Generalized Wasserstein Flow Matching: Nested Couplings for Point Cloud Generation

Moritz Piening[1,2] **Richard Duong**[1] **Gabriele Steidl**[1]
[1]Department of Mathematics, Technische Universität Berlin, Germany
[2]Institute of Computer Science, Faculty of Mathematics and Computer Science
University of Göttingen, Germany
{piening,duong,steidl}@math.tu-berlin.de moritz.piening@uni-goettingen.de

Abstract

We formulate flow matching on probability measures over probability measures. Directly from the Wasserstein-on-Wasserstein (WoW) continuity equation, we show that admissible nested outer and inner couplings induce a valid metameasure flow and non-local velocity field. Validity requires neither optimal plans nor a smooth manifold. Practically, this opens up the design space of multisample flow matching: exact Wasserstein plans are distinguished by energy and straightness, while sliced and lazy linear plans yield scalable alternatives. Point-cloud experiments, including controlled Bures–Wasserstein and Euclidean minibatch baselines, demonstrate the resulting geometry–cost trade-off.[1]

1 Introduction

Among the wide range of generative modelling approaches, flow matching (FM) has emerged as a particularly flexible and conceptually simple framework [Albergo and Vanden-Eijnden, 2023, Lipman et al., 2023, Liu et al., 2023]. Given a fixed target probability measure, FM learns an implicitly defined time-dependent vector field that transports samples from a simple source distribution (e.g., a standard Gaussian) to the target. In contrast to stochastic generative models such as diffusion models [Ho et al., 2020, Yang et al., 2023], flow matching constructs a deterministic continuous normalizing flow. This is achieved by regressing the vector field onto a conditional probability flow induced by a prescribed coupling between source and target measures. Instead of costly stochastic simulations, this results in solving a deterministic ordinary differential equation at sampling time.

While originally introduced in Euclidean spaces, the flow matching framework extends naturally to more general geometric settings. Notably, Riemannian flow matching replaces linear interpolations with geodesic paths, enabling generative modelling on manifold-valued data [Chen and Lipman, 2024]. Beyond the Riemannian setting, recent developments have further generalized the framework to function spaces [Kerrigan et al., 2024, Holderrieth et al., 2025], discrete spaces [Gat et al., 2024, Qin et al., 2025] or quotient spaces [Klein et al., 2023, Ruscelli, 2026]. In this work, we are particularly interested in the case of *Wasserstein* flow matching on the space of probability measures [Atanackovic et al., 2025, Haviv et al., 2025]. Equipping this space with the Wasserstein metric results in the so-called Wasserstein manifold. Importantly, this terminology is somewhat misleading, as the space fails to satisfy the exact manifold definition due to the presence of singularities. Particularly, the subset of fixed-size empirical measures is rather a Euclidean quotient space containing permutation-based equivalence classes [Boumal, 2023].

From a theoretical perspective, flow matching can be interpreted as learning curves of probability measures connecting a tractable source measure to a target distribution given by samples [Lipman et al., 2023, Wald and Steidl, 2025]. This viewpoint naturally places flow matching within the framework of optimal transport (OT) and Wasserstein geometry, where probability measures are treated as points in the metric Wasserstein space. In this context, our learned transport field corresponds

[1]Code: https://github.com/MoePien/wow-flow-matching

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

to an interpolation between measures in the Wasserstein space. Since the coupling (or transport plan) between the source and target can be arbitrarily chosen, this theoretical background further provides a natural design choice via *multisample* flow matching [Pooladian et al., 2023]. Since straighter velocity fields improve numerical sampling performance [Liu et al., 2023], it is possible to leverage the dynamic OT formulation to learn straight OT velocity fields [Chemseddine et al., 2025, Mousavi-Hosseini et al., 2026, Tong et al., 2024].

In this work, we build on this perspective and extend flow matching beyond classical probability measures to metameasures, i.e., probability measures over probability measures. Using the Wasserstein structure of the base space, we equip the space of metameasures with a second-order Wasserstein geometry and study flow-based interpolations in this Wasserstein-on-Wasserstein (WoW) space. This provides a natural bridge between recent work on the theory of the WoW space [Beiglböck et al., 2025, Bonet et al., 2025, Emami and Pass, 2025, Huesmann and Müller, 2025, Pinzi and Savaré, 2025], computational methods for hierarchical OT [Bonet et al., 2026, Nguyen et al., 2025, Piening and Beinert, 2025, 2026b] and generative methods on the Wasserstein manifolds [Atanackovic et al., 2025, Haviv et al., 2025, Jiang et al., 2025, Tang et al., 2026]. Notably, once we employ the natural discretization of the Wasserstein space via empirical measures, these methods are closely related to general point cloud generative models for unordered sets [Li et al., 2025, Lüdke et al., 2025] and point clouds [Luo and Hu, 2021, Yang et al., 2019]. Similarly to the use of OT couplings in Euclidean flow matching, we propose to leverage the tools of computational OT to balance efficient training with numerically stable velocity fields.

We make a conceptual contribution with practical consequences: the WoW continuity equation shows that arbitrary admissible nested plans define valid flows, without invoking a smooth "Wasserstein manifold." Optimal plans minimize transport energy and give straight WoW geodesics. Sliced, lazy linear, and independent plans remain valid with different accuracy–cost trade-offs.

WFM [Haviv et al., 2025] uses optimal inner maps and, in its multisample variant, Gaussian Fréchet/Bures outer matching. Our framework recovers these cases and permits full-Wasserstein, sliced, and lazy linear outer geometry. NSOT [Hui et al., 2025] instead constructs an offline coupling from high-resolution noise supersets. Appendix D discusses the distinction and an adapted reproduction. Empirically, Bures and flattened Euclidean minibatch OT test whether reordering alone suffices, while our geometry-aware plans target few-step generation on point clouds and finite molecular sets. Summarizing our **contributions**,

- we prove that nested outer and inner transport plans induce metameasure flows directly through the WoW continuity equation, and identify their non-local velocity fields as least-squares minimizers (Theorem 2.5).
- we show that coupling validity is more general than WoW optimality and design sliced and lazy linear constructions that trade exact geometry for scalability.
- we validate the coupling principle through controlled Bures–Wasserstein and Euclidean-minibatch baselines on 2D point clouds, 3D shapes, and finite molecular sets.

All novel proofs are given in Appendix A.

2 Curves between Probability Measures

In flow matching generative models, parametrized measure curves are constructed that connect an easy-to-sample source measure with a target measure which is available only through samples [Lipman et al., 2023, Wald and Steidl, 2025]. Notably, this description is not necessarily limited to Euclidean probability measures [Chen and Lipman, 2024]. After a short recap on the Wasserstein geometry [Ambrosio et al., 2005, Santambrogio, 2015, Wald and Steidl, 2025], we will indeed see how we can lift this construction to stochastic measures, i.e., metameasures.

2.1 Wasserstein Geometry

Let (X, d) be a Polish metric space and $\mathcal{P}_2(X)$ the set of Borel probability measures on X with finite second moments, i.e. $\int_X d(x, x_0)^2 \,\mathrm{d}\mu(x) < \infty$ for some $x_0 \in X$. The set $\mathcal{P}_2(X)$ becomes a complete metric space with the *Wasserstein distance* W_2 which is given for any $\mu, \nu \in \mathcal{P}_2(X)$ by

$$\mathrm{W}_2^2(\mu, \nu) \coloneqq \min_{\pi \in \mathrm{c}(\mu,\nu)} \int_{X \times X} d(x, x')^2 \,\mathrm{d}\pi(x, x'). \tag{1}$$

The *couplings* or *transport plans* are given by $\mathrm{c}(\mu, \nu) = \{\pi \in \mathcal{P}_2(X \times X) : \mathrm{proj}^0_\sharp \pi = \mu, \mathrm{proj}^1_\sharp \pi = \nu\}$ with $\mathrm{proj}^i : X \times X \to X$, $(x_0, x_1) \mapsto x_i$. By $\mathrm{c}^{\mathrm{opt}}(\mu, \nu)$ we denote *optimal* couplings, where the

minimum in (1) is attained. For $I = [0, 1]$, let $C_I(X)$ denote the set of continuous curves mapping from I to X, and $AC_I(X)$ the set of absolutely continuous ones.
We are interested in two examples for (X, d), namely the Euclidean setting $(\mathbb{R}^d, \|\cdot\|)$, and the so-called *Wasserstein-on-Wasserstein* (WoW) case $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. For visual distinction, we use capital/bold letters in the latter case $X = \mathcal{P}_2(\mathbb{R}^d)$, i.e. for any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ we define

$$\mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \min_{\Pi \in \mathrm{C}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d)} \mathrm{W}_2^2(\mu, \nu) \, \mathrm{d}\Pi(\mu, \nu), \tag{2}$$

where the *outer plans* are given by

$$\mathrm{C}(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \{\Pi \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d)) : \mathrm{Proj}^0_\sharp \Pi = \boldsymbol{\mu}, \ \mathrm{Proj}^1_\sharp \Pi = \boldsymbol{\nu}\}.$$

In contrast, we address elements $\pi \in \mathrm{c}(\mu, \nu)$ in the case $X = \mathbb{R}^d$ as *inner plans*. By $\mathrm{C}^{\mathrm{opt}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ we again denote the couplings which minimize (2).

2.2 Wasserstein-on-Euclidean

In the following, we briefly recall the basic idea of flow matching in the Euclidean case, see, e.g. [Lipman et al., 2023, Wald and Steidl, 2025] and generalize it to the WoW setting in the next subsection.

Theorem 2.1. *Let $(\mu_t)_{t \in I} \in C_I(\mathcal{P}(\mathbb{R}^d))$ and $v : I \times \mathbb{R}^d \to \mathbb{R}^d$ be a sufficiently smooth[2] Borel vector field satisfying $\int_I \int_{\mathbb{R}^d} \|v\| \, \mathrm{d}\mu_t(x) \, \mathrm{d}t < \infty$. Assume that (μ_t, v_t) solves the continuity equation*

$$\partial_t \mu_t + \mathrm{div}(v_t \mu_t) = 0 \tag{3}$$

in the sense of distributions. Then, for μ_0-a.e. $x \in \mathbb{R}^d$ the ODE

$$\dot{\gamma}(t, x) = v(t, \gamma(t, x)), \ t \in (0, 1), \quad \gamma(0, x) = x,$$

admits a unique solution $\gamma(\cdot, x)$ on I, and with $\gamma_t : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto \gamma(t, x)$, it holds $\mu_t = \gamma_{t,\sharp} \mu_0$.

Once such a velocity field $v : I \times \mathbb{R}^d \to \mathbb{R}^d$ is known, we can use any ODE solver to transport samples from μ_0 to μ_1 via the relation $\mu_t = \gamma_{t,\sharp} \mu_0$.
In flow matching, this velocity field is learned from samples of μ_1 using a neural network. A simple construction is based on the following observation.

Theorem 2.2. *For any coupling $\pi \in c(\mu, \nu)$, the curve $(\mu_t)_{t \in I}$ defined by $\mu_t \coloneqq \mathrm{proj}^t_\sharp \pi$, $\mathrm{proj}^t(x, x') \coloneqq (1 - t)x + tx'$, fulfills the continuity equation* (3) *together with a vector field v_t^π which minimizes*

$$J(v) \coloneqq \int_I \int_{\mathbb{R}^d \times \mathbb{R}^d} \|v_t(\mathrm{proj}^t(x, x')) - (x' - x)\|^2 \, \mathrm{d}\pi(x, x') \, \mathrm{d}t.$$

If $\pi \in \mathrm{c}^{\mathrm{opt}}(\mu, \nu)$, then it holds $\int_I \int_{\mathbb{R}^d} \|v_t^\pi(x)\|^2 \, \mathrm{d}\mu_t(x) \, \mathrm{d}t = W_2^2(\mu, \nu)$.

Therefore, a neural network v^θ is trained to minimize the empirical loss $\mathcal{J}(\theta) \coloneqq J(v^\theta)$.

2.3 Wasserstein-on-Wasserstein

Our goal is to extend the flow matching framework to the *WoW space* $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, the space of *metameasures*. This space can be equipped with a meaningful geometry via (2). The following nested superposition result was proved in [Pinzi and Savaré, 2025].

Theorem 2.3. *Let $(M_t)_{t \in I} \in C_I(\mathcal{P}(\mathcal{P}(\mathbb{R}^d)))$ and $V : I \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ be a sufficiently smooth[3] Borel non-local vector field satisfying $\int_I \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|V(t, x, \mu)\| \, \mathrm{d}\mu(x) \, \mathrm{d}M_t(\mu) \, \mathrm{d}t < \infty$. Assume that (M_t, V_t) satisfies the continuity equation*

$$\partial_t M_t + \mathrm{div}_{\mathcal{P}}(V_t M_t) = 0 \tag{4}$$

in the sense of duality with cylindric functions, see Definition A.1 in Appendix A. Then, for M_0-a.e. $\mu_0 \in \mathcal{P}(\mathbb{R}^d)$ the collection of ODEs, indexed by $x \in \mathrm{supp}\, \mu_0$,

$$\dot{\gamma}(t, x, \mu_0) = V(t, \gamma(t, x, \mu_0), \gamma_{t,\mu_0,\sharp} \mu_0), \ t \in (0, 1), \ \gamma(0, x, \mu_0) = x, \tag{5}$$

admits a unique collection of solutions $\gamma(\cdot, x, \mu_0)$ on I, where we set $\gamma_{t,\mu_0} : \mathrm{supp}\, \mu_0 \to \mathbb{R}^d$, $x \mapsto \gamma(t, x, \mu_0)$. It holds that for M_0-a.e. μ_0, the curve $\mu_t \coloneqq \gamma_{t,\mu_0,\sharp} \mu_0$ solves $\partial_t \mu_t + \mathrm{div}(V_t(\cdot, \mu_t)\mu_t) = 0$ in the sense of distributions. Furthermore, with $\Gamma_t : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d)$, $\mu \mapsto \gamma_{t,\mu,\sharp} \mu$, we have $M_t = \Gamma_{t,\sharp} M_0$.

[2] For concrete regularity conditions, see [Ambrosio et al., 2005, Proposition 8.1.8]
[3] For details, see [Pinzi and Savaré, 2025, Theorem 6.2]

Once such a non-local vector field $V : I \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ is known, we can transfer a sample μ_0 from $\boldsymbol{\mu} = M_0$ to one of $\boldsymbol{\nu} = M_1$ by solving the above system of ODEs (5) starting in a collection of samples $x \sim \mu_0$. Importantly, during integration, each single curve $\gamma(t_n, x, \mu_0)$ at time step t_n depends (non-locally) on the collective information of all curves $\gamma_{t_{n-1},\mu_0,\sharp}\mu_0$ at the prior time step. In order to formulate a result on how to approximate the non-local vector field similar to Theorem 2.2, we fix $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ and introduce the set of *random couplings* by

$$\mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu}) \coloneqq \{\boldsymbol{\Pi} \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)) : \mathbf{Proj}^0_\sharp \boldsymbol{\Pi} = \boldsymbol{\mu},\, \mathbf{Proj}^1_\sharp \boldsymbol{\Pi} = \boldsymbol{\nu}\}$$

where $\mathbf{Proj}^i : \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \to \mathcal{P}(\mathbb{R}^d),\ \pi \mapsto \mathrm{proj}^i_\sharp \pi$. Such random couplings are easily constructed in a two-step procedure based on a given outer plan and inner plan selection, i.e., mappings $\mathrm{OP} : (\boldsymbol{\mu}, \boldsymbol{\nu}) \mapsto \mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}} \in \mathrm{C}(\boldsymbol{\mu}, \boldsymbol{\nu})$, and $\mathrm{IP} : (\mu, \nu) \mapsto \mathrm{IP}_{\mu,\nu} \in \mathrm{c}(\mu, \nu)$.
This procedure as well as the connection to the WoW distance (2) is described in the following lemma, extending the results of [Pinzi and Savaré, 2025].

