ONE-STEP GENERATIVE MODELING VIA WASSERSTEIN GRADIENT FLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models and flow-based methods have shown impressive generative capability, especially for images, but their sampling is expensive because it requires many iterative updates. We introduce *W-Flow*, a framework for training a generator that transforms samples from a simple reference distribution into samples from a target data distribution in a single step. This is achieved in two steps: we first define an evolution from the reference distribution to the target distribution through a Wasserstein gradient flow that minimizes an energy functional; second, we train a static neural generator to compress this evolution into one-step generation. We instantiate the energy functional with the Sinkhorn divergence, which yields an efficient optimal-transport-based update rule that captures global distributional discrepancy and improves coverage of the target distribution. We further prove that the finite-sample training dynamics converge to the continuous-time distributional dynamics under suitable assumptions. Empirically, W-Flow sets a new state of the art for one-step ImageNet 256×256 generation, achieving 1.29 FID, with improved mode coverage and domain transfer. Compared to multi-step diffusion models with similar FID scores, our method yields approximately 100× faster sampling. These results show that Wasserstein gradient flows provide a principled and effective foundation for fast and high-fidelity generative modeling.

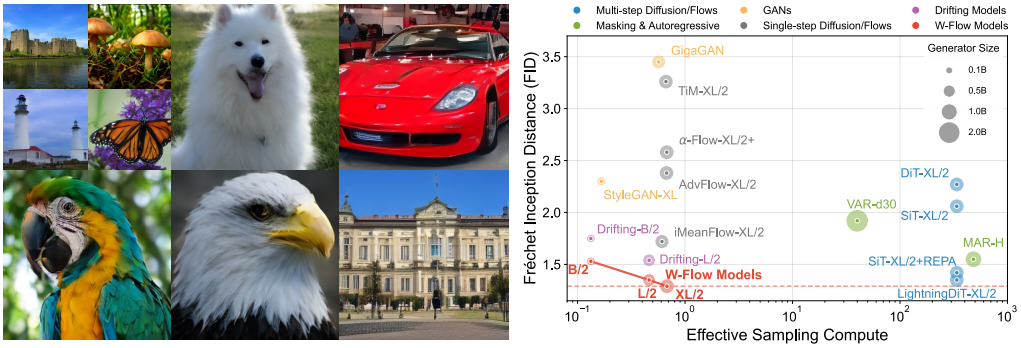


Figure 1: *(Left)* **1-NFE samples** from W-Flow-L/2 trained from scratch on ImageNet-256×256. *(Right)* Sample quality (measured by FID) vs. effective sampling compute (Lu & Song, 2024) (billion parameters × number of function evaluations during sampling) evaluated on ImageNet 256×256.

1 INTRODUCTION

Generative modeling can be viewed as transporting a simple reference distribution (often referred to as a prior) to a complex target distribution (e.g., data distribution). In modern high-dimensional settings, this transport is rarely learned *from scratch* as a single global map. Instead, the dominant paradigm constructs it through a sequence of simpler updates. Diffusion models (Sohl-Dickstein et al., 2015; Song et al., 2020; Song & Ermon, 2019; Ho et al., 2020) learn a sequence of 'infinitesimal' denoising steps and then unroll them over *many* inference steps. Flow models (Lipman et al., 2022; Liu et al., 2022) learn a sequence of velocity fields and likewise unroll the dynamics over many steps. Despite high sample quality, multi-step inference leads to significant latency and computational cost.

It would be convenient to use an iterative procedure *during training* to specify the evolution of the generated data distribution, and then compress this evolution into a static map. In this way, the resulting generator would directly transform an input from the reference distribution into an output from the target distribution in one step. This would combine the efficiency of one-step generation with the flexibility of a distributional evolution during training. The fundamental question would thus be:

What principles should govern the training dynamics of one-step generators?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches fall short of providing a satisfactory answer. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Radford et al., 2015; Arjovsky et al., 2017) evolve the model via discriminator-induced gradients, yet inherit the well-known instability of minimax optimization. More recently, Drifting Model (Deng et al., 2026) achieves strong empirical performance by iteratively displacing particles[1] via handcrafted attractive and repulsive fields. Yet, the underlying dynamics remain heuristic and lack a clear and principled interpretation, leaving the model vulnerable to unpredictable convergence and mode collapse (see Fig. 6).

In this paper, we propose W-Flow, a new framework for one-step generative modeling. W-Flow prescribes how the *entire model distribution* should evolve toward the target distribution. This evolution is given by a Wasserstein gradient flow (WGF): at each training step, the current generated distribution moves along the steepest descent direction of a chosen energy functional to the target distribution. We instantiate this energy functional with the Sinkhorn divergence, an optimal-transport-based distance that can be estimated efficiently from mini-batches. The resulting Sinkhorn update combines two batch-level transport plans: a generated-to-real transport plan and a generated-to-generated self-transport plan. This makes each update depend on the batch-level transport structure, rather than on a single matched sample or nearest neighbor. We then train a generator to imitate these updates, thereby compressing the multi-step training evolution into a single one-step map at inference time. This fundamentally differs from few/one-step diffusion methods (Song et al., 2023; Zhou et al., 2025), which typically distill or follow a path originally designed for iterative reverse diffusion.

Notably, W-Flow achieves a new state of the art for one-step generation on ImageNet 256×256, obtaining 1.29 FID at XL scale and 1.35 at L scale, significantly outperforming Deng et al. (2026) while also surpassing many multi-step diffusion and flow models (see Fig. 1). Beyond class-conditional generation, experiments on FFHQ further demonstrate improved mode coverage and strong domain transfer capability, enabled by the globally coordinated OT dynamics. Overall, W-Flow establishes a new framework for designing the training dynamics of generators, showing that WGFs can serve as a powerful foundation for fast and high-fidelity generative modeling.

2 Related Work

(Few/one-step) diffusion and flow-based models. Diffusion (Sohl-Dickstein et al., 2015; Song et al., 2020; Ho et al., 2020) and flow models (Lipman et al., 2022; Liu et al., 2022) learn to transport a simple reference distribution to a complex target distribution through a forward noising process and its time-reverse. However, they require many sequential steps for sampling, leading to substantial computational cost and latency. Many works distill multi-step teachers into few/one-step models via variational score distillation (Yin et al., 2024b;a; Wang et al., 2023; Salimans et al., 2024) or learning shortcuts along the diffusion ODE specified by the teacher (Song et al., 2023; Song & Dhariwal, 2023; Lu & Song, 2024; Salimans & Ho, 2022). Efforts have also been made to directly learn a few/one-step generator from scratch, typically by enforcing certain self-consistency conditions on the trajectory (Geng et al., 2025a;b; Boffi et al., 2025; Song et al., 2023) or the intermediate marginals (Zhou et al., 2025). These methods largely inherit their training signal from a predefined diffusion transport path, which is originally designed for iterative generation. We instead directly define the training-time evolution of a one-step generator via WGFs.

One-step generator mappings. An alternative is to directly learn a static map that takes an input from the reference distribution and outputs a sample from the target distribution. The most studied is the family of GANs (Goodfellow et al., 2014; Arjovsky et al., 2017; Radford et al., 2015; Genevay et al., 2018), which train the generator using discriminator-driven gradients in a minimax game. However, adversarial training is often unstable and requires considerable effort to stabilize. Recently, Drifting Model (Deng et al., 2026) guides the generator via an empirically constructed drifting field using kernel interactions and achieves strong performance (Li et al., 2026; Gao et al., 2026). Yet, the design of this technique remains largely heuristic, lacking a principled understanding of the induced dynamics. Subsequent works (Lai et al., 2026; Turan & Ovsjanikov, 2026) connect Drifting Model to score-based formulations under Gaussian kernels, while He et al. (2026); Cao et al. (2026) explore links to Sinkhorn divergence and WGFs. Concurrent Sinkhorn-Drifting (He et al., 2026) shares our population Sinkhorn velocity and stop-gradient regression objective. Our distinctions lie

[1] A particle is one candidate sample/image as it evolves through the algorithm.

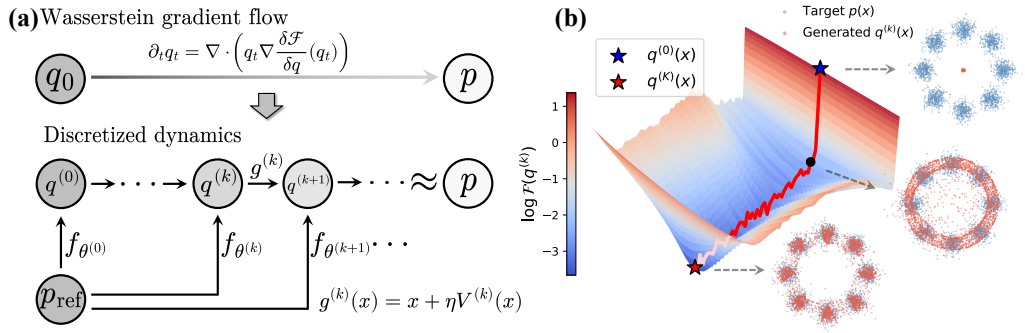


Figure 2: **(a)** The conceptual diagram of W-Flow. **(b)** Visualization of the training dynamics projected onto the Sinkhorn divergence landscape on 8 Gaussian mixtures, shown on a logarithmic scale.

in independent two-batch self-transport estimation, velocity-level guidance, finite-horizon particle-trajectory analysis, and ImageNet-scale validation; see Appendix A.5.

Optimal transport in generative models. Existing applications of OT in generative models broadly fall into two categories. The first uses OT-inspired discrepancies as the training objectives. Wasserstein GANs (Arjovsky et al., 2017; Gulrajani et al., 2017; Tolstikhin et al., 2017) optimize a dual form of the Wasserstein distance via an adversarial critic, while Sinkhorn GANs (Genevay et al., 2018) leverage OT-based divergence computed by Sinkhorn iterations. The second employs OT to guide dynamic transport paths, including Schrödinger bridges (Alouadi et al., 2026; De Bortoli et al., 2021; Ma et al., 2025) and OT-inspired flow models (Albergo & Vanden-Eijnden, 2022; Lipman et al., 2022; Liu et al., 2022; Lin et al., 2025). Yet, these approaches use OT either as a generator loss or to construct a transport path, while we instead use WGF velocities to supervise a static map.

Wasserstein gradient flows in generative models. WGFs provide a principled description of the evolution of probability measures. A standard discretization is the Jordan–Kinderlehrer–Otto (JKO) scheme (Jordan et al., 1998), which has inspired several recent approaches and theoretical frameworks (Mokrov et al., 2021; Choi et al., 2024; Fan et al., 2021; Xie & Cheng, 2025). However, JKO-based methods require solving a complex inner optimization problem at each step, often involving adversarial training over dual potentials, which makes them difficult to scale in practice. Neural Sinkhorn Gradient Flow (NSGF) (Zhu et al., 2024) also adopts the population Sinkhorn-divergence velocity under quadratic cost. However, NSGF learns a time-dependent velocity field from minibatch flow trajectories and integrates it at inference, while our W-Flow instead learns a static one-step map directly from local transport targets. NSGF also establishes weak mean-field convergence, whereas our analysis gives a quantitative finite-horizon $\mathcal{W}_2$ bound for the empirical Euler dynamics.

3 Method

We introduce W-Flow, a deterministic *one-step* generator with its training dynamics guided by a WGF. The main idea is to separate the design of training dynamics from the parameterization of the generator: we first prescribe how the model distribution should evolve toward the data distribution during training, and then learn a neural network to realize this evolution through particle updates.

3.1 W-Flow: Learning one-step transport map under Wasserstein gradient flow

One-step transport maps. Given a reference distribution $p_{\text{ref}} \in \mathcal{P}(\mathbb{R}^m)$ and a target distribution $p \in \mathcal{P}(\mathbb{R}^n)$, the goal is to learn a mapping $f : \mathbb{R}^m \to \mathbb{R}^n$ such that if an input z is sampled from the reference distribution, the output $f(z)$ follows the target distribution p. In distribution space, this is denoted by the pushforward operator $f_\#$, meaning we aim to learn a map f such that $f_\# p_{\text{ref}} \approx p$. In practice, p is only observed through a finite dataset of samples.

We adopt an incremental construction of f, which breaks this complex transport into a sequence of simpler steps. Specifically, let $f^{(0)} : \mathbb{R}^m \to \mathbb{R}^n$ be an initial map and set $q^{(0)} = f^{(0)}_\# p_{\text{ref}}$. We then build the generator through a sequence of local transport maps $\{g^{(k)}\}_{k=0}^{K-1} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying

$$q^{(k+1)} = g^{(k)}_\# q^{(k)}, \qquad f^{(k+1)} = g^{(k)} \circ f^{(k)} = g^{(k)} \circ \cdots \circ g^{(0)} \circ f^{(0)}, \tag{1}$$

so that the final map $f^{(K)}$ pushes p_{ref} to $q^{(K)} \approx p$ (see Fig. 2(a)). Explicitly, $f^{(K)}$ is a *static* map obtained by composing stepwise transport maps $\{g^{(k)}\}_{k=0}^{K-1}$ and the initial $f^{(0)}$. The role of this sequence is *not* to define a standalone multi-step generative model, but rather to provide an incremental procedure for constructing the target one-step map $f^{(K)}$ used at inference time.

Velocity field. We parameterize the maps $g^{(k)}$ to reflect incremental *displacements* between iterations:

$$g^{(k)}(x) \coloneqq x + \eta V^{(k)}(x), \tag{2}$$

where $V^{(k)} : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous velocity field, also referred to as a drifting field (Deng et al., 2026), and $\eta > 0$ is the step size. Under this parameterization, each iteration moves a data sample $x \sim q^{(k)}$ only locally, while the composition of these maps yields a flexible global transport, and the field $V^{(k)}$ determines how the intermediate distributions evolve.

In principle, the recursion in (1) would make the generator an increasingly long composition, so exact sampling from $f^{(k)}_{\#} p_{\text{ref}}$ becomes prohibitively costly when k is large. To keep the model one-step throughout training, we represent the generator by a neural network and learn $f_{\theta^{(k+1)}}$ to match the updated map $f^{(k+1)}$. Using $f_{\theta^{(k)}}$ as a surrogate for $f^{(k)}$, we optimize

$$\mathcal{L}^{(k)}(\theta) \coloneqq \mathbb{E}_{z\sim p_{\text{ref}}} \left\| f_\theta(z) - g^{(k)} \circ \tilde{f}^{(k)}(z) \right\|^2, \qquad \tilde{f}^{(k)}(z) \coloneqq \text{sg}(f_{\theta^{(k)}}(z)), \tag{3}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The parameter $\theta^{(k+1)}$ is then obtained by taking a gradient step in (3). At $\theta = \theta^{(k)}$, the loss scale is $\mathbb{E}_{z\sim p_{\text{ref}}} \left\| \eta V^{(k)}(f_{\theta^{(k)}}(z)) \right\|^2$, hence $\mathcal{L}^{(k)} = 0$ implies $V^{(k)} = 0$.

A central challenge, however, is to design the vector field $V^{(k)}$ to guide the iteration in (1) towards the target distribution. Existing approaches, such as Deng et al. (2026), are largely heuristic and guarantee only that the target distribution is a stationary point of the dynamics: $V^{(k)} = 0$ if $q^{(k)} = p$. However, the converse generally fails. In particular, during training, observing that $V^{(k)} \to 0$ does not imply that $q^{(k)} \to p$. Moreover, the intermediate distributional dynamics $\{q^{(k)}\}$ induced by such velocity fields remain largely uncharacterized, and convergence guarantees are generally unavailable.

Continuous-time Wasserstein velocity field. To obtain a principled and well-behaved velocity field $V^{(k)}$ in Eq. (2), we take a step back and first specify the desired evolution of the model distribution in continuous time, and then derive the corresponding discrete updates. Specifically, we model the evolution of $\{q^{(k)}\}$ via a WGF. Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^n) \to \mathbb{R}$ be an energy functional defined on probability distributions. The associated WGF is given by

$$\partial_t q_t = \nabla \cdot \left(q_t \nabla \frac{\delta \mathcal{F}}{\delta q}(q_t) \right), \tag{4}$$

where $\frac{\delta \mathcal{F}}{\delta q}(q_t)$ denotes the first variation of $\mathcal{F}$ at q_t. Eq. (4) defines a *continuous* evolution of distributions $q_t \in \mathcal{P}(\mathbb{R}^n)$ that decreases the energy $\mathcal{F}$. Equivalently, this evolution can be expressed through the continuity equation with a specific form of the velocity field

$$\partial_t q_t + \nabla \cdot (q_t V_t) = 0, \quad V_t(x) = -\nabla \frac{\delta \mathcal{F}}{\delta q}(q_t)(x). \tag{5}$$

This characterization provides a clear variational interpretation: the distribution q_t evolves along the *steepest descent* direction of $\mathcal{F}$ in Wasserstein space (see Fig. 2(b) for an illustration). To drive the dynamics toward the target distribution p, we choose the energy functional $\mathcal{F}(q) = \mathcal{D}(q \,\|\, p)$ where $\mathcal{D}$ is a suitable divergence such that $p = \arg\min_q \mathcal{F}(q)$. Under this choice, the induced velocity field in (5) transports q_t to progressively reduce its discrepancy from p.

Time discretization. Discretizing the continuous-time evolution provides a concrete prescription for the stepwise transport maps introduced earlier. Applying an explicit Euler scheme to the WGF over time steps $\{t^{(k)} : t^{(k)} = k\eta,\ k \in \mathbb{N}_{\geq 0}\}$ yields

$$g^{(k)}(x) \coloneqq x + \eta V^{(k)}(x), \qquad V^{(k)}(x) = -\nabla \frac{\delta \mathcal{F}}{\delta q}(q^{(k)})(x), \tag{6}$$

where $q^{(k)} = q_{t^{(k)}}$. This recovers exactly the stepwise velocity-based update in Eq. (2), now with the velocity field $V^{(k)}$ specified by the underlying gradient flow. Each map $g^{(k)}$ operates directly on individual data points in $\mathbb{R}^n$. By moving these points, the resulting empirical distribution approximates the continuous transport trajectory in distribution space, while composing these local updates yields a map $f^{(k)}$ that approximates the flow at time $t^{(k)}$. In sum, the gradient flow defines the *training-time dynamics* in distribution space, whereas the learned model is ultimately a *static map* obtained by compressing these local transports into a single generator. The diagram is provided in Fig. 2 (a).

3.2 Instantiation via Sinkhorn divergence

Choices of the energy functional $\mathcal{F}$. Common choices include Maximum Mean Discrepancy (MMD) (Li et al., 2015) and KL divergence (Table 1). However, both of them face significant limitations in practice. MMD induces kernel-based interactions whose gradients can vanish when q_t is far from p. KL requires approximating the generally intractable score $\nabla \log q_t(x)$, introducing additional bias and instability. Our ablations in Table 2a confirm that both choices lead to inferior performance. We adopt the Sinkhorn divergence, an OT-based discrepancy that avoids explicit score estimation and is directly computable on samples. While the convergence of WGFs to a global minimizer is not automatic, a recent result in Hardion & Lacombe (2026) shows that, under Gaussian distributional assumptions, the WGF of the Sinkhorn divergence converges to its global minimizer.

