ONE-STEP GENERATIVE MODELING VIA UNBALANCED OPTIMAL TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Drifting models enable one-step generation by amortizing distribution transport into training, but this efficiency places greater demands on the transport field estimated at each update. In large-scale training, the field is computed from finite mini-batches of generated and real samples, which provide only imperfect approximations to the underlying distributions. Balanced optimal transport enforces exact mass matching within every mini-batch, making the estimated field sensitive to the particular composition of the real-data batch. We find that generated and real samples should be treated asymmetrically: letting the mass assigned to real samples adapt while keeping every generated sample fully transported improves generation across six feature-space metrics in controlled ablations, and is more robust to the relaxation strength than relaxing both marginals simultaneously, which falls below balanced transport under stronger relaxation. Motivated by this observation, we propose Unbalanced Optimal Transport Gradient Flow (UOT-GF), which keeps the generated-sample marginal fixed and relaxes only the real-data marginal. Under identical settings at DiT-B/2 on ImageNet-256, UOT-GF improves Fréchet Inception Distance (FID) from 1.53 to 1.46 over the balanced W-Flow baseline; scaling the same recipe yields 1.34 and 1.22 FID at L/2 and XL/2, the best FID among the one-step models we compare. We further derive the induced UOT transport force, establish a kinetic Vlasov–Fokker–Planck formulation whose overdamped zero-temperature limit recovers the drifting dynamics, and characterize non-target stationary states together with sufficient conditions for convergence.

1 INTRODUCTION

Generative modeling seeks to transform samples from a simple reference distribution into samples from the data distribution. Diffusion and flow-based models achieve high generation quality but often require repeated network evaluations at inference time (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Liu et al., 2023). GANs reduce sampling to a single forward pass, but introduce adversarial optimization and its associated training difficulties (Goodfellow et al., 2014; Karras et al., 2019). Autoregressive and bridge-based approaches offer further alternatives, with different trade-offs between generation speed, modeling flexibility, and sample quality (Li et al., 2024; Tian et al., 2024; Shi et al., 2023; Bortoli et al., 2021; Pariset et al., 2023). Drifting models use a different strategy: rather than integrating a long trajectory at test time, they learn how generated particles should move toward the data distribution during training, enabling one-step sampling (Deng et al., 2026). W-Flow (Han et al., 2026) interprets these dynamics as a Wasserstein gradient flow induced by the balanced Sinkhorn divergence and reports strong one-step performance on ImageNet-256.

The practical accuracy of such a drifting update depends on how well its transport direction can be estimated during training. At each iteration, only a limited set of generated and real examples participates in the transport problem (Han et al., 2026). For large-scale datasets with substantial intra-class or local variation, the samples present in one batch need not reflect the relative prevalence of the underlying data patterns: some may be over-represented, while others may be missing altogether. ImageNet provides a representative example, where individual classes can exhibit considerable variation in appearance, pose, viewpoint, and background (Deng et al., 2009; Russakovsky et al., 2015; Tsipras et al., 2020). Balanced transport still assigns a prescribed amount of mass to every sample in the batch. Consequently, the resulting coupling may be influenced by incidental batch composition (Fatras et al., 2021) rather than only by the population-level structure one ultimately

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.

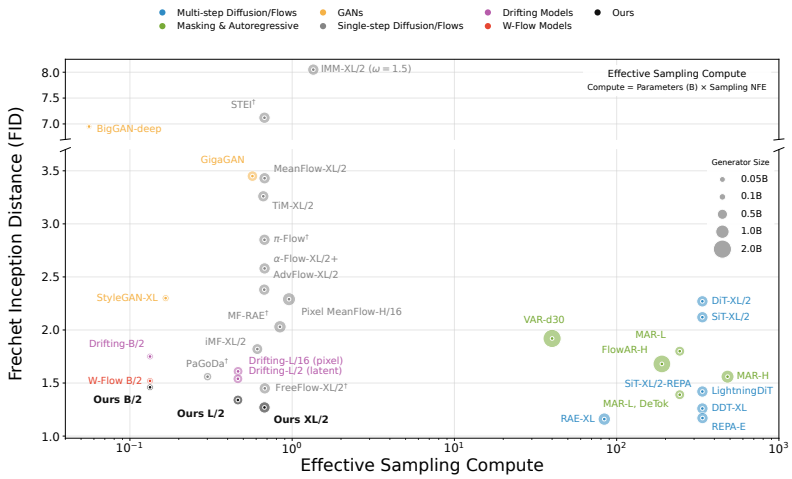


Figure 1: Fréchet Inception Distance (FID) versus effective sampling compute (parameters × NFE (Song et al., 2023)) on ImageNet-256. 1-NFE samples are shown in Appendix J.

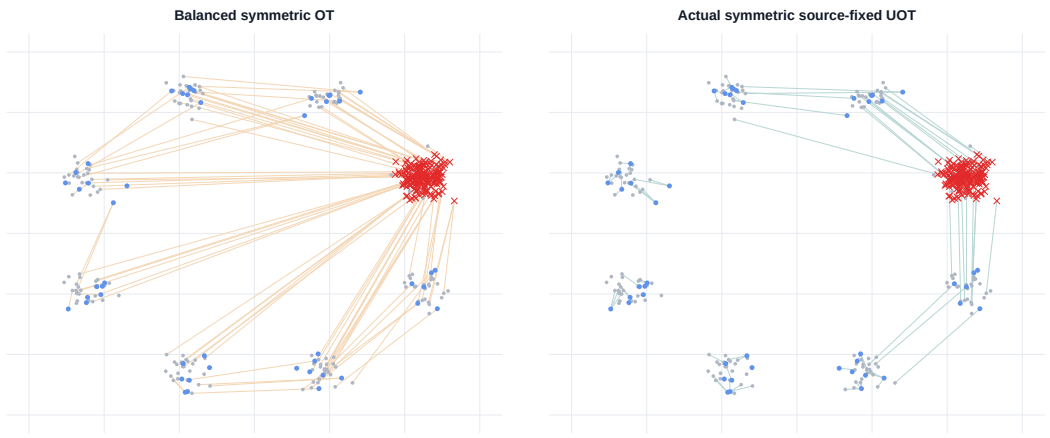
wishes to learn. This raises a natural question: should the generated and real samples be subject to the same marginal constraints in drifting-based training?

To address this issue, we propose Unbalanced Optimal Transport Gradient Flow (UOT-GF). Unlike semi-dual, adversarial, or JKO-type uses of unbalanced transport (Balaji et al., 2020; Choi et al., 2023; 2024), it enters as the geometry of an explicit, simulation-free drifting update, making the marginal relaxation a design axis: it defines a family of transport energies, each inducing its own flow, with balanced Sinkhorn drifting recovered as the $\tau \to 1$ member. Our central finding is that *which marginal is relaxed matters*: relaxing the real-data marginal alone is at least as effective as relaxing both, and is more robust to the relaxation strength—at stronger relaxation, relaxing both marginals degrades to worse than balanced transport while symmetrized source-fixed transport still improves on it. Each directed transport problem fixes its source marginal and relaxes its target marginal, and the full interaction averages contributions from both directions. Generated and real samples therefore each appear once on the fixed side and once on the relaxed side. The effective cross-interaction plan permits marginal reweighting while retaining at least half of each distribution's reference mass, as shown in Section 3.1. This structural property does not by itself guarantee target coverage, which we evaluate empirically in Section 5.Our contributions are as follows:

- We study marginal relaxation in gradient-flow drifting and show empirically that symmetrized source-fixed transport improves generation over balanced transport across six feature-space metrics, and is markedly more robust to the relaxation strength than relaxing both marginals, which falls below balanced transport under stronger relaxation.
- We develop a unified kinetic theory connecting stochastic dynamics to deterministic drifting, with balanced Sinkhorn drifting as a special case. We establish free-energy dissipation, reveal non-target stationary states inherent to the drifting paradigm, and give sufficient conditions for reaching the target.
- On ImageNet-256, relaxation improves balanced transport (Han et al., 2026) at B/2 under identical settings (1.53 → 1.46 FID), and scaling one fixed recipe yields 1.22 FID at XL/2, a new one-step state of the art among models trained from scratch (Figure 1).

2 Preliminary

Generative modeling can be formulated as learning a transport map $f : \mathbb{R}^{d_0} \to \mathbb{R}^d$ that pushes a simple reference distribution $p_{\text{ref}} \in \mathcal{P}(\mathbb{R}^{d_0})$ toward the data distribution $p := p_{\text{data}} \in \mathcal{P}(\mathbb{R}^d)$. Writing $f_\#$ for the pushforward operator, the goal is to learn f such that $f_\# p_{\text{ref}} \approx p$. Generative models parameterize f by a neural network $f_{\boldsymbol{\theta}} : \mathbb{R}^{d_0} \to \mathbb{R}^d$ and define the induced model distribution $q_{\boldsymbol{\theta}} := (f_{\boldsymbol{\theta}})_\# p_{\text{ref}}$, so that sampling $\mathbf{z} \sim p_{\text{ref}}$ and computing $\mathbf{x} = f_{\boldsymbol{\theta}}(\mathbf{z})$ yields $\mathbf{x} \sim q_{\boldsymbol{\theta}}$. Training seeks to minimize the discrepancy between $q_{\boldsymbol{\theta}}$ and p.



(a) Balanced OT (b) Symmetrized source-fixed UOT

Figure 2: **Oversampled mode in a minibatch.** One mode supplies 80% of the targets (red crosses). Balanced OT (a) must deliver full mass to them, pulling sources from distant modes; symmetrized source-fixed UOT (b) reweights them and keeps most couplings local. Gray: sources; blue: remaining targets. More examples in Appendix G.9.

2.1 Wasserstein Gradient Flow

Beyond the static pushforward formulation above, a generative model can be viewed at the distributional level as a dynamical system that continuously evolves the model distribution toward p. Let $\mathcal{F} : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ be an energy functional and write $\frac{\delta \mathcal{F}}{\delta q}(q)$ for its first variation at q (Definition 1), and let $\mathbf{v}_q := -\nabla \frac{\delta \mathcal{F}}{\delta q}(q)$ denote the induced velocity field. A curve of densities $\{q_t\}_{t \geq 0}$ is a *Wasserstein gradient flow* (WGF) of $\mathcal{F}$ if it solves the continuity equation

$$\partial_t q_t = -\nabla \cdot (q_t \, \mathbf{v}_{q_t}), \tag{1}$$

that is, mass is transported along the steepest-descent direction of $\mathcal{F}$ in the 2-Wasserstein geometry. Along any sufficiently regular solution the energy dissipates as

$$\frac{d}{dt}\mathcal{F}[q_t] = -\mathbb{E}_{q_t}\left[\|\mathbf{v}_{q_t}\|^2\right] \leq 0. \tag{2}$$

To drive the flow toward p, one takes $\mathcal{F}[q] = \mathsf{D}(q \,\|\, p)$ for a nonnegative, definite divergence D, i.e., $\mathsf{D}(q\|p) \geq 0$ with equality exactly when $q = p$, so that p is the unique global minimizer and $\mathbf{v}_p = \mathbf{0}$. Equation 2 guarantees that $\mathcal{F}[q_t]$ is non-increasing, but not that it reaches zero: Wasserstein stationarity requires $\mathbf{v}_{q_t} = \mathbf{0}$ only on $\mathrm{supp}(q_t)$, and non-target stationary states can exist. We return to this in Section 3.3.

The framework recovers familiar dynamics. For $\mathcal{F} = \mathrm{KL}(\cdot \,\|\, p)$, equation 1 is the Fokker–Planck equation $\partial_t q_t = -\nabla \cdot (q_t \nabla \log p) + \Delta q_t$, whose drift is the score of p; score-based models learn the scores of a time-indexed family of perturbed targets and integrate this equation numerically at inference time (Song & Ermon, 2019; Ho et al., 2020; Song et al., 2021; 2023).

2.2 Drifting Models

Unlike score-based diffusion and ODE-based flow models, which numerically integrate the sample dynamics over multiple steps at inference time, drifting models amortize this transport into training: the generator $f_{\boldsymbol{\theta}}$ is updated so that $q_{\boldsymbol{\theta}} = (f_{\boldsymbol{\theta}})_{\#} p_{\mathrm{ref}}$ approaches p, and sampling afterwards requires a single evaluation of $f_{\hat{\boldsymbol{\theta}}}$.

Let $f_n := f_{\boldsymbol{\theta}_n}$ and $q_n := (f_n)_{\#} p_{\mathrm{ref}}$ denote the generator and its induced distribution at iteration n. For a fixed $\mathbf{z} \sim p_{\mathrm{ref}}$, updating $\boldsymbol{\theta}_n$ to $\boldsymbol{\theta}_{n+1}$ displaces the generated sample by $\Delta \mathbf{x}_n := f_{n+1}(\mathbf{z}) - f_n(\mathbf{z})$. Drifting models train this displacement to follow a prescribed velocity field $\mathbf{v}_{q_n}$ that depends on the current model distribution:

$$f_{n+1}(\mathbf{z}) \approx f_n(\mathbf{z}) + h_n \, \mathbf{v}_{q_n}\big(f_n(\mathbf{z})\big), \tag{3}$$

with step size $h_n > 0$. If the regression target is realized exactly, equation 3 is an explicit Euler step of the mean-field dynamics $\dot{\boldsymbol{x}}_t = \mathbf{v}_{q_t}(\boldsymbol{x}_t)$, and when $\mathbf{v}_q$ is the WGF velocity equation 1 of an energy $\mathcal{F}$, the induced q_t follows the corresponding WGF. The idealized update is stationary at an equilibrium $\mathbf{v}_{q_{\hat{\boldsymbol{\theta}}}} = \mathbf{0}$ on $\mathrm{supp}(q_{\hat{\boldsymbol{\theta}}})$, i.e., $f_{\hat{\boldsymbol{\theta}}}$ is a fixed point of equation 3; by Section 3.3, such an equilibrium need not satisfy $q_{\hat{\boldsymbol{\theta}}} = p$. In practice $\mathbf{v}_{q_n}$ is estimated from mini-batches of generated and real samples (Section 4), so the field actually followed is a finite-sample surrogate of equation 1.

The choice of $\mathbf{v}_q$ is what distinguishes existing instances. The original Drifting model (Deng et al., 2026) prescribes it heuristically through kernel-based attraction toward p and repulsion from q. W-Flow (Han et al., 2026) instead takes $\mathcal{F}$ to be the debiased entropic Sinkhorn divergence $S_\varepsilon(q, p)$ (Feydy et al., 2018), whose WGF velocity is the difference of two barycentric projections,

$$\mathbf{v}_q(\mathbf{x}) = B_{q,p}(\mathbf{x}) - B_{q,q}(\mathbf{x}), \qquad B_{q,\alpha}(\mathbf{x}) := \mathbb{E}_{\pi^{q,\alpha}}[\mathbf{y} \mid \mathbf{x}], \tag{4}$$

where $\pi^{q,\alpha}$ is the optimal plan of *balanced* entropic transport between q and α, in which both marginals are matched exactly. This is the field we generalize: Section 3.1 replaces the balanced energy by an unbalanced one and derives the resulting force, and Section 3.2 recovers the mean-field dynamics underlying equation 3 as the overdamped, zero-temperature limit of a kinetic lifting.

3 Theoretical analysis

Several recent works have interpreted drifting models through Wasserstein gradient flows (He et al., 2026; Caucheteux et al., 2026; Cao et al., 2026; Gretton et al., 2026; Dumont et al., 2026; Han et al., 2026). In particular, W-Flow (Han et al., 2026) formulates drifting as the gradient flow of the balanced Sinkhorn divergence. Our method instead relaxes only the real-data marginal while keeping the generated-sample marginal fixed. Because this asymmetric transport changes both the energy and the induced force field, the existing balanced-flow analysis does not directly apply.

To accommodate this setting, we first derive the transport force induced by a generalized UOT energy and show that balanced W-Flow is recovered in the hard-constraint limit (Section 3.1). We then develop a nonlinear Vlasov–Fokker–Planck formulation, establish free-energy dissipation, and recover the first-order drifting dynamics through overdamped and zero-temperature limits (Section 3.2). Finally, we analyze long-time behavior, identifying possible non-target stationary states and giving sufficient conditions for convergence to the target distribution (Section 3.3).

3.1 UOT–Sinkhorn Energy and Its First Variation

Let $p, q \in \mathcal{P}(\Omega)$, $\Omega \subset \mathbb{R}^d$, and $c(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$. A marginal penalty is generated either by the convex indicator ι ($\iota(1) = 0$, $+\infty$ otherwise), which enforces a marginal exactly, or by a strictly convex $\varphi \in C^2((0, \infty))$ with $\varphi(1) = 0$. For generators (φ_1, φ_2) and $\rho, \varepsilon > 0$,

$$\mathrm{UOT}^{\varphi_1,\varphi_2}(\alpha_1, \alpha_2) := \inf_{\pi \in \mathcal{M}_+(\Omega^2)} \int c\, d\pi + \rho D_{\varphi_1}(\pi_1 \| \alpha_1) + \rho D_{\varphi_2}(\pi_2 \| \alpha_2) + \varepsilon\, \mathrm{KL}(\pi \| \alpha_1 \otimes \alpha_2), \tag{5}$$

with D_φ the Csiszár divergence (Definition 3) and $\tau := \rho/(\rho + \varepsilon)$ the relaxation strength ($\tau \to 1$ is the hard constraint). Three members matter: balanced (ι, ι), underlying W-Flow (Han et al., 2026); symmetric (φ, φ) (Séjourné et al., 2019); and source-fixed (ι, KL), used here (Figure 2). Let $\mathcal{A}(\alpha_1, \alpha_2) := \frac{1}{2}[\mathrm{UOT}^{\varphi_1,\varphi_2}(\alpha_1, \alpha_2) + \mathrm{UOT}^{\varphi_1,\varphi_2}(\alpha_2, \alpha_1)]$ and define the configurational energy $\mathcal{E}_p(q) := \mathcal{A}(q, p) - \frac{1}{2}\mathcal{A}(q, q) - \frac{1}{2}\mathcal{A}(p, p)$ (Definition 7), which coincides on $\mathcal{P}(\Omega)$ with the unbalanced Sinkhorn divergence of Séjourné et al. (2019) for (φ, φ) and with $S_\varepsilon(q, p)$ (Feydy et al., 2018; Han et al., 2026) for (ι, ι). Let $\Phi_q := \frac{\delta \mathcal{E}_p}{\delta q}(q)$ and $F_p[q] := -\nabla \Phi_q$. By construction $\mathcal{E}_p(p) = 0$ and $F_p[p] = \mathbf{0}$ for every choice of generators.[1]

Proposition 1 (Force). *Under Assumption 1, assume that all optimal plans appearing below are unique. For a pair (α_1, α_2) with optimal plan π of $\mathrm{UOT}^{\varphi_1,\varphi_2}(\alpha_1, \alpha_2)$ and a slot $k \in \{1, 2\}$, let $r^{(k)}_{\alpha_1,\alpha_2} := d\pi_k / d\alpha_k$ and $B^{(k)}_{\alpha_1,\alpha_2}(\mathbf{x}) := \mathbb{E}_\pi[\mathbf{x}_{-k} \mid \mathbf{x}_k = \mathbf{x}]$, with $r^{(k)} \equiv 1$ if $\varphi_k = \iota$ and $r^{(k)}$ positive and C^1 otherwise, and define the force on the k-th marginal*

$$F^{(k)}_{\alpha_1,\alpha_2}(\mathbf{x}) := r^{(k)}_{\alpha_1,\alpha_2}(\mathbf{x})\big(B^{(k)}_{\alpha_1,\alpha_2}(\mathbf{x}) - \mathbf{x}\big). \tag{6}$$

[1] The asymmetric divergence of Séjourné et al. (2019, §4.7.1) debiases with symmetric self-terms, which gives positivity but $F_p[p] \neq \mathbf{0}$. A drifting energy needs the target stationary, so we debias with the same asymmetric cost.

Then $F^{(k)}_{\alpha_1,\alpha_2}$ is the negative spatial gradient of the first variation of $\mathrm{UOT}^{\varphi_1,\varphi_2}(\alpha_1,\alpha_2)$ *with respect to α_k, and*

$$F_p[q] = \tfrac{1}{2}\big[F^{(1)}_{q,p} + F^{(2)}_{p,q}\big] - \tfrac{1}{2}\big[F^{(1)}_{q,q} + F^{(2)}_{q,q}\big]. \tag{7}$$

For (φ,φ) this is $r_{q,p}(B_{q,p}-\mathbf{x}) - r_{q,q}(B_{q,q}-\mathbf{x})$; for (ι,ι) it is $B_{q,p}-B_{q,q}$, the W-Flow field equation 4; for (ι,KL), $r^{(1)}\equiv 1$, $r^{(2)}>0$, and $\int_\Omega r^{(2)}\,d\alpha_2 = 1$. All force identities hold almost everywhere with respect to the corresponding input measure.

Averaging the forward and transposed reverse source-fixed plans gives an effective coupling $\bar\pi$ with $\bar\pi_1 \geq q/2$ and $\bar\pi_2 \geq p/2$: marginals may be reweighted but keep at least half their mass, with no guarantee of nonzero force or target coverage (Appendix G.3). On ImageNet, forward-plan reweighting is mild (0.92–1.03) yet stable (Figure 4).

3.2 Kinetic Lifting and the Drifting Flow

Proposition 1 turns the static transport problem into a self-consistent force field. We lift the positional dynamics to phase space, taking $\Omega = \mathbb{R}^d$ here, and then eliminate momentum to recover the first-order flow of Section 2.1. Let $\mu^{m,\beta}_t(\mathbf{x},\mathbf{u})$ be the joint density of position and momentum with positional marginal $q^{m,\beta}_t := \int \mu^{m,\beta}_t\,d\mathbf{u}$. For inertia $m>0$, friction $\gamma>0$, and inverse temperature $\beta>0$, consider the nonlinear Vlasov–Fokker–Planck equation

$$\partial_t \mu^{m,\beta}_t + \nabla_{\mathbf{x}}\cdot\Big(\frac{\mathbf{u}}{m}\mu^{m,\beta}_t\Big) + \nabla_{\mathbf{u}}\cdot\Big(F_p[q^{m,\beta}_t]\,\mu^{m,\beta}_t\Big) = \frac{\gamma}{m}\nabla_{\mathbf{u}}\cdot\big(\mathbf{u}\,\mu^{m,\beta}_t\big) + \frac{\gamma}{\beta}\Delta_{\mathbf{u}}\mu^{m,\beta}_t, \tag{8}$$

whose left-hand side is conservative transport driven by the UOT force and whose right-hand side is friction and thermal diffusion in momentum. The lifting is variational: with $q_\mu := \int \mu\,d\mathbf{u}$, the free energy $\mathcal{H}_{m,\beta}(\mu) := \mathcal{E}_p(q_\mu) + \int \frac{\|\mathbf{u}\|^2}{2m}\mu + \frac{1}{\beta}\int \mu\log\mu$ is a Lyapunov functional.

Proposition 2 (Free-energy dissipation)**.** *Under Assumption 2, every sufficiently regular solution of equation 8 satisfies $\frac{d}{dt}\mathcal{H}_{m,\beta}(\mu^{m,\beta}_t) = -\gamma\int \mu^{m,\beta}_t\big\|\frac{\mathbf{u}}{m} + \frac{1}{\beta}\nabla_{\mathbf{u}}\log\mu^{m,\beta}_t\big\|^2 \leq 0$.*

Equation 8 is thus a kinetic variational lifting of $\mathcal{E}_p$ rather than a gradient flow; the latter emerges as momentum relaxes.