Lemma 2.4. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$. Then, for any $\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}} \in C(\boldsymbol{\mu}, \boldsymbol{\nu})$ and any $\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}$-measurable selection $\mathrm{IP} : (\mu, \nu) \mapsto \mathrm{IP}_{\mu,\nu} \in c(\mu, \nu)$, it holds that $\boldsymbol{\Pi} \coloneqq \mathrm{IP}_\sharp\, \mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$. Furthermore, the WoW distance* (2) *satisfies*

$$\mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) = \inf_{\boldsymbol{\Pi} \in \mathrm{RC}(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^2 \,\mathrm{d}\pi(x, x') \,\mathrm{d}\boldsymbol{\Pi}(\pi).$$

Now, we construct a metameasure curve $(M_t)_{t\in[0,1]}$ in $\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ that connects given metameasures $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, and the corresponding non-local vector field $V(t, x, \mu)$. In particular, we need to make sure that (M_t, V_t) fulfills the WoW continuity equation (4). To this aim, we employ the established flow matching framework similar to Theorem 2.2, for the proof see Appendix A.

Theorem 2.5. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ and $\boldsymbol{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$, and define the curve $(M_t)_{t\in[0,1]}$ by $M_t \coloneqq (\mathbf{Proj}^t_\sharp \boldsymbol{\Pi})$, $\mathbf{Proj}^t \pi \coloneqq \mathrm{proj}^t_\sharp \pi$. Then, $\boldsymbol{\Pi} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d))$, i.e. it holds the integrability*

$$\int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x\|^2 + \|x'\|^2 \,\mathrm{d}\pi(x, x') \,\mathrm{d}\boldsymbol{\Pi}(\pi) < \infty. \tag{6}$$

Furthermore, $(M_t)_{t\in[0,1]} \in AC_I(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)))$ and the functional

$$J(V) \coloneqq \int_I \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|V(t, \mathrm{proj}^t(x, x'), \mathbf{Proj}^t(\pi)) - (x' - x)\|^2 \,\mathrm{d}\pi(x, x') \,\mathrm{d}\boldsymbol{\Pi}(\pi) \,\mathrm{d}t \tag{7}$$

admits a unique minimizer V_t, such that (M_t, V_t) fulfills the continuity equation (4)*. Moreover, the kinetic energy of V satisfies*

$$\mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \le \int_I \int \int \|V(t, x, \mu)\|^2 \,\mathrm{d}\mu(x) \,\mathrm{d}M_t(\mu) \,\mathrm{d}t \le \int \int \|x' - x\|^2 \,\mathrm{d}\pi(x, x') \,\mathrm{d}\boldsymbol{\Pi}(\pi), \tag{8}$$

with equality in both bounds if $\boldsymbol{\Pi} = \mathrm{IP}^*_\sharp\, \mathrm{OP}^*$ with $\mathrm{OP}^* \in \mathrm{C}^{\mathrm{opt}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathrm{IP}^*_{\mu,\nu} \in \mathrm{c}^{\mathrm{opt}}(\mu, \nu)$.

Together with Lemma 2.4, we can construct metameasure curves between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. By (8), cost-efficient couplings guarantee low kinetic energy, easing numerical integration. The *Wasserstein flow matching* approach [Haviv et al., 2025] is recovered for appropriate IP and OP (Section 3.1).

3 Wasserstein-on-Wasserstein Flow Matching

Theorem 2.5 yields a principled training objective for learning the WoW velocity field similar to the classical Euclidean flow matching approach. However, enabling the generation of arbitrary, non-parametric probability measures requires discretizing them. Here, we employ a scalable free-support approach [Lindheim, 2023] that uses fixed uniform weights for parametric locations, as opposed to fixed-support approaches with parametric weights for fixed locations [Lin et al., 2020] or approaches based on a parametric family of distributions [Jiang et al., 2025].
We aim to parametrize our velocity field via a network $V_\theta : [0, 1] \times \mathbb{R}^{d\times N} \to \mathbb{R}^{d\times N}$, $N \in \{\underline{N}, \dots, \overline{N}\}$, where V_θ is potentially able to process data for non-fixed N, and the network is permutation-equivariant (or S_N-equivariant), i.e., $V_\theta(t, P_\sigma \mathbf{x}) = P_\sigma\, V_\theta(t, \mathbf{x})$ for all $\sigma \in S_N$. Here, $\mathbf{x} = (x_1, \dots, x_N)$, S_N is the group of N-permutations and P_σ is the permutation matrix associated with σ. This corresponds to encoding empirical measures as elements of the quotient

space $\mathbb{R}^{d\times N}/S_N$ composed of equivalence classes of permutable vectors [Boumal, 2023, Ch. 9]. Denoting equality up to reparametrization via $\simeq$, we obtain the parametrization

$$[\mathbf{x}] \in \mathbb{R}^{d\times N}/S_N \simeq \hat{\mu} = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_j} \in \mathcal{P}_2^N(\mathbb{R}^d), \quad V_\theta(t, x_i, \hat{\mu}) \in \mathbb{R}^d \simeq V_\theta(t, \mathbf{x})_i \in \mathbb{R}^d. \quad (9)$$

Hence, the network predicts the velocity of particle $x_i \in \mathbb{R}^d$ as $V_\theta(t, \mathbf{x})_i \in \mathbb{R}^d$ at time t based on a vector-encoded $\hat{\mu}$. Throughout, $\|\cdot\|$ on $\mathbb{R}^{d\times N}$ denotes the normalized Frobenius norm $\|\mathbf{x}\|^2 := \frac{1}{N}\sum_{j=1}^{N}\|x_j\|^2$. Notably, natural choices for the parametrization are transformer architectures without positional encodings [Guo et al., 2021, Lee et al., 2019] or a shared feedforward network for all points in combination with global information [Qi et al., 2017].

Given such a parametrization, our training objective is then the minimization of the loss $\mathcal{L}_{\text{WoW-FM}}(\theta) = J(V_\theta)$ in (7). Since the actual integrals appearing in this loss are intractable, we follow the standard flow matching procedure and train the network using a stochastic Monte Carlo approximation $\widehat{\mathcal{L}}_{\text{WoW-FM}}(\theta)$, where we rely on two batches of B source and target measures with (potentially random) discretization N. Given estimators for the outer plan via $\widehat{\text{OP}}$ and the inner plans via $\widehat{\text{IP}}$, we compute a Monte Carlo loss estimator for every gradient step via $t \sim \text{Unif}[0,1]$, $N \sim \text{Unif}(\{\underline{N}, \dots, \overline{N}\})$, $\hat{\boldsymbol{\mu}} = \frac{1}{B}\sum_{i=1}^{B}\delta_{\hat{\mu}_i}$, $\hat{\boldsymbol{\nu}} = \frac{1}{B}\sum_{i'=1}^{B}\delta_{\hat{\nu}_{i'}}$ and

$$\{(\hat{\mu}_b, \hat{\nu}_b)\}_{b=1}^{B} \overset{\text{iid}}{\sim} \widehat{\text{OP}}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}} = \sum_{i=1}^{B}\sum_{i'=1}^{B} \hat{\Pi}^{\text{OP}}_{ii'}\, \delta_{(\hat{\mu}_i, \hat{\nu}_{i'})}, \qquad \hat{\Pi}^{\text{OP}}_{ii'} \in \left[0, \frac{1}{B}\right],$$

$$(\mathbf{x}_b, \mathbf{x}'_b) = \left((x_{b,j}, x'_{b,j})\right)_{j=1}^{N} \overset{\text{iid}}{\sim} \widehat{\text{IP}}(\hat{\mu}_b, \hat{\nu}_b) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{\pi}^{\text{IP}}_{b,ii'}\, \delta_{(x_{b,i}, x'_{b,i'})}, \qquad \hat{\pi}^{\text{IP}}_{b,ii'} \in \left[0, \frac{1}{N}\right],$$

$$\widehat{\mathcal{L}}_{\text{WoW-FM}}(\theta) = \frac{1}{B}\sum_{b=1}^{B} \left\| V_\theta\big(t, \text{proj}^t(\mathbf{x}_b, \mathbf{x}'_b)\big) - (\mathbf{x}'_b - \mathbf{x}_b)\right\|^2.$$

Note that the augmented vector $\text{proj}^t(\mathbf{x}_b, \mathbf{x}'_b)$ represents $\text{proj}^t(x, x')$ and $\mathbf{Proj}^t(\pi)$ simultaneously, with $(\mathbf{x}_b, \mathbf{x}'_b)$ given by N i.i.d. draws from $\widehat{\text{IP}}(\hat{\mu}_b, \hat{\nu}_b)$. For permutation-valued inner plans this reduces to repeated sampling among the matched pairs $(x_i, x'_{\sigma(i)})$. Under the assumption of a parametric source sampling procedure for $\hat{\boldsymbol{\mu}}$ similar to the classical Gaussian sampling in generative modelling, we can produce new target samples by encoding $\hat{\mu} \sim \hat{\boldsymbol{\mu}}$ as a vector $\mathbf{x}_{\hat{\mu}}$ and numerically integrating the flow ODE (5) produced by V_θ. In particular, we may directly parametrize a discretized metameasure via a vectorized measure $\overrightarrow{\boldsymbol{\mu}} \simeq \hat{\boldsymbol{\mu}}$, i.e., $\mathbf{x}_{\hat{\mu}} \sim \overrightarrow{\boldsymbol{\mu}} \in \mathcal{P}_2(\mathbb{R}^{d\times N})$.

Importantly, any admissible choice for our transport plan estimators would result in a valid flow, mirroring the flexibility of Euclidean flow matching [Pooladian et al., 2023, Wald and Steidl, 2025]. This design space contains standard point-cloud WFM $(\widehat{\text{OP}}_{\text{ind}}, \widehat{\text{IP}}_{\text{W}})$ [Haviv et al., 2025], multi-sample Gaussian/Bures matching $(\widehat{\text{OP}}_{\text{BW}}, \widehat{\text{IP}}_{\text{W}})$, and flattened Euclidean minibatch OT. The latter tests reordering alone. W, SW, and LLW instead use permutation-invariant cloud geometry.

3.1 Constructing Transport Plans

Even though flow matching is theoretically valid for any transport plan, the specific choice affects numerical performance. Besides the standard independent coupling $\mu \otimes \nu$, a common alternative is the Wasserstein coupling $\pi^* = \arg\min_{\pi\in\text{c}(\mu,\nu)} \int \|x - x'\|^2 \,\text{d}\pi$ [Chemseddine et al., 2025, Tong et al., 2024]. By construction, the latter approximates the optimal-transport straight-line interpolation and often yields more stable ODE integration, at the cost of more expensive training due to the complexity of OT solvers. For nested couplings, (8) makes this precise: the kinetic energy of the induced velocity field is at most the coupling cost.

This issue is even more pronounced in our setting, where both inner and outer transport plans must be chosen. While our framework allows for straight-line WoW interpolations for metameasures via (8), approximating the loss requires solving the (batched) WoW problem (2) at each step. For batch size B, this requires B^2 Euclidean Wasserstein distances and solving a linear program of complexity $\mathcal{O}(B^3)$. With exact Wasserstein solvers of complexity $\mathcal{O}(N^3)$ for discretization size N, this yields an overall complexity of $\mathcal{O}(B^2N^3 + B^3)$. Even for Sinkhorn-based solvers, the resulting $\mathcal{O}(B^2N^2\log N + B^2\log B)$ complexity remains substantial. As an alternative to WoW solvers, we therefore consider practical alternative constructions for inner and outer transport plans.

Independent Couplings. The simplest admissible choice is given by independent couplings based on the product measure. At the outer level, this corresponds to $\widehat{\mathrm{OP}}_{\mathrm{ind}} = \hat{\boldsymbol{\mu}} \otimes \hat{\boldsymbol{\nu}}$ and at the inner level to $\widehat{\mathrm{IP}}_{\mathrm{ind}}(\hat{\mu}, \hat{\nu}) = \hat{\mu} \otimes \hat{\nu}$. This choice is computationally trivial, but may lead to curved and unstable interpolations. Choosing $\widehat{\mathrm{IP}}_{\mathrm{ind}}$ corresponds to standard Euclidean flow matching on randomly ordered point clouds with permutation-based data augmentation, see [Hui et al., 2025].

Wasserstein Couplings. A straightforward alternative is the classical use of standard OT solvers. For the outer plan, this means choosing $\widehat{\mathrm{OP}}_{\mathbf{W}}$ according to $\arg\min_{\mathrm{OP}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}} \in \mathrm{C}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}})} \int \mathrm{W}_2^2(\mu,\nu)\, \mathrm{dOP}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}}$ based on a precomputed $B \times B$ Euclidean Wasserstein cost matrix. As the inner coupling, we set $\widehat{\mathrm{IP}}_{\mathrm{W}}(\hat{\mu}, \hat{\nu})$ according to $\arg\min_{\mathrm{IP}_{\hat{\mu},\hat{\nu}} \in \mathrm{c}(\hat{\mu},\hat{\nu})} \int \|x - x'\|^2 \, \mathrm{dIP}_{\hat{\mu},\hat{\nu}}$. Combining $\widehat{\mathrm{OP}}_{\mathbf{W}}$ and $\widehat{\mathrm{IP}}_{\mathrm{W}}$ allows for approximating the dynamic WoW formulation, see Theorem 2.5. Inner Wasserstein couplings appear in S_N-equivariant flow matching [Klein et al., 2023] for point clouds [Hui et al., 2025] and the choice of $\widehat{\mathrm{IP}}_{\mathrm{W}}$ and $\widehat{\mathrm{OP}}_{\mathrm{ind}}$ reflects the training setup of Wasserstein flow matching [Haviv et al., 2025].

Euclidean Minibatch Coupling. This performs outer minibatch OT using squared Euclidean costs between vectorized point clouds, followed by exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$. While order-dependent, its expected cost under independent random orderings is the squared mean distance plus both variances.

Bures–Wasserstein Coupling. Let $\hat{m}_\mu$ and $\hat{\Sigma}_\mu$ denote the empirical mean and covariance of $\hat{\mu}$. Outer minibatch OT uses $c_{\mathrm{BW}}(\hat{\mu}, \hat{\nu}) = \|\hat{m}_\mu - \hat{m}_\nu\|^2 + \mathrm{tr}(\hat{\Sigma}_\mu + \hat{\Sigma}_\nu - 2(\hat{\Sigma}_\mu^{1/2}\hat{\Sigma}_\nu\hat{\Sigma}_\mu^{1/2})^{1/2})$, followed by exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$. Thus, outer matching only considers the first two moments.

Sliced Wasserstein Couplings. Since the computational burden of $\widehat{\mathrm{OP}}_{\mathbf{W}}$ is mainly the result of the cost matrix computation, we can simply replace it via a related, but more efficient distance estimator. A popular Wasserstein alternative is the sliced Wasserstein distance [Bonnotte, 2013] given via $\mathrm{SW}_2^2(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_2^2\big(\mathrm{proj}^\theta_\#\mu, \mathrm{proj}^\theta_\#\nu\big)\, \mathrm{d}S(\theta)$, where $\mathrm{proj}^\theta(x) = \langle \theta, x \rangle$, and S is the uniform probability measure on $\mathbb{S}^{d-1}$. Now, we choose $\widehat{\mathrm{OP}}_{\mathrm{SW}}$ according to $\arg\min_{\mathrm{OP}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}} \in \mathrm{C}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}})} \int \mathrm{SW}_2^2(\mu,\nu)\ \mathrm{d}\,\mathrm{OP}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}}$ based on a $B \times B$ sliced Wasserstein cost matrix [Piening and Beinert, 2025]. Since each $\theta \in \mathbb{S}^{d-1}$ gives a 1D transport plan between projected measures, canonically lifting to a plan $\mathrm{IP}_\theta(\hat{\mu}, \hat{\nu}) \in \mathrm{c}(\hat{\mu}, \hat{\nu})$, we can similarly employ inner sliced transport plans $\mathrm{IP}_{\mathrm{SW}}(\hat{\mu}, \hat{\nu}) = \int_{\mathbb{S}^{d-1}} \mathrm{IP}_\theta(\hat{\mu}, \hat{\nu})\, \mathrm{d}S(\theta)$.