Table 1: WGF energies and induced velocity fields. Derivations in Appendix A.1.

Energy Functional $\mathcal{F}(\rho_t)$	Continuous Velocity Field $V_t(x)$
$\frac{1}{2}\mathrm{MMD}^2(q_t, p)$	$\int \nabla_x k(x,y)p(dy) - \int \nabla_x k(x,y)q_t(dy)$
$D_{\mathrm{KL}}(q_t\|p)$	$\nabla \log p(x) - \nabla \log q_t(x)$
$S_\varepsilon(q_t, p)$	$T^\varepsilon_{q_t,p}(x) - T^\varepsilon_{q_t,q_t}(x)$ (Eq. 10)

Entropic optimal transport and Sinkhorn divergence. A natural candidate is the entropically regularized OT cost, which is computationally attractive due to Sinkhorn iterations. However, it is biased: in general, $\mathrm{OT}_\varepsilon(p,p) \neq 0$, so $\mathrm{OT}_\varepsilon(\cdot, p)$ does not define an energy whose minimizer is exactly the target distribution. This motivates the use of the *debiased* Sinkhorn divergence. Let $\mathcal{P}_2(\mathbb{R}^n)$ denote probability measures with finite second moment and $\Pi(q,p)$ the set of couplings of $q, p \in \mathcal{P}_2(\mathbb{R}^n)$. The entropic optimal transport (EOT) functional under quadratic cost is

$$\mathrm{OT}_\varepsilon(q,p) := \inf_{\pi \in \Pi(q,p)} \left\{ \int \tfrac{1}{2}\|x-y\|^2 \, \pi(dx,dy) + \varepsilon \, D_{\mathrm{KL}}(\pi \,\|\, q \otimes p) \right\}, \tag{7}$$

where $\varepsilon > 0$ is the entropic regularization parameter. The associated Sinkhorn divergence is given by

$$S_\varepsilon(q,p) = \mathrm{OT}_\varepsilon(q,p) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(q,q) - \tfrac{1}{2}\mathrm{OT}_\varepsilon(p,p). \tag{8}$$

Velocity field induced by Sinkhorn divergence. Let $\pi^\varepsilon_{q,p} \in \Pi(q,p)$ and $\pi^\varepsilon_{q,q} \in \Pi(q,q)$ be the optimal couplings for $\mathrm{OT}_\varepsilon(q,p)$ and $\mathrm{OT}_\varepsilon(q,q)$. With the conditional distributions in $\pi^\varepsilon_{q,p}(dx,dy) = q(dx)\,\pi^\varepsilon_{q,p}(dy \mid x)$ and $\pi^\varepsilon_{q,q}(dx,dx') = q(dx)\,\pi^\varepsilon_{q,q}(dx' \mid x)$, we define the barycentric projections

$$T^\varepsilon_{q,p}(x) := \int y \, \pi^\varepsilon_{q,p}(dy \mid x), \qquad T^\varepsilon_{q,q}(x) := \int x' \, \pi^\varepsilon_{q,q}(dx' \mid x). \tag{9}$$

It is well known (see, *e.g.*, Zhu et al. (2024); Genevay et al. (2018); Peyré & Cuturi (2019); Nutz & Wiesel (2022) and the proof in Appendix A.1) that the velocity field coincides with the Wasserstein gradient direction of the Sinkhorn divergence

$$V^\varepsilon_{q,p}(x) = -\nabla \frac{\delta S_\varepsilon(\cdot,p)}{\delta q}(q)(x) = T^\varepsilon_{q,p}(x) - T^\varepsilon_{q,q}(x). \tag{10}$$

The velocity field in Eq. (10) admits a natural interpretation as a balance between *attraction* through $T^\varepsilon_{q,p}(x)$ and *self-interaction* through $T^\varepsilon_{q,q}(x)$ that enforces consistency by a global mass-conserving coupling. For smooth, strictly positive densities on a connected compact domain, $V^\varepsilon_{q,p} \equiv 0$ if and only if $q = p$ (Zhu et al., 2024; He et al., 2026). In this population setting, the only equilibrium is the target distribution, unlike local drifting dynamics that can stall at some $q \neq p$ (see Sec. 4.3).

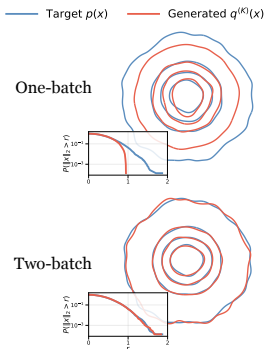


Figure 3: One-batch and two-batch estimators learning a 2D Gaussian.

3.3 Interactive particle dynamics

In practice, we operate on finite mini-batches. Given N i.i.d. samples $\{x_i\}_{i=1}^N \in \mathbb{R}^n$ from q_t and M i.i.d. samples $\{y_j\}_{j=1}^M \in \mathbb{R}^n$ from p, let the weighted empirical measures be $\widehat{q}_t = \sum_{i=1}^N a_i \, \delta_{x_i}$, $\widehat{p} = \sum_{j=1}^M b_j \, \delta_{y_j}$, with $a_i, b_j > 0$, and $\sum_i a_i = \sum_j b_j = 1$. In this discrete setting, the general definition in Eq. (7) reduces to the finite-dimensional problem over the coupling matrix $\pi \in \Pi(\widehat{q}_t, \widehat{p})$:

$$\mathrm{OT}_\varepsilon\big(\widehat{q}_t, \widehat{p}\big) := \min_{\pi \in \Pi(\widehat{q}_t, \widehat{p})} \sum_{i,j} \frac{1}{2}\|x_i - y_j\|^2 \pi_{ij} + \varepsilon \sum_{i,j} \pi_{ij}(\log \pi_{ij} - 1), \tag{11}$$

where $\Pi(\widehat{q}_t, \widehat{p})$ is the set of matrices with prescribed marginals. Denote the optimal solution $\pi^{\varepsilon,*}_{\widehat{q}_t,\widehat{p}}$.

Algorithm 1 W-Flow Training

Input: Generator f_θ, reference distribution p_{ref}, target distribution p, step size η, regularization ε, iterations L

repeat
  Sample input $\{z_i\}_{i=1}^N \sim p_{\text{ref}}$, $\{z'_l\}_{l=1}^N \sim p_{\text{ref}}$
  Sample target $\{y_j\}_{j=1}^M \sim p$
  $x_i \leftarrow f_\theta(z_i)$, $x'_l \leftarrow \text{sg}(f_\theta(z'_l))$
  $\{T^\varepsilon_{\widehat{q}_\theta,\widehat{p}}(x_i)\} \leftarrow$ SINKHORN$(\{x_i\},\{y_j\},\varepsilon,L)$
  $\{T^\varepsilon_{\widehat{q}_\theta,\widehat{q}'_\theta}(x_i)\} \leftarrow$ SINKHORN$(\{x_i\},\{x'_l\},\varepsilon,L)$
  $V^\varepsilon(x_i) \leftarrow T^\varepsilon_{\widehat{q}_\theta,\widehat{p}}(x_i) - T^\varepsilon_{\widehat{q}_\theta,\widehat{q}'_\theta}(x_i)$
  $\tilde{x}_i \leftarrow \text{sg}(x_i + \eta V^\varepsilon(x_i))$
  Take a gradient descent step on $\nabla_\theta \frac{1}{N}\sum_{i=1}^N \|f_\theta(z_i) - \tilde{x}_i\|^2$
until converged

Algorithm 2 Sinkhorn Projection (SINKHORN)

Input: Source particles $\{x_i\}_{i=1}^N$, target particles $\{y_j\}_{j=1}^M$, regularization ε, Sinkhorn iterations L

$a \leftarrow \frac{1}{N}\mathbf{1}_N$, $b \leftarrow \frac{1}{M}\mathbf{1}_M$ ▷ Uniform weights
$C_{ij} \leftarrow \frac{1}{2}\|x_i - y_j\|^2$ ▷ Cost matrix
$K_{ij} = \exp(-C_{ij}/\varepsilon)$ ▷ Gibbs kernel
$v^{(0)} \leftarrow \mathbf{1}_M$ ▷ Initialization
for $\ell = 0,\dots,L-1$ **do** ▷ Sinkhorn-Knopp
  $u^{(\ell+1)} \leftarrow a \oslash (Kv^{(\ell)})$
  $v^{(\ell+1)} \leftarrow b \oslash (K^\top u^{(\ell+1)})$
end for
$\Pi^\varepsilon \leftarrow \text{diag}(u^{(L)})\, K\, \text{diag}(v^{(L)})$ ▷ Coupling
$T^\varepsilon(x_i) = \frac{1}{a_i}\sum_{j=1}^M \Pi^\varepsilon_{ij} y_j$ ▷ Projection
return $\{T^\varepsilon(x_i)\}_{i=1}^N$

Two-batch estimate for self-transport. Naïvely estimating the self-entropic OT term $\text{OT}_\varepsilon(\widehat{q}_t,\widehat{q}_t)$ from a single empirical batch introduces a self-matching artifact: since each particle can be matched to itself at zero transport cost, the resulting coupling tends to overemphasize the diagonal entries π_{ii}. To mitigate it, we adopt a *two-batch* strategy. We draw an independent second batch $\{x'_l\}_{l=1}^N$ from q_t and define the empirical measure $\widehat{q}'_t = \sum_{l=1}^N a_l \delta_{x'_l}$. We then compute the coupling between $\widehat{q}_t$ and $\widehat{q}'_t$, yielding the optimal plan $\pi^{\varepsilon,*}_{\widehat{q}_t,\widehat{q}'_t}$. This construction removes the need for the diagonal masking heuristic used in Deng et al. (2026) and admits a cleaner theoretical interpretation. Fig. 3 compares the two-batch strategy with the naïve one-batch estimate on learning a 2D Gaussian. The one-batch strategy fails to capture the Gaussian tails, while the two-batch strategy recovers the target faithfully. See more illustrations in Appendix A.2. The discrete counterparts of the barycentric maps in (9) are

$$T^\varepsilon_{\widehat{q}_t,\widehat{p}}(x_i) = \frac{1}{a_i}\sum_j (\pi^{\varepsilon,*}_{\widehat{q}_t,\widehat{p}})_{ij} y_j, \quad T^\varepsilon_{\widehat{q}_t,\widehat{q}'_t}(x_i) = \frac{1}{a_i}\sum_l (\pi^{\varepsilon,*}_{\widehat{q}_t,\widehat{q}'_t})_{il} x'_l, \tag{12}$$

where in practice the optimal couplings $\pi^{\varepsilon,*}_{\widehat{q}_t,\widehat{p}}$ and $\pi^{\varepsilon,*}_{\widehat{q}_t,\widehat{q}'_t}$ can be efficiently approximated using the Sinkhorn-Knopp algorithm (Sinkhorn & Knopp, 1967). We present the subroutine in Alg. 2. With these, the particle-based estimate of the velocity field in Eq. (10) is

$$V^\varepsilon_{\widehat{q}_t,\widehat{p}}(x_i)^2 = T^\varepsilon_{\widehat{q}_t,\widehat{p}}(x_i) - T^\varepsilon_{\widehat{q}_t,\widehat{q}'_t}(x_i). \tag{13}$$

Convergence of the particle dynamics. The following theorem shows that the finite-particle, discrete-time transport mechanism used in our method is a consistent approximation of the intended population-level WGF. The convergence result holds for more general velocity fields, but we present below using Sinkhorn velocity. The full statement, assumptions, and proof are deferred to Appendix B.

Theorem 3.1 (Informal). *Let $\widehat{p}^M$ be the empirical measure based on M i.i.d. samples $\{y_j\}_{j=1}^M$ from the target measure p. The particles $\{x_i^{(k)}\}_{i=1}^N$ evolve as $x_i^{(k+1)} = x_i^{(k)} + \eta\, V^\varepsilon_{\widehat{q}_k^N,\widehat{p}^M}(x_i^{(k)})$, where $\widehat{q}_k^N$ denotes the empirical measure of $\{x_i^{(k)}\}_{i=1}^N$. Define the continuous-time interpolation as $x_i^\eta(t) := x_i^{(k)} + (t-t_k)\, V_{\widehat{q}_k^N,\widehat{p}^M}(x_i^{(k)})$, $t \in [t_k, t_{k+1})$, with the empirical measure $\widehat{q}_t^{N,M,\eta} := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^\eta(t)}$. Assume suitable regularity conditions (specified in Assumption B.1). Then for every $T > 0$, there exists a unique $q \in C([0,T];\mathcal{P}_2(\mathbb{R}^d))$ that solves the continuity equation $\partial_t q_t + \nabla\cdot(q_t\, V_{q_t,p}) = 0$ in the weak sense. Moreover, as $\eta \to 0$ and $N, M \to \infty$, it holds that $\sup_{t\in[0,T]} \mathcal{W}_2(\widehat{q}_t^{N,M,\eta}, q_t) \to 0$.*

3.4 Implementation

Training. Empirically, we construct the training loss by learning the generator f_θ to simulate the discrete particle dynamics driven by the velocity in Eq. (13). This can be viewed as a finite-sample counterpart of the population-level objective in (3), where expectations with respect to $q^{(k)}$ and p are replaced by empirical measures. Specifically, at each training step, we optimize

$$\mathcal{L}_{\text{W-Flow}}(\theta) = \frac{1}{N}\sum_{i=1}^N \left\| x_i - \text{sg}\big(x_i + \eta\, V^\varepsilon_{\widehat{q}_\theta,\widehat{p}}(x_i)\big)\right\|^2, \quad x_i = f_\theta(z_i), \tag{14}$$

[2]Here we omit the second-batch $\widehat{q}'$ in the subscript of V for brevity.

where $\{z_i\}_{i=1}^N$ are i.i.d. samples from p_{ref}. Crucially, the stop-gradient operator $\text{sg}(\cdot)$ precisely aligns with the discrete WGF by treating the regression target as fixed, which also avoids differentiation through the OT plan and stabilizes optimization. We summarize the training procedure in Alg. 1.

Classifier-free guidance. We bake classifier-free guidance (CFG) (Ho & Salimans, 2022) into f_θ at *training time*, using a small batch of unconditional reference particles to introduce an extra repulsive effect. We introduce guidance directly at the *velocity field level*. Specifically, we define

$$\tilde{V}^{\varepsilon,w}_{q_\theta,p}(x_i) := \Big(T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|c)}(x_i) - T^{\varepsilon}_{q_\theta(\cdot|c),q_\theta(\cdot|c)}(x_i)\Big) + w\underline{\Big(T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|c)}(x_i) - T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|\varnothing)}(x_i)\Big)}, \tag{15}$$

where $T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|\varnothing)}$ denotes the Sinkhorn barycentric projection (Alg. 2) between $q_\theta(\cdot|c)$ and the unconditional data distribution $p(\cdot|\varnothing)$. This construction injects guidance via the underlined term that computes the difference between the conditional and unconditional velocity fields (see Appendix A.3). We refer to Eq. (15) as *velocity guidance*, in contrast to the *distribution guidance* of Deng et al. (2026); see Appendix A.3 for the comparison. Although the effective target distribution induced by Eq. (15) is generally not available in closed form, it admits a natural geometric interpretation through a barycentric construction. Moreover, we show in Appendix A.3 that, when the KL divergence is used as $\mathcal{F}$, this formulation exactly recovers the standard exponentially tilted target. This provides a conceptual justification for adopting the same design under the Sinkhorn divergence, whose empirical benefits are demonstrated in Sec. 4.1.

Sinkhorn divergence in feature space. Rather than evaluating the Sinkhorn divergence directly in the target space $\mathbb{R}^n$, we may lift particles into a feature space through $\phi : \mathbb{R}^n \to \mathbb{R}^d$. The velocity field and training loss are then computed using the embedded particles $\{\phi(x_i)\}, \{\phi(x'_l)\}, \{\phi(y_j)\}$, while gradients are backpropagated through ϕ to update the generator. Following Deng et al. (2026), we use multiple feature maps and sum the corresponding losses.

4 Experiments

Experiment setup. We conduct our main experiments on ImageNet (Deng et al., 2009) generation with 256×256 resolution. We directly adopt the DiT-style (Peebles & Xie, 2023) generator architecture in Deng et al. (2026). We also use the officially released pretrained MAE feature extractors from Deng et al. (2026). More details and the full set of hyperparameters are deferred to Appendix C.

4.1 Ablation Study

Ablation setting. To enable an informative comparison in a controlled setting, we follow the default ablation setup of Deng et al. (2026). All experiments are conducted in the SD-VAE (Rombach et al., 2022) latent space, using a B/2 DiT generator with the same latent-MAE feature encoder. The generator is trained for 100 epochs, and FID (Heusel et al., 2017) is evaluated on 50K generated images. We set $\varepsilon = 0.05$ by default with additional detailed hyperparameters provided in Appendix C.

Energy functionals in gradient flow. We compare three different energy functionals: squared MMD, KL divergence, and Sinkhorn divergence (Eq. (8)). Results in Table 2a confirm that Sinkhorn divergence performs much better than MMD and KL due to the OT-driven dynamics. Notably, it also significantly outperforms the heuristic design of Drifting Model when both use distribution guidance.

Classifier-free guidance. We further equip the Sinkhorn W-Flow with our proposed velocity-guidance CFG (denoted as `velo.`). As reported in Table 2b, this guidance mechanism is fully compatible with Sinkhorn W-Flow and further improves the FID to 7.08, outperforming distribution guidance (`dist.`). We adopt it as the default configuration in the subsequent studies.

Entropy regularization. Table 2b shows that W-Flow is fairly robust to ε, with $\varepsilon = 0.05$ performing best among the values considered. Detailed choices for the main experiments are in Table 6.

The cost function in OT. We compare our default choice of quadratic cost $c(x, y) = \frac{1}{2}\|x - y\|^2$ with the ℓ_2 distance cost $c(x, y) = \|x - y\|$. The results in Table 2c demonstrate that the quadratic cost yields superior performance, in line with our theoretical assumptions in Eq. (7).

Velocity field estimation. We validate the two-batch estimator in Sec. 3.3 for the self-transport map. As shown in Table 2c, estimating $T^{\varepsilon}_{\widehat{q}_\theta,\widehat{q}_\theta}$ from a single batch leads to a substantially worse FID of 17.57,

Divergence	FID, 1-NFE
Drifting	8.46
MMD	10.40
KL	10.17
Sinkhorn	**7.29**

(a) **Choice of the divergence.** Sinkhorn achieves the lowest FID.

ε	CFG	FID, 1-NFE
0.05	dist.	7.29
0.05	velo.	**7.08**
0.025	velo.	7.11
0.075	velo.	7.26

(b) **CFG and ε.** Velocity guidance with $\varepsilon = 0.05$ performs best.

Cost	Mask	2-batch	FID, 1-NFE
$\|\cdot\|$	-	✓	7.16
$\frac{1}{2}\|\cdot\|^2$	-	✓	**7.08**
$\frac{1}{2}\|\cdot\|^2$	-	-	17.57
$\frac{1}{2}\|\cdot\|^2$	✓	-	7.45

(c) **OT cost and V_t estimation.** 2-batch quadratic cost is most effective.

