Geometric Optimal Transport for Generative Modeling without Hallucinations

Ryan Engel
Department of Computer Science
Stony Brook University

Yazheng Chen
Department of Computer Science
Stony Brook University

Ben Xin Gu
Cadence

Shing-Tung Yau
Mathematical Science Center
Tsinghua University

Xianfeng Gu
Department of Computer Science
Stony Brook University

***Abstract*—Modern generative models aim to learn a transport map from a data distribution to a prior distribution. While the true objective is to map the support of the source measure onto that of the target measure, widely used flow-based models, Normalizing Flows (NF), Continuous Normalizing Flows (CNF), and Flow Matching (FM), are restricted to learning diffeomorphic transformations, i.e., smooth and invertible maps over the ambient space. Denoising Diffusion (DDPM) must similarly reconstruct manifold structure from noise without explicit coverage guarantees. However, when the data distribution has disconnected support, the regularity theory of the Monge–Ampère equation implies that the optimal transport map is inherently discontinuous. As a result, these models fail to fully cover the target support, leaving zero-density regions that lead to hallucinations in generated samples.**

The geometric variational formulation of optimal transport directly solves a discrete Monge–Ampère equation and constructs transport maps that exactly cover the target support without gaps, thereby eliminating hallucination and mode-mixing issues. We first illustrate this coverage property using two-dimensional embeddings of MNIST, and then provide quantitative comparisons (FID) across four benchmark datasets in a 100-dimensional latent space. Although the exact geometric construction becomes intractable in high dimensions, we introduce a Monte Carlo–based approximation that bypasses this limitation while preserving the essential geometric structure. Experimental results demonstrate that the proposed method achieves competitive generative performance while eliminating hallucinations, an advantage not attainable with conventional generative models.

I. Introduction

In generative AI, a training data set is modeled as a probability distribution on data manifolds embedded in higher dimensional ambient space. The data manifold is mapped onto the latent space, and the data distribution is pushed forward to the latent data distribution. Generative modeling aims to learn a transportation map T from a simple prior distribution (Ω, μ) (e.g., uniform or Gaussian) to a complex data distribution (Ω^*, ν) either in the latent space or in the ambient space. The transportation map is learned by a deep neural network parameterized by θ. The approximated map, denoted as T_θ, pushes the source measure μ forward. The generative model optimizes θ to minimize the distance between the push-forward measure $(T_\theta)_{\#}\mu$ and the data distribution ν, where the distance could be KL divergence, Wasserstein distance or other metrics. Ideally, at the optimum the push-forward measure equals the target measure $(T_\theta)_{\#}\mu = \nu$, if T_θ is C^1, then the following Jacobian equation holds

$$\det(DT_\theta) = \frac{f(x)}{g \circ T_\theta(x)}, \tag{1}$$

where $d\mu(x) = f(x)dx$ and $d\nu(y) = g(y)dy$ are density functions.

Recent approaches, Normalizing Flows [1], Continuous Normalizing Flows [2], Flow Matching [3], and Denoising Diffusion Probabilistic Models [4], parameterize this map as a neural network trained to be smooth and invertible. These diffeomorphic maps are mathematically elegant and have achieved impressive sample quality across many domains. However, these models have intrinsic flaws: the real goal is to achieve transportation maps which map the support of the source measure to the support of the target measure, whereas these models find maps transforming the whole latent or ambient space to itself. In particular, the regularity of the transport map $T : (\Omega, \mu) \to (\Omega^*, \nu)$ itself depends on the regularity of the densities, the curvature condition of the cost function and especially the geometric properties of the domains. In practice, the source domain Ω is always simply connected and convex for simplicity, but the data domain Ω^* may be multiply connected or concave, then the transport map may not be continuous. This shows the fundamental limitation of these approaches: if the learned latent space consists of disconnected clusters separated by zero-density regions, or with large concavities filled with zero-density regions, as is typical for class-conditional data like MNIST, then any smooth, invertible map from a connected prior to this latent space must preserve those gaps; otherwise the Jacobian equation will not hold. In the inference process, points sampled from the gaps decode into meaningless images (hallucinations), degrading generation quality, and introducing mode collapse.

Optimal Transport (OT) based on the geometric variational approach [5] provides a principled alternative. In practice, the target distribution ν is represented as a discrete empirical measure $\nu = \frac{1}{N}\Sigma_{i=1}^{N}\delta(y - y_i)$, where y_i are data samples. The method computes the unique optimal transport map $T : (\Omega, \mu) \to \nu$ by solving a convex optimization problem without

introducing an approximation error. Importantly, even when the transport map is discontinuous, the geometric formulation can still represent the map and explicitly identify the singularity set in Ω. Moreover, the preimage of any zero-density region in the target domain has zero measure on the transport map. These properties provide a theoretical mechanism for avoiding hallucinations and mode collapse in generative models.

The geometric variational formulation constructs power diagrams and weighted Delaunay triangulations during optimization. However, the spatial complexity of the convex hull of N points in $\mathbb{R}^d$ is $O(N^{\lfloor d/2 \rfloor})$, which makes the exact computation intractable in high dimensions. To address this difficulty, An et al. [6] proposed the Monte Carlo Optimal Transport (MC-OT) method, which extends the geometric variational framework to arbitrary dimensions. In this formulation, the gradient of the functional energy depends on the volumes of the power cells. While computing the exact geometry of power cells becomes exponentially expensive as the dimension increases, their volumes can be efficiently estimated using Monte Carlo sampling. As is well known, when integrals are approximated via Monte Carlo sampling, the estimation error decreases proportionally to the inverse square root of the number of samples.

There are many methods to solve optimal transportation problem. The Earth Mover's Distance (EMD) [7] method is to solve Kantorovich problem to find the optimal transportation scheme via linear programming. Although it also gives a precise solution, it involves N^2 unknown variables, in contrast the geometric OT algorithm only tackles N variables. Therefore, the EMD method cannot scale easily. The Sinkhorn [8] algorithm minimizes an entropy-regularized version of the optimal transport objective function, the summation of the transportation cost and the KL-divergence of the transportation plan and $\mu \otimes \nu$, which trades off computational cost and accuracy. It smooths the transportation plan and loses the singularity set information. Sliced Wasserstein Distance (SWD) method is used to approximate high-dimensional optimal transport by projecting distributions onto one-dimensional lines. But it doesn't provide a transportation map directly, which limits its usage in generative AI. Recent techniques are introduced to lift and average multiple 1D plans to create an approximate high-dimensional plan [7].

In this paper, we present a comprehensive comparison of semi-discrete OT, both geometric version and the Monte-Carlo version, against a broad set of generative methods: four diffeomorphic frameworks (Normalizing Flows, Continuous Normalizing Flows, Flow Matching, and Denoising Diffusion), two alternative OT approaches (Earth Mover's Distance and Sliced Wasserstein Distance), a standard autoencoder baseline, and a Variational Autoencoder. We evaluate these methods through two complementary experiments:

1) **Visual evidence** (Section V-C): We embed MNIST images into a 2-dimensional latent space using UMAP and transport these embeddings toward a uniform distribution on $[0, 1]^2$ using NF, FM, CNF, DDPM, OT-EMD, OT-SWD, and exact semi-discrete OT. An autoencoder and VAE are included as additional baselines for visual comparison.
2) **Quantitative evaluation** (Section V-D): We compare generation quality via FID scores across MNIST, Fashion-MNIST, CIFAR-10, and CelebA with 100-dimensional latent space, using AE-OT, AE, OT-SWD, NF, FM, CNF, DDPM, and VAE, confirming that semi-discrete OT consistently outperforms all other methods.

II. Theoretic Foundation

In this section, we briefly review the basic theorems of optimal transportation and Monge-Ampère equation, and the variational approach to solve the semi-discrete optimal transportation problem.

A. Minkowski Problem and Optimal Transportation

a) Minkowski Problem: The classical Minkowski problem asks for the existence and uniqueness of a closed strictly convex hypersurface whose Gaussian curvature (or more generally surface area measure) is prescribed as a function of the outer normal on the unit sphere. The discrete version for convex polytopes was first solved by Minkowski, who proved that a convex polytope with given face normals and face areas exists if and only if the equilibrium condition $\sum_i A_i n_i = 0$ holds, and that the solution is unique up to translation [9]. The general formulation for arbitrary measures on the sphere was later established by Alexandrov, who showed that every suitable measure satisfying the same balance condition arises as the surface area measure of a convex body, again uniquely up to translation [10]. For smooth curvature functions, Pogorelov obtained fundamental curvature estimates and regularity results for the associated Monge–Ampère equation [11]. Nirenberg solved the smooth Minkowski problem in dimension two using nonlinear elliptic PDE techniques [12]. Cheng and Yau later solved the problem in all dimensions by establishing a priori estimates and applying the continuity method to the Monge–Ampère equation on the sphere [13]. Modern treatments and extensions of the Minkowski problem can be found in the monograph of Schneider and in later developments by Oliker and Guan–Li [14], [8], [15].

The Prescribed Gauss Curvature Problem is the local (graph) version of the Minkowski problem: finding a convex function $u(x)$ on a domain Ω given its Gaussian curvature $K(x)$. The governing equation for the prescribed Gauss curvature problem is the Monge-Ampère equation:

$$\frac{\det(D^2 u)}{(1 + |\nabla u|^2)^2} = K(x), \tag{2}$$

here $\det(D^2 u)$ is the determinant of the Hessian matrix, ∇u is the gradient, $K(x)$ the Gauss curvature. Let $u^*(y)$ be the Legendre dual of $u(x)$,

$$u^*(y) := \sup_{x \in \Omega} \langle x, y \rangle - u(x), \tag{3}$$

then $y = \nabla u(x)$, $(D^2u(x))^{-1} = D^2u^*(y)$, and the Gauss curvature of the graph of u^* is $K*(y)$, satisfying

$$K(x)K^*(y) = \frac{1}{(1+|y|^2)^2(1+|x|^2)^2}, \quad y = \nabla u(x).$$

b) Optimal Transportation: The Minkowski problem and the Monge-Ampère equation are closely related to the optimal transportation problem, which seeks the most economical way to transport one probability measure to the other. Suppose the source distribution is μ with a convex support $\Omega \subset \mathbb{R}^n$, the target distribution is ν with support $\Omega^* \subset \mathbb{R}^n$. A map $T : \Omega \to \Omega^*$ is measure-preserving, denoted as $T_{\#}\mu = \nu$, if for any Borel set $B \subset \Omega^*$,

$$\int_{T^{-1}(B)} d\mu(x) = \int_B d\nu(y).$$

The cost function $c : \Omega \times \Omega^* \to \mathbb{R}$ measures the cost for moving a unit mass from $x \in \Omega$ to $T(x) \in \Omega^*$. The optimal transportation problem, originally posed by Gaspard Monge in 1781, seeks the measure-preserving map with the minimal total transportation cost:

$$\min_{T_{\#}\mu=\nu} \int_\Omega c(x, T(x)) d\mu(x).$$

Kantorovich relaxed the transportation map to the transportation scheme (or plan) $\rho : \Omega \times \Omega^* \to \mathbb{R}$, which is a joint distribution with marginal distributions equal to μ and ν respectively, $(\pi_x)_{\#}\rho = \mu$ and $(\pi_y)_{\#}\rho = \nu$, where π_x, π_y are projection maps. The Kantorovich problem is

$$\min_\rho \int_{\Omega\times\Omega^*} c(x,y) d\rho(x,y), \quad s.t. (\pi_x)_{\#}\rho = \mu, (\pi_y)_{\#}\rho = \nu.$$

By introducing Lagrange multipliers, Kantorovich formulated the dual problem:

$$\max_\varphi \int_\Omega \varphi(x) d\mu(x) + \int_{\Omega^*} \varphi^c(y) d\nu(y), \tag{4}$$

where $\varphi^c : \Omega^* \to \mathbb{R}$ is the c-transform of $\varphi : \Omega \to \mathbb{R}$,

$$\varphi^c(y) := \inf_{x\in\Omega} c(x,y) - \varphi(x).$$

When the cost is the squared Euclidean distance $c(x,y) = \frac{1}{2}|x-y|^2$, and the source density function is absolutely continuous, then Brenier proved in [16] that the optimal transport map T is the gradient map of a convex function $u : \Omega \to \mathbb{R}$, $T = \nabla u$, u satisfies the Monge-Ampère equation:

$$\det(D^2u)(x) = \frac{f(x)}{g \circ \nabla u(x)}, \tag{5}$$

with natural boundary condition $\nabla u(\Omega) = \Omega^*$, where f, g are the density functions of μ and ν. Compare Eqn. 2 and Eqn. 5, one can see the intrinsic relation between the Minkowski problem and the optimal transportation.