Lazy Linear Wasserstein Couplings. Alternatively, linearized Wasserstein distances have been introduced precisely to accelerate pairwise Wasserstein distance calculation [Wang et al., 2013], as needed for computing $\widehat{\mathrm{OP}}_{\mathbf{W}}$. Focusing on our use case of empirical measures with N fixed, this technique replaces the Wasserstein distance with the linearized divergence

$$\mathrm{LW}_2^2(\hat{\mu}, \hat{\nu}) = \int \|x - x'\|^2 \ \mathrm{d}\big((\mathrm{id}, T^{-1}_{\hat{\nu},\hat{\rho}} \circ T_{\hat{\mu},\hat{\rho}})_\sharp \hat{\mu}\big)(x, x') = \|P_{\mathbf{x}_{\hat{\mu}},\mathbf{x}_{\hat{\rho}}}\mathbf{x}_{\hat{\mu}} - P_{\mathbf{x}_{\hat{\nu}},\mathbf{x}_{\hat{\rho}}}\mathbf{x}_{\hat{\nu}}\|^2 \quad (10)$$

based on precomputed transport maps $T_{\hat{\mu},\hat{\rho}}, T_{\hat{\nu},\hat{\rho}} : \mathbb{R}^d \to \mathbb{R}^d$ and permutation matrices $P_{\mathbf{x}_{\hat{\mu}},\mathbf{x}_{\hat{\rho}}}, P_{\mathbf{x}_{\hat{\nu}},\mathbf{x}_{\hat{\rho}}} \in \mathbb{R}^{N \times N}$ characterized via

$$\mathrm{W}_2^2(\hat{\mu}, \hat{\rho}) = \int \|x - x'\|^2 \ \mathrm{d}\big((\mathrm{id}, T_{\hat{\mu},\hat{\rho}})_\sharp \hat{\mu}\big)(x, x') = \|\mathbf{x}_{\hat{\rho}} - P_{\mathbf{x}_{\hat{\mu}},\mathbf{x}_{\hat{\rho}}}\mathbf{x}_{\hat{\mu}}\|^2,$$
$$\mathrm{W}_2^2(\hat{\nu}, \hat{\rho}) = \int \|x - x'\|^2 \ \mathrm{d}\big((\mathrm{id}, T_{\hat{\nu},\hat{\rho}})_\sharp \hat{\nu}\big)(x, x') = \|\mathbf{x}_{\hat{\rho}} - P_{\mathbf{x}_{\hat{\nu}},\mathbf{x}_{\hat{\rho}}}\mathbf{x}_{\hat{\nu}}\|^2.$$

Importantly, the existence of such maps that (i) realize the discrete Wasserstein distance and (ii) are represented as invertible permutations, as well as the vector representations $\mathbf{x}_{\hat{\mu}_b}$, $\mathbf{x}_{\hat{\nu}_{b'}}$, and $\mathbf{x}_{\hat{\rho}} \in \mathbb{R}^{d \times N}$, relies on the formulation of empirical measures $\hat{\mu}_b$, $\hat{\nu}_{b'}$, and $\hat{\rho} \in \mathcal{P}_2^N(\mathbb{R}^d)$ of fixed size as in (9). While linearization allows efficient approximation of pairwise Wasserstein distances for a suitably chosen reference $\hat{\rho}$, employing (10) as an inner plan estimator would require solving two OT problems per gradient step: Instead of the single solve required by $\widehat{\mathrm{IP}}_{\mathrm{W}}$, we would solve for $T_{\hat{\mu},\hat{\rho}}$ and $T_{\hat{\nu},\hat{\rho}}$. To avoid this overhead, we instead exploit the local stability of discrete OT maps to construct a lazy linearization strategy based on a suitable source metameasure $\hat{\boldsymbol{\mu}}$.

Proposition 3.1. *Given the mapping* $\mathrm{Vec2M} : \mathbb{R}^{d\times N} \to \mathcal{P}_2^N(\mathbb{R}^d), \quad \mathbf{x} \mapsto \hat{\mu} = \frac{1}{N}\sum_{j=1}^N \delta_{x_j}$, *a reference measure* $\hat{\rho} \in \mathcal{P}_2^N(\mathbb{R}^d)$*, and some* $\mathbf{x}_{\hat{\rho}} \in \mathbb{R}^{d\times N}$ *with* $\mathbf{x}_{\hat{\rho},i} \neq \mathbf{x}_{\hat{\rho},i'}$ *for all* $i \neq i'$ *and* $\mathrm{Vec2M}(\mathbf{x}_{\hat{\rho}}) = \hat{\rho}$*, then for (Lebesgue-)almost every* $\mathbf{x}' \in \mathbb{R}^{d\times N}$*, there exists* $R_{\mathbf{x}'} > 0$ *such that for all* $\mathbf{x}$ *with* $\|\mathbf{x} - \mathbf{x}_{\hat{\rho}}\| < R_{\mathbf{x}'}$*, it holds that*

$$\mathrm{W}_2^2(\mathrm{Vec2M}(\mathbf{x}), \mathrm{Vec2M}(\mathbf{x}')) = \|\mathbf{x} - P_{\mathrm{Vec2M}(\mathbf{x}'),\mathbf{x}_{\hat{\rho}}}\mathbf{x}'\|^2.$$

The elementary proof is contained in Section A and is based on the fact that all Wasserstein distances in $\mathcal{P}_2^N(\mathbb{R}^d)$ are realized via permutations [Peyré and Cuturi, 2019]. Since the set of permutations is finite, the optimal permutation is almost everywhere unique and therefore locally stable around the reference. This statement implies a simple corollary motivating an accelerated linear OT regime. Note that it follows directly from Proposition 3.1, since the minimum of B positive constants is again a positive constant.

Corollary 3.2. *Under the setting of Proposition 3.1, fix a right-inverse* $\mathrm{M2Vec} : \mathcal{P}_2^N(\mathbb{R}^d) \to \mathbb{R}^{d\times N}$ *of* Vec2M*, i.e.,* $\mathrm{Vec2M} \circ \mathrm{M2Vec} = \mathrm{id}$*, and define the* $\hat{\rho}$*-alignment map* $\Phi_{\hat{\rho}} : \mathbb{R}^{d\times N} \to \mathbb{R}^{d\times N}$, $\Phi_{\hat{\rho}}(\mathbf{x}) := P_{\mathbf{x},\mathbf{x}_{\hat{\rho}}}\,\mathbf{x}$*, which re-orders* $\mathbf{x}$ *so that its entries are optimally matched against* $\mathbf{x}_{\hat{\rho}} := \mathrm{M2Vec}(\hat{\rho})$*. Then, for almost every target metameasure* $\hat{\boldsymbol{\nu}} = \frac{1}{B}\sum_{i=1}^B \delta_{\frac{1}{N}\sum_{j=1}^N \delta_{x_{i,j}}} \in \mathcal{P}_2(\mathcal{P}_2^N(\mathbb{R}^d))$ *(in the Lebesgue sense on* $\mathbb{R}^{d\times N\times B}$*), there exists* $R_{\hat{\boldsymbol{\nu}}} > 0$ *such that whenever* $\mathrm{supp}(\mathrm{M2Vec}_\sharp\,\hat{\boldsymbol{\mu}}) \subset B(\mathbf{x}_{\hat{\rho}}, R_{\hat{\boldsymbol{\nu}}}) := \{\mathbf{x} \in \mathbb{R}^{d\times N} : \|\mathbf{x} - \mathbf{x}_{\hat{\rho}}\| \leq R_{\hat{\boldsymbol{\nu}}}\}$*, it holds*

$$\mathbf{W}_2^2(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\nu}}) = \mathrm{W}_2^2\big(\mathrm{M2Vec}_\sharp\,\hat{\boldsymbol{\mu}},\ (\Phi_{\hat{\rho}})_\sharp\,\mathrm{M2Vec}_\sharp\,\hat{\boldsymbol{\nu}}\big).$$

This statement applies directly to the practical setting, where $\hat{\boldsymbol{\nu}}$ is our full training dataset. It says that for a source metameasure obtained by sampling pre-ordered vectors from $\mathrm{M2Vec}_\sharp\,\hat{\boldsymbol{\mu}}$ centred around $\mathrm{M2Vec}(\hat{\rho})$ (in the Euclidean sense), we *only* need to permute our empirical measure training samples with respect to $\mathrm{M2Vec}(\hat{\rho})$, once before training. Notably, $\mathrm{M2Vec}_\sharp\,\hat{\boldsymbol{\mu}}$ simply describes sampling via a vectorized $\overrightarrow{\boldsymbol{\mu}} \in \mathcal{P}_2(\mathbb{R}^{d\times N})$ and $\mathbf{x}_{\hat{\rho}} = \mathrm{M2Vec}(\hat{\rho}) \simeq \hat{\rho}$ as described in Section 3.
Once we employ this discretization for $\overrightarrow{\boldsymbol{\mu}}$ with $\mathrm{supp}(\overrightarrow{\boldsymbol{\mu}}) \subset B(\mathbf{x}_{\hat{\rho}}, R_{\hat{\boldsymbol{\nu}}})$, we are in a classical Euclidean flow matching regime on $\mathbb{R}^{d\times N}$. If we choose the standard independent coupling here, this corresponds to the Wasserstein flow matching regime of employing $\widehat{\mathrm{OP}}_{\mathbf{ind}}$ and $\widehat{\mathrm{IP}}_{\mathrm{W}}$. Opting for classical OT flow matching here, results in approximating the WoW flow via $\widehat{\mathrm{OP}}_{\mathbf{W}}$ and $\widehat{\mathrm{IP}}_{\mathrm{W}}$.
Inspired by this observation, we design our lazy linear plans. Based on a reference measure $\hat{\rho} \in \mathcal{P}_2^N(\mathbb{R}^d)$ with vectorization $\mathbf{x}_{\hat{\rho}} \simeq \hat{\rho}$ satisfying $\mathbf{x}_{\hat{\rho},i} \neq \mathbf{x}_{\hat{\rho},i'}$ for $i \neq i'$, and a reference metameasure $\hat{\boldsymbol{\mu}}_{\hat{\rho}} := \mathrm{Vec2M}_\sharp\,\overrightarrow{\boldsymbol{\mu}}$ with $\overrightarrow{\boldsymbol{\mu}} \in \mathcal{P}_2(\mathbb{R}^{d\times N})$ heavily centred around $\mathbf{x}_{\hat{\rho}}$, we introduce our lazy linear plan and divergence as

$$\widehat{\mathrm{IP}}_{\mathrm{LLW}}(\hat{\mu}, \hat{\nu}) := \frac{1}{N}\sum_{j=1}^N \delta_{((\mathbf{x}_{\hat{\mu}})_j,\,(P_{\mathbf{x}_{\hat{\nu}},\mathbf{x}_{\hat{\rho}}}\mathbf{x}_{\hat{\nu}})_j)} \in \mathrm{c}(\hat{\mu}, \hat{\nu}),$$

$$\mathrm{LLW}_2^2(\hat{\mu}, \hat{\nu}) := \int \|x - x'\|^2 \,\mathrm{d}\widehat{\mathrm{IP}}_{\mathrm{LLW}}(\hat{\mu}, \hat{\nu})(x, x') = \|\mathbf{x}_{\hat{\mu}} - P_{\mathbf{x}_{\hat{\nu}},\mathbf{x}_{\hat{\rho}}}\,\mathbf{x}_{\hat{\nu}}\|^2,$$

where again $P_{\mathbf{x}_{\hat{\nu}},\mathbf{x}_{\hat{\rho}}}$ is the permutation aligning $\mathbf{x}_{\hat{\nu}}$ to $\mathbf{x}_{\hat{\rho}}$, and the source vector $\mathbf{x}_{\hat{\mu}}$ is already ordered like $\mathbf{x}_{\hat{\rho}}$ by construction of $\overrightarrow{\boldsymbol{\mu}}$. Crucially, this permutation depends only on $\hat{\rho}$ and $\hat{\nu}$ and not on $\hat{\mu}$. Thus, it can be precomputed once per sample $\hat{\nu}$ before training.
Due to Corollary 3.2, we know that $\mathrm{LLW}_2^2(\hat{\mu}, \hat{\nu}) = \mathrm{W}_2^2(\hat{\mu}, \hat{\nu})$ for $\hat{\boldsymbol{\mu}}_{\hat{\rho}} := \mathrm{Vec2M}_\sharp\,\overrightarrow{\boldsymbol{\mu}}$ with 'high' probability if $\overrightarrow{\boldsymbol{\mu}}$ is sufficiently centred around $\mathbf{x}_{\hat{\rho}}$. Now, we may use this to approximate WoW transport plans via optimal outer and inner plans. Consequently for suitably concentrated $\hat{\mu}_i \sim \hat{\boldsymbol{\mu}}_{\hat{\rho}}$ with randomly batched $\hat{\boldsymbol{\mu}} = 1/B\sum_{i=1}^B \delta_{\hat{\mu}_i}$, we set our transport plan estimator $\widehat{\mathrm{OP}}_{\mathbf{LLW}}$ according to $\arg\min_{\mathrm{OP}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}} \in \mathrm{C}(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}})} \int \mathrm{LLW}_2^2(\mu, \nu)\,\mathrm{dOP}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}}$ and use the inner plan $\widehat{\mathrm{IP}}_{\mathrm{LLW}}(\hat{\mu}, \hat{\nu})$ defined above. Unlike previous estimators, the lazy linear estimator quality depends on the source distribution, see Appendix C.

4 Experiments

Note that our general approach already covers the Wasserstein flow matching model [Haviv et al., 2025], which has recently demonstrated SOTA performance in point cloud generation. Throughout our experiments, we therefore focus on explicit 'apples-to-apples' comparison: we aim to compare our couplings based on a fixed architecture and training setup. Mainly presenting experiments with fixed N, a multi-resolution experiment is deferred to Appendix B and ablations to Appendix C.

4.1 Implementation and Evaluation

Networks. To reduce any bias induced by a specific architecture, we employ two distinct architectural choices. As a simple **baseline model**, we employ a shared MLP_θ conditioned on per-point local features $F_\ell(\mathbf{x})_i$ [Piening and Beinert, 2026a, Sato et al., 2020], i.e., sorted pairwise distances and neighbour positions, and global features $F_g(\mathbf{x})$, i.e., mean and covariance of $(\mathbf{x}_1, \dots, \mathbf{x}_N)$, together with a sinusoidal time embedding $\gamma(t)$, followed by a global self-attention layer A_θ. This results in $V_\theta(t, \mathbf{x}_t) = \mathrm{A}_\theta(M_1, \dots, M_N)$ with $M_i = \mathrm{MLP}_\theta((\mathbf{x}_t)_i, F_\ell(\mathbf{x}_t)_i, F_g(\mathbf{x}_t), \gamma(t))$. As a more standard architecture, we additionally employ a **transformer** [Vaswani et al., 2017] without positional encodings. As highlighted in Section 3, both architectures are S_N-equivariant. More details can be found in Appendix E.

Reference Measure. Similarly to most existing linearized OT methods [Beier et al., 2022, Nguyen and Tsuda, 2023, Sarrazin and Schmitzer, 2024, Wang et al., 2013], we rely on established solvers [Flamary et al., 2021] to approximate the (finite) barycenter $\hat{\rho} \in \arg\min_{\rho \in \mathcal{P}_2^N(\mathbb{R}^d)} \sum_{b=1}^{B_{\mathrm{ref}}} \mathrm{W}_2^2(\rho, \hat{\nu}_b)$, where $\hat{\nu}_b \sim \hat{\boldsymbol{\nu}}$, and we generally employ $B_{\mathrm{ref}} = 8$.

Source Metameasure. Due to the importance of the latent source for flow-based models [Chemseddine et al., 2026, Hagemann and Neumayer, 2021, Pandey et al., 2025], we investigate three potential choices for our source metameasure $\hat{\boldsymbol{\mu}}$ parametrized via $\overrightarrow{\boldsymbol{\mu}}$. As i) a baseline **pure noise metameasure**, we employ independent and identically distributed (i.i.d.) entries $\mathbf{x}_{i,j} \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, N$ and $j = 1, \dots, d$, for $\mathbf{x} \sim \overrightarrow{\boldsymbol{\mu}}$. Here, $\sigma \sim \mathrm{Unif}[\underline{\sigma}, \overline{\sigma}]$ is sampled independently per cloud, see supplementary Experiment C.1. However, to enable our lazy linear approach, we theoretically require concentration around $\mathbf{x}_{\hat{\rho}}$, see Corollary 3.2. Therefore, we alternatively use ii) our **barycentric noise metameasure** $\overrightarrow{\boldsymbol{\mu}}$. In the MNIST and ShapeNet implementations, its entries are sampled as $\mathbf{x}_{i,j} = (\mathbf{x}_{\hat{\rho}})_{i,j} + \sigma_0 Z_{i,j} U_{i,j}$, where $Z_{i,j} \sim \mathcal{N}(0, 1)$ and $U_{i,j} \sim \mathrm{Unif}[0.5, 1.5]$ independently. Thus, every coordinate receives Gaussian noise with an independently rescaled standard deviation. See Experiments 4.3 and 4.4. In line with neural OT [Kim et al., 2024, Korotin et al., 2023], we further employ iii) **empirical metameasures** to investigate unsupervised couplings between data distributions, see Experiments 4.2 and B.1.

