UNBALANCED W-FLOW: UNBALANCED WASSERSTEIN DRIFTING UNDER FINITE-BATCH MARGINAL FLUCTUATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Teacher-free one-step generative models such as W-Flow learn directly from distributional objectives, with the Sinkhorn divergence providing the strongest empirical performance among the energies studied in W-Flow. However, Sinkhorn divergence is based on balanced optimal transport, whose exact marginal constraints can induce undesirable transport under finite minibatch fluctuations. We introduce **UW-Flow**, which replaces the balanced Sinkhorn energy with a debiased unbalanced Sinkhorn divergence and derives the corresponding Wasserstein velocity. The resulting dynamics retain W-Flow's cross-minus-self structure, recover the balanced limit, and preserve target identifiability under suitable assumptions. We further characterize the finite-batch velocity error of balanced and unbalanced transport by comparing each minibatch field with its population counterpart. Our analysis identifies conditions under which marginal relaxation yields smaller minibatch-to-population velocity error, and shows that for separated matched modes, balanced sensitivity grows with inter-mode separation while the corresponding UOT (Unbalanced Optimal Transport) sensitivity remains controlled by local transport geometry. Controlled experiments on fixed-support and continuous two-mode distributions support these predictions. On MNIST, CIFAR-10, and ImageNet-100, UW-Flow improves generation under matched training budgets, with the largest gains in unconditional generation and consistent improvements in the class-conditional setting.

1 INTRODUCTION

Recent progress in teacher-free one-step generative modeling has shown that high-quality generators can be learned directly from distributional objectives (Zhou et al., 2025; Deng et al., 2026; Han et al., 2026). W-Flow (Han et al., 2026) is a representative example that defines a Wasserstein gradient flow from a distributional energy and trains a one-step generator to follow the resulting particle updates. Among the distributional energies studied in W-Flow, the Sinkhorn divergence (Genevay et al., 2018; Feydy et al., 2019) achieves the strongest empirical performance.

The Sinkhorn divergence used by W-Flow is built on *balanced* optimal transport (Cuturi, 2013), whose coupling must exactly reproduce both input marginals. In practice, however, the corresponding velocity field is estimated from finite minibatches (Zhu et al., 2024; Han et al., 2026), whose empirical composition fluctuates from one update to another. Exact marginal matching can therefore force transport between geometrically unrelated regions when, for example, the sampled proportions of different modes do not agree (Fatras et al., 2021; Nguyen et al., 2022). Unbalanced optimal transport (UOT) relaxes these hard marginal constraints through soft penalties (Liero et al., 2018; Chizat et al., 2018), allowing the transported marginals to deviate when exact matching would require undesirable transport. Prior work has shown that such marginal relax-

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.

ation can improve robustness to minibatch composition mismatch (Fatras et al., 2021; Nguyen et al., 2022) and outliers (Balaji et al., 2020).

Motivated by this, we introduce **UW-Flow**, which uses the debiased unbalanced Sinkhorn divergence (Séjourné et al., 2019) as its distributional energy, and derive the resulting Wasserstein velocity. In contrast to prior uses of UOT for minibatch objectives (Fatras et al., 2021), transport maps or couplings (Yang & Uhler, 2018; Choi et al., 2023; Eyring et al., 2024), and scalable JKO schemes (Choi et al., 2024), our formulation uses the debiased unbalanced Sinkhorn divergence as the population energy that directly defines W-Flow's one-step velocity. In W-Flow, the transport objective directly determines the velocity field used for one-step learning, and thus replacing the balanced Sinkhorn energy changes the underlying population dynamics rather than merely modifying a coupling within a fixed objective. We show that UW-Flow retains the cross-minus-self structure of W-Flow, recovers the balanced dynamics in the hard-marginal limit (Séjourné et al., 2019), and preserves the desired target-identification property under suitable assumptions. Marginal relaxation therefore defines alternative Wasserstein dynamics while retaining the key structural properties required for distribution matching.

We next characterize how finite-batch sampling perturbs the velocity fields induced by balanced and unbalanced transport. Specifically, we measure how far each minibatch velocity deviates from its own population counterpart. For balanced Sinkhorn transport, exact marginal preservation imposes an unavoidable finite-batch velocity error. For UW-Flow, we characterize the leading sensitivity to empirical marginal perturbations and identify conditions under which the UOT minibatch-to-population velocity error is smaller than the corresponding balanced error. In particular, for separated matched modes, the balanced sensitivity grows with inter-mode separation, whereas the corresponding UOT sensitivity remains controlled by the local transport geometry. We further test these predictions on fixed-support and continuous two-mode toy distributions by varying batch size, mode separation, and marginal relaxation.

We evaluate UW-Flow on MNIST (LeCun et al., 1998), CIFAR-10 (Krizhevsky et al., 2009), and ImageNet-100 (Deng et al., 2009; Tian et al., 2020) under matched training budgets. The gains are largest in unconditional generation, where multimodal composition mismatch is more substantial, while class-conditional generation also benefits from marginal relaxation.

2 Related Work

Wasserstein flows for one-step generation. Wasserstein gradient flows provide generative models with particle velocities induced by a distributional energy. NSGF (Zhu et al., 2024) studies flows driven by the Sinkhorn divergence, while W-Flow (Han et al., 2026) trains a one-step generator to regress onto the resulting particle updates. Optimal transport (OT) has also been used as a training loss (Genevay et al., 2018), to construct generators (Rout et al., 2021), and to pair samples for flow matching (Tong et al., 2024; Pooladian et al., 2023). In W-Flow, the velocity is estimated from balanced minibatch couplings rather than population distributions. Because these couplings must match the sampled marginals exactly, fluctuations in regional sample proportions can alter the transport plan (Fatras et al., 2019; 2021; Nguyen et al., 2021; 2022) and perturb the velocity used as the regression target.

Relaxed transport. Unbalanced transport relaxes exact marginal matching and has been used to improve the robustness of minibatch couplings (Séjourné et al., 2019; Fatras et al., 2021). Related approaches use partial transport to limit mismatched minibatch pairings (Nguyen et al., 2022), learn maps with reweighted marginals (Choi et al., 2023), incorporate UOT into flow matching (Eyring et al., 2024), or formulate scalable JKO-based flows (Choi et al., 2024). We incorporate an unbalanced Sinkhorn divergence into the W-Flow (Han et al., 2026) framework while retaining the target distribution as an equilibrium of the induced velocity field.

Statistical analysis of transport estimates. Prior analyses study empirical entropic OT maps (Rigollet & Stromme, 2025; Goldfeld et al., 2024) and minibatch OT objectives (Fatras et al., 2019; 2021). Fatras et al. (2021) also demonstrate minibatch UOT particle flows, but their concentration and gradient results concern the expected minibatch loss. To our knowledge, none directly compares the minibatch-to-population error of the cross-minus-self velocity under balanced and unbalanced marginal constraints. We make this comparison and identify a separated-mode setting in which the leading unbalanced error is smaller.

3 PRELIMINARIES

We first review Wasserstein gradient flows as a framework for defining distributional dynamics and particle updates (Sec. 3.1). We then describe how W-Flow instantiates this framework with the balanced Sinkhorn divergence and learns the resulting updates with a one-step generator (Sec. 3.2).

3.1 WASSERSTEIN DYNAMICS INDUCED BY DISTRIBUTIONAL ENERGIES

Let $p \in \mathcal{P}_2(\mathbb{R}^d)$ be a target distribution and $q \in \mathcal{P}_2(\mathbb{R}^d)$ an evolving distribution, where $\mathcal{P}_2(\mathbb{R}^d)$ denotes the space of probability measures with finite second moment. Given a differentiable distributional energy $\mathcal{F}(q; p)$ measuring the discrepancy between q and p, its formal 2-Wasserstein gradient flow (Ambrosio et al., 2005) satisfies

$$\partial_t q_t = \nabla \cdot \left(q_t \nabla \frac{\delta \mathcal{F}}{\delta q}(q_t; p) \right). \tag{1}$$

Equivalently, this evolution satisfies the continuity equation

$$\partial_t q_t + \nabla \cdot (q_t V_t) = 0, \qquad V_t(x) = -\nabla_x \frac{\delta \mathcal{F}}{\delta q}(q_t; p)(x). \tag{2}$$

The choice of energy thus determines the velocity field through which the distribution evolves. Under suitable regularity conditions, the same dynamics can be described at the particle level. A particle initialized at $X_0 = x$ follows

$$\frac{dX_t}{dt} = V_{q_t,p}(X_t), \qquad q_t = (\Phi_t)_{\#} q_0, \tag{3}$$

where $\Phi_t(x) = X_t$ denotes the resulting flow map. When a distribution is represented by a generator, its evolution can be learned by fitting the generator to particle displacements prescribed by this flow. We next describe the energy and velocity used by W-Flow for this purpose.

3.2 SINKHORN-INDUCED DYNAMICS AND ONE-STEP LEARNING

W-Flow (Han et al., 2026) instantiates the Wasserstein dynamics in Sec. 3.1 using the balanced Sinkhorn divergence as its distributional energy. This choice provides a transport-based velocity that can be computed from samples through entropic optimal transport. Although choosing this energy alone does not guarantee convergence of the gradient flow to the target, W-Flow demonstrates that the resulting dynamics can guide one-step generative modeling at scale.

For $q, p \in \mathcal{P}_2(\mathbb{R}^d)$ and the quadratic ground cost $c(x, y) = \frac{1}{2}\|x - y\|_2^2$, the entropically regularized optimal transport cost is

$$\mathrm{OT}_\varepsilon(q, p) = \inf_{\pi \in \Pi(q,p)} \left\{ \int c(x, y)\, d\pi(x, y) + \varepsilon\, \mathrm{KL}(\pi \,\|\, q \otimes p) \right\}, \tag{4}$$

where $\Pi(q, p)$ is the set of couplings with marginals q and p. Entropic regularization enables computation through Sinkhorn iterations (Cuturi, 2013), but generally gives a nonzero self-cost, $\mathrm{OT}_\varepsilon(p, p) \neq 0$. The Sinkhorn divergence (Feydy et al., 2019) removes this self-cost:

$$\mathcal{S}_\varepsilon(q, p) = \mathrm{OT}_\varepsilon(q, p) - \frac{1}{2}\mathrm{OT}_\varepsilon(q, q) - \frac{1}{2}\mathrm{OT}_\varepsilon(p, p). \tag{5}$$

Let $\pi_{q,p}$ and $\pi_{q,q}$ be optimal couplings for the cross and self transport problems, respectively, and define their barycentric projections as $T_{q,p}(x) = \int y\, \pi_{q,p}(dy \mid x)$ and $T_{q,q}(x) = \int x'\, \pi_{q,q}(dx' \mid x)$. Suppressing the dependence on ε in this notation, the Wasserstein gradient of $\mathcal{S}_\varepsilon(q, p)$ induces the velocity

$$V_{q,p}(x) = T_{q,p}(x) - T_{q,q}(x). \tag{6}$$

At $q = p$, the cross and self terms coincide, so $V_{p,p} = 0$. Han et al. (2026) approximate the continuous particle dynamics in Eq. 3 with an explicit Euler step. At a current distribution q, this gives

$$g_{q,p}(x) = x + \eta V_{q,p}(x), \qquad q^+ = (g_{q,p})_{\#} q, \tag{7}$$

where $\eta > 0$ is the step size. To learn these updates with a one-step generator, let p_{ref} be a reference distribution and $f_{\bar\theta}$ the current generator, inducing $q_{\bar\theta} = (f_{\bar\theta})_{\#} p_{\mathrm{ref}}$. With $\bar\theta$ fixed when constructing the target, Han et al. (2026) minimize

$$\mathcal{L}_{\text{W-Flow}}(\theta) = \mathbb{E}_{z\sim p_{\mathrm{ref}}}\left[\left\| f_\theta(z) - \mathrm{sg}\big[f_{\bar\theta}(z) + \eta V_{q_{\bar\theta},p}\big(f_{\bar\theta}(z)\big)\big]\right\|_2^2\right], \tag{8}$$

where sg denotes stop-gradient. The generator learns the particle updates during training and produces samples in a single forward pass at inference. In practice, the cross and self barycentric terms are estimated from finite minibatches; we examine how this estimation perturbs the population velocity field in Sec. 4.

4 UW-FLOW

The balanced Sinkhorn field in Eq. 6 is defined at the population level, while in practice it is estimated from finite mini-batches (Han et al., 2026). We first examine how empirical marginal fluctuations distort the balanced transport geometry (Sec. 4.1). Motivated by this, we introduce **UW-Flow**, which relaxes the hard marginal constraints of balanced OT using unbalanced transport penalties (Sec. 4.2). We then analyze how this relaxation affects finite-batch stability relative to balanced OT (Sec. 4.3).