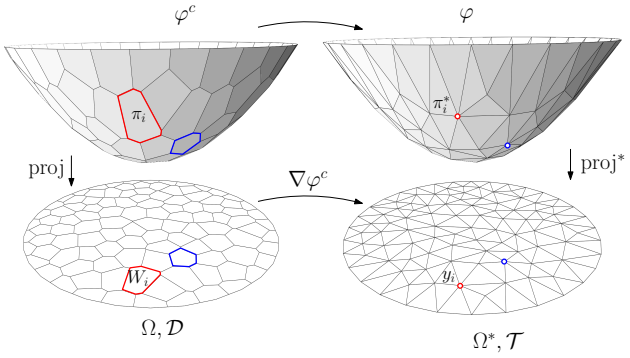


Fig. 1: Variational method for optimal transport.

B. Geometric Variational Method

In [5], Yau et al developed a geometric variational method to solve semi-discrete optimal transport problem, which directly solved the prescribed Gauss curvature problem 2. For the squared Euclidean distance cost $c(x,y) = \frac{1}{2}|x-y|^2$, under the condition $T_{\#}\mu = \nu$,

$$\int_\Omega |T(x)|^2 d\mu(x) = \int_{\Omega^*} |y|^2 d\nu(y)$$

which is independent of the transport map T. Hence the original Monge's problem can be reformulated

$$\min_{T_{\#}\mu=\nu} \int_\Omega \frac{1}{2}|x-T(x)|^2 d\mu(x) \iff \max_{T_{\#}\mu=\nu} \int_\Omega \langle x, T(x)\rangle d\mu(x).$$

with cost function $c(x,y) = \langle x, y\rangle$. The Kantorovich dual problem 4 becomes

$$\min_\varphi \int_\Omega \varphi^c(x) d\mu(x) + \int_{\Omega^*} \varphi(y) d\nu(y) \tag{6}$$

where the c-transform becomes conventional Legendre dual 3.

In practice, the target measure is approximated as the sum of Dirac measures, as shown in Fig. 1,

$$\nu = \sum_{i=1}^n \nu_i \delta(y - y_i), \quad \Omega^* = \{y_1, y_2, \ldots, y_n\}.$$

The Kantorovich potential function

$$\varphi(y) = \sum_{i=1}^n h_i \delta(y - y_i),$$

Let $h = (h_1, h_2, \ldots, h_n)$, the graph of $\varphi(y)$ is the convex hull of the points $\{(y_1, h_1), \ldots, (y_n, h_n)\}$, denoted as $\mathrm{Conv}(h)$. Its Legendre dual

$$\varphi^c(x) = \sup_{y\in\Omega^*} \langle x, y\rangle - \varphi(y) = \max_{i=1}^n \{\langle x, y_i\rangle - h_i\}$$

The graph of $\varphi^c(x)$ is the upper envelope of the planes $\pi_i(x) := \langle x, y_i\rangle - h_i$, $i = 1, \ldots, n$, denoted as $\mathrm{Env}(h)$.

The projection of the upper envelope induced a power diagram of $\mathcal{D}(h)$ Ω, $\Omega = \bigcup_{i=1}^n W_i(h)$,

$$W_i(h) := \{x \in \Omega : \langle x, y_i\rangle - h_i \geq \langle x, y_j\rangle - h_j, \forall j\}$$

The dual of the power diagram is the power Delaunay triangulation $\mathcal{T}(h)$ of Ω^*, each cell $W_i(h)$ is dual to the point y_i, each edge $e_{ij} := W_i(h) \cap W_j(h)$ is dual to the edge $\bar{e}_{ij}$ in the Delaunay triangulation connecting y_i and y_j.

The Kantorovich dual energy 6 has simple form due to the piecewise linearity of φ^c,

$$E(h) = \sum_{i=1}^{n} \int_{W_i(h)} (\langle x, y_i \rangle - h_i) d\mu(x) + \sum_{j=1}^{n} h_j \nu_j \qquad (7)$$

The gradient of the energy is

$$\nabla E(h) = (\nu_1 - w_1(h), \nu_2 - w_2(h), \ldots, \nu_n - w_n(h))$$

here $w_i(h)$ is the μ-measure of the power cell $W_i(h)$,

$$w_i(h) = \mu(W_i(h)) = \int_{W_i(h)} f(x) dx$$

Furthermore, the Hessian matrix of the energy has explicit geometric interpretation: for off-diagonal elements

$$\frac{\partial^2 E(h)}{\partial h_i \partial h_j} = -\frac{\partial w_i(h)}{\partial h_j} = -\frac{\mu(e_{ij})}{|y_j - y_i|}$$

here $\mu(e_{ij})$ is unconventionally defined as

$$\mu(e_{ij}) := \int_{W_i(h) \cap W_j(h)} f(x) dx$$

for diagonal elements

$$\frac{\partial^2 E(h)}{\partial h_i^2} = -\sum_{j=1}^{n} \frac{\partial^2 E(h)}{\partial h_i \partial h_j} = \sum_{j=1}^{n} \frac{\partial w_i(h)}{\partial h_j} > 0$$

Hence the Hessian matrix has one dimensional null space $\lambda(1, 1, \ldots, 1)^T$, and is diagonal dominant, it is positive definite on the complementary space. We define the admissible solution space as

$$\mathcal{H} := \left\{ h \in \mathbb{R}^n : \sum_{i=1}^{n} h_i = 0 \right\} \cap \{ W_i(h) \neq \emptyset, \forall i = 1, \ldots, n \}.$$

The energy 7 is strictly convex on $\mathcal{H}$. If $h_1, h_2 \in \mathcal{H}$, then

$$W_i((1-t)h_1 + th_2) = (1-t)W_i(h_1) \oplus tW_i(h_2),$$

where $\oplus$ is the Minkowski sum. By the Brunn-Minkowski inequality, the volume of $W_i((1-t)h_1 + th_2)$ is a concave function of t, therefore its volume is greater than 0, hence $(1-t)h_1 + th_2 \in \mathcal{H}$, $\mathcal{H}$ is convex. At a boundary point $h \in \partial\overline{\mathcal{H}}$, some cells are degenerated, $w_i(h) = 0$, so the gradient $v_i - w_i(h)$ points to the interior of $\mathcal{H}$, therefore the unique optimum is achieved at some interior point $h^* \in \mathcal{H}$. At the critical point h^*, the gradient $E(h)$ is zero, for each cell the μ-volume $w_i(h^*)$ equals to the prescribed measure ν_i. The optimal transport map T maps each cell $W_i(h^*)$ to y_i. Geometrically, if μ is the Lebesgue measure, then the convex hull $\mathrm{Conv}(h^*)$ is the solution to the prescribed Gauss curvature problem, namely the solution to the Monge-Ampère equation 5.

Fig. 2 shows one computational example. The female face model is shown in the top row. The surface is conformally mapped on the planar unit disk shown in the middle left frame. The conformal factor is treated as the target measure ν, the source measure μ is the Lebesgue measure. The image of the optimal transport map is shown in the middle right frame. The Kantorovich potentials $\varphi^c(x)$ and $\varphi(y)$ are shown in the bottom row.

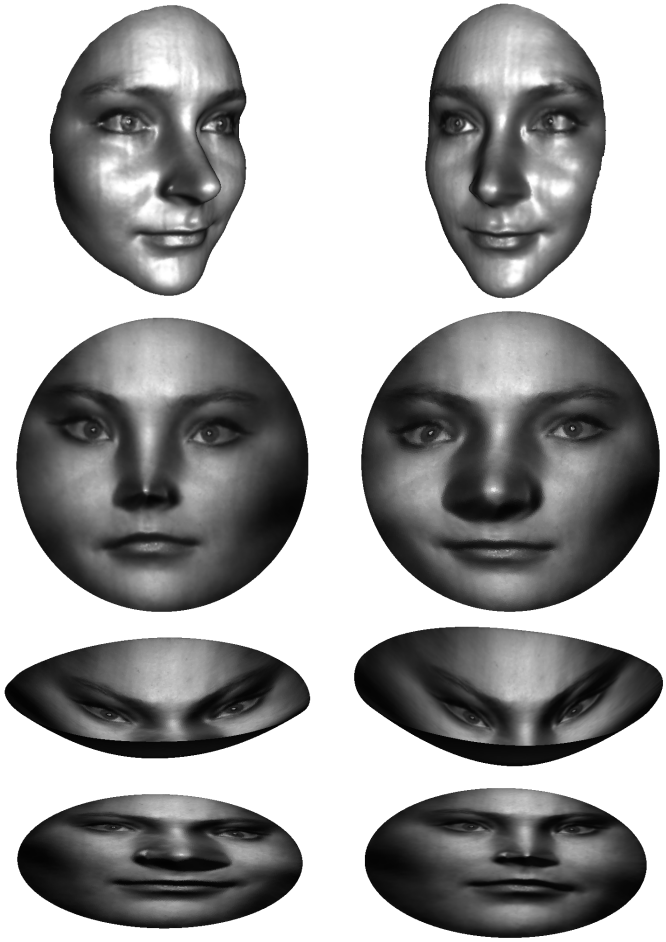
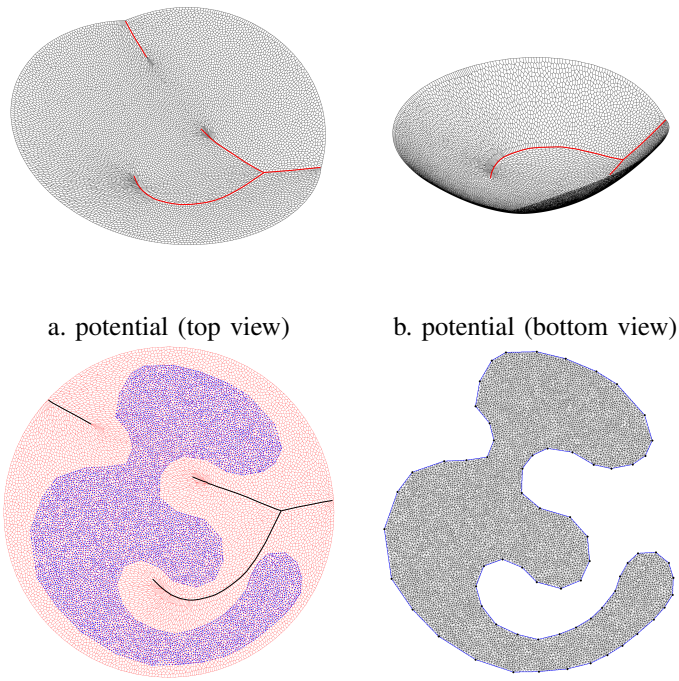


Fig. 2: Optimal transport map for a human face model.