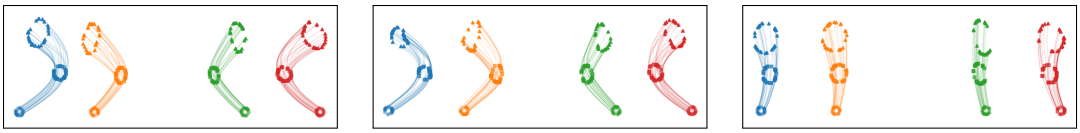
Evaluation. Since we primarily focus on the case $d = 2$ or $d = 3$, we follow the evaluation protocol in point cloud generation [Haviv et al., 2025, Hui et al., 2025, Yang et al., 2019]. Thus, we are mainly interested in the accuracy of a nearest neighbour classifier based on the Chamfer ('Chamfer-NNA') or the Wasserstein distance ('OT-NNA'). Here, we generate $\{\mathbf{x}^b\}_{b=1}^B$ point clouds and compare them to $\{\mathbf{x}^b\}_{b=B+1}^{2B}$ real test set point clouds. Based on the pairwise distance matrix $(d(\mathbf{x}^b, \mathbf{x}^{b'}))_{b,b'=1}^{2B}$, we then predict whether $\mathbf{x}^b$ is generated or real based on the labels of all other point clouds. Averaging results for all $b = 1, \dots, 2B$ gives the nearest neighbour accuracy, where Chamfer-NNA is obtained via the Chamfer distance $d(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^N \min_j \|\mathbf{x}_i - \mathbf{x}'_j\|^2 + \sum_{i=1}^N \min_j \|\mathbf{x}'_i - \mathbf{x}_j\|^2$ and OT-NNA via $d(\mathbf{x}, \mathbf{x}') = \mathrm{W}_2^2(\mathrm{Vec2M}(\mathbf{x}), \mathrm{Vec2M}(\mathbf{x}'))$. The best NNA is 0.5, and the worst one is 1.0.
We evaluate each NNA with 512 generated and 512 real inputs and report the average NNA and its standard deviation over 5 repetitions. Moreover, we perform this evaluation with 5, 25 and 125 Euler steps for the generation to investigate the stability of numerical ODE integration.

4.2 Circular Point Clouds (2D, Baseline Model, Resolution: 30)

We experiment on a synthetic 2D metameasure transport task to illustrate the importance of the outer coupling. Circular source clouds $\hat{\mu} \sim \hat{\boldsymbol{\mu}}$ are drawn as $\hat{\mu} = \frac{1}{30} \sum_{i=1}^{30} \delta_{(h,\,0) + r_s\, \varepsilon_i / \|\varepsilon_i\|}$, for i.i.d. $\varepsilon_i \sim \mathcal{N}(0, I_2)$ and $h \sim \mathrm{Unif}[-20, 20]$, i.e., $N = 30$ points placed on a ring of radius $r_s = 0.5$ centred at the point $(h, 0)$. Target clouds $\hat{\nu} \sim \hat{\boldsymbol{\nu}}$ follow the same construction but with radius $r_t = 2.0$ and a vertical offset of 10, i.e., vertically translated circles with bigger radius. After $15{,}000$ steps (baseline model, Adam, step size $5 \cdot 10^{-4}$, $B = 8$), only the WoW coupling yields nearly straight paths, see Figure 1. Inner OT alone does not suffice.

4.3 MNIST Point Clouds (2D, Baseline Model, Resolution: 64)

Similar to [Haviv et al., 2025], we experiment on point clouds obtained from MNIST images [LeCun et al., 1998]. Therefore, we view the normalized pixel values of each MNIST image as the intensities of a 2D histogram on $[0, 1]^2$, allowing us to sample $N = 64$ points from it. This results in fixed-size point clouds encoded as $\mathbf{x} \in \mathbb{R}^{2 \times 64}$. We train our baseline model for 50 epochs using



(a) $(\widehat{\mathrm{OP}}_{\mathrm{ind}}, \widehat{\mathrm{IP}}_{\mathrm{ind}})$ (b) $(\widehat{\mathrm{OP}}_{\mathrm{ind}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ (c) $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$

Figure 1: Learned paths from source circles (bottom) to larger, offset targets (top), $h \in \{-20, -10, 10, 20\}$ by colour. Only the WoW coupling yields straight paths.

		Chamfer-NNA ↓			OT-NNA ↓			Coupling/step (ms)
OP	**IP**	Euler-5	Euler-25	Euler-125	Euler-5	Euler-25	Euler-125	
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{ind}}$	$0.94_{\pm 0.00}$	$0.70_{\pm 0.01}$	$0.65_{\pm 0.01}$	$0.95_{\pm 0.01}$	$0.65_{\pm 0.01}$	$0.61_{\pm 0.01}$	<0.1
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$0.77_{\pm 0.01}$	$0.69_{\pm 0.03}$	$0.66_{\pm 0.02}$	$0.71_{\pm 0.01}$	$0.64_{\pm 0.01}$	$0.61_{\pm 0.02}$	≈ 1.2
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{SW}}$	$0.79_{\pm 0.01}$	$\underline{0.63_{\pm 0.01}}$	$0.62_{\pm 0.01}$	$0.75_{\pm 0.01}$	$0.61_{\pm 0.01}$	$0.59_{\pm 0.02}$	≈ 0.4
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{LLW}}$	$0.75_{\pm 0.01}$	$0.64_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.69_{\pm 0.01}$	$\underline{0.59_{\pm 0.01}}$	$0.58_{\pm 0.02}$	<0.1
$\widehat{\mathrm{OP}}_{\mathbf{W}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$\underline{0.72_{\pm 0.01}}$	$0.65_{\pm 0.02}$	$\underline{0.62_{\pm 0.02}}$	$0.66_{\pm 0.01}$	$\mathbf{0.59_{\pm 0.03}}$	$\underline{0.58_{\pm 0.01}}$	≈ 13.7
$\widehat{\mathrm{OP}}_{\mathrm{BW}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$0.73_{\pm 0.01}$	$0.66_{\pm 0.01}$	$0.65_{\pm 0.01}$	$\underline{0.65_{\pm 0.01}}$	$0.61_{\pm 0.01}$	$0.60_{\pm 0.02}$	≈ 5.5
Euclidean	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$0.74_{\pm 0.00}$	$0.64_{\pm 0.02}$	$\mathbf{0.62_{\pm 0.01}}$	$0.68_{\pm 0.01}$	$0.60_{\pm 0.03}$	$0.59_{\pm 0.01}$	≈ 1.6
$\widehat{\mathrm{OP}}_{\mathbf{SW}}$	$\widehat{\mathrm{IP}}_{\mathrm{SW}}$	$0.79_{\pm 0.02}$	$0.65_{\pm 0.01}$	$0.64_{\pm 0.01}$	$0.76_{\pm 0.01}$	$0.63_{\pm 0.02}$	$0.61_{\pm 0.01}$	≈ 3.1
$\widehat{\mathrm{OP}}_{\mathbf{LLW}}$	$\widehat{\mathrm{IP}}_{\mathrm{LLW}}$	$\mathbf{0.72_{\pm 0.01}}$	$\mathbf{0.63_{\pm 0.02}}$	$0.62_{\pm 0.02}$	$\mathbf{0.65_{\pm 0.00}}$	$0.59_{\pm 0.02}$	$\mathbf{0.57_{\pm 0.02}}$	<0.1

Table 1: MNIST results for barycentric noise ($B = 8$, $N = 64$). Bold and underline mark the two best scores before rounding. Coupling times are CPU-timed per step, see Appendix C.6.

Adam (step size: $5 \cdot 10^{-4}$). Evaluating our barycentric noise source with base scale $\sigma_0 = 0.1$ and coordinate-wise factors $U_{i,j} \in [0.5, 1.5]$, the results displayed in Table 1 clearly show performance gains from OT-based plans on both levels. Even for this small dataset, we see a computational explosion of the coupling time from $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$. In contrast, our linearized $(\widehat{\mathrm{OP}}_{\mathbf{LLW}}, \widehat{\mathrm{IP}}_{\mathrm{LLW}})$ gives comparable or better results, while being significantly faster to train. While falling behind the other two OT solvers, $(\widehat{\mathrm{OP}}_{\mathbf{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ balances speed and performance, while not being tied to a specific source metameasure. Notably, the independent coupling has the worst performance, but reduces the performance gap for enough Euler steps. Exact WoW and LLW remain within standard deviation here, consistent with LLW approximating concentrated sources.
Under the same protocol, Euclidean minibatch OT improves over independent pairing at five steps but trails W, BW, and LLW in OT-NNA. The gap largely vanishes with more Euler steps, locating the main benefit in few-step trajectories. Exact W and LLW differ chiefly in online cost. While Table 1 fixes $B = 8$, minibatch OT benefits from larger batches [Tong et al., 2024], where exact WoW scales quadratically (Appendix C.6) but LLW stays negligible. At $B = 64$ (140k steps, otherwise same setting), LLW reaches a five-step OT-NNA of $0.61_{\pm 0.01}$, outperforming all $B = 8$ couplings and the slower, equally trained $(\text{Euclidean}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ $(0.65_{\pm 0.02})$.

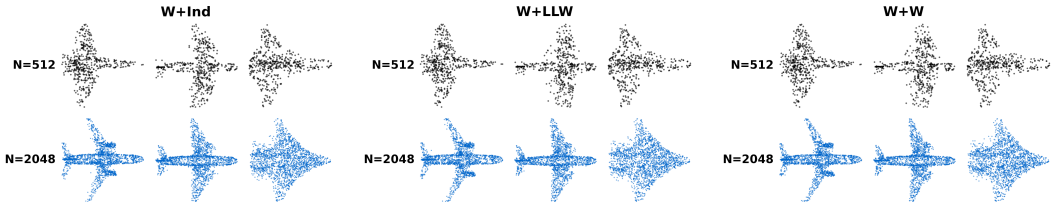
4.4 ShapeNet Airplanes (3D, Transformer, Resolution: 512)

Next, we investigate 3D plane-shaped point clouds in the form of planes obtained from the ShapeNet dataset [Yi et al., 2016]. We use a standard transformer and resolution $N = 512$ resulting in min-max normalized point clouds encoded as $\mathbf{x} \in [0, 1]^{3\times 512} \subset \mathbb{R}^{3\times 512}$. Again, we employ barycentric noise with base scale $\sigma_0 = 0.3$ and coordinate-wise factors $U_{i,j} \in [0.5, 1.5]$ and train each model for 500 epochs (Adam, step size: $5 \cdot 10^{-4}$). We evaluate different outer transport plans while fixing the inner transport plan as IP_{W}.

Outer matching	Chamfer-NNA ↓ Euler-5	OT-NNA ↓ Euler-5	Coupling/step (s)
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$ Euclidean	$0.742_{\pm .02}$ $0.738_{\pm .01}$	$0.741_{\pm .01}$ $0.715_{\pm .01}$	≈ 0.14 ≈ 0.15
$\widehat{\mathrm{OP}}_{\mathrm{BW}}$	$\underline{0.702_{\pm .01}}$	$\underline{0.671_{\pm .01}}$	≈ 0.16
$\widehat{\mathrm{OP}}_{\mathbf{W}}$	$\mathbf{0.695_{\pm .02}}$	$\mathbf{0.661_{\pm .01}}$	≈ 1.31
$\widehat{\mathrm{OP}}_{\mathrm{SW}}$	$0.739_{\pm .01}$	$0.724_{\pm .01}$	≈ 0.15
$\widehat{\mathrm{OP}}_{\mathrm{LLW}}$	$0.713_{\pm .01}$	$0.685_{\pm .01}$	≈ 0.15

Table 2: ShapeNet-Airplane controlled outer-matching comparison with $\widehat{\mathrm{IP}}_{\mathrm{W}}$ fixed. Training and coupling timings use $B = 8$ and $N = 512$. NNA uses five Euler steps.

Although trained at $N = 512$, our discretization-invariant transformer generalizes to high-resolution settings, see Figure 2 for $N = 2048$. After 25 Euler steps, the independent, LLW, and exact W outer couplings induce similar endpoints.



(a) $(\widehat{\mathrm{OP}}_{\mathbf{ind}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ (b) $(\widehat{\mathrm{OP}}_{\mathbf{LLW}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ (c) $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$

Figure 2: ShapeNet planes from identical barycentric-source samples after 25 Euler steps. The outer couplings induce similar endpoints.

Table 2 compares outer plans at the training batch size $B = 8$ under exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$, EMA, a cosine schedule, and equal budgets. NNA clouds are downsampled to $N = 256$ [Piening and Beinert, 2026b]. Exact W performs best, while BW and LLW retain most of its five-step gain and LLW scales like flattened Euclidean matching. SW, with eight slices in $d = 3$, only matches flattened Euclidean matching here.

4.5 QM9 Molecular Sets (7D, Baseline Model, Resolution: 9)

Beyond resampled surfaces, we use QM9 geometries [Ramakrishnan et al., 2014] with nine heavy atoms. Centred 3D coordinates and one-hot types in $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}\}$ give shuffled sets in $\mathbb{R}^{9\times 7}$. We use the baseline network, $B = 64$, exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$, and barycentric noise with $\sigma_0 = 0.1$ and $U_{i,j} \equiv 1$. LLW improves five-step quality over Euclidean matching, with a narrower gap at 25 steps. BW was prohibitively slow at $B = 64$ ($\approx$6 s vs. 0.1 s per step for LLW, i.e., $>$1 day vs. 27 min). Offline superset couplings such as NSOT (Appendix D) do not apply here, since molecules cannot be upsampled by adding atoms.

Outer matching	Chamfer-NNA ↓		OT-NNA ↓	
	Euler-5	Euler-25	Euler-5	Euler-25
Euclidean	$0.825_{\pm.00}$	$0.685_{\pm.01}$	$0.850_{\pm.00}$	$0.659_{\pm.01}$
$\widehat{\mathrm{OP}}_{\mathrm{LLW}}$	$\mathbf{0.752}_{\pm.01}$	$\mathbf{0.671}_{\pm.01}$	$\mathbf{0.754}_{\pm.01}$	$\mathbf{0.652}_{\pm.01}$

Table 3: QM9 fixed-$N = 9$ outer-matching comparison with exact inner plans (3 evaluation repeats, see Appendix E.2).

4.6 MNIST→USPS Transport (2D, Transformer, Multi-Resolution: [64, 128, 256])

To illustrate differences in the induced flow maps, we transport point clouds of handwritten MNIST digits to the bolder digit style of USPS [Hull, 1994], i.e., between two empirical metameasures. Figure 3 shows kernel density estimates of the learned paths. At $t = 0.5$, $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ and $(\widehat{\mathrm{OP}}_{\mathrm{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ preserve the digit structure, while independent coupling yields amorphous densities, hinting that the sliced flow map is closer to the WoW one. The complete experiment is given in Appendix B.1.

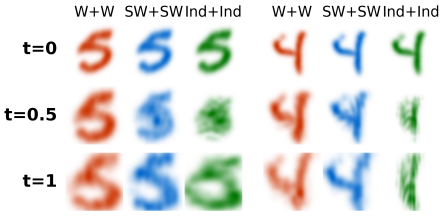


Figure 3: Estimated densities along the learned MNIST→USPS paths.