Table 2: **Ablation study on ImageNet 256×256 generation.** We use a B/2 backbone with 100-epoch training from scratch. The FID on 50K images is reported. The default settings are marked in gray.

Table 4: **Class-conditional generation on ImageNet 256×256.** FID is computed on 50K images. All methods are reported with CFG if applicable. "#Params" denotes the number of parameters of "generator + decoder". All generators are trained *from scratch* without distilling from a teacher model.

	#Params	NFE	FID↓	IS↑
Multi-step Diffusion/Flows				
ADM-G (Dhariwal & Nichol, 2021)	554M	250×2	4.59	186.7
DiT-XL/2 (Peebles & Xie, 2023)	675M+49M	250×2	2.27	278.2
SiT-XL/2 (Ma et al., 2024)	675M+49M	250×2	2.06	270.3
SiT-XL/2+REPA (Yu et al., 2024)	675M+49M	250×2	1.42	305.7
LightningDiT-XL/2 (Yao et al., 2025)	675M+70M	250×2	1.35	295.3
RAE+DiT^{DH}-XL/2 (Zheng et al., 2025)	839M+415M	50×2	**1.13**	262.6
Masking & Autoregressive				
MaskGIT (Chang et al., 2022)	227M+31M	8	6.18	182.1
VAR-$d30$ (Tian et al., 2024)	2B+61M	10×2	1.92	323.1
MAR-H (Li et al., 2024)	943M+41M	256×2	1.55	303.7
GANs				
BigGAN (Brock et al., 2018)	112M	1	6.95	152.8
GigaGAN (Kang et al., 2023)	569M	1	3.45	225.5
StyleGAN-XL (Sauer et al., 2022)	166M	1	2.30	265.1
Single-step Diffusion/Flows				
iCT-XL/2 (Song & Dhariwal, 2023)	675M+49M	1	34.24	–
Shortcut-XL/2 (Frans et al., 2024)	675M+49M	1	10.60	–
MeanFlow-XL/2 (Geng et al., 2025a)	676M+49M	1	3.43	–
TiM-XL/2 (Wang et al., 2025)	664M+49M	1	3.26	210.3
α-Flow-XL/2+ (Zhang et al., 2025)	676M+49M	1	2.58	–
AdvFlow-XL/2 (Lin et al., 2025)	673M+49M	1	2.38	284.2
iMeanFlow-XL/2 (Geng et al., 2025b)	610M+49M	1	1.72	282.0
Drifting Models				
Drifting Model, B/2 (Deng et al., 2026)	133M+49M	1	1.75	263.2
Drifting Model, L/2 (Deng et al., 2026)	463M+49M	1	1.54	258.9
W-Flow, B/2	133M+49M	1	**1.52**	271.8
W-Flow, L/2	463M+49M	1	**1.35**	272.5
W-Flow, XL/2	679M+49M	1	**1.29**	265.4

due to the intrinsic self-matching bias of this estimator. Our two-batch strategy also outperforms the mask-diagonal heuristic used in Deng et al. (2026) while admitting a more principled interpretation.

Sample size. Table 3 studies the effect of sample size N and M when maintaining the same batch size. W-Flow shows remarkably stronger robustness to the reduction in sample size compared with Drifting Model.

Table 3: **Ablation on sample size.**

(N, M)	(8,8)	(16,16)	(32,32)	(64,64)
Drifting	11.82	10.16	9.32	8.46
W-Flow	8.92	7.73	7.31	**7.08**

Batch size. Table 8 in Appendix E separately ablates the effect of the data particle sample size M and the effective batch size, where W-Flow demonstrates stronger performance consistently.

4.2 Main Results on ImageNet

Setup. Putting together the key components validated in the ablation study, we scale up training to achieve state-of-the-art results on ImageNet 256×256 generation. We retain the generator architecture, feature encoder, and most hyperparameters of Deng et al. (2026) (see Appendix C for details).

Results. Table 4 compares W-Flow with state-of-the-art one-step generators and multi-step diffusion and flow models. W-Flow achieves an FID of 1.29 at XL scale and 1.35 at L scale, outperforming MeanFlow-based methods and Drifting Models. With only 133M generator parameters, W-Flow B/2 achieves an FID of 1.52, surpassing the 463M-parameter Drifting Model L/2. Furthermore, despite 1-NFE sampling, W-Flow outperforms most diffusion models requiring up to 250 steps, such as LightningDiT-XL/2. See Appendix D for a detailed inference throughput comparison. More image samples produced by W-Flow are provided in Appendix F.

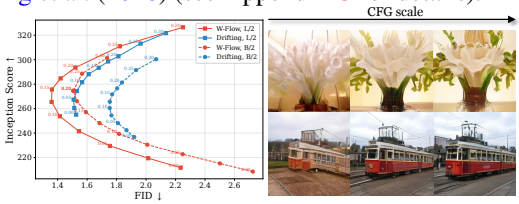


Figure 4: **Classifier-free guidance.** *Left:* FID and Inception Score with varying CFG scales. *Right:* Samples by W-Flow, L/2 with CFG 0.0 to 1.0.

Convergence speed. Table 5 shows that W-Flow converges substantially faster than Drifting Model: both B/2 and L/2 trained for only 384 epochs closely match the FID of Drifting Model trained for 1280 epochs.

Table 5: FID vs. epoch.

	Epoch	B/2	L/2
Drifting	1280	1.75	1.54
W-Flow	384	1.75	1.55
	640	1.63	1.46
	1280	1.52	1.35

The effect of CFG. Fig. 4 shows the FID and Inception Score with different CFG scales at inference time; W-Flow clearly achieves a better

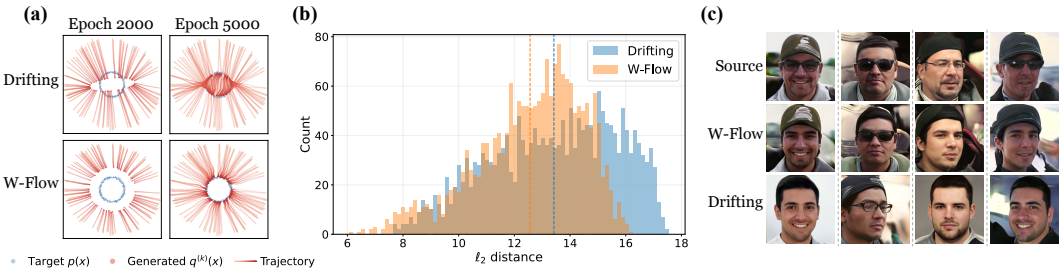


Figure 5: **(a) Oval-to-circle domain transfer.** Source and target are constructed by sampling angles uniformly from $[0, 2\pi)$ with parametric curves corrupted by Gaussian noise. **(b) & (c) One-step facial age translation on FFHQ, mapping older faces to younger ones.** **(b)** Histogram of the latent ℓ_2 distance between 2,000 source images and their generated targets. **(c)** Visual comparison.

tradeoff compared with Drifting. Samples in Fig. 4 demonstrate that increasing CFG leads to clearer class-specific details and higher perceptual quality, which aligns with the expected CFG behavior.

4.3 Domain Transfer and Mode Coverage

Domain transfer. Unlike standard generative models that typically map Gaussian noise to data, W-Flow can instead define a WGF between arbitrary distributions. This makes it naturally suited for domain transfer tasks where the source and target share the same semantic space. By setting p_{ref} to be the source data distribution, it learns a one-step domain-transfer map, *e.g.*, from senior to young-adult faces. To preserve source semantics, we use a residual map (He et al., 2016), $f_\theta(z) = z + \mathcal{N}_\theta(z)$, with a zero-initialized final layer, such that $f_{\theta_0}(z) = z$.

In a 2D oval-to-circle toy example in Fig. 5(a), Drifting Model shows erratic intermediate states and often moves particles far from their initial positions. In contrast, the OT-induced velocity in W-Flow produces coordinated, nearly straight, and much shorter trajectories toward the target.

We further evaluate one-step age translation on FFHQ (Karras et al., 2019), mapping senior faces (ages 55–100) to young-adult faces (ages 18–30) in the 512-dimensional latent space of a pretrained autoencoder (Pidhorskyi et al., 2020). We train a 4-layer MLP that maps source latent codes to the target domain. Fig. 5(b) and 5(c) show that W-Flow achieves shorter latent ℓ_2 transport distances and better preserves source identity.

Mode coverage. We provide a 2D Gaussian mixture experiment in Appendix A.4. We further extend this observation to image generation on FFHQ under an imbalanced training distribution, where target batches contain 95% senior faces (ages 55–100) and 5% child faces (ages 0–12). The PCA visualization in Fig. 6 shows that Drifting Model almost entirely drops the minority child mode, whereas W-Flow successfully captures it while maintaining high fidelity on the dominant senior mode. See Appendix F for representative samples. Furthermore, on ImageNet, Table 14 in Appendix E also shows modest recall gains over same-size Drifting Models, together with improved precision.

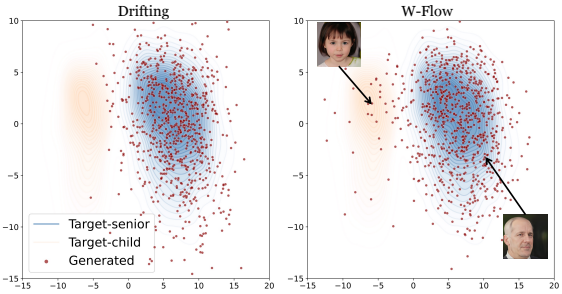


Figure 6: PCA scatter plot of generated latent codes on imbalanced FFHQ (95% senior faces, 5% child faces).

5 Conclusion

We introduced W-Flow for one-step generative modeling that guides the training dynamics of image generators by Wasserstein gradient flow. Our method derives the velocity field via the steepest descent of Sinkhorn divergence, yielding a globally coordinated OT-based update rule with consistent particle dynamics. Empirically, W-Flow sets a new state of the art for one-step ImageNet 256×256 generation with 1.29 FID. The results establish the WGF dynamics as a foundation for fast, high-fidelity generation. Future work includes exploring a broader class of energy functionals within the framework and extending this paradigm to more complex settings such as text-to-image generation.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing methods. We have not used generative AI tools to generate synthetic data sets, formulate mathematical claims, help develop theoretical models or conceptual frameworks, and design or provide feedback on research methodology or experiments. Additionally, we used generative AI tools to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We propose a framework for one-step generation from scratch via Wasserstein gradient flows. Our work aims to contribute to the advancement of the field of machine learning. We do not identify particular ethical concerns that are worth highlighting.

REPRODUCIBILITY STATEMENT

We fully commit to ensure reproducibility of this work. We have included the training algorithm in Alg. 1, the experiment setup and the detailed hyperparameter selection in Appendix C. We also provide the detailed theoretical proof in Appendix B and the code in the supplementary material,.

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022. 3

Alexandre Alouadi, Pierre Henry-Labordère, Grégoire Loeper, Othmane Mazhar, Huyên Pham, and Nizar Touzi. Lightsbb-m: Bridging Schrödinger and bass for generative diffusion modeling. *arXiv preprint arXiv:2601.19312*, 2026. 3

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Birkhäuser, 2008. 24

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017. 2, 3

Nicholas M Boffi, Michael S Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. *arXiv preprint arXiv:2505.18825*, 2025. 2

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 8

Jiarui Cao, Zixuan Wei, and Yuxin Liu. Gradient flow drifting: Generative modeling via Wasserstein gradient flows of KDE-approximated divergences. *arXiv preprint arXiv:2603.10592*, 2026. 2

Guillaume Carlier, Lénaïc Chizat, and Maxime Laborde. Displacement smoothness of entropic optimal transport. *ESAIM: Control, Optimisation and Calculus of Variations*, 30:25, 2024. 31

René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I*. Springer, 2018. 24

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022. 8

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable Wasserstein gradient flow for generative modeling through unbalanced optimal transport. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8629–8650. PMLR, 21–27 Jul 2024. 3, 23

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *arXiv preprint arXiv:2106.01357*, 2021. 3

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE, 2009. 7

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026. 2, 4, 6, 7, 8, 19, 20, 22, 23, 32, 35

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *NeurIPS*, 34:8780–8794, 2021. 8, 20

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. *arXiv preprint arXiv:2112.02424*, 2021. 3

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *arXiv preprint arXiv:2410.12557*, 2024. 8

Yuxuan Gao, Yedong Shen, Shiqi Zhang, Wenhao Yu, Yifan Duan, Jiajia Wu, Jiajun Deng, Yanyong Zhang, et al. Drift-based policy optimization: Native one-step policy learning for online robot control. *arXiv preprint arXiv:2604.03540*, 2026. 2

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018. 2, 3, 5, 23

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025a. 2, 8

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. *arXiv preprint arXiv:2512.02012*, 2025b. 2, 8

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014. 2

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. *Advances in neural information processing systems*, 30, 2017. 3

Mathis Hardion and Théo Lacombe. The Wasserstein gradient flow of the Sinkhorn divergence between gaussian distributions. *arXiv preprint arXiv:2602.10726*, 2026. 5

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016. 9

Ping He, Om Khangaonkar, Hamed Pirsiavash, Yikun Bai, and Soheil Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026. 2, 5, 23, 24

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *NeurIPS*, 2017. 7

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 7, 20

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851, 2020. 1, 2

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998. 3

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, pp. 10124–10134, 2023. 8

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019. 9

Chieh-Hsin Lai, Bac Nguyen, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yuki Mitsufuji, Stefano Ermon, and Molei Tao. A unified view of drifting and score-based models. *arXiv preprint arXiv:2603.07514*, 2026. 2

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *NeurIPS*, 37:56424–56445, 2024. 8

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *ICML*, pp. 1718–1727. PMLR, 2015. 5, 23

Zirong Li, Siyuan Mei, Weiwen Wu, Andreas Maier, Lina Gölz, and Yan Xia. Generative drifting for conditional medical image generation. *arXiv preprint arXiv:2604.19736*, 2026. 2

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Adversarial flow models. *arXiv preprint arXiv:2511.22475*, 2025. 3, 8

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 1, 2, 3

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 1, 2, 3

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *arXiv preprint arXiv:2410.11081*, 2024. 1, 2

Jin Ma, Ying Tan, and Renyuan Xu. Schrödinger bridge for generative ai: Soft-constrained formulation and convergence analysis. *arXiv preprint arXiv:2510.11829*, 2025. 3

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, pp. 23–40. Springer, 2024. 8, 34

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, and Evgeny Burnaev. Large-scale Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 34:15243–15256, 2021. 3

Marcel Nutz and Johannes Wiesel. Entropic optimal transport: Convergence of potentials. *Probability Theory and Related Fields*, 184(1):401–424, 2022. 5

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *CVPR*, pp. 4195–4205, 2023. 7, 8

Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019. 5

Benedetto Piccoli and Francesco Rossi. Transport equation with nonlocal velocity in Wasserstein spaces: convergence of numerical schemes. *Acta applicandae mathematicae*, 124(1):73–105, 2013. 24

Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14104–14113, 2020. 9

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022. 7

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 2

Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. *NeurIPS*, 37:36046–36070, 2024. 2

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *SIGGRAPH*, pp. 1–10, 2022. 8

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 6

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pp. 2256–2265. PMLR, 2015. 1, 2

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. *arXiv preprint arXiv:2310.14189*, 2023. 2, 8

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019. 1

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 1, 2

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 2

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *NeurIPS*, 37:84839–84865, 2024. 8

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017. 3

Erkan Turan and Maks Ovsjanikov. Generative drifting is secretly score matching: a spectral and variational perspective. *arXiv preprint arXiv:2603.09936*, 2026. 2

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023. 2

Zidong Wang, Yiyuan Zhang, Xiaoyu Yue, Xiangyu Yue, Yangguang Li, Wanli Ouyang, and Lei Bai. Transition models: Rethinking the generative learning objective. *arXiv preprint arXiv:2509.04394*, 2025. 8

Yao Xie and Xiuyuan Cheng. Flow-based generative models as iterative algorithms in probability space. *arXiv preprint arXiv:2502.13394*, 2025. 3

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *CVPR*, pp. 15703–15712, 2025. 8, 34

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a. 2

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, pp. 6613–6623, 2024b. 2

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. *arXiv preprint arXiv:2410.06940*, 2024. 8

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. AlphaFlow: Understanding and improving MeanFlow models. *arXiv preprint arXiv:2510.20771*, 2025. 8

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 8

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. *arXiv preprint arXiv:2503.07565*, 2025. 2, 23

Huminhao Zhu, Fangyikang Wang, Chao Zhang, Hanbin Zhao, and Hui Qian. Neural sinkhorn gradient flow. *arXiv preprint arXiv:2401.14069*, 2024. 3, 5

Appendix

Table of Contents

A Additional discussions **15**
- A.1 Wasserstein gradient flows of energy functionals 15
- A.2 More discussion on the estimators for self-transport 19
- A.3 More discussions on velocity guidance 20
- A.4 Mode coverage on an imbalanced Gaussian mixture 22
- A.5 More discussions on related work 22

B Proofs **24**
- B.1 Complete statement and the proof of Theorem 3.1 24
- B.2 Assumption verification for Sinkhorn 30

C Additional implementation details **32**

D Sampling throughput comparison **34**

E Additional experimental results **34**

F Additional generated samples **36**

A Additional discussions

A.1 Wasserstein gradient flows of energy functionals

In this section, we derive the WGFs induced by the functionals listed in Table 1. For MMD and KL divergence, we also provide their particle-based estimators in the same spirit as the Sinkhorn estimator used in Sec. 3.3. The estimators are then used in the ablation study in Table 2a.

Throughout this section, p denotes the target data distribution. Given an energy functional $\mathcal{F}$, recall that the corresponding WGF follows

$$\partial_t q_t = \nabla \cdot \left(q_t \nabla \frac{\delta \mathcal{F}}{\delta q}(q_t) \right),$$

or equivalently, through the continuity equation, we have

$$\partial_t q_t + \nabla \cdot (q_t V_t) = 0, \qquad V_t(x) = -\nabla \frac{\delta \mathcal{F}}{\delta q}(q_t)(x).$$

Gradient flow of Sinkhorn divergence. To compute the first variation of $S_\varepsilon(q, p)$ under quadratic cost with respect to q, we first consider the dual formulation of the EOT problem:

$$\begin{aligned} \mathrm{OT}_\varepsilon(q, p) = \max_{u,v} & \int_{\mathbb{R}^d} u(x) dq(x) + \int_{\mathbb{R}^d} v(y) dp(y) \\ & - \varepsilon \iint_{\mathbb{R}^d \times \mathbb{R}^d} \exp\left(\frac{u(x) + v(y) - \frac{1}{2}\|x - y\|^2}{\varepsilon} \right) dq(x) dp(y) + \varepsilon, \end{aligned} \tag{16}$$

where $u(x)$ and $v(y)$ are the dual potentials.