Theorem 1 (Quantitative Kramers–Smoluchowski limit)**.** *Fix $\gamma,\beta,T>0$ and assume Assumption 3. Suppose that the initial laws share a positional marginal $q_0\in\mathcal{P}_2(\mathbb{R}^d)$ and satisfy $\int\|\mathbf{u}\|^2\,d\mu^{m,\beta}_0 \leq C_0 m$ uniformly for $0<m\leq 1$. Then $\sup_{t\in[0,T]}\mathcal{W}_2(q^{m,\beta}_t, q^\beta_t) \leq C_{T,\beta}\sqrt{m}$, where $C_{T,\beta}$ is independent of m and $q^\beta\in C([0,T];\mathcal{P}_2(\mathbb{R}^d))$ is the unique weak solution of*

$$\partial_t q^\beta_t = \frac{1}{\gamma}\nabla\cdot\big(q^\beta_t\nabla\Phi_{q^\beta_t}\big) + \frac{1}{\gamma\beta}\Delta q^\beta_t, \qquad q^\beta_0 = q_0. \tag{9}$$

Equation 9 is the WGF equation 1 of $\mathcal{E}_p + \beta^{-1}\int q\log q\,d\mathbf{x}$ with mobility $1/\gamma$. As $\beta\to\infty$, Corollary 1 gives convergence, uniformly on finite time intervals in $\mathcal{W}_2$, to

$$\partial_t q_t = -\frac{1}{\gamma}\nabla\cdot\big(q_t F_p[q_t]\big), \qquad q_{t=0} = q_0. \tag{10}$$

This is equation 1 with $\mathcal{F}=\mathcal{E}_p$ and $\mathbf{v}_q = \gamma^{-1}F_p[q]$. Its characteristic Euler update is $\mathbf{x}^{n+1} = \mathbf{x}^n + (h/\gamma)F_p[q_n](\mathbf{x}^n)$, with the mobility absorbed into the drifting step size $\eta = h/\gamma$ (cf. equation 3). For (ι,ι), the force is the W-Flow field.

Remark 1. For finite β, diffusion yields an absolutely continuous law at every positive time, whereas equation 10 transports measures without diffusion. On $\mathbb{R}^d$, the zero-temperature stationary states of Section 3.3 do not remain stationary at finite temperature. Other finite-temperature equilibria may nevertheless exist and need not coincide with p.

3.3 Long-Time Behavior: Obstruction and Conditional Convergence

Along any regular solution of equation 10 the energy dissipates as

$$\frac{d}{dt}\mathcal{E}_p(q_t) = -\frac{1}{\gamma}\mathcal{D}(q_t), \qquad \mathcal{D}(q) := \int_\Omega \|\nabla\Phi_q\|^2\,dq, \tag{11}$$

but, as noted in Section 2.1, this does not force $q_t \to p$. Throughout this subsection $\Omega \subset \mathbb{R}^d$ is compact and the kernel $e^{-c/\varepsilon}$ is positive and universal. The obstruction is elementary: take a target symmetric about the origin and the single atom $q = \delta_{\mathbf{0}}$. Every optimal plan inherits the symmetry, so every barycenter at the origin vanishes and $F_p[\delta_{\mathbf{0}}](\mathbf{0}) = \mathbf{0}$; the continuity equation cannot split an atom, so the model stays at $\delta_{\mathbf{0}}$.

Proposition 3 (Non-target stationary state)**.** *Let $\Omega = [-1, 1]$, let $p \in \mathcal{P}(\Omega)$ satisfy $\mathsf{R}_{\#}p = p$ for $\mathsf{R}(x) = -x$ with $p \neq \delta_0$, and assume Φ_{δ_0} admits a C^2 representative on Ω. Then, for every choice of generators, $q_t \equiv \delta_0$ is a stationary distributional solution of equation 10 and the metric slope $|\partial \mathcal{E}_p|(\delta_0)$ vanishes, although $\delta_0 \neq p$.*

Wasserstein stationarity requires $\nabla \Phi_q = \mathbf{0}$ only q-almost everywhere, not on all of Ω. In particular, for (ι, ι) the flow is W-Flow, whose velocity field vanishes identically only at $q = p$ (Han et al., 2026); the proposition shows that vanishing on $\mathrm{supp}(q)$, which is all the flow sees, is strictly weaker. The obstruction is a property of the zero-temperature limit (Remark 1), not of the transport geometry. Convergence therefore needs a quantitative link between the energy gap and its dissipation.

Convergence to p itself requires that the flow keep covering the target. Appendix F gives sufficient conditions (convexity of $\mathcal{E}_p$ along mixtures, bounded $\chi^2(p\|q_t)$, and a uniform Poincaré inequality along the trajectory) under which the energy–dissipation inequality $\mathcal{E}_p(q_t)^2 \leq A\, \mathcal{D}(q_t)$ holds and yields $\mathcal{E}_p(q_t) = \mathcal{O}(t^{-1})$ and $\mathcal{W}_2(q_t, p) \to 0$ (Proposition 5). The stationary state of Proposition 3 violates the coverage condition, since $\chi^2(p\|\delta_{\mathbf{0}}) = \infty$; without it the flow still converges along a subsequence to a Wasserstein-stationary state (Proposition 4).

The Dirac stationary example alone does not establish whether collapse can occur from a nondegenerate initialization. Whether the coverage condition in Assumption 4 holds throughout training remains open (Appendix I). We next instantiate equation 10 as a training algorithm.

4 Implementation

Mini-batch approximation. Given independent mini-batches $\{\mathbf{x}_i\}_{i=1}^N \sim q_t$ and $\{\mathbf{y}_j\}_{j=1}^M \sim p$ with uniform weights, a coupling is a matrix $\mathbf{P} \in \mathbb{R}_+^{N \times M}$ and equation 5 becomes

$$\inf_{\mathbf{P} \geq 0} \langle \mathbf{C}, \mathbf{P} \rangle + \rho D_{\varphi_1}\big(\mathbf{P}\mathbf{1}_M \,\big\|\, \tfrac{1}{N}\mathbf{1}_N\big) + \rho D_{\varphi_2}\big(\mathbf{P}^\top \mathbf{1}_N \,\big\|\, \tfrac{1}{M}\mathbf{1}_M\big) + \varepsilon\, \mathrm{KL}\big(\mathbf{P} \,\big\|\, \tfrac{1}{NM}\mathbf{1}_N\mathbf{1}_M^\top\big), \quad (12)$$

with $C_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$. For $\varphi_1 = \iota$ the first penalty is the hard constraint $\mathbf{P}\mathbf{1}_M = \frac{1}{N}\mathbf{1}_N$.

Training procedure. The remaining components follow W-Flow (Han et al., 2026); we summarize them here and defer pseudocode to Appendix G. Transport is carried out in the space of frozen feature blocks ψ_ℓ rather than in latent space. At iteration n, for each block ℓ, we compute $\mathbf{x}_i = \psi_\ell(f_{\boldsymbol{\theta}_n}(\mathbf{z}_i))$, build the costs $C_{ij} = \frac{1}{2}\|\mathbf{x}_i - \psi_\ell(\mathbf{y}_j)\|^2$ against a real batch and, with a second independent generated batch $\tilde{\mathcal{X}}_n$ in place of the real one, the self-interaction costs, so that the self term avoids trivial diagonal matches. The plans solving equation 12 yield the mini-batch estimate $\widehat{\mathbf{v}}_i^\ell$ of $F_p[q_n]$ at $\mathbf{x}_i$ (Appendix G.3), and the generator regresses onto an explicit Euler target,

$$\bar{\mathbf{x}}_i^\ell = \mathrm{sg}\left(\mathbf{x}_i + \eta\, \widehat{\mathbf{v}}_i^\ell\right), \qquad \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n_\psi N} \sum_{\ell, i} \left\|\psi_\ell(f_{\boldsymbol{\theta}}(\mathbf{z}_i)) - \bar{\mathbf{x}}_i^\ell\right\|^2, \quad (13)$$

where η absorbs the mobility γ^{-1} and the stop-gradient avoids differentiating through the transport solve. Guidance acts on the velocity, $\widehat{\mathbf{v}}_{\mathsf{c}} - \widehat{\mathbf{v}}_{\mathrm{self}} + w\,(\widehat{\mathbf{v}}_{\mathsf{c}} - \widehat{\mathbf{v}}_{\varnothing})$, with w sampled during training and fixed at inference (Appendix G.4). The departure from Han et al. (2026) is that the plans solve equation 12 rather than balanced OT; at $\tau = 1$ our symmetrized estimator reduces to theirs (the balanced reverse plan is the transposed forward plan), and our independently trained $\tau = 1$ model reproduces W-Flow-B (Han et al., 2026) theoretically.

5 Experiments

Setup. We evaluate on class-conditional ImageNet 256×256. Following W-Flow (Han et al., 2026), a DiT-style generator operates in the SD-VAE latent space (Rombach et al., 2022), transport

Table 1: Quantitative comparison of balanced OT, fully unbalanced OT, and source-fixed UOT across marginal divergences and solver settings. The six feature-specific FDr scores are followed by their mean, FDr^6. KID and KD-DINOv2 are scaled by 100. Lower is better except for IS. Red, blue, and green shading denotes the best, second-best, and third-best results in each metric across the entire table, respectively; ties share the same rank.

Model / setting	FDr ↓							Other metrics			
	Incep.	ConvNeXt	DINOv2	MAE	SigLIP	CLIP	Mean FDr^6	KID ↓	KD-DINOv2 ↓	FID ↓	IS ↑
BALANCED OT *(both marginals fixed)*											
W-FLOW baseline, our retraining	4.21	8.92	27.13	15.35	50.16	39.50	24.21	0.1321	40.70	7.07	140.37 ± 1.64
FULLY UNBALANCED OT *(both marginals relaxed)*											
Séjourné KL-UOT τ=0.95	4.27	9.17	28.17	15.61	51.46	40.74	24.90	**0.1273**	43.12	7.17	**145.28 ± 1.87**
Séjourné KL-UOT τ=0.975	4.16	8.77	26.99	15.05	49.81	39.02	23.97	0.1303	40.40	7.00	143.28 ± 3.12
Séjourné KL-UOT τ=0.985	4.18	8.69	26.79	15.17	49.32	38.46	23.77	0.1362	39.75	7.03	142.57 ± 2.63
Séjourné KL-UOT τ=0.99	4.16	8.67	26.74	15.06	49.23	38.34	23.70	0.1414	39.77	7.00	142.65 ± 2.83
FORWARD-ONLY SOURCE-FIXED UOT *(source fixed; target relaxed)*											
KL τ=0.985, i=10	4.23	8.82	26.73	15.20	49.20	38.71	23.81	0.1313	40.04	7.11	140.83 ± 2.72
SYMMETRIZED SOURCE-FIXED UOT *(source fixed; target relaxed in each directional solve)*											
Pearson-χ^2 τ=0.95	4.20	8.50	26.57	15.39	49.36	38.78	23.80	0.1377	39.55	7.06	144.07 ± 2.45
Hellinger τ=0.95	4.20	8.36	26.29	**15.04**	48.83	38.26	23.50	0.1375	38.65	7.05	144.90 ± 1.50
KL: Marginal relaxation τ variants											
KL τ=0.95	**4.13**	8.53	26.64	15.12	48.93	38.41	23.63	0.1328	39.46	6.94	145.08 ± 2.29
KL τ=0.975	4.17	8.71	26.84	15.19	49.31	38.67	23.81	0.1327	39.92	7.00	143.10 ± 2.42
KL τ:0.95→0.985	**4.13**	8.61	26.79	15.12	49.28	38.93	23.81	**0.1268**	40.04	**6.93**	142.82 ± 2.66
KL τ=0.99	4.25	8.79	26.87	15.13	49.53	38.80	23.89	0.1375	39.96	7.14	142.40 ± 3.14
KL τ=0.985	**4.09**	8.57	26.53	**15.05**	48.64	38.34	23.54	0.1343	39.32	**6.86**	143.28 ± 2.49
KL: Step size η variant											
KL τ=0.985, η=0.5	4.23	9.00	27.07	15.44	50.20	39.31	24.21	0.1374	40.31	7.10	141.49 ± 1.99
KL: Sinkhorn iteration i variants											
KL τ=0.985, i=1	**4.00**	**7.44**	**24.86**	15.28	**47.28**	**37.35**	**22.70**	0.1495	**35.89**	**6.72**	**156.59 ± 2.70**
KL τ=0.985, i=5	4.19	8.60	26.49	**15.09**	48.75	38.43	23.59	0.1364	39.51	7.04	144.05 ± 2.85
KL τ=0.985, i=15	4.30	8.90	26.89	15.13	49.34	38.70	23.88	0.1387	39.96	7.22	141.37 ± 3.41
KL: Entropic regularization ϵ variants											
KL τ=0.985, ϵ=0.01	4.15	**8.26**	**25.33**	15.14	**47.38**	**37.34**	**22.93**	0.1564	**36.43**	6.98	144.18 ± 2.25
KL τ=0.985, ϵ=0.01, i=1	4.16	**6.47**	**22.85**	16.04	**44.78**	**34.77**	**21.51**	0.2236	**31.28**	6.99	**164.49 ± 2.40**
KL τ=0.985, ϵ=0.1	4.49	9.08	28.62	16.04	52.49	41.03	25.29	**0.1297**	43.52	7.54	143.13 ± 3.52

costs use the released pretrained MAE encoders (Deng et al., 2026), and guidance acts on the velocity with its weight sampled during training (Appendix G.4); sampling takes one forward pass. By default we use ten Sinkhorn iterations and $\varepsilon = 0.05$ at every scale, since for our method ten outperform one at B/2 and L/2 (Table 3); W-Flow instead uses one iteration at L/2 and XL/2 and $\varepsilon = 0.01$ at XL/2, and its other training settings are kept (Appendix G). Ablations use DiT-B/2 for 100 epochs. FID (Heusel et al., 2017) uses 50K samples under the ADM protocol (Dhariwal & Nichol, 2021).

Run-to-run variation. Each row of Table 1 is a single training run, so we calibrate differences against independent retrainings of the same configuration. Evaluation is deterministic given a checkpoint and sampling seed; across three sampling seeds FID varies by $\pm$ 0.009. Our from-scratch retraining of the balanced baseline gives 7.07 FID and 24.21 FDr^6 at 30K steps, against 7.08 and 24.18 for the 30K checkpoint released by Han et al. (2026) under the same pipeline (they report 7.08 under their protocol). At 200K steps, our $\tau = 1$ B/2 model and the released W-Flow-B checkpoint give 1.53 vs. 1.52 FID and 11.64 vs. 11.71 FDr^6 (Table 3). The two retrainings thus differ by 0.01 FID and 0.03 FDr^6 at 30K steps and by 0.01 FID and 0.07 FDr^6 at 200K steps, well below the differences discussed next. Over three independent training seeds, balanced and $\tau = 0.985$ give 7.07 ± 0.05 and 6.86 ± 0.06 FID, and 24.21 ± 0.04 and 23.54 ± 0.06 FDr^6.

Marginal relaxation. Relaxing the target marginal helps: all seven symmetrized source-fixed runs in Table 1 improve on balanced transport in FDr_6 (23.50–23.89 vs. 24.21) and in each of the six feature spaces, and so does the two-sided KL-UOT of Séjourné et al. (2019) at $\tau = 0.975$ and 0.985 (23.97 and 23.77). Source-fixed transport is nonetheless preferable: at the respective best settings it reaches 6.86 FID and 23.54 FDr_6 against 7.00 and 23.77 for two-sided; at matched $\tau = 0.985$ it is better in all six spaces; and it is more robust to the relaxation strength, still improving on balanced at $\tau = 0.95$ (23.63, 6.94) where two-sided falls below it. In FID the differences are smaller and more

Table 2: Density, coverage, precision, and recall on ImageNet 256×256. All results use 50K generated images and the ImageNet validation set as reference. Density and coverage use $k = 5$; precision and recall use $k = 3$. DINOv2 results use CLS-token features. Higher is better for all metrics. Within each of the two comparison blocks independently, red, blue, and green denote the best, second-best, and third-best results in each metric column, respectively; ties share the same rank. Subsections within a block share the same ranking.

Model / setting	Step	CFG	Inception				DINOv2-CLS			
			Density ↑	Coverage ↑	Precision ↑	Recall ↑	Density ↑	Coverage ↑	Precision ↑	Recall ↑
30K Ablations *(ranked within this block)*										
Balanced OT: both marginals fixed										
W-FLOW baseline, our retraining	30K	1.50	1.275 972	0.879 240	0.766 760	0.391 320	0.360 704	0.237 440	0.441 700	0.159 980
Fully unbalanced OT: both marginals relaxed										
Séjourné KL-UOT τ=0.95	30K	1.50	1.274 408	0.876 160	0.766 900	**0.398180**	0.356 112	0.227 020	0.436 980	0.162 640
Séjourné KL-UOT τ=0.975	30K	1.50	1.282 540	0.880 580	0.766 400	0.391 820	0.357 960	0.235 980	0.439 280	0.165 760
Séjourné KL-UOT τ=0.985	30K	1.50	1.280 780	0.883 520	0.763 340	**0.394160**	0.353 650	0.236 880	0.438 820	**0.168260**
Séjourné KL-UOT τ=0.99	30K	1.50	1.289 264	0.883 520	0.768 060	0.390 340	0.355 656	0.238 200	0.443 180	0.167 400
Forward-only source-fixed UOT: source fixed; target relaxed										
KL τ=0.985, i=10	30K	1.50	1.285 796	0.880 000	0.763 980	0.391 540	**0.369696**	0.238 960	0.443 620	0.159 780
Symmetrized source-fixed UOT: source fixed; target relaxed in each directional solve										
Pearson-χ^2 τ=0.95	30K	1.50	1.318 908	0.885 160	0.773 060	0.382 380	**0.368712**	0.242 340	**0.445860**	0.149 680
Hellinger τ=0.95	30K	1.50	**1.325380**	**0.886320**	**0.774840**	0.377 060	**0.375876**	**0.245320**	**0.450580**	0.153 000
KL τ=0.95	30K	1.50	1.297 344	0.881 000	0.768 580	0.384 120	0.365 412	0.239 720	0.443 620	0.164 260
KL τ=0.975	30K	1.50	1.294 996	0.883 800	0.769 840	0.387 900	0.362 096	0.237 740	0.440 280	0.156 080
KL τ:0.95→0.985	30K	1.50	1.295 212	0.882 240	0.766 400	0.389 620	0.358 616	0.240 260	0.441 440	0.159 420
KL τ=0.99	30K	1.50	1.286 124	0.885 620	0.767 260	0.383 880	0.360 840	0.238 480	0.443 160	0.162 020
KL τ=0.985	30K	1.50	1.295 844	**0.886740**	0.769 360	0.393 320	0.356 080	0.237 660	0.442 120	0.160 920
KL τ=0.985, η=0.5	30K	1.50	1.266 760	0.878 880	0.760 660	0.393 480	0.361 552	0.236 480	0.441 680	0.152 500
KL τ=0.985, i=1	30K	1.50	**1.342936**	**0.888300**	**0.783180**	0.372 300	0.359 704	**0.253880**	**0.447560**	0.153 800
KL τ=0.985, i=5	30K	1.50	1.288 708	0.883 800	0.772 880	0.384 600	0.358 848	0.243 140	0.441 600	0.156 920
KL τ=0.985, i=15	30K	1.50	1.299 728	0.880 700	0.766 340	0.386 240	0.366 136	0.236 440	0.445 840	0.167 740
KL τ=0.985, ϵ=0.01	30K	1.50	1.232 700	0.875 160	0.758 220	**0.405240**	0.344 216	0.244 160	0.436 800	**0.175820**
KL τ=0.985, ϵ=0.01, i=1	30K	1.50	1.258 084	0.878 020	0.772 440	0.388 720	0.348 040	**0.257740**	0.436 420	**0.183040**
KL τ=0.985, ϵ=0.1	30K	1.50	**1.327708**	0.877 540	**0.780740**	0.359 080	0.368 236	0.224 580	0.445 700	0.140 040
Scaled and Official Models *(ranked within this block)*										
Scaled symmetrized source-fixed UOT models										
KL B/2 τ=0.95	200K	1.19	1.067 172	0.947 400	0.721 460	0.611 280	0.497 324	0.553 880	0.544 420	0.428 020
KL B/2 τ=0.985	200K	1.19	1.064 864	0.948 600	**0.723980**	0.615 300	0.502 468	0.563 280	0.547 880	0.428 720
KL B/2 τ=0.985, i=1	200K	1.19	**1.079880**	0.948 760	**0.731400**	0.602 500	0.508 476	0.563 320	0.553 580	0.429 280
KL L/2 τ=0.985	200K	1.15	1.040 424	0.948 140	0.712 040	**0.627460**	0.501 716	0.576 740	0.544 980	0.467 820
KL L/2 τ=0.985, i=1	200K	1.15	1.048 880	0.947 580	0.721 940	0.617 800	0.510 300	0.580 980	0.556 860	0.461 920
KL XL/2 τ=0.985	200K	1.14	1.040 012	**0.949280**	0.714 240	**0.638660**	**0.561476**	**0.644860**	**0.585140**	**0.502080**
Official checkpoints and external reference										
Pixel Drift-L Deng et al. (2026)	100K	1.00	**1.091920**	**0.951360**	**0.732360**	0.595 420	**0.569008**	**0.658880**	**0.580220**	**0.485960**
W-Flow B/2 Han et al. (2026)	200K	1.19	**1.072436**	**0.950140**	0.721 940	0.613 040	0.505 844	0.561 520	0.548 800	0.428 560
W-Flow L/2 Han et al. (2026)	200K	1.14	1.041 888	0.946 180	0.711 580	0.626 180	0.511 296	0.582 060	0.556 000	0.463 300
W-Flow XL/2 Han et al. (2026)	180K	1.09	0.935 748	0.937 380	0.680 620	**0.667500**	**0.538504**	**0.639100**	**0.569640**	**0.516040**

τ-sensitive for both formulations. Forward-only transport matches the symmetrized family in FDr_6 (23.81) but not in FID (7.11 vs. 6.86). Regularization ε also affects performance.

Scaling. The solver setting interacts with the marginal treatment: at 30K steps, balanced transport improves from one to ten Sinkhorn iterations whereas source-fixed UOT is best at one (6.72 vs. 6.86; Table 1); at 200K steps UOT is best at ten (Table 3). We therefore compare each method at its best setting. UOT-GF ($\tau = 0.985$) achieves FIDs of $\mathbf{1.46}/\mathbf{1.34}/\mathbf{1.22}$ at B/L/XL (Table 3); the XL result is a new one-step state of the art in FID among models trained from scratch. Two comparisons are controlled: at B/2, against balanced transport under identical settings (1.53; the released W-Flow-B gives 1.52), and at XL/2, where balanced transport under our recipe reaches 1.39 FID at 100K steps against 1.26 for UOT-GF with otherwise identical settings, so the gain persists at scale. The released W-Flow-L and W-Flow-XL checkpoints (Han et al., 2026) (1.36/1.33) differ in solver settings (and W-Flow-XL trains 180K steps); at W-Flow's L/2 setting (one iteration) relaxation does not help (1.40 vs. 1.36), and balanced transport at ten iterations is untested at L/2. UOT-GF remains competitive with multi-step models at substantially lower sampling compute (Figure 1; Figure 3 and Figure 8).

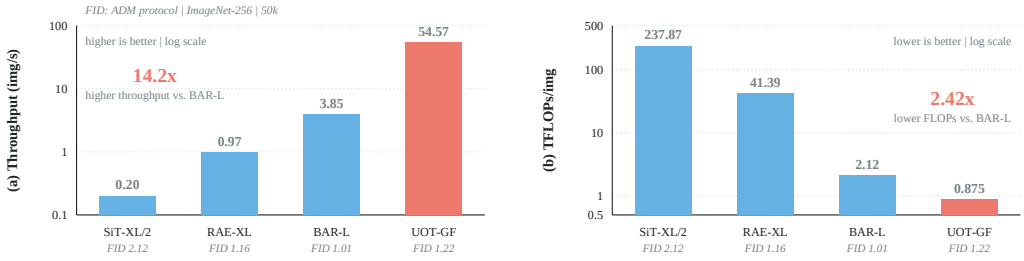


Figure 3: Sampling efficiency on ImageNet-256, measured on a single Nvidia H20 under identical settings (log scale). (a) Throughput in images per second; (b) compute per image in TFLOPs. UOT-GF uses one network evaluation. Measurement details are in Appendix H.