5 Conclusion

We extended multisample flow matching [Pooladian et al., 2023] from Euclidean space to the hierarchical WoW space. The WoW continuity equation shows that any admissible nested outer and inner coupling induces a valid metameasure flow whose kinetic energy is bounded by the coupling cost: exact WoW plans attain the minimal energy and straight paths, while sliced and lazy linear plans trade exactness for scalability. Our experiments compare couplings within fixed architectures rather than across differently designed point cloud generators. This isolates the coupling, showing that permutation-invariant geometry, not reordering alone, improves few-step generation, but leaves comparisons across generators open. Further limitations are the restriction to fixed-size empirical measures in low dimensions, the quadratic batch-size scaling of exact WoW couplings, and the reliance of LLW on a source concentrated around the barycenter. Natural extensions include conditional and nested measures [Beiglböck et al., 2025, Chemseddine et al., 2025] and parameterizations beyond empirical distributions [Haviv et al., 2025, Jiang et al., 2025].

Acknowledgments

M.P. acknowledges funding from the German Research Foundation (DFG, GRK2260 BIOQIC, project 289347353). Further, M.P. thanks Gregor Kornhardt for fruitful discussions. G.S. and R.D. acknowledge funding from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, EXC-2046/2, project ID: 390685689). Further, G.S. thanks CNRS for support within the CNRS Ambassador Fellowship.

References

Michael Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2005.

Lazar Atanackovic, Xi Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the Wasserstein manifold. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Florian Beier, Robert Beinert, and Gabriele Steidl. On a linear Gromov–Wasserstein distance. *IEEE Transactions on Image Processing*, 31:7292–7305, 2022.

Mathias Beiglböck, Gudmund Pammer, and Stefan Schrott. A Brenier theorem on $(\mathcal{P}_2(\dots\mathcal{P}_2(H)\dots),\mathcal{W}_2)$ and applications to adapted transport. *arXiv preprint arXiv:2509.03506*, 2025.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing datasets with Wasserstein over Wasserstein gradient flows. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Clément Bonet, Elsa Cazelles, Lucas Drumetz, and Nicolas Courty. Busemann functions in the Wasserstein space: existence, closed-forms, and applications to slicing. In *Proceedings of the AISTATS'26*, volume 300, pages 1972–1980. PMLR, 2026.

Nicolas Bonnotte. *Unidimensional and Evolution Methods for Optimal Transportation*. PhD thesis, Université Paris Sud–Paris XI, Orsay, France, 2013. PhD Thesis.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Jannis Chemseddine, Paul Hagemann, Gabriele Steidl, and Christian Wald. Conditional Wasserstein distances with applications in Bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

Jannis Chemseddine, Gregor Kornhardt, Richard Duong, and Gabriele Steidl. Adapting noise to data by quantile learning. In *Proceedings of the ICLR'26 2nd Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*. OpenReview.net, 2026. URL `https://openreview.net/forum?id=E0pD1gcaX7`.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *Proceedings of the ICLR'24*. OpenReview.net, 2024.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, 2013.

Pedram Emami and Brendan Pass. Optimal transport with optimal transport cost: the Monge–Kantorovich problem on Wasserstein spaces. *Calculus of Variations and Partial Differential Equations*, 64(2):43, 2025.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78): 1–8, 2021. Software available at `https://pythonot.github.io/`.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, 2024.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. PCT: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

Paul Hagemann and Sebastian Neumayer. Stabilizing invertible neural networks using mixture models. *Inverse Problems*, 37(8):085002, 2021.

Doron Haviv, Aram-Alexandre Pooladian, Dana Pe'er, and Brandon Amos. Wasserstein flow matching: Generative modeling over families of distributions. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Peter Holderrieth, Marton Havasi, Jason Yim, Neta Shaul, Itai Gat, Tommi Jaakkola, Brian Karrer, Ricky T. Q. Chen, and Yaron Lipman. Generator matching: Generative modeling with arbitrary Markov processes. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Martin Huesmann and Bastian Müller. A Benamou–Brenier formula for transport distances between stationary random measures. *Stochastic Processes and their Applications*, 185:104633, 2025.

Ka-Hei Hui, Chao Liu, Xiaohui Zeng, Chi-Wing Fu, and Arash Vahdat. Not-so-optimal transport flows for 3D point cloud generation. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. doi: 10.1109/34.291440.

Keyue Jiang, Jiahao Cui, Xiaowen Dong, and Laura Toni. Bures-Wasserstein flow matching for graph generation. In *Proceedings of the Generative AI and Biology ICML Workshop*. OpenReview.net, 2025.

Gavin Kerrigan, Giosue Migliorini, and Padhraic Smyth. Functional flow matching. In *Proceedings of the AISTATS'24*, pages 3934–3942. PMLR, 2024.

Beomsu Kim, Gihyun Kwon, Kwanyoung Kim, and Jong Chul Ye. Unpaired image-to-image translation via neural Schrödinger bridge. In *Proceedings of the ICLR'24*. OpenReview.net, 2024.

Leon Klein, Andreas Krämer, and Frank Noé. Equivariant flow matching. *Advances in Neural Information Processing Systems*, 36:59886–59910, 2023.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the ICML'19*, pages 3744–3753. PMLR, 2019.

Yangming Li, Chaoyu Liu, and Carola-Bibiane Schönlieb. Generative unordered flow for set-structured data generation. *arXiv preprint arXiv:2501.17770*, 2025.

Tianyi Lin, Nhat Ho, Xi Chen, Marco Cuturi, and Michael Jordan. Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm. In *Advances in Neural Information Processing Systems*, volume 33, pages 5368–5380. Curran Associates, 2020.

Johannes von Lindheim. Simple approximative algorithms for free-support Wasserstein barycenters. *Computational Optimization and Applications*, 85(1):213–246, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proceedings of the ICLR'23*. OpenReview.net, 2023.

Xinran Liu, Rocio Diaz Martin, Yikun Bai, Ashkan Shahbazi, Matthew Thorpe, Akram Aldroubi, and Soheil Kolouri. Expected sliced transport plans. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

David Lüdke, Enric Rabasseda Raventós, Marcel Kollovieh, and Stephan Günnemann. Unlocking point processes through point set diffusion. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Shitong Luo and Wei Hu. Diffusion probabilistic models for 3D point cloud generation. In *Proceedings of the CVPR'21*, pages 2837–2845. IEEE, 2021.

Alireza Mousavi-Hosseini, Stephen Y Zhang, Michal Klein, and Marco Cuturi. Flow matching with semidiscrete couplings. In *Proceedings of the ICLR'26*. OpenReview.net, 2026.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Simsekli. Statistical and topological properties of sliced probability divergences. In *Advances in Neural Information Processing Systems*, volume 33, pages 15657–15669. Curran Associates, 2020.

Dai Hai Nguyen and Koji Tsuda. On a linear fused Gromov–Wasserstein distance for graph structured data. *Pattern Recognition*, 138:109351, 2023.

Khai Nguyen, Hai Nguyen, Tuan Pham, and Nhat Ho. Lightspeed geometric dataset distance via sliced optimal transport. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Trung Nguyen, Quang-Hieu Pham, Tam Le, Tung Pham, Nhat Ho, and Binh-Son Hua. Point-set distances for learning representations of 3D point clouds. In *Proceedings of the ICCV'21*, pages 10458–10467. IEEE, 2021.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *Proceedings of the ICLR'25*. OpenReview.net, 2025.

Emanuel Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*, volume 11 of *Foundations and Trends® in Machine Learning*. Now Publishers, 2019. ISBN 978-1-68083-550-2. doi: 10.1561/2200000073.

Moritz Piening and Robert Beinert. Slicing the Gaussian mixture Wasserstein distance. *Transactions on Machine Learning Research*, 2025.

Moritz Piening and Robert Beinert. A novel sliced fused Gromov-Wasserstein distance. In *Proceedings of the AAAI'26*, volume 40, pages 24828–24836. AAAI Press, 2026a.

Moritz Piening and Robert Beinert. Slicing Wasserstein over Wasserstein via functional optimal transport. In *Proceedings of the ICLR'26*. OpenReview.net, 2026b.

Alessandro Pinzi and Giuseppe Savaré. Nested superposition principle for random measures and the geometry of the Wasserstein on Wasserstein space. *arXiv preprint arXiv:2510.07523*, 2025.

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 ICML'23*. OpenReview.net, 2023.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the CVPR'17*, pages 652–660. IEEE, 2017.

Yiming Qin, Manuel Madeira, Dorina Thanou, and Pascal Frossard. DeFoG: Discrete flow matching for graph generation. In *Proceedings of the ICML'25*. OpenReview.net, 2025.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, August 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL `https://doi.org/10.1038/sdata.2014.22`.

Francesco Ruscelli. Flow matching on homogeneous spaces. *arXiv preprint arXiv:2603.24829*, 2026.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs and Modeling*. Birkhäuser, Cham, 2015.

Clément Sarrazin and Bernhard Schmitzer. Linearized optimal transport on manifolds. *SIAM Journal on Mathematical Analysis*, 56(4):4970–5016, 2024.

Ryoma Sato, Marco Cuturi, Makoto Yamada, and Hisashi Kashima. Fast and robust comparison of probability measures in heterogeneous spaces, 2020. arXiv:2002.01615.

David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, 2 edition, 2015. ISBN 978-0-471-69755-8.

Xiangjun Tang, Biao Zhang, and Peter Wonka. Generative human geometry distribution. In *Proceedings of the ICLR'26*. OpenReview.net, 2026.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2009.

Christian Wald and Gabriele Steidl. Flow matching: Markov kernels, stochastic processes and transport plans. *Variational and Information Flows in Machine Learning and Optimal Transport*, pages 185–254, 2025.

Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013.

Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. PointFlow: 3D point cloud generation with continuous normalizing flows. In *Proceedings of the ICCV'19*, pages 4541–4550. IEEE, 2019.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3D shape collections. *ACM Transactions on Graphics (ToG)*, 35(6):1–12, 2016.

A Background and Proofs

In order to describe the continuity equation (4) in the WoW setting, we need to define test functions suitable for testing on the infinite dimensional space $\mathcal{P}(\mathbb{R}^d)$.

Definition A.1. *Consider the cylindric functions $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ defined by*

$$F(\mu) \coloneqq \Psi\Big(\int_{\mathbb{R}^d} \phi_1 \,\mathrm{d}\mu, \dots, \int_{\mathbb{R}^d} \phi_k \,\mathrm{d}\mu\Big), \quad \Psi \in C_b^1(\mathbb{R}^k),\ \phi_i \in C_c^1(\mathbb{R}^d),\ i = 1, \dots, k.$$

For $I = [0,1]$, let $(M_t)_{t\in I} \in C_I(\mathcal{P}(\mathcal{P}(\mathbb{R}^d)))$ and $V : I \times \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}^d$ be a Borel non-local vector field satisfying the integrability condition

$$\int_I \int_{\mathcal{P}(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|V(t,x,\mu)\| \,\mathrm{d}\mu(x) \,\mathrm{d}M_t(\mu) \,\mathrm{d}t < \infty. \tag{11}$$

Then, it is said that (M_t, V_t) satisfies the continuity equation (4) *if for all cylindric functions F it holds*

$$\frac{d}{dt}\int_{\mathcal{P}} F(\mu) \,\mathrm{d}M_t(\mu) = \int_{\mathcal{P}} \int_{\mathbb{R}^d} (\nabla_W F)(x,\mu) \cdot V_t(x,\mu) \,\mathrm{d}\mu(x) \,\mathrm{d}M_t(\mu)$$

in the sense of distributions on I, where $(\nabla_W F)(x,\mu) \coloneqq \sum_{i=1}^k \partial_i \Psi(\int \phi_1 \,\mathrm{d}\mu, \dots, \int \phi_k \,\mathrm{d}\mu)\nabla\phi_i(x)$.

Here, we prove Lemma 2.4 on how to construct random couplings from inner and outer plans.

Lemma A.2. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$. Then, for any $\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}} \in C(\boldsymbol{\mu},\boldsymbol{\nu})$ and any $\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}$-measurable selection $\mathrm{IP} : (\mu,\nu) \mapsto \mathrm{IP}_{\mu,\nu} \in c(\mu,\nu)$, it holds that $\boldsymbol{\Pi} \coloneqq \mathrm{IP}_\sharp \,\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}} \in \mathrm{RC}(\boldsymbol{\mu},\boldsymbol{\nu})$. Furthermore, the WoW distance* (2) *satisfies*

$$\mathbf{W}_2^2(\boldsymbol{\mu},\boldsymbol{\nu}) = \inf_{\boldsymbol{\Pi}\in\mathrm{RC}(\boldsymbol{\mu},\boldsymbol{\nu})} \int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - x'\|^2 \,\mathrm{d}\pi(x,x') \,\mathrm{d}\boldsymbol{\Pi}(\pi).$$

Proof. 1. Concerning the first claim, we get for any bounded measurable $f : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ that

$$\begin{aligned}
\int f(\mu) \,\mathrm{d}(\mathbf{Proj}^0_\sharp \boldsymbol{\Pi})(\mu) &= \int f(\mathbf{Proj}^0(\pi)) \,\mathrm{d}\boldsymbol{\Pi}(\pi) \\
&= \int f(\mathbf{Proj}^0(\mathrm{IP}_{\mu,\nu})) \,\mathrm{d}\,\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}(\mu,\nu) \\
&= \int f(\mathrm{proj}^0_\sharp \,\mathrm{IP}_{\mu,\nu}) \,\mathrm{d}\,\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}(\mu,\nu) \\
&= \int f(\mu) \,\mathrm{d}\,\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}(\mu,\nu) = \int f(\mu) \,\mathrm{d}\boldsymbol{\mu}(\mu),
\end{aligned}$$

where the second equality uses $\boldsymbol{\Pi} = \mathrm{IP}_\sharp \,\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}$, the third one relies on $\mathbf{Proj}^0(\pi) = \mathrm{proj}^0_\sharp \pi$, and the fourth one on $\mathrm{proj}^0_\sharp \,\mathrm{IP}_{\mu,\nu} = \mu$. The last equality holds since $\mathrm{OP}_{\boldsymbol{\mu},\boldsymbol{\nu}}$ has first marginal $\boldsymbol{\mu}$. An analogous computation shows $\mathbf{Proj}^1_\sharp \boldsymbol{\Pi} = \boldsymbol{\nu}$, hence $\boldsymbol{\Pi} \in \mathrm{RC}(\boldsymbol{\mu},\boldsymbol{\nu})$.

2. We prove an upper and lower bound estimate, resulting in the claimed identity. Note that the upper bound estimate is due to [Pinzi and Savaré, 2025, Proposition 3.1]. For completeness, we include it here.

Upper bound '$\leq$'. Let $\mathrm{OP}^* \in \mathrm{C}(\boldsymbol{\mu},\boldsymbol{\nu})$ realize the WoW distance via

$$\mathbf{W}_2^2(\boldsymbol{\mu},\boldsymbol{\nu}) = \int_{\mathcal{P}_2(\mathbb{R}^d)\times\mathcal{P}_2(\mathbb{R}^d)} \mathrm{W}_2^2(\mu,\nu) \,\mathrm{d}\,\mathrm{OP}^*(\mu,\nu).$$

By a measurable selection theorem [Villani, 2009, Cor. 5.22], there exists a OP^*-a.e. measurable map $\mathrm{IP}^* \colon (\mu,\nu) \mapsto \mathrm{IP}^*_{\mu,\nu} \in c(\mu,\nu)$ such that

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} \|x - x'\|^2 \,\mathrm{d}\,\mathrm{IP}^*_{\mu,\nu}(x,x') = \mathrm{W}_2^2(\mu,\nu).$$

Define

$$\boldsymbol{\Pi}^* \coloneqq \mathrm{IP}^*_\sharp \,\mathrm{OP}^*.$$

Then it holds $\mathbf{\Pi}^* \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$ by the first part. Taking the infimum over all admissible $\mathbf{\Pi}$ gives the bound

$$\inf_{\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int \int \|x - x'\|^2 \, \mathrm{d}\pi \, \mathrm{d}\mathbf{\Pi} \leq \int \int \|x - x'\|^2 \, \mathrm{d}\pi \, \mathrm{d}\mathbf{\Pi}^* = \mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

Lower bound '$\geq$'. Let $\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and define

$$\widetilde{\mathrm{OP}} \coloneqq (\mathbf{Proj}^0, \mathbf{Proj}^1)_\sharp \mathbf{\Pi} \in \mathrm{C}(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

Then for $\mathbf{\Pi}$-a.e. $\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ we obtain by definition of W_2 that

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^2 \, \mathrm{d}\pi(x, x') \;\geq\; \mathrm{W}_2^2(\mathrm{proj}^0_\sharp \pi, \mathrm{proj}^1_\sharp \pi).$$

By definition of the push-forward and $\mathbf{W}_2$, we get

$$\begin{aligned}
\int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - x'\|^2 \, \mathrm{d}\pi \, \mathrm{d}\mathbf{\Pi}(\pi) &\geq \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \mathrm{W}_2^2(\mathrm{proj}^0_\sharp \pi, \mathrm{proj}^1_\sharp \pi) \, \mathrm{d}\mathbf{\Pi}(\pi) \\
&= \int_{\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d)} \mathrm{W}_2^2(\mu, \nu) \, \mathrm{d}((\mathbf{Proj}^0, \mathbf{Proj}^1)_\sharp \mathbf{\Pi})(\mu, \nu) \\
&= \int_{\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d)} \mathrm{W}_2^2(\mu, \nu) \, \mathrm{d}\widetilde{\mathrm{OP}}(\mu, \nu) \\
&\geq \inf_{\mathrm{OP} \in \mathrm{C}(\boldsymbol{\mu}, \boldsymbol{\nu})} \int_{\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d)} \mathrm{W}_2^2(\mu, \nu) \, \mathrm{d}\,\mathrm{OP}(\mu, \nu) \\
&= \mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}).
\end{aligned}$$

Taking the infimum over $\mathbf{\Pi}$ gives the second inequality, and the proof is finished. □

We now come to the proof of Theorem 2.5 which characterizes the non-local velocity fields of certain metameasure curves as minimizers of the loss function (7).