By the envelope theorem, the first variation of the EOT cost Eq. (16) with respect to the marginal distribution q is exactly the optimal dual potential $u^{\varepsilon}_{q,p}(x)$:

$$\frac{\delta \mathrm{OT}_{\varepsilon}(q,p)}{\delta q}(x) = u^{\varepsilon}_{q,p}(x).$$

Consequently, the first variation of $S_{\varepsilon}(\cdot, p)$ evaluated at q is:

$$\frac{\delta S_{\varepsilon}(\cdot,p)}{\delta q}(q)(x) = u^{\varepsilon}_{q,p}(x) - u^{\varepsilon}_{q,q}(x).$$

From the optimality conditions of the dual problem, $u^{\varepsilon}_{q,p}(x)$ satisfies the following structural equation:

$$u^{\varepsilon}_{q,p}(x) = -\varepsilon \log \int_{\mathbb{R}^d} \exp\left(\frac{v^{\varepsilon}_{q,p}(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right) dp(y). \tag{17}$$

Taking the gradient with respect to x yields:

$$\begin{aligned}
\nabla_x u^{\varepsilon}_{q,p}(x) =& \frac{\int_{\mathbb{R}^d} \nabla_x \frac{1}{2}\|x-y\|^2 \exp\left(\frac{v^{\varepsilon}_{q,p}(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right) dp(y)}{\int_{\mathbb{R}^d} \exp\left(\frac{v^{\varepsilon}_{q,p}(y) - \frac{1}{2}\|x-y\|^2}{\varepsilon}\right) dp(y)} \\
=& \int_{\mathbb{R}^d} \nabla_x \frac{1}{2}\|x-y\|^2 \pi^{\varepsilon}_{q,p}(dy|x) \\
=& \int_{\mathbb{R}^d} (x-y) \pi^{\varepsilon}_{q,p}(dy|x) \\
=& x - T^{\varepsilon}_{q,p}(x).
\end{aligned} \tag{18}$$

Combining the previous results, we obtain

$$\begin{aligned}
V^{\varepsilon}_{q_t,p}(x) &= -\nabla_x \frac{\delta S_{\varepsilon}(\cdot,p)}{\delta q}(q_t)(x) \\
&= -\nabla_x \left(u^{\varepsilon}_{q_t,p}(x) - u^{\varepsilon}_{q_t,q_t}(x)\right) \\
&= -\left(x - T^{\varepsilon}_{q_t,p}(x)\right) + \left(x - T^{\varepsilon}_{q_t,q_t}(x)\right) \\
&= T^{\varepsilon}_{q_t,p}(x) - T^{\varepsilon}_{q_t,q_t}(x).
\end{aligned}$$

This recovers the Sinkhorn divergence velocity field reported in Table 1.

Gradient flow of MMD. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a differentiable positive definite kernel. For probability measures $p, q \in \mathcal{P}(\mathbb{R}^d)$, we consider the energy

$$\mathcal{F}_{\mathrm{MMD}}(q) := \frac{1}{2}\mathrm{MMD}^2(q,p), \tag{19}$$

where

$$\mathrm{MMD}^2(q,p) = \iint k(x,y)q(dx)q(dy) - 2\iint k(x,y)q(dx)p(dy) + \iint k(x,y)p(dx)p(dy). \tag{20}$$

Equivalently,

$$\mathcal{F}_{\mathrm{MMD}}(q) = \frac{1}{2}\iint k(x,y)(q-p)(dx)(q-p)(dy).$$

To compute the first variation, consider a perturbation $q_s = q + s\,\chi$ with signed measure χ satisfying $\int \chi(dx) = 0$. Then

$$\begin{aligned}
\frac{d}{ds}\mathcal{F}_{\mathrm{MMD}}(q_s)\Big|_{s=0} &= \iint k(x,y)\chi(dx)(q-p)(dy) \\
&= \int \left[\int k(x,y)(q-p)(dy)\right]\chi(dx).
\end{aligned}$$

Therefore, up to an additive constant,

$$\frac{\delta \mathcal{F}_{\mathrm{MMD}}}{\delta q}(q)(x) = \int k(x,y)q(dy) - \int k(x,y)p(dy).$$

Taking the negative spatial gradient gives the Wasserstein velocity field

$$\begin{aligned} V_t^{\mathrm{MMD}}(x) &= -\nabla_x \frac{\delta \mathcal{F}_{\mathrm{MMD}}}{\delta q}(q_t)(x) \\ &= \int \nabla_x k(x,y)p(dy) - \int \nabla_x k(x,y)q_t(dy), \end{aligned}$$

with q_t the corresponding flow under the initial condition q_0. This is the MMD velocity field reported in Table 1.

Particle-based estimator for MMD. Given generated particles $\{x_i\}_{i=1}^N \sim q_t$ and target particles $\{y_j\}_{j=1}^M \sim p$, define the empirical measures

$$\widehat{q}_t = \sum_{i=1}^N a_i \delta_{x_i}, \qquad \widehat{p} = \sum_{j=1}^M b_j \delta_{y_j}, \qquad \sum_i a_i = \sum_j b_j = 1.$$

As in the Sinkhorn self-transport estimator, we draw an independent second batch $\{x'_\ell\}_{\ell=1}^{N'} \sim q_t$ and define

$$\widehat{q}'_t = \sum_{\ell=1}^{N'} a'_\ell \delta_{x'_\ell}.$$

Then the particle estimator of the MMD-induced velocity at x_i is

$$\widehat{V}^{\mathrm{MMD}}_{\widehat{q}_t,\widehat{p}}(x_i) = \sum_{j=1}^M b_j \nabla_x k(x_i,y_j) - \sum_{\ell=1}^{N'} a'_\ell \nabla_x k(x_i,x'_\ell).$$

For uniform weights, this becomes

$$\widehat{V}^{\mathrm{MMD}}_{\widehat{q}_t,\widehat{p}}(x_i) = \frac{1}{M}\sum_{j=1}^M \nabla_x k(x_i,y_j) - \frac{1}{N'}\sum_{\ell=1}^{N'} \nabla_x k(x_i,x'_\ell).$$

If $k(\cdot,\cdot)$ is the Gaussian kernel

$$k_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right),$$

then

$$\nabla_x k_\sigma(x,y) = \frac{y-x}{\sigma^2} k_\sigma(x,y),$$

and hence

$$\widehat{V}^{\mathrm{MMD}}_{\widehat{q}_t,\widehat{p}}(x_i) = \frac{1}{M\sigma^2}\sum_{j=1}^M (y_j - x_i) k_\sigma(x_i,y_j) - \frac{1}{N'\sigma^2}\sum_{\ell=1}^{N'} (x'_\ell - x_i) k_\sigma(x_i,x'_\ell).$$

The corresponding particle update follows:

$$x_i \leftarrow x_i + \eta \widehat{V}^{\mathrm{MMD}}_{\widehat{q}_t,\widehat{p}}(x_i).$$

Gradient flow of KL divergence. For $p, q \in \mathcal{P}(\mathbb{R}^d)$, we next consider the KL energy

$$\mathcal{F}_{\mathrm{KL}}(q) := D_{\mathrm{KL}}(q\|p) = \int q(x) \log \frac{q(x)}{p(x)} dx,$$

assuming that q and p admit smooth positive densities with respect to the Lebesgue measure. For a perturbation $q_s = q + s\,\chi$ with $\int \chi(x)dx = 0$, we have

$$\frac{d}{ds}\mathcal{F}_{\mathrm{KL}}(q_s)\Big|_{s=0} = \int (\log q(x) - \log p(x) + 1)\,\chi(x)dx.$$

Therefore, up to an additive constant,

$$\frac{\delta \mathcal{F}_{\mathrm{KL}}}{\delta q}(q)(x) = \log q(x) - \log p(x).$$

The induced Wasserstein velocity field is

$$\begin{aligned} V_t^{\mathrm{KL}}(x) &= -\nabla_x \frac{\delta \mathcal{F}_{\mathrm{KL}}}{\delta q}(q_t)(x) \\ &= \nabla_x \log p(x) - \nabla_x \log q_t(x), \end{aligned}$$

with q_t the corresponding flow under the initial condition q_0. This recovers the KL velocity field reported in Table 1. The first term attracts particles toward high-density regions of the target distribution, while the second term is a repulsive score term induced by the current model distribution.

Particle-based estimator for KL. Unlike MMD and Sinkhorn divergence, the KL velocity field requires access to the score functions $\nabla \log p$ and $\nabla \log q_t$. In the absence of analytic scores, a particle-based estimator can be constructed through kernel density estimation (KDE).

Let $k_\sigma(\cdot, \cdot)$ be a smooth kernel with bandwidth $\sigma > 0$, for example,

$$k_\sigma(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right).$$

Using target particles $\{y_j\}_{j=1}^M \sim p$ and an independent generated batch $\{x'_\ell\}_{\ell=1}^{N'} \sim q_t$, define

$$\widehat{p}_\sigma(x) = \sum_{j=1}^{M} b_j k_\sigma(x, y_j), \qquad \widehat{q}_{t,\sigma}(x) = \sum_{\ell=1}^{N'} a'_\ell k_\sigma(x, x'_\ell).$$

The KDE-based score estimators are

$$\widehat{s}_p(x) := \nabla_x \log \widehat{p}_\sigma(x) = \frac{\sum_{j=1}^M b_j \nabla_x k_\sigma(x, y_j)}{\sum_{j=1}^M b_j k_\sigma(x, y_j)},$$

and

$$\widehat{s}_{q_t}(x) := \nabla_x \log \widehat{q}_{t,\sigma}(x) = \frac{\sum_{\ell=1}^{N'} a'_\ell \nabla_x k_\sigma(x, x'_\ell)}{\sum_{\ell=1}^{N'} a'_\ell k_\sigma(x, x'_\ell)}.$$

The particle estimator of the KL-induced velocity is

$$\widehat{V}^{\mathrm{KL}}_{\widehat{q}_t, \widehat{p}}(x_i) = \widehat{s}_p(x_i) - \widehat{s}_{q_t}(x_i).$$

For the Gaussian kernel, we have

$$\nabla_x k_\sigma(x, y) = \frac{y - x}{\sigma^2} k_\sigma(x, y).$$

Therefore,

$$\widehat{s}_p(x_i) = \frac{1}{\sigma^2}\left(\frac{\sum_{j=1}^M b_j k_\sigma(x_i, y_j) y_j}{\sum_{j=1}^M b_j k_\sigma(x_i, y_j)} - x_i\right),$$

and similarly

$$\widehat{s}_{q_t}(x_i) = \frac{1}{\sigma^2}\left(\frac{\sum_{\ell=1}^{N'} a'_\ell k_\sigma(x_i, x'_\ell) x'_\ell}{\sum_{\ell=1}^{N'} a'_\ell k_\sigma(x_i, x'_\ell)} - x_i\right).$$

The particle-based estimate of the velocity field simplifies to the difference between two kernel-weighted local means:

$$\widehat{V}^{\mathrm{KL}}_{\widehat{q}_t, \widehat{p}}(x_i) = \frac{1}{\sigma^2}\left[\frac{\sum_{j=1}^M b_j k_\sigma(x_i, y_j) y_j}{\sum_{j=1}^M b_j k_\sigma(x_i, y_j)} - \frac{\sum_{\ell=1}^{N'} a'_\ell k_\sigma(x_i, x'_\ell) x'_\ell}{\sum_{\ell=1}^{N'} a'_\ell k_\sigma(x_i, x'_\ell)}\right].$$

The corresponding particle update then follows:

$$x_i \leftarrow x_i + \eta \widehat{V}^{\mathrm{KL}}_{\widehat{q}_t, \widehat{p}}(x_i).$$

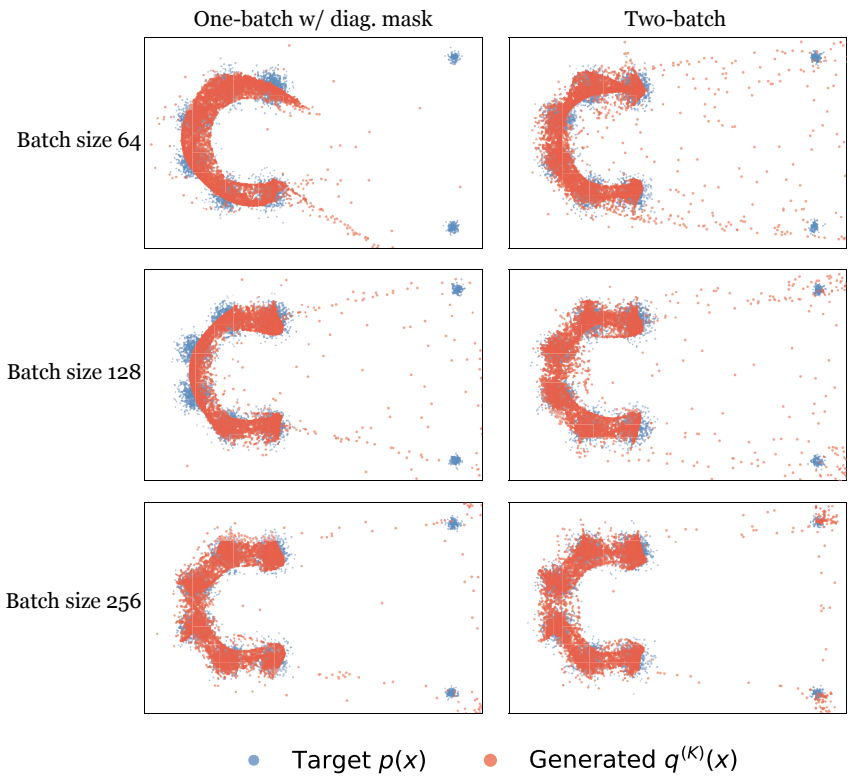


Figure 7: Evaluation of self-transport estimators on a 2D Gaussian mixture dataset featuring six dominant modes and two distant minority modes. "One-batch" refers to the single batch estimator with diagonal masking. "Two-batch" refers to our two-batch estimator with resampling.

A.2 More discussion on the estimators for self-transport

As introduced in Sec. 3.3, estimating the self-interaction velocity field $T^{\varepsilon}_{q,q}(x)$ requires computing the OT plan of the model distribution q with itself. A naive empirical approximation based on a single mini-batch of size B, i.e., $\{x_i\}_{i=1}^{B}$, inherently suffers from a self-matching bias, as shown in Fig. 3. Following Deng et al. (2026), which addresses a similar issue, one may instead apply a diagonal mask to the cost matrix (setting $C_{ii} = \infty$), which explicitly prevents any particle from matching with itself.

However, this masking strategy introduces a new source of bias that is sensitive to the batch size B. For any given particle x_i, the diagonal mask effectively forces it to transport its mass to an empirical target distribution composed of the remaining particles:

$$\widehat{q}_{-i} = \frac{1}{B-1} \sum_{j \neq i} \delta_{x_j}.$$

Because x_i is excluded from its own target space, the particle is artificially forced to move a strictly positive distance towards its neighbors, even if it is already situated perfectly at a high-density peak of the true distribution q. The magnitude of this structural bias scales proportionally to $\mathcal{O}(1/B)$, effectively meaning that a fraction of the mass corresponding to roughly $(B-1)/B$ is mapped sub-optimally. As B becomes smaller, this forced redistribution dominates the velocity field, pushing particles away from their true local density centers (as observed in the left column of Fig. 7).

By drawing an independent second batch $\{x'_l\}_{l=1}^{B} \sim q_t$, our two-batch estimator effectively avoids this issue. The OT plan is computed between two independent empirical measures, allowing a particle x_i to match freely with a particle x'_l that happens to be infinitesimally close to it, without any artificial ∞-cost penalty. This yields a consistent estimate of the continuous self-transport map that remains robust even at small batch sizes.

A.3 More discussions on velocity guidance

We first recall distribution guidance, then show that velocity guidance recovers the standard exponentially tilted target under the KL energy. This motivates the analogous Sinkhorn construction in Eq. (15).

Distribution guidance. A natural implementation in Deng et al. (2026) is to modify the repulsive distribution as

$$\tilde{q}_\theta^w(x|c) := \frac{1}{w+1} q_\theta(x|c) + \frac{w}{w+1} p(x|\varnothing), \tag{21}$$

where $w \geq 0$ is the guidance scale and $p(x|\varnothing)$ is the unconditional (marginal) distribution. Operationally, this amounts to sampling additional particles from $p(x|\varnothing)$ with importance weight w *w.r.t.* q_θ. However, this approach implicitly *alters* the target being learned. One can show that it corresponds to learning a *linearly* extrapolated target distribution $\tilde{p}^w(x|c) = p(x|c) + w(p(x|c) - p(x|\varnothing))$, which deviates from the standard CFG formulation based on exponential tilting (Ho & Salimans, 2022; Dhariwal & Nichol, 2021).

Let $p(\cdot|c)$ denote the conditional data distribution, $p(\cdot|\varnothing)$ denote the unconditional data distribution, and $w \geq 0$ be the guidance weight. In analogy with Eq. (15), we construct the velocity-guidance field by adding a term scaled by w that points from the unconditional to the conditional distribution:

$$\widetilde{V}_t^{\mathrm{KL},w}(x) := \underbrace{\Big(\nabla_x \log p(x|c) - \nabla_x \log q_t(x|c)\Big)}_{V^{\mathrm{KL}}_{q_t(\cdot|c),p(\cdot|c)}} + w \underbrace{\Big(\nabla_x \log p(x|c) - \nabla_x \log p(x|\varnothing)\Big)}_{V^{\mathrm{KL}}_{q_t(\cdot|c),p(\cdot|c)} - V^{\mathrm{KL}}_{q_t(\cdot|c),p(\cdot|\varnothing)}}. \tag{22}$$

We formalize the connection between this constructed velocity field and CFG in the following proposition.

Proposition A.1. *The velocity-guidance field $\widetilde{V}_t^{\mathrm{KL},w}$ in Eq. (22) is exactly the WGF velocity field of the combined energy functional*

$$\mathcal{F}^w_{\mathrm{KL}}(q_t(\cdot|c)) = (1+w) D_{\mathrm{KL}}\big(q_t(\cdot|c) \| p(\cdot|c)\big) - w D_{\mathrm{KL}}\big(q_t(\cdot|c) \| p(\cdot|\varnothing)\big).$$

Furthermore, the unique global minimizer of $\mathcal{F}_w$ is the exponentially tilted target distribution $p_w(\cdot|c)$, defined as

$$p_w(x|c) := \frac{1}{Z_w(c)} p(x|c) \left(\frac{p(x|c)}{p(x|\varnothing)}\right)^w,$$

where $Z_w(c)$ is the normalizing constant.

Proof. We first expand the combined energy functional $\mathcal{F}^w_{\mathrm{KL}}(q_t(\cdot|c))$:

$$\begin{aligned}
\mathcal{F}^w_{\mathrm{KL}}(q_t) &= (1+w) \int q_t \log \frac{q_t}{p(\cdot|c)} - w \int q_t \log \frac{q_t}{p(\cdot|\varnothing)} \\
&= \int q_t \log q_t - \int q_t \log \Big(p(\cdot|c)^{1+w} p(\cdot|\varnothing)^{-w}\Big) \\
&= \int q_t \log \frac{q_t}{p_w(\cdot|c)} - \log Z_w(c) \\
&= D_{\mathrm{KL}}\big(q_t(\cdot|c) \| p_w(\cdot|c)\big) - \log Z_w(c).
\end{aligned}$$

Since $\log Z_w(c)$ is a constant with respect to the distribution q_t, it is evident that $\mathcal{F}^w_{\mathrm{KL}}(q_t(\cdot|c))$ achieves its unique global minimum when $q_t(\cdot|c) = p_w(\cdot|c)$, which proves the second part of the proposition.