C. Regularity of Transport Maps

The regularity theory of optimal transport is closely connected to the regularity theory of the Monge–Ampère equation. The analysis of this equation builds on the weak solution theory introduced by Aleksandrov through the Monge–Ampère measure [17]. Caffarelli established fundamental interior regularity results showing that if the densities are bounded away from zero and infinity then convex solutions are locally $C^{1,\alpha}$, and if the densities are Hölder continuous then the potential is locally $C^{2,\alpha}$ [18], [19]. For the global problem, if the domains Ω and Ω^* are bounded, uniformly convex, and have C^2 boundaries, and if the densities satisfy $0 < \lambda \leq \rho, \sigma \leq \Lambda$ with $\rho, \sigma \in C^\alpha$, then Caffarelli proved boundary regularity and obtained $u \in C^{2,\alpha}(\overline{\Omega})$, implying that the optimal transport map is a $C^{1,\alpha}$ diffeomorphism between the domains [20].



a. potential (top view) b. potential (bottom view)

c. singularity set d. target domain

Fig. 3: The singularity set of an optimal transport map.

For more general cost functions, Ma, Trudinger, and Wang discovered a curvature condition on the cost function, now known as the MTW condition, which ensures regularity of optimal transport maps and extends the smooth theory beyond the quadratic cost case [21]. Loeper subsequently clarified the geometric role of the MTW condition and proved that it is essentially necessary for continuity of optimal transport maps, while also establishing regularity results under this condition [22]. Comprehensive treatments of these results and their geometric structure can be found in the monographs of Villani and Figalli [23], [24].

Convexity of the target domain remains essential for global regularity in the quadratic case, as it guarantees obliqueness of the boundary condition and prevents degeneracy of the Monge–Ampère equation near the boundary. The second boundary condition $\nabla u(\Omega) = \Omega^*$ implies $\nabla u(\partial\Omega) \subset \partial\Omega^*$. The obliqueness condition requires

$$\langle n(x), n^*(\nabla u(x)) \rangle > 0$$

where $n(x), n^*(y)$ are the outward normals of Ω and Ω^* respectively. Figalli showed that when the target domain is not convex, global regularity may fail: even for smooth positive densities, singularities of the transport map can appear, although the map remains smooth outside a lower-dimensional singular set.

The smoothness of optimal transport maps depends on the geometric properties of the domains, the regularity of the densities, and curvature conditions on the cost function. Fig. 3 shows the singularity set of an optimal transport map. The source measure (μ, Ω) is the uniform distribution on the unit disk, the target measure (ν, Ω^*) is also the uniform distribution on a concave domain. The Brenier potential $\varphi^c : \Omega \to \mathbb{R}$ is shown in the top row, the singularity set is the red curve, where the potential is only C^0 not C^1. The singularity set on the disk is shown as black curves, where the optimal transport map T is discontinuous.

III. Related Work

a) Normalizing Flows: Normalizing Flows [1], [25] learn an invertible map between data and a base distribution (typically Gaussian) by composing simple bijective transformations. The change-of-variables formula provides exact log-likelihood, but the diffeomorphic constraint limits expressiveness: topological features of the data manifold are necessarily preserved.

b) Continuous Normalizing Flows: CNFs [2], [26] parameterize the transformation as the flow of an ODE, using the instantaneous change of variables formula. The Hutchinson trace estimator enables scalable training, but the map remains a diffeomorphism.

c) Flow Matching: Flow Matching [3] trains a velocity field by regressing on conditional probability paths, avoiding the simulation required by CNFs. The resulting ODE flow is still diffeomorphic.

d) Denoising Diffusion: DDPMs [4], [27] learn the reverse of a fixed forward noising process. The forward process progressively adds Gaussian noise until the data distribution is destroyed; the reverse process learns to denoise, generating samples from noise. While the stochastic reverse process is not strictly diffeomorphic, it shares the same practical limitation: it must reconstruct manifold structure from noise without explicit coverage guarantees.

e) Optimal Transport: Classical OT [23] finds the minimum-cost map between distributions. Computational approaches include discrete solvers [28], [7], neural OT approximations [29], and sliced methods [30]. Semi-discrete OT [5], [31] partitions the continuous target into power cells assigned to discrete source points, providing exact full-coverage guarantees. AE-OT [6] scales this to high dimensions via Monte Carlo volume estimation.

IV. Computational Methods

A. Semi-Discrete Optimal Transport

Given a dataset $\{x_i\}_{i=1}^N$, we train an autoencoder to obtain latent codes $Y = \{y_1, y_2, \ldots, y_n\} \subset \mathbb{R}^d$. The generative task is to learn a map from a source domain $\Omega \subset \mathbb{R}^d$ equipped with a continuous measure μ (uniform or Gaussian) to the discrete target measure $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. As established in the geometric variational framework, the optimal height vector h^* is found by minimizing the Kantorovich dual energy (7), whose gradient is the volume error $(\nu_i - w_i(h))$. At the optimum, every power cell has equal μ-volume $w_i(h^*) = \nu_i$, and the cells tile Ω with no gaps and no overlaps, guaranteeing full coverage by construction.

Algorithm 1 Exact Semi-Discrete OT (2D)

Require: Source points $Y = \{y_i\}_{i=1}^n \subset \mathbb{R}^2$, target volumes $\nu_i = 1/n$
Ensure: Optimal height vector h^*
Initialize $h \leftarrow \mathbf{0} \in \mathbb{R}^n$
repeat
  Lift points to paraboloid: $\hat{y}_i \leftarrow (a_i, b_i, a_i^2 + b_i^2 + h_i)$
  Compute lower convex hull of $\{\hat{y}_i\}$
  Project to obtain power diagram $\{W_i(h)\}$
  Clip each cell to $[0,1]^2$; compute exact areas $w_i(h)$
  Compute gradient: $g_i \leftarrow \nu_i - w_i(h)$
  Assemble Hessian H from shared boundary lengths
  Newton update: $h \leftarrow h - \alpha\, H^{-1} g$
  Center: $h \leftarrow h - \mathrm{mean}(h)$
until $\|g\|_\infty < \epsilon$
return h

Algorithm 2 Semi-Discrete OT Map [6]

Require: Latent codes $Y = \{y_i\}_{i \in I}$, empirical distribution $\nu = \frac{1}{|I|}\sum_{i \in I} \delta_{y_i}$, number of MC samples N, positive integer s
Ensure: Optimal transport map $T(\cdot)$
Initialize $h = (h_1, h_2, \ldots, h_{|I|}) \leftarrow (0, 0, \ldots, 0)$
repeat
  Generate N uniformly distributed samples $\{x_j\}_{j=1}^N$
  Calculate $\nabla h = (\nu_i - \hat{w}_i(h))^T$
  $h \leftarrow h - \mathrm{mean}(h)$
  Update h by Adam with $\beta_1 = 0.9$, $\beta_2 = 0.5$
  if $E(h)$ has not decreased for s steps **then**
    $N \leftarrow N \times 2$
  end if
until converge
OT map $T(\cdot) \leftarrow \nabla(\max_i \langle \cdot, y_i \rangle - h_i)$

For the 2D visualization experiments (Section V-C), we use the exact solver of Gu et al. [5]. In $\mathbb{R}^2$, the power diagram corresponds to the lower envelope of a convex hull in $\mathbb{R}^3$. Each source point $y_i = (a_i, b_i)$ is lifted to the paraboloid:

$$\hat{y}_i = \big(a_i,\ b_i,\ a_i^2 + b_i^2 + h_i\big). \tag{8}$$

The lower convex hull of $\{\hat{y}_i\}$ projects onto Ω as the power diagram. Cell areas and adjacencies are computed exactly via computational geometry (Sutherland-Hodgman clipping to $[0,1]^2$, shoelace formula for areas). The Hessian of $E(h)$ is assembled from the lengths of shared boundaries between adjacent cells, enabling Newton's method with quadratic convergence. The procedure is summarized in Algorithm 1.

This converges in typically 10–20 Newton iterations and produces the power diagram visualizations in Fig. 4, where the cell structure is directly visible.

B. Monte Carlo Semi-Discrete OT for High Dimensions

In dimensions $d \gg 2$, exact power diagram construction is computationally intractable (the number of cell faces grows exponentially with d). An et al. [6] bypass this limitation by estimating cell volumes via Monte Carlo sampling. The key observation is that $w_i(h)$ can be estimated by drawing N random samples $\{x_j\}_{j=1}^N \overset{\text{i.i.d.}}{\sim} \mu$ and counting assignments:

$$\hat{w}_i(h) = \frac{\#\{j : x_j \in W_i(h)\}}{N}, \tag{9}$$

where $x_j \in W_i(h)$ is determined by $i = \arg\max_i \{\langle x_j, y_i \rangle - h_i\}$. This assignment step reduces to a matrix multiplication $XY^\top - h$, which is efficiently parallelized on GPU regardless of dimension. The gradient is then approximated as $\nabla E \approx (\nu_i - \hat{w}_i(h))^T$. When N is large, $\hat{w}_i(h)$ converges to $w_i(h)$.

To balance precision and speed, An et al. employ an adaptive strategy: if $E(h)$ stops decreasing for s consecutive steps, N is doubled. The full procedure is given in Algorithm 2.