4.1 FINITE-BATCH VELOCITY ERROR IN BALANCED SINKHORN FIELDS

Balanced transport must match empirical marginals exactly, so sampling fluctuations can perturb the barycentric field. To examine this effect, draw three mutually independent batches of B i.i.d. samples, $x_1, \dots, x_B \sim q$, $y_1, \dots, y_B \sim p$, and $x'_1, \dots, x'_B \sim q$, and define

$$\widehat{q}_B = \frac{1}{B}\sum_{i=1}^{B} \delta_{x_i}, \qquad \widehat{p}_B = \frac{1}{B}\sum_{j=1}^{B} \delta_{y_j}, \qquad \widehat{q}'_B = \frac{1}{B}\sum_{i=1}^{B} \delta_{x'_i}. \tag{9}$$

The regression targets in Eq. 8 use $\widehat{V}_B := T_{\widehat{q}_B,\widehat{p}_B} - T_{\widehat{q}_B,\widehat{q}'_B}$. While prior work studies finite-sample convergence of entropic OT maps (Rigollet & Stromme, 2025; Goldfeld et al., 2024), we measure the error of this cross-minus-self field at the sampled source points:

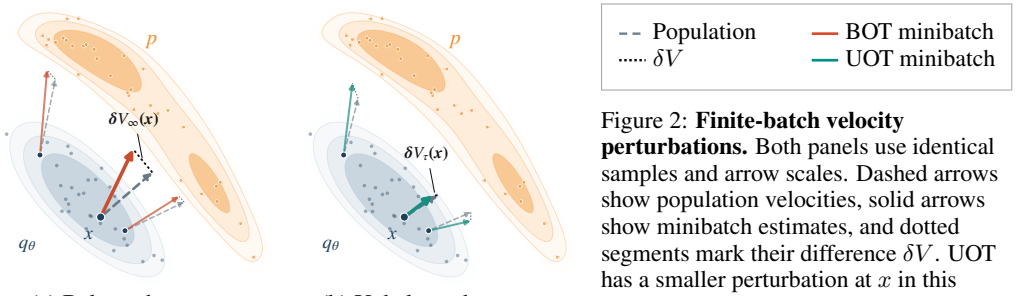
$$\mathcal{E}_{\mathrm{BOT}}(B) := \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B}\left\|\widehat{V}_B(x_i) - V_{q,p}(x_i)\right\|_2^2\right]. \tag{10}$$

The following proposition gives a lower bound on finite-batch velocity error under exact marginal matching.

Proposition 1 (Balanced finite-batch velocity error)**.** *Assume the cross and self couplings are balanced and the following covariances are finite. For every B,*

$$\mathcal{E}_{\mathrm{BOT}}(B) \geq \frac{L^2_{\mathrm{BOT}}}{B}, \qquad L^2_{\mathrm{BOT}} := \mathrm{Tr}\big[\mathrm{Cov}_p(Y) + \mathrm{Cov}_q(X) + \mathrm{Cov}_q\big(V_{q,p}(X)\big)\big]. \tag{11}$$

The proof is in Appendix B. Thus, even when the balanced OT field is computed exactly for each minibatch, fluctuations in the empirical means contribute an unavoidable variance component to the velocity error.



(a) Balanced transport (b) Unbalanced transport

Figure 2: **Finite-batch velocity perturbations.** Both panels use identical samples and arrow scales. Dashed arrows show population velocities, solid arrows show minibatch estimates, and dotted segments mark their difference δV. UOT has a smaller perturbation at x in this illustration.

4.2 UW-Flow: Relaxing Marginal Constraints

The finite-batch error in Sec. 4.1 motivates relaxing empirical marginal constraints, allowing transported mass to adapt to sampling fluctuations and potentially reducing velocity perturbations. We therefore use unbalanced optimal transport (UOT) (Chizat et al., 2018), which replaces hard marginal constraints with soft penalties. For $\tau > 0$, we use

$$\mathrm{UOT}_{\varepsilon,\tau}(q,p) = \inf_{\pi \geq 0} \left\{ \int c(x,y)\, d\pi(x,y) + \tau D_{\mathrm{KL}}(\pi_1 \| q) + \tau D_{\mathrm{KL}}(\pi_2 \| p) + \varepsilon D_{\mathrm{KL}}(\pi \,\|\, q \otimes p) \right\}, \quad (12)$$

where π_1 and π_2 are marginals of π, and D_{KL} denotes the generalized KL divergence. Larger τ penalizes deviations from q and p, while smaller τ allows trade-offs between marginal mismatch and transport cost. Following Eq. 5, we use the debiased unbalanced Sinkhorn divergence (Séjourné et al., 2019),

$$\mathcal{S}^{\mathrm{U}}_{\varepsilon,\tau}(q,p) = \mathrm{UOT}_{\varepsilon,\tau}(q,p) - \frac{1}{2}\mathrm{UOT}_{\varepsilon,\tau}(q,q) - \frac{1}{2}\mathrm{UOT}_{\varepsilon,\tau}(p,p), \quad (13)$$

as our distributional energy, $\mathcal{F}_\tau(q;p) := \mathcal{S}^{\mathrm{U}}_{\varepsilon,\tau}(q,p)$. The mass-correction term in the general unbalanced Sinkhorn divergence vanishes here because q and p both have unit mass.

Let $\pi^{\varepsilon,\tau}_{q,\mu}$ be the optimal UOT coupling between q and μ, where $\mu \in \{p, q\}$. Unlike a balanced coupling, its first marginal need not equal q. Writing $\pi^{\varepsilon,\tau}_{q,\mu}(dx, dy) = q(dx)\, \pi^{\varepsilon,\tau}_{q,\mu}(dy \mid x)$, define the local transported-mass ratio and normalized barycentric projection by

$$s^\tau_{q,\mu}(x) := \frac{d(\pi^{\varepsilon,\tau}_{q,\mu})_1}{dq}(x), \qquad T^\tau_{q,\mu}(x) := \frac{\int y\, \pi^{\varepsilon,\tau}_{q,\mu}(dy \mid x)}{s^\tau_{q,\mu}(x)}. \quad (14)$$

Here, $\pi^{\varepsilon,\tau}_{q,\mu}(dy \mid x)$ is a possibly unnormalized measure, and the barycenter is defined wherever $s^\tau_{q,\mu}(x) > 0$. The Wasserstein gradient of $\mathcal{F}_\tau$ yields

$$V^\tau_{q,p}(x) = s^\tau_{q,p}(x)\big(T^\tau_{q,p}(x) - x\big) - s^\tau_{q,q}(x)\big(T^\tau_{q,q}(x) - x\big). \quad (15)$$

The derivation from the UOT dual and first variation is in Appendix A.1. This field retains the cross-minus-self structure in Eq. 6, with each displacement weighted by transported mass. Under the assumptions stated in Appendix A.2, it satisfies target identifiability: $V^\tau_{q,p} = 0$ q-almost everywhere if and only if $q = p$. The proof appears there. Together with marginal relaxation, this property motivates the hypothesis that UOT may improve one-step training when minibatch-induced marginal mismatch perturbs the balanced velocity field. As $\tau \to \infty$, $\mathcal{S}^{\mathrm{U}}_{\varepsilon,\tau}(q,p) \to \mathcal{S}_\varepsilon(q,p)$ and $V^\tau_{q,p} \to V_{q,p}$, recovering balanced W-Flow (Han et al., 2026).

4.3 Finite-Batch Velocity Error under Marginal Relaxation

We now compare the finite-batch estimation errors of balanced and unbalanced velocity fields. To this end, define the mass-weighted transport term $U^\tau_{q,\mu}(x) := s^\tau_{q,\mu}(x)\big(T^\tau_{q,\mu}(x) - x\big)$. The population field in Eq. 15 is then $V^\tau_{q,p} = U^\tau_{q,p} - U^\tau_{q,q}$, and the empirical field is $\widehat{V}^\tau_B := U^\tau_{\widehat{q}_B,\widehat{p}_B} - U^\tau_{\widehat{q}_B,\widehat{q}'_B}$.

We measure its finite-batch error relative to the population field $V^{\tau}_{q,p}$ as

$$\mathcal{E}_\tau(B) := \mathbb{E}\left[\frac{1}{B}\sum_{i=1}^{B}\left\|\widehat{V}^{\tau}_{B}(x_i) - V^{\tau}_{q,p}(x_i)\right\|_2^2\right]. \tag{16}$$

Consistent with the hard-marginal limit $\tau \to \infty$, we write the balanced quantities from Sec. 4.1 as $\mathcal{E}_\infty(B) := \mathcal{E}_{\mathrm{BOT}}(B)$ and $L_\infty := L_{\mathrm{BOT}}$. Note that each error is measured relative to its own population field.

Finite-batch sensitivity of UOT. For the first-order analysis, we assume fixed finite supports with strictly positive population masses, so minibatch sampling induces multinomial fluctuations in the empirical masses. Unlike balanced OT, UOT allows the transported-mass ratio s_i to vary. We analyze how these mass fluctuations affect both s_i and the barycentric destination T_i.

Proposition 2 (Finite-batch error of UW-Flow)**.** *Under the fixed-support assumptions above, for any fixed $0 < \tau < \infty$, as $B \to \infty$,*

$$\mathcal{E}_\tau(B) = \frac{A_\tau}{B} + O(B^{-3/2}). \tag{17}$$

The $1/B$ factor reflects the scale of finite-batch sampling noise, while $A_\tau < \infty$ is a B-independent coefficient that quantifies the leading sensitivity of the UOT velocity field to empirical marginal perturbations. The derivation is deferred to Appendix C. Combining this expansion with the balanced lower bound from Sec. 4.1 yields the following comparison theorem.

Theorem 1 (Finite-batch stability comparison)**.** *From the finite-batch error characterizations of UOT in Eq. 17 and balanced OT in Eq. 11, if $A_\tau < L_\infty^2$, then there exists $B_0 < \infty$ such that*

$$A_\tau < L_\infty^2 \quad \Longrightarrow \quad \mathcal{E}_\tau(B) < \mathcal{E}_\infty(B), \qquad \forall B \geq B_0. \tag{18}$$

Under the stated condition, the theorem shows that, for sufficiently large B, the empirical UOT field is closer to its own population field than the balanced field is to its population counterpart. Although this comparison does not rank the population fields or guarantee better generation, Sec. 4.4 establishes the condition in a separated two-mode setting and tests the predicted stability gap with finite-batch experiments. Sec. 5 then reports generation gains in more complex settings.

4.4 Separate-Mode Analysis

Separated mode analysis. To verify the sufficient condition $A_\tau < L_\infty^2$, we consider a separated two-mode setting. The source and target distributions each consist of two matched modes whose within-pair geometry remains uniformly bounded as the two mode pairs become increasingly separated. Let the target modes have masses ω and $1-\omega$, with conditional means μ_1 and μ_2, and define their separation as $d_\mu := \|\mu_1 - \mu_2\|$.

For balanced OT, the target-mode separation directly contributes to the finite-batch error through the target covariance. In particular, the balanced coefficient satisfies

$$L_\infty^2 \geq \omega(1-\omega)d_\mu^2. \tag{19}$$

In contrast, for any fixed finite τ, UOT can absorb mode-count mismatch through marginal relaxation, so that transport between unmatched modes vanishes as the modes separate while the local transport geometry remains bounded. Consequently,

Corollary 1 (Bounded UOT sensitivity under mode separation)**.** *For any fixed $0 < \tau < \infty$, under the separated matched-mode assumptions above,*

$$A_\tau(d_\mu) = O(1), \qquad d_\mu \to \infty. \tag{20}$$

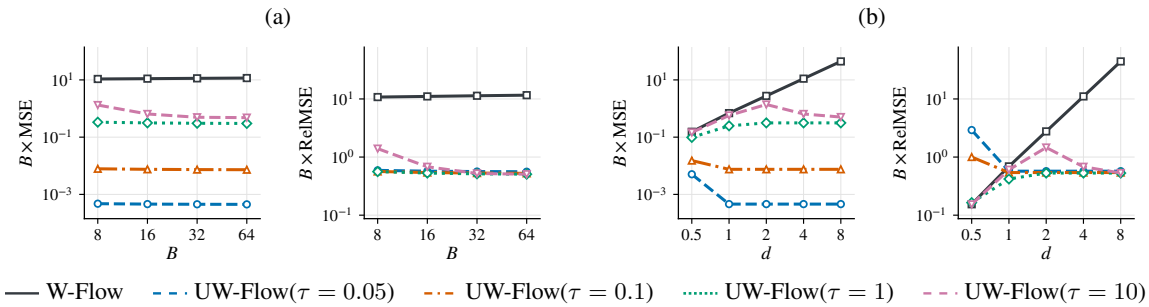


Figure 3: **Minibatch velocity error on fixed supports.** Each method is evaluated against its own population velocity. Panel (a) sweeps B at $d = 4$, while panel (b) sweeps d at $B = 16$. Within each panel, the left plot shows $B \times \mathrm{MSE}$ and the right plot shows $B \times \mathrm{RelMSE}$. RelMSE is MSE divided by the corresponding population-field energy.