To derive the induced WGF, we compute the first variation of $\mathcal{F}^w_{\mathrm{KL}}(q_t(\cdot|c))$. Because the additive constant $-\log Z_w(c)$ vanishes upon differentiation, the first variation is identical to that of $D_{\mathrm{KL}}\big(q_t(\cdot|c) \| p_w(\cdot|c)\big)$:

$$\frac{\delta \mathcal{F}^w_{\mathrm{KL}}}{\delta q}(q_t(\cdot|c))(x) = \log q_t(x|c) - \log p_w(x|c) + \text{const}.$$

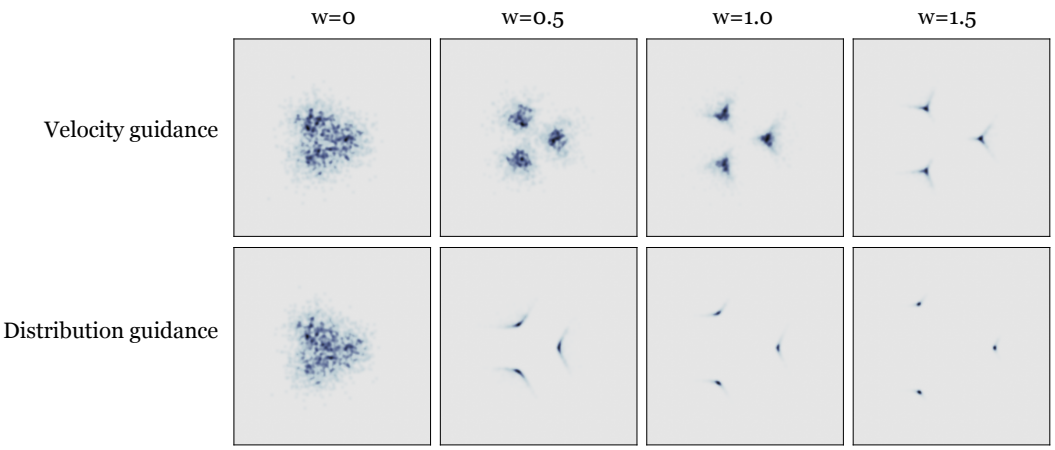


Figure 8: Comparison of velocity guidance and distribution guidance for conditional generation on a 2D 3-mode Gaussian mixture dataset.

The corresponding Wasserstein velocity field is the negative spatial gradient of the first variation:

$$\begin{aligned}-\nabla_x \frac{\delta \mathcal{F}^w_{\mathrm{KL}}}{\delta q}(q_t(\cdot|c))(x) &= \nabla_x \log p_w(x|c) - \nabla_x \log q_t(x|c) \\ &= \nabla_x \Big((1+w)\log p(x|c) - w \log p(x|\varnothing) - \log Z_w(c)\Big) - \nabla_x \log q_t(x|c) \\ &= (1+w)\nabla_x \log p(x|c) - w\nabla_x \log p(x|\varnothing) - \nabla_x \log q_t(x|c).\end{aligned}$$

Rearranging the terms, we obtain:

$$-\nabla_x \frac{\delta \mathcal{F}^w_{\mathrm{KL}}}{\delta q}(q_t(\cdot|c))(x) = \Big(\nabla_x \log p(x|c) - \nabla_x \log q_t(x|c)\Big) + w\Big(\nabla_x \log p(x|c) - \nabla_x \log p(x|\varnothing)\Big),$$

which is exactly the constructed velocity-guidance field $\widetilde{V}_t^{\mathrm{KL},w}(x)$ defined in Eq. (22). □

Therefore, under the KL energy, our velocity-guidance formulation is precisely equivalent to running the WGF toward the exponentially tilted distribution $p_w(x|c)$. This corresponds exactly to the standard CFG target form, where the conditional distribution is amplified relative to the unconditional distribution by the density ratio $p(x|c)/p(x|\varnothing)$. The results in Appendix E verify that velocity guidance yields enhanced performance over distribution guidance when KL divergence is leveraged as the energy functional.

Connection to the Sinkhorn velocity guidance. For the Sinkhorn divergence, the same closed-form target distribution is generally unavailable, because the velocity field is expressed through barycentric projections rather than score functions. Nevertheless, the KL case shows that adding the velocity-level correction

$$V_{q_t(\cdot|c),p(\cdot|c)} - V_{q_t(\cdot|c),p(\cdot|\varnothing)}$$

has the correct interpretation in the score-based limit: it changes the effective target from $p(x|c)$ to the exponentially tilted distribution $p(x|c)(p(x|c)/p(x|\varnothing))^w$. This motivates applying the analogous guidance design to the Sinkhorn-induced velocity field. Specifically, we have

$$\tilde{V}^{\varepsilon,w}_{q_\theta,p}(x_i) = \underbrace{\Big(T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|c)}(x_i) - T^{\varepsilon}_{q_\theta(\cdot|c),q_\theta(\cdot|c)}(x_i)\Big)}_{V^{\epsilon}_{q_\theta(\cdot|c),p(\cdot|c)}} + w \underbrace{\Big(T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|c)}(x_i) - T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|\varnothing)}(x_i)\Big)}_{V^{\epsilon}_{q_\theta(\cdot|c),p(\cdot|c)} - V^{\epsilon}_{q_\theta(\cdot|c),p(\cdot|\varnothing)}},$$

where

$$V^{\epsilon}_{q_\theta(\cdot|c),p(\cdot|c)}(x_i) = T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|c)}(x_i) - T^{\varepsilon}_{q_\theta(\cdot|c),q_\theta(\cdot|c)}(x_i), \tag{23}$$

$$V^{\epsilon}_{q_\theta(\cdot|c),p(\cdot|\varnothing)}(x_i) = T^{\varepsilon}_{q_\theta(\cdot|c),p(\cdot|\varnothing)}(x_i) - T^{\varepsilon}_{q_\theta(\cdot|c),q_\theta(\cdot|c)}(x_i), \tag{24}$$



Figure 9: Mode coverage on a 2D Gaussian mixture with six dominant modes and two distant minority modes.

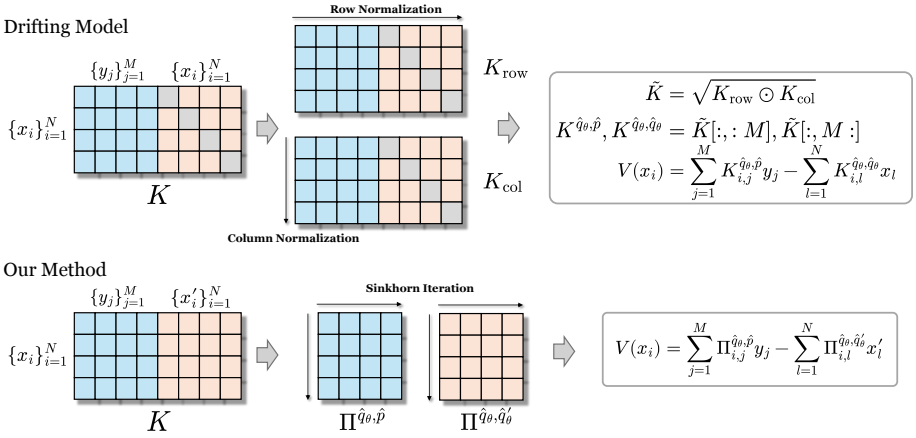


Figure 10: Illustrations of the difference in the velocity field computation between Drifting Model (Alg. 2 in Deng et al. (2026)) and our approach. "$\sqrt{\cdot}$" refers to the element-wise square root, and the gray blocks refer to the diagonal mask adopted in Deng et al. (2026). Drifting Model adopts a heuristic construction with separate row-wise and column-wise Softmax normalizations. Our approach leverages Sinkhorn iteration to compute the barycentric projection.

which recovers Eq. (15).

Toy illustration. To empirically compare our proposed velocity guidance with distribution guidance, we evaluate conditional generation on a 2D 3-mode Gaussian mixtures. Here, each mode is treated as a distinct class. We train a generator $x = f_\theta(z, c, w)$ where the guidance scale w is randomly sampled during training.

As illustrated in Fig. 8 (top row), velocity guidance reflects the behavior of an exponentially tilted target distribution $p_w(x|c)$. As w increases, the probability mass is amplified at the highest-density regions of the conditional mode. In contrast, the distribution guidance implicitly targets a linear mixture between the conditional and unconditional distributions. Fig. 8 (bottom row) demonstrates that increasing w causes particles to scatter away from the modes.

A.4 Mode coverage on an imbalanced Gaussian mixture

We construct a 2D Gaussian mixture dataset with six large, tightly concentrated modes and two small distant minority modes. As shown in Fig. 9, Drifting Model greedily collapses on the dominant clusters, while the Sinkhorn barycentric projections in W-Flow enforce global marginal constraints and preserve all modes. This complements the imbalanced FFHQ results in Sec. 4.3.

A.5 More discussions on related work

Our proposed W-Flow framework is related to several prominent directions in generative modeling. We highlight the key distinctions below.

Connection to Drifting Model. Our formulation shares the forward, iterative pushforward structure with Drifting Model (Deng et al., 2026). The main difference lies in the construction of the velocity field. Drifting Model employs heuristic attraction and repulsion terms (V_p^+ and V_q^-) based on localized kernel interactions, resulting in a mean-shift-type dynamics driven by local proximity. Such local updates can lead to unstable behavior and require careful tuning to avoid mode collapse.

In contrast, W-Flow derives the velocity field from a WGF and instantiates it via Sinkhorn divergence. The resulting dynamics are governed by global OT couplings, which enforce marginal constraints and coordinate particle movement at the distribution level. This yields a principled and globally consistent transport mechanism.

Connection to Sinkhorn-Drifting. Concurrent Sinkhorn-Drifting (He et al., 2026) shares both the population velocity in Eq. (10) and the stop-gradient regression objective in Eq. (3). While its empirical flow uses a same-support self-coupling (Eq. (15) therein), we explicitly estimate self-transport with an independent second generated batch (Sec. 3.3). In our ablation, removing this second batch increases FID from 7.08 to 17.57 (Table 2c). We further introduce velocity-level classifier-free guidance and validate the combined framework on ImageNet 256×256. On the theoretical side, their statistical analysis bounds instantaneous field-estimation error, whereas Theorem B.4 controls the empirical Euler trajectory over a finite horizon, separating initial-particle, target-sampling, and discretization errors. This guarantee concerns particle dynamics rather than neural regression error.

Distinction from JKO-based WGF models. Another line of work formulates generative processes as WGFs using the JKO scheme (Choi et al., 2024), which corresponds to an implicit time discretization. However, each JKO step requires solving a nontrivial optimization problem, which in practice often necessitates adversarial training, thereby introducing training instability and limiting scalability.

Distinct from Choi et al. (2024), W-Flow adopts an explicit Euler discretization of the flow. By computing the Sinkhorn velocity field V^ε via efficient Sinkhorn scaling, the learning problem reduces to a simple regression objective, avoiding inner optimization loops and eliminating the need for adversarial training. This enables scalable learning of a one-step pushforward map.

Distinction from Sinkhorn GANs. Sinkhorn GAN (Genevay et al., 2018) directly uses the Sinkhorn divergence as a generator loss. Differentiating through unrolled Sinkhorn iterations can incur memory overhead, although gradients at optimal couplings can also be computed via an envelope argument without differentiating through the solver (He et al., 2026).

W-Flow uses this divergence to define WGF velocities and regress toward detached local transport targets. The distinction is this training formulation; avoiding backpropagation through Sinkhorn alone is not unique to our method.

Distinction from moment matching. Generative Moment Matching Networks (GMMNs) (Li et al., 2015) learn generators through particle updates driven by the squared MMD, which can be viewed as an instance of the MMD-based variant of W-Flow in Table 1. However, MMD relies on kernel-based interactions between particles, which can suffer from the curse of dimensionality and yield vanishing gradients when the current pushforward distribution is far from the target, as also reflected in our ablation experiments. Inductive Moment Matching (Zhou et al., 2025) performs kernel-based moment matching along a prescribed static interpolation between the prior and target distributions. As a result, it inherits the same kernel-related limitations and typically requires multiple pushforward sampling steps to simulate transitions along the interpolator. W-Flow instead leverages the OT-based dynamics and directly parametrizes a one-step map that does not require any intermediate simulations at inference time.

Limitations. While W-Flow achieves strong one-step generation performance, several limitations remain. First, our empirical evaluation is focused on ImageNet 256×256 and FFHQ; extending the framework to higher-resolution, text-conditioned, video, or other multimodal generation settings remains future work. Second, training still requires pretrained feature encoders or autoencoders in our large-scale experiments, and the choice of the feature encoders remains heuristic. Third, our theoretical convergence result analyzes the particle dynamics under regularity and asymptotic assumptions, and does not fully characterize finite-network optimization or finite-compute training in high-dimensional neural generators.

B PROOFS

In this section, we present the detailed statements of assumptions and theorems discussed in the main paper, along with their proofs.

The main result is formulated for the W-Flow setting, capturing both the two-measure velocity field $V^{\varepsilon}_{q,p}$ and the empirical target approximation. Its proof uses stability and particle-approximation techniques for nonlocal continuity equations. These techniques are largely inspired by related arguments in the literature, though the assumptions and settings are different; see Piccoli & Rossi (2013). Relatedly, Proposition 3.8 of He et al. (2026) studies the statistical approximation of the Sinkhorn drifting field, whereas our result establishes the convergence of the particle dynamics under a more general setting (see Assumption B.1).

In the main text, Theorem 3.1 is stated for the Sinkhorn velocity $V^{\varepsilon}_{q,p}$ defined in (10). For generality, here we prove the result under Assumption B.1 for a generic velocity field

$$V : \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d, \qquad (q, p, x) \mapsto V_{q,p}(x),$$

where q is the evolving source distribution and p is the fixed target distribution. Proposition B.7 then verifies the assumption for the Sinkhorn velocity when q_0 and p have bounded support.

B.1 COMPLETE STATEMENT AND THE PROOF OF THEOREM 3.1

For $N \in \mathbb{N}$ and step size $\eta > 0$, let particles $\{x_i^{(k)}\}_{i=1}^N$ evolve according to

$$x_i^{(k+1)} = x_i^{(k)} + \eta\, V_{\widehat{q}_k^N, \widehat{p}^M}(x_i^{(k)}), \qquad \widehat{q}_k^N := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^{(k)}}, \qquad t_k := k\eta, \tag{25}$$

with the initial data points $x_i^{(0)}$ sampled i.i.d. from q_0.

Define the piecewise-linear interpolation $x_i^{\eta} : [0, \infty) \to \mathbb{R}^d$ by

$$x_i^{\eta}(t) := x_i^{(k)} + (t - t_k)\, V_{\widehat{q}_k^N, \widehat{p}^M}(x_i^{(k)}), \quad t \in [t_k, t_{k+1}), \quad i = 1, 2, \cdots, N, \tag{26}$$

and the associated empirical measure

$$\widehat{q}_t^{N,M,\eta} := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^{\eta}(t)}.$$

We study the limiting behavior of (26) under a joint asymptotic regime in which $N, M \to \infty$ and $\eta \to 0$.

Let $m_2(\mu)$ be the root second moment of $\mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$m_2(\mu) := \left(\int_{\mathbb{R}^d} \|x\|^2 \, d\mu(x)\right)^{1/2}.$$

Assumption B.1. *Assume that $V_{q,p}$ is Borel measurable and belongs to $L^2(q)$ for every $q, p \in \mathcal{P}_2(\mathbb{R}^d)$. Assume that there exist constants $L_0, L_1, L_2, L_3 < \infty$ such that, for all $q, \widetilde{q}, p, \widetilde{p} \in \mathcal{P}_2(\mathbb{R}^d)$,*

$$\|V_{q,p}\|_{L^2(q)} \le L_0\big(m_2(q) + m_2(p)\big), \tag{27}$$

$$\|V_{q,p}(x) - V_{q,p}(x')\| \le L_1\|x - x'\|, \qquad x, x' \in \mathbb{R}^d, \tag{28}$$

$$\|V_{q,p} - V_{\widetilde{q},p}\|_{L^2(\widetilde{q})} \le L_2\mathcal{W}_2(q, \widetilde{q}), \tag{29}$$

$$\|V_{q,p} - V_{q,\widetilde{p}}\|_{L^2(q)} \le L_3\mathcal{W}_2(p, \widetilde{p}). \tag{30}$$

Remark B.2 (Some examples on $\mathcal{P}_2(\mathbb{R}^d)$)**.** The conditions in Assumption B.1 are in the spirit of the Lipschitz and linear-growth assumptions commonly used for stability of continuity equations and mean-field particle systems; see, e.g., Ambrosio et al. (2008); Carmona & Delarue (2018). Assumption B.1 can be verified directly for several natural velocity fields on the full space $\mathbb{R}^d$. One example is the velocity induced by the squared MMD energy. More precisely, for a positive-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, recall the corresponding energy potential defined in (19). If

$k(x,y) = \kappa(x-y)$ is translation invariant and $\nabla\kappa : \mathbb{R}^d \to \mathbb{R}^d$ is globally Lipschitz, then the corresponding Wasserstein velocity becomes

$$V_{q,p}(x) = \int_{\mathbb{R}^d} \kappa(x-y)\, p(dy) - \int_{\mathbb{R}^d} \kappa(x-y)\, q(dy).$$

In this case, Assumption B.1 holds on $\mathcal{P}_2(\mathbb{R}^d)$, with constants depending only on the Lipschitz constant of $\nabla\kappa$. For example, the Gaussian kernel $\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ belongs to this class. Another simple example is the moment-matching velocity

$$V_{q,p}(x) = \bar{p} - \bar{q}, \quad \text{with} \quad \bar{q} := \int_{\mathbb{R}^d} y\, q(dy), \ \ \bar{p} := \int_{\mathbb{R}^d} y\, p(dy),$$

which also satisfies Assumption B.1 on $\mathcal{P}_2(\mathbb{R}^d)$.

Remark B.3 (Localization on bounded domains)**.** When the initial distribution and the target distribution have bounded supports, the Sinkhorn velocity satisfies the conditions in Assumption B.1 along the relevant flow trajectory. Indeed, the Sinkhorn flow and its empirical approximations remain supported on a bounded domain over every finite time interval $[0,T]$; see Proposition B.7. Therefore, it is enough to verify the estimates in Assumption B.1 on this bounded class of measures, rather than globally on $\mathcal{P}_2(\mathbb{R}^d)$.

Under these conditions, one can establish a quantitative convergence result showing that the interpolated empirical measure $\widehat{q}_t^{N,M,\eta}$ converges to the unique solution of a nonlinear PDE as $N, M \to \infty$ and $\eta \to 0$.

Now we present the *formal version* of Theorem 3.1 below, followed by the detailed proof.