C. Sample Generation via Extended OT Map

The semi-discrete OT map $T = \nabla u_h$ maps all $x \in \Omega$ to the training latent codes $\{y_i\}$ and does not generate new samples by itself. An et al. [6] extend T to a piecewise-linear (PL) map $\tilde{T}$ as follows. Each cell $W_i(h)$ is represented by its μ-mass center $\hat{c}_i = \frac{1}{\mu(W_i)} \int_{W_i} x\, d\mu(x)$, estimated in practice as the mean of MC samples falling in W_i. The adjacency structure of the power diagram induces a triangulation of the centers: if $W_i \cap W_j \neq \emptyset$, then $\hat{c}_i$ and $\hat{c}_j$ are connected. Given a random sample x, the $d+1$ nearest centers $\{\hat{c}_{i_0}, \ldots, \hat{c}_{i_d}\}$ form a simplex, and the generated code is the weighted interpolation:

$$\tilde{T}(x) = \sum_{k=0}^{d} \lambda_k\, T(\hat{c}_{i_k}), \quad \lambda_k = \frac{d^{-1}(x, \hat{c}_{i_k})}{\sum_{j=0}^{d} d^{-1}(x, \hat{c}_{i_j})}, \tag{10}$$

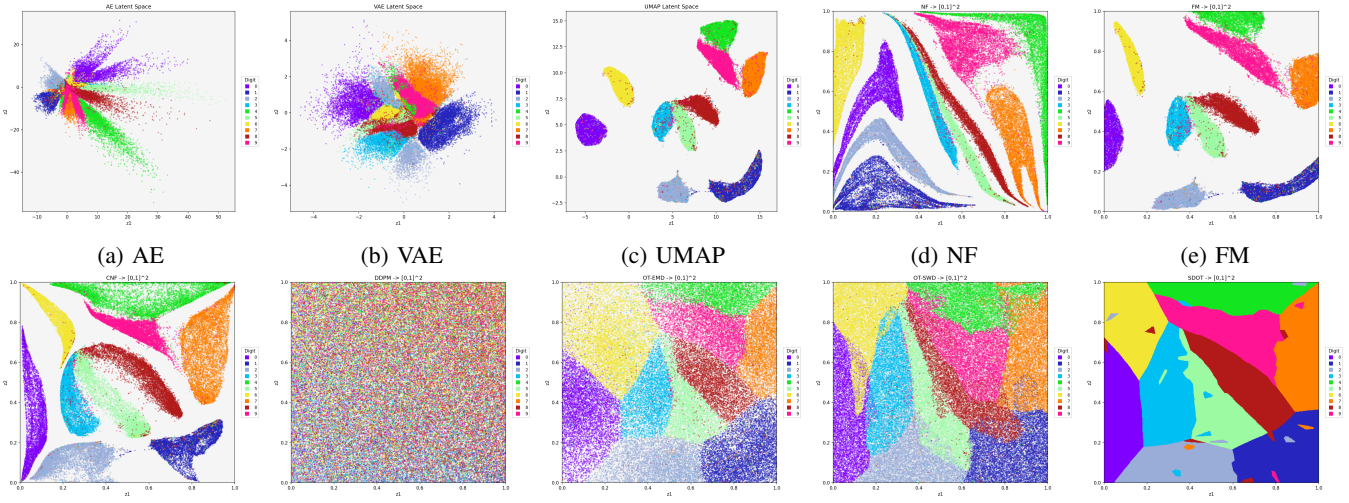
where $d^{-1}(x, \hat{c}_{i_k})$ denotes the inverse Euclidean distance from x to center $\hat{c}_{i_k}$.

a) Singular set detection: By Figalli's regularity theory [32], if the target distribution has multiple modes, the Brenier potential u_h has ridges where it is continuous but not differentiable, making the transport map discontinuous. Interpolating across these *singular sets* produces spurious samples (mode mixture). To detect singularities, we compute the angle θ_{ik} between neighboring cells i_0 and i_k. If all angles exceed a threshold $\hat{\theta}$, the sample lies near the singular set and is discarded. Otherwise, only the $d' + 1 \leq d + 1$ neighbors with $\theta_{ik} \leq \hat{\theta}$ are retained, and Eq. (10) is applied over the filtered set. This is summarized in Algorithm 3.

The generated codes are decoded through the shared autoencoder to produce images. By discarding samples near the singular set, the method avoids both mode collapse (all source points are used in the triangulation) and mode mixture (cross-mode interpolation is filtered out).

V. Experiments

First we present results on classical 2D datasets: swissroll and checkerboard. We organize our experiments into two groups. Group A provides visual evidence of the coverage



(a) AE (b) VAE (c) UMAP (d) NF (e) FM

(f) CNF (g) DDPM (h) OT-EMD (i) OT-SWD (j) SD-OT

Fig. 4: 2D MNIST visualization (70k points colored by digit). Top row: (a-c) latent space of different encoders, (d–f) flow based methods preserve gaps. Bottom row: (g) diffusion method, (h) OT-EMD exact discrete assignment, (i) OT-SWD approximate coverage, (j) semi-discrete OT partitions the entire domain with no gaps.

Algorithm 3 Generate Latent Code [6]

Require: OT map $T(\cdot)$, number of samples to generate n, angle threshold $\hat{\theta}$
Ensure: Generated latent codes P

Compute cell centers $\hat{c}_i$ by Monte Carlo method
repeat
  Sample $x \sim \mu$; find the $d+1$ nearest centers $\{\hat{c}_{i_0}, \hat{c}_{i_1}, \ldots, \hat{c}_{i_d}\}$ sorted by distance
  Compute angles θ_{i_k} between cells i_0 and i_k
  Select i_k with $\theta_{i_k} \leq \hat{\theta}$, yielding $\{\hat{i}_k\}_{k=0}^{d'}$
  if $\forall k,\ \theta_{i_k} > \hat{\theta}$ **then**
    Abandon x {singular set}
  else
    Generate: $\tilde{T}(x) = \sum_{k=0}^{d'} \lambda_k T(\hat{c}_{\hat{i}_k})$ with $\lambda_k = \frac{d^{-1}(x, \hat{c}_{\hat{i}_k})}{\sum_j d^{-1}(x, \hat{c}_{\hat{i}_j})}$
  end if
until n new latent codes generated
return P

problem using 2D MNIST embeddings, where the contrast between OT and conventional methods is immediately apparent. Group B provides quantitative FID evaluation across four datasets in a 100-dimensional latent space, confirming that the visual intuition extends to practical generation quality.

A. Swiss Roll and Checkerboard

We evaluate SD-OT against three generative model baselines, DDPM [4], Vanilla Flow Matching (FM) [3], and OT-Flow Matching (OT-FM) [33] on two canonical 2-D benchmarks: the Swiss Roll spiral manifold and the Checkerboard multi-modal distribution.

Both datasets use N = 10,000 discrete target points. For evaluation, 5,000 samples are generated per method using a shared fixed Gaussian noise pool $\{x_k\}_{k=1}^{5000} \sim \mathcal{N}(0, I)$ (seed = 0), ensuring that any transport-cost comparison is fair. For 2-D distributions we use the following metrics in place of FID:

- $\mathbf{W_2}$: exact Wasserstein-2 distance computed via the EMD solver in the POT library [7], subsampled to 2,000 points per distribution.
- **SWD**: Sliced Wasserstein Distance with 200 random projections.

a) Baselines: All baselines use a 4-layer MLP with hidden width 256 and SiLU activations.

- **DDPM**: cosine noise schedule, $T = 1{,}000$ diffusion steps, sinusoidal time embedding of dimension 64. Inference uses the full 1,000-step ancestral sampler.
- **Flow Matching (FM)**: straight conditional flows with independent Gaussian–data coupling; Euler ODE integration with $\Delta t = 1/50$.
- **OT-Flow Matching (OT-FM)**: same architecture as FM, but each mini-batch is re-paired via the Hungarian algorithm (mini-batch OT coupling) before computing the Flow Matching loss. Inference identical to FM (50 Euler steps).

B. Results on Swiss Roll and Checkerboard

SD-OT achieves the best SWD (0.0068) with a single forward pass (NFE = 1), matching FM's leading W_2 within 2% while requiring 50× fewer function evaluations.

The trajectory visualisations confirm this theoretically: SD-OT paths are straight and non-crossing (consistent with Brenier's theorem), DDPM paths resemble a random walk, and FM / OT-FM paths are straighter than DDPM but still exhibit crossing.

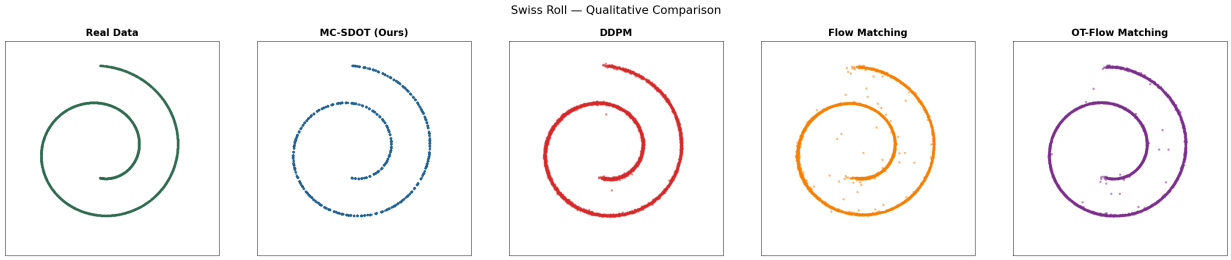


Fig. 5: **Qualitative comparison on Swiss Roll.** Generated samples from each method overlaid on the true data (grey). MC-OT produces razor-sharp spiral arms with no off-manifold "hallucinated" points, whereas DDPM shows manifold bleeding in the inter-arm gaps and FM / OT-FM exhibit mild boundary blur at the edges.

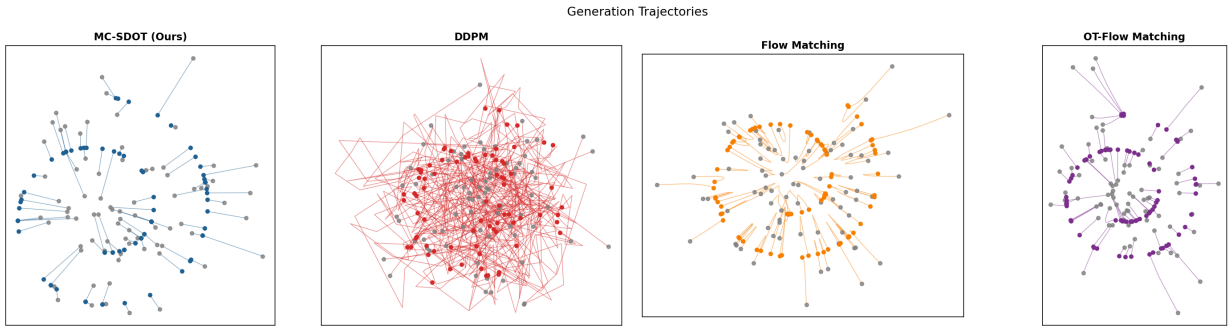


Fig. 6: **Generation trajectories on Swiss Roll.** Each curve traces a single particle from Gaussian source ($t = 0$) to generated sample ($t = 1$). SD-OT paths are straight and non-crossing (Brenier's theorem); DDPM paths resemble a random walk with large detours; FM / OT-FM paths are straighter but still exhibit crossings.