Theorem 2.5. *Let $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ and $\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$, and define the curve $(M_t)_{t \in [0,1]}$ by $M_t \coloneqq (\mathbf{Proj}^t_\sharp \mathbf{\Pi})$, $\mathbf{Proj}^t \pi \coloneqq \mathrm{proj}^t_\sharp \pi$. Then, $\mathbf{\Pi} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d))$, i.e. it holds the integrability*

$$\int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x\|^2 + \|x'\|^2 \, \mathrm{d}\pi(x, x') \, \mathrm{d}\mathbf{\Pi}(\pi) < \infty. \tag{6}$$

Furthermore, $(M_t)_{t \in [0,1]} \in AC_I(\mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)))$ and the functional

$$J(V) \coloneqq \int_I \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|V(t, \mathrm{proj}^t(x, x'), \mathbf{Proj}^t(\pi)) - (x' - x)\|^2 \, \mathrm{d}\pi(x, x') \, \mathrm{d}\mathbf{\Pi}(\pi) \, \mathrm{d}t \tag{7}$$

admits a unique minimizer V_t, such that (M_t, V_t) fulfills the continuity equation (4)*. Moreover, the kinetic energy of V satisfies*

$$\mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}) \leq \int_I \int \int \|V(t, x, \mu)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}M_t(\mu) \, \mathrm{d}t \leq \int \int \|x' - x\|^2 \, \mathrm{d}\pi(x, x') \, \mathrm{d}\mathbf{\Pi}(\pi), \tag{8}$$

with equality in both bounds if $\mathbf{\Pi} = \mathrm{IP}^*_\sharp \, \mathrm{OP}^*$ with $\mathrm{OP}^* \in \mathrm{C}^{\mathrm{opt}}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\mathrm{IP}^*_{\mu,\nu} \in \mathrm{c}^{\mathrm{opt}}(\mu, \nu)$.

Proof. First, we prove essential properties of the curve $(M_t)_{t \in [0,1]}$.

Correct endpoints of the measure curve: By the assumption $\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$, the curve $(M_t)_{t \in [0,1]}$ indeed has the given endpoints $M_0 = \boldsymbol{\mu}$ and $M_1 = \boldsymbol{\nu}$.

Integrability property (6)**:** First, by the assumptions $\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$ and $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, it holds for $\mathbf{\Pi}$-a.e. $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ that $\mathrm{proj}^0_\sharp \pi, \mathrm{proj}^1_\sharp \pi \in \mathcal{P}_2(\mathbb{R}^d)$. Thus,

$$\int_{\mathbb{R}^d \times \mathbb{R}^d} \|x\|^2 + \|x'\|^2 \, \mathrm{d}\pi(x, x') = \int_{\mathbb{R}^d} \|z\|^2 \, \mathrm{d}\mathrm{proj}^0_\sharp \pi(z) + \int_{\mathbb{R}^d} \|z\|^2 \, \mathrm{d}\mathrm{proj}^1_\sharp \pi(z) < \infty,$$

which means that $\mathbf{\Pi}$ is supported on $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$.
Second, note that for all $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ it holds the estimates

$$W_2^2\big(\mu,\nu\big) \leq 2\int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x) + 2\int_{\mathbb{R}^d} \|y\|^2 \,\mathrm{d}\nu(y) \tag{12}$$

and

$$\int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x) \leq 2W_2^2\big(\mu,\nu\big) + 2\int_{\mathbb{R}^d} \|y\|^2 \,\mathrm{d}\nu(y) \tag{13}$$

by the triangle and Young's inequality, and using any coupling $\pi \in \mathrm{c}(\mu,\nu)$ in (12), and an optimal coupling $\pi^* \in \mathrm{c}^{\mathrm{opt}}(\mu,\nu)$ in (13). In particular, it holds for all $M \in \mathcal{P}(\mathcal{P}_2(\mathbb{R}^d))$ the equivalence

$$M \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d)) \iff \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}M(\mu) < \infty. \tag{14}$$

Therefore, by the assumption $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$, we have

$$\int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\boldsymbol{\mu}(\mu) + \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}\boldsymbol{\nu}(\mu) < \infty.$$

Now, $\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu},\boldsymbol{\nu})$ immediately implies that the above sum is equal to

$$\int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x\|^2 + \|x'\|^2 \,\mathrm{d}\pi(x,x')\,\mathrm{d}\mathbf{\Pi}(\pi),$$

which proves the integrability bound (6), i.e. $\mathbf{\Pi} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d))$.

Absolute continuity of the curve: We first show that $M_t \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ for all $t \in [0,1]$. Indeed, since $\mathbf{\Pi}$-a.e. π is in $\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, we have $\mathrm{proj}^t_\sharp \pi \in \mathcal{P}_2(\mathbb{R}^d)$, meaning that M_t is supported on $\mathcal{P}_2(\mathbb{R}^d)$. Then, it holds

$$\begin{aligned}
&\int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}M_t(\mu) = \int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d} \|x\|^2 \,\mathrm{d}(\mathrm{proj}^t_\sharp\pi)(x)\,\mathrm{d}\mathbf{\Pi}(\pi)\\
=&\int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|(1-t)x + tx'\|^2 \,\mathrm{d}\pi(x,x')\,\mathrm{d}\mathbf{\Pi}(\pi)\\
\leq&\int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} 2(1-t)^2\|x\|^2 + 2t^2\|x'\|^2 \,\mathrm{d}\pi(x,x')\,\mathrm{d}\mathbf{\Pi}(\pi),
\end{aligned}$$

which is finite by the already shown estimate (6) and $t \in [0,1]$. Hence, $M_t \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ by (14). Next, for all $s,t \in [0,1]$ and every $\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, the coupling induced by the map $(x,x') \mapsto (\mathrm{proj}^s(x,x'), \mathrm{proj}^t(x,x'))$ gives the bound

$$W_2^2\big(\mathrm{proj}^s_\#\pi, \mathrm{proj}^t_\#\pi\big) \leq |t-s|^2 \int_{\mathbb{R}^d\times\mathbb{R}^d} |x'-x|^2 \,\mathrm{d}\pi(x,x'). \tag{15}$$

With the mapping $T : \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d), \pi \mapsto (\mathrm{proj}^s_\sharp\pi, \mathrm{proj}^t_\sharp\pi)$, we obtain a coupling $T_\sharp\mathbf{\Pi} \in \mathrm{C}(M_s, M_t)$: the endpoint relations are clear by definition, and we have

$$\begin{aligned}
&\int_{\mathcal{P}_2(\mathbb{R}^d)\times\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu,\tilde\mu) + W_2^2(\nu,\tilde\nu)\,\mathrm{d}T_\sharp\mathbf{\Pi}(\mu,\nu)\\
=&\int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} W_2^2(\mathrm{proj}^s_\sharp\pi,\tilde\mu) + W_2^2(\mathrm{proj}^t_\sharp\pi,\tilde\nu)\,\mathrm{d}\mathbf{\Pi}(\pi)\\
=&\int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu,\tilde\mu)\,\mathrm{d}M_s(\mu) + \int_{\mathcal{P}_2(\mathbb{R}^d)} W_2^2(\mu,\tilde\nu)\,\mathrm{d}M_t(\mu) < \infty
\end{aligned}$$

for any $\tilde\mu, \tilde\nu \in \mathcal{P}_2(\mathbb{R}^d)$, since $M_s, M_t \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$. Hence, $T_\sharp\mathbf{\Pi} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d))$. Therefore we verified $T_\sharp\mathbf{\Pi} \in \mathrm{C}(M_s, M_t)$, and the above estimate (15) yields

$$\mathbf{W}_2^2(M_s, M_t) \leq |t-s|^2 \int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x'-x\|^2 \,\mathrm{d}\pi(x,x')\,\mathrm{d}\mathbf{\Pi}(\pi).$$

By the proven estimate (6), the last double integral is finite. Hence $t \mapsto M_t$ is Lipschitz, in particular absolutely continuous with respect to $\mathbf{W}_2$.

Square-integrable conditional vector field: Define the two-level measure Q on $\mathcal{X} \coloneqq \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}^d$ via

$$\mathrm{d}Q(\pi, x, x') \coloneqq \mathrm{d}\pi(x, x')\, \mathrm{d}\mathbf{\Pi}(\pi).$$

Next, consider the measurable map $P^t : \mathcal{X} \to \mathcal{P}_2(\mathbb{R}^d) \times \mathbb{R}^d$ given by

$$P^t : (\pi, x, x') \mapsto \big(\mathrm{proj}^t_{\#}\pi, \mathrm{proj}^t(x, x')\big) = (\tilde{\mu}, \tilde{x}).$$

Now we have $\mathrm{d}(P^t_{\#}Q)(\tilde{\mu}, \tilde{x}) = \mathrm{d}\tilde{\mu}(\tilde{x})\, \mathrm{d}M_t(\tilde{\mu})$. Since all concerned spaces are Polish, disintegration of Q with regard to P_t defines a regular conditional probability

$$\Gamma_{t,\tilde{\mu},\tilde{x}} \in \mathcal{P}\big(\mathcal{X}\big),$$

such that

$$\mathrm{d}Q(\pi, x, x') = \mathrm{d}\Gamma_{t,\tilde{\mu},\tilde{x}}(\pi, x, x')\ \mathrm{d}\tilde{\mu}(\tilde{x})\ \mathrm{d}M_t(\tilde{\mu}).$$

Then we define the *conditional vector field*

$$V(t, \tilde{x}, \tilde{\mu}) \coloneqq \int_{\mathcal{X}} (x' - x)\ \mathrm{d}\Gamma_{t,\tilde{\mu},\tilde{x}}(\pi, x, x').$$

Let us verify that V satisfies a square-integrability condition. By Jensen's inequality applied to the probability measure $\Gamma_{t,\tilde{\mu},\tilde{x}}$, we obtain

$$\begin{aligned} |V(t, \tilde{x}, \tilde{\mu})|^2 &= \left| \int_{\mathcal{X}} (x' - x)\ \mathrm{d}\Gamma_{t,\tilde{\mu},\tilde{x}}(\pi, x, x') \right|^2 \\ &\leq \int_{\mathcal{X}} \|x' - x\|^2\ \mathrm{d}\Gamma_{t,\tilde{\mu},\tilde{x}}(\pi, x, x'). \end{aligned}$$

Integrating with respect to $\mathrm{d}\tilde{\mu}(\tilde{x})\ \mathrm{d}M_t(\tilde{\mu})\ \mathrm{d}t$ and using the disintegration identity for Q, we find

$$\begin{aligned} &\int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} |V(t, \tilde{x}, \tilde{\mu})|^2\ \mathrm{d}\tilde{\mu}(\tilde{x})\ \mathrm{d}\, M_t(\tilde{\mu})\ \mathrm{d}t \qquad (16) \\ &\leq \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \int_{\mathcal{X}} \|x' - x\|^2\ \mathrm{d}\Gamma_{t,\tilde{\mu},\tilde{x}}(\pi, x, x')\ \mathrm{d}\tilde{\mu}(\tilde{x})\ \mathrm{d}M_t(\tilde{\mu})\ \mathrm{d}t \\ &= \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x' - x\|^2\ \mathrm{d}\pi(x, x')\ \mathrm{d}\mathbf{\Pi}(\pi)\ \mathrm{d}t < \infty, \end{aligned}$$

where the last term is finite by the bound (6). In particular, V satisfies the integrability condition (11) of Definition A.1.

Metameasure flow is generated via conditional vector field: We claim that $\partial_t M_t + \mathrm{div}_{\mathcal{P}}(V_t M_t) = 0$ in the sense of Definition A.1. To prove this, let

$$F(\mu) = \Psi\Big(\int_{\mathbb{R}^d} \phi_1\ \mathrm{d}\mu, \ldots, \int_{\mathbb{R}^d} \phi_k\ \mathrm{d}\mu\Big)$$

with $\phi_i \in C^1_c(\mathbb{R}^d), i = 1, \ldots, k$, and $\Psi \in C^1_b(\mathbb{R}^k)$. Since $M_t = \mathbf{Proj}^t_{\#}\mathbf{\Pi}$, we have

$$\int_{\mathcal{P}_2(\mathbb{R}^d)} F(\mu)\ \mathrm{d}M_t(\mu) = \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} F\big(\mathrm{proj}^t_{\#}\pi\big)\ \mathrm{d}\mathbf{\Pi}(\pi).$$

For fixed $\pi \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$, the map $t \mapsto F(\mathrm{proj}^t_\# \pi)$ is differentiable. Thus, we may apply the chain rule to reformulate

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\, F\big(\mathrm{proj}^t_\# \pi\big) \\
&= \frac{\mathrm{d}}{\mathrm{d}t}\, \Psi\Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_1(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x'), \dots, \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_k(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x')\Big) \\
&= \sum_{i=1}^k \partial_i \Psi\Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_1(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x'), \dots, \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_k(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x')\Big) \\
&\qquad \cdot \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_i(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x') \\
&= \sum_{i=1}^k \partial_i \Psi\Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_1(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x'), \dots, \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_k(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x')\Big) \\
&\qquad \cdot \int_{\mathbb{R}^d\times\mathbb{R}^d} \nabla\phi_i(\mathrm{proj}^t(x,x')) \cdot (x'-x)\, \mathrm{d}\pi(x,x') \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{i=1}^k \partial_i \Psi\Big(\int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_1(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x'), \dots, \int_{\mathbb{R}^d\times\mathbb{R}^d} \phi_k(\mathrm{proj}^t(x,x'))\, \mathrm{d}\pi(x,x')\Big) \\
&\qquad \cdot \nabla\phi_i(\mathrm{proj}^t(x,x')) \cdot (x'-x)\, \mathrm{d}\pi(x,x') \\
&= \int_{\mathbb{R}^d\times\mathbb{R}^d} \nabla_W F\big(\mathrm{proj}^t(x,x'), \mathrm{proj}^t_\#\pi\big) \cdot (x'-x)\, \mathrm{d}\pi(x,x').
\end{aligned}
$$

Here, we applied the definition for the first equation and the classical chain rule for the second one. The third equality is justified by differentiation under the integral sign, since $\phi_i \in C^1_c(\mathbb{R}^d)$, the chain rule and Cauchy-Schwarz inequality imply

$$
|\partial_t \phi_i(\mathrm{proj}_t(x,x'))| \le \|\nabla\phi_i\|_\infty \, \|x'-x\|,
$$

and the right-hand side is π-integrable by assumption. In the end, we employed the definition of $\nabla_W F$ after extracting the integral.
The above calculation gives the estimate

$$
\left|\frac{\mathrm{d}}{\mathrm{d}t} F\big(\mathrm{proj}^t_\#\pi\big)\right| \le C_F \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x'-x\|\, \mathrm{d}\pi(x,x'),
$$

for some constant $0 < C_F < \infty$ which is finite by the boundedness of the derivatives of Ψ and ϕ_i. In particular, by the estimate (6) the mapping $\pi \mapsto \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x'-x\|\, \mathrm{d}\pi(x,x')$ is in $L^1(\mathbf{\Pi})$. Thus, we may again differentiate under the integral sign with respect to $\mathbf{\Pi}$. Finally, by the construction of V and M_t and using that $(\mathrm{proj}^t_\#\pi, \mathrm{proj}^t(x,x')) = P^t(\pi,x,x') = (\tilde\mu,\tilde x)$ for $\Gamma_{t,\tilde\mu,\tilde x}$-a.e. (π,x,x'), we obtain

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathcal{P}_2(\mathbb{R}^d)} F(\mu)\, \mathrm{d}M_t(\mu) \\
&= \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \int_{\mathcal{X}} \nabla_W F\big(\mathrm{proj}^t(x,x'), \mathrm{proj}^t_\#\pi\big) \cdot (x'-x)\, \mathrm{d}\Gamma_{t,\tilde\mu,\tilde x}(\pi,x,x')\, \mathrm{d}\tilde\mu(\tilde x)\, \mathrm{d}M_t(\tilde\mu) \\
&= \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \nabla_W F\big(\tilde x,\tilde\mu\big) \cdot \int_{\mathcal{X}} (x'-x)\, \mathrm{d}\Gamma_{t,\tilde\mu,\tilde x}(\pi,x,x')\, \mathrm{d}\tilde\mu(\tilde x)\, \mathrm{d}M_t(\tilde\mu) \\
&= \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \nabla_W F(\tilde x,\tilde\mu) \cdot V(t,\tilde x,\tilde\mu)\, \mathrm{d}\tilde\mu(\tilde x)\, \mathrm{d}M_t(\tilde\mu).
\end{aligned}
$$

Hence, (M_t, V_t) solves the WoW continuity equation $\partial_t M_t + \mathrm{div}_{\mathcal{P}}(V_t M_t) = 0$ in the sense of Definition A.1.