A formal derivation is provided in Appendix D. Hence, for sufficiently large d_μ, the condition $A_\tau < L_\infty^2$ holds.

We use two controlled settings to examine (i) whether the finite-batch error follows the predicted $1/B$ sensitivity, (ii) how their error gap varies with mode separation, (iii) whether the observed advantage extends beyond fixed supports. We vary B, d, and τ in a fixed-support two-particle model and a continuously sampled two-Gaussian extension. Experimental details are provided in Appendix E.

Fixed-support perturbations. Fig. 3 shows approximately constant $B \times \mathrm{MSE}$ across the tested batch sizes, consistent with $1/B$ error scaling for both methods. The absolute error gap widens with mode separation, consistent with Eq. 19 and Eq. 20. The $B \times \mathrm{RelMSE}$ curves show that UOT's advantage at larger separations persists after accounting for differences in population-field magnitude. Full results, including cosine similarity, appear in Appendix E.2.

Beyond fixed supports. Appendix E.3 extends the comparison to Gaussian mixtures, where both mode proportions and sample locations fluctuate. UOT generally exhibits lower relative error and higher directional agreement over the tested configurations.

5 Experimental Results

We evaluate UW-Flow on synthetic targets and image-generation benchmarks. Within each comparison, we replace W-Flow's (Han et al., 2026) balanced Sinkhorn energy with its unbalanced counterpart while keeping the generator architecture, optimization protocol, and training budget matched. Each transport group uses $N_{\mathrm{gen}} = N_{\mathrm{pos}} = B$ samples from the generated, target, and self-reference distributions, respectively. In transport-batch ablations, we vary B while adjusting the number of groups N_c to keep the number of generated samples per optimizer update fixed at $B_{\mathrm{opt}} = N_c B$.

5.1 Toy Analysis: Synthetic Distribution Matching

We evaluate distribution matching beyond the controlled two-mode settings of Sec. 4.4. We train W-Flow and UW-Flow on three two-dimensional targets: checkerboard, two moons, and Gaussian ring. Both methods use $B = 128$ and $B_{\mathrm{opt}} = 512$, with $\tau = 0.1$ for UW-Flow. Experimental details and additional ablations over B and τ are provided in Appendix. F.1.

Qualitatively, UW-Flow captures the characteristic structures of all three targets (Fig. 4). The quantitative results in Tab. 1 generally favor UW-Flow over W-Flow, as measured by squared sliced Wasserstein distance (SW^2) and squared maximum mean discrepancy (MMD^2).

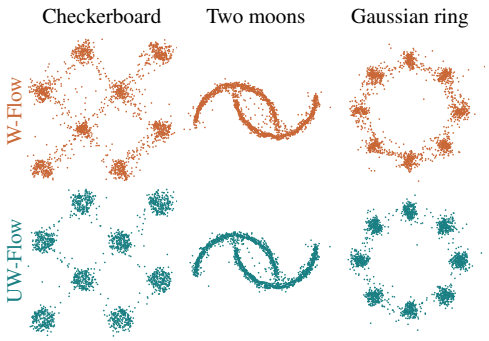


Figure 4: W-Flow (top) and UW-Flow (bottom) across three distributions.

Table 1: Toy distribution matching with W-Flow and UW-Flow. Values are $\times 10^{-3}$; best in bold.

Metric	Method	Distribution		
		Checker-board	Two moons	Gaussian ring
$SW^2 \downarrow$	W-Flow	2.970	**0.825**	4.280
	UW-Flow	**2.600**	0.852	**3.760**
$MMD^2 \downarrow$	W-Flow	1.490	0.483	0.866
	UW-Flow	**0.700**	**0.407**	**0.501**

5.2 Fixed-Budget Minibatch Discrepancy Ablation

We examine whether marginal relaxation improves unconditional generation with limited transport support on MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009). We vary $B \in \{4, 8, 16, 32, 64\}$ while keeping B_{opt} fixed at 64 for MNIST and 128 for CIFAR-10. We report Fréchet Inception Distance (FID) (Heusel et al., 2017), precision, and recall (Kynkäänniemi et al., 2019) on both datasets, and Inception Score (IS) (Salimans et al., 2016) on CIFAR-10. Further experimental details are provided in Appendix F.

Tab. 2 and Tab. 3 show the generation results as the transport support size B varies under a fixed per-update budget. FID decreases with increasing B for both W-Flow and the reported UW-Flow configurations. We evaluate $\tau \in \{0.1, 1, 10\}$ at each support size and report a selected UW-Flow configuration in the main tables; the complete results are provided in Appendix F.2. UW-Flow consistently achieves lower FID than W-Flow at every tested support size on both datasets. At the smallest support sizes, $B = 4$ and $B = 8$, UW-Flow achieves the largest absolute FID reductions over W-Flow, by 15.02 and 14.69 points, respectively. It also maintains a lower FID at $B = 64$ (4.51 versus 8.10). These results show that an appropriate choice of τ can improve one-step generation by relaxing the hard marginal constraints of balanced OT.

Table 2: MNIST results with $B_{\text{opt}} = 64$.

N_c	N_{gen}	N_{pos}	Method	FID ↓	Prec. ↑	Rec. ↑
16	4	4	W-Flow	19.70	**0.54**	0.06
			UW-Flow	**17.98**	0.46	**0.12**
8	8	8	W-Flow	13.69	0.46	0.24
			UW-Flow	**12.37**	**0.48**	**0.28**
4	16	16	W-Flow	9.86	**0.49**	0.43
			UW-Flow	**9.22**	0.47	**0.51**
2	32	32	W-Flow	8.72	0.58	0.32
			UW-Flow	**7.33**	**0.60**	**0.34**
1	64	64	W-Flow	7.03	0.57	0.53
			UW-Flow	**5.67**	**0.59**	**0.57**

Table 3: CIFAR-10 results with $B_{\text{opt}} = 128$.

N_c	N_{gen}	N_{pos}	Method	FID ↓	IS ↑	Prec. ↑	Rec. ↑
32	4	4	W-Flow	30.58	9.75	**0.55**	0.01
			UW-Flow	**15.56**	**10.60**	0.52	**0.07**
16	8	8	W-Flow	24.39	8.99	**0.58**	0.01
			UW-Flow	**9.70**	**10.53**	0.56	**0.17**
8	16	16	W-Flow	12.49	**10.09**	0.59	0.07
			UW-Flow	**10.03**	9.87	**0.61**	**0.15**
4	32	32	W-Flow	9.57	10.32	**0.63**	0.08
			UW-Flow	**5.01**	**10.43**	0.61	**0.39**
2	64	64	W-Flow	8.10	**10.37**	**0.60**	0.27
			UW-Flow	**4.51**	**10.37**	**0.60**	**0.45**

Fig. 5 compares FID trajectories across three transport support sizes. UW-Flow with $\tau = 0.1$ or $\tau = 1$ generally reaches lower FID in fewer training epochs than W-Flow. At $B = 4$, UW-Flow with $\tau = 0.1$ reaches the FID attained by W-Flow at epoch 500 after approximately 55 epochs, requiring about $9.1\times$ fewer epochs; at $B = 64$, it reaches the corresponding FID after 80 epochs, requiring $6.25\times$ fewer epochs. As τ increases to 10, the trajectories generally move closer to that of W-Flow.

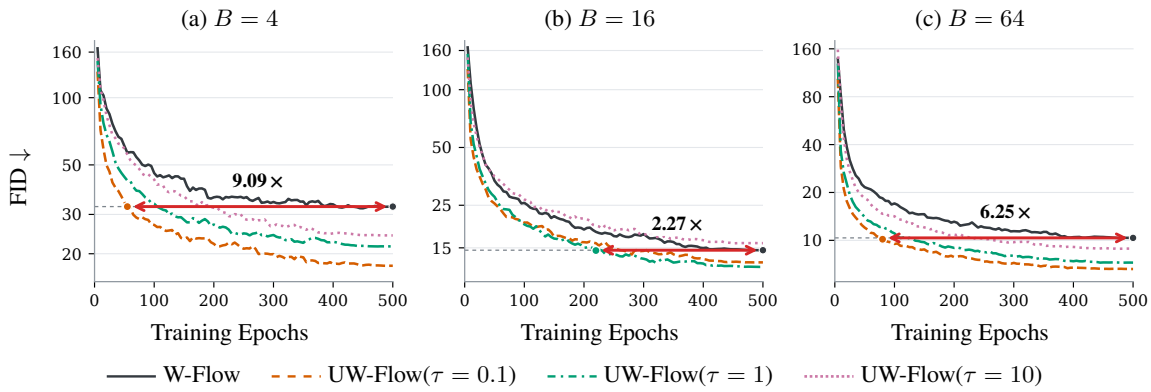


Figure 5: **CIFAR-10 training dynamics across transport support sizes.** FID versus training epoch for W-Flow and UW-Flow with $\tau \in \{0.1, 1, 10\}$. FID is computed using 10,000 generated samples at each evaluation. Panels show $B = 4$, 16, and 64, with the sample budget per optimizer update fixed at $B_{\text{opt}} = N_c B = 128$.

5.3 Scaling to Complex Visual Distributions

To evaluate UW-Flow on a more complex target distribution, we use ImageNet-100, a 100-class subset of ImageNet (Deng et al., 2009), and test both unconditional and class-conditional generation. We compare W-Flow (Han et al., 2026) and UW-Flow under matched generator architectures, feature encoders, sampling configurations, and optimization budgets. We report FID (Heusel et al., 2017), IS (Salimans et al., 2016), precision, and recall (Kynkäänniemi et al., 2019). The experimental setup and results for all tested values of τ are provided in Appendix F.4.

Table 4: ImageNet-100 results at matched budgets.

Setting	Method	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Uncond	W-Flow	55.25	23.89	**0.42**	0.14
Uncond	UW-Flow	**46.17**	**27.71**	0.40	**0.29**
Cond	W-Flow	9.43	60.31	0.74	0.26
Cond	UW-Flow	**9.11**	**61.34**	**0.75**	**0.26**

Tab. 4 reports the results for unconditional and class-conditional ImageNet-100 generation. We evaluate $\tau \in \{0.1, 1, 10\}$ and report the configuration with the lowest FID for each setting. For unconditional generation, UW-Flow with $\tau = 0.1$ reduces FID from 55.25 to 46.17 (16.4%), with higher IS and recall but lower precision. For class-conditional generation, $\tau = 1$ reduces FID from 9.43 to 9.11 (3.4%) and improves IS and precision, while recall remains unchanged at the reported precision. The smaller conditional gain is consistent with class-wise transport reducing opportunities for between-class mismatch, although these results do not directly measure coupling mismatch. Full results for all tested values of τ are provided in Appendix F.4.

6 Conclusion

We introduced UW-Flow to reduce the sensitivity of W-Flow's velocity field to finite-batch marginal fluctuations. Our analysis characterizes the minibatch-to-population error and identifies a separated-mode setting in which marginal relaxation improves finite-batch stability. Controlled experiments support this analysis, while results on MNIST, CIFAR-10, and ImageNet-100 show gains in generation quality under finite transport support.

Limitations. Our stability comparison measures each minibatch field against its own population field. It does not compare the population velocities of UOT and balanced transport or guarantee better generation for arbitrary distributions. We also do not provide a method for selecting τ for a given setting.

AI USE STATEMENT

Generative AI tools were used in this work only for two supporting tasks: revising the manuscript for clarity and readability, and proofreading the mathematical derivations for typographical and notational errors. The research itself was carried out by the authors. This includes the idea of using the debiased unbalanced Sinkhorn divergence as the energy for W-Flow, the derivation of the UW-Flow velocity and its identifiability property, the finite-batch error analysis of balanced and unbalanced transport. The authors were also solely responsible for designing the experiments, implementing the method, selecting datasets, baselines, and metrics, interpreting the results, and surveying and positioning the related literature. AI tools were not used for any of these tasks.

Every AI-suggested edit was checked by the authors, and any change affecting a mathematical expression was verified against the original derivation before being adopted. The authors take full responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available datasets (MNIST, CIFAR-10, and ImageNet-100) and synthetic distributions, and it does not involve human-subject experiments or the collection of personally identifiable information. As with other generative models, models trained with our method may inherit biases present in the training data and could be misused to produce misleading synthetic images. We do not identify ethical concerns beyond those generally associated with generative modeling.