Theorem B.4. *Assume Assumption B.1 holds and $q_0, p \in \mathcal{P}_2(\mathbb{R}^d)$. For every $T > 0$, there exists a unique $q \in C([0,T]; \mathcal{P}_2(\mathbb{R}^d))$ solving the nonlinear continuity equation*

$$\partial_t q_t + \nabla \cdot \big(q_t\, V_{q_t,p}\big) = 0 \tag{31}$$

in the weak sense

$$\int_{\mathbb{R}^d} \varphi\, dq_t - \int_{\mathbb{R}^d} \varphi\, dq_0 = \int_0^t \int_{\mathbb{R}^d} \nabla\varphi(x) \cdot V_{q_s,p}(x)\, dq_s(x)\, ds, \qquad \forall \varphi \in C_c^1(\mathbb{R}^d).$$

Moreover, for all N, M and all $\eta > 0$,

$$\begin{aligned}\sup_{t\in[0,T]} \mathcal{W}_2\big(\widehat{q}_t^{N,M,\eta}, q_t\big) \le\ & e^{(L_1+L_2)T} \mathcal{W}_2(\widehat{q}_0^N, q_0) \\ & + \frac{L_3}{L_1+L_2}\big(e^{(L_1+L_2)T} - 1\big)\mathcal{W}_2(\widehat{p}^M, p) \\ & + C_T\, \eta\big[m_2(\widehat{q}_0^N) + m_2(\widehat{p}^M)\big],\end{aligned}$$

where $C_T < \infty$ depends only on T, L_0, L_1, L_2. Consequently, as $\eta \to 0$ and $N, M \to \infty$,

$$\sup_{t\in[0,T]} \mathcal{W}_2\big(\widehat{q}_t^{N,M,\eta}, q_t\big) \to 0, \qquad \textit{almost surely.} \tag{32}$$

Proof. For generic $\boldsymbol{x} = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N$, define

$$\widehat{q}^N := \frac{1}{N}\sum_{i=1}^N \delta_{x_i}, \quad V^{N,M}(\boldsymbol{x}) := \big(V_{\widehat{q}^N,\widehat{p}^M}(x_1), \dots, V_{\widehat{q}^N,\widehat{p}^M}(x_N)\big).$$

Equip $(\mathbb{R}^d)^N$ with the norm

$$\|\boldsymbol{x}\|_{2,N} := \left(\frac{1}{N}\sum_{i=1}^N \|x_i\|^2\right)^{1/2}.$$

Let $\boldsymbol{x}' = (x_1', \dots, x_N')$, and set $\widehat{q}^{N\prime} := N^{-1}\sum_{i=1}^N \delta_{x_i'}$. By (28) and (29),

$$\begin{aligned}\|V^{N,M}(\boldsymbol{x}) - V^{N,M}(\boldsymbol{x}')\|_{2,N} &\le \left(\frac{1}{N}\sum_{i=1}^N \left\|V_{\widehat{q}^N,\widehat{p}^M}(x_i) - V_{\widehat{q}^N,\widehat{p}^M}(x_i')\right\|^2\right)^{1/2} \\ &\quad + \left(\frac{1}{N}\sum_{i=1}^N \left\|V_{\widehat{q}^N,\widehat{p}^M}(x_i') - V_{\widehat{q}^{N\prime},\widehat{p}^M}(x_i')\right\|^2\right)^{1/2} \\ &\le L_1\|\boldsymbol{x}-\boldsymbol{x}'\|_{2,N} + L_2\mathcal{W}_2(\widehat{q}^N, \widehat{q}^{N\prime}) \\ &\le (L_1+L_2)\|\boldsymbol{x}-\boldsymbol{x}'\|_{2,N},\end{aligned}$$

where

$$\mathcal{W}_2(\widehat{q}^N, \widehat{q}^{N\prime}) \le \left(\frac{1}{N}\sum_{j=1}^N \|x_j - x_j'\|^2\right)^{1/2} = \|\boldsymbol{x}-\boldsymbol{x}'\|_{2,N}$$

by coupling x_i with x_i'. Moreover, by (27),

$$\|V^{N,M}(\boldsymbol{x})\|_{2,N} = \|V_{\widehat{q}^N,\widehat{p}^M}\|_{L^2(\widehat{q}^N)} \le L_0\big(m_2(\widehat{q}^N) + m_2(\widehat{p}^M)\big).$$

Thus the finite-particle vector field is globally Lipschitz and has linear growth in $\|\cdot\|_{2,N}$.

Consider now the ODE

$$\dot{\boldsymbol{x}}^{N,M}(t) = V^{N,M}(\boldsymbol{x}^{N,M}(t)), \qquad \boldsymbol{x}^{N,M}(0) = (x_1^{(0)}, \dots, x_N^{(0)}), \tag{33}$$

with the initialization $x_i^{(0)}$ defined below (25). Since $V^{N,M}$ is globally Lipschitz in $\|\cdot\|_{2,N}$, it holds that (33) has a unique global solution by the Picard–Lindelöf theorem.

Let

$$\widehat{q}_t^{N,M} := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^{N,M}(t)}.$$

For any $\varphi \in C_c^1(\mathbb{R}^d)$,

$$\begin{aligned}\frac{d}{dt}\int \varphi\, d\widehat{q}_t^{N,M} &= \frac{d}{dt}\frac{1}{N}\sum_{i=1}^N \varphi(x_i^{N,M}(t)) = \frac{1}{N}\sum_{i=1}^N \nabla\varphi(x_i^{N,M}(t))\cdot \dot{x}_i^{N,M}(t) \\ &= \int \nabla\varphi(x)\cdot V_{\widehat{q}_t^{N,M},\widehat{p}^M}(x)\, d\widehat{q}_t^{N,M}(x).\end{aligned}$$

Thus $\widehat{q}_t^{N,M}$ is a weak solution of

$$\partial_t \widehat{q}_t^{N,M} + \nabla\cdot\left(\widehat{q}_t^{N,M} V_{\widehat{q}_t^{N,M},\widehat{p}^M}\right) = 0.$$

We next derive the stability estimate. Let $\mu, \nu \in C([0,T];\mathcal{P}_2(\mathbb{R}^d))$ be weak solutions of

$$\partial_t\mu_t + \nabla\cdot(\mu_t V_{\mu_t,p}) = 0, \qquad \partial_t\nu_t + \nabla\cdot(\nu_t V_{\nu_t,\widetilde{p}}) = 0,$$

where $p, \widetilde{p} \in \mathcal{P}_2(\mathbb{R}^d)$ are fixed. First note that (27) and (28) imply the pointwise linear-growth bound

$$\|V_{r,a}(x)\| \le L_1\|x\| + (L_0+L_1)m_2(r) + L_0 m_2(a), \qquad r, a \in \mathcal{P}_2(\mathbb{R}^d). \tag{34}$$

Indeed,

$$\|V_{r,a}(0)\| \le \left(\int \|V_{r,a}(0) - V_{r,a}(z)\|^2\, dr(z)\right)^{1/2} + \|V_{r,a}\|_{L^2(r)} \le L_1 m_2(r) + L_0\big(m_2(r)+m_2(a)\big).$$

Since $\mu, \nu \in C([0,T];\mathcal{P}_2(\mathbb{R}^d))$, for these fixed solution trajectories we have

$$M_T := \sup_{t\in[0,T]}\big(m_2(\mu_t) + m_2(\nu_t)\big) < \infty.$$

Let $\gamma_0 \in \Pi(\mu_0, \nu_0)$ be optimal for $\mathcal{W}_2(\mu_0, \nu_0)$. For each $x_0, y_0 \in \mathbb{R}^d$, define $X(\cdot; x_0)$ and $Y(\cdot; y_0)$ as the solutions on $[0, T]$ of

$$\begin{cases} \frac{d}{dt} X(t; x_0) = V_{\mu_t, p}(X(t; x_0)), \\ X(0; x_0) = x_0, \end{cases} \qquad \begin{cases} \frac{d}{dt} Y(t; y_0) = V_{\nu_t, \widetilde{p}}(Y(t; y_0)), \\ Y(0; y_0) = y_0. \end{cases}$$

The global well-posedness of these ODEs on $[0, T]$ follows from (34) and (28), since $\mu, \nu \in C([0, T]; \mathcal{P}_2(\mathbb{R}^d))$. Since μ_t and ν_t solve the corresponding continuity equations, their characteristic representations give

$$X(t; \cdot)_{\#} \mu_0 = \mu_t, \qquad Y(t; \cdot)_{\#} \nu_0 = \nu_t.$$

Therefore

$$\big(X(t; \cdot), Y(t; \cdot)\big)_{\#} \gamma_0 \in \Pi(\mu_t, \nu_t).$$

Define

$$D(t) := \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} \|X(t; x_0) - Y(t; y_0)\|^2 \, d\gamma_0(x_0, y_0) \right)^{1/2}.$$

Since $\big(X(t; \cdot), Y(t; \cdot)\big)_{\#} \gamma_0 \in \Pi(\mu_t, \nu_t)$, we have

$$\mathcal{W}_2(\mu_t, \nu_t) \le D(t).$$

For $\delta > 0$, set

$$D_\delta(t) := (D(t)^2 + \delta)^{1/2}.$$

For a.e. t,

$$\begin{aligned} \frac{d}{dt} D_\delta(t) &= \frac{1}{D_\delta(t)} \int \big(X(t; x_0) - Y(t; y_0)\big) \cdot \Big[V_{\mu_t, p}(X(t; x_0)) - V_{\nu_t, \widetilde{p}}(Y(t; y_0))\Big] \, d\gamma_0(x_0, y_0) \\ &\le \left(\int \Big\| V_{\mu_t, p}(X(t; x_0)) - V_{\nu_t, \widetilde{p}}(Y(t; y_0)) \Big\|^2 \, d\gamma_0(x_0, y_0) \right)^{1/2} \\ &\le \left(\int \Big\| V_{\mu_t, p}(X(t; x_0)) - V_{\mu_t, p}(Y(t; y_0)) \Big\|^2 \, d\gamma_0 \right)^{1/2} \\ &\quad + \left(\int \Big\| V_{\mu_t, p}(Y(t; y_0)) - V_{\nu_t, p}(Y(t; y_0)) \Big\|^2 \, d\gamma_0 \right)^{1/2} \\ &\quad + \left(\int \Big\| V_{\nu_t, p}(Y(t; y_0)) - V_{\nu_t, \widetilde{p}}(Y(t; y_0)) \Big\|^2 \, d\gamma_0 \right)^{1/2} \\ &\le L_1 D(t) + \|V_{\mu_t, p} - V_{\nu_t, p}\|_{L^2(\nu_t)} + \|V_{\nu_t, p} - V_{\nu_t, \widetilde{p}}\|_{L^2(\nu_t)} \\ &\le (L_1 + L_2) D(t) + L_3 \mathcal{W}_2(p, \widetilde{p}). \end{aligned}$$

Here we used that the law of $Y(t; y_0)$ under γ_0 is ν_t, together with $\mathcal{W}_2(\mu_t, \nu_t) \le D(t)$. Integrating the preceding inequality and letting $\delta \downarrow 0$ gives

$$D(t) \le D(0) + \int_0^t \Big((L_1 + L_2) D(s) + L_3 \mathcal{W}_2(p, \widetilde{p}) \Big) \, ds.$$

Since $D(0) = \mathcal{W}_2(\mu_0, \nu_0)$, Gronwall's inequality yields

$$\mathcal{W}_2(\mu_t, \nu_t) \le e^{(L_1 + L_2)t} \mathcal{W}_2(\mu_0, \nu_0) + \frac{L_3}{L_1 + L_2} \big(e^{(L_1 + L_2)t} - 1 \big) \mathcal{W}_2(p, \widetilde{p}). \tag{35}$$

In particular, taking $\widetilde{p} = p$ gives uniqueness.

We now prove existence. Let $(\widehat{q}_0^N)_N$ and $(\widehat{p}^M)_M$ be the empirical measures so that

$$\widehat{q}_0^N \to q_0, \qquad \widehat{p}^M \to p \qquad \text{in } \mathcal{W}_2$$

almost surely. Applying (35) to $\widehat{q}_t^{\,N,M}$ and $\widehat{q}_t^{\,N',M'}$ yields

$$\sup_{t \in [0,T]} \mathcal{W}_2(\widehat{q}_t^{N,M}, \widehat{q}_t^{N',M'}) \le e^{(L_1 + L_2)T} \mathcal{W}_2(\widehat{q}_0^N, \widehat{q}_0^{N'}) + \frac{L_3}{L_1 + L_2} \big(e^{(L_1 + L_2)T} - 1 \big) \mathcal{W}_2(\widehat{p}^M, \widehat{p}^{M'}).$$

Since $\widehat{q}_0^N \to q_0$ and $\widehat{p}^M \to p$ in $\mathcal{W}_2$ almost surely, the right-hand side tends to 0 as $N, N', M, M' \to \infty$. Hence $(\widehat{q}^{N,M})_{N,M}$ is a Cauchy family in $C([0,T]; \mathcal{P}_2(\mathbb{R}^d))$, where $\mathcal{P}_2(\mathbb{R}^d)$ is equipped with $\mathcal{W}_2$. Since $(\mathcal{P}_2(\mathbb{R}^d), \mathcal{W}_2)$ is complete, there exists

$$\overline{q} \in C([0,T]; \mathcal{P}_2(\mathbb{R}^d))$$

such that

$$\sup_{t\in[0,T]} \mathcal{W}_2(\widehat{q}_t^{N,M}, \overline{q}_t) \to 0 \qquad \text{almost surely} \qquad \text{as } N, M \to \infty. \tag{36}$$

In particular, $\overline{q}_0 = q_0$.

We also need a uniform second-moment bound. Set

$$m_2^{N,M}(t) := \left(\int \|x\|^2 \, d\widehat{q}_t^{N,M}(x) \right)^{1/2}.$$

For a.e. t, using the ODE and Assumption B.1,

$$\frac{d}{dt} m_2^{N,M}(t) \le \|V_{\widehat{q}_t^{N,M}, \widehat{p}^M}\|_{L^2(\widehat{q}_t^{N,M})} \le L_0 \big(m_2^{N,M}(t) + m_2(\widehat{p}^M) \big).$$

Therefore, by Gronwall's inequality,

$$m_2^{N,M}(t) \le e^{L_0 t} m_2(\widehat{q}_0^N) + \big(e^{L_0 t} - 1\big) m_2(\widehat{p}^M), \qquad 0 \le t \le T. \tag{37}$$

Since $\widehat{q}_0^N \to q_0$ and $\widehat{p}^M \to p$ in $\mathcal{W}_2$ almost surely, the sequences $m_2(\widehat{q}_0^N)$ and $m_2(\widehat{p}^M)$ are bounded. Hence there exists $C_T > 0$ such that

$$\sup_{N,M} \sup_{t\in[0,T]} \int \|x\|^2 \, d\widehat{q}_t^{N,M}(x) \le C_T.$$

From (37) and the convergence in (36), passing to the limit also gives

$$\sup_{t\in[0,T]} \int \|x\|^2 \, d\overline{q}_t(x) \le C_T.$$

It remains to pass to the limit in the weak formulation. Fix $\varphi \in C_c^1(\mathbb{R}^d)$. For each $s \in [0,T]$, let $\gamma_s^{N,M}$ be an optimal coupling of $\widehat{q}_s^{N,M}$ and $\overline{q}_s$. Write

$$I_{N,M}(s) := \int \nabla\varphi(x) \cdot V_{\widehat{q}_s^{N,M}, \widehat{p}^M}(x) \, d\widehat{q}_s^{N,M}(x),$$
$$I(s) := \int \nabla\varphi(y) \cdot V_{\overline{q}_s, p}(y) \, d\overline{q}_s(y).$$

Then

$$|I_{N,M}(s) - I(s)| \le \left| \int \big(\nabla\varphi(x) - \nabla\varphi(y)\big) \cdot V_{\widehat{q}_s^{N,M}, \widehat{p}^M}(x) \, d\gamma_s^{N,M}(x,y) \right|$$
$$+ \|\nabla\varphi\|_\infty \left(\int \big\| V_{\widehat{q}_s^{N,M}, \widehat{p}^M}(x) - V_{\overline{q}_s, p}(y) \big\|^2 \, d\gamma_s^{N,M}(x,y) \right)^{1/2}.$$

The second term is bounded by

$$\|\nabla\varphi\|_\infty \Big[(L_1 + L_2) \mathcal{W}_2(\widehat{q}_s^{N,M}, \overline{q}_s) + L_3 \mathcal{W}_2(\widehat{p}^M, p) \Big],$$

and converges to zero uniformly in s. For the first term, let ω_φ be a modulus of continuity of $\nabla\varphi$. For every $r > 0$, Cauchy–Schwarz and the optimality of $\gamma_s^{N,M}$ give

$$\left| \int \big(\nabla\varphi(x) - \nabla\varphi(y)\big) \cdot V_{\widehat{q}_s^{N,M}, \widehat{p}^M}(x) \, d\gamma_s^{N,M}(x,y) \right|$$
$$\le \omega_\varphi(r) \, \|V_{\widehat{q}_s^{N,M}, \widehat{p}^M}\|_{L^1(\widehat{q}_s^{N,M})} + 2\|\nabla\varphi\|_\infty \|V_{\widehat{q}_s^{N,M}, \widehat{p}^M}\|_{L^2(\widehat{q}_s^{N,M})} \frac{\mathcal{W}_2(\widehat{q}_s^{N,M}, \overline{q}_s)}{r}.$$

The velocity norms are uniformly bounded on $[0,T]$ by (27) and (37). Taking $N, M \to \infty$ and then $r \downarrow 0$ proves that $I_{N,M} \to I$ uniformly on $[0,T]$. Passing to the limit in the weak formulation gives

$$\int \varphi \, d\overline{q}_t - \int \varphi \, dq_0 = \int_0^t \int \nabla\varphi(x) \cdot V_{\overline{q}_s,p}(x) \, d\overline{q}_s(x) \, ds.$$

Thus $\overline{q}$ is a weak solution of (31). By uniqueness, $\overline{q}$ is the unique weak solution; we denote it by q.

Applying the stability estimate in (35) with

$$\mu_t = q_t, \qquad \nu_t = \widehat{q}_t^{N,M}, \qquad \widetilde{p} = \widehat{p}^M$$

gives, for all $t \in [0,T]$,

$$\mathcal{W}_2(\widehat{q}_t^{N,M}, q_t) \le e^{(L_1+L_2)t}\mathcal{W}_2(\widehat{q}_0^N, q_0) + \frac{L_3}{L_1+L_2}(e^{(L_1+L_2)t}-1)\mathcal{W}_2(\widehat{p}^M, p). \tag{38}$$

Taking the supremum over $t \in [0,T]$ gives the corresponding bound for the continuous-time particle system.