Table 3: System-level comparison on ImageNet 256×256. NFE and Params. denote sampling evaluations and model size. All metrics for all methods are computed by us under a unified evaluation pipeline; numbers may differ slightly from the original papers.

Method	NFE	Params.	Incep.	ConvNeXt	DINOv2	MAE	SigLIP	FDr-CLIP	FID↓	IS↑	FDr^6 ↓	Precision↑	Recall↑
50k validation images	N/A	N/A	1.00	1.00	1.00	1.00	1.00	1.00	1.68	232.2	1.00	0.75	0.66
discrete													
VAR-d30 (Tian et al., 2024)	10×2	2B	1.18	1.70	5.31	6.83	11.89	13.31	1.97	304.6	6.70	**0.82**	0.59
BAR-L (Yu et al., 2026)	256×2×4	1.1B	**0.61**	**0.78**	3.29	4.20	6.60	**5.98**	**1.01**	281.9	3.57	0.77	**0.68**
latent, multi-step													
without semantic distillation													
SiT-XL/2 (Ma et al., 2024)	250×2	675M	1.26	2.02	7.89	5.62	16.14	17.69	2.12	256.7	8.44	0.81	0.60
MAR-L (Li et al., 2024)	256×2×100	478M	1.07	1.10	6.09	4.38	12.67	14.78	1.80	293.4	6.68	0.80	0.60
FlowAR-H (Ren et al., 2025)	50×2	1.9B	1.00	1.30	4.68	4.59	12.75	12.49	1.68	274.1	6.13	0.80	0.62
MAR-H (Li et al., 2024)	256×2×100	942M	0.93	0.95	4.95	3.71	10.07	13.02	1.56	299.5	5.61	0.80	0.62
MAR-L, DeTok (Yang et al., 2026)	256×2×100	478M	0.83	1.36	4.66	4.40	9.57	12.12	1.39	306.2	5.49	0.81	0.62
with semantic distillation													
REG (Wu et al., 2025)	250×2	685M	0.92	1.14	3.45	3.02	8.42	10.86	1.54	302.9	4.64	0.78	0.62
SiT-XL/2-REPA (Yu et al., 2025)	250×2	675M	0.85	1.22	4.27	3.85	9.87	12.65	1.42	306.1	5.45	0.80	0.65
LightningDiT (Yao et al., 2025)	250×2	675M	0.85	1.09	3.76	3.02	8.47	10.21	1.42	294.3	4.57	0.80	0.64
DDT-XL (Wang et al., 2026b)	250×2	675M	0.75	1.02	4.26	4.11	10.16	13.86	1.26	309.3	5.70	0.79	0.66
REPA-E (Leng et al., 2025)	250×2	676M	0.70	1.28	2.44	**2.52**	5.04	6.28	1.17	298.3	**3.04**	0.79	0.66
RAE-XL (Zheng et al., 2026)	50×2	839M	0.69	1.79	**2.11**	3.30	**3.79**	7.87	1.16	261.0	3.26	0.77	0.67
latent, one-step													
Drift-L (Deng et al., 2026)	1	463M	0.91	2.03	10.35	6.51	24.12	21.59	1.53	257.2	10.92	0.79	0.63
iMF-XL (Geng et al., 2026)	1	610M	1.09	1.72	7.30	6.09	17.02	17.14	1.82	278.9	8.39	0.78	0.63
iMF-XL (Geng et al., 2026)	2	610M	0.96	1.54	6.31	5.62	14.91	15.52	1.61	289.1	7.48	0.79	0.63
W-Flow-B (Han et al., 2026)	1	133M	0.91	2.10	11.48	7.54	25.67	22.54	1.52	271.9	11.71	0.72	0.61
W-Flow-L (Han et al., 2026)	1	463M	0.81	1.81	10.12	6.94	22.71	20.59	1.36	271.7	10.49	0.71	0.63
W-Flow-XL (Han et al., 2026)	1	679M	0.79	1.68	8.89	6.69	20.48	18.71	1.33	265.6	9.54	0.68	0.67
pixel, multi-step													
PixNerd-XL (Wang et al., 2026a)	250×2	1.0B	1.25	1.21	3.57	3.56	9.12	11.36	2.10	**318.8**	5.01	0.81	0.59
JiT-L (Li & He, 2026)	250×2	459M	1.54	3.49	6.10	8.07	19.37	25.82	2.59	288.5	10.73	0.79	0.59
JiT-H (Li & He, 2026)	250×2	953M	1.18	2.52	4.28	5.65	11.91	20.40	1.97	296.0	7.66	0.78	0.63
pixel, one-step													
Drift-L (Deng et al., 2026)	1	465M	0.87	1.26	3.88	3.44	8.70	10.48	1.46	298.7	4.40	0.81	0.60
pMF-L (Lu et al., 2026)	1	410M	1.62	1.36	6.70	9.72	20.34	14.81	2.72	261.7	9.09	0.81	0.56
pMF-H (Lu et al., 2026)	1	935M	1.37	1.15	5.43	6.25	15.33	11.68	2.29	267.2	6.87	0.80	0.59
*Unbalanced optimal transport (**ours**)*													
KL $\tau = 1$ - B	1	133M	0.91	2.10	11.48	7.51	25.47	22.39	1.53	271.1	11.64	0.72	0.61
KL $\tau = 0.95$ - B	1	133M	0.90	2.01	11.40	7.55	25.39	22.38	1.51	277.2	11.61	0.72	0.61
KL $\tau = 0.985$ - B	1	133M	0.87	2.05	11.37	7.47	25.19	22.13	1.46	271.5	11.51	0.72	0.62
KL $\tau = 0.985$, $i = 1$ - B	1	133M	0.92	1.92	11.15	7.54	25.47	22.41	1.54	280.5	11.56	0.73	0.60
KL $\tau = 0.985$ - L	1	463M	0.80	1.91	10.57	7.14	23.90	21.11	1.34	268.4	10.91	0.71	0.63
KL $\tau = 0.985$, $i = 1$ - L	1	463M	0.84	1.76	10.21	7.03	23.41	20.92	1.40	280.6	10.70	0.72	0.62
KL $\tau = 0.985$ - XL - 100k	1	679M	0.75	1.59	9.16	6.34	20.74	19.29	1.26	287.6	9.65	0.72	0.63
KL $\tau = 0.985$ - XL - 200k	1	679M	0.74	1.60	9.00	6.20	20.26	19.05	1.22	284.2	9.47	0.71	0.64

On other feature-space metrics the picture is mixed: at B/2, where settings and guidance match, precision, recall and coverage change by less than 0.003 (Table 2); the lower XL recall (0.639 vs. 0.668) coincides with stronger guidance (1.14 vs. 1.09).

6 Conclusion

We introduced UOT-GF using symmetrized source-fixed transport, allowing marginal reweighting with lower bounds on both interaction marginals. In our experiments, it improves generation over balanced transport. Our kinetic analysis recovers drifting, including the balanced $\tau \to 1$ limit, and identifies non-target stationary states and sufficient conditions for target convergence. On ImageNet 256, UOT-GF achieves a one-step state-of-the-art FID of **1.22**. These results motivate extensions to higher resolutions and broader conditional settings.

REFERENCES

David Alvarez-Melis, Yair Schiff, and Youssef Mroueh. Optimizing functionals on the space of probabilities with input convex neural networks. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=dpOYN7o8Jm.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Abdul Fatir Ansari, Ming Liang Ang, and Harold Soh. Refining deep generative models via discriminator gradient flow. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Zbc-ue9p_rE.

Michal Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:184486871.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12934–12944. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9719a00ed0c5709d80dfef33795dcef3-Paper.pdf.

Krishnakumar Balasubramanian. Finite-particle convergence rates for conservative and non-conservative drifting models. *arXiv preprint arXiv:2605.22795*, 2026.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=9BnCwiXB0ty.

Charlotte Bunne, Laetitia Papaxanthos, Andreas Krause, and Marco Cuturi. Proximal optimal transport modeling of population dynamics. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 6511–6528. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/bunne22a.html.

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.

Paul Caucheteux, Clément Bonet, and Anna Korba. A unifying view of variational generative Wasserstein flows. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=sJ7ngz2eQx.

Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv: Optimization and Control*, 2016. URL https://api.semanticscholar.org/CorpusID:119312616.

Hyundo Choi, Junhyeong An, Jinseong Park, and Jaewoong Choi. Unlearning for one-step generative models via unbalanced optimal transport. *arXiv preprint arXiv:2603.16489*, 2026.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

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*, ICML'24. JMLR.org, 2024.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Théo Dumont, Théo Lacombe, and François-Xavier Vialard. Learning Monge maps with constrained drifting models. *arXiv e-prints*, art. arXiv:2603.25182, March 2026. doi: 10.48550/arXiv.2603.25182.

Maria Esteban-Casadevall, Jorge Carrasco-Pollo, Max Welling, Jan-Willem van de Meent, Erik J Bekkers, and Floor Eijkelboom. Kernel-gradient drifting models. *arXiv preprint arXiv:2605.10727*, 2026.

Jiaojiao Fan, Shu Liu, Shaojun Ma, Yongxin Chen, and Hao-Min Zhou. Scalable computation of Monge maps with general costs. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022a. URL https://openreview.net/forum?id=rEnGR3VdDW5.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6185–6215. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/fan22d.html.

Kilian Fatras, Younes Zine, Rémi Flamary, Remi Gribonval, and Nicolas Courty. Learning with minibatch Wasserstein: asymptotic and gradient properties. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2131–2141. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/fatras20a.html.

Kilian Fatras, Thibault Sejourne, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3186–3197. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fatras21a.html.

Lan Feng, Wuyang Li, Eloi Zablocki, Matthieu Cord, and Alexandre Alahi. Representation distribution matching for one-step visual generation. *arXiv preprint arXiv:2607.02375*, 2026.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018. URL https://api.semanticscholar.org/CorpusID:84834062.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=OlzB6LnXcS.

Leonard T Franz, Sebastian Hoffmann, Tim Weiland, Bernhard Schölkopf, and Georg Martius. Drifting fields are not conservative. *arXiv preprint arXiv:2604.06333*, 2026.

Yuan Gao, Yuling Jiao, Yang Wang, Yao Wang, Can Yang, and Shunkang Zhang. Deep generative learning via variational gradient flow. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2093–2101. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/gao19b.html.

Aude Genevay, Lénaïc Chizat, Francis R. Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 2018a. URL https://api.semanticscholar.org/CorpusID:53641639.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with Sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 09–11 Apr 2018b. URL https://proceedings.mlr.press/v84/genevay18a.html.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=uWj4s7rMnR.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30467–30476, June 2026.

Pierre Glaser, Michael Arbel, and Arthur Gretton. KALE flow: A relaxed KL gradient flow for probabilities with disjoint support. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=ZBeCVICs1Ua.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pp. 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

Arthur Gretton, Li Kevin Wenliang, Alexandre Galashov, James Thornton, Valentin De Bortoli, and Arnaud Doucet. On the Wasserstein gradient flow interpretation of drifting models. *arXiv preprint arXiv:2605.05118*, 2026.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5769–5779, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jiaqi Han, Puheng Li, Qiushan Guo, Renyuan Xu, Stefano Ermon, and Emmanuel J Candès. One-step generative modeling via Wasserstein gradient flows. *arXiv preprint arXiv:2605.11755*, 2026.

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

Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative sliced MMD flows with Riesz kernels. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VdkGRV1vcf.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local Nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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. doi: 10.1137/S0036141096303359. URL https://doi.org/10.1137/S0036141096303359.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

Alexander Korotin, Vage Egiazarian, Arip Asadulaev, Alexander Safin, and Evgeny Burnaev. Wasserstein-2 generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=bEoxzW_EXsa.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=d8CBRlWNkqH.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:118648975.

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 e-prints*, pp. arXiv–2603, 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. REPA-E: Unlocking VAE for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272, 2025. doi: 10.1109/ICCV51701.2025.01697.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, June 2026.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=VNBIF0gmkb.

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.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211:969–1117, 2015. URL https://api.semanticscholar.org/CorpusID:119148789.

Shanchuan Lin, Ceyuan Yang, Zhijie Lin, Hao Chen, and Haoqi Fan. Adversarial flow models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=3GPPUHlJU9.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LyJi5ugyJx.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=WUK8JIeetF.

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 *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, pp. 23–40, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-72979-9. doi: 10.1007/978-3-031-72980-5_2. URL https://doi.org/10.1007/978-3-031-72980-5_2.

Yian Ma, Yi-An Ma, Niladri S. Chatterji, Xiang Cheng, Nicolas Flammarion, Peter L. Bartlett, and Michael I. Jordan. Is there an analog of Nesterov acceleration for MCMC? *ArXiv*, abs/1902.00996, 2019. URL https://api.semanticscholar.org/CorpusID:59599742.

Ashok Makkuva, Amirhossein Taghvaei, Sewoong Oh, and Jason Lee. Optimal transport mapping via input convex neural networks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6672–6681. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/makkuva20a.html.

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5acdc9ca5d99ae66afdfe1eea0e3b26b-Paper.pdf.

Petr Mokrov, Alexander Korotin, Lingxiao Li, Aude Genevay, Justin Solomon, and Evgeny Burnaev. Large-scale Wasserstein gradient flows. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=nlLjIuHsMHp.

Youssef Mroueh and Truyen Nguyen. On the convergence of gradient descent in GANs: MMD GAN as a gradient flow. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1720–1728. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/mroueh21a.html.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26:101–174, 01 2001. doi: 10.1081/PDE-100002243.

Matteo Pariset, Ya-Ping Hsieh, Charlotte Bunne, Andreas Krause, and Valentin De Bortoli. Unbalanced diffusion Schrödinger bridge. *arXiv preprint arXiv:2306.09099*, 2023.

G. Peyré and M. Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and trends in machine learning. Now Publishers, 2019. ISBN 9781680835502. URL https://books.google.com.hk/books?id=J0BiwgEACAAJ.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. FlowAR: Scale-wise autoregressive image generation meets flow matching. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=JfLgvNe1tj.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=5JdLZg346Lw.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkQkBnJAb.

F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications. Springer International Publishing, 2015. ISBN 9783319208282. URL https://books.google.com.hk/books?id=UOHHCgAAQBAJ.

Filippo Santambrogio. Euclidean, metric, and Wasserstein gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 03 2017. doi: 10.1007/s13373-017-0101-1.

Thibault Séjourné, Jean Feydy, Franccois-Xavier Vialard, Alain Trouv'e, and Gabriel Peyr'e. Sinkhorn divergences for unbalanced optimal transport. *ArXiv*, abs/1910.12958, 2019. URL https://api.semanticscholar.org/CorpusID:204949797.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger bridge matching. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=qy07OHsJT5.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: scalable image generation via next-scale prediction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Alexander Tong, Kilian FATRAS, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=CD9Snc73AW. Expert Certification.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Mądry. From imagenet to image classification: contextualizing progress on benchmarks. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009a.

Cédric Villani. Hypocoercivity. *Memoirs of the American Mathematical Society*, 202, 07 2009b. doi: 10.1090/S0065-9266-09-00567-5.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=BDnOrExHmt.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40633–40642, June 2026b.

Ge Wu, Shen Zhang, Ruijing Shi, Shanghua Gao, Zhenyuan Chen, Lei Wang, Zhaowei Chen, Hongcheng Gao, Yao Tang, jian Yang, Ming-Ming Cheng, and Xiang Li. Representation entanglement for generation: Training diffusion transformers is much easier than you think. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=koEALFNBj1.

Liang Xu, Diego Caviedes-Nozal, W Bastiaan Kleijn, Longfei Felix Yan, and Rasmus Kongsgaard Olsson. Speech enhancement based on drifting models. *arXiv preprint arXiv:2604.24199*, 2026.

Jiawei Yang, Tianhong Li, Lijie Fan, Yonglong Tian, and Yue Wang. Latent denoising makes good tokenizers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1jBsi98fVe.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15703–15712, 2025. URL https://api.semanticscholar.org/CorpusID:275211938.

Qihang Yu, Qihao Liu, Ju He, Xinyang Zhang, Yang Liu, Liang-Chieh Chen, and Xi Chen. Autoregressive image generation with masked bit modeling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=jALSox02QD.

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. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DJSZGGZYVi.

Guoqiang Zhang, Kenta Niwa, and W Bastiaan Kleijn. Lookahead drifting model. *arXiv preprint arXiv:2605.04060*, 2026a.

Jiawei Zhang, Mengfei Xia, Gen Li, and Yuantao Gu. Distilling drifting transformers with representation autoencoders. *arXiv preprint arXiv:2606.15553*, 2026b.

Qi Zhang. Positive-only drifting policy optimization. *arXiv preprint arXiv:2604.16519*, 2026.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0u1LigJaab.

AI USE STATEMENT

We used large language models as writing and engineering assistants: to polish the prose and draft text from the authors' outlines, to search for related work (every reference was checked against the original source by the authors), to help modify and debug experimental code, to typeset tables and algorithms, to check the presentation and notational consistency of the mathematical statements, and to obtain critical feedback on drafts. They were not used to generate research ideas, to design the method, to run or analyze experiments, or to produce any reported result or figure. All content produced with their assistance was reviewed and verified by the authors, who take full responsibility for the paper.

A NOTATION

This appendix collects the notation used throughout the paper. Vectors are bold lowercase, matrices bold uppercase, and random vectors bold uppercase as well; the two never appear in the same context. Generic measures are written α, α'; q is always the model distribution and p the target. Symbols that appear only inside a single proof are listed in the last block and are not reused elsewhere.

Numbers and Arrays

a	A scalar
$\mathbf{x}$	A vector (bold lowercase)
$\mathbf{P}$	A matrix (bold uppercase)
$\mathbf{1}_N$	All-ones vector of length N
$\mathbf{x}^\top$, $\mathbf{P}^\top$	Transpose
$\|\mathbf{x}\|$	Euclidean norm
$\langle \mathbf{C}, \mathbf{P} \rangle$	Frobenius inner product, $\sum_{ij} C_{ij} P_{ij}$
$\delta_{\mathbf{x}}$	Dirac mass at $\mathbf{x}$

Sets and Spaces

$\mathbb{R}$, $\mathbb{R}_+$	Real numbers; nonnegative reals
d, d_0	Dimension of the latent (data) space; of the noise space
$\Omega \subset \mathbb{R}^d$	Domain of the latent space
$\{1, \dots, N\}$	Integers from 1 to N
$[a, b]$, $(a, b]$	Closed interval; half-open interval
$\mathcal{P}(\Omega)$, $\mathcal{P}_2(\mathbb{R}^d)$	Probability measures on Ω; those with finite second moment
$\mathcal{M}_+(\Omega^2)$	Nonnegative finite measures on $\Omega \times \Omega$
$\mathrm{supp}(q)$	Support of a measure q
$C^k(\Omega)$, $C_c^\infty(\Omega)$	k-times differentiable functions; smooth compactly supported functions
$L^1(\alpha)$, $L^2(\alpha)$	Integrable and square-integrable functions w.r.t. α
$\mathcal{K}$	$\mathcal{W}_2$-compact set in Assumption 4

Indexing

x_i, P_{ij}	Element i of vector $\mathbf{x}$; element (i, j) of matrix $\mathbf{P}$
i, j	Indices of generated and real samples in a mini-batch
n	Training iteration
t	Continuous time

$k \in \{1, 2\}$	Marginal slot of a transport problem
ℓ	Feature-block index
$\mathbf{x}_k$, $\mathbf{x}_{-k}$	The k-th coordinate of a pair; the other coordinate
$(\cdot)^{(k)}$	Quantity attached to the k-th marginal
$(\cdot)_{\alpha_1,\alpha_2}$	Quantity attached to the pair (α_1, α_2)
$\pi_{1,i}$, $\pi_{2,j}$	Row and column sums of a mini-batch plan, $\sum_j P_{ij}$ and $\sum_i P_{ij}$

Calculus and Operators

∂_t	Partial derivative in time
∇, $\nabla_{\mathbf{x}}$, $\nabla_{\mathbf{u}}$	Gradient; gradient in position; gradient in momentum
$\nabla\cdot$	Divergence
Δ, $\Delta_{\mathbf{u}}$	Laplacian; Laplacian in momentum
$\frac{d}{dt}$	Total time derivative along a curve
$\lvert\partial\mathcal{F}\rvert$	Metric slope of $\mathcal{F}$ in $(\mathcal{P}(\Omega), \mathcal{W}_2)$
$\frac{\delta\mathcal{F}}{\delta q}$	First variation of a functional $\mathcal{F}$ at q (Definition 1)
$\int f \, d\alpha$	Integral of f against the measure α
$\frac{d\alpha}{d\alpha'}$	Radon–Nikodym derivative
$\mathrm{LSE}_j(\cdot)$	Log-sum-exp over index j
$\mathcal{O}(\cdot)$	Big-O notation
$\mathrm{sg}(\cdot)$	Stop-gradient

Probability and Information Theory

$\mathbf{x} \sim q$	$\mathbf{x}$ is distributed according to q
$\mathbb{E}_q[f]$	Expectation of f under q
$\mathbb{E}_\pi[\mathbf{x}_{-k} \mid \mathbf{x}_k]$	Conditional expectation under a plan π
$\mathrm{Var}_q(h)$	Variance of h under q
$\mathcal{N}(0, \mathbf{I})$	Standard Gaussian
$\mathrm{Law}(\mathbf{X})$	Probability law of a random vector $\mathbf{X}$
$\mathbf{X}$, $\mathbf{Y}$, $\mathbf{U}$, $\mathbf{B}$	Random vectors in $\mathbb{R}^d$
$\mathrm{KL}(\alpha\|\alpha')$	Generalized Kullback–Leibler divergence (Definition 4)
$D_\varphi(\alpha\|\alpha')$	Csiszár φ-divergence (Definition 3)
$\chi^2(p\|q)$	Chi-square divergence
$\mathcal{W}_2(\alpha, \alpha')$	2-Wasserstein distance
$\alpha \ll \alpha'$	α is absolutely continuous w.r.t. α'
$\alpha \otimes \alpha'$	Product measure
$f_\#\alpha$	Pushforward of α by a map f
$\alpha(\Omega)$	Total mass of α
$C_{\mathrm{PI}}(q)$	Poincaré constant of q

Functions

$f : \mathcal{A} \to \mathcal{B}$	Function with domain $\mathcal{A}$ and codomain $\mathcal{B}$

$\log x$	Natural logarithm
x^+	Positive part, $\max(0, x)$
φ'	Derivative of a scalar function φ
φ'_∞	Recession constant, $\lim_{r\to\infty} \varphi(r)/r$
ι	Convex indicator of $\{1\}$: 0 at 1, $+\infty$ elsewhere
R	Reflection $\mathsf{R}(x) = -x$ (Proposition 3)

Distributions and Measures

p	Target (data) distribution, $p = p_{\mathrm{data}}$
p_{ref}	Reference (noise) distribution
$q_{\boldsymbol{\theta}}$	Model distribution, $(f_{\boldsymbol{\theta}})_{\#}p_{\mathrm{ref}}$
q_t, q_n	Model distribution at time t; at iteration n
$\widehat{q}$, $\widehat{p}$	Empirical mini-batch measures
α, α', α_1, α_2	Generic measures (arguments of a transport cost)
π, π^{α_1,α_2}	Transport plan; optimal plan for the pair (α_1, α_2)
π_1, π_2	First and second marginals of π
$\kappa^{(k)}$	Conditional kernel of π given the k-th coordinate
$\mu_t^{m,\beta}(\mathbf{x}, \mathbf{u})$	Phase-space density
$q_t^{m,\beta}$, q_t^{β}	Positional marginal of $\mu_t^{m,\beta}$; its overdamped limit
q_μ	Positional marginal of a phase-space density μ