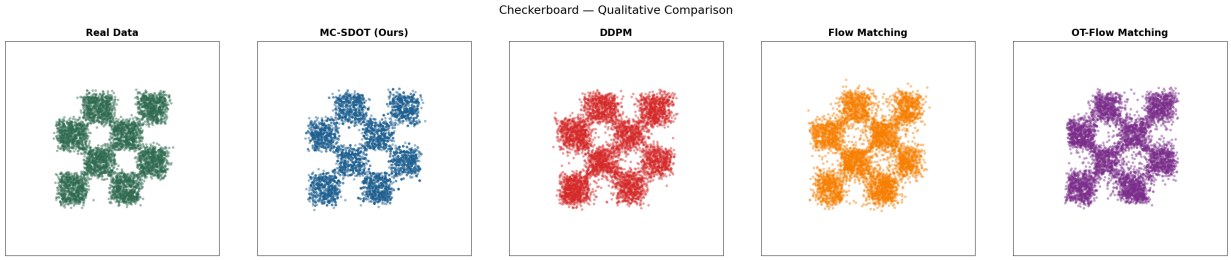


Fig. 7: **Qualitative comparison on Checkerboard.** Each panel shows 10,000 generated samples. SD-OT fills each of the eight cells cleanly with negligible leakage into the white inter-cell gaps. DDPM over-concentrates mass in the corner cell $(0, 0)$ and scatters stray points across the forbidden white region.

TABLE I: Quantitative results on the 2-D Swiss Roll dataset ($N = 10{,}000$). All metrics averaged over 5,000 generated samples. ↓: lower is better; ↑: higher is better. **Bold**: best value in column.

Method	NFE ↓	W_2 ↓	SWD ↓
SD-OT (Ours)	**1**	0.0677	**0.0070**
Flow Matching	50	**0.0622**	0.0191
OT-Flow Matching	50	0.0694	0.0121
DDPM	1,000	0.0692	0.0247

TABLE II: Quantitative results on the 2-D Checkerboard dataset ($N = 10{,}000$). OT Cost is computed with the shared source noise pool via EMD (independent of each model's own sample pairing).

Method	NFE ↓	W_2 ↓	SWD ↓
SD-OT (Ours)	**1**	**0.0702**	**0.0105**
Flow Matching	50	0.0726	0.0286
OT-Flow Matching	50	0.0785	0.0191
DDPM	1,000	0.0907	0.0462

All methods achieve 8/8 mode coverage, confirming that Checkerboard does not induce mode dropping under our training budget. However, the *uniformity* of coverage differs substantially: SD-OT distributes samples within $\pm 0.5\%$ of the $\frac{1}{8}$ target across all eight cells, while DDPM exhibits a systematic mode imbalance. SD-OT's SWD advantage is most

pronounced on this dataset, outperforming all baselines.

C. Group A: 2D Visualization

We embed the full 70,000 MNIST image dataset into 2D using UMAP [34], then apply each transport method to map the embeddings toward a uniform distribution on $[0, 1]^2$. An autoencoder (AE) and a variational autoencoder (VAE) are also trained for comparison. The results are visualized as scatter plots (or power diagrams) colored by digit class.

a) Setup: All methods except for the AE and VAE receive the same UMAP embedding as input. Fig. 4 shows the 2D embeddings and all mapping results. The UMAP embedding reveals the characteristic cluster structure of MNIST with gaps between digit classes. The AE and VAE produce tightly clustered latent codes with large empty regions and blurred class boundaries.

b) Mapping methods: The flow-based methods (NF, FM, CNF) are diffeomorphic and preserve the inter-cluster gaps. For DDPM, we apply the forward diffusion process, which progressively adds noise, then map the resulting distribution to $[0, 1]^2$ via the empirical CDF. OT-EMD finds the exact discrete assignment; OT-SWD minimizes an approximate OT loss via random 1D projections but provides no structural coverage guarantee. Only the semi-discrete OT power diagram partitions the entire unit square into cells, every point in $[0, 1]^2$ maps to a valid digit with no gaps.

D. Group B: 100-dim Generation Quality

a) Generated sample comparison: Figs. 8 to 15 show generated image grids from each method across all four datasets. AE-OT produces sharp, recognizable samples, while all other methods produce blurry or distorted images and illustrate strong examples of mode collapse and mode mixture.

b) Setup: We evaluate generation quality using Fréchet Inception Distance (FID) [35], which measures the distributional similarity between real and generated images. We use four datasets: MNIST, Fashion-MNIST, CIFAR-10, and CelebA. All methods except VAE share the same autoencoder backbone (InfoGAN architecture [36] following Lucic et al. [37]), trained with MSE reconstruction loss and L1 latent regularization at a latent dimension of 100.

This shared-backbone design isolates the transport method as the sole experimental variable. We note that this architecture is not optimized for any particular dataset; the goal is to evaluate coverage and mode collapse across methods rather than to maximize reconstruction fidelity. Additionally, we note that OT-EMD is intractable at higher dimensions due to its $O(N^3)$ computational cost, further motivating the use of the semi-discrete OT approach.

The eight methods compared are:

- **AE**: Sample from a Gaussian fitted to the latent codes.
- **AE-OT**: Semi-discrete OT via Monte Carlo sampling.
- **OT-SWD**: Sliced Wasserstein Distance, which approximates the transport cost by projecting distributions onto random 1D slices and solving independently per slice.
- **NF**: RealNVP normalizing flow trained on latent codes.
- **FM**: Flow matching with learned velocity field.
- **CNF**: Continuous normalizing flow.
- **DDPM**: Denoising diffusion, reverse process generates latent codes from noise.
- **VAE**: Variational autoencoder with its own encoder/decoder.

For each method, we sweep the learning rate over $\{10^{-4}, 5\times10^{-4}, 10^{-3}, 2\times10^{-3}\}$ and select the configuration with the lowest final loss. The AE-OT solver uses a fixed step size of 0.05 following An et al. [6]. Convergence curves for all methods across each dataset are shown in Figs. 16 to 19. FID is computed using 10k generated vs. 10k real images with the TF-ported InceptionV3 weights [35] via `pytorch-fid`.

c) Results: Table III reports FID scores across all datasets. AE-OT achieves the best FID on every dataset, often by a substantial margin. On MNIST, AE-OT attains an FID of 4.4, more than $13\times$ lower than the next-best non-OT method (VAE at 23.5). This pattern is consistent across all four datasets.

The flow-based methods (NF, FM, CNF) yield FID scores ranging from 45.7 to 102.9 when applied to latent codes, indicating that diffeomorphic maps systematically fail to cover the latent distribution. DDPM performs comparably (40.9–111.1), as the reverse denoising process must reconstruct the latent manifold structure from noise without coverage guarantees. OT-SWD, despite using an optimal transport loss (Sliced Wasserstein Distance), performs comparably to these methods rather than to AE-OT. This suggests that minimizing a transport cost alone is insufficient; the structural guarantee of the power diagram partition, which assigns every point in the target domain to a source cell, is what distinguishes semi-discrete OT from approximate OT methods. The AE baseline (sampling from a fitted Gaussian) performs poorly across all datasets. The VAE achieves moderate FID on MNIST (23.5) and Fashion-MNIST (42.4), but degrades on CIFAR-10 (110.4), likely reflecting architectural limitations rather than a fundamental advantage of KL regularization. On CelebA, the gap between methods narrows: AE-OT achieves 24.2 compared to 32.2 for NF, but OT still holds a clear lead. The narrowing is expected, CelebA's latent space is more smoothly distributed so the coverage problem is less severe than for discrete-class datasets like MNIST.

VI. Discussion

A diffeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$ preserves connected components and cannot merge disconnected clusters into a single connected region. When the latent space has k well-separated clusters, T^{-1} must map them to k disconnected regions of the prior, leaving the complement unmapped. Any sample from these gaps produces an out-of-distribution latent code leading to mode collapse and nonsensical generated samples. The 2D visualizations in Section V-C make this failure mode directly visible, and the FID results in Section V-D confirm that it persists at higher dimensions.

The semi-discrete OT map is discontinuous at power cell boundaries. This discontinuity is precisely what enables the

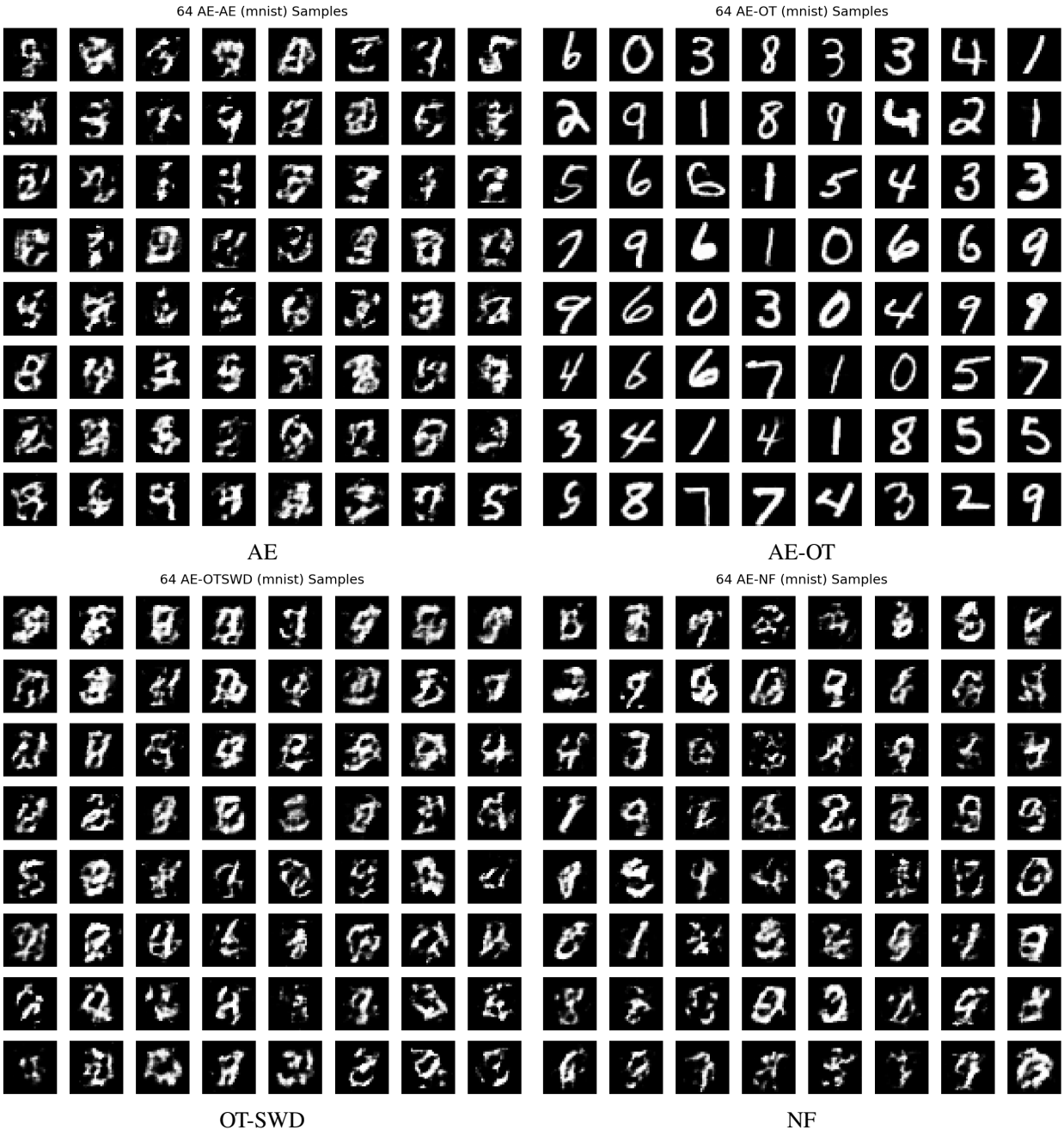


Fig. 8: MNIST generated samples (100-dim latent).

map to cut the prior into N pieces and assign each piece to a source point, regardless of the topology of the source distribution. The power diagram tiles the entire target domain with no overlaps and no gaps, so every sample from the prior decodes into a meaningful image.