REPRODUCIBILITY STATEMENT

All theoretical results are stated with their assumptions in the main text, and complete proofs are provided in Appendix. Also describes the setup of the velocity-error experiments, and Appendix F provides the full training configurations, hyperparameters, and complete results for the generation experiments. All datasets used in this work are publicly available, and we will release our code upon acceptance.

REFERENCES

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

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of computation*, 87(314):2563–2609, 2018.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. In *NeurIPS*, 2023.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. In *ICML*, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

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

Luca Eyring, Dominik Klein, Théo Uscidda, Giovanni Palla, Niki Kilbertus, Zeynep Akata, and Fabian Theis. Unbalancedness in neural monge maps improves unpaired domain translation. In *ICLR*, 2024.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Kilian Fatras, Thibault Séjourné, Rémi Flamary, and Nicolas Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In *ICML*, 2021.

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 *AISTATS*, 2019.

Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International conference on artificial intelligence and statistics*, pp. 1608–1617. PMLR, 2018.

Ziv Goldfeld, Kengo Kato, Gabriel Rioux, and Ritwik Sadhu. Limit theorems for entropic optimal transport maps and sinkhorn divergence. *Electronic Journal of Statistics*, 2024.

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.

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

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 *NeurIPS*, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 2019.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Khai Nguyen, Dang Nguyen, Quoc Nguyen, Tung Pham, Hung Bui, Dinh Phung, Trung Le, and Nhat Ho. On transportation of mini-batches: A hierarchical approach. *arXiv preprint arXiv:2102.05912*, 2021.

Khai Nguyen, Dang Nguyen, The-Anh Vu-Le, Tung Pham, and Nhat Ho. Improving mini-batch optimal transport via partial transportation. In *ICML*, 2022.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *ICML*, 2023.

Philippe Rigollet and Austin J Stromme. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 2025.

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

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. *arXiv preprint arXiv:2110.02999*, 2021.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*. Springer, 2020.

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. *TMLR*, 2024.

Karren D Yang and Caroline Uhler. Scalable unbalanced optimal transport using generative adversarial networks. *arXiv preprint arXiv:1810.11447*, 2018.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *ICML*, 2025.

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

A Energy Functional as an Unbalanced Sinkhorn Divergence

This section establishes the two population-level properties used in Sec. 4.2. We first derive the Wasserstein velocity induced by the debiased unbalanced Sinkhorn divergence and then show that the resulting field preserves the desired population target.

A.1 From the UOT Energy to the Wasserstein Velocity

For positive measures $\mu \ll \nu$, define the generalized KL divergence

$$D_{\mathrm{KL}}(\mu\|\nu) = \int \log\left(\frac{d\mu}{d\nu}\right) d\mu - m(\mu) + m(\nu). \tag{21}$$

Using the convex conjugate of $h(r) = r\log r - r + 1$, namely $h^\star(z) = e^z - 1$, the entropic KL-UOT problem admits the dual

$$\begin{aligned}\mathrm{UOT}_{\varepsilon,\tau}(q,p) = \sup_{f,g}\Bigg\{ & \tau\int\left(1 - e^{-f(x)/\tau}\right) dq(x) \\ & + \tau\int\left(1 - e^{-g(y)/\tau}\right) dp(y) \\ & - \varepsilon\iint\left[\exp\left(\frac{f(x)+g(y)-c(x,y)}{\varepsilon}\right) - 1\right] dq(x)dp(y)\Bigg\}.\end{aligned} \tag{22}$$

The optimal coupling has Gibbs form

$$d\pi^{\varepsilon,\tau}_{q,p}(x,y) = \exp\left(\frac{f_{q,p}(x) + g_{q,p}(y) - c(x,y)}{\varepsilon}\right) dq(x)dp(y). \tag{23}$$

Differentiating the dual objective with respect to $f_{q,p}(x)$ gives

$$e^{-f_{q,p}(x)/\tau} = \int \exp\left(\frac{f_{q,p}(x) + g_{q,p}(y) - c(x,y)}{\varepsilon}\right) dp(y). \tag{24}$$

The right-hand side is the density of the first coupling marginal with respect to q. Define

$$s^{\tau}_{q,p}(x) := \frac{d(\pi^{\varepsilon,\tau}_{q,p})_1}{dq}(x) = e^{-f_{q,p}(x)/\tau}. \tag{25}$$

Let

$$T^{\tau}_{q,p}(x) := \frac{\int y\, \pi^{\varepsilon,\tau}_{q,p}(dy \mid x)}{s^{\tau}_{q,p}(x)}. \tag{26}$$

Thus $s^{\tau}_{q,p}(x)$ determines how much mass is transported from x, whereas $T^{\tau}_{q,p}(x)$ determines its normalized barycentric destination.

By the envelope theorem,

$$\frac{\delta \mathrm{UOT}_{\varepsilon,\tau}(q,p)}{\delta q}(x) = (\tau+\varepsilon)\left(1 - s^{\tau}_{q,p}(x)\right). \tag{27}$$

For the debiased energy

$$\mathcal{F}_{\tau}(q;p) := \mathcal{S}^{\mathrm{U}}_{\varepsilon,\tau}(q,p), \tag{28}$$

symmetry of the self term gives

$$\frac{\delta \mathcal{F}_{\tau}}{\delta q}(q;p)(x) = (\tau+\varepsilon)\left(s^{\tau}_{q,q}(x) - s^{\tau}_{q,p}(x)\right). \tag{29}$$

For $c(x,y) = \frac{1}{2}\|x-y\|_2^2$, differentiating the optimality condition yields

$$\nabla f_{q,p}(x) = \frac{\tau}{\tau+\varepsilon}\left(x - T^{\tau}_{q,p}(x)\right). \tag{30}$$

Since $s^{\tau}_{q,p}(x) = e^{-f_{q,p}(x)/\tau}$,

$$\nabla s^{\tau}_{q,p}(x) = \frac{s^{\tau}_{q,p}(x)}{\tau+\varepsilon}\left(T^{\tau}_{q,p}(x) - x\right). \tag{31}$$

Therefore

$$V^{\tau}_{q,p}(x) = -\nabla_x \frac{\delta \mathcal{F}_{\tau}}{\delta q}(q;p)(x) \tag{32}$$

takes the form

$$V^{\tau}_{q,p}(x) = s^{\tau}_{q,p}(x)\left(T^{\tau}_{q,p}(x) - x\right) - s^{\tau}_{q,q}(x)\left(T^{\tau}_{q,q}(x) - x\right). \tag{33}$$

A.2 Population Target Consistency

Let

$$\phi_q(x) := \frac{\delta \mathcal{F}_\tau}{\delta q}(q;p)(x), \qquad V^\tau_{q,p}(x) = -\nabla \phi_q(x). \tag{34}$$

Assume that p and q are supported on a compact, piecewise-C^1, path-connected set $\mathcal{X} \subset \mathbb{R}^d$, that q has full support on $\mathcal{X}$, ϕ_q is C^1 on a neighborhood of $\mathcal{X}$, and $\mathcal{F}_\tau(\cdot; p)$ is convex, nonnegative, and definite on $\mathcal{P}(\mathcal{X})$.

Suppose first that

$$V^\tau_{q,p} = 0 \qquad q\text{-a.e.} \tag{35}$$

Full support and continuity imply

$$\nabla \phi_q(x) = 0, \qquad \forall x \in \mathcal{X}. \tag{36}$$

Path connectivity therefore gives

$$\phi_q(x) = C, \qquad x \in \mathcal{X}. \tag{37}$$

By convexity, for every probability measure r supported on $\mathcal{X}$,

$$\mathcal{F}_\tau(r;p) \geq \mathcal{F}_\tau(q;p) + \int \phi_q(x)\, d(r-q)(x) \tag{38}$$

$$= \mathcal{F}_\tau(q;p), \tag{39}$$

because r and q have the same total mass. Hence q is a global minimizer over $\mathcal{P}(\mathcal{X})$. Since

$$\mathcal{F}_\tau(p;p) = 0 \tag{40}$$

and $\mathcal{F}_\tau$ is nonnegative,

$$\mathcal{F}_\tau(q;p) = 0. \tag{41}$$

Definiteness then implies

$$q = p. \tag{42}$$

Conversely, if $q = p$, the cross and self terms in Eq. 33 coincide, so

$$V^\tau_{q,p} = 0. \tag{43}$$

Therefore

$$V^\tau_{q,p} = 0 \quad q\text{-a.e.} \qquad \Longleftrightarrow \qquad q = p. \tag{44}$$

B Finite-Batch Error of Balanced Sinkhorn Fields

This section derives a non-asymptotic lower bound on the velocity MSE of balanced Sinkhorn fields. The key observation is that exact marginal preservation determines the source-averaged velocity through the empirical sample means. Their sampling fluctuations therefore impose an unavoidable component of the finite-batch error.

B.1 A Non-Asymptotic Lower Bound on Velocity MSE

Mean preservation under balanced transport. Let π^c denote the balanced coupling between $\widehat{q}_B$ and $\widehat{p}_B$. Since each empirical atom has mass $1/B$, its barycentric projection is $T_{\widehat{q}_B,\widehat{p}_B}(x_i) = B \sum_j \pi^c_{ij} y_j$. The exact target marginal constraint, $\sum_i \pi^c_{ij} = 1/B$, then gives

$$\frac{1}{B}\sum_{i=1}^{B} T_{\widehat{q}_B,\widehat{p}_B}(x_i) = \sum_j y_j \sum_i \pi^c_{ij} = \frac{1}{B}\sum_{j=1}^{B} y_j = \overline{Y}_B.$$

Applying the same argument to the self coupling gives $\frac{1}{B}\sum_i T_{\widehat{q}_B,\widehat{q}'_B}(x_i) = \overline{X}'_B$. Consequently,

$$\int \widehat{V}_{\infty,B}(x)\, d\widehat{q}_B(x) = \overline{Y}_B - \overline{X}'_B. \tag{45}$$

Thus, although individual barycentric assignments depend on the transport geometry, their source average is fixed by the empirical marginals.

From the mean error to a lower bound. Let $V_\infty^{\mathrm{pop}} = T_{q,p} - T_{q,q}$ denote the population field, and define $\Delta_B(x) := \widehat{V}_{\infty,B}(x) - V_\infty^{\mathrm{pop}}(x)$. Write $\overline{\Delta}_B := \int \Delta_B\, d\widehat{q}_B$ for its empirical source average. Since $\widehat{q}_B$ is a probability measure, Jensen's inequality yields

$$\mathcal{E}_\infty(B) := \mathbb{E}\big[\|\Delta_B\|^2_{\widehat{q}_B}\big] \geq \mathbb{E}\big[\|\overline{\Delta}_B\|_2^2\big]\,, \tag{46}$$

where $\|\Delta_B\|^2_{\widehat{q}_B} = \frac{1}{B}\sum_i \|\Delta_B(x_i)\|_2^2$. It therefore suffices to compute the second moment of the batch-averaged error.

By Eq. 45,

$$\overline{\Delta}_B = \overline{Y}_B - \overline{X}'_B - \frac{1}{B}\sum_{i=1}^{B} V_\infty^{\mathrm{pop}}(x_i). \tag{47}$$

The same marginal-preservation identity at the population level gives

$$\mathbb{E}_q[V_\infty^{\mathrm{pop}}(X)] = \mathbb{E}_p[Y] - \mathbb{E}_q[X].$$

Hence $\mathbb{E}[\overline{\Delta}_B] = 0$. Moreover, the three sample averages in Eq. 47 depend on mutually independent batches. Using the covariance of an i.i.d. sample mean, we obtain

$$\begin{aligned}\mathbb{E}\|\overline{\Delta}_B\|_2^2 &= \mathrm{Tr}\,\mathrm{Cov}(\overline{\Delta}_B)\\ &= \frac{1}{B}\,\mathrm{Tr}[\mathrm{Cov}_p(Y) + \mathrm{Cov}_q(X) + \mathrm{Cov}_q(V_\infty^{\mathrm{pop}}(X))] =: \frac{L_\infty^2}{B}.\end{aligned} \tag{48}$$

Combining this identity with Eq. 46 proves

$$\mathcal{E}_\infty(B) \geq \frac{L_\infty^2}{B}, \qquad B \geq 1. \tag{49}$$

This bound captures the component of velocity error that remains after averaging over source particles. It follows from exact marginal preservation and batch independence.

C Finite-Batch Error of UW-Flow

This section derives the finite-batch velocity MSE of UW-Flow. Because UOT does not preserve the input marginals exactly, the mean-preservation argument in Appendix B.1 does not directly apply. Instead, we analyze how empirical mass fluctuations perturb the mass factors and barycentric projections that define the field. We first derive the linear response of a single UOT field, then combine the cross and self responses to obtain the velocity MSE, and finally bound its leading coefficient.