On the other hand, let $\boldsymbol{x}^{N,M,\eta}(t)$ be the piecewise linear Euler interpolation of (33), namely

$$\boldsymbol{x}^{N,M,\eta}(t_{k+1}) = \boldsymbol{x}^{N,M,\eta}(t_k) + \eta \, V^{N,M}(\boldsymbol{x}^{N,M,\eta}(t_k)),$$

and

$$\boldsymbol{x}^{N,M,\eta}(t) = \boldsymbol{x}^{N,M,\eta}(t_k) + (t-t_k)V^{N,M}(\boldsymbol{x}^{N,M,\eta}(t_k)), \qquad t \in [t_k, t_{k+1}).$$

Let

$$\widehat{q}_t^{N,M,\eta} := \frac{1}{N}\sum_{i=1}^N \delta_{x_i^{N,M,\eta}(t)}.$$

Set

$$e(t) := \|\boldsymbol{x}^{N,M,\eta}(t) - \boldsymbol{x}^{N,M}(t)\|_{2,N}.$$

Since $V^{N,M}$ is globally Lipschitz with constant L_1+L_2, the standard Euler estimate gives

$$\sup_{t\in[0,T]} e(t) \le \frac{1}{2}\big(e^{(L_1+L_2)T}-1\big)\eta \sup_{s\in[0,T]} \|V^{N,M}(\boldsymbol{x}^{N,M}(s))\|_{2,N}.$$

Moreover, by (27) and (37), for every $s \in [0,T]$,

$$\begin{aligned}\|V^{N,M}(\boldsymbol{x}^{N,M}(s))\|_{2,N} &= \|V^{\varepsilon}_{\widehat{q}_s^{N,M},\widehat{p}^M}\|_{L^2(\widehat{q}_s^{N,M})} \\ &\le L_0\big(m_2(\widehat{q}_s^{N,M}) + m_2(\widehat{p}^M)\big) \\ &\le L_0 e^{L_0 T}\big[m_2(\widehat{q}_0^N) + m_2(\widehat{p}^M)\big].\end{aligned}$$

Therefore,

$$\sup_{t\in[0,T]} e(t) \le \frac{L_0 e^{L_0 T}}{2}\big(e^{(L_1+L_2)T}-1\big)\eta\big[m_2(\widehat{q}_0^N) + m_2(\widehat{p}^M)\big]. \tag{39}$$

Since the coupling that matches $x_i^{N,M,\eta}(t)$ with $x_i^{N,M}(t)$ is admissible, we have

$$\mathcal{W}_2(\widehat{q}_t^{N,M,\eta}, \widehat{q}_t^{N,M}) \le \|\boldsymbol{x}^{N,M,\eta}(t) - \boldsymbol{x}^{N,M}(t)\|_{2,N} = e(t).$$

Combining this estimate with (38) and (39), and taking the supremum over $t \in [0,T]$, proves the quantitative bound in the theorem.

Finally, since $q_0, p \in \mathcal{P}_2(\mathbb{R}^d)$, the empirical measures satisfy

$$\mathcal{W}_2(\widehat{q}_0^N, q_0) \to 0, \qquad \mathcal{W}_2(\widehat{p}^M, p) \to 0, \qquad \text{almost surely.}$$

Moreover,

$$\int_{\mathbb{R}^d} \|x\|^2 \, d\widehat{q}_0^N(x) \to \int_{\mathbb{R}^d} \|x\|^2 \, dq_0(x), \qquad \int_{\mathbb{R}^d} \|x\|^2 \, d\widehat{p}^M(x) \to \int_{\mathbb{R}^d} \|x\|^2 \, dp(x), \qquad \text{almost surely.}$$

Letting $N, M \to \infty$ and $\eta \downarrow 0$, we obtain

$$\sup_{t\in[0,T]} \mathcal{W}_2(\widehat{q}_t^{N,M,\eta}, q_t) \to 0, \qquad \text{almost surely.}$$

□

B.2 Assumption verification for Sinkhorn

Recall that the Sinkhorn velocity in (10):

$$V^{\varepsilon}_{q,p}(x) = T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{q,q}(x),$$

where $T^{\varepsilon}_{q,p}$ denotes the barycentric projection of the entropic optimal coupling between q and p for the quadratic cost.

Before verifying Assumption B.1, we first localize the dynamics. If $\mathrm{supp}(q_0) \subset B_{R_0}$ and $\mathrm{supp}(p) \subset B_R$, then the WGF under Sinkhorn divergence remains supported in a bounded ball on every finite time interval; see Proposition B.5. On this bounded domain, we then use compact-domain stability and regularity estimates for Sinkhorn barycentric projections, collected in Lemma B.6. Combining these two ingredients allows us to verify Assumption B.1 for the Sinkhorn velocity along the localized flow and its finite-particle approximations; see Proposition B.7.

Proposition B.5 (Bounded support propagation for the Sinkhorn flow)**.** *Assume that*

$$\mathrm{supp}(q_0) \subset B_{R_0}, \qquad \mathrm{supp}(p) \subset B_R,$$

where $B_R := \{x \in \mathbb{R}^d : \|x\| \le R\}$. *Let* q_t *solve*

$$\partial_t q_t + \nabla \cdot (q_t V^{\varepsilon}_{q_t,p}) = 0.$$

Then, on every finite time interval $[0, T]$,

$$\mathrm{supp}(q_t) \subset B_{e^t R_0 + (e^t - 1)R}.$$

The same support bound holds for the discrete finite-particle dynamics in (25) and for the piecewise-linear interpolation in (26), provided the initial particles are supported in B_{R_0} and the target particles are supported in B_R.

Proof. Let us define the radius

$$\rho(t) := \sup\{\|x\| : x \in \mathrm{supp}(q_t)\}.$$

Since $T^{\varepsilon}_{q_t,p}(x)$ is the barycenter of a probability measure supported in $\mathrm{supp}(p) \subset B_R$, and $T^{\varepsilon}_{q_t,q_t}(x)$ is the barycenter of a probability measure supported in $\mathrm{supp}(q_t)$, we have

$$\|T^{\varepsilon}_{q_t,p}(x)\| \le R, \qquad \|T^{\varepsilon}_{q_t,q_t}(x)\| \le \rho(t).$$

Therefore,

$$\|V^{\varepsilon}_{q_t,p}(x)\| \le \|T^{\varepsilon}_{q_t,p}(x)\| + \|T^{\varepsilon}_{q_t,q_t}(x)\| \le R + \rho(t), \qquad x \in \mathrm{supp}(q_t).$$

Equivalently, for any particle trajectory $x(t)$ in the support of the flow,

$$\frac{d}{dt}\|x(t)\| \le \|V^{\varepsilon}_{q_t,p}(x(t))\| \le R + \rho(t).$$

Taking the supremum over all such trajectories gives

$$\rho(t) \le R_0 + \int_0^t (R + \rho(s))\, ds.$$

By Gronwall's inequality,

$$\rho(t) \le e^t R_0 + (e^t - 1)R.$$

This proves

$$\mathrm{supp}(q_t) \subset B_{e^t R_0 + (e^t - 1)R}.$$

For the continuous-time finite-particle dynamics, the same argument applies with q_t replaced by the empirical measure. For the Euler scheme, if $\widehat{q}^N_k$ is supported in B_{ρ_k}, then

$$\left\|x_i^{(k+1)}\right\| \le \left\|x_i^{(k)}\right\| + \eta \left\|V^{\varepsilon}_{\widehat{q}^N_k, \widehat{p}^M}(x_i^{(k)})\right\| \le \rho_k + \eta(R + \rho_k).$$

Hence

$$\rho_{k+1} \le (1 + \eta)\rho_k + \eta R.$$

Iterating this recursion gives

$$\rho_k \le (1 + \eta)^k R_0 + \big((1 + \eta)^k - 1\big)R \le e^{t_k} R_0 + (e^{t_k} - 1)R.$$

The same estimate controls the piecewise-linear interpolation between t_k and t_{k+1}. □

Lemma B.6 (Compact-domain stability of Sinkhorn barycentric projections)**.** *Let $K \subset \mathbb{R}^d$ be compact, let $D_K := \mathrm{diam}(K)$, and fix $\varepsilon > 0$. There exists $C_{K,\varepsilon} < \infty$ such that, for all $q, \widetilde{q}, p, \widetilde{p} \in \mathcal{P}(K)$,*

$$\sup_{x \in K} \left\| T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{\widetilde{q},\widetilde{p}}(x) \right\| \le C_{K,\varepsilon} \big(\mathcal{W}_2(q, \widetilde{q}) + \mathcal{W}_2(p, \widetilde{p}) \big), \tag{40}$$

and, for all $x, x' \in K$,

$$\left\| T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{q,p}(x') \right\| \le \frac{D_K^2}{4\varepsilon} \|x - x'\|. \tag{41}$$

Proof. We first prove (40). By the compact-domain stability result for entropic OT potentials (Carlier et al., 2024, Corollary 2.4), applied to the rescaled quadratic cost $(x, y) \mapsto \|x - y\|^2/(2\varepsilon)$ on $K \times K$, there exists $C_{K,\varepsilon} < \infty$ such that, for all $q, \widetilde{q}, p, \widetilde{p} \in \mathcal{P}(K)$,

$$\sup_{x \in K} \left\| \nabla u^{\varepsilon}_{q,p}(x) - \nabla u^{\varepsilon}_{\widetilde{q},\widetilde{p}}(x) \right\| \le C_{K,\varepsilon} \big(\mathcal{W}_2(q, \widetilde{q}) + \mathcal{W}_2(p, \widetilde{p}) \big),$$

where $u^{\varepsilon}_{q,p}$ denotes the Schrödinger potential introduced in (17). Since K is compact and the quadratic cost is smooth on $K \times K$, the assumptions required for (Carlier et al., 2024, Corollary 2.4) to hold are satisfied. Using the identity in (18), we immediately obtain (40).

It remains to relate the Schrödinger potential to the barycentric projection. Recall from (18) that $\nabla u^{\varepsilon}_{q,p}(x) = x - T^{\varepsilon}_{q,p}(x)$. Therefore,

$$\begin{aligned} \sup_{x \in K} \left\| T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{\widetilde{q},\widetilde{p}}(x) \right\| &= \sup_{x \in K} \left\| \nabla u^{\varepsilon}_{q,p}(x) - \nabla u^{\varepsilon}_{\widetilde{q},\widetilde{p}}(x) \right\| \\ &\le C_{K,\varepsilon} \big(\mathcal{W}_2(q, \widetilde{q}) + \mathcal{W}_2(p, \widetilde{p}) \big). \end{aligned}$$

This proves (40).

It remains to prove the spatial estimate. Since

$$T^{\varepsilon}_{q,p}(x) = \int_K y \, \pi^{\varepsilon}_{q,p}(dy \mid x),$$

differentiating the conditional mean with respect to x gives

$$\nabla_x T^{\varepsilon}_{q,p}(x) = \frac{1}{\varepsilon} \mathrm{Cov}_{\pi^{\varepsilon}_{q,p}(dy|x)}(Y).$$

For any probability measure supported on K, the operator norm of its covariance matrix is at most $D_K^2/4$. Therefore

$$\|\nabla_x T^{\varepsilon}_{q,p}(x)\|_{\mathrm{op}} \le \frac{D_K^2}{4\varepsilon}, \qquad x \in K.$$

This proves (41). □

Proposition B.7 (Verification of Assumption B.1 for bounded-support Sinkhorn flow)**.** *Assume*

$$\mathrm{supp}(q_0) \subset B_{R_0}, \qquad \mathrm{supp}(p) \subset B_R,$$

and fix $T > 0$. Then all measures appearing in the Sinkhorn flow and in the finite-particle approximations on $[0, T]$ are supported in $B_{\bar{R}_T}$ with

$$\bar{R}_T := \max\{R,\ e^T R_0 + (e^T - 1)R\}.$$

Moreover, for the Sinkhorn velocity $V^{\varepsilon}_{q,p}(x)$ defined in (10), Assumption B.1 holds on $\mathcal{P}(B_{\bar{R}_T})$, with constants depending only on (T, R_0, R, ε). More explicitly, there exist $L_0, L_1, L_2, L_3 < \infty$, depending only on $(\bar{R}_T, \varepsilon)$, such that (27)–(30) hold for all $q, \widetilde{q}, p, \widetilde{p} \in \mathcal{P}(B_{\bar{R}_T})$, with the spatial estimate required only for $x, x' \in B_{\bar{R}_T}$.

Proof. The support statement follows from Proposition B.5. It remains to verify Assumption B.1 on $B_{\bar{R}_T}$.

The L^2-growth estimate follows from Jensen's inequality. Specifically, if $(X, Y) \sim \pi^{\varepsilon}_{q,p}$, then

$$\int \|T^{\varepsilon}_{q,p}(x)\|^2 \, dq(x) = \int \|\mathbb{E}[Y \mid X = x]\|^2 \, dq(x) \le \int \|y\|^2 \, dp(y).$$

Similarly,

$$\int \|T^{\varepsilon}_{q,q}(x)\|^2 \, dq(x) \le \int \|y\|^2 \, dq(y).$$

Hence

$$\|V^{\varepsilon}_{q,p}\|_{L^2(q)} \le \|T^{\varepsilon}_{q,p}\|_{L^2(q)} + \|T^{\varepsilon}_{q,q}\|_{L^2(q)} \le m_2(p) + m_2(q),$$

so (27) holds with $L_0 = 1$.

By Lemma B.6, for $x, x' \in B_{\bar{R}_T}$,

$$\begin{aligned}\|V^{\varepsilon}_{q,p}(x) - V^{\varepsilon}_{q,p}(x')\| &\le \|T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{q,p}(x')\| + \|T^{\varepsilon}_{q,q}(x) - T^{\varepsilon}_{q,q}(x')\| \\ &\le \frac{\operatorname{diam}(B_{\bar{R}_T})^2}{2\varepsilon}\|x - x'\|.\end{aligned}$$

Thus (28) holds on $B_{\bar{R}_T}$.

For the q-stability estimate, Lemma B.6 gives

$$\begin{aligned}\sup_{x \in B_{\bar{R}_T}} \|V^{\varepsilon}_{q,p}(x) - V^{\varepsilon}_{\widetilde{q},p}(x)\| &\le \sup_{x \in B_{\bar{R}_T}} \|T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{\widetilde{q},p}(x)\| \\ &\quad + \sup_{x \in B_{\bar{R}_T}} \|T^{\varepsilon}_{q,q}(x) - T^{\varepsilon}_{\widetilde{q},\widetilde{q}}(x)\| \\ &\le 3C_{B_{\bar{R}_T},\varepsilon}\mathcal{W}_2(q, \widetilde{q}).\end{aligned}$$

Integrating with respect to $\widetilde{q}$ gives (29).

For the p-stability estimate, the self-interaction term cancels:

$$V^{\varepsilon}_{q,p}(x) - V^{\varepsilon}_{q,\widetilde{p}}(x) = T^{\varepsilon}_{q,p}(x) - T^{\varepsilon}_{q,\widetilde{p}}(x).$$

Therefore, by Lemma B.6,

$$\sup_{x \in B_{\bar{R}_T}} \|V^{\varepsilon}_{q,p}(x) - V^{\varepsilon}_{q,\widetilde{p}}(x)\| \le C_{B_{\bar{R}_T},\varepsilon}\mathcal{W}_2(p, \widetilde{p}).$$

Integrating with respect to q gives (30). □

C Additional implementation details

ImageNet experiments. We provide detailed hyperparameters and configurations in Table 6. Note that we largely follow the configurations and setups adopted in Deng et al. (2026), without extensive hyperparameter tuning. We conduct experiments on 8 Nvidia H100 GPU nodes, where each GPU has 80 GB of memory. The training on L/2 takes around 3 days.

FFHQ experiments. For the FFHQ experiments, we operate in the 512-dimensional latent space of a pretrained Adversarial Latent Autoencoder (ALAE), utilizing a training split of 60,000 images. The generator is parameterized as a 4-layer MLP with hidden dimensions of 1024, utilizing LayerNorm and SiLU activations in the hidden blocks. Models are trained using the AdamW optimizer with a constant learning rate of 2×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and zero weight decay. We apply gradient clipping at 5.0 and maintain an exponential moving average (EMA) of the weights with a decay rate of 0.999. During training, we use a batch size of $N = 64$ for the generated particles and $M = 64$ for the target particles, with a fixed step size of $\eta = 1$. For the Sinkhorn divergence, we use the quadratic cost with an entropy regularization parameter $\varepsilon = 0.02$, with 10 Sinkhorn iterations. Consistent with Sec. 3.3, we explicitly use the two-batch debiased self-transport estimator, avoiding any heuristic diagonal masking.

We evaluate this setup on two distinct tasks: domain transfer and mode coverage, as shown in Sec. 4.3. For the domain transfer task, we map a source distribution of senior faces (ages 55–100) to a target distribution of young adult faces (ages 18–30). The initial particles are drawn directly from the source data distribution, and we apply the residual identity initialization by zero-initializing the final layer of the MLP to learn the residual update. For the mode coverage experiment, we perform unconditional generation starting from a standard Gaussian prior. The target batches are constructed to be highly imbalanced, comprising 95% senior faces and 5% child faces (ages 0–12). In this setting, the standard network initialization is used without the residual connection.

Table 6: Configurations and hyperparameters for ImageNet experiments.

	Ablation setup (Table 2)	**W-Flow B/2** (Table 4)	**W-Flow L/2** (Table 4)	**W-Flow XL/2** (Table 4)
Generator Architecture				
arch	DiT-B/2	DiT-B/2	DiT-L/2	DiT-XL/2
input size	32×32×4	32×32×4	32×32×4	32×32×4
patch size	2×2	2×2	2×2	2×2
hidden dim	768	768	1024	1152
depth	12	12	24	28
register tokens	16	16	16	16
style embedding tokens	32	32	32	32
Feature Encoder				
architecture	ResNet	ResNet	ResNet	ResNet
SSL pre-train method	latent-MAE	latent-MAE	latent-MAE	latent-MAE
ResNet: input size	32×32×4	32×32×4	32×32×4	32×32×4
ResNet: conv_1 stride	1	1	1	1
ResNet: base width	256	640	640	640
ResNet: block type	bottleneck			
ResNet: blocks / stage	[3, 4, 6, 3]			
ResNet: size / stage	$[32^2, 16^2, 8^2, 4^2]$			
MAE: masking ratio	50%			
MAE: pre-train epochs	192	1280	1280	1280
classification finetune	No	3k steps	3k steps	3k steps
Generator Optimizer				
optimizer	AdamW (β_1 = 0.9, β_2 = 0.95)			
learning rate	2e-4	4e-4	4e-4	3e-4
weight decay	0.01	0.0	0.01	0.01
warmup steps	5k	10k	10k	10k
gradient clip	2.0	2.0	2.0	2.0
training steps	30k	200k	200k	200k
training epochs	100	1280	1280	1280
EMA decay	0.999	0.999	0.999	0.999
Training Loss Computation				
class labels N_c	64	128	128	128
positive samples N_{pos}	64	64	64	128
generated samples N_{neg}	64	64	64	64
effective batch B ($N_c \times N_{neg}$)	4096	8192	8192	8192
iteration L	10	10	1	1
regularization ε	0.05	0.05	0.05	0.01
step size η	1.0	1.0	1.0	1.0
CFG Configuration				
train: CFG w range	$[0, 3]$	$[0, 3]$	$[0, 3]$	$[0, 3]$
train: CFG w sampling	$p(w) \propto (w+1)^{-3}$	$p(w) \propto (w+1)^{-5}$	50%: w=0, 50%: $p(w) \propto (w+1)^{-3}$	20%: w=0, 80%: $p(w) \propto (w+1)^{-4}$
train: uncond samples N_{uncond}	16	32	32	32
inference: CFG w search	[0.0, 2.5]			

Settings for plotting Fig. 2 (b). We train W-Flow with the Sinkhorn divergence on toy datasets for 20,000 iterations, saving a checkpoint every 400 iterations. This yields a sequence of parameter vectors $\{\theta_k\}_{k=0}^{K}$, where $K = 50$. We visualize the Sinkhorn-divergence energy landscape over a two-dimensional affine slice of the parameter space. Each point (x, y) on the x-y plane is mapped to a parameter vector $\theta_{x,y}$, and the corresponding height is given by $S_\varepsilon(q_{\theta_{x,y}}, p)$, where $q_{\theta_{x,y}}$ denotes the model distribution induced by $\theta_{x,y}$.