Figalli's regularity theory [32] predicts that when the target distribution has multiple modes, the Brenier potential u_h develops ridges, points where it is continuous but not differentiable, and the gradient map (the transport map) is discontinuous across these ridges. The projection of these ridges onto the source domain forms the singular set Σ. Conventional generators based on DNNs cannot represent discontinuous maps, so they either collapse modes or blend across Σ, producing spurious inter-mode samples (mode mixture).

The AE-OT framework avoids this by never asking a neural network to represent the transport map. Instead, the discontinuous map is computed explicitly via the power diagram, and the singular set is detected geometrically through angles between adjacent cells (Algorithm 3). Samples near Σ are discarded rather than interpolated, preventing mode mixture while preserving full mode coverage.

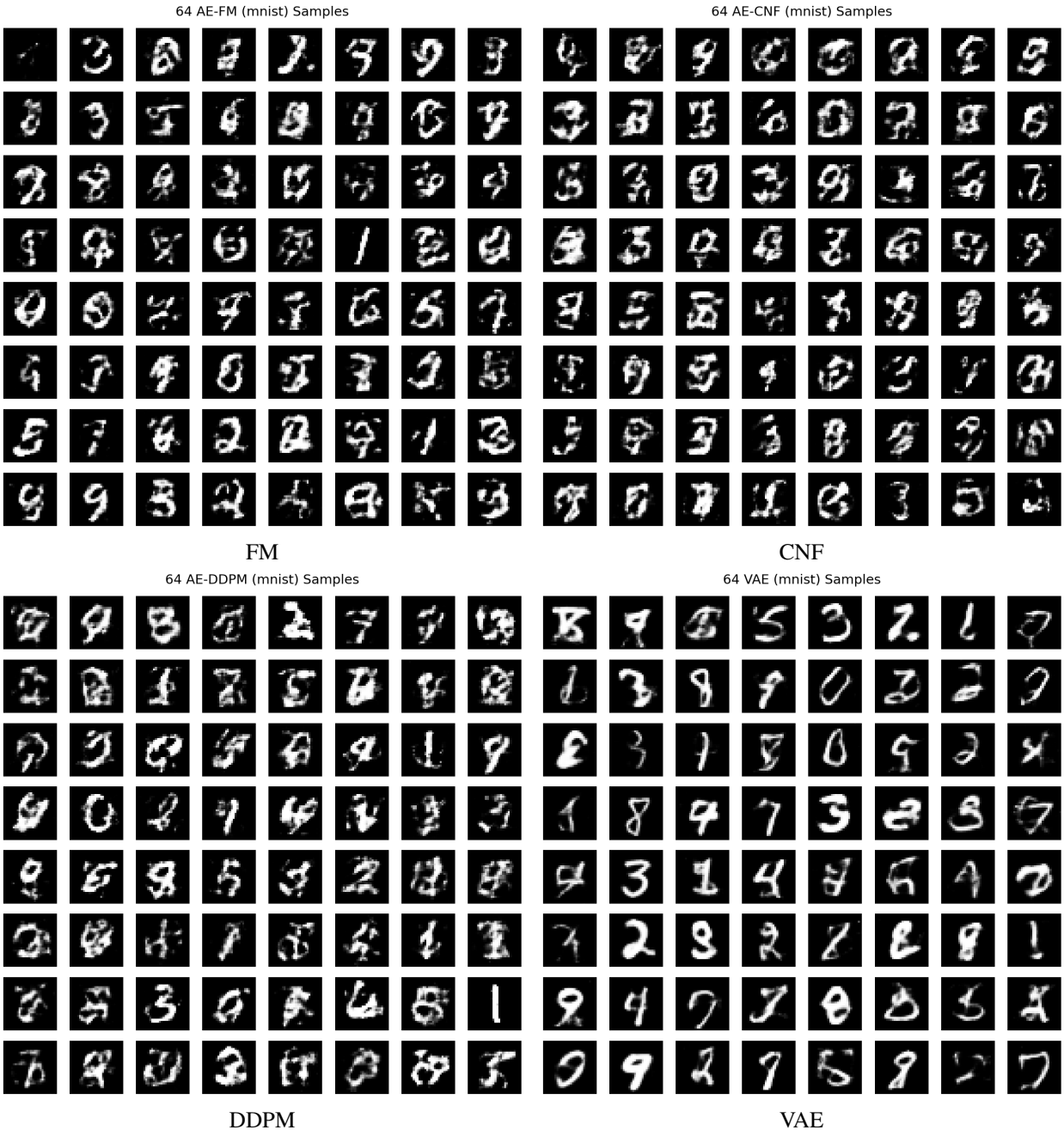


Fig. 9: MNIST generated samples (100-dim latent).

VII. Conclusion

We have presented a systematic comparison of semi-discrete optimal transport against seven alternative generative methods: three flow-based (Normalizing Flows, Continuous Normalizing Flows, Flow Matching), Denoising Diffusion, Sliced Wasserstein OT, an Autoencoder baseline, and a Variational Autoencoder, all evaluated under controlled conditions with a shared autoencoder backbone. Our 2D visualization experiments provide direct visual evidence that flow-based methods preserve the topological gaps inherent in latent space cluster structure, while DDPM destroys manifold structure through its forward noising process. Semi-discrete OT eliminates these gaps by construction through its power diagram partition. Quantitative FID evaluation across four standard datasets confirms that AE-OT consistently achieves the best generation quality, outperforming all baselines. Notably, approximate OT methods such as Sliced Wasserstein perform no better than flow-based or diffusion methods, highlighting that the full-coverage guarantee of the power diagram is the critical factor. These results demonstrate that Monte Carlo volume estimation makes semi-discrete OT practical in high-dimensional latent

Fig. 10: Fashion-MNIST generated samples (100-dim latent).

spaces, enabling competitive generation quality through guaranteed full coverage of the target domain.

REFERENCES

[1] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real-NVP," in *International Conference on Learning Representations (ICLR)*, 2017.

[2] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[3] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in *International Conference on Learning Representations (ICLR)*, 2023.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[5] X. Gu, F. Luo, J. Sun, and S.-T. Yau, "Variational principles for Minkowski type problems, discrete optimal transport, and discrete Monge–Ampère equations," *Asian Journal of Mathematics*, vol. 20, no. 2, pp. 383–398, 2016.

[6] D. An, Y. Guo, N. Lei, F. Luo, S.-T. Yau, and X. Gu, "AE-OT: A new generative model based on extended semi-discrete optimal transport," in *International Conference on Learning Representations (ICLR)*, 2020.

[7] R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, L. Gautheron, N. T. Gayraud, H. Janati, A. Rakotomamonjy, I. Redko, A. Rolet, A. Schutz, V. Seguy, D. J. Sutherland, R. Tavenard, A. Tong, and T. Vayer, "Pot: Python optimal transport," *Journal of Machine*

Fig. 11: Fashion-MNIST generated samples (100-dim latent).

Learning Research, vol. 22, no. 78, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-451.html

[8] V. Oliker, "Hypersurfaces in $\mathbb{R}^{n+1}$ with prescribed gaussian curvature and related equations of monge–ampère type," *Communications in Partial Differential Equations*, vol. 9, pp. 807–838, 1984.

[9] H. Minkowski, "Allgemeine lehrsätze über die konvexen polyeder," *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen*, pp. 198–219, 1897.

[10] A. D. Alexandrov, "On the theory of mixed volumes of convex bodies," *Matematicheskii Sbornik*, vol. 2, pp. 947–972, 1938.

[11] A. V. Pogorelov, *The Minkowski Multidimensional Problem*. Wiley, 1978.

[12] L. Nirenberg, "The weyl and minkowski problems in differential geometry in the large," *Communications on Pure and Applied Mathematics*, vol. 6, pp. 337–394, 1953.

[13] S.-Y. Cheng and S.-T. Yau, "On the regularity of the solution of the n-dimensional minkowski problem," *Communications on Pure and Applied Mathematics*, vol. 29, pp. 495–516, 1976.

[14] R. Schneider, *Convex Bodies: The Brunn–Minkowski Theory*, 2nd ed. Cambridge University Press, 2014.

[15] P. Guan and J. Li, "The christoffel–minkowski problem i: convexity of solutions of a hessian equation," *Inventiones Mathematicae*, vol. 151, pp. 553–577, 2009.

[16] Y. Brenier, "Polar factorization and monotone rearrangement of vector-valued functions," *Communications on Pure and Applied Mathematics*, vol. 44, no. 4, pp. 375–417, 1991.

[17] A. D. Alexandrov, *Convex Polyhedra*. Springer, 1958.

[18] L. A. Caffarelli, "Interior $w^{2,p}$ estimates for solutions of monge–ampère