Generator

$f_{\boldsymbol{\theta}}$	Generator network with parameters $\boldsymbol{\theta}$
f_n	Generator at training iteration n, $f_{\boldsymbol{\theta}_n}$
$\mathbf{z}$	Latent noise, $\mathbf{z} \sim p_{\mathrm{ref}}$
$\mathbf{x}$	Generated sample, $f_{\boldsymbol{\theta}}(\mathbf{z})$
$\Delta\mathbf{x}_n$	Displacement of a sample between iterations
ψ_ℓ	Frozen feature encoder, block ℓ
n_ψ	Number of feature blocks
c	Class label
$\mathcal{X}$, $\mathcal{Y}$	Mini-batches of generated and real samples

Transport

$c(\mathbf{x}, \mathbf{y})$	Ground cost, $\frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$
$\mathbf{C}$	Cost matrix, $C_{ij} = c(\mathbf{x}_i, \mathbf{y}_j)$
$\mathbf{P}$	Coupling matrix on a mini-batch
$\mathbf{K}$	Log-kernel, $-\mathbf{C}/\varepsilon$
ε	Entropic regularization strength
ρ	Marginal-penalty weight
τ	Relaxation strength, $\rho/(\rho + \varepsilon)$; $\tau \to 1$ is balanced
φ_1, φ_2	Marginal-penalty generators for source and target
$\mathrm{UOT}_{\varepsilon,\rho}^{\varphi_1,\varphi_2}$	Entropic unbalanced transport cost (Definition 5)
$\mathcal{A}(\alpha_1, \alpha_2)$	Symmetrized transport cost

S_ε, $S^{U,\varphi}_{\varepsilon,\rho}$	Balanced Sinkhorn divergence; symmetric unbalanced Sinkhorn divergence
$r^{(k)}_{\alpha_1,\alpha_2}$	Transported-mass ratio of the k-th marginal, $d\pi_k/d\alpha_k$
$B^{(k)}_{\alpha_1,\alpha_2}$	Conditional barycenter of the other coordinate given the k-th
a_i, b_j	Sinkhorn scaling factors in Algorithm 2, entries of $\mathbf{a} \in \mathbb{R}^N$, $\mathbf{b} \in \mathbb{R}^M$

Energies and Forces

$\mathcal{F}$	Generic energy functional on $\mathcal{P}(\Omega)$
$\mathsf{D}(q\|p)$	Generic nonnegative, definite divergence (Section 2.1)
$\mathcal{E}_p$	Configurational energy (Definition 7)
$\mathcal{E}^\beta_p$	Finite-temperature energy, $\mathcal{E}_p + \frac{1}{\beta}\int q \log q$
$\mathcal{H}_{m,\beta}$	Kinetic free energy on phase space
Φ_q	First variation of $\mathcal{E}_p$ at q
$F^{(k)}_{\alpha_1,\alpha_2}$	Force on the k-th marginal, $r^{(k)}(B^{(k)} - \mathbf{x})$
$F_p[q]$	Force induced by $\mathcal{E}_p$, $-\nabla\Phi_q$
$\mathbf{v}_q$	Wasserstein gradient-flow velocity
$\mathcal{D}(q)$	Dissipation, $\int \|\nabla\Phi_q\|^2 dq$
A	Uniform nondegeneracy constant in Assumption 4

Dynamics

$\mathbf{u}$	Momentum (phase-space coordinate)
m, γ, β	Inertia, friction, inverse temperature
h_n	Drifting step size at iteration n
η	Euler step size, h/γ
T	Time horizon
$L_{\mathbf{x}}$, $L_{\mathcal{W}}$, Λ	Lipschitz constants of the force in position and in law; their sum
K_0	Force at the origin, $\|F_p[\delta_{\mathbf{0}}](\mathbf{0})\|$
C_0, $C_{T,\beta}$, C_F	Constants in Theorem 1

Implementation

N, M	Number of generated and real samples in a mini-batch
N_{c}	Classes per training step
N_{pos}, N_{neg}	Real and generated samples per class
N_{unc}	Unconditional real samples per step
L	Number of Sinkhorn iterations
$\mathbf{P}^{\rightarrow}$, $\mathbf{P}^{\leftarrow}$	Forward and reverse source-fixed plans
$\widehat{\mathbf{v}}^{\rightarrow}$, $\widehat{\mathbf{v}}^{\leftarrow}$	Forward and reverse velocity estimates
$\widehat{\mathbf{v}}_{\mathsf{c}}$, $\widehat{\mathbf{v}}_{\mathrm{self}}$, $\widehat{\mathbf{v}}_{\varnothing}$	Estimated velocities toward class, self, and unconditional batches
$\widehat{\mathbf{v}}^{\ell}_i$	Estimated velocity of sample i in feature block ℓ
w, s	Guidance weight; reported scale $s = w + 1$
$\mathcal{L}(\boldsymbol{\theta})$	Regression loss

Proof-Local Symbols

$\mathcal{U}$, $\Psi^{(k)}$	Shorthand for $\mathrm{UOT}^{\varphi_1,\varphi_2}$ and its first variation in slot k
$\mathcal{T}_t$, S	Perturbation diffeomorphism; coordinate swap (Proposition 1)
Ξ, ϑ_R, ν, σ	Free-energy density, cutoff, growth and decay exponents (Proposition 2)
$\varkappa_m$, e, e_β	Momentum kernel; coupling errors (Theorem 1, Corollary 1)
$\mathcal{G}$, $\mathcal{S}$, ς, λ, C_Φ	Cross and self terms, plan mass, mixing parameter, Hessian bound (Proposition 3)

B Related Work

Optimal transport and its relaxations. Optimal transport lifts a ground cost on samples to a geometry on distributions (Villani, 2009a; Santambrogio, 2015; Peyré & Cuturi, 2019). Entropic regularization makes it tractable via Sinkhorn iterations (Cuturi, 2013) and improves sample complexity from the curse-of-dimensionality rate of plain OT to a parametric rate (Genevay et al., 2018a; Mena & Niles-Weed, 2019); the entropic bias is removed by the Sinkhorn divergence, which is positive, definite, and metrizes weak convergence (Genevay et al., 2018b; Feydy et al., 2018). Unbalanced OT replaces the hard marginal constraints by Csiszár penalties, allowing mass to be created or destroyed (Liero et al., 2015; Chizat et al., 2016), and Séjourné et al. (2019) extend Sinkhorn divergences to this setting, proving positivity for symmetric penalties. In the mini-batch regime, balanced OT is known to produce couplings that the population plan would never form, and unbalanced mini-batch OT was proposed as the remedy (Fatras et al., 2020; 2021). Our energy is a source-fixed member of the unbalanced family, debiased so that the target is stationary (Section 3.1).

Wasserstein gradient flows. Gradient flows of energy functionals in the Wasserstein geometry originate in the JKO scheme and Otto calculus (Jordan et al., 1998; Otto, 2001), with the metric-space theory of Ambrosio et al. (2005); see Santambrogio (2017) for an overview. In machine learning they have been instantiated with kernel discrepancies (Arbel et al., 2019; Mroueh & Nguyen, 2021; Glaser et al., 2021; Hertrich et al., 2024), variational and discriminator-based divergences (Gao et al., 2019; Ansari et al., 2021; Fan et al., 2022b), JKO steps parameterized by input-convex networks (Mokrov et al., 2021; Alvarez-Melis et al., 2022; Bunne et al., 2022), and unbalanced transport with an inner adversarial solve (Choi et al., 2024). These methods integrate the flow at inference time or through a costly inner optimization; drifting models instead amortize it into the generator's training trajectory. Kinetic lifts of Wasserstein flows and their overdamped limits are classical in PDE theory (Villani, 2009b) and have been studied as accelerated sampling schemes (Ma et al., 2019); we use them to place drifting inside a Langevin-type hierarchy (Section 3.2).

Optimal transport in generative modeling. The Wasserstein distance entered generative modeling as a GAN critic (Arjovsky et al., 2017; Gulrajani et al., 2017; Salimans et al., 2018) and as a Sinkhorn-divergence loss (Genevay et al., 2018b). A second line learns the transport map itself, through input-convex potentials (Makkuva et al., 2020; Korotin et al., 2021) or neural dual solvers (Korotin et al., 2023; Rout et al., 2022; Fan et al., 2022a). Unbalanced transport was introduced for robustness to outliers via robust OT (Balaji et al., 2020), machine unlearning generative models (Choi et al., 2026) and semi-dual UOT maps (Choi et al., 2023); these rely on adversarial or semi-dual formulations and have been demonstrated up to CelebA-HQ. OT couplings also straighten flow-matching paths (Tong et al., 2024; Pooladian et al., 2023) and underlie Schrödinger-bridge samplers (Bortoli et al., 2021; Shi et al., 2023). We use unbalanced transport differently: not as a loss to be minimized by an inner solver, but as the geometry that defines the velocity of an explicit drifting update.

Drifting models and their interpretation. Drifting models (Deng et al., 2026) train a one-step generator by regressing each sample onto a stop-gradient target displaced along a kernel-based attraction–repulsion field, reaching 1.54 FID on ImageNet-256. Several works have since interpreted the field. With Gaussian kernels it is a difference of kernel-density score estimates, connecting drifting

to score matching and to Wasserstein flows of smoothed divergences (Lai et al., 2026; Cao et al., 2026; Gretton et al., 2026); Franz et al. (2026) show the original field is not conservative, and Caucheteux et al. (2026) give a unifying variational account of generative Wasserstein flows. Finite-particle rates for conservative and non-conservative fields are established by Balasubramanian (2026). On the algorithmic side, He et al. (2026) replace one-sided kernel normalization by Sinkhorn coupling, restoring identifiability; W-Flow (Han et al., 2026) realizes drifting as the Wasserstein gradient flow of the balanced Sinkhorn divergence with a convergence guarantee for the particle dynamics and state-of-the-art one-step results; Esteban-Casadevall et al. (2026) use kernel gradients, Zhang et al. (2026a) regress onto multi-step drifts, Dumont et al. (2026) constrain the dynamics to transport maps, and Feng et al. (2026); Zhang et al. (2026b) match distributions in representation space. Drifting has also been applied to speech enhancement, medical imaging, and policy optimization (Xu et al., 2026; Li et al., 2026; Zhang, 2026). All of these inherit balanced transport, or a kernel surrogate of it, as the geometry of the field; to our knowledge, our work is the first to relax its marginal constraints, showing that a source-fixed unbalanced geometry improves on it under identical settings on ImageNet.

One-step generation. Beyond drifting, one-step generators are obtained by distilling multi-step models (Song et al., 2023; Lu & Song, 2025; Frans et al., 2025), by learning average velocities (Geng et al., 2025), or adversarially (Karras et al., 2019; Lin et al., 2026). Drifting-type methods differ in requiring neither a teacher nor a discriminator.

C Definition

Definition 1 (First variation). *Let $\mathcal{F} : \mathcal{P}(\Omega) \to \mathbb{R}$ be a functional. A measurable function $\Phi_q : \Omega \to \mathbb{R}$ is called a first variation of $\mathcal{F}$ at q if, for every finite signed Borel measure ω with $\Phi_q \in L^1(|\omega|)$ and $q + t\omega \in \mathcal{P}(\Omega)$ for all $t \in [0, t_0)$ and some $t_0 > 0$,*

$$\mathcal{F}[q + t\omega] = \mathcal{F}[q] + t \int_\Omega \Phi_q \, d\omega + o(t) \qquad \text{as } t \to 0^+. \tag{14}$$

We write $\Phi_q = \frac{\delta \mathcal{F}}{\delta q}(q)$. Since admissible perturbations have zero total mass, Φ_q is determined only up to an additive constant.

Applying Definition 1 to $\mathcal{E}_p$ from Definition 7, we write $\Phi_q := \frac{\delta \mathcal{E}_p}{\delta q}(q)$ and $F_p[q] := -\nabla \Phi_q$.

Definition 2 (The 2-Wasserstein distance). *Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the space of probability measures with finite second moments. For $\alpha, \alpha' \in \mathcal{P}_2(\mathbb{R}^d)$, their 2-Wasserstein distance is*

$$\mathcal{W}_2^2(\alpha, \alpha') := \inf_{\pi \in \Pi(\alpha, \alpha')} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|\mathbf{x} - \mathbf{y}\|^2 \, d\pi(\mathbf{x}, \mathbf{y}), \tag{15}$$

where $\Pi(\alpha, \alpha')$ denotes the set of couplings whose marginals are α and α'.

Definition 3 (φ-divergence). *Let $\varphi : \mathbb{R}_+ \to \mathbb{R} \cup \{+\infty\}$ be a convex lower-semicontinuous function satisfying $\varphi(1) = 0$. For two nonnegative finite measures $\alpha, \alpha' \in \mathcal{M}_+(\Omega)$, let*

$$\alpha = \frac{d\alpha}{d\alpha'} \alpha' + \alpha^\perp \tag{16}$$

be the Lebesgue decomposition of α with respect to α'. The Csiszár φ-divergence from α to α' is defined by

$$D_\varphi(\alpha \| \alpha') := \int_\Omega \varphi\Big(\frac{d\alpha}{d\alpha'}\Big) d\alpha' + \varphi'_\infty \, \alpha^\perp(\Omega), \tag{17}$$

where $\varphi'_\infty := \lim_{r \to +\infty} \varphi(r)/r$ is the recession constant of φ, with the convention $0 \cdot (+\infty) = 0$.

In particular, for the convex indicator ι one has $\varphi'_\infty = +\infty$ and $D_\iota(\alpha \| \alpha') = 0$ if $\alpha = \alpha'$ and $+\infty$ otherwise, so the corresponding marginal penalty acts as a hard constraint.

Definition 4 (Generalized Kullback–Leibler divergence). *For $\alpha, \alpha' \in \mathcal{M}_+(\Omega)$ with $\alpha \ll \alpha'$, the generalized Kullback–Leibler divergence is*

$$\mathrm{KL}(\alpha \| \alpha') := \int_\Omega \log\Big(\frac{d\alpha}{d\alpha'}\Big) d\alpha - \alpha(\Omega) + \alpha'(\Omega), \tag{18}$$

and $\mathrm{KL}(\alpha \| \alpha') = +\infty$ whenever $\alpha \not\ll \alpha'$.

Definition 5 (Entropic unbalanced optimal transport)**.** *Let $c : \Omega \times \Omega \to \mathbb{R}_+$ be a nonnegative ground cost, let $\varepsilon, \rho > 0$, and let φ_1, φ_2 be marginal-penalty generators, each either the convex indicator ι or a strictly convex $\varphi \in C^2((0,\infty))$ with $\varphi(1) = 0$. For $\alpha_1, \alpha_2 \in \mathcal{M}_+(\Omega)$,*

$$\mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}(\alpha_1,\alpha_2) := \inf_{\pi\in\mathcal{M}_+(\Omega^2)} \Big\{ \int_{\Omega^2} c(\mathbf{x},\mathbf{y})\, d\pi(\mathbf{x},\mathbf{y}) + \rho D_{\varphi_1}(\pi_1\|\alpha_1) + \rho D_{\varphi_2}(\pi_2\|\alpha_2) + \varepsilon\,\mathrm{KL}(\pi\|\alpha_1\otimes\alpha_2)\Big\}, \tag{19}$$

where π_1, π_2 are the marginals of π. The case $\varphi_k = \iota$ imposes the hard constraint $\pi_k = \alpha_k$; a C^2 generator penalizes deviation from it. We write $\tau := \rho/(\rho+\varepsilon)$.

Definition 6 (Symmetric unbalanced Sinkhorn divergence (Séjourné et al., 2019))**.** *For a single generator φ applied to both marginals, the debiased divergence*

$$S^{U,\varphi}_{\varepsilon,\rho}(\alpha_1,\alpha_2) := \mathrm{UOT}^{\varphi,\varphi}_{\varepsilon,\rho}(\alpha_1,\alpha_2) - \tfrac{1}{2}\mathrm{UOT}^{\varphi,\varphi}_{\varepsilon,\rho}(\alpha_1,\alpha_1) - \tfrac{1}{2}\mathrm{UOT}^{\varphi,\varphi}_{\varepsilon,\rho}(\alpha_2,\alpha_2) + \tfrac{\varepsilon}{2}\big(\alpha_1(\Omega)-\alpha_2(\Omega)\big)^2 \tag{20}$$

is nonnegative, definite, and metrizes weak convergence when the kernel $e^{-c/\varepsilon}$ is positive and universal. On $\mathcal{P}(\Omega)$ the mass correction vanishes, and for $\varphi = \iota$ it reduces to the Sinkhorn divergence S_ε of Feydy et al. (2018).

Definition 7 (Configurational energy)**.** *Let $p \in \mathcal{P}(\Omega)$ be the fixed target. With the symmetrized cost $\mathcal{A}(\alpha_1,\alpha_2) := \frac{1}{2}\big[\mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}(\alpha_1,\alpha_2) + \mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}(\alpha_2,\alpha_1)\big]$, the configurational energy of $q \in \mathcal{P}(\Omega)$ is*

$$\mathcal{E}_p(q) := \mathcal{A}(q,p) - \tfrac{1}{2}\mathcal{A}(q,q) - \tfrac{1}{2}\mathcal{A}(p,p). \tag{21}$$

For (φ,φ) this coincides on $\mathcal{P}(\Omega)$ with Definition 6. For (ι,KL) the self-term $\mathcal{A}(q,q)$ uses the same asymmetric cost as the cross-term, so that $\mathcal{E}_p(p) = 0$ and $F_p[p] = \mathbf{0}$ (Section 3.1); this construction is not covered by the positivity proof of Séjourné et al. (2019), see Assumption 4.

Definition 8 (Transported-mass ratio and conditional barycenter)**.** *Let π be an optimal plan of $\mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}(\alpha_1,\alpha_2)$ with $\pi_k \ll \alpha_k$ for $k \in \{1,2\}$, and let $\pi = \pi_k \otimes \kappa^{(k)}$ be its disintegration with respect to the k-th marginal, where $\kappa^{(k)}(\cdot \mid \mathbf{x})$ is a probability kernel. Define*

$$r^{(k)}_{\alpha_1,\alpha_2}(\mathbf{x}) := \frac{d\pi_k}{d\alpha_k}(\mathbf{x}), \qquad B^{(k)}_{\alpha_1,\alpha_2}(\mathbf{x}) := \int_\Omega \mathbf{x}_{-k}\,\kappa^{(k)}(d\mathbf{x}_{-k} \mid \mathbf{x}), \tag{22}$$

the transported-mass ratio of the k-th marginal and the conditional barycenter of the other coordinate given $\mathbf{x}_k = \mathbf{x}$. If $\varphi_k = \iota$ then $r^{(k)} \equiv 1$.

D Assumptions

Assumption 1 (Regularity of the transport cost)**.** *Write $\mathcal{U} := \mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}$. For every pair (α_1,α_2) under consideration, $\mathcal{U}(\alpha_1,\alpha_2)$ is finite, its optimal plan has finite second moment, and for each slot $k \in \{1,2\}$ the first variation $\Psi^{(k)}_{\alpha_1,\alpha_2} := \frac{\delta\mathcal{U}}{\delta\alpha_k}(\alpha_1,\alpha_2)$ admits a C^1 representative satisfying the transport chain rule*

$$\mathcal{U}\big|_{\alpha_k\to(\mathcal{T}_t)_\#\alpha_k} = \mathcal{U}(\alpha_1,\alpha_2) + t\int_\Omega \nabla\Psi^{(k)}_{\alpha_1,\alpha_2}\cdot\boldsymbol{\xi}\, d\alpha_k + o(t) \qquad \text{as } t\to 0, \tag{23}$$

for every $\boldsymbol{\xi} \in C^\infty_c(\mathrm{int}\,\Omega;\mathbb{R}^d)$, where $\mathcal{T}_t := \mathrm{id} + t\boldsymbol{\xi}$ and the left-hand side denotes $\mathcal{U}$ with only its k-th argument replaced by $(\mathcal{T}_t)_\#\alpha_k$.

Assumption 2 (Variational regularity)**.** *$\mathcal{E}_p$ admits a first variation $\Phi_q \in C^1(\mathbb{R}^d)$ with $F_p[q] = -\nabla_{\mathbf{x}}\Phi_q$, and along every smooth curve of densities $t \mapsto q_t$ under consideration the chain rule $\frac{d}{dt}\mathcal{E}_p(q_t) = \int_{\mathbb{R}^d}\Phi_{q_t}\,\partial_t q_t\, d\mathbf{x}$ holds with an absolutely convergent integral. Moreover, with $\mathbf{z} = (\mathbf{x},\mathbf{u})$, there exist $\nu \geq 0$ and $\sigma > 2d + 2\nu$ such that, uniformly on compact time intervals,*

$$|\Phi_q| + \|\nabla\Phi_q\| + \|F_p[q]\| \leq C(1+\|\mathbf{x}\|)^\nu, \qquad |\log\mu^{m,\beta}_t| + \|\nabla_{\mathbf{z}}\log\mu^{m,\beta}_t\| \leq C(1+\|\mathbf{z}\|)^\nu, \tag{24}$$

$$\mu^{m,\beta}_t + \|\nabla_{\mathbf{z}}\mu^{m,\beta}_t\| + |\partial_t\mu^{m,\beta}_t| \leq C(1+\|\mathbf{z}\|)^{-\sigma}. \tag{25}$$

Assumption 3 (Force stability and initial data)**.** *Fix $\gamma > 0$ and the target p, and write $F[\alpha] := F_p[\alpha]$, assumed independent of m and β and jointly Borel measurable in $(\alpha, \mathbf{x})$. There exist $L_{\mathbf{x}}, L_{\mathcal{W}} \geq 0$ such that, for all $\alpha, \alpha' \in \mathcal{P}_2(\mathbb{R}^d)$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,*

$$\|F[\alpha](\mathbf{x}) - F[\alpha](\mathbf{y})\| \leq L_{\mathbf{x}} \|\mathbf{x} - \mathbf{y}\|, \qquad \Big(\int \|F[\alpha] - F[\alpha']\|^2 \, d\alpha' \Big)^{1/2} \leq L_{\mathcal{W}} \, \mathcal{W}_2(\alpha, \alpha'); \tag{26}$$

set $\Lambda := L_{\mathbf{x}} + L_{\mathcal{W}}$ and $K_0 := \|F[\delta_{\mathbf{0}}](\mathbf{0})\| < \infty$. The initial laws $\mu_0^{m,\beta} \in \mathcal{P}_2(\mathbb{R}^{2d})$ share the positional marginal $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and satisfy $\int \|\mathbf{u}\|^2 \, d\mu_0^{m,\beta} \leq C_0 m$ for $0 < m \leq 1$, with C_0 independent of m.

Assumption 4 (Long-time convergence)**.** *There exists a $\mathcal{W}_2$-compact set $\mathcal{K} \subset \mathcal{P}_2(\Omega)$ containing p and the trajectory $(q_t)_{t \geq 0}$ such that $\mathcal{E}_p$ is lower semicontinuous on $\mathcal{K}$ and has p as its unique minimizer there. Moreover, for some $A > 0$,*

$$\mathcal{E}_p(q_t)^2 \leq A \, \mathcal{D}(q_t) \qquad \text{for a.e. } t \geq 0. \tag{27}$$

Lemma 1 (Sufficient condition for Assumption 4(ii))**.** *Suppose $\mathcal{E}_p$ is convex along linear mixtures, $p \ll q$ with $\frac{dp}{dq} \in L^2(q)$, and q satisfies the Poincaré inequality $\mathrm{Var}_q(h) \leq C_{\mathrm{PI}}(q) \int \|\nabla h\|^2 dq$ for admissible h. Then*

$$\mathcal{E}_p(q)^2 \leq C_{\mathrm{PI}}(q) \, \chi^2(p\|q) \, \mathcal{D}(q). \tag{28}$$

In particular, Assumption 4(ii) holds with $A = \operatorname{ess\,sup}_{t \geq 0} C_{\mathrm{PI}}(q_t) \, \chi^2(p\|q_t)$ whenever the right-hand side is finite.

