 ± 1.30	1.04 ± 0.60

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Table 6)

Applications of OT-CFM

- Schrödinger Bridge
  - Learns probability flow between distributions

- accurately approximates SB solutions, faster than diffusion-based approaches

Table 3: Schrödinger bridge flow comparison, showing average error over flow time to ground truth averaged over 5 models for SB-CFM and 5 dynamics from DSB (De Bortoli et al., 2021).

Dataset ↓ Alg. →	SB-CFM	DSB
$\mathcal{N}$→8gaussians	**0.454 ± 0.164**	1.440 ± 0.720
moons→8gaussians	**1.377 ± 0.229**	2.407 ± 1.025
$\mathcal{N}$→moons	**0.283 ± 0.048**	0.333 ± 0.129
$\mathcal{N}$→scurve	**0.297 ± 0.064**	0.383 ± 0.134

Source: Tong et al., "Improving and Generalizing Flow-Based Generative Models with Minibatch Optimal Transport", 2023 (Table 3)

Paper Strengths

- Simulation-free training
- Supports general source distributions
- Theoretically grounded (OT + SB)
- Faster training + inference
- Strong empirical results

Paper Weaknesses

- Requires closed-form conditional paths
- Minibatch OT is approximate
- OT still adds computational overhead
- Performance depends on batch quality

Open Questions

- Can we learn OT maps directly?
- How does minibatch OT scale to very high dimensions?
- Can we combine this with diffusion models?
- Better approximations to OT?

Final Takeaways

- CFM = simulation-free training for CNFs
- OT-CFM = smarter pairing using optimal transport
- Minibatch OT = makes it scalable
- Result:
  - better flows
  - faster training
  - faster inference