Throughout, we fix $\varepsilon > 0$ and $0 < \tau < \infty$ and consider probability measures on fixed finite supports:

$$q = \sum_{i=1}^{M_q} a_i \delta_{x_i}, \qquad p = \sum_{j=1}^{M_p} b_j \delta_{y_j}, \qquad a_i, b_j \geq m_0 > 0, \tag{50}$$

where $\sum_i a_i = \sum_j b_j = 1$. We use the quadratic cost $C_{ij} = \frac{1}{2}\|x_i - y_j\|_2^2$. For a nonnegative weight vector ν, write $\|z\|_\nu^2 := \sum_i \nu_i \|z_i\|_2^2$; unsubscripted norms are Euclidean norms of the stacked vectors.

C.1 First-Order Perturbation of the UOT Field

Population coupling and conditional transport. The population KL-UOT coupling has the form

$$\pi_{ij} = a_i u_i K_{ij} b_j v_j, \qquad K_{ij} = e^{-C_{ij}/\varepsilon}, \qquad \lambda = \frac{\tau}{\tau+\varepsilon} \in (0,1), \tag{51}$$

where the generalized Sinkhorn scalings satisfy

$$u = [K(b \odot v)]^{-\lambda}, \qquad v = [K^\top (a \odot u)]^{-\lambda}. \tag{52}$$

All powers, logarithms, and products denoted by $\odot$ are entrywise.

Let $m^\pi = \pi \mathbf{1}$ and $n^\pi = \pi^\top \mathbf{1}$ be the transported marginals, and define the conditional coupling matrices

$$P_{ij} = \frac{\pi_{ij}}{m_i^\pi}, \qquad Q_{ji} = \frac{\pi_{ij}}{n_j^\pi}. \tag{53}$$

Thus, P averages target quantities conditional on a source atom, while Q averages source quantities conditional on a target atom.

The identity $m_i^\pi P_{ij} = \pi_{ij} = n_j^\pi Q_{ji}$ shows that $Q = P^*$ in the weighted spaces $L^2(m^\pi)$ and $L^2(n^\pi)$. Conditional Jensen's inequality also makes both operators contractions. Consequently,

$$0 \preceq QP \preceq I, \qquad \left\|(I - \lambda^2 QP)^{-1}\right\|_{n^\pi \to n^\pi} \le \frac{1}{1-\lambda^2}. \tag{54}$$

Response of the Sinkhorn scalings. We perturb the source and target masses while keeping their locations fixed. Let

$$r = \log u, \qquad t = \log v, \qquad D_a = \mathrm{diag}(a), \qquad D_b = \mathrm{diag}(b),$$

and define the relative perturbations

$$\rho_a = D_a^{-1}\delta a, \qquad \rho_b = D_b^{-1}\delta b. \tag{55}$$

Differentiating the logarithmic form of Eq. 52 gives

$$\delta r = -\lambda P(\rho_b + \delta t), \qquad \delta t = -\lambda Q(\rho_a + \delta r). \tag{56}$$

The quantity $w := \rho_b + \delta t = \delta \log(b \odot v)$ collects the change in the effective target weights. Eliminating δr yields

$$w = (I - \lambda^2 QP)^{-1}(\rho_b - \lambda Q \rho_a). \tag{57}$$

The inverse is well defined by Eq. 54.

Response of the transport field. Define the transported-mass factor and barycentric projection by

$$s_i = \frac{m_i^\pi}{a_i}, \qquad T_i = \sum_j P_{ij} y_j, \qquad U_{\tau,i} = s_i(T_i - x_i). \tag{58}$$

Marginal perturbations affect both the conditional destination T_i and the mass factor s_i. Their first-order changes are

$$\begin{aligned} \delta P_{ij} &= P_{ij}\big[w_j - (Pw)_i\big], \\ \delta T_i &= \sum_j P_{ij}(y_j - T_i) w_j, \\ \delta s_i &= s_i(1-\lambda)(Pw)_i. \end{aligned} \tag{59}$$

For the last identity, we used $s_i = [K(b \odot v)]_i^{1-\lambda}$. Applying the product rule to $U_{\tau,i} = s_i(T_i - x_i)$ gives

$$\delta U_{\tau,i} = s_i \left[\sum_j P_{ij}(y_j - T_i) w_j + (1-\lambda)(T_i - x_i)(Pw)_i \right]. \tag{60}$$

Let H denote this linear map, so that $\delta U_\tau = Hw$. The two terms in Eq. 60 account for changes in the barycentric projection and the transported-mass factor, respectively.

C.2 FINITE-BATCH VELOCITY MSE

Sampling fluctuations and the debiased response. Under i.i.d. sampling from Eq. 50, the empirical mass vectors satisfy

$$B\widehat{a}_B \sim \mathrm{Multinomial}(B, a), \qquad B\widehat{b}_B \sim \mathrm{Multinomial}(B, b), \qquad B\widehat{a}'_B \sim \mathrm{Multinomial}(B, a), \tag{61}$$

independently. These vectors define $\widehat{q}_B$, $\widehat{p}_B$, and $\widehat{q}'_B$ on the fixed supports. The population and minibatch fields are

$$\begin{aligned} V_\tau^{\mathrm{pop}} &= U_\tau(q, p) - U_\tau(q, q), \\ \widehat{V}_{\tau,B} &= U_\tau(\widehat{q}_B, \widehat{p}_B) - U_\tau(\widehat{q}_B, \widehat{q}'_B). \end{aligned} \tag{62}$$

To separate the batch-size dependence from the field response, define

$$\begin{aligned} \xi_a &= \sqrt{B}\, D_a^{-1}(\widehat{a}_B - a), \\ \xi_b &= \sqrt{B}\, D_b^{-1}(\widehat{b}_B - b), \\ \xi'_a &= \sqrt{B}\, D_a^{-1}(\widehat{a}'_B - a). \end{aligned} \tag{63}$$

These vectors are centered and mutually independent, with

$$\mathrm{Cov}(\xi_a) = \mathrm{Cov}(\xi'_a) = D_a^{-1} - \mathbf{1}\mathbf{1}^\top, \qquad \mathrm{Cov}(\xi_b) = D_b^{-1} - \mathbf{1}\mathbf{1}^\top. \tag{64}$$

In particular, their covariances are independent of B.

Use subscripts c and s for quantities associated with the population cross coupling (q, p) and self coupling (q, q), respectively. Applying Eq. 57 to the scaled fluctuations gives

$$\begin{aligned} w_c &= (I - \lambda^2 Q_c P_c)^{-1}(\xi_b - \lambda Q_c \xi_a), \\ w_s &= (I - \lambda^2 P_s^2)^{-1}(\xi'_a - \lambda P_s \xi_a), \end{aligned} \tag{65}$$

where $P_s = Q_s$ follows from symmetry of the population self coupling. The scaled first-order velocity response and its second moment are

$$L_\tau = H_c w_c - H_s w_s, \qquad A_\tau := \mathbb{E}\|L_\tau\|_a^2. \tag{66}$$

The cross and self responses share the source fluctuation ξ_a; their dependence is therefore retained in A_τ. Since L_τ is linear in noises with the covariances in Eq. 64, A_τ is independent of B.

From linear response to expected MSE. Let

$$\delta\theta_B := (\widehat{a}_B - a, \widehat{b}_B - b, \widehat{a}'_B - a), \qquad e_B := \widehat{V}_{\tau,B} - V_\tau^{\mathrm{pop}}.$$

Multinomial sampling gives

$$\mathbb{E}\|\delta\theta_B\|^2 = O(B^{-1}), \qquad \mathbb{E}\|\delta\theta_B\|^4 = O(B^{-2}). \tag{67}$$

The simultaneous log-scaling map is a λ-contraction in the sup norm and is smooth for positive input masses. Its fixed point, and hence the velocity map, is therefore smooth in the interior of the probability simplices. On the event that every empirical mass is at least $m_0/2$, a second-order Taylor expansion yields

$$e_B = \frac{1}{\sqrt{B}} L_\tau + R_B, \qquad \|R_B\| \le C\|\delta\theta_B\|^2. \tag{68}$$

We also account for batches with small or zero counts. The probability of the complementary event is at most $C_0 e^{-cB}$. Moreover, the strictly positive kernels on fixed finite supports give a bounded extension of the field to all probability weights. Indeed, for input weights (α, β), set

$$z_i = [K(\beta \odot v)]_i, \qquad s_i = z_i^{1-\lambda}, \qquad P_{ij} = \frac{K_{ij}\beta_j v_j}{z_i}.$$

These expressions define $U_i = s_i(T_i - x_i)$ even when $\alpha_i = 0$. The log-scaling contraction bounds the scalings uniformly over (α, β) for the fixed kernel, so this extension is bounded. Its values at unobserved source atoms do not affect the empirical norm.

Combining this bounded extension with Eq. 67, the Taylor bound, and local Lipschitz continuity gives

$$\mathbb{E}\|R_B\|^2 = O(B^{-2}), \qquad \mathbb{E}\|e_B\|^4 = O(B^{-2}). \tag{69}$$

Consequently, Cauchy–Schwarz yields

$$\begin{aligned} \mathbb{E}\|e_B\|_a^2 &= \frac{A_\tau}{B} + \frac{2}{\sqrt{B}}\mathbb{E}\langle L_\tau, R_B\rangle_a + \mathbb{E}\|R_B\|_a^2 \\ &= \frac{A_\tau}{B} + O(B^{-3/2}). \end{aligned} \tag{70}$$

Finally, replacing the population weights by the empirical weights changes the expectation by at most

$$\begin{aligned} \mathbb{E}\left|\|e_B\|_{\widehat{a}_B}^2 - \|e_B\|_a^2\right| &\le \left(\mathbb{E}\|\widehat{a}_B - a\|_\infty^2\right)^{1/2}\left(\mathbb{E}\|e_B\|^4\right)^{1/2} \\ &= O(B^{-3/2}). \end{aligned} \tag{71}$$

Thus,

$$\mathcal{E}_\tau(B) := \mathbb{E}\|e_B\|_{\widehat{a}_B}^2 = \frac{A_\tau}{B} + O(B^{-3/2}). \tag{72}$$

The coefficient A_τ measures the leading velocity response to sampling fluctuations, while $1/B$ captures their batch-size dependence.

C.3 An Upper Bound on UOT Sensitivity

We now bound A_τ through the conditional transport distances

$$H_{\tau,c} := \max_i \sum_j P_{ij}^c \|y_j - x_i\|_2^2, \qquad H_{\tau,s} := \max_i \sum_k P_{ik}^s \|x_k - x_i\|_2^2. \tag{73}$$

These quantities measure the squared transport distances seen from each source atom. We first establish a bound for one coupling and then combine the cross and self terms.

Controlling the mass and noise response. For a population coupling, let $M_\pi := \sum_{ij} \pi_{ij}$ and $\overline{s} := \max_i m_i^\pi / a_i$. The nonnegative cost and probability input masses imply $M_\pi \le 1$. To see this, write $\pi = M_\pi \gamma$ with $\sum_{ij} \gamma_{ij} = 1$. The KL-UOT objective takes the form

$$M_\pi G(\gamma) + (\varepsilon + 2\tau)(M_\pi \log M_\pi - M_\pi + 1),$$

where $G(\gamma) \ge 0$ is the cost-plus-KL objective evaluated at the normalized plan γ. Optimality in M_π therefore gives

$$M_\pi = e^{-G(\gamma)/(\varepsilon+2\tau)} \le 1, \qquad \overline{s} \le \frac{M_\pi}{m_0} \le \frac{1}{m_0}. \tag{74}$$

For the scaled relative noises, Eq. 64 implies

$$\mathbb{E}\|\xi_a\|_{m^\pi}^2 \le \frac{M_\pi}{m_0}, \qquad \mathbb{E}\|\xi_b\|_{n^\pi}^2 \le \frac{M_\pi}{m_0}.$$

Using the resolvent bound in Eq. 54 and the contraction property of Q, we obtain

$$\begin{aligned}\big(\mathbb{E}\|w\|_{n^\pi}^2\big)^{1/2} &\le \frac{\big(\mathbb{E}\|\xi_b\|_{n^\pi}^2\big)^{1/2} + \lambda\big(\mathbb{E}\|\xi_a\|_{m^\pi}^2\big)^{1/2}}{1-\lambda^2} \\ &\le \frac{\sqrt{M_\pi/m_0}}{1-\lambda}.\end{aligned} \tag{75}$$

Controlling the velocity response. For a single coupling, write $H_{\mathrm{geom}} := \max_i \sum_j P_{ij}\|y_j - x_i\|_2^2$. Let $d_{ij} := y_j - x_i$ and $\overline{d}_i := \sum_j P_{ij} d_{ij} = T_i - x_i$. Then Eq. 60 becomes

$$(Hw)_i = s_i \sum_j P_{ij}(d_{ij} - \lambda \overline{d}_i) w_j.$$

Since

$$\begin{aligned}\sum_j P_{ij}\|d_{ij} - \lambda\overline{d}_i\|_2^2 &= \sum_j P_{ij}\|d_{ij}\|_2^2 - \lambda(2-\lambda)\|\overline{d}_i\|_2^2 \\ &\le H_{\mathrm{geom}},\end{aligned}$$

conditional Cauchy–Schwarz gives

$$\begin{aligned}\|Hw\|_a^2 &\le H_{\mathrm{geom}} \sum_i a_i s_i^2 \sum_j P_{ij} w_j^2 \\ &\le \overline{s}\, H_{\mathrm{geom}} \|w\|_{n^\pi}^2.\end{aligned} \tag{76}$$

Combining this with Eq. 75 yields

$$\mathbb{E}\|Hw\|_a^2 \le \frac{H_{\mathrm{geom}}}{m_0^2(1-\lambda)^2}. \tag{77}$$

Applying this bound to the cross and self couplings and using Minkowski's inequality gives

$$A_\tau \le \left(\frac{\sqrt{H_{\tau,c}} + \sqrt{H_{\tau,s}}}{m_0(1-\lambda)}\right)^2. \tag{78}$$

This step does not require independence between the cross and self responses.