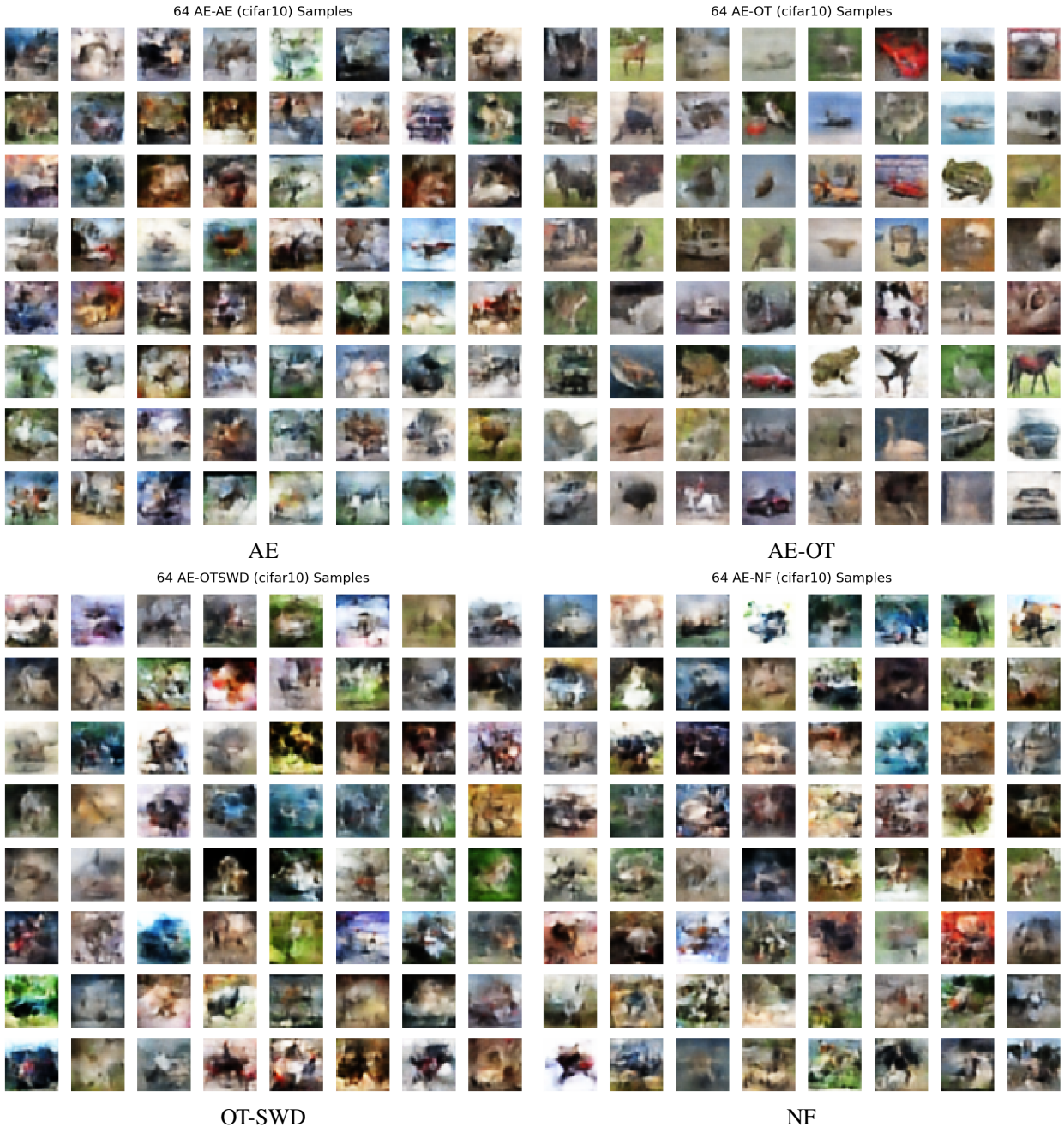


Fig. 12: CIFAR-10 generated samples (100-dim latent).

equations," *Annals of Mathematics*, vol. 131, pp. 135–150, 1990.

[19] ——, "The regularity of mappings with a convex potential," *Journal of the American Mathematical Society*, vol. 5, pp. 99–104, 1992.

[20] ——, "Boundary regularity of maps with convex potentials," *Communications on Pure and Applied Mathematics*, vol. 45, pp. 1141–1151, 1992.

[21] X.-N. Ma, N. S. Trudinger, and X.-J. Wang, "Regularity of potential functions of the optimal transportation problem," *Archive for Rational Mechanics and Analysis*, vol. 177, pp. 151–183, 2005.

[22] G. Loeper, "On the regularity of solutions of optimal transportation problems," *Acta Mathematica*, vol. 202, pp. 241–283, 2009.

[23] C. Villani, *Optimal Transport: Old and New*. Springer, 2009.

[24] A. Figalli, *The Monge–Ampère Equation and Its Applications*. European Mathematical Society, 2017.

[25] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[26] W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud, "FFJORD: Free-form continuous dynamics for scalable reversible generative models," in *International Conference on Learning Representations (ICLR)*, 2019.

[27] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations (ICLR)*, 2021.

[28] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

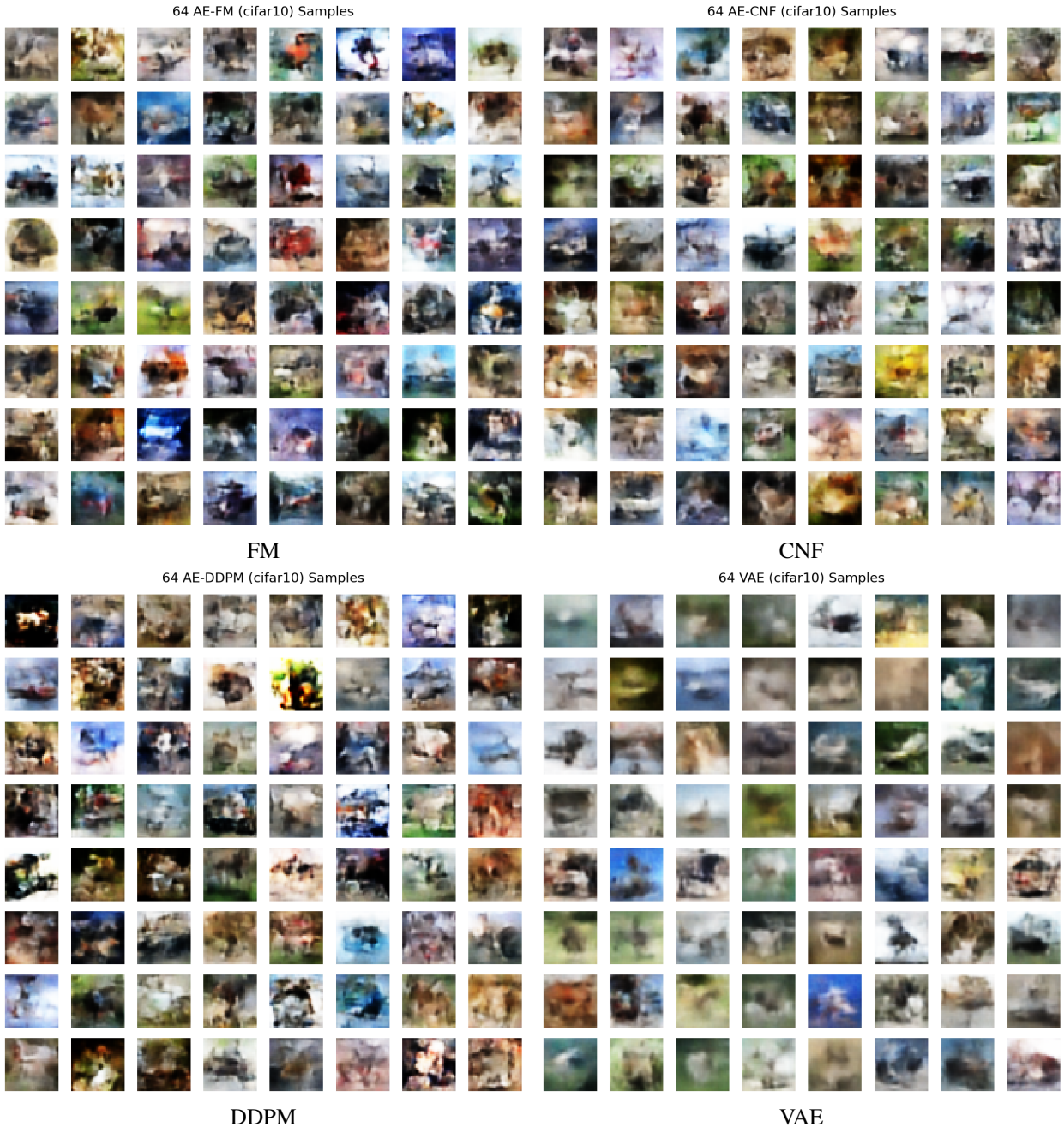


Fig. 13: CIFAR-10 generated samples (100-dim latent).

[29] A. V. Makkuva, A. Taghvaei, S. Oh, and J. Lee, "Optimal transport mapping via input convex neural networks," in *International Conference on Machine Learning (ICML)*, 2020.

[30] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister, "Sliced and Radon Wasserstein barycenters of measures," vol. 51, no. 1, 2015, pp. 22–45.

[31] Q. Mérigot, "A multiscale approach to optimal transport," *Computer Graphics Forum*, vol. 30, no. 5, pp. 1583–1592, 2011.

[32] A. Figalli, "Regularity properties of optimal maps between nonconvex domains in the plane," *Communications in Partial Differential Equations*, vol. 35, no. 3, pp. 465–479, 2010.

[33] A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio, "Improving and generalizing flow-matching with minibatch optimal transport," *Transactions on Machine Learning Research*, 2024, expert Certification. [Online]. Available: https://openreview.net/forum?id=CD9Snc73AW

[34] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[36] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[37] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are GANs created equal? A large-scale study," in *Advances in Neural*

Fig. 14: CelebA generated samples (100-dim latent).

Information Processing Systems (NeurIPS), 2018.

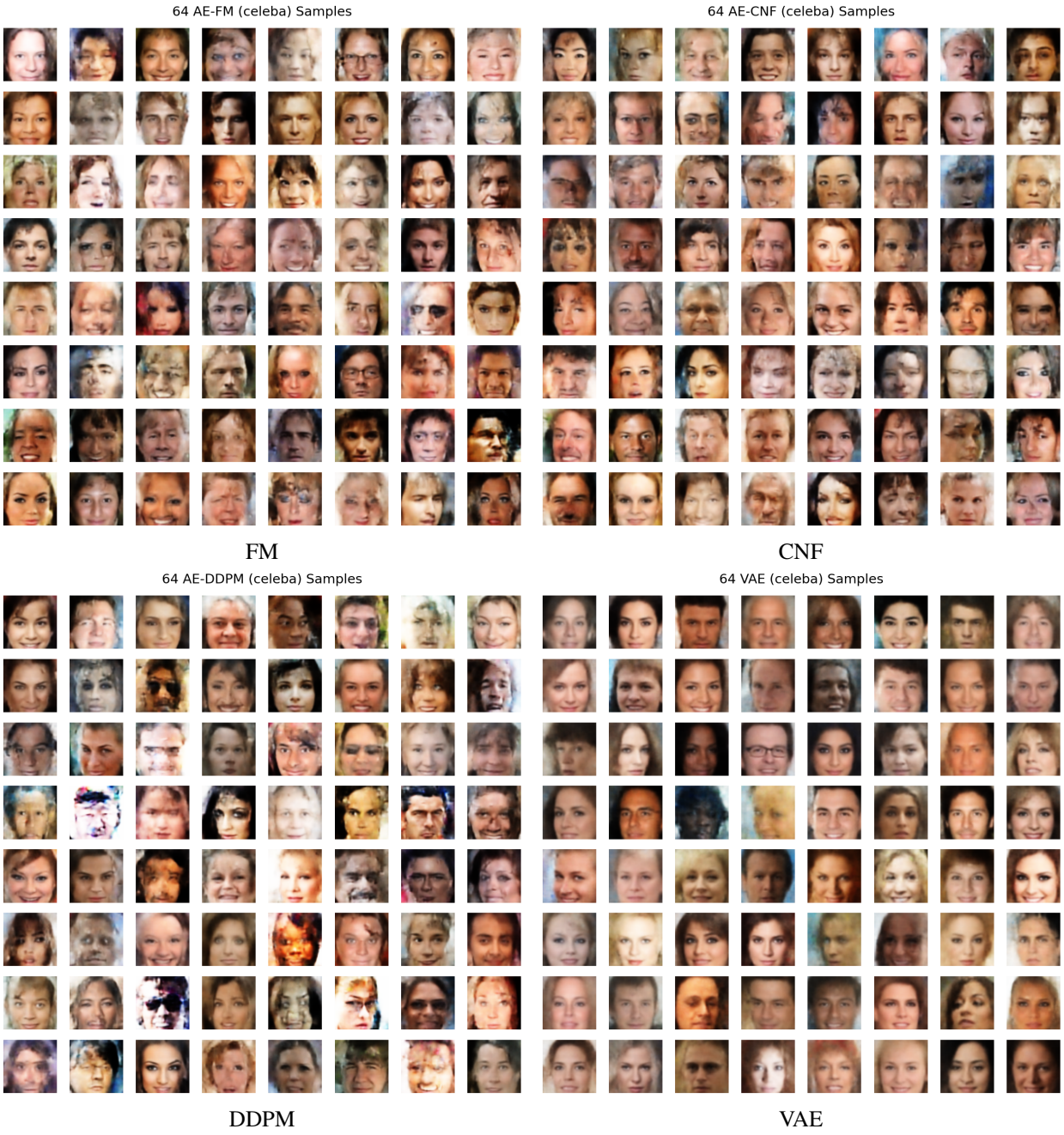
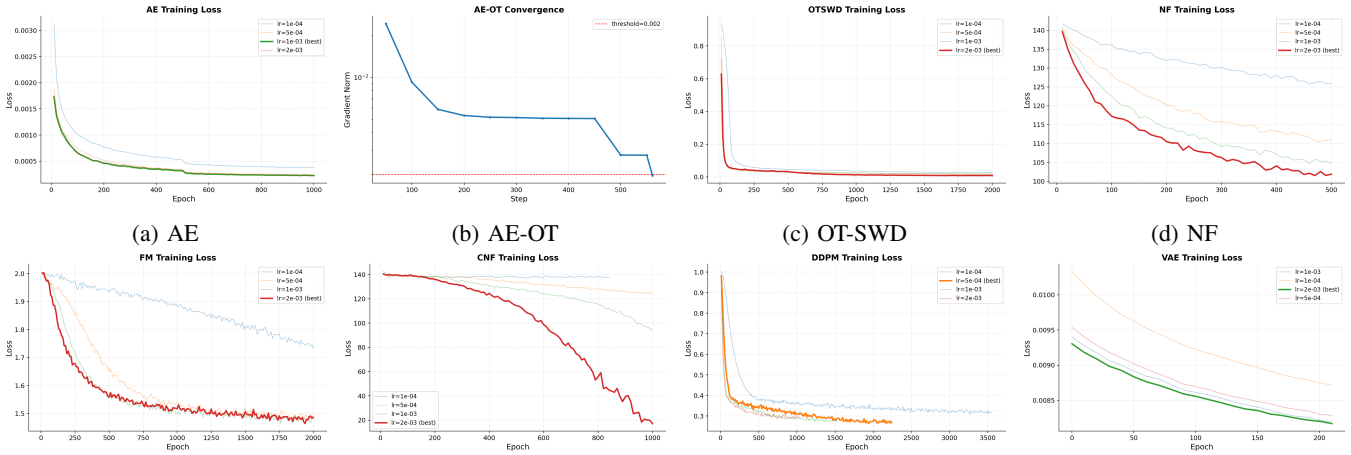