Proof. Convexity and $\mathcal{E}_p(p) = 0$ give $\mathcal{E}_p(q) \leq \int \Phi_q \, d(q - p) = \int (\Phi_q - \bar{\Phi}_q)\big(1 - \frac{dp}{dq}\big) \, dq$ with $\bar{\Phi}_q := \int \Phi_q \, dq$, the constant dropping out because both measures have unit mass. Cauchy–Schwarz bounds this by $\sqrt{\mathrm{Var}_q(\Phi_q)} \, \sqrt{\chi^2(p\|q)}$, and the Poincaré inequality gives $\mathrm{Var}_q(\Phi_q) \leq C_{\mathrm{PI}}(q) \int \|\nabla \Phi_q\|^2 dq = C_{\mathrm{PI}}(q) \, \mathcal{D}(q)$. □

E PROOFS

Proposition 1 (Force). Under Assumption 1, assume that all optimal plans appearing below are unique. For a pair (α_1, α_2) with optimal plan π of $\mathrm{UOT}^{\varphi_1, \varphi_2}(\alpha_1, \alpha_2)$ and a slot $k \in \{1, 2\}$, let $r^{(k)}_{\alpha_1, \alpha_2} := d\pi_k / d\alpha_k$ and $B^{(k)}_{\alpha_1, \alpha_2}(\mathbf{x}) := \mathbb{E}_\pi[\mathbf{x}_{-k} \mid \mathbf{x}_k = \mathbf{x}]$, with $r^{(k)} \equiv 1$ if $\varphi_k = \iota$ and $r^{(k)}$ positive and C^1 otherwise, and define the force on the k-th marginal

$$F^{(k)}_{\alpha_1, \alpha_2}(\mathbf{x}) := r^{(k)}_{\alpha_1, \alpha_2}(\mathbf{x}) \big(B^{(k)}_{\alpha_1, \alpha_2}(\mathbf{x}) - \mathbf{x} \big). \tag{29}$$

Then $F^{(k)}_{\alpha_1, \alpha_2}$ is the negative spatial gradient of the first variation of $\mathrm{UOT}^{\varphi_1, \varphi_2}(\alpha_1, \alpha_2)$ with respect to α_k, and

$$F_p[q] = \tfrac{1}{2} \big[F^{(1)}_{q,p} + F^{(2)}_{p,q} \big] - \tfrac{1}{2} \big[F^{(1)}_{q,q} + F^{(2)}_{q,q} \big]. \tag{30}$$

For (φ, φ) this is $r_{q,p}(B_{q,p} - \mathbf{x}) - r_{q,q}(B_{q,q} - \mathbf{x})$; for (ι, ι) it is $B_{q,p} - B_{q,q}$, the W-Flow field equation 4; for (ι, KL), $r^{(1)} \equiv 1$, $r^{(2)} > 0$, and $\int_\Omega r^{(2)} \, d\alpha_2 = 1$. All force identities hold almost everywhere with respect to the corresponding input measure.

Proof. Write

$$\mathcal{U}(\alpha_1, \alpha_2) := \mathrm{UOT}^{\varphi_1, \varphi_2}(\alpha_1, \alpha_2), \tag{31}$$

and denote its first variation in the k-th argument by

$$\Psi^{(k)}_{\alpha_1, \alpha_2} := \frac{\delta \, \mathcal{U}}{\delta \alpha_k}(\alpha_1, \alpha_2). \tag{32}$$

We first establish the force identity for the first argument. Let $\boldsymbol{\xi} \in C^\infty_c(\Omega; \mathbb{R}^d)$ and, for sufficiently small $|t|$, define the diffeomorphism

$$\mathcal{T}_t(\mathbf{x}) := \mathbf{x} + t \, \boldsymbol{\xi}(\mathbf{x}). \tag{33}$$

Set

$$\alpha_{1,t} := (\mathcal{T}_t)_\# \alpha_1, \qquad \pi_t := (\mathcal{T}_t, \mathrm{id})_\# \pi. \tag{34}$$

The marginals of π_t are

$$(\pi_t)_1 = (\mathcal{T}_t)_{\#}\pi_1, \qquad (\pi_t)_2 = \pi_2. \tag{35}$$

Csiszár divergences are invariant under simultaneous pushforward of both arguments by a measurable bijection. Consequently,

$$D_{\varphi_1}\big((\pi_t)_1 \| \alpha_{1,t}\big) = D_{\varphi_1}(\pi_1 \| \alpha_1). \tag{36}$$

This identity also holds for the hard marginal constraint. Moreover, since

$$\alpha_{1,t} \otimes \alpha_2 = (\mathcal{T}_t, \mathrm{id})_{\#}(\alpha_1 \otimes \alpha_2), \tag{37}$$

the entropic regularization satisfies

$$\mathrm{KL}\big(\pi_t \| \alpha_{1,t} \otimes \alpha_2\big) = \mathrm{KL}\big(\pi \| \alpha_1 \otimes \alpha_2\big). \tag{38}$$

The second marginal penalty is unchanged as well.

Using π_t as a competitor for $\mathcal{U}(\alpha_{1,t}, \alpha_2)$ and expanding the quadratic cost therefore yields

$$\mathcal{U}(\alpha_{1,t}, \alpha_2) - \mathcal{U}(\alpha_1, \alpha_2) \le \int_{\Omega^2} \big[c(\mathcal{T}_t(\mathbf{x}), \mathbf{y}) - c(\mathbf{x}, \mathbf{y})\big]\, d\pi(\mathbf{x}, \mathbf{y}) \tag{39}$$

$$= t \int_{\Omega^2} (\mathbf{x} - \mathbf{y}) \cdot \boldsymbol{\xi}(\mathbf{x})\, d\pi(\mathbf{x}, \mathbf{y}) + \frac{t^2}{2} \int_{\Omega} \|\boldsymbol{\xi}(\mathbf{x})\|^2\, d\pi_1(\mathbf{x}). \tag{40}$$

Equality holds at $t = 0$. Since the left-hand side is differentiable at $t = 0$, dividing by t and taking the limits from both sides gives

$$\frac{d}{dt}\mathcal{U}(\alpha_{1,t}, \alpha_2)\bigg|_{t=0} = \int_{\Omega^2} (\mathbf{x} - \mathbf{y}) \cdot \boldsymbol{\xi}(\mathbf{x})\, d\pi(\mathbf{x}, \mathbf{y}). \tag{41}$$

Disintegrating π with respect to π_1 and using $d\pi_1 = r^{(1)}_{\alpha_1,\alpha_2}\, d\alpha_1$, we obtain

$$\frac{d}{dt}\mathcal{U}(\alpha_{1,t}, \alpha_2)\bigg|_{t=0} = \int_{\Omega} r^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x})\big(\mathbf{x} - B^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x})\big) \cdot \boldsymbol{\xi}(\mathbf{x})\, d\alpha_1(\mathbf{x}). \tag{42}$$

On the other hand, the transport chain rule gives

$$\frac{d}{dt}\mathcal{U}(\alpha_{1,t}, \alpha_2)\bigg|_{t=0} = \int_{\Omega} \nabla\Psi^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x}) \cdot \boldsymbol{\xi}(\mathbf{x})\, d\alpha_1(\mathbf{x}). \tag{43}$$

Since $\boldsymbol{\xi}$ is arbitrary, it follows that

$$-\nabla\Psi^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x}) = r^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x})\big(B^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x}) - \mathbf{x}\big) = F^{(1)}_{\alpha_1,\alpha_2}(\mathbf{x}) \tag{44}$$

for α_1-almost every $\mathbf{x}$. Perturbing the second coordinate instead proves the same identity for $k = 2$. Thus, for either slot,

$$-\nabla\frac{\delta\mathcal{U}}{\delta\alpha_k}(\alpha_1, \alpha_2) = F^{(k)}_{\alpha_1,\alpha_2}. \tag{45}$$

Next, the definition of $\mathcal{A}$ implies

$$\mathcal{E}_p(q) = \frac{1}{2}\mathcal{U}(q, p) + \frac{1}{2}\mathcal{U}(p, q) - \frac{1}{2}\mathcal{U}(q, q) - \frac{1}{2}\mathcal{U}(p, p). \tag{46}$$

Applying the chain rule to both occurrences of q in $\mathcal{U}(q, q)$, we obtain, up to an additive spatial constant,

$$\Phi_q = \frac{1}{2}\left[\Psi^{(1)}_{q,p} + \Psi^{(2)}_{p,q}\right] - \frac{1}{2}\left[\Psi^{(1)}_{q,q} + \Psi^{(2)}_{q,q}\right]. \tag{47}$$

Taking the negative spatial gradient yields

$$F_p[q] = \frac{1}{2}\left[F^{(1)}_{q,p} + F^{(2)}_{p,q}\right] - \frac{1}{2}\left[F^{(1)}_{q,q} + F^{(2)}_{q,q}\right], \tag{48}$$

which proves equation 7.

Suppose now that $\varphi_1 = \varphi_2 = \varphi$. Let $\mathsf{S}(\mathbf{x}, \mathbf{y}) := (\mathbf{y}, \mathbf{x})$. By symmetry of the cost and of the marginal penalties, together with uniqueness of the optimal plans,

$$\pi^{p,q} = \mathsf{S}_{\#}\pi^{q,p}, \qquad \pi^{q,q} = \mathsf{S}_{\#}\pi^{q,q}. \tag{49}$$

Hence

$$F^{(2)}_{p,q} = F^{(1)}_{q,p}, \qquad F^{(2)}_{q,q} = F^{(1)}_{q,q}. \tag{50}$$

Writing $r_{\alpha,\alpha'} := r^{(1)}_{\alpha,\alpha'}$ and $B_{\alpha,\alpha'} := B^{(1)}_{\alpha,\alpha'}$, we conclude that

$$F_p[q](\mathbf{x}) = r_{q,p}(\mathbf{x})\big(B_{q,p}(\mathbf{x}) - \mathbf{x}\big) - r_{q,q}(\mathbf{x})\big(B_{q,q}(\mathbf{x}) - \mathbf{x}\big). \tag{51}$$

In the balanced case (ι, ι), both marginal density ratios are identically one, and this reduces to

$$F_p[q](\mathbf{x}) = B_{q,p}(\mathbf{x}) - B_{q,q}(\mathbf{x}),$$

as claimed.

Finally, in the source-fixed case (ι, KL), the hard constraint gives $\pi_1 = \alpha_1$, hence $r^{(1)} \equiv 1$. Since both input measures are probabilities,

$$\int_\Omega r^{(2)}\, d\alpha_2 = \pi_2(\Omega) = \pi_1(\Omega) = \alpha_1(\Omega) = 1.$$

Together with the assumed positivity of $r^{(2)}$, this establishes the final assertion. □

Proposition 2 (Free-energy dissipation). Under Assumption 2, every sufficiently regular solution of equation 8 satisfies $\frac{d}{dt}\mathcal{H}_{m,\beta}(\mu_t^{m,\beta}) = -\gamma \int \mu_t^{m,\beta} \big\| \frac{\mathbf{u}}{m} + \frac{1}{\beta}\nabla_{\mathbf{u}} \log \mu_t^{m,\beta} \big\|^2 \le 0$.

Proof. Write $\mu = \mu_t^{m,\beta}$, $q = q_t^{m,\beta}$, $\Phi = \delta\mathcal{E}_p/\delta q(q)$, and $F = F_p[q] = -\nabla_{\mathbf{x}}\Phi$. All integrals below are over $\mathbb{R}^{2n}$ unless otherwise specified, with $d\mathbf{z} = d\mathbf{x}\, d\mathbf{u}$.

Define

$$\Xi(\mathbf{x}, \mathbf{u}) := \Phi(\mathbf{x}) + \frac{\|\mathbf{u}\|^2}{2m} + \frac{1}{\beta}(1 + \log \mu(\mathbf{x}, \mathbf{u})). \tag{52}$$

Then

$$\nabla_{\mathbf{u}}\Xi = \frac{\mathbf{u}}{m} + \frac{1}{\beta}\nabla_{\mathbf{u}} \log \mu, \tag{53}$$

and equation 8 can be written as

$$\partial_t \mu = -\nabla_{\mathbf{x}} \cdot \Big(\frac{\mathbf{u}}{m}\mu\Big) - \nabla_{\mathbf{u}} \cdot (F\mu) + \gamma \nabla_{\mathbf{u}} \cdot \big(\mu \nabla_{\mathbf{u}}\Xi\big). \tag{54}$$

By Assumption 2, with constants allowed to change from line to line,

$$|\Xi| + \|\nabla_{\mathbf{z}}\Xi\| + \|F\| + \frac{\|\mathbf{u}\|}{m} \le C(1 + \|\mathbf{z}\|)^{\nu}. \tag{55}$$

In particular,

$$|\Xi\, \partial_t \mu| + \mu \left(\frac{\|\mathbf{u}\|}{m} + \|F\| \right) \|\nabla_{\mathbf{z}}\Xi\| + \mu \|\nabla_{\mathbf{u}}\Xi\|^2 \le C(1 + \|\mathbf{z}\|)^{2\nu - \sigma} \in L^1(\mathbb{R}^{2d}), \tag{56}$$

since $\sigma > 2d + 2\nu$.

To justify integration by parts on the unbounded phase space, choose $\vartheta \in C_c^\infty(\mathbb{R}^{2d})$ such that $0 \le \vartheta \le 1$, $\vartheta = 1$ on B_1, and $\operatorname{supp}\vartheta \subset B_2$. For $R \ge 1$, set

$$\vartheta_R(\mathbf{z}) := \vartheta(\mathbf{z}/R), \qquad A_R := B_{2R} \setminus B_R, \tag{57}$$

where B_R is the ball of radius R in $\mathbb{R}^{2d}$. Then

$$\operatorname{supp} \nabla_{\mathbf{z}}\vartheta_R \subset A_R, \qquad \|\nabla_{\mathbf{z}}\vartheta_R\|_{L^\infty} \le \frac{C}{R}. \tag{58}$$

Multiplying the equation by $\vartheta_R \Xi$ and integrating by parts with compact support gives

$$\int \vartheta_R \Xi \, \partial_t \mu \, d\mathbf{z} = \int \vartheta_R \mu \left(\frac{\mathbf{u}}{m} \cdot \nabla_{\mathbf{x}} \Xi + F \cdot \nabla_{\mathbf{u}} \Xi \right) d\mathbf{z} - \gamma \int \vartheta_R \mu \|\nabla_{\mathbf{u}} \Xi\|^2 \, d\mathbf{z} + \mathcal{R}_R, \tag{59}$$

where

$$\mathcal{R}_R := \int \mu \Xi \left(\frac{\mathbf{u}}{m} \cdot \nabla_{\mathbf{x}} \vartheta_R + F \cdot \nabla_{\mathbf{u}} \vartheta_R - \gamma \nabla_{\mathbf{u}} \Xi \cdot \nabla_{\mathbf{u}} \vartheta_R \right) d\mathbf{z}. \tag{60}$$

The decay and growth bounds imply

$$|\mathcal{R}_R| \leq \frac{C}{R} \int_{A_R} (1 + \|\mathbf{z}\|)^{2\nu - \sigma} \, d\mathbf{z} \leq C R^{2d + 2\nu - \sigma - 1} \longrightarrow 0 \qquad \text{as} \quad R \to \infty. \tag{61}$$

Here we used $|A_R| \leq C R^{2d}$.

It remains to treat the conservative contribution. Since $F = -\nabla_{\mathbf{x}} \Phi$, its integrand satisfies the pointwise identity

$$\mu \left(\frac{\mathbf{u}}{m} \cdot \nabla_{\mathbf{x}} \Xi + F \cdot \nabla_{\mathbf{u}} \Xi \right) = \frac{1}{\beta} \left(\frac{\mathbf{u}}{m} \cdot \nabla_{\mathbf{x}} \mu + F \cdot \nabla_{\mathbf{u}} \mu \right). \tag{62}$$

Moreover, $\nabla_{\mathbf{x}} \cdot (\mathbf{u}/m) = 0$ and $\nabla_{\mathbf{u}} \cdot F = 0$. Another compactly supported integration by parts therefore yields

$$\int \vartheta_R \mu \left(\frac{\mathbf{u}}{m} \cdot \nabla_{\mathbf{x}} \Xi + F \cdot \nabla_{\mathbf{u}} \Xi \right) d\mathbf{z} = -\frac{1}{\beta} \int \mu \left(\frac{\mathbf{u}}{m} \cdot \nabla_{\mathbf{x}} \vartheta_R + F \cdot \nabla_{\mathbf{u}} \vartheta_R \right) d\mathbf{z}. \tag{63}$$

Its absolute value is bounded by

$$\frac{C}{R} \int_{A_R} (1 + \|\mathbf{z}\|)^{\nu - \sigma} \, d\mathbf{z} \leq C R^{2d + \nu - \sigma - 1} \longrightarrow 0. \tag{64}$$

Passing to the limit $R \to \infty$ in equation 59, using dominated convergence and equation 56, gives

$$\int \Xi \, \partial_t \mu \, d\mathbf{z} = -\gamma \int \mu \|\nabla_{\mathbf{u}} \Xi\|^2 \, d\mathbf{z}. \tag{65}$$

Finally, the assumed chain rule for $\mathcal{E}_p$, together with differentiation under the integral sign for the kinetic and entropy terms, gives

$$\frac{d}{dt} \mathcal{H}_{m,\beta}(\mu) = \int \Xi \, \partial_t \mu \, d\mathbf{z}. \tag{66}$$

The latter differentiations are justified by the uniform integrable bounds

$$\left(1 + \|\mathbf{u}\|^2 + |\log \mu|\right) |\partial_t \mu| \leq C (1 + \|\mathbf{z}\|)^{\nu - \sigma} \in L^1(\mathbb{R}^{2d}). \tag{67}$$

Consequently,

$$\frac{d}{dt} \mathcal{H}_{m,\beta}(\mu_t^{m,\beta}) = -\gamma \int_{\mathbb{R}^{2d}} \mu_t^{m,\beta} \left\| \frac{\mathbf{u}}{m} + \frac{1}{\beta} \nabla_{\mathbf{u}} \log \mu_t^{m,\beta} \right\|^2 d\mathbf{x} \, d\mathbf{u} \leq 0. \tag{68}$$

□

Theorem 1 (Quantitative Kramers–Smoluchowski limit). Fix $\gamma, \beta, T > 0$ and assume Assumption 3. Suppose that the initial laws share a positional marginal $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$ and satisfy $\int \|\mathbf{u}\|^2 \, d\mu_0^{m,\beta} \leq C_0 m$ uniformly for $0 < m \leq 1$. Then $\sup_{t \in [0,T]} \mathcal{W}_2(q_t^{m,\beta}, q_t^{\beta}) \leq C_{T,\beta} \sqrt{m}$, where $C_{T,\beta}$ is independent of m and $q^{\beta} \in C([0,T]; \mathcal{P}_2(\mathbb{R}^d))$ is the unique weak solution of

$$\partial_t q_t^{\beta} = \frac{1}{\gamma} \nabla \cdot \left(q_t^{\beta} \nabla \Phi_{q_t^{\beta}} \right) + \frac{1}{\gamma \beta} \Delta q_t^{\beta}, \qquad q_0^{\beta} = q_0. \tag{69}$$

Proof. For an $\mathbb{R}^d$-valued random variable $\mathbf{X}$, we denote its probability law by $\mathrm{Law}(\mathbf{X})$, namely,

$$\mathrm{Law}(\mathbf{X})(\Gamma) := \mathbb{P}(\mathbf{X} \in \Gamma), \qquad \Gamma \subset \mathbb{R}^d \text{ Borel.} \tag{70}$$

For square-integrable random vectors, we write

$$\|\mathbf{X}\|_{L^2} := \big(\mathbb{E}\|\mathbf{X}\|^2\big)^{1/2}. \tag{71}$$

For any square-integrable random vectors $\mathbf{X}, \mathbf{Y}$ defined on the same probability space, the force stability assumptions give

$$\begin{aligned}\|F[\mathrm{Law}(\mathbf{X})](\mathbf{X}) - F[\mathrm{Law}(\mathbf{Y})](\mathbf{Y})\|_{L^2} &\le L_{\mathbf{x}}\|\mathbf{X}-\mathbf{Y}\|_{L^2} + L_{\mathcal{W}}\,\mathcal{W}_2\big(\mathrm{Law}(\mathbf{X}),\mathrm{Law}(\mathbf{Y})\big)\\ &\le \Lambda\|\mathbf{X}-\mathbf{Y}\|_{L^2}.\end{aligned} \tag{72}$$

Taking $\mathbf{Y} \equiv \mathbf{0}$ also yields

$$\|F[\mathrm{Law}(\mathbf{X})](\mathbf{X})\|_{L^2} \le K_0 + \Lambda\|\mathbf{X}\|_{L^2}. \tag{73}$$

These Lipschitz and growth estimates ensure well-posedness of the following McKean–Vlasov equations by Picard iteration.

Choose $(\mathbf{X}_0, \mathbf{U}_0^m)$ with law $\mu_0^{m,\beta}$, and let $\mathbf{B}_t$ be a standard d-dimensional Brownian motion independent of the initial data. Define

$$\begin{aligned} d\mathbf{X}_t^m &= \frac{\mathbf{U}_t^m}{m}\,dt,\\ d\mathbf{U}_t^m &= F[q_t^{m,\beta}](\mathbf{X}_t^m)\,dt - \frac{\gamma}{m}\mathbf{U}_t^m\,dt + \sqrt{\frac{2\gamma}{\beta}}\,d\mathbf{B}_t,\end{aligned} \tag{74}$$

where $q_t^{m,\beta} = \mathrm{Law}(\mathbf{X}_t^m)$. The joint law of $(\mathbf{X}_t^m, \mathbf{U}_t^m)$ solves equation 8.

Using the same Brownian motion and the same initial position, define

$$d\mathbf{Y}_t = \frac{1}{\gamma}F[q_t^\beta](\mathbf{Y}_t)\,dt + \sqrt{\frac{2}{\gamma\beta}}\,d\mathbf{B}_t, \qquad \mathbf{Y}_0 = \mathbf{X}_0, \qquad q_t^\beta = \mathrm{Law}(\mathbf{Y}_t). \tag{75}$$

Its law is the unique weak solution in $C([0,T];\mathcal{P}_2(\mathbb{R}^d))$ of

$$\partial_t q_t^\beta = -\frac{1}{\gamma}\nabla\cdot\big(q_t^\beta F[q_t^\beta]\big) + \frac{1}{\gamma\beta}\Delta q_t^\beta. \tag{76}$$

Since $F[q] = -\nabla\Phi_q$, this is equation 9, interpreted in the weak sense.

An overdamped moment bound. Let $M(t) := \mathbb{E}\|\mathbf{Y}_t\|^2$. By Itô's formula and equation 73,

$$M'(t) = \frac{2}{\gamma}\mathbb{E}\big[\mathbf{Y}_t \cdot F[q_t^\beta](\mathbf{Y}_t)\big] + \frac{2n}{\gamma\beta} \le \frac{1+2\Lambda}{\gamma}M(t) + \frac{K_0^2}{\gamma} + \frac{2n}{\gamma\beta}. \tag{77}$$

Gronwall's inequality gives $\sup_{0\le t\le T} M(t) < \infty$. Consequently,

$$C_F := \sup_{0\le t\le T}\mathbb{E}\|F[q_t^\beta](\mathbf{Y}_t)\|^2 < \infty. \tag{78}$$

The constant C_F is independent of m, because the overdamped process does not depend on m.