Bounded sensitivity under mode separation. Consider a fixed finite collection of matched source and target blocks with centers $c_k(d)$:

$$x_i = c_{k(i)}(d) + \widetilde{x}_i, \qquad y_j = c_{\ell(j)}(d) + \widetilde{y}_j, \qquad \|\widetilde{x}_i\|_2, \|\widetilde{y}_j\|_2 \le R_0. \tag{79}$$

Distinct centers are at least d apart, and every block contains both source and target atoms. Assume that all within-block cross and self distances are bounded by R_{loc}. The atom masses, block structure, R_0, and R_{loc} remain fixed as d increases.

Define the distance-weighted off-block leakage by

$$\ell_c(d) := \max_i \sum_{j:\, \ell(j)\neq k(i)} P_{ij}^c \|y_j - x_i\|_2^2, \tag{80}$$

and define $\ell_s(d)$ analogously for the self coupling. Then

$$H_{\tau,c} \le R_{\mathrm{loc}}^2 + \ell_c(d), \qquad H_{\tau,s} \le R_{\mathrm{loc}}^2 + \ell_s(d). \tag{81}$$

It remains to control the scalings, since decay of the Gibbs kernel alone does not imply decay of the conditional coupling. Set $k_{\mathrm{loc}} := e^{-R_{\mathrm{loc}}^2/(2\varepsilon)}$. The existence of a local partner of mass at least m_0 gives

$$m_0 k_{\mathrm{loc}} \le (Kb)_i \le 1, \qquad m_0 k_{\mathrm{loc}} \le (K^\top a)_j \le 1.$$

The λ-contraction of the log-scaling map therefore bounds its fixed point by

$$\|(\log u, \log v)\|_\infty \le \frac{\lambda}{1-\lambda}\left(\log\frac{1}{m_0} + \frac{R_{\mathrm{loc}}^2}{2\varepsilon}\right) =: L_\star, \tag{82}$$

independently of d. Consequently,

$$P_{ij}^c = \frac{K_{ij} b_j v_j}{\sum_\ell K_{i\ell} b_\ell v_\ell} \le C_\star b_j e^{-\|x_i - y_j\|_2^2/(2\varepsilon)}, \qquad C_\star := \frac{e^{2L_\star}}{m_0 k_{\mathrm{loc}}}. \tag{83}$$

For off-block pairs, $\|x_i - y_j\|_2 \ge d - 2R_0$. Since $r^2 e^{-r^2/(2\varepsilon)}$ decreases for $r \ge \sqrt{2\varepsilon}$, sufficiently large d gives

$$\ell_c(d) \le C_\star (d - 2R_0)^2 e^{-(d-2R_0)^2/(2\varepsilon)} \longrightarrow 0. \tag{84}$$

The same argument applies to the self coupling, so $\ell_s(d) \to 0$. Combining Eq. 81 with Eq. 78 proves

$$A_\tau(d) \le \frac{4R_{\mathrm{loc}}^2 + 2\ell_c(d) + 2\ell_s(d)}{m_0^2(1-\lambda)^2} = O(1), \qquad d \to \infty. \tag{85}$$

Thus, for fixed finite τ and ε, the leading UOT sensitivity remains bounded as matched modes separate.

D Comparison of Balanced and Unbalanced Finite-Batch Errors

This section directly compares the finite-batch velocity errors derived for balanced Sinkhorn transport in Appendix B and UW-Flow in Appendix C. The comparison shows that both errors operate at the same leading $1/B$ sampling scale, while marginal relaxation can reduce the corresponding sensitivity coefficient.

Comparison of the leading finite-batch errors. Appendix B gives the non-asymptotic lower bound

$$\mathcal{E}_\infty(B) \ge \frac{L_\infty^2}{B}, \qquad B \ge 1, \tag{86}$$

where L_∞^2 is the variance contribution imposed by exact marginal preservation.

For finite τ, Appendix C gives

$$\mathcal{E}_\tau(B) = \frac{A_\tau}{B} + O(B^{-3/2}), \tag{87}$$

where A_τ is the leading sensitivity of the UOT velocity field to empirical marginal perturbations.

Suppose

$$A_\tau < L_\infty^2. \tag{88}$$

Let

$$g := L_\infty^2 - A_\tau > 0. \tag{89}$$

Multiplying Eq. 87 by B gives

$$B\mathcal{E}_\tau(B) = A_\tau + O(B^{-1/2}). \tag{90}$$

Hence, there exists $B_0 < \infty$ such that

$$|B\mathcal{E}_\tau(B) - A_\tau| < \frac{g}{2}, \qquad B \geq B_0. \tag{91}$$

For such B,

$$B\mathcal{E}_\tau(B) < A_\tau + \frac{g}{2} < L_\infty^2 \leq B\mathcal{E}_\infty(B), \tag{92}$$

and therefore

$$\mathcal{E}_\tau(B) < \mathcal{E}_\infty(B), \qquad B \geq B_0. \tag{93}$$

Thus, the comparison does not rely on a different convergence rate in B: both errors scale at the leading $1/B$ order. The advantage arises when marginal relaxation reduces the leading finite-batch sensitivity from the balanced level.

Implication under separated modes. We now specialize the coefficient comparison to the paired-mode setting considered in Sec. 4.4. For any fixed finite τ, Appendix C shows that

$$A_\tau(d) = O(1) \qquad \text{as } d \to \infty, \tag{94}$$

provided the local source–target geometry remains bounded.

In contrast, Appendix B gives

$$L_\infty^2 \geq \omega(1-\omega)d_\mu^2. \tag{95}$$

Therefore, if the separation of the target modes satisfies

$$d_\mu = \Omega(d), \tag{96}$$

then

$$L_\infty^2(d) = \Omega(d^2). \tag{97}$$

Consequently, for sufficiently large mode separation, $L_\infty^2(d)$ exceeds the bounded UOT coefficient $A_\tau(d)$, so the condition in Eq. 88 is satisfied. It follows that

$$\exists d_0\ \forall d \geq d_0\ \exists B_0(d)\ \forall B \geq B_0(d): \qquad \mathcal{E}_\tau(B) < \mathcal{E}_\infty(B). \tag{98}$$

Hence, in the separated-mode regime, exact marginal preservation can make the balanced field increasingly sensitive to finite-batch mass fluctuations, while finite-τ UOT can keep the leading sensitivity controlled by the local transport geometry.

E Additional Velocity-Field Experiments

E.1 Evaluation Protocol and Metrics

Metrics. For both experiments, we compare the minibatch velocity $\widehat{V}_{\tau,B}$ with its corresponding population field V_τ^{pop} using

$$B\,\mathrm{MSE}, \qquad \mathrm{RelMSE}, \qquad \text{Cosine Similarity}.$$

We report the batch-scaled velocity error

$$B\,\mathrm{MSE}_\tau(B) = B\left\|\widehat{V}_{\tau,B} - V_\tau^{\mathrm{pop}}\right\|_{\widehat{q}_B}^2,$$

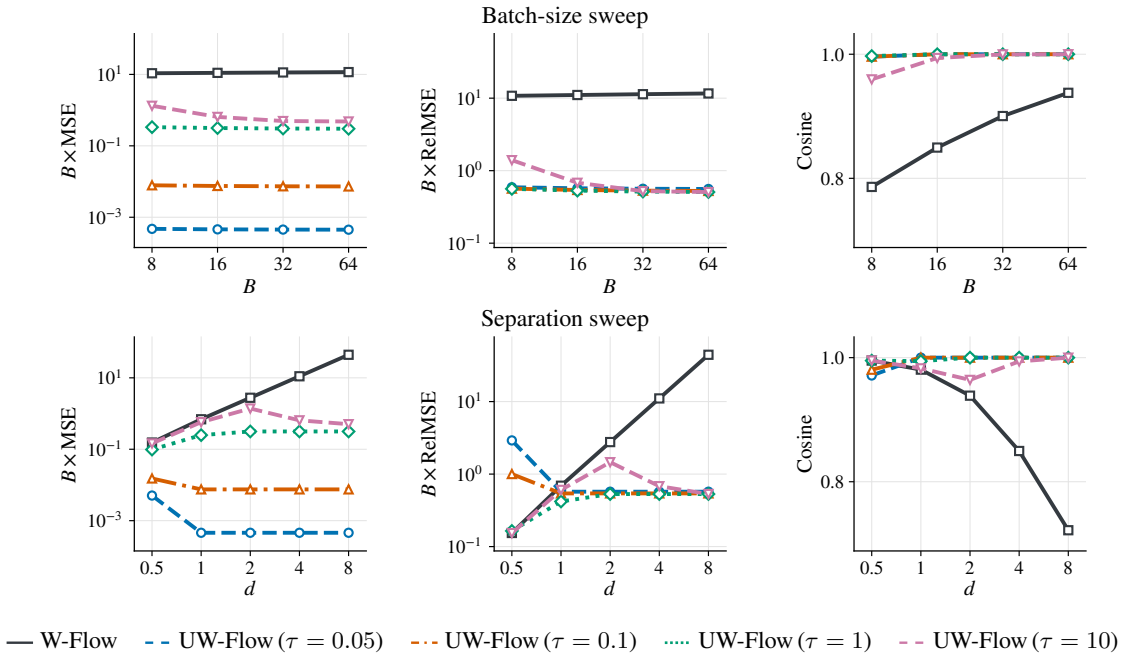


Figure 6: **Full fixed-support velocity-error results.** Columns show $B \times \mathrm{MSE}$, $B \times \mathrm{RelMSE}$, and cosine similarity. The top and bottom rows vary B and d, respectively.

which factors out the leading $1/B$ sampling dependence and facilitates comparison of finite-batch sensitivity across batch sizes.

Since a smaller population velocity can also reduce the absolute error, we report the relative MSE

$$\mathrm{RelMSE}_\tau(B) = \frac{\left\| \widehat{V}_{\tau,B} - V_\tau^{\mathrm{pop}} \right\|_{\widehat{q}_B}^2}{\|V_\tau^{\mathrm{pop}}\|_{\widehat{q}_B}^2},$$

which measures the error relative to the population-field magnitude. Finally, cosine similarity measures directional agreement between the minibatch and population velocity fields independently of their scale.

E.2 Fixed-Support Setting

We consider two matched source–target pairs with fixed supports

$$x_\pm = (\pm d/2, 0), \qquad y_\pm = (\pm d/2, 1),$$

and assign probability mass $1/2$ to each point. The support locations are fixed across trials, while independent minibatches $\widehat{q}_B$, $\widehat{p}_B$, and $\widehat{q}'_B$ induce fluctuations only in the empirical masses. This directly matches the fixed-support multinomial perturbation setting analyzed in Sec. 4.3.

For the separation sweep, we fix $B = 16$ and use $d \in \{0.5, 1, 2, 4, 8\}$. We compare UOT with $\tau \in \{0.05, 0.1, 1, 10\}$ against balanced OT, corresponding to $\tau = \infty$. Fig. 6 presents the full results for $B \times \mathrm{MSE}$, $B \times \mathrm{RelMSE}$, and cosine similarity.