Fig. 15: CelebA generated samples (100-dim latent).

TABLE III: FID scores (↓) on four datasets with 100-dim latent space. All AE-* methods share the same autoencoder backbone per dataset. Best result per dataset in **bold**.

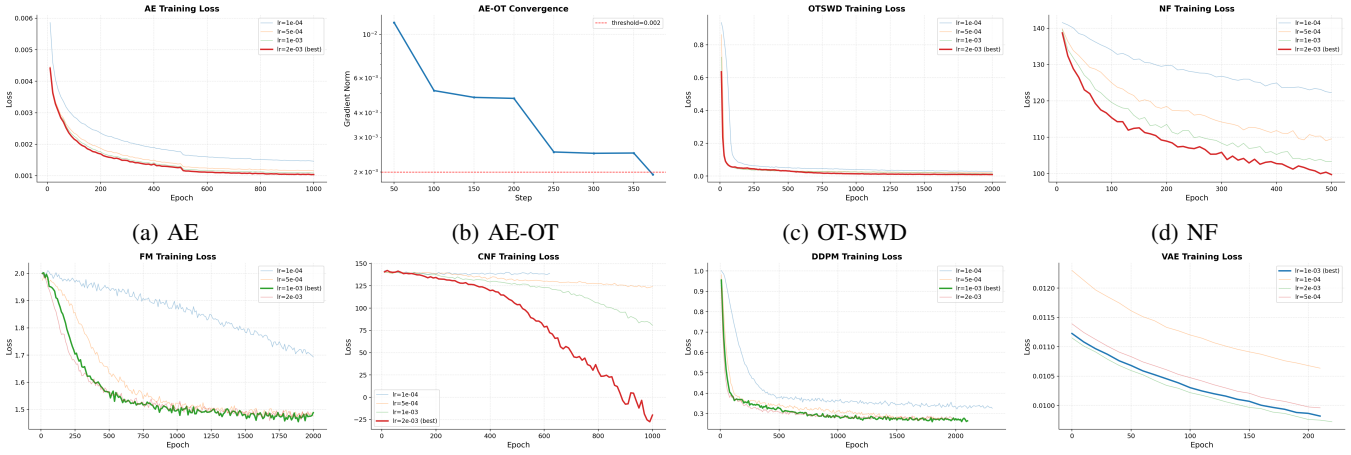
Dataset	AE	AE-OT	OT-SWD	NF	FM	CNF	DDPM	VAE
MNIST	124.5	**4.4**	115.2	58.8	79.2	102.9	111.1	23.5
F-MNIST	101.4	**11.3**	90.4	45.7	57.4	96.7	79.2	42.4
CIFAR-10	91.7	**61.8**	92.6	86.6	89.4	91.3	87.6	110.4
CelebA	40.4	**24.2**	41.0	32.2	34.3	67.0	40.9	36.0



(a) AE (b) AE-OT (c) OT-SWD (d) NF

(e) FM (f) CNF (g) DDPM (h) VAE

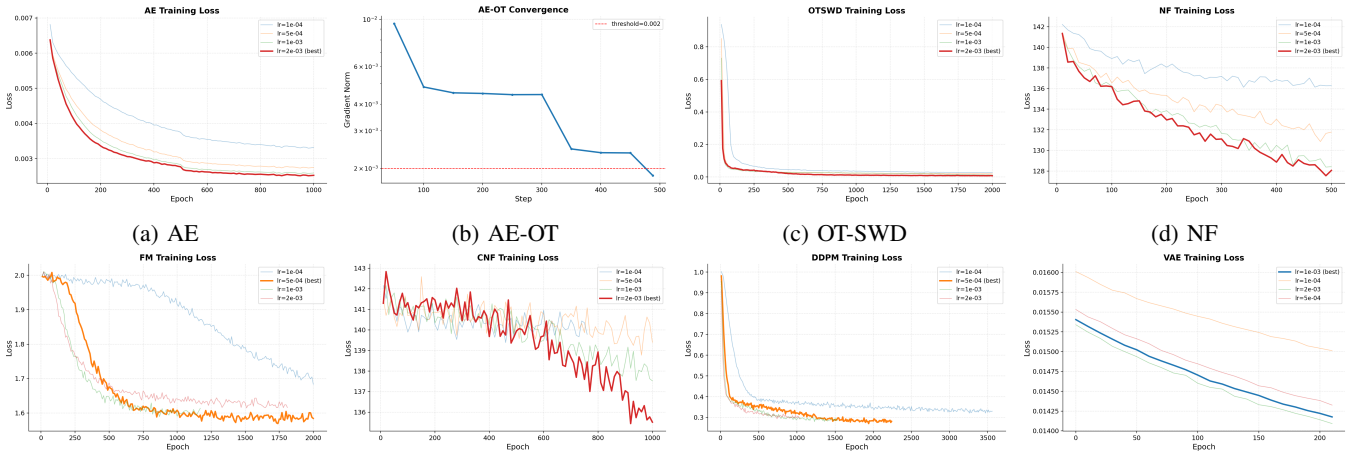
Fig. 16: MNIST convergence curves. Methods (a,c–h) show training loss vs. epoch for each learning rate; the best LR is highlighted. (b) AE-OT shows gradient norm vs. optimization step.



(a) AE (b) AE-OT (c) OT-SWD (d) NF

(e) FM (f) CNF (g) DDPM (h) VAE

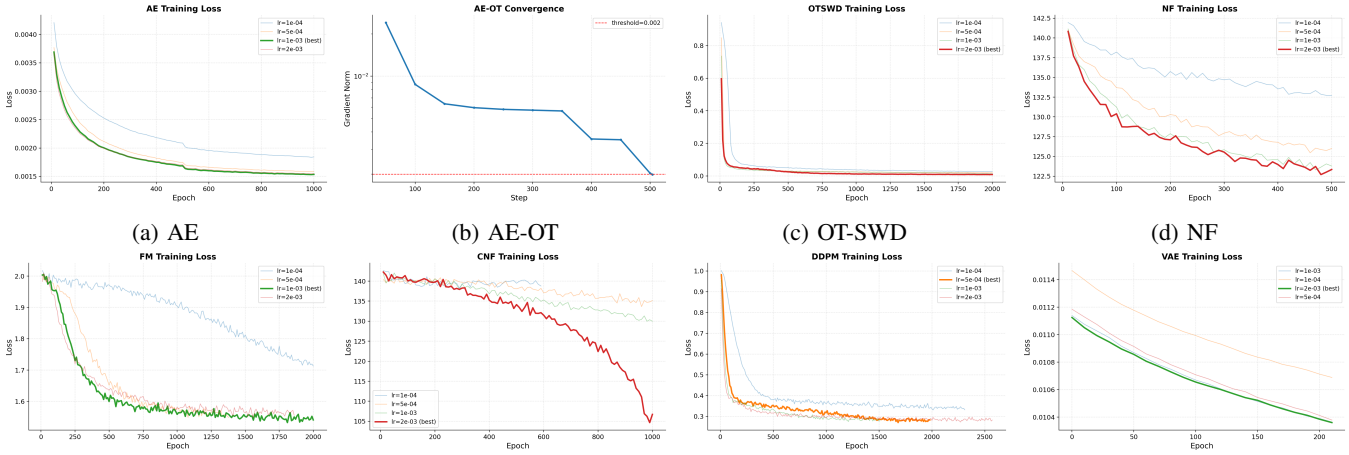
Fig. 17: Fashion-MNIST convergence curves. Methods (a,c–h) show training loss vs. epoch for each learning rate; the best LR is highlighted. (b) AE-OT shows gradient norm vs. optimization step.



(a) AE (b) AE-OT (c) OT-SWD (d) NF

(e) FM (f) CNF (g) DDPM (h) VAE

Fig. 18: CIFAR-10 convergence curves. Methods (a,c–h) show training loss vs. epoch for each learning rate; the best LR is highlighted. (b) AE-OT shows gradient norm vs. optimization step.



(a) AE (b) AE-OT (c) OT-SWD (d) NF

(e) FM (f) CNF (g) DDPM (h) VAE

Fig. 19: CelebA convergence curves. Methods (a,c–h) show training loss vs. epoch for each learning rate; the best LR is highlighted. (b) AE-OT shows gradient norm vs. optimization step.