The small-mass estimate. Set

$$\varkappa_m(t) := e^{-\gamma t/m}. \tag{79}$$

Solving the linear momentum equation and integrating in time gives

$$\begin{aligned}\mathbf{X}_t^m = \mathbf{X}_0 &+ \frac{1-\varkappa_m(t)}{\gamma}\mathbf{U}_0^m\\ &+ \frac{1}{\gamma}\int_0^t \big(1-\varkappa_m(t-s)\big)F[q_s^{m,\beta}](\mathbf{X}_s^m)\,ds + \sqrt{\frac{2}{\gamma\beta}}\int_0^t\big(1-\varkappa_m(t-s)\big)\,d\mathbf{B}_s.\end{aligned} \tag{80}$$

Define

$$\mathbf{D}_t := \mathbf{X}_t^m - \mathbf{Y}_t, \qquad \Delta F_s := F[q_s^{m,\beta}](\mathbf{X}_s^m) - F[q_s^\beta](\mathbf{Y}_s). \tag{81}$$

Subtracting the integral equation for $\mathbf{Y}_t$ from equation 80 yields

$$\begin{aligned}\mathbf{D}_t = {} & \frac{1-\varkappa_m(t)}{\gamma}\mathbf{U}_0^m \\ & + \frac{1}{\gamma}\int_0^t \big(1-\varkappa_m(t-s)\big)\Delta F_s\, ds - \frac{1}{\gamma}\int_0^t \varkappa_m(t-s)F[q_s^\beta](\mathbf{Y}_s)\, ds - \sqrt{\frac{2}{\gamma\beta}}\int_0^t \varkappa_m(t-s)\, d\mathbf{B}_s.\end{aligned} \tag{82}$$

Write $\mathsf{e}(t) := \mathbb{E}\|\mathbf{D}_t\|^2$. By equation 72,

$$\mathbb{E}\|\Delta F_s\|^2 \le \Lambda^2 \mathsf{e}(s). \tag{83}$$

Moreover,

$$0 \le \varkappa_m(t) \le 1, \qquad \int_0^t \varkappa_m(t-s)\, ds \le \frac{m}{\gamma}, \qquad \int_0^t \varkappa_m(t-s)^2\, ds \le \frac{m}{2\gamma}. \tag{84}$$

Using $\|\boldsymbol{\xi}_1+\boldsymbol{\xi}_2+\boldsymbol{\xi}_3+\boldsymbol{\xi}_4\|^2 \le 4\sum_{j=1}^4 \|\boldsymbol{\xi}_j\|^2$, Cauchy–Schwarz, and Itô's isometry in equation 82, we obtain

$$\mathsf{e}(t) \le \frac{4C_0 m}{\gamma^2} + \frac{4\Lambda^2 T}{\gamma^2}\int_0^t \mathsf{e}(s)\, ds + \frac{4C_F m^2}{\gamma^4} + \frac{4dm}{\gamma^2\beta}. \tag{85}$$

Since $0 < m \le 1$, Gronwall's inequality gives

$$\sup_{0\le t\le T} \mathsf{e}(t) \le \frac{4}{\gamma^2}\left(C_0 + \frac{n}{\beta} + \frac{C_F}{\gamma^2}\right) e^{4\Lambda^2 T^2/\gamma^2}\, m. \tag{86}$$

The pair $(\mathbf{X}_t^m, \mathbf{Y}_t)$ is a coupling of $q_t^{m,\beta}$ and q_t^β. Therefore,

$$\mathcal{W}_2(q_t^{m,\beta}, q_t^\beta)^2 \le \mathsf{e}(t), \tag{87}$$

and hence

$$\sup_{0\le t\le T} \mathcal{W}_2(q_t^{m,\beta}, q_t^\beta) \le C_{T,\beta}\sqrt{m}, \tag{88}$$

with

$$C_{T,\beta} = \frac{2}{\gamma}\left(C_0 + \frac{n}{\beta} + \frac{C_F}{\gamma^2}\right)^{1/2} e^{2\Lambda^2 T^2/\gamma^2}, \tag{89}$$

which is independent of m. □

Corollary 1 (Zero-temperature limit)**.** *Under the force stability conditions of Assumption 3, let $q_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Let q_t^β and q_t be the unique weak solutions of equation 9 and equation 10, respectively, with the same initial law q_0. Then, for every $T > 0$,*

$$\sup_{t\in[0,T]} \mathcal{W}_2(q_t^\beta, q_t) \le \sqrt{\frac{2dT}{\gamma}}\, e^{\Lambda T/\gamma}\, \beta^{-1/2}, \qquad \Lambda := L_{\mathbf{x}} + L_{\mathcal{W}}. \tag{90}$$

In particular, $q^\beta \to q$ uniformly on every finite time interval in $\mathcal{W}_2$ as $\beta \to \infty$.

Proof. Let $\mathbf{B}_t$ be a standard d-dimensional Brownian motion. Couple the processes $\mathbf{Y}_t^\beta$ and $\mathbf{Y}_t$ through

$$\begin{aligned} d\mathbf{Y}_t^\beta &= \frac{1}{\gamma}F_p[q_t^\beta](\mathbf{Y}_t^\beta)\, dt + \sqrt{\frac{2}{\gamma\beta}}\, d\mathbf{B}_t, \\ d\mathbf{Y}_t &= \frac{1}{\gamma}F_p[q_t](\mathbf{Y}_t)\, dt, \end{aligned} \tag{91}$$

with a common initial random variable of law q_0, independent of $\mathbf{B}$. Their time-marginal laws are q_t^β and q_t, respectively.

Set

$$\mathsf{e}_\beta(t) := \mathbb{E}\|\mathbf{Y}_t^\beta - \mathbf{Y}_t\|^2. \tag{92}$$

Since $(\mathbf{Y}_t^\beta, \mathbf{Y}_t)$ is a coupling of q_t^β and q_t,

$$\mathcal{W}_2(q_t^\beta, q_t)^2 \le \mathsf{e}_\beta(t). \tag{93}$$

The force stability assumptions therefore imply

$$\begin{aligned}\Big(\mathbb{E}\|F_p[q_t^\beta](\mathbf{Y}_t^\beta) - F_p[q_t](\mathbf{Y}_t)\|^2\Big)^{1/2} &\le L_{\mathbf{x}}\,\mathsf{e}_\beta(t)^{1/2} + L_{\mathcal{W}}\,\mathcal{W}_2(q_t^\beta, q_t)\\ &\le \Lambda\,\mathsf{e}_\beta(t)^{1/2}.\end{aligned} \tag{94}$$

Applying Itô's formula and Cauchy–Schwarz gives, for almost every $t \in [0,T]$,

$$\begin{aligned}\mathsf{e}_\beta'(t) &= \frac{2}{\gamma}\mathbb{E}\Big[(\mathbf{Y}_t^\beta - \mathbf{Y}_t)\cdot\big(F_p[q_t^\beta](\mathbf{Y}_t^\beta) - F_p[q_t](\mathbf{Y}_t)\big)\Big] + \frac{2d}{\gamma\beta}\\ &\le \frac{2\Lambda}{\gamma}\mathsf{e}_\beta(t) + \frac{2d}{\gamma\beta}.\end{aligned} \tag{95}$$

Since $\mathsf{e}_\beta(0) = 0$, Gronwall's inequality yields

$$\mathsf{e}_\beta(t) \le \frac{2d}{\gamma\beta}\int_0^t e^{2\Lambda(t-s)/\gamma}\,ds \le \frac{2dt}{\gamma\beta}e^{2\Lambda t/\gamma}. \tag{96}$$

Combining this estimate with equation 93 and taking the supremum over $t \in [0,T]$ proves the claim. □

Proposition 3 (Non-target stationary state). Let $\Omega = [-1,1]$, let $p \in \mathcal{P}(\Omega)$ satisfy $\mathsf{R}_\# p = p$ for $\mathsf{R}(x) = -x$ with $p \neq \delta_0$, and assume Φ_{δ_0} admits a C^2 representative on Ω. Then, for every choice of generators, $q_t \equiv \delta_0$ is a stationary distributional solution of equation 10 and the metric slope $|\partial\mathcal{E}_p|(\delta_0)$ vanishes, although $\delta_0 \neq p$.

Proof. Write

$$\mathcal{U}(\alpha,\alpha') := \mathrm{UOT}^{\varphi_1,\varphi_2}(\alpha,\alpha'), \qquad \mathcal{G}(q) := \frac{1}{2}\big[\mathcal{U}(q,p) + \mathcal{U}(p,q)\big], \qquad \mathcal{S}(q) := \mathcal{U}(q,q). \tag{97}$$

Then

$$\mathcal{E}_p(q) = \mathcal{G}(q) - \frac{1}{2}\mathcal{S}(q) - \frac{1}{2}\mathcal{U}(p,p). \tag{98}$$

Set $\Phi := \Phi_{\delta_0}$. The first variation is understood in the sense that, for every $q \in \mathcal{P}(\Omega)$,

$$\lim_{\lambda\downarrow 0}\frac{\mathcal{E}_p((1-\lambda)\delta_0 + \lambda q) - \mathcal{E}_p(\delta_0)}{\lambda} = \int_\Omega \big(\Phi(x) - \Phi(0)\big)\,dq(x). \tag{99}$$

Reflection symmetry and stationarity. Simultaneous reflection of both coordinates preserves the quadratic cost and all divergence terms. Hence

$$\mathcal{U}(\mathsf{R}_\#\alpha, \mathsf{R}_\#\alpha') = \mathcal{U}(\alpha,\alpha'). \tag{100}$$

Since $\mathsf{R}_\# p = p$, it follows that

$$\mathcal{E}_p(\mathsf{R}_\# q) = \mathcal{E}_p(q). \tag{101}$$

This argument does not interchange the two marginal slots and therefore does not require $\varphi_1 = \varphi_2$.

Apply equation 101 to $q_\lambda = (1-\lambda)\delta_0 + \lambda\delta_x$ and differentiate at $\lambda = 0+$. By equation 99,

$$\Phi(x) - \Phi(0) = \Phi(-x) - \Phi(0). \tag{102}$$

Thus Φ is even, and its differentiability implies

$$\Phi'(0) = 0, \qquad F_p[\delta_0](0) = -\Phi'(0) = 0. \tag{103}$$

Consequently, for every $\zeta \in C_c^\infty(\mathbb{R})$, the constant curve $q_t = \delta_0$ satisfies

$$\frac{d}{dt}\int_\Omega \zeta\,dq_t = 0 = \frac{1}{\gamma}\int_\Omega \zeta'(x)F_p[\delta_0](x)\,d\delta_0(x). \tag{104}$$

It is therefore a stationary distributional solution of equation 10.

A bound for the self-interaction term. Define

$$H(\varsigma) := \rho\varphi_1(\varsigma) + \rho\varphi_2(\varsigma) + \varepsilon(\varsigma\log\varsigma - \varsigma + 1), \qquad \varsigma \geq 0, \tag{105}$$

with $0\log 0 := 0$, and let ς_* be a minimizer of H. Such a minimizer exists by lower semicontinuity and coercivity. If either generator is ι, the constraint enforces $\varsigma_* = 1$.

For any finite-cost plan π with total mass $\varsigma = \pi(\Omega^2)$, Jensen's inequality gives

$$D_{\varphi_k}(\pi_k\|q) \geq \varphi_k(\varsigma), \quad k \in \{1,2\}, \qquad \mathrm{KL}(\pi\|q\otimes q) \geq \varsigma\log\varsigma - \varsigma + 1. \tag{106}$$

Since the transport cost is nonnegative,

$$\mathcal{S}(q) \geq H(\varsigma_*) = \mathcal{S}(\delta_0). \tag{107}$$

Conversely, the competitor $\pi = \varsigma_*\, q\otimes q$ gives

$$\begin{aligned}\mathcal{S}(q) &\leq H(\varsigma_*) + \frac{\varsigma_*}{2}\int_{\Omega^2}(x-y)^2\, dq(x)\, dq(y)\\ &= \mathcal{S}(\delta_0) + \varsigma_*\left(\int_\Omega x^2\, dq(x) - \left(\int_\Omega x\, dq(x)\right)^2\right).\end{aligned} \tag{108}$$

Therefore,

$$0 \leq \mathcal{S}(q) - \mathcal{S}(\delta_0) \leq \varsigma_*\int_\Omega x^2\, dq(x). \tag{109}$$

A quadratic lower bound for the energy. The maps $q \mapsto \mathcal{U}(q,p)$ and $q \mapsto \mathcal{U}(p,q)$ are convex. Indeed, the primal objective is jointly convex in the plan and the varying input measure, by joint convexity of Csiszár divergences and the affine dependence of the product reference measure on that input. Taking the infimum over the plan preserves convexity. Thus $\mathcal{G}$ is convex.

Fix $q \in \mathcal{P}(\Omega)$ and set $q_\lambda = (1-\lambda)\delta_0 + \lambda q$. By equation 98, equation 109, and convexity of $\mathcal{G}$,

$$\begin{aligned}\mathcal{E}_p(q_\lambda) - \mathcal{E}_p(\delta_0) &= \mathcal{G}(q_\lambda) - \mathcal{G}(\delta_0) - \frac{1}{2}\big(\mathcal{S}(q_\lambda) - \mathcal{S}(\delta_0)\big)\\ &\leq \mathcal{G}(q_\lambda) - \mathcal{G}(\delta_0)\\ &\leq \lambda\big(\mathcal{G}(q) - \mathcal{G}(\delta_0)\big).\end{aligned} \tag{110}$$

Dividing by λ and letting $\lambda \downarrow 0$ yields

$$\mathcal{G}(q) - \mathcal{G}(\delta_0) \geq \int_\Omega \big(\Phi(x) - \Phi(0)\big)\, dq(x). \tag{111}$$

Using equation 109 once more, we obtain

$$\mathcal{E}_p(q) - \mathcal{E}_p(\delta_0) \geq \int_\Omega \big(\Phi(x) - \Phi(0)\big)\, dq(x) - \frac{\varsigma_*}{2}\int_\Omega x^2\, dq(x). \tag{112}$$

Let $C_\Phi := \|\Phi''\|_{L^\infty(\Omega)} < \infty$. Since $\Phi'(0) = 0$, Taylor's theorem gives

$$\Phi(x) - \Phi(0) \geq -\frac{C_\Phi}{2}x^2. \tag{113}$$

Consequently,

$$\mathcal{E}_p(q) - \mathcal{E}_p(\delta_0) \geq -\frac{C_\Phi + \varsigma_*}{2}\int_\Omega x^2\, dq(x) = -\frac{C_\Phi + \varsigma_*}{2}\mathcal{W}_2(q,\delta_0)^2. \tag{114}$$

Vanishing metric slope. By the definition of the descending metric slope,

$$\begin{aligned}|\partial\mathcal{E}_p|(\delta_0) &:= \limsup_{\substack{q\to\delta_0 \text{ in } \mathcal{W}_2\\ q\neq\delta_0}} \frac{[\mathcal{E}_p(\delta_0) - \mathcal{E}_p(q)]_+}{\mathcal{W}_2(q,\delta_0)}\\ &\leq \limsup_{q\to\delta_0}\frac{C_\Phi + \varsigma_*}{2}\mathcal{W}_2(q,\delta_0) = 0.\end{aligned} \tag{115}$$

The slope is nonnegative, so it equals zero. Finally, $\delta_0 \neq p$ by assumption. □

F LONG-TIME BEHAVIOR

This appendix contains the two long-time results referenced in Section 3.3: convergence to a stationary state without further assumptions, and algebraic convergence to the target under Assumption 4.

Proposition 4 (Convergence to a stationary state)**.** *Let $\Omega \subset \mathbb{R}^d$ be compact and let $(q_t)_{t\geq 0} \subset \mathcal{P}(\Omega)$ be a global regular solution of equation 10 such that $E(t) := \mathcal{E}_p(q_t)$ is absolutely continuous on every bounded interval with $E'(t) = -\frac{1}{\gamma}\mathcal{D}(q_t)$ for a.e. $t \geq 0$. Suppose $\mathcal{D} : \mathcal{P}(\Omega) \to [0,\infty]$ is lower semicontinuous with respect to $\mathcal{W}_2$. Then there exist $t_j \uparrow \infty$ and $q_\infty \in \mathcal{P}(\Omega)$ such that $\mathcal{W}_2(q_{t_j}, q_\infty) \to 0$ and $\mathcal{D}(q_\infty) = 0$; in particular, $\nabla\Phi_{q_\infty} = \mathbf{0}$ q_∞-almost everywhere, so q_∞ is a Wasserstein-stationary point of $\mathcal{E}_p$.*

Proof. Step 1: boundedness of the energy. Since Ω is compact, $c_{\max} := \sup_{\Omega^2} c < \infty$. For $\alpha_1, \alpha_2 \in \mathcal{P}(\Omega)$, all terms in equation 19 are nonnegative, so $\mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}(\alpha_1,\alpha_2) \geq 0$; the product coupling $\pi = \alpha_1 \otimes \alpha_2$ satisfies both marginal constraints exactly and has zero entropic penalty, so $\mathrm{UOT}^{\varphi_1,\varphi_2}_{\varepsilon,\rho}(\alpha_1,\alpha_2) \leq \int c\, d(\alpha_1 \otimes \alpha_2) \leq c_{\max}$. Hence $0 \leq \mathcal{A} \leq c_{\max}$ on $\mathcal{P}(\Omega)^2$ and $|\mathcal{E}_p(q)| \leq c_{\max}$ for every $q \in \mathcal{P}(\Omega)$.

Step 2: integrability of the dissipation. By hypothesis $E' \leq 0$ a.e., so E is nonincreasing and, by Step 1, bounded below; thus $E_\infty := \lim_{t\to\infty} E(t)$ exists in $[-c_{\max}, c_{\max}]$. For every $T > 0$, absolute continuity gives $\frac{1}{\gamma}\int_0^T \mathcal{D}(q_t)\, dt = E(0) - E(T)$, and letting $T \to \infty$ by monotone convergence,

$$\int_0^\infty \mathcal{D}(q_t)\, dt = \gamma\big(E(0) - E_\infty\big) \leq 2\gamma c_{\max} < \infty. \tag{116}$$

Step 3: a sequence along which the dissipation vanishes. Fix $j \in \mathbb{N}$. The set $\{t \geq 0 : \mathcal{D}(q_t) > 1/j\}$ has finite Lebesgue measure by equation 116, so its complement is unbounded. Choose inductively $t_j > t_{j-1} + 1$ with $\mathcal{D}(q_{t_j}) \leq 1/j$; then $t_j \uparrow \infty$ and $\mathcal{D}(q_{t_j}) \to 0$.

Step 4: compactness. Since Ω is compact, $(\mathcal{P}(\Omega), \mathcal{W}_2)$ is compact (Villani, 2009a). Hence $(q_{t_j})_j$ admits a subsequence, not relabeled, and a limit $q_\infty \in \mathcal{P}(\Omega)$ with $\mathcal{W}_2(q_{t_j}, q_\infty) \to 0$.

Step 5: stationarity of the limit. Lower semicontinuity of $\mathcal{D}$ and Step 3 give $0 \leq \mathcal{D}(q_\infty) \leq \liminf_{j\to\infty} \mathcal{D}(q_{t_j}) = 0$. Since $\mathcal{D}(q_\infty) = \int_\Omega \|\nabla\Phi_{q_\infty}\|^2\, dq_\infty$, this forces $\nabla\Phi_{q_\infty} = \mathbf{0}$ q_∞-a.e., which is the Wasserstein stationarity condition for equation 10. □

Proposition 5 (Conditional algebraic decay)**.** *Let $\Omega \subset \mathbb{R}^d$ be compact and $(q_t)_{t\geq 0}$ a global regular solution of equation 10 satisfying equation 11 and Assumption 4. With $E_0 := \mathcal{E}_p(q_0)$,*

$$\mathcal{E}_p(q_t) \leq \frac{\gamma A E_0}{\gamma A + t E_0} \quad \forall t \geq 0, \tag{117}$$

and $\mathcal{W}_2(q_t, p) \to 0$ as $t \to \infty$. Assumption 4(ii) holds with $A = \operatorname{ess\,sup}_t C_{\mathrm{PI}}(q_t)\, \chi^2(p\|q_t)$ whenever $\mathcal{E}_p$ is convex along mixtures, $p \ll q_t$ with $\frac{dp}{dq_t} \in L^2(q_t)$, and q_t satisfies a Poincaré inequality with constant $C_{\mathrm{PI}}(q_t)$ (Lemma 1).

Proof. Set $E(t) := \mathcal{E}_p(q_t)$. Combining equation 11 with Assumption 4(ii) gives $E'(t) \leq -E(t)^2/(\gamma A)$ for a.e. t, and $E \geq 0$ since p minimizes $\mathcal{E}_p$ with $\mathcal{E}_p(p) = 0$. If $E_0 = 0$ the bound is trivial. Otherwise, on any interval where $E > 0$, the chain rule for absolutely continuous functions yields $\frac{d}{dt}\big(1/E\big) \geq 1/(\gamma A)$, hence $1/E(t) \geq 1/E_0 + t/(\gamma A)$, which is equation 117; if E reaches zero at some t_*, monotonicity and nonnegativity keep it at zero thereafter, so the bound holds for all t. In particular $E(t) \to 0$. Since $\mathcal{P}(\Omega)$ is $\mathcal{W}_2$-compact, every sequence $t_j \to \infty$ has a subsequence with $q_{t_j} \to q_*$, and lower semicontinuity gives $\mathcal{E}_p(q_*) \leq \liminf_j E(t_j) = 0$, so $q_* = p$ by uniqueness of the minimizer. As every subsequence has a further subsequence converging to p, $\mathcal{W}_2(q_t, p) \to 0$. □

G IMPLEMENTATION DETAILS

ImageNet experiments. Detailed hyperparameters and configurations are given in Table 5. We largely follow the setup of W-Flow (Han et al., 2026) without extensive hyperparameter tuning; the

Algorithm 1: One UOT-GF training step

Input: generator $f_{\boldsymbol{\theta}}(\cdot; \mathsf{c}, w)$; frozen feature blocks $\{\psi_\ell\}_{\ell=1}^{n_\psi}$; ε, τ, η; guidance distribution $p(w)$; Sinkhorn iterations L

Sample $w \sim p(w)$ and N_{c} classes; for each class c: real batch $\mathcal{Y}_{\mathsf{c}} = \{\mathbf{y}_{\mathsf{c},j}\}_{j=1}^{N_{\text{pos}}}$, noise $\{\mathbf{z}_{\mathsf{c},i}\}_{i=1}^{N_{\text{neg}}}$ and $\{\tilde{\mathbf{z}}_{\mathsf{c},i}\}_{i=1}^{N_{\text{neg}}}$;

Sample an unconditional real batch $\mathcal{Y}_\varnothing$ (N_{unc});

$\mathcal{X}_{\mathsf{c}} \leftarrow \{f_{\boldsymbol{\theta}}(\mathbf{z}_{\mathsf{c},i}; \mathsf{c}, w)\}_i, \quad \tilde{\mathcal{X}}_{\mathsf{c}} \leftarrow \{f_{\boldsymbol{\theta}}(\tilde{\mathbf{z}}_{\mathsf{c},i}; \mathsf{c}, w)\}_i$; `// no gradient`

for *each block* ℓ *and class* c **do**

$\mathbf{x}_i \leftarrow \psi_\ell(f_{\boldsymbol{\theta}}(\mathbf{z}_{\mathsf{c},i}; \mathsf{c}, w))$ for all i;

$\widehat{\mathbf{v}}_{\mathsf{c}} \leftarrow \textsc{Velocity}(\{\mathbf{x}_i\}, \psi_\ell(\mathcal{Y}_{\mathsf{c}})), \quad \widehat{\mathbf{v}}_{\text{self}} \leftarrow \textsc{Velocity}(\{\mathbf{x}_i\}, \psi_\ell(\tilde{\mathcal{X}}_{\mathsf{c}}))$,

$\widehat{\mathbf{v}}_\varnothing \leftarrow \textsc{Velocity}(\{\mathbf{x}_i\}, \psi_\ell(\mathcal{Y}_\varnothing))$;

$\widehat{\mathbf{v}}^{\ell}_{\mathsf{c},i} \leftarrow (\widehat{\mathbf{v}}_{\mathsf{c},i} - \widehat{\mathbf{v}}_{\text{self},i}) + w\,(\widehat{\mathbf{v}}_{\mathsf{c},i} - \widehat{\mathbf{v}}_{\varnothing,i})$;

$\bar{\mathbf{x}}^{\ell}_{\mathsf{c},i} \leftarrow \mathrm{sg}\big(\mathbf{x}_i + \eta\, \widehat{\mathbf{v}}^{\ell}_{\mathsf{c},i}\big)$;

$$\mathcal{L}(\boldsymbol{\theta}) \leftarrow \frac{1}{n_\psi N_{\mathsf{c}} N_{\text{neg}}} \sum_{\ell,\mathsf{c},i} \big\| \psi_\ell(f_{\boldsymbol{\theta}}(\mathbf{z}_{\mathsf{c},i}; \mathsf{c}, w)) - \bar{\mathbf{x}}^{\ell}_{\mathsf{c},i} \big\|^2;$$

Update $\boldsymbol{\theta}$ by a gradient step on $\mathcal{L}$;

Algorithm 2: Source-fixed Sinkhorn

Input: cost $\mathbf{C} \in \mathbb{R}^{N \times M}$; ε; $\tau = \rho/(\rho + \varepsilon)$; iterations L

Output: plan $\mathbf{P}$ with $\mathbf{P}\mathbf{1}_M = \frac{1}{N}\mathbf{1}_N$

$\mathbf{K} \leftarrow -\mathbf{C}/\varepsilon; \quad \log \mathbf{b} \leftarrow \mathbf{0}$;

for $t = 1, \ldots, L$ **do**

$\log a_i \leftarrow \log \frac{1}{N} - \mathrm{LSE}_j\left(K_{ij} + \log b_j\right)$; `// exact source projection`

$\log b_j \leftarrow \log \frac{1}{M} - \tau\, \mathrm{LSE}_i\left(K_{ij} + \log a_i\right)$; `// relaxed target projection`

$\log a_i \leftarrow \log \frac{1}{N} - \mathrm{LSE}_j\left(K_{ij} + \log b_j\right)$; `// final projection`

$P_{ij} \leftarrow \exp\left(\log a_i + K_{ij} + \log b_j\right)$;

only new hyperparameters are the marginal penalties (φ_1, φ_2), the relaxation strength τ, and the symmetrization of the velocity estimator. Experiments are conducted on Nvidia H20 GPUs with 141 GB of memory each. The B/2, L/2, and XL/2 models are trained on 8 nodes of 8 GPUs; the ablations use a single node of 8 GPUs. Training L/2 takes about 5 days.