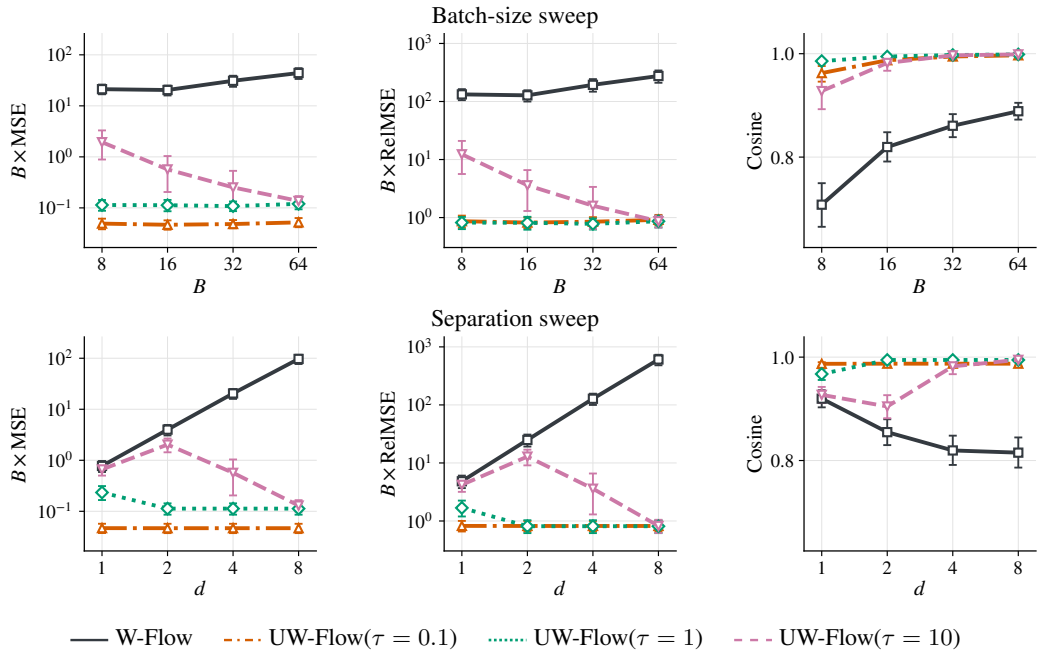


Figure 7: **Velocity errors under Gaussian-mixture sampling.** Columns show $B \times$ MSE, $B \times$ RelMSE, and cosine similarity. The top and bottom rows vary B and d, respectively.

E.3 Continuous Gaussian-Mixture Setting

To move beyond fixed supports, we replace each point above with an isotropic Gaussian component. Specifically, the source and target distributions are equal-weight mixtures centered at

$$(\pm d/2, 0) \qquad \text{and} \qquad (\pm d/2, h),$$

respectively, with covariance $\sigma^2 I$. We use $h = 1$ and $\sigma = 0.15$. Fresh samples are drawn at every trial, so minibatch variation affects both the empirical mode proportions and the sampled support locations.

We use $B \in \{8, 16, 32, 64\}$ at $d = 4$ for the batch-size sweep and $d \in \{0.5, 1, 2, 4, 8\}$ at $B = 16$ for the separation sweep. We compare UOT with $\tau \in \{0.1, 1, 10\}$ against balanced OT. Each configuration is evaluated over 64 independent trials. Fig. 7 reports the corresponding $B \times$ MSE, $B \times$ RelMSE, and cosine similarity.

F Experimental Details and Additional Generation Results

This appendix provides the setup and full $B \times \tau$ ablations for toy distribution matching (Sec. F.1), followed by image-generation hyperparameters (Sec. F.3). We also report detailed settings and complete results for the fixed-budget minibatch ablation (Sec. F.2) and ImageNet-100 generation (Sec. F.4), together with ablations of the Sinkhorn iteration count and entropic coefficient (Sec. F.5).

F.1 Generation on Synthetic Distributions

We provide experimental details and full $B \times \tau$ ablations for the synthetic generation experiments in Sec. 5.1. We consider three two-dimensional target distributions: checkerboard, two moons, and Gaussian ring.

Training configuration. W-Flow and UW-Flow use the same Fourier-feature generator with an effective optimizer batch size of $B_{\text{opt}} = 512$. We vary the transport batch size $B \in \{8, 32, 128, 512\}$ and evaluate UW-Flow with $\tau \in \{0.01, 0.1, 1, 10\}$. All runs use 30,000 optimizer updates, 30 Sinkhorn iterations, $\varepsilon = 0.01$, and random seed 0. The reported ε and τ denote coefficients in the normalized training loss.

Full qualitative ablations. Fig. 8, Fig. 9, and Fig. 10 show that UW-Flow with $\tau \in 0.01, 0.1$ generally captures the target structures more clearly than W-Flow at smaller B. As B increases, both methods improve and the qualitative gap tends to narrow. Larger τ generally produces results closer to W-Flow.

F.2 Fixed-Budget Minibatch Ablation

Each transport group contains B source samples, B target samples, and B independently drawn self-reference samples. Cross and self transport couplings are computed separately within each group.

We vary $B \in \{4, 8, 16, 32, 64\}$ while keeping

$$B_{\text{opt}} = N_c B$$

fixed at 64 for MNIST and 128 for CIFAR-10. Accordingly, we set $N_c = B_{\text{opt}}/B$. The corresponding values of N_c are $\{16, 8, 4, 2, 1\}$ for MNIST and $\{32, 16, 8, 4, 2\}$ for CIFAR-10. Regression losses from all groups are aggregated for each optimizer update.

For each dataset and support size, W-Flow and UW-Flow use the same model, feature representation, optimizer, training schedule, and entropic regularization parameter. They differ in the transport objective used to construct the velocity. We compute the evaluation metrics using 50,000 generated samples for CIFAR-10 and 10,000 for MNIST. Additional hyperparameters and evaluation settings are provided in Appendix F.3. The complete results corresponding to Tab. 2 and Tab. 3 are reported in Tab. 5 and Tab. 6, respectively.

F.3 Image-Generation Hyperparameters

Tab. 7 reports the model, transport, and optimization settings for the MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky et al., 2009) experiments in Sec. 5.2 and the ImageNet-100 (Deng et al., 2009) experiments in Sec. 5.3. The ImageNet-100 experiments are based on the ImageNet ablation setup of W-Flow (Han et al., 2026). For the fixed-budget ablation, we vary N_c and the samples per transport group as detailed in Appendix F.2.

For Sec. 5.2, we vary N_c and the number of samples per transport group while fixing the per-update budget B_{opt}, as detailed in Appendix F.2. The other dataset-specific hyperparameters follow Tab. 7 unless stated otherwise. For ImageNet-100, we follow the ImageNet ablation setup of W-Flow (Han et al., 2026), changing only the number of classes to 100.

F.4 ImageNet-100 Generation

Dataset and training setup. We use the 100 ImageNet classes specified by the ImageNet-100 class list released with Contrastive Multiview Coding (Tian et al., 2020). We follow the ImageNet ablation setup of W-Flow (Han et al., 2026), adapting it to the 100-class subset. We train a DiT-B/2 generator in the SD-VAE (Rombach et al., 2022) latent space and compute transport costs using features from a frozen ResNet (He et al., 2016) encoder pretrained with latent-MAE. Both methods use the same architecture,

Table 5: Full results of the fixed-budget minibatch discrepancy ablation on unconditional MNIST, with fixed $B_{\text{opt}} = N_c N_{\text{gen}} = 64$.

N_c	N_{gen}	N_{pos}	Method	FID ↓	Prec. ↑	Rec. ↑
16	4	4	W-Flow	19.70	0.536	0.059
			UW-Flow ($\tau = 10$)	22.54	0.562	0.081
			UW-Flow ($\tau = 1$)	17.98	0.464	0.116
			UW-Flow ($\tau = 0.1$)	18.66	0.419	0.253
8	8	8	W-Flow	13.69	0.461	0.239
			UW-Flow ($\tau = 10$)	13.50	0.545	0.110
			UW-Flow ($\tau = 1$)	12.37	0.477	0.281
			UW-Flow ($\tau = 0.1$)	13.56	0.492	0.351
4	16	16	W-Flow	9.86	0.487	0.430
			UW-Flow ($\tau = 10$)	10.14	0.513	0.355
			UW-Flow ($\tau = 1$)	9.97	0.517	0.218
			UW-Flow ($\tau = 0.1$)	9.22	0.470	0.507
2	32	32	W-Flow	8.72	0.576	0.317
			UW-Flow ($\tau = 10$)	7.33	0.596	0.343
			UW-Flow ($\tau = 1$)	7.85	0.517	0.435
			UW-Flow ($\tau = 0.1$)	7.78	0.528	0.530
1	64	64	W-Flow	7.03	0.572	0.532
			UW-Flow ($\tau = 10$)	6.11	0.589	0.578
			UW-Flow ($\tau = 1$)	5.67	0.585	0.568
			UW-Flow ($\tau = 0.1$)	6.18	0.559	0.603

encoder weights, and optimization protocol within each setting. We use $N_c = 64$ transport groups with $N_{\text{gen}} = N_{\text{pos}} = 64$ samples per group, giving $B_{\text{opt}} = N_c N_{\text{gen}} = 4096$. Additional hyperparameters are listed in Tab. 7.

Parameter selection and evaluation. We evaluate UW-Flow with $\tau \in \{0.1, 1, 10\}$ in both settings. For class-conditional evaluation, we follow W-Flow's CFG selection protocol, testing guidance scales in $\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$ and reporting the metrics at the scale with the lowest FID for each method. Tab. 8 reports the complete results across τ.

F.5 Sinkhorn Parameter Ablations

Following W-Flow (Han et al., 2026), we examine the effects of Sinkhorn iterations L and the entropic coefficient ε on UW-Flow. We use the CIFAR-10 configuration with $N_c = 2$, $N_{\text{gen}} = N_{\text{pos}} = 64$, and $\tau = 0.1$, varying one parameter at a time. W-Flow selects $\varepsilon = 0.05$ based on its ablation study. We retain this value for the results reported in the main text to compare W-Flow and UW-Flow under the same entropic regularization; the default iteration count is $L = 30$.

As shown in Tab. 9, FID ranges from 4.51 to 5.18 across $L \in \{5, 10, 20, 30, 40\}$, with no monotonic improvement as L increases. The choice of ε has a larger effect in this setting: at $L = 30$, reducing ε from 0.05 to 0.025 improves FID from 4.51 to 3.75 and recall from 0.45 to 0.52. The main-text results use $\varepsilon = 0.05$ for the matched comparison; the lower FID at $\varepsilon = 0.025$ is reported here as an additional ablation result.

Table 6: Full results of the fixed-budget minibatch discrepancy ablation on unconditional CIFAR-10, with fixed $B_{\text{opt}} = N_c N_{\text{gen}} = 128$.

N_c	N_{gen}	N_{pos}	Method	FID ↓	IS ↑	Prec. ↑	Rec. ↑
32	4	4	W-Flow	30.58	9.75	0.55	0.01
			UW-Flow ($\tau = 10$)	22.03	9.58	0.57	0.02
			UW-Flow ($\tau = 1$)	19.41	10.10	0.56	0.02
			UW-Flow ($\tau = 0.1$)	15.56	10.60	0.52	0.07
16	8	8	W-Flow	24.39	8.99	0.58	0.01
			UW-Flow ($\tau = 10$)	14.64	10.39	0.59	0.04
			UW-Flow ($\tau = 1$)	13.71	10.82	0.56	0.05
			UW-Flow ($\tau = 0.1$)	9.70	10.53	0.56	0.17
8	16	16	W-Flow	12.49	10.09	0.59	0.07
			UW-Flow ($\tau = 10$)	10.39	10.16	0.59	0.13
			UW-Flow ($\tau = 1$)	10.03	9.87	0.61	0.15
			UW-Flow ($\tau = 0.1$)	13.97	10.09	0.52	0.24
4	32	32	W-Flow	9.57	10.32	0.63	0.08
			UW-Flow ($\tau = 10$)	8.68	10.19	0.61	0.23
			UW-Flow ($\tau = 1$)	8.43	10.20	0.60	0.16
			UW-Flow ($\tau = 0.1$)	5.01	10.43	0.61	0.39
2	64	64	W-Flow	8.10	10.37	0.60	0.27
			UW-Flow ($\tau = 10$)	6.67	10.30	0.60	0.31
			UW-Flow ($\tau = 1$)	5.11	10.23	0.62	0.39
			UW-Flow ($\tau = 0.1$)	4.51	10.37	0.60	0.45

Table 7: Hyperparameters for image-generation experiments. N_q and N_p are the numbers of generated and target samples per transport group; the total generated batch size is $N_c N_q$. The listed ε and τ values are before scaling by the feature dimension D.