We construct the x-y grid as follows. First, we place the training trajectory along the x-axis by assigning checkpoint θ_k to coordinate $(k, 0)$ for $k = 0, \ldots, K$. Let

$$d_k = \theta_{k+1} - \theta_k, \qquad k = 0, \ldots, K-1,$$

denote the displacement directions along the checkpoint trajectory. We then sample a random direction v in parameter space and project it onto the orthogonal complement of $\text{span}\{d_0, \ldots, d_{K-1}\}$, followed by normalization. This direction v is used as a common transverse direction for the y-axis.

For each interval $x \in [k, k+1]$, we interpolate along the local trajectory direction d_k and extend transversely along v:

$$\theta_{x,y} = \theta_k + (x-k)d_k + yv, \qquad k \le x \le k+1, \quad -Y \le y \le Y.$$

This defines the landscape grid over $0 \le x \le K$ and $-Y \le y \le Y$, while ensuring that the saved checkpoints lie exactly on the x-axis.

To visualize the landscape beyond the training trajectory, we further extend the x-axis to $-X \le x \le K + X$. For the left extension $-X \le x \le 0$, we use the initial trajectory direction $d_0 = \theta_1 - \theta_0$:

$$\theta_{x,y} = \theta_0 + xd_0 + yv.$$

For the right extension $K \le x \le K + X$, we use the final trajectory direction $d_{K-1} = \theta_K - \theta_{K-1}$:

$$\theta_{x,y} = \theta_K + (x-K)d_{K-1} + yv.$$

Together, this construction yields a continuous two-dimensional slice of the parameter space on which we evaluate and plot the Sinkhorn-divergence energy.

D SAMPLING THROUGHPUT COMPARISON

We compare our one-step W-Flow against multi-step diffusion model SiT (Ma et al., 2024) and LightningDiT (Yao et al., 2025) in terms of actual sampling throughput measured by wall-clock time. We benchmark on a single Nvidia H100 GPU with a batch size of 4 for all models. For SiT-XL/2 and LightningDiT-XL/2, we adopt 250 sampling steps with CFG enabled and strictly follow the configuration in the original papers (Ma et al., 2024; Yao et al., 2025) that leads to a reported FID of 2.06 and 1.35, respectively. Note that we also include the latent preparation and VAE decoding time when recording the actual runtime and calculating the throughput. The results are depicted in Table 7, where W-Flow clearly yields significant speedup while achieving competitive generation quality measured by FID on ImageNet 256×256.

Table 7: **Generation quality and throughput comparison.** Throughput is measured in *images per second*. The time is recorded as the *actual wallclock time*. The speedup is computed *w.r.t.* SiT-XL/2.

	FID ↓	Throughput (images/sec) ↑	Speedup ↑
W-Flow, B/2	1.52	107.76	115.87×
W-Flow, L/2	1.35	84.88	91.27×
W-Flow, XL/2	1.29	77.88	83.74×
LightningDiT-XL/2 (Yao et al., 2025)	1.35	1.14	1.22×
SiT-XL/2 (Ma et al., 2024)	2.06	0.93	–

E ADDITIONAL EXPERIMENTAL RESULTS

Ablation study on the empirical batch size. We systematically study the finite-sample behavior of W-Flow in Table 8. Let C denote the number of classes sampled at each training step, and N and M the numbers of generated (negative) and data (positive) particles per class, respectively. We independently investigate the number of available positive particles, the sample size used to estimate each class-conditional Sinkhorn coupling, and the effective training batch size.

Number of positive particles. As shown in Table 8(a), W-Flow consistently outperforms Drifting Models across the entire sweep of M. Notably, with only $M = 16$ positive particles per class, W-Flow achieves an FID of 8.61, comparable to the 8.46 FID obtained by Drifting Models using $M = 64$. This suggests that the distribution-level transport supervision remains effective even when positive samples are limited.

Sinkhorn solver sample size. Table 8(b) varies the number of samples used to estimate each class-conditional Sinkhorn coupling while keeping the overall training batch size fixed. W-Flow remains robust even with very small empirical measures: using only $N = M = 8$ particles per class yields an FID of 8.92, outperforming Drifting Models with $N = M = 32$ (9.32 FID). Thus, finite-sample estimation of the Sinkhorn velocity does not cause substantial empirical degradation even at relatively small sample sizes.

Effective batch size. Finally, Table 8(c) studies the overall training batch size. Reducing it from 4096 to 2048 changes FID only from 7.08 to 7.78, while still outperforming Drifting Models trained with batch size 4096 (8.46 FID). Moreover, at a fixed effective batch size, allocating the sampling budget across more classes generally performs at least as well as allocating more particles to each class.

Sinkhorn iterations. We provide more ablation results on number of Sinkhorn iterations L used when computing the Sinkhorn barycentric projection (Alg. 2). We follow the same ablation setup in Sec. 4.1 with distribution guidance. The results in Table 9 show that performance improves as the number of Sinkhorn iterations increases, plateauing at around 10 iterations.

We further investigate the convergence of the Sinkhorn iteration. Using the trained W-Flow B/2 model, we sample 32 empirical batches and run the same OT pipeline used during training while

Table 8: **Finite-batch ablations of W-Flow.** We study the effects of (a) the number of positive particles M, (b) the sample size (N, M) used for each class-conditional Sinkhorn problem, and (c) the effective training batch size. Lower FID is better.

(a) Positive particles M

Method	M	N	FID↓
Drifting	1	64	20.43
	16	64	10.39
	32	64	8.97
	64	64	8.46
W-Flow	1	64	15.27
	16	64	8.61
	32	64	7.25
	64	64	7.08

(b) Sinkhorn sample size

Method	C	N	M	FID↓
Drifting	512	8	8	11.82
	256	16	16	10.16
	128	32	32	9.32
	64	64	64	8.46
W-Flow	512	8	8	8.92
	256	16	16	7.73
	128	32	32	7.31
	64	64	64	7.08

(c) Effective batch size

C	N	Batch	FID↓
64	8	512	11.83
8	64	512	12.68
64	16	1024	9.12
16	64	1024	9.51
64	32	2048	7.78
32	64	2048	7.79
64	64	4096	7.08

Table 9: Ablation on Sinkhorn iterations L.

L	FID, 1-NFE ↓
1	7.92
5	7.45
10	7.29
20	7.33

Table 10: Velocity guidance vs. distribution guidance applied to KL divergence.

CFG	FID, 1-NFE ↓
`dist.`	10.17
`velo.`	7.68

Table 11: Ablation on ε with ℓ_2 OT cost.

ε	FID, 1-NFE ↓
0.025	7.19
0.05	7.16
0.075	7.40

recording, at each Sinkhorn iteration, the marginal constraint violation and the error of the resulting barycentric transport target. We use the solution obtained after 40 iterations as the reference target.

As shown in Table 12, the Sinkhorn solution converges rapidly. By iteration 10, the average marginal violation decreases from 0.29 to 0.0041, corresponding to roughly two orders of magnitude, while the barycentric transport target differs from the 40-iteration reference by only 0.30%. Increasing the number of iterations to 20 further reduces the relative error to 0.05%. Thus, although additional Sinkhorn iterations continue to improve numerical convergence, the transport target used for training is already effectively converged after approximately 10 iterations. Consequently, additional iterations produce only negligible changes to the training signal, consistent with the observed saturation in downstream performance.

Velocity guidance. We further show in Table 10 that the proposed velocity guidance not only enhances the performance of Sinkhorn divergence-driven WGF, but also that of KL divergence. This observation aligns with our theoretical analysis in Appendix A.3, where we draw a connection to the exponentially tilted target distribution when using velocity guidance with KL divergence. Note that KL divergence with velocity guidance also outperforms the design in Deng et al. (2026) with 8.46 FID, further highlighting the significance of our principled WGF framework.

Entropic regularization parameter ε. We provide more results under the ablation study setting for the entropic regularization parameter ε with the ℓ_2 distance cost, under velocity guidance. As shown in Table 11, the ℓ_2 distance OT cost yields consistently higher FIDs than the quadratic OT cost, while both achieve the best performance under $\varepsilon = 0.05$.

Computational and memory analysis. We decompose the computation at each training step into three components: generator forward/backward, feature encoding, and the Sinkhorn solver with the OT loss. We systematically profile the computational cost of each component, measured in TFLOPs per GPU, under the large-scale setting used in Table 3. The results are summarized in Table 13.

The Sinkhorn solver and OT loss constitute only a negligible fraction of the total computation, while the generator forward/backward pass and feature encoding dominate the overall cost. In particular, the Sinkhorn component requires only 0.82 TFLOPs per GPU across all model scales, corresponding to less than 0.6% of the total computation. This demonstrates that solving the transport coupling introduces minimal computational overhead relative to the remaining training pipeline.

Table 12: **Convergence of the Sinkhorn solver.** We report the marginal constraint violation and the error of the barycentric transport target relative to the 40-iteration solution, averaged over 32 empirical batches.

Iterations	Marginal violation	Absolute error	Relative error
1	0.29	1.76	9.1%
5	0.019	0.278	1.2%
10	0.0041	0.082	0.30%
20	0.00072	0.017	0.05%

Table 13: **Computational cost per training step.** We report TFLOPs per GPU for different generator scales. The Sinkhorn solver and OT loss account for only a small fraction of the overall computation.

Stage	B/2	L/2	XL/2
Generator forward/backward	55.2	140.9	192.3
Feature encoding	103.7	103.7	103.7
Sinkhorn solver + OT loss	0.82	0.82	0.82
Total	159.7	245.4	296.8

The memory overhead is similarly small. We compute the Sinkhorn coupling under `no_grad` using detached features, such that the Sinkhorn iterations do not contribute to the autograd graph. Consequently, the peak training memory is primarily determined by the generator and feature-encoder activations. The coupling matrices themselves occupy only approximately 0.1GB of memory and are released before backpropagation. Therefore, the Sinkhorn solver introduces negligible additional memory overhead.

Precision and recall evaluation. To further assess mode coverage on ImageNet, we evaluate generation precision and recall following the standard protocol and compare W-Flow against iMeanFlow and Drifting Models. The results are summarized in Table 14.

Table 14: **Precision and recall evaluation.** W-Flow improves recall while maintaining competitive precision, together with consistent gains in FID.

	FID↓	Precision↑	Recall↑
iMeanFlow, XL/2	1.82	0.78	0.63
Drifting, B/2	1.75	0.78	0.61
Drifting, L/2	1.54	0.78	0.62
W-Flow, B/2	1.52	0.80	0.62
W-Flow, L/2	1.35	0.80	0.64
W-Flow, XL/2	1.29	0.78	0.67

W-Flow consistently improves the quality–coverage trade-off. Across comparable model scales, it matches or improves precision while achieving higher recall, with the largest recall reaching 0.67 for the XL/2 model. These gains are accompanied by consistently lower FID. The improved recall provides additional evidence that the Sinkhorn-induced velocity field promotes broader coverage of the target distribution rather than improving sample quality alone.

F Additional generated samples

In this section, we provide additional generated image samples. We provide uncurated samples on ImageNet 256×256 in Fig. 11 and Fig. 12, and additional samples in Fig. 13 with a larger guidance scale. We provide uncurated samples on FFHQ for the mode coverage experiment in Fig. 14 and Fig. 15 for Drifting Model and our W-Flow, respectively.



Class 012: house finch, linnet, Carpodacus mexicanus



Class 031: tree frog, tree-frog



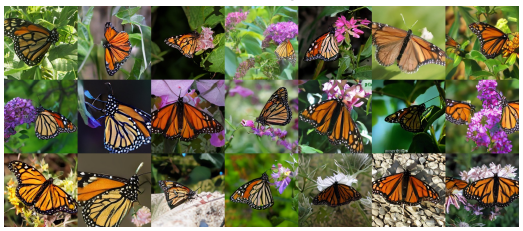
Class 022: bald eagle, American eagle, Haliaeetus leucocephalus



Class 992: agaric



Class 698: palace



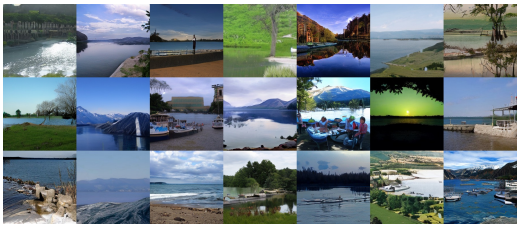
Class 323: monarch, …, Danaus plexippus



Class 927: trifle



Class 985: daisy



Class 975: lakeside, lakeshore



Class 001: goldfish, Carassius auratus



Class 113: snail



Class 980: volcano

Figure 11: Uncurated samples generated by W-Flow, L/2 with CFG $w = 0.15$.

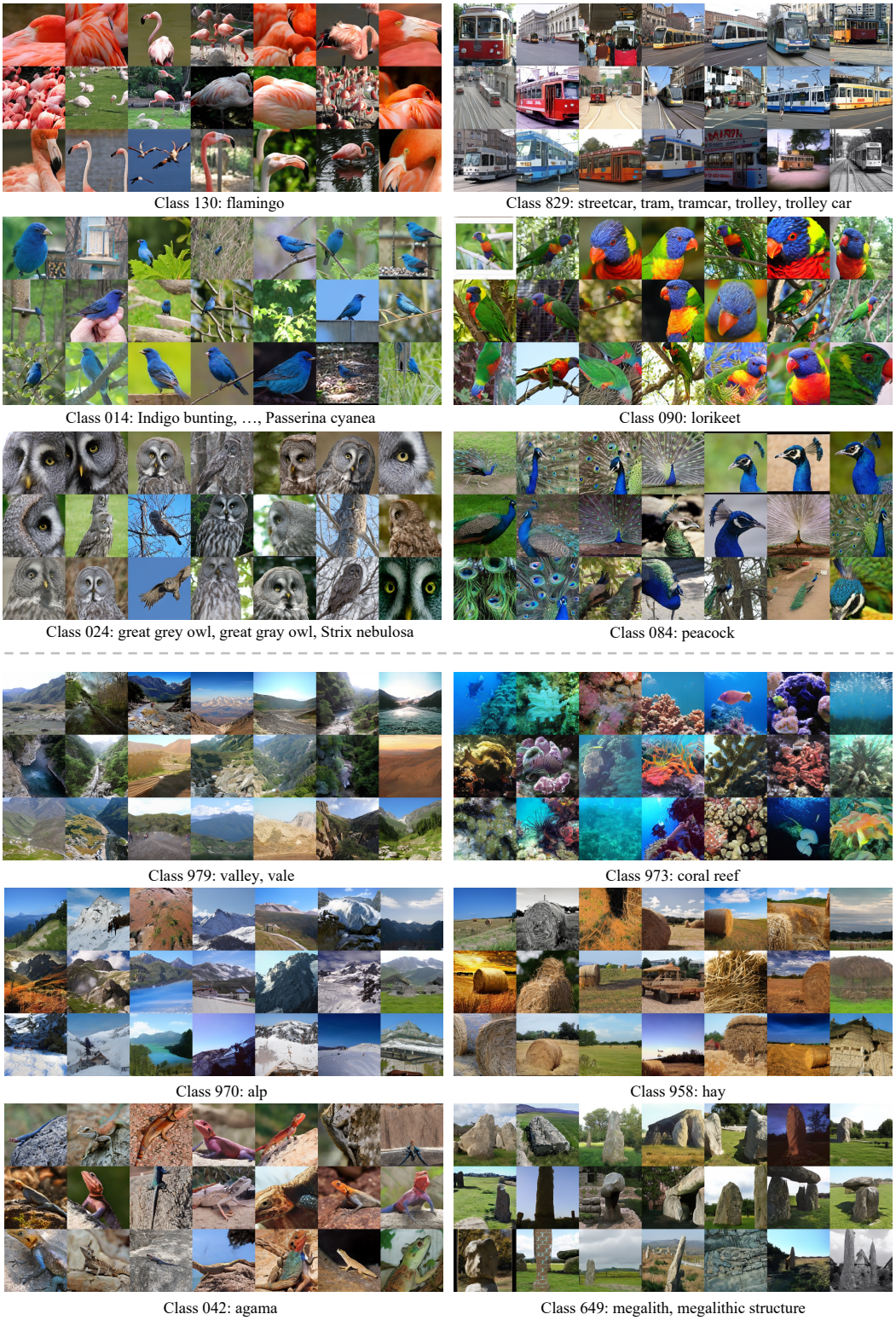


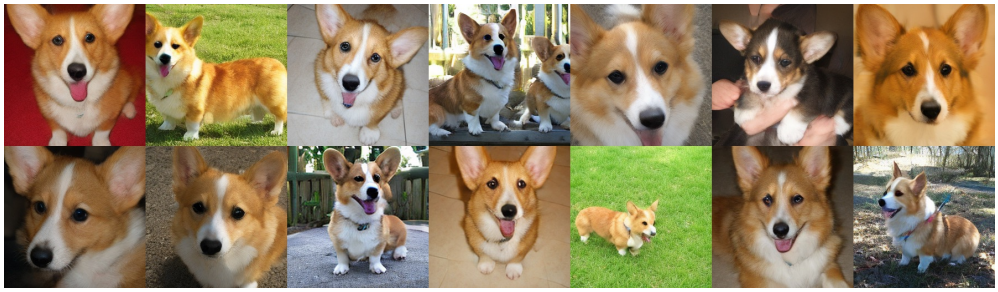
Figure 12: Uncurated samples generated by W-Flow, XL/2 *(Top)* and B/2 *(Bottom)* with CFG $w = 0.15$.



Class 829: streetcar, tram, tramcar, trolley, trolley car



Class 291: lion, king of beasts, Panthera leo



Class 263: Pembroke, Pembroke Welsh corgi



Class 817: sports car, sport car



Class 145: king penguin, Aptenodytes patagonica

Figure 13: Uncurated samples generated by W-Flow, XL/2 with CFG $w = 2.0$.



Figure 14: Uncurated samples generated by Drifting Model in the mode coverage experiment (Sec. 4.3). None of the 100 randomly generated images is a child's face, showing failure in covering the minority mode.



Figure 15: Uncurated samples generated by W-Flow in the mode coverage experiment (Sec. 4.3). 4 of the 100 randomly generated images are child faces, which is close to the 5% child face ratio in the down-sampled population.