Conditional vector field minimizes $\mathcal{J}$: In line with the derivation of the classical flow matching loss, we now employ the decomposition

$$\begin{aligned}
&\big\|\tilde{V}\big(t, \mathrm{proj}^t(x,x'), \mathrm{proj}^t_{\#}\pi\big) - (x'-x)\big\|^2 \\
&= \big\|\tilde{V}\big(t, \mathrm{proj}^t(x,x'), \mathrm{proj}^t_{\#}\pi\big)\big\|^2 - 2\Big\langle \tilde{V}\big(t, \mathrm{proj}^t(x,x'), \mathrm{proj}^t_{\#}\pi\big), x'-x\Big\rangle \\
&\qquad + \|x'-x\|^2.
\end{aligned}$$

For the first term, by disintegration with respect to

$$P^t : (\pi, x, x') \mapsto \big(\mathrm{proj}^t_{\#}\pi, \mathrm{proj}^t(x,x')\big) = (\tilde{\mu}, \tilde{x}),$$

we obtain

$$\begin{aligned}
&\int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \big\|\tilde{V}\big(t, \mathrm{proj}^t(x,x'), \mathrm{proj}^t_{\#}\pi\big)\big\|^2 \,\mathrm{d}\pi(x,x') \,\mathrm{d}\mathbf{\Pi}(\pi)\,\mathrm{d}t \\
&\qquad = \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|\tilde{V}(t,\tilde{x},\tilde{\mu})\|^2 \,\mathrm{d}\tilde{\mu}(\tilde{x})\,\mathrm{d}M_t(\tilde{\mu})\,\mathrm{d}t.
\end{aligned}$$

For the second mixed term, using again the disintegration kernel $\Gamma_{t,\tilde{\mu},\tilde{x}}$ and the definition of V, we conclude

$$\begin{aligned}
&\int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \Big\langle \tilde{V}\big(t, \mathrm{proj}^t(x,x'), \mathrm{proj}^t_{\#}\pi\big), x'-x\Big\rangle \,\mathrm{d}\pi(x,x')\,\mathrm{d}\mathbf{\Pi}(\pi)\,\mathrm{d}t \\
&= \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \Big\langle \tilde{V}(t,\tilde{x},\tilde{\mu}), \int_{\mathcal{X}} (x'-x)\,\mathrm{d}\Gamma_{t,\tilde{\mu},\tilde{x}}(\pi,x,x')\Big\rangle \,\mathrm{d}\tilde{\mu}(\tilde{x})\,\mathrm{d}M_t(\tilde{\mu})\,\mathrm{d}t \\
&= \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \langle \tilde{V}(t,\tilde{x},\tilde{\mu}), V(t,\tilde{x},\tilde{\mu})\rangle \,\mathrm{d}\tilde{\mu}(\tilde{x})\,\mathrm{d}M_t(\tilde{\mu})\,\mathrm{d}t.
\end{aligned}$$

Therefore, setting

$$C := \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x'-x\|^2 \,\mathrm{d}\pi(x,x')\,\mathrm{d}\mathbf{\Pi}(\pi)\,\mathrm{d}t,$$

we arrive at

$$\begin{aligned}
\mathcal{J}(\tilde{V}) &= \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|\tilde{V}(t,x,\mu)\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}M_t(\mu)\,\mathrm{d}t \\
&\qquad - 2\int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \langle \tilde{V}(t,x,\mu), V(t,x,\mu)\rangle \,\mathrm{d}\mu(x)\,\mathrm{d}M_t(\mu)\,\mathrm{d}t + C \\
&= \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|\tilde{V}(t,x,\mu) - V(t,x,\mu)\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}M_t(\mu)\,\mathrm{d}t \\
&\qquad + \Big(C - \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|V(t,x,\mu)\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}M_t(\mu)\,\mathrm{d}t\Big).
\end{aligned}$$

Hence it follows

$$\mathcal{J}(\tilde{V}) = \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|\tilde{V}(t,x,\mu) - V(t,x,\mu)\|^2 \,\mathrm{d}\mu(x)\,\mathrm{d}M_t(\mu)\,\mathrm{d}t + C_V,$$

where C_V is a constant independent of $\tilde{V}$. Since the first term is nonnegative, we obtain

$$\mathcal{J}(\tilde{V}) \geq \mathcal{J}(V),$$

with equality if and only if

$$\tilde{V} = V \qquad \text{in } L^2([0,1]\times\mathcal{P}_2(\mathbb{R}^d)\times\mathbb{R}^d;\ \mathrm{d}\mu\,\mathrm{d}M_t(\mu)\,\mathrm{d}t;\ \mathbb{R}^d).$$

Consequently, V is the unique minimizer of $\mathcal{J}$ up to L^2-equivalence.

Optimality. Now let $\mathbf{\Pi} = \mathrm{IP}^*_\sharp \, \mathrm{OP}^*$. As established in (16) and in the proof of Lemma 2.4, it holds that

$$\int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|V(t,x,\mu)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}M_t(\mu) \, \mathrm{d}t$$
$$\leq \int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x' - x\|^2 \, \mathrm{d}\pi(x,x') \, \mathrm{d}\mathbf{\Pi}^*(\pi) \, \mathrm{d}t = \mathbf{W}_2^2(\boldsymbol{\mu}, \boldsymbol{\nu}).$$

On the other hand, the velocity field V satisfies the WoW continuity equation $\partial_t M_t + \mathrm{div}_{\mathcal{P}}(V_t M_t) = 0$ as shown above. Hence it is an admissible velocity field in the dynamic WoW formulation, see [Pinzi and Savaré, 2025, Thm. 1.3], giving

$$\int_0^1 \int_{\mathcal{P}_2(\mathbb{R}^d)} \int_{\mathbb{R}^d} \|V(t,x,\mu)\|^2 \, \mathrm{d}\mu(x) \, \mathrm{d}M_t(\mu) \, \mathrm{d}t \geq \mathbf{W}_2^2(M_0, M_1).$$

Combining both bounds results in equality in (8). For general $\mathbf{\Pi} \in \mathrm{RC}(\boldsymbol{\mu}, \boldsymbol{\nu})$, the upper bound in (8) is (16), and the lower bound follows by the same dynamic argument, which concludes the proof. □

In the following we prove Proposition 3.1 justifying the lazy linearization strategy proposed in Section 3.1.

Proposition A.3. *Given the mapping* $\mathrm{Vec2M} : \mathbb{R}^{d\times N} \to \mathcal{P}_2^N(\mathbb{R}^d), \quad \mathbf{x} \mapsto \hat{\mu} = \frac{1}{N}\sum_{j=1}^N \delta_{x_j}$, *a reference measure* $\hat{\rho} \in \mathcal{P}_2^N(\mathbb{R}^d)$*, and some* $\mathbf{x}_{\hat{\rho}} \in \mathbb{R}^{d\times N}$ *with* $\mathbf{x}_{\hat{\rho},i} \neq \mathbf{x}_{\hat{\rho},i'}$ *for all* $i \neq i'$ *and* $\mathrm{Vec2M}(\mathbf{x}_{\hat{\rho}}) = \hat{\rho}$*, then for (Lebesgue-)almost every* $\mathbf{x}' \in \mathbb{R}^{d\times N}$*, there exists* $R_{\mathbf{x}'} > 0$ *such that for all* $\mathbf{x}$ *with* $\|\mathbf{x} - \mathbf{x}_{\hat{\rho}}\| < R_{\mathbf{x}'}$*, it holds that*

$$\mathrm{W}_2^2(\mathrm{Vec2M}(\mathbf{x}), \mathrm{Vec2M}(\mathbf{x}')) = \|\mathbf{x} - P_{\mathrm{Vec2M}(\mathbf{x}'),\mathbf{x}_{\hat{\rho}}}\mathbf{x}'\|^2.$$

Proof. **Almost every permutation minimizer is unique:** For $\mathbf{y}, \mathbf{z} \in \mathbb{R}^{d\times N}$, the measures $\mathrm{Vec2M}(\mathbf{y})$ and $\mathrm{Vec2M}(\mathbf{z})$ are empirical measures with N atoms of equal mass. Hence,

$$\mathrm{W}_2^2(\mathrm{Vec2M}(\mathbf{y}), \mathrm{Vec2M}(\mathbf{z})) = \min_{P\in\mathrm{Perm}(N)} \|\mathbf{y} - P\mathbf{z}\|^2, \tag{17}$$

where $\mathrm{Perm}(N) \subset \mathbb{R}^{N\times N}$ denotes the set of permutation matrices [Peyré and Cuturi, 2019, Prop. 2.1].
We first show that for almost every $\mathbf{x}' \in \mathbb{R}^{d\times N}$, the minimizer of

$$P \mapsto \|\mathbf{x}_{\hat{\rho}} - P\mathbf{x}'\|^2$$

over $P \in \mathrm{Perm}(N)$ is unique.
To this aim, let $P, Q \in \mathrm{Perm}(N)$ with $P \neq Q$. Using that permutations are orthogonal and hence preserve the norm, i.e., $\|P\mathbf{x}'\|^2 = \|Q\mathbf{x}'\|^2 = \|\mathbf{x}'\|^2$, we obtain

$$\begin{aligned}\|\mathbf{x}_{\hat{\rho}} - P\mathbf{x}'\|^2 - \|\mathbf{x}_{\hat{\rho}} - Q\mathbf{x}'\|^2 &= -2\langle \mathbf{x}_{\hat{\rho}}, P\mathbf{x}' - Q\mathbf{x}'\rangle \\ &= -2\langle (P-Q)^T\mathbf{x}_{\hat{\rho}}, \mathbf{x}'\rangle.\end{aligned}$$

Thus, equality of the two costs is equivalent to

$$\langle (P-Q)^T\mathbf{x}_{\hat{\rho}}, \mathbf{x}'\rangle = 0.$$

Since $\mathbf{x}_{\hat{\rho},i} \neq \mathbf{x}_{\hat{\rho},i'}$ for all $i \neq i'$, distinct permutations $P \neq Q$ satisfy $P\mathbf{x}_{\hat{\rho}} \neq Q\mathbf{x}_{\hat{\rho}}$, and thus

$$(P-Q)^T\mathbf{x}_{\hat{\rho}} \neq 0.$$

Therefore, for every pair $P \neq Q$, the set

$$H_{P,Q} := \left\{\mathbf{x}' \in \mathbb{R}^{d\times N} \;:\; \|\mathbf{x}_{\hat{\rho}} - P\mathbf{x}'\|^2 = \|\mathbf{x}_{\hat{\rho}} - Q\mathbf{x}'\|^2\right\}$$

is a hyperplane in $\mathbb{R}^{d\times N}$ and hence has Lebesgue measure zero. As $\mathrm{Perm}(N)$ is finite, the union of all such tie sets also has Lebesgue measure zero. Consequently, for almost every $\mathbf{x}' \in \mathbb{R}^{d\times N}$, there exists a unique permutation matrix

$$P_* = P_{\mathrm{Vec2M}(\mathbf{x}'),\mathbf{x}_{\hat{\rho}}}$$

such that

$$\|\mathbf{x}_{\hat{\rho}} - P_*\mathbf{x}'\|^2 < \|\mathbf{x}_{\hat{\rho}} - P\mathbf{x}'\|^2 \qquad \text{for all } P \in \mathrm{Perm}(N) \setminus \{P_*\}.$$

The permutation minimizer is locally stable: Now fix such an $\mathbf{x}'$. Since $\mathrm{Perm}(N)$ is finite, the quantity

$$\gamma := \min_{P \in \mathrm{Perm}(N)\setminus\{P_*\}} \left(\|\mathbf{x}_{\hat{\rho}} - P\mathbf{x}'\|^2 - \|\mathbf{x}_{\hat{\rho}} - P_*\mathbf{x}'\|^2\right)$$

is strictly positive.
For $P \in \mathrm{Perm}(N)$, define

$$f_P(\mathbf{x}) := \|\mathbf{x} - P\mathbf{x}'\|^2.$$

Then, for any $\mathbf{x} \in \mathbb{R}^{d\times N}$ and $P \in \mathrm{Perm}(N)$,

$$\big(f_P(\mathbf{x}) - f_{P_*}(\mathbf{x})\big) - \big(f_P(\mathbf{x}_{\hat{\rho}}) - f_{P_*}(\mathbf{x}_{\hat{\rho}})\big) = 2\langle \mathbf{x} - \mathbf{x}_{\hat{\rho}}, (P_* - P)\mathbf{x}'\rangle.$$

Hence, by Cauchy–Schwarz,

$$\big|\big(f_P(\mathbf{x}) - f_{P_*}(\mathbf{x})\big) - \big(f_P(\mathbf{x}_{\hat{\rho}}) - f_{P_*}(\mathbf{x}_{\hat{\rho}})\big)\big| \le 2\|\mathbf{x} - \mathbf{x}_{\hat{\rho}}\|\,\|(P_* - P)\mathbf{x}'\|.$$

Again using finiteness of $\mathrm{Perm}(N)$, we may set

$$M := \max_{P \in \mathrm{Perm}(N)\setminus\{P_*\}} \|(P_* - P)\mathbf{x}'\| < \infty.$$

Choose $R_{\mathbf{x}'} > 0$ such that $2R_{\mathbf{x}'}M < \gamma$. Then for every $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{x}_{\hat{\rho}}\| < R_{\mathbf{x}'}$, we have for all $P \neq P_*$ that

$$\begin{aligned} f_P(\mathbf{x}) - f_{P_*}(\mathbf{x}) &\ge f_P(\mathbf{x}_{\hat{\rho}}) - f_{P_*}(\mathbf{x}_{\hat{\rho}}) - 2R_{\mathbf{x}'}M \\ &\ge \gamma - 2R_{\mathbf{x}'}M \\ &> 0. \end{aligned}$$

Thus, for all $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{x}_{\hat{\rho}}\| < R_{\mathbf{x}'}$,

$$f_{P_*}(\mathbf{x}) < f_P(\mathbf{x}) \qquad \text{for all } P \in \mathrm{Perm}(N) \setminus \{P_*\},$$

so P_* remains the unique minimizer of

$$P \mapsto \|\mathbf{x} - P\mathbf{x}'\|^2.$$

Applying (17), we conclude that

$$\mathrm{W}_2^2(\mathrm{Vec2M}(\mathbf{x}), \mathrm{Vec2M}(\mathbf{x}')) = \|\mathbf{x} - P_*\mathbf{x}'\|^2 = \|\mathbf{x} - P_{\mathrm{Vec2M}(\mathbf{x}'),\mathbf{x}_{\hat{\rho}}}\mathbf{x}'\|^2,$$

which proves the claim. □

B Further Experiments

B.1 MNIST→USPS Transport (2D, Transformer, Multi-Resolution: [64, 128, 256])

OP	IP	Chamfer-NNA ↓ Euler-5	Chamfer-NNA ↓ Euler-25	Chamfer-NNA ↓ Euler-125	OT-NNA ↓ Euler-5	OT-NNA ↓ Euler-25	OT-NNA ↓ Euler-125	Train
$\widehat{\mathrm{OP}}_{\mathbf{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{ind}}$	$1.00_{\pm 0.00}$	$0.79_{\pm 0.01}$	$0.70_{\pm 0.00}$	$0.99_{\pm 0.00}$	$0.67_{\pm 0.01}$	$\underline{0.61_{\pm 0.01}}$	50–60h
$\widehat{\mathrm{OP}}_{\mathbf{W}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$\mathbf{0.64_{\pm 0.01}}$	$\mathbf{0.60_{\pm 0.01}}$	$\mathbf{0.60_{\pm 0.01}}$	$\mathbf{0.62_{\pm 0.02}}$	$\mathbf{0.60_{\pm 0.02}}$	$\mathbf{0.61_{\pm 0.02}}$	100–110h
$\widehat{\mathrm{OP}}_{\mathbf{SW}}$	$\widehat{\mathrm{IP}}_{\mathrm{SW}}$	$\underline{0.94_{\pm 0.01}}$	$\underline{0.65_{\pm 0.01}}$	$\underline{0.62_{\pm 0.01}}$	$\underline{0.84_{\pm 0.01}}$	$\underline{0.64_{\pm 0.01}}$	$0.63_{\pm 0.01}$	50–60h

Table 4: Generation quality for MNIST→USPS point clouds.