Hyperparameter	MNIST	CIFAR-10	ImageNet-100
Model and representation			
Generator	DiT-based, patch 2	DiT-based, patch 2	DiT-B/2
Training shape / space	$28\times28\times1$	$32\times32\times3$	$32\times32\times4$
Feature encoder (frozen)	ConvNeXt V2–Tiny	ConvNeXt V2–Tiny	Latent-MAE ResNet (width 256)
Feature normalization	Distance-based scaling	Distance-based scaling	Distance-based scaling
Velocity normalization	RMS normalization	RMS normalization	RMS normalization
Transport			
Per-update budget $B_{\text{opt}} = N_c N_q$	64	128	4,096
Samples per group $N_q = N_p$ †	$\{4, 8, 16, 32, 64\}$	$\{4, 8, 16, 32, 64\}$	64
Transport groups N_c	B_{opt}/N_q	B_{opt}/N_q	64
Entropic coefficient ε	0.05	0.05	0.05
Marginal penalty coefficient τ	$\{0.1, 1, 10\}$	$\{0.1, 1, 10\}$	$\{0.1, 1, 10\}$
Sinkhorn iterations L	30	30	10
Drift step size η	1.0	1.0	1.0
Optimization			
Optimizer	AdamW	AdamW	AdamW
(β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Learning rate	2×10^{-4}	2×10^{-4}	2×10^{-4}
Weight decay	0.01	0.01	0.01
LR schedule (after warmup)	Cosine	Cosine	Constant
EMA decay	0.999	0.999	0.999
Training budget	100 epochs	500 epochs	100 epochs

†For MNIST and CIFAR-10, N_q and N_c are varied jointly with B_{opt} fixed.

Table 8: Full ImageNet-100 generation results following the ImageNet ablation setup and CFG selection protocol of W-Flow (Han et al., 2026). Both settings use $N_c = 64$, $N_{\text{gen}} = N_{\text{pos}} = 64$, and $B_{\text{opt}} = 4096$. For class-conditional generation, we select the CFG scale with the lowest FID for each method from $\{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$.

Setting	N_c	N_{gen}	N_{pos}	B_{opt}	Method	FID ↓	IS ↑	Prec. ↑	Rec. ↑
Uncond	64	64	64	4096	W-Flow	55.25	23.89	0.42	0.14
					UW-Flow ($\tau = 10$)	55.00	25.01	0.40	0.16
					UW-Flow ($\tau = 1$)	51.08	25.47	0.41	0.19
					UW-Flow ($\tau = 0.1$)	46.17	27.71	0.40	0.29
Cond	64	64	64	4096	W-Flow	9.43	60.31	0.74	0.26
					UW-Flow ($\tau = 10$)	9.47	61.08	0.75	0.25
					UW-Flow ($\tau = 1$)	9.11	61.34	0.75	0.26
					UW-Flow ($\tau = 0.1$)	9.59	58.35	0.78	0.25

Table 9: Sinkhorn parameter ablations on CIFAR-10. We fix $N_c = 2$, $N_{\text{gen}} = N_{\text{pos}} = 64$, $B_{\text{opt}} = 128$, and $\tau = 0.1$. The default configuration uses $L = 30$ and $\varepsilon = 0.05$; each table varies one parameter while fixing the other.

Sinkhorn iterations L				
L	FID ↓	IS ↑	Precision ↑	Recall ↑
5	5.04	10.39	0.57	0.41
10	4.73	**10.54**	0.58	**0.45**
20	5.16	10.26	0.59	0.40
30	**4.51**	10.37	**0.60**	**0.45**
40	5.18	10.29	0.57	**0.45**

Entropic regularization ε				
ε	FID ↓	IS ↑	Precision ↑	Recall ↑
0.01	3.77	10.21	0.60	**0.52**
0.025	**3.75**	10.39	**0.61**	**0.52**
0.05	4.51	10.37	0.60	0.45
0.075	5.47	**10.60**	0.57	0.39
0.1	8.35	10.54	0.54	0.23

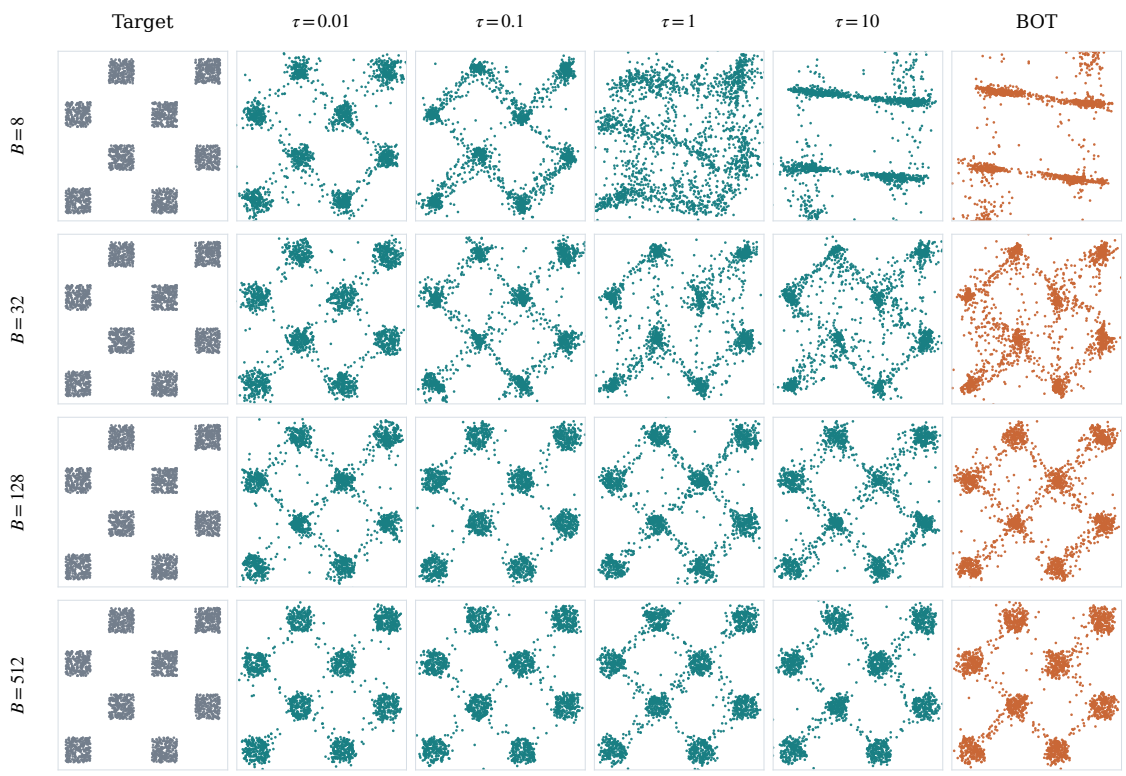


Figure 8: Qualitative $B \times \tau$ ablation on checkerboard. Rows vary B; columns show the target, UW-Flow with increasing τ, and W-Flow.

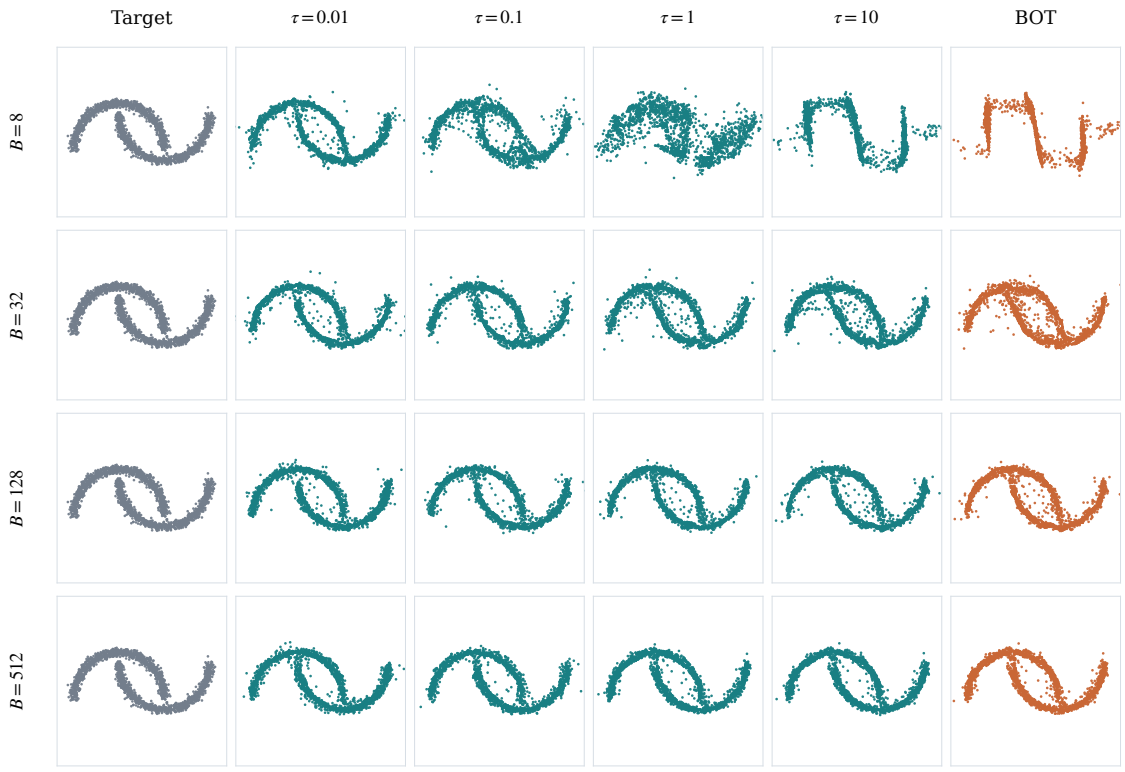


Figure 9: Qualitative $B \times \tau$ ablation on two moons. Rows vary B; columns show the target, UW-Flow with increasing τ, and W-Flow.

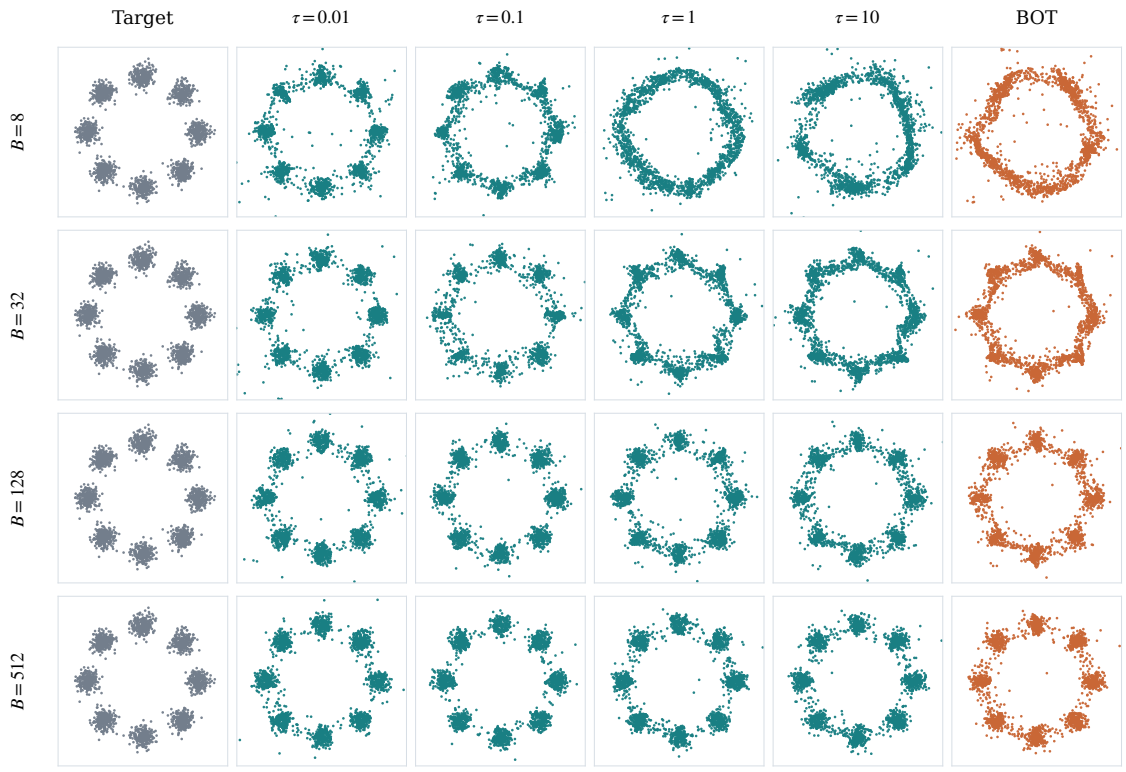


Figure 10: Qualitative $B \times \tau$ ablation on Gaussian ring. Rows vary B; columns show the target, UW-Flow with increasing τ, and W-Flow.