G.1 Training Algorithm

Algorithm 1 gives one training step. All transport quantities are computed under `no_grad`; gradients reach the generator only through the regression loss, so the generator regresses onto a stop-gradient target displaced along the UOT force.

G.2 Source-Fixed Sinkhorn Solver

For a cost matrix $\mathbf{C} \in \mathbb{R}^{N \times M}$ with uniform weights $\frac{1}{N}\mathbf{1}_N$, $\frac{1}{M}\mathbf{1}_M$, the source-fixed problem equation 12 with $(\varphi_1, \varphi_2) = (\iota, \mathrm{KL})$ is solved in the log domain. With $\mathbf{K} := -\mathbf{C}/\varepsilon$ and $\tau = \rho/(\rho + \varepsilon)$, initialize $\log \mathbf{b} = \mathbf{0}$ and iterate L times

$$\log a_i \leftarrow \log \tfrac{1}{N} - \mathrm{LSE}_j\left(K_{ij} + \log b_j\right), \qquad \log b_j \leftarrow \log \tfrac{1}{M} - \tau\, \mathrm{LSE}_i\left(K_{ij} + \log a_i\right), \quad (118)$$

where LSE is log-sum-exp. The first update is the exact source projection ($\varphi_1 = \iota$); the second is the KL-relaxed target projection, whose exponent $\tau < 1$ is the standard damping for a $\rho\,\mathrm{KL}$ penalty (Séjourné et al., 2019). After the loop one further source update is applied so that the returned plan $P_{ij} = \exp(\log a_i + K_{ij} + \log b_j)$ satisfies $\mathbf{P}\mathbf{1}_M = \frac{1}{N}\mathbf{1}_N$ exactly for any finite L; this also keeps the particle velocities correctly scaled. Setting $\tau = 1$ recovers balanced Sinkhorn and hence W-Flow.

G.3 VELOCITY ESTIMATOR

$\textsc{Velocity}(\mathbf{x}, \mathbf{y})$ estimates the symmetrized cross-interaction contribution in Proposition 1 from the empirical plans. Let $\mathbf{P}^{\rightarrow} \in \mathbb{R}_+^{N\times M}$ be the plan of $\mathrm{UOT}(\widehat{q}, \widehat{p})$ ($\mathbf{x}$ source-fixed), and $\mathbf{P}^{\leftarrow} \in \mathbb{R}_+^{M\times N}$ that of $\mathrm{UOT}(\widehat{p}, \widehat{q})$ ($\mathbf{y}$ source-fixed, $\mathbf{x}$ relaxed). The barycenters and mass ratios below are computed from these empirical plans:

$$\text{forward:}\quad \widehat{\mathbf{v}}_i^{\rightarrow} = \frac{\sum_j P_{ij}^{\rightarrow}\mathbf{y}_j}{\sum_j P_{ij}^{\rightarrow}} - \mathbf{x}_i = B_{q,p}^{(1)}(\mathbf{x}_i) - \mathbf{x}_i, \tag{119}$$

$$\text{reverse:}\quad \widehat{\mathbf{v}}_i^{\leftarrow} = N\Big(\sum_j P_{ji}^{\leftarrow}\mathbf{y}_j - \pi_{2,i}^{\leftarrow}\mathbf{x}_i\Big) = r_{p,q}^{(2)}(\mathbf{x}_i)\big(B_{p,q}^{(2)}(\mathbf{x}_i) - \mathbf{x}_i\big), \qquad \pi_{2,i}^{\leftarrow} := \sum_j P_{ji}^{\leftarrow}. \tag{120}$$

We use $\textsc{Velocity}(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\widehat{\mathbf{v}}^{\rightarrow} + \widehat{\mathbf{v}}^{\leftarrow})$. In equation 119, the row mass equals $1/N$ exactly, so $r^{(1)} \equiv 1$; in equation 120, the incoming mass $\pi_{2,i}^{\leftarrow}$ may deviate from $1/N$, giving $r_i^{(2)} = N\pi_{2,i}^{\leftarrow}$, which averages to one over the batch but may be below or above one locally. The same estimator with $\mathbf{y}$ drawn from a second generated batch gives the self-interaction term $\widehat{\mathbf{v}}_{\mathrm{self}}$; with $\mathbf{y}$ drawn from unconditional data, it gives the unconditional term $\widehat{\mathbf{v}}_{\varnothing}$ for guidance (Appendix G.4).

Effective interaction and marginal bounds. Define the effective cross-interaction matrix

$$\overline{\mathbf{P}} := \tfrac{1}{2}\big(\mathbf{P}^{\rightarrow} + (\mathbf{P}^{\leftarrow})^{\top}\big).$$

The estimator can equivalently be written as

$$\textsc{Velocity}(\mathbf{x}, \mathbf{y})_i = N\sum_j \overline{P}_{ij}(\mathbf{y}_j - \mathbf{x}_i).$$

Since each directed solve fixes its source marginal,

$$\mathbf{P}^{\rightarrow}\mathbf{1}_M = \tfrac{1}{N}\mathbf{1}_N, \qquad \mathbf{P}^{\leftarrow}\mathbf{1}_N = \tfrac{1}{M}\mathbf{1}_M.$$

Consequently,

$$\overline{\mathbf{P}}\mathbf{1}_M = \tfrac{1}{2}\Big(\tfrac{1}{N}\mathbf{1}_N + (\mathbf{P}^{\leftarrow})^{\top}\mathbf{1}_M\Big) \geq \tfrac{1}{2N}\mathbf{1}_N, \tag{121}$$

$$\overline{\mathbf{P}}^{\top}\mathbf{1}_N = \tfrac{1}{2}\Big((\mathbf{P}^{\rightarrow})^{\top}\mathbf{1}_N + \tfrac{1}{M}\mathbf{1}_M\Big) \geq \tfrac{1}{2M}\mathbf{1}_M, \tag{122}$$

where the inequalities are componentwise. Equivalently, the associated measure $\bar{\pi} = \sum_{i,j}\overline{P}_{ij}\delta_{(\mathbf{x}_i,\mathbf{y}_j)}$ satisfies

$$\bar{\pi}_1 \geq \tfrac{1}{2}\widehat{q}, \qquad \bar{\pi}_2 \geq \tfrac{1}{2}\widehat{p}.$$

Thus, the effective cross interaction allows marginal reweighting while retaining at least half of each reference marginal. These bounds also hold for finitely many Sinkhorn iterations, up to floating-point error, because both directed solves end with a source projection. They concern cross-interaction mass and guarantee neither a nonzero net force nor target coverage.

Symmetry of the construction. For the symmetric cost and common regularization parameters used here, transposing a transport plan gives

$$\mathrm{UOT}^{\phi_1,\phi_2}(q, p) = \mathrm{UOT}^{\phi_2,\phi_1}(p, q).$$

Exchanging the two marginal penalties therefore leaves the symmetrized cost A, the energy E_p, and its exact gradient field unchanged. Here "source-fixed" describes each directed solve; the full construction uses symmetrized one-sided relaxation. This symmetry concerns the exact transport objectives, whereas the marginal bounds above also apply to the finite-iteration plans.

G.4 GUIDANCE

Following Han et al. (2026), the guidance weight w is sampled at each training step from the distribution in Table 5 and the generator is conditioned on it, so a single model covers $w \in [0, 3]$; at

inference a fixed w is chosen by search and sampling remains one forward pass. We report guidance as $s := w + 1$ in the main text. For a class c and weight w, three velocities are estimated from the same generated batch: $\widehat{\mathbf{v}}_{\mathsf{c}}$ toward real samples of class c, $\widehat{\mathbf{v}}_{\mathrm{self}}$ toward a second generated batch of class c, and $\widehat{\mathbf{v}}_{\varnothing}$ toward N_{unc} real samples drawn from all classes. The regression target uses

$$\widehat{\mathbf{v}}_w = (\widehat{\mathbf{v}}_c - \widehat{\mathbf{v}}_{\mathrm{self}}) + w\,(\widehat{\mathbf{v}}_c - \widehat{\mathbf{v}}_{\varnothing}), \tag{123}$$

in which the self-interaction term cancels from the guidance difference. No label dropout is used; the unconditional direction comes from the unconditional data batch, not from an unconditional generator.

G.5 Feature-Space Transport

Costs are squared Euclidean distances between features of the released pretrained MAE encoder (Deng et al., 2026), $C^{\ell}_{ij} = \frac{1}{2}\|\psi_\ell(\mathbf{x}_i) - \psi_\ell(\mathbf{y}_j)\|^2$, computed independently for each selected block ℓ. Plans, velocities, and the regression target are formed per block, and the per-block losses are averaged. Block selection and feature normalization follow Han et al. (2026).

G.6 Hyperparameters

Table 5 lists the configuration. Unless stated, values are inherited from W-Flow (Han et al., 2026).

Ablations (Section 5) use DiT-B/2 with $N_c = 64$ (effective batch 4096), 30K steps, $\varepsilon = 0.05$, and guidance $s = 1.5$.

G.7 Diagnostics

The solver logs, per plan: the source residual $\max_i |\pi_{1,i} - 1/N|$ with $\pi_{1,i} = \sum_j P_{ij}$ (zero up to floating point after the final projection); the target $\mathrm{KL}(\mathbf{P}^\top \mathbf{1}_N \,\|\, \frac{1}{M}\mathbf{1}_M)$; the target effective sample size fraction $\big(\sum_j \pi_{2,j}\big)^2 / \big(M \sum_j \pi_{2,j}^2\big)$ with $\pi_{2,j} = \sum_i P_{ij}$, which measures how much target mass is reallocated; the range $[\min_j, \max_j]$ of $M\pi_{2,j}$, the empirical $r^{(2)}$; and the reverse fraction $\|\widehat{\mathbf{v}}^{\leftarrow}\|_{\mathrm{rms}} / \|\widehat{\mathbf{v}}^{\rightarrow}\|_{\mathrm{rms}}$. Figure 4 reports $M\pi_{2,j}$ on real ImageNet targets for our checkpoint.

G.8 Evaluation

All metrics use 50K generated images against the ImageNet validation reference. FID in the main text follows the ADM evaluation suite (Dhariwal & Nichol, 2021). W-Flow's published numbers use the JiT protocol (Li & He, 2026); Table 6 reports both protocols for W-Flow and for UOT-GF. Precision/recall use $k = 3$ and density/coverage $k = 5$ (Kynkäänniemi et al., 2019; Naeem et al., 2020), in both Inception and DINOv2-CLS feature spaces. Guidance scales for reported FIDs are those of Table 5; Figure 12 gives the FID–IS trade-off.

G.9 Toy example

Synthetic transport diagnostics. We construct four two-dimensional source–target problems to compare balanced OT with symmetrized source-fixed UOT under target perturbations. Let $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\mathbf{y}_j\}_{j=1}^M$ denote the source and target samples with uniform weights $\frac{1}{N}\mathbf{1}_N$ and $\frac{1}{M}\mathbf{1}_M$. In every experiment the ground cost is $C_{ij} = \frac{1}{2}\|\mathbf{x}_i - \mathbf{y}_j\|^2$ and the entropic regularization is $\varepsilon = 0.05$. The balanced baseline minimizes

$$\langle \mathbf{C}, \mathbf{P}\rangle + \varepsilon\,\mathrm{KL}\big(\mathbf{P} \,\big\|\, \tfrac{1}{NM}\mathbf{1}_N\mathbf{1}_M^\top\big) \quad \text{subject to} \quad \mathbf{P}\mathbf{1}_M = \tfrac{1}{N}\mathbf{1}_N, \qquad \mathbf{P}^\top\mathbf{1}_N = \tfrac{1}{M}\mathbf{1}_M.$$

The source-fixed unbalanced variant instead solves

$$\min_{\mathbf{P}\geq 0,\ \mathbf{P}\mathbf{1}_M = \frac{1}{N}\mathbf{1}_N} \langle \mathbf{C}, \mathbf{P}\rangle + \varepsilon\,\mathrm{KL}\big(\mathbf{P} \,\big\|\, \tfrac{1}{NM}\mathbf{1}_N\mathbf{1}_M^\top\big) + \rho\,\mathrm{KL}\big(\mathbf{P}^\top\mathbf{1}_N \,\big\|\, \tfrac{1}{M}\mathbf{1}_M\big),$$

where $\tau = \rho/(\rho+\varepsilon) = 0.985$, corresponding to $\rho \approx 3.283$. The displayed optimization problems define one directed solve. We solve each problem in both directions, obtaining $\mathbf{P}^{\rightarrow} \in \mathbb{R}_+^{N\times M}$ for source-to-target transport and $\mathbf{P}^{\leftarrow} \in \mathbb{R}_+^{M\times N}$ for target-to-source transport, with the cost matrix

	Ablation (Table 1)	**UOT-GF B/2**	**UOT-GF L/2**	**UOT-GF XL/2**
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	bottleneck	bottleneck	bottleneck
ResNet: blocks / stage	[3, 4, 6, 3]	[3, 4, 6, 3]	[3, 4, 6, 3]	[3, 4, 6, 3]
ResNet: size / stage	$[32^2, 16^2, 8^2, 4^2]$	$[32^2, 16^2, 8^2, 4^2]$	$[32^2, 16^2, 8^2, 4^2]$	$[32^2, 16^2, 8^2, 4^2]$
MAE: masking ratio	50%	50%	50%	50%
MAE: pre-train epochs	192	1280	1280	1280
classification finetune	No	3k steps	3k steps	3k steps
Generator Optimizer				
optimizer	AdamW	AdamW	AdamW	AdamW
(β_1, β_2)	(0.9, 0.95)	(0.9, 0.95)	(0.9, 0.95)	(0.9, 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 ($N_{\text{c}}\times N_{\text{neg}}$)	4096	8192	8192	8192
marginal penalties (φ_1, φ_2)	(ι, KL)	(ι, KL)	(ι, KL)	(ι, KL)
relaxation τ	0.985	0.985	0.985	0.985
regularization ε	0.05	0.05	0.05	0.05
Sinkhorn iterations L	10	10	10	10
symmetrized velocity	yes	yes	yes	yes
self-transport diagonal mask	no	no	no	no
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	$(w{+}1)^{-3}$	$(w{+}1)^{-5}$	50%: w=0; 50%: $(w{+}1)^{-3}$	20%: w=0; 80%: $(w{+}1)^{-4}$
train: uncond samples N_{unc}	16	32	32	32
inference: CFG w search	$[0.0, 2.5]$	$[0.0, 2.5]$	$[0.0, 2.5]$	$[0.0, 2.5]$
inference: selected $s{=}w{+}1$	1.50	1.19	1.15	1.14

Table 5: Configurations and hyperparameters. All rows except the transport rows ((φ_1, φ_2), τ, L, ϵ, symmetrization, diagonal mask) follow Han et al. (2026, Table 5). CFG sampling entries abbreviate $p(w) \propto (\cdot)$.

Table 6: FID under the JiT and ADM evaluation protocols.

Model / setting	CFG	FID(JiT) ↓	FID(ADM) ↓
Drifting Pixel-L (Deng et al., 2026)	1.00	1.597925	1.463253
W-Flow-B (Han et al., 2026)	1.19	1.518592	1.521997
W-Flow-L (Han et al., 2026)	1.14	1.352408	1.357978
W-Flow-XL (Han et al., 2026)	1.09	1.295583	1.333832
UOT-GF-B ($\tau = 0.985$)	1.19	1.469181	1.461928
UOT-GF-B ($\tau = 0.985,\ i = 1$)	1.19	1.548405	1.539864
UOT-GF-L ($\tau = 0.985$)	1.15	1.348107	1.341342
UOT-GF-L ($\tau = 0.985,\ i = 1$)	1.15	1.405842	1.400581
UOT-GF-XL 100k	1.15	1.261174	1.261597
UOT-GF-XL 200k	1.14	1.225218	1.219600

transposed and the reference weights exchanged in the reverse solve. The effective cross-interaction

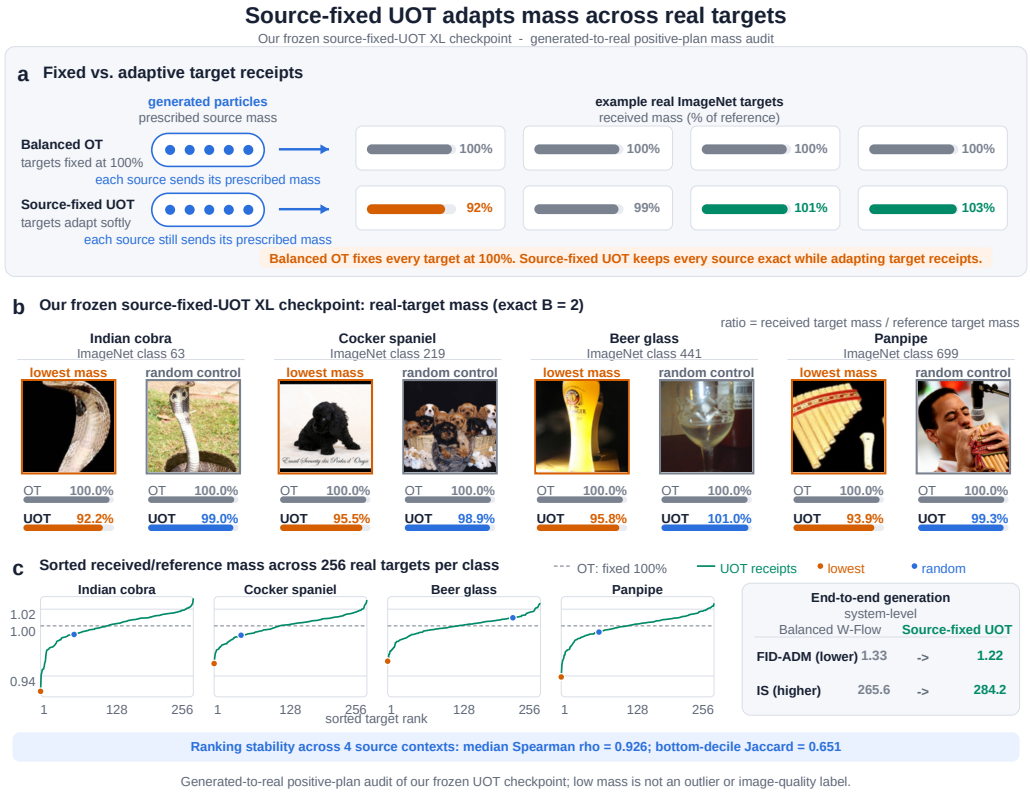


Figure 4: **Forward source-fixed UOT reweights real targets. (a)** Balanced OT fixes both marginals; source-fixed UOT fixes the generated marginal and KL-relaxes the real-data marginal. **(b)** Lowest-mass targets and class-matched controls, where r_j is received mass relative to the uniform target marginal. **(c)** Rankings over 256 targets per class are stable across four contexts ($\rho = 0.926$, Jaccard $= 0.651$). Lower mass indicates reduced weight in the forward interaction, not lower image quality. Right: final-checkpoint FID and IS.

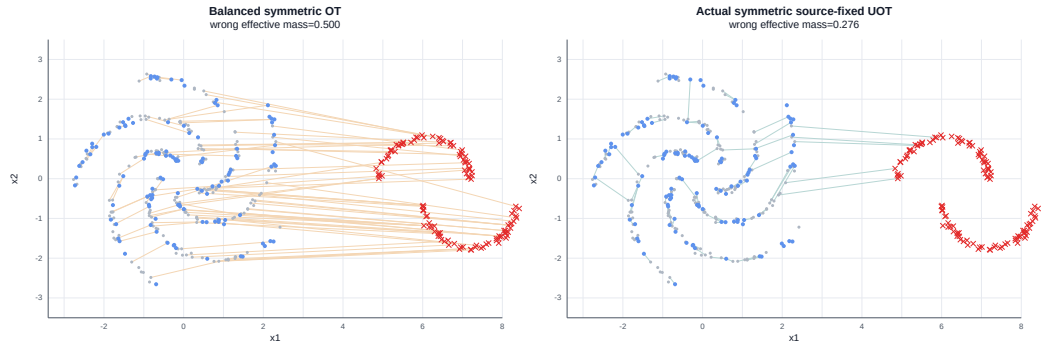


Figure 5: **Wrong-manifold targets.** Part of the target mass (red crosses) lies on a corrupted manifold. Balanced OT (left) must transport full mass to it; symmetrized source-fixed UOT (right) reduces the mass sent there. Gray: sources; blue: nominal targets.

plan is

$$\overline{\mathbf{P}} := \tfrac{1}{2}\big(\mathbf{P}^{\rightarrow} + (\mathbf{P}^{\leftarrow})^{\top}\big).$$

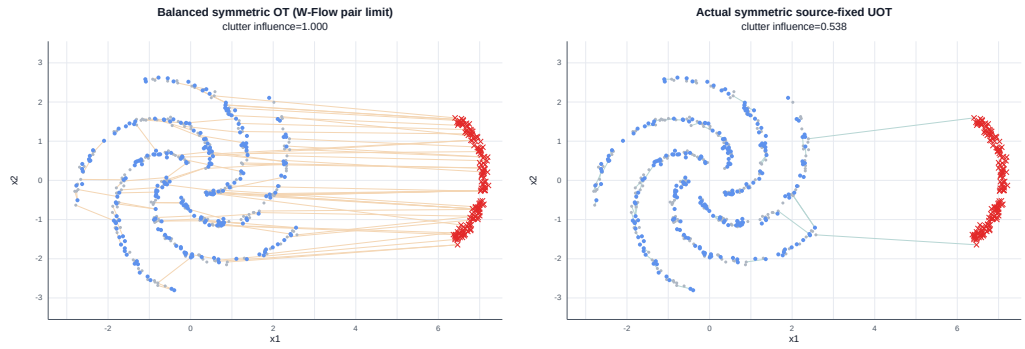


Figure 6: **Spurious-island targets.** A cluster of spurious targets (red crosses) appears away from the data. Balanced OT (left) couples sources to it at full weight; symmetrized source-fixed UOT (right) reduces its influence.

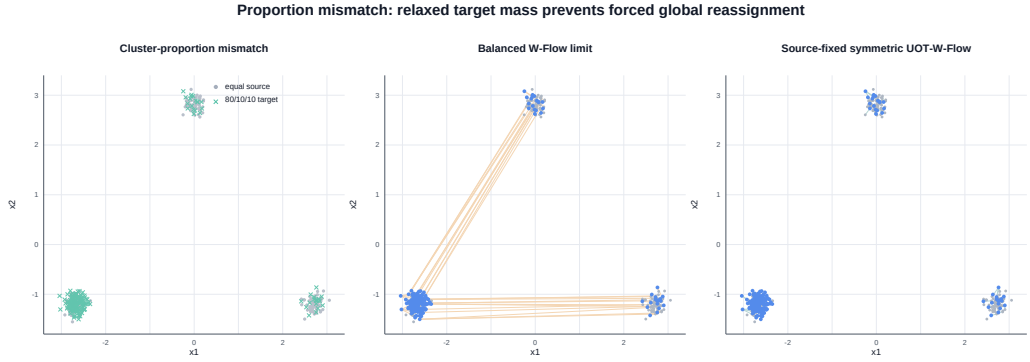


Figure 7: **Mode imbalance.** Equal-weight sources (gray dots) against an 80/10/10 target (teal crosses), left. Balanced OT (middle) reassigns source mass across clusters to match the proportions; symmetrized source-fixed UOT (right) keeps sources on nearby clusters.

For source-fixed UOT, $\mathbf{P}^{\rightarrow}\mathbf{1}_M = \frac{1}{N}\mathbf{1}_N$ and $\mathbf{P}^{\leftarrow}\mathbf{1}_N = \frac{1}{M}\mathbf{1}_M$. Consequently,

$$\overline{\mathbf{P}}\mathbf{1}_M \geq \frac{1}{2N}\mathbf{1}_N, \qquad \overline{\mathbf{P}}^{\top}\mathbf{1}_N \geq \frac{1}{2M}\mathbf{1}_M,$$

componentwise. Thus, each directed solve fixes its own source marginal, while both effective marginals may be reweighted. The balanced comparison uses $\tau = 1$ in both directions.

The plans are computed with Algorithm 2 in double precision; the balanced baseline sets $\tau = 1$, and no diagonal mask is applied. The four problems use $L = 3000/80, 3000/150, 400/400$, and $3000/80$ iterations for the balanced/source-fixed solvers, respectively. All random samples are reproducible from the base seed 20260831, with separate deterministic seed offsets for each source, target, and perturbation.

Wrong-manifold replacement (Figure 5). We sample 192 source points and 120 clean target points from a five-arm pinwheel. For each point, the radius is uniform on $[0.45, 2.75]$, and the angle is $2\pi k/5 + 1.17r + \eta$, where k is the arm index and $\eta \sim \mathcal{N}(0, 0.045^2)$; isotropic Cartesian noise with standard deviation 0.03 is then added. Another 120 target points are sampled from a shifted, noisy two-moons distribution. These distractors constitute half of the observed target sample.

Spurious target island (Figure 6). The source contains 240 points from a five-arm pinwheel, and the observed target contains 240 independently sampled clean pinwheel points plus 160

remote crescent-shaped distractors. The latter therefore account for 40% of the 400 target points. The clean pinwheel uses radii uniform on $[0.45, 2.85]$, angular coefficient 1.16, angular noise of standard deviation 0.045, and Cartesian noise of standard deviation 0.03. For the distractor crescent, $\theta \sim \mathrm{Uniform}[-1.28, 1.28]$ and $R \sim \mathcal{N}(1.02, 0.045^2)$, with coordinates $(6.10 + R\cos\theta,\ 0.05 + 1.55R\sin\theta)$ and additional isotropic noise of standard deviation 0.025.

Cluster-proportion mismatch (Figure 7). Source and target points are sampled around the same three centers, $(-2.7, -1.2)$, $(2.7, -1.2)$, and $(0, 2.8)$, with isotropic Gaussian noise of standard deviation 0.12. The 180 source points are divided equally among the three components (60/60/60), whereas the 180 target points have counts 144/18/18, giving an 80/10/10 empirical target proportion.

Oversampled-mode minibatch (Figure 2). Eight Gaussian modes are placed on a ring of radius 2.45, at angles $2\pi k/8 + \pi/8$. All points have isotropic Gaussian noise of standard deviation 0.16. The 192 source points are equally divided among the modes. Of the 240 independently sampled target points, 80% belong to one mode, with the remainder distributed across the other seven. This is a near-duplicate *mode burst*, not a set of point-identical copies.

The panels display the effective cross-interaction plans $\overline{\mathbf{P}}$. For each method, we connect evenly spaced source points to their largest-coupling targets,

$$j^*(i) = \arg\max_j \overline{P}_{ij}.$$

The four problems display 96, 100, 54, and 96 such links, respectively. Colors follow Figure 2, and link width does not encode transported mass. These plots visualize couplings constructed from both transport directions, rather than integrated particle trajectories.

H Additional results

Efficiency evaluation. We evaluated all methods at 256×256 resolution on NVIDIA H20 GPUs, assigning one GPU to each process. Inference used FP32 tensors with TF32 enabled. We retained each model's high-fidelity sampling configuration: SiT-XL/2 used a 250-step Euler SDE sampler with CFG 1.5 and the official EMA VAE; RAE-XL used a 50-step Euler ODE sampler with autoguidance 1.42; BAR-L used guidance 5.3, temperature 3.0, token allocation $[2, 2, 5, 7]$, and KV caching; and UOT-GF used its 200k-step XL checkpoint with CFG 1.14 and one generator evaluation.

For throughput, we used a batch size of 125, discarded 10 warm-up batches, and timed the following 390 batches. Throughput is the number of timed images divided by the elapsed generation-and-decoding time. For single-image latency, we fixed the conditioning label to ImageNet class 0, discarded 20 warm-up requests, and timed 200 batch-one requests individually, synchronizing CUDA immediately before and after each request. We report the mean latency; checkpoint loading and file I/O are excluded from both timing measurements.

We estimated FLOPs per image by running the complete sampling and decoding path under PyTorch's `FlopCounterMode` and dividing the counted operations by the batch size. These are theoretical estimates for supported operators, not measurements of hardware utilization or energy consumption. The speedup in panel is each model's batch-125 throughput divided by SiT-XL/2's batch-125 throughput.

Controlled visualization of target-marginal relaxation. Figure 9 isolates the transport mechanism by applying balanced OT and source-fixed UOT to the *same* generated sources and real ImageNet targets. The 128 source particles are generated with a fixed seed from our frozen EMA W-Flow-XL/2 checkpoint at 180k steps (class 178, CFG 1.14). The 128 targets consist of 96 distinct class-178 images, 16 exact copies of four of those images, and 16 foreign-class images (class 0). We encode the targets as 4096-dimensional global SD-VAE latents and use the training-time input scaling and quadratic cost for both operators, with $\epsilon = 0.05$ ($\epsilon_{\text{eff}} = 204.8$) and 10 Sinkhorn iterations; source-fixed UOT additionally uses $\tau = 0.985$. Only the positive-forward coupling is shown. The upper plots report each target's received mass relative to its uniform reference mass: balanced OT keeps this ratio near one, whereas source-fixed UOT preserves the source marginal while allowing target-specific deviations through a soft KL penalty. The lower heatmaps show row-normalized coupling

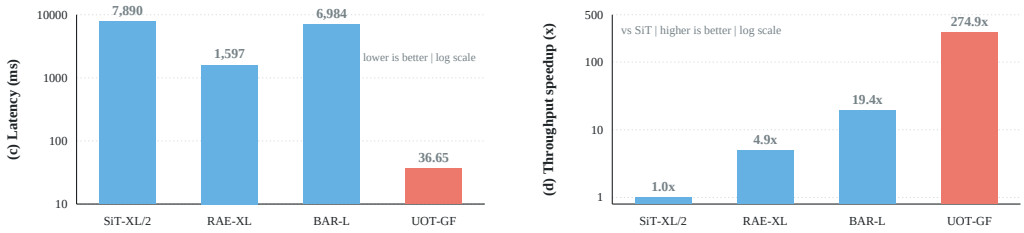


Figure 8: Single-H20 inference efficiency at 256×256 resolution. (c) Mean batch-one generation-and-decoding latency over 200 timed requests. (d) Batch-125 throughput speedup relative to SiT-XL/2. Both panels use logarithmic vertical axes and model-specific high-fidelity sampling settings. The speedup in (d) is not computed from batch-one latency.

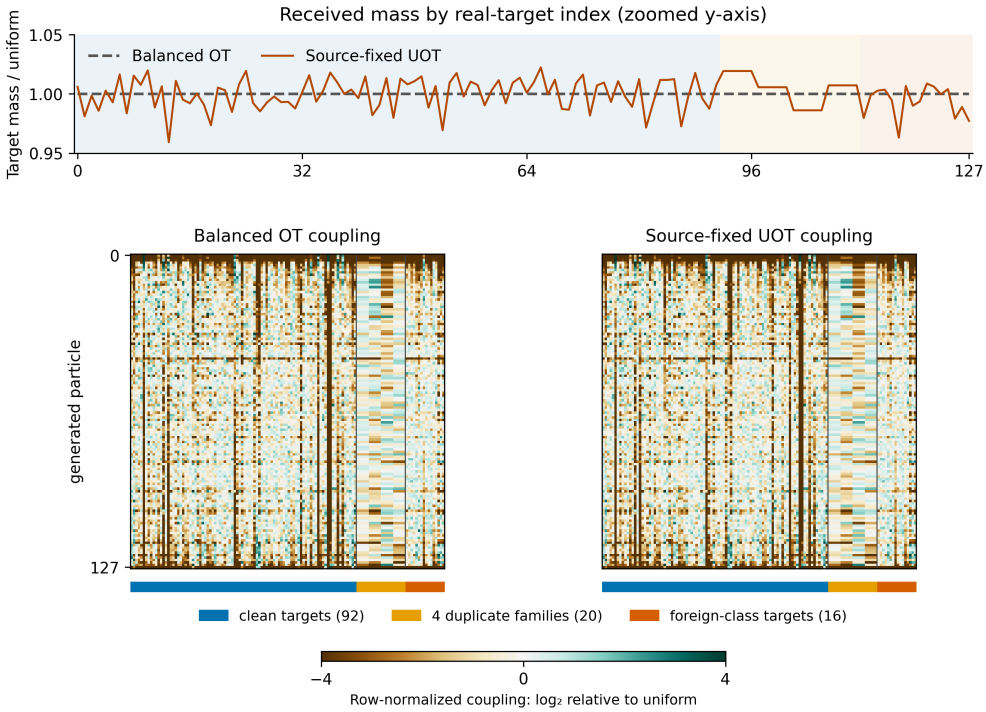


Figure 9: Controlled comparison of balanced OT and source-fixed UOT on identical frozen XL sources and real ImageNet targets. Top: received target mass relative to the uniform reference. Bottom: row-normalized positive-forward couplings; the colored strips distinguish 92 clean targets, 20 members of four duplicate families, and 16 foreign-class targets.

density relative to uniform, with target columns grouped into clean images, duplicate families, and foreign-class images. This controlled probe visualizes marginal adaptation, rather than comparing separately trained generators or establishing a difference in image quality.

Representation-space updates and couplings. We additionally visualize a frozen-checkpoint replay of the source-fixed UOT training update for ImageNet classes 178 and 208 (Figs. 10 and 11). The shared PCA projection shows generated particles, reference features, and the particles after one Euler update; arrows indicate the projected update direction. The accompanying heatmaps display the coupling components used to construct that update. Transport and velocity are computed in the full 4096-dimensional global latent space; PCA is used only for visualization. These figures are diagnostics of the update mechanism, not balanced-OT comparisons or measures of generation quality.

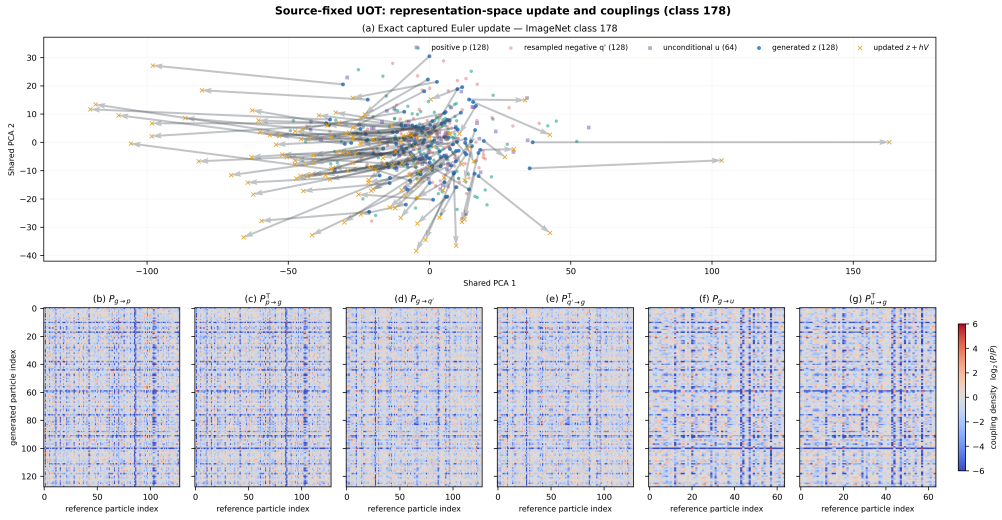


Figure 10: Source-fixed UOT update for ImageNet class 178 using the frozen XL checkpoint. (a) Shared PCA projection of the latent particles and one Euler update; arrows connect generated particles to their updated positions. (b–g) Coupling components used to form the update. Computation uses 4096-dimensional global latents, $\epsilon_{\text{eff}} = 204.8$, $\tau = 0.985$, and 10 Sinkhorn iterations.

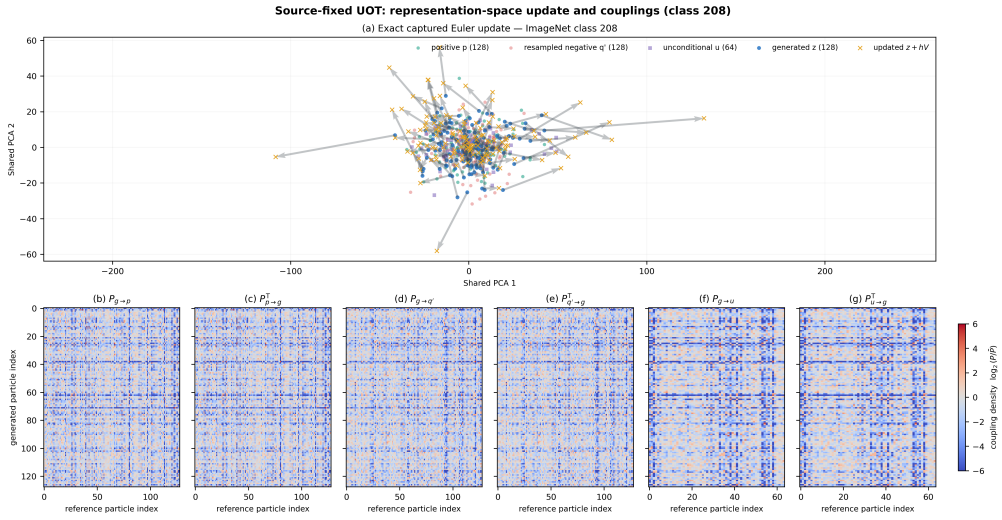
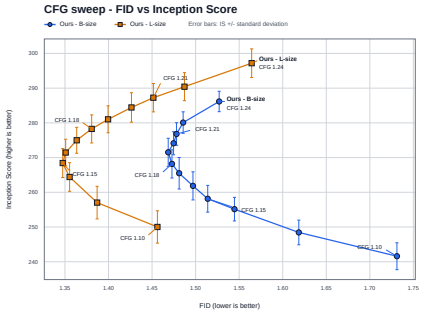


Figure 11: The same frozen-checkpoint source-fixed UOT probe for ImageNet class 208. (a) Shared PCA projection and one Euler update. (b–g) Corresponding coupling components. The PCA coordinates are for display only; all transport computations use the full latent representation.

I Limitations and Broader Impact

Limitations. Our analysis establishes force identities, kinetic lifting, and stationary-state examples under the stated assumptions. Positive definiteness is known for symmetric generators under the conditions of Séjourné et al. (2019), but remains open for the source-fixed energy, as do the trajectory assumptions required for convergence. The convergence statements concern the population flow; the mini-batch update with finitely many Sinkhorn iterations is the next object to analyze. Empirically, why the mild forward-plan reweighting observed on ImageNet (Figure 4) improves generation remains open, and the controlled comparison with balanced transport is at B/2; L/2 and XL/2 are system-level results with a fixed transport recipe. Experiments are on ImageNet-256 in the SD-VAE latent space with DiT generators; other datasets and latent spaces remain to be explored.



CFG scales: 1.10–1.17

Model	Metric	1.10	1.12	1.14	1.15	1.16	1.17
B/2	FID ↓	1.73	1.61	1.54	1.51	1.49	1.48
	IS ↑	241.59±3.87	248.44±3.56	255.14±3.40	258.14±3.85	261.85±4.05	265.49±4.54
L/2	FID ↓	1.45	1.38	1.35	**1.34**	1.35	1.36
	IS ↑	250.01±4.66	257.03±4.69	264.41±4.12	268.43±4.18	271.45±3.83	274.98±3.74

CFG scales: 1.18–1.24

Model	Metric	1.18	1.19	1.20	1.21	1.22	1.24
B/2	FID ↓	1.47	**1.46**	1.47	1.47	1.48	1.52
	IS ↑	268.18±4.06	271.52±4.05	274.13±3.32	276.79±3.25	280.09±3.09	286.16±2.94
L/2	FID ↓	1.38	1.39	1.42	1.45	1.48	1.56
	IS↑	278.26±4.05	281.02±3.89	284.44±4.24	287.25±4.09	290.42±4.04	297.19±4.13

Figure 12: CFG-scale analysis for our B/2 and L/2 models. The left panel shows the FID–IS trade-off, while the right panel reports the corresponding numerical results.

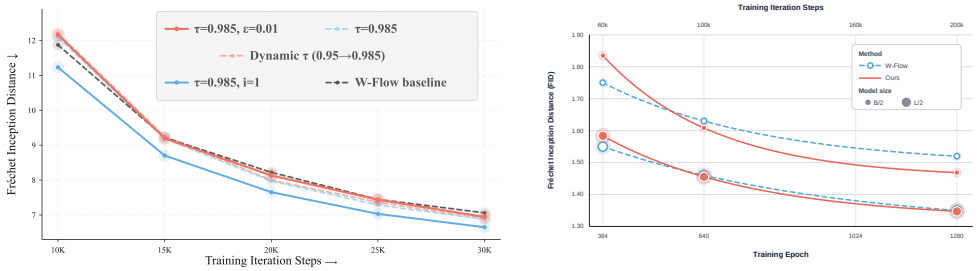


Figure 13: Training dynamics and FID convergence. (Left) Comparison of our ablation variants with baseline methods. (Right) Comparison of our B/2 and L/2 models with baseline methods.



Figure 14: Selected 1-NFE samples from UOT-GF-XL/2 trained from scratch on ImageNet-256 (guidance scale 1.14).

Broader impact. One-step generation produces an image in a single forward pass, making high-quality synthesis cheaper to serve, easier to embed in interactive settings, and less energy-intensive at inference than diffusion and flow models. It carries the familiar hazards of generative imagery: fabricated or misleading content, reproduction of training-data stereotypes, and likenesses generated



(a) House finch (Class 012)

(b) Tree frog (Class 031)



(c) Bald eagle (Class 022)



(d) Agaric (Class 992)



(e) Palace (Class 698)



(f) Monarch butterfly (Class 323)



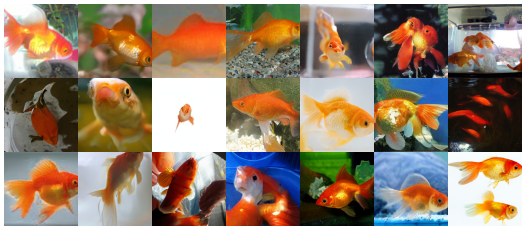
(g) Trifle (Class 927)



(h) Daisy (Class 985)



(i) Lakeside (Class 975)



(j) Goldfish (Class 001)

Figure 15: Random class-conditional ImageNet samples generated by our UOT-XL model at 200K training steps with CFG $= 1.14$. Class indices follow the zero-based ImageNet-1K label mapping.

without consent. Our method adds one consideration: marginal reweighting may affect the representation of underrepresented samples, and aggregate precision and recall do not establish subgroup coverage, so subgroup-level evaluation remains necessary. We support standard release practices: content provenance or watermarking, documented training data and intended use, and access controls or output filtering outside research contexts.

J Additional generated samples



(a) Flamingo (Class 130)



(b) Streetcar (Class 829)



(c) Indigo bunting (Class 014)



(d) Lorikeet (Class 090)



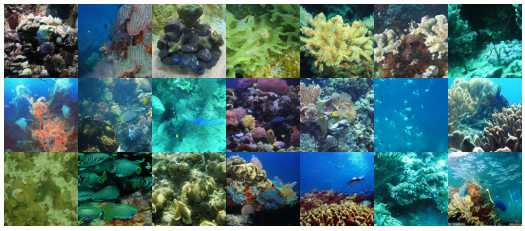
(e) Great grey owl (Class 024)



(f) Peacock (Class 084)



(g) Valley (Class 979)



(h) Coral reef (Class 973)



(i) Alp (Class 970)



(j) Hay (Class 958)



(k) Agama (Class 042)



(l) Megalith (Class 649)

Figure 16: Random class-conditional ImageNet samples generated by our UOT-XL model at 200K training steps with CFG $= 1.14$. Class indices follow the zero-based ImageNet-1K label mapping.



(a) W-Flow-B (balanced OT)



(b) Ours-B (source-fixed UOT)

Figure 17: Qualitative comparison for lion (ImageNet Class 291): official W-Flow-B (Han et al., 2026) (top) and our UOT-B (bottom). Both models use 200K-step EMA checkpoints, CFG $= 1.19$, and sampling seed $= 42$. Corresponding positions use the same initial noise and class label. All 21 samples per model from this seed are shown without filtering.



(a) W-Flow-B (balanced OT)



(b) Ours-B (source-fixed UOT)

Figure 18: Qualitative comparison for Pembroke Welsh corgi (ImageNet Class 263): official W-Flow-B (Han et al., 2026) (top) and our UOT-B (bottom). Both models use 200K-step EMA checkpoints, CFG $= 1.19$, and sampling seed $= 42$. Corresponding positions use the same initial noise and class label. All 21 samples per model from this seed are shown without filtering.



(a) W-Flow-B (balanced OT)



(b) Ours-B (source-fixed UOT)

Figure 19: Qualitative comparison for king penguin (ImageNet Class 145): official W-Flow-B (Han et al., 2026) (top) and our UOT-B (bottom). Both models use 200K-step EMA checkpoints, CFG $= 1.19$, and sampling seed $= 42$. Corresponding positions use the same initial noise and class label. All 21 samples per model from this seed are shown without filtering.



(a) CFG = 2.0



(b) CFG = 3.0

Figure 20: Effect of classifier-free guidance scale on samples from our XL model. Both panels show sports cars (ImageNet Class 817), generated using the same 200K-step EMA checkpoint and sampling seed = 0. Each panel contains 21 samples arranged in a 3×7 grid; corresponding positions use the same initial noise.