Beyond generation from noise, we evaluate our framework on a cross-domain transport task. We use point clouds sampled from MNIST as the empirical source metameasure, and point clouds sampled from USPS [Hull, 1994] as the target metameasure, treating both datasets analogously to the MNIST setting above. In particular, normalized pixel intensities define a 2D histogram on $[0,1]^2$ from which $N = 64$,128, or 256 points are sampled. Thus, we randomly obtain $\mathbf{x} \in \mathbb{R}^{2\times 64}$, $\mathbf{x} \in \mathbb{R}^{2\times 128}$ or $\mathbf{x} \in \mathbb{R}^{2\times 256}$. We train a standard transformer for 350 epochs using Adam (step size: $5 \cdot 10^{-4}$) using different coupling strategies. Notably, a domain-adaptive model should learn that the USPS digits are bolder than the MNIST digits.

Evaluating the quality of translated samples at $N = 256$, Table 4 shows the clear advantage of OT-based couplings. Especially for few-step generation (Euler-5), the approximation of the WoW plan clearly gives the best results. In this few-step case, the sliced coupling drastically improves upon the independent coupling as well, while adding almost no computational training overhead in contrast to the WoW approximation. For sufficiently discretized flow ODEs, the NNA metrics stabilize, see the comparison between Euler-25 and Euler-125, and the independent coupling becomes competitive.

Figure 4 visualizes learned interpolations via kernel density estimation [Parzen, 1962], see Appendix E.9. While $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ and $(\widehat{\mathrm{OP}}_{\mathrm{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ preserve digit structure at the midpoint ($t = 0.5$), the independent coupling $(\widehat{\mathrm{OP}}_{\mathbf{ind}}, \widehat{\mathrm{IP}}_{\mathrm{ind}})$ yields amorphous intermediate densities.



(a) $N = 128$ (b) $N = 4096$

Figure 4: Learned interpolation for MNIST→USPS with $N = 128$ and out-of-training $N = 4096$ for (i) $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$, (ii) $(\widehat{\mathrm{OP}}_{\mathrm{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ and (iii) $(\widehat{\mathrm{OP}}_{\mathbf{ind}}, \widehat{\mathrm{IP}}_{\mathrm{ind}})$ visualized via estimated densities. MNIST digits ($t = 0$) are mapped to emboldened USPS-like digits ($t = 1$) for all couplings. However, $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ and $(\widehat{\mathrm{OP}}_{\mathrm{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ midpoints ($t = 0.5$) maintain digit structure for $N = 4096$, whereas $(\widehat{\mathrm{OP}}_{\mathbf{ind}}, \widehat{\mathrm{IP}}_{\mathrm{ind}})$ midpoints appear amorphous. Similar trajectories emerge for low-resolution (4a) and high-resolution (4b).

C Ablation Studies

C.1 Ablation of Latent Source Metameasure

Based on the MNIST generation experiment in Section 4.3, we investigate the importance of the latent noise metameasure. Since the barycentric latent employed in the experiment is closer to the data distribution than raw noise, we repeat the experiment under the same conditions with our pure noise latent ($\underline{\sigma} = 0.05$, $\overline{\sigma} = 0.15$). The results are presented in Table 5. In comparison to the results based on the barycentric latent in Table 1, exact W and SW perform comparably or slightly worse, while the quality clearly declines for our lazy linear estimators. This is expected by our theory, since the pure noise latent is not covered by Corollary 3.2 because it is centred around the origin. In contrast, the independent coupling benefits from pure noise in few-step generation (five-step OT-NNA 0.82 vs. 0.95). Thus, the barycentric source mainly pays off in combination with OT-based couplings.

		Chamfer-NNA ↓			OT-NNA ↓			**Train**
OP	**IP**	Euler-5	Euler-25	Euler-125	Euler-5	Euler-25	Euler-125	
$\widehat{\mathrm{OP}}_{\mathbf{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{ind}}$	$0.84_{\pm 0.01}$	$0.67_{\pm 0.02}$	$0.65_{\pm 0.01}$	$0.82_{\pm 0.01}$	$0.62_{\pm 0.01}$	$0.61_{\pm 0.02}$	2–3h
$\widehat{\mathrm{OP}}_{\mathbf{W}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$\mathbf{0.72_{\pm 0.02}}$	$0.65_{\pm 0.01}$	$0.66_{\pm 0.03}$	$\mathbf{0.67_{\pm 0.02}}$	$\underline{0.60_{\pm 0.01}}$	$\underline{0.59_{\pm 0.03}}$	5–6h
$\widehat{\mathrm{OP}}_{\mathbf{SW}}$	$\widehat{\mathrm{IP}}_{\mathrm{SW}}$	$\underline{0.78_{\pm 0.02}}$	$\underline{0.65_{\pm 0.02}}$	$\mathbf{0.64_{\pm 0.01}}$	$0.74_{\pm 0.01}$	$0.62_{\pm 0.02}$	$0.62_{\pm 0.01}$	3–4h
$\widehat{\mathrm{OP}}_{\mathbf{LLW}}$	$\widehat{\mathrm{IP}}_{\mathrm{LLW}}$	$0.78_{\pm 0.01}$	$\mathbf{0.65_{\pm 0.01}}$	$\underline{0.64_{\pm 0.01}}$	$\underline{0.71_{\pm 0.01}}$	$\mathbf{0.59_{\pm 0.02}}$	$\mathbf{0.58_{\pm 0.01}}$	3–4h

Table 5: MNIST generation quality and training time for pure noise.

C.2 Locality of the Lazy Linear Approximation

Proposition 3.1 is local, so we quantify how LLW degrades away from the reference. On ShapeNet-Airplane, we draw a source cloud from the barycenter plus fixed-scale Gaussian noise and compare the LLW assignment with exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$ over 300 target clouds. This diagnostic deliberately sweeps a clean fixed σ. The training source instead uses the coordinate-wise uniform rescaling described in Section 4.1, with base scale 0.3.

σ	exact assignments (LLW / random)	$\mathrm{cost}_{\mathrm{LLW}} / \mathrm{W}_2^2$	$\mathrm{cost}_{\mathrm{rand}} / \mathrm{W}_2^2$
0.001	$0.93/0.002$	1.00	8.7
0.01	$0.57/0.002$	1.01	8.7
0.15	$0.02/0.002$	2.1	6.7
0.30	$0.01/0.002$	2.2	3.7

Table 6: ShapeNet LLW locality diagnostic. Cost 1 is optimal. The random assignment rate is approximately $1/512$.

For $\sigma \leq 0.01$, LLW is essentially exact, as predicted. Exact-permutation agreement then falls rapidly, but this strict statistic treats a nearby alternative partner as entirely wrong. At $\sigma = 0.3$, LLW still has substantially lower realized cost than random matching. In the ShapeNet experiment LLW is used only for the outer cost and selected cloud pairs still use exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$. Preserving pairwise cost structure and ranking therefore matters more than recovering the exact inner permutation. With the actual training source, LLW pairwise costs have Pearson correlation 0.70 with exact W_2^2, compared with 0.28 for flattened Euclidean costs. On MNIST the corresponding correlations are 0.76 and 0.31. This supports interpreting LLW as a cheap geometry-aware cost estimator beyond the strictly exact local regime, not as exact OT at the operating noise scale.

C.3 Ablation of Batch Size

In general, minibatch OT methods are heavily influenced by the batch size [Mousavi-Hosseini et al., 2026]. While a large batch size B becomes computationally infeasible, we would still expect the straightness of trajectories to improve for larger batches. Therefore, we repeat the experiment from Section 4.2 for one epoch each with $B = 4, 8, 16$ for our $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ estimator, i.e., with the same number of training samples ($15{,}000$, $7{,}500$, and $3{,}750$ updates). Looking at Figure 5, we

clearly see that increasing the batch size improves straightness, even though larger batches receive fewer updates.



(a) $B = 4$ (b) $B = 8$ (c) $B = 16$

Figure 5: Learned paths for $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ and $B = 4, 8, 16$. Larger batches lead to straighter paths.

C.4 Ablation of Sliced Wasserstein Solver

The quality of any sliced Wasserstein approximation is directly linked to the number of random projections $\sharp\,\mathrm{Slices}$ in (18). In order to study its impact, we repeat the experiment from Section 4.2 for 7,500 steps, but now with our sliced transport plan estimators $(\widehat{\mathrm{OP}}_{\mathbf{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$, and we vary $\sharp\,\mathrm{Slices} = 1, 8, 32$. Figure 6 clearly shows that more projections lead to straighter paths.

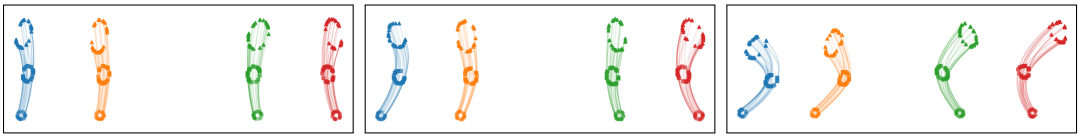


(a) $\sharp\,\mathrm{Slices} = 1$ (b) $\sharp\,\mathrm{Slices} = 8$ (c) $\sharp\,\mathrm{Slices} = 32$

Figure 6: Learned paths for $(\widehat{\mathrm{OP}}_{\mathbf{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ and $\sharp\,\mathrm{Slices} = 1, 8, 32$. More projections lead to straighter paths.

C.5 Ablation of Wasserstein Solver

Notably, we mainly relied on exact linear program solvers for our Wasserstein OT estimators due to the availability of highly efficient implementations [Flamary et al., 2021]. Especially for large-scale problems, such exact solvers are often replaced by Sinkhorn solvers that solve an entropically regularized OT problem [Cuturi, 2013, Peyré and Cuturi, 2019]. These solvers are governed by an entropic regularization parameter $\mathrm{reg} > 0$, where $\mathrm{reg} \to 0$ recovers an exact OT plan and $\mathrm{reg} \to \infty$ simply recovers the independent coupling between the input measures. To investigate the impact of replacing the exact OT solver with a Sinkhorn solver, we repeat the experiment from Section 4.2 for 7,500 steps with $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$, but replace the inner and outer `ot.emd` solvers from the Python Optimal Transport library [Flamary et al., 2021], with the default `ot.sinkhorn` solvers on cost matrices normalized by their maximum for $\mathrm{reg} = 0.01, 0.1, 1$. For our computational setup, training takes around ∼3:50 minutes using the `ot.emd` solver and $\mathrm{reg} = 0.1, 1$ and increases to ∼5 minutes for $\mathrm{reg} = 0.01$. While we observe no runtime advantage for the Sinkhorn solver in comparison to the exact solver in this case, computational savings are expected for settings with more points, GPU acceleration or limited Sinkhorn iterations. The learned paths visualized in Figure 7 clearly show that lowering the entropy parameter reg reduces curvature.



(a) $\mathrm{reg} = 0.01$ (b) $\mathrm{reg} = 0.1$ (c) $\mathrm{reg} = 1$

Figure 7: Learned paths for $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ with Sinkhorn regularizer $\mathrm{reg} = 0.01, 0.1, 1$. Less regularization leads to straighter paths, but increases training time.

C.6 Ablation of Transport Solver Runtime

To quantify the computational overhead of different transport solver combinations, we measure the wall-clock time required by the transport solver per training step for varying cloud size N and batch size B. Concretely, we simulate B independent point clouds of dimension $d = 2$ with points drawn uniformly at random from $[0, 1]^2$ and solve the necessary transport problems. We report the mean CPU runtime (in milliseconds) and standard deviation over 5 runs. Results are shown in Table 7.

As expected, the WoW approximation $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ becomes prohibitively expensive even at moderate N, already exceeding 75 seconds per step at $N = 1024$, $B = 32$. The mixed coupling $(\widehat{\mathrm{OP}}_{\mathrm{ind}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ used in Wasserstein flow matching reduces this cost substantially but still scales with N, reaching $\sim$2.3 seconds at $N = 1024$, $B = 32$. Similarly, $(\widehat{\mathrm{OP}}_{\mathbf{SW}}, \widehat{\mathrm{IP}}_{\mathrm{SW}})$ with $\sharp\,\mathrm{Slices} = 8$ projections (as used in our main experiments) incurs a moderate and predictable overhead that grows more slowly with N and B. In contrast, $(\widehat{\mathrm{OP}}_{\mathrm{LLW}}, -)$ introduces only negligible overhead, as the Euclidean cost on preordered clouds based on the barycenter requires no inner resorting.

Note that the use of the lazy linear strategy requires solving OT problems once during a preprocessing step. The computational overhead of this preprocessing step is in line with the computational cost of solving $(\widehat{\mathrm{OP}}_{\mathrm{ind}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ once for the whole training dataset.

OP	**IP**	$B = 8$			$B = 16$			$B = 32$		
		$N = 64$	$N = 256$	$N = 1024$	$N = 64$	$N = 256$	$N = 1024$	$N = 64$	$N = 256$	$N = 1024$
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{ind}}$	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1
$\widehat{\mathrm{OP}}_{\mathrm{SW}}$	$\widehat{\mathrm{IP}}_{\mathrm{SW}}$	$3.1_{\pm 0.4}$	$10.3_{\pm 0.3}$	$46.0_{\pm 0.8}$	$11.2_{\pm 0.4}$	$38.7_{\pm 0.8}$	$165.2_{\pm 3.0}$	$39.5_{\pm 0.3}$	$142.2_{\pm 0.8}$	$594.4_{\pm 1.5}$
$\widehat{\mathrm{OP}}_{\mathrm{W}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$13.7_{\pm 0.5}$	$141.7_{\pm 1.5}$	$5128.0_{\pm 31.1}$	$51.9_{\pm 3.2}$	$565.9_{\pm 3.6}$	$19244.6_{\pm 91.7}$	$196.6_{\pm 1.8}$	$2264.8_{\pm 20.0}$	$75792.9_{\pm 489.4}$
$\widehat{\mathrm{OP}}_{\mathrm{LLW}}$	–	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	<0.1	$1.3_{\pm 0.3}$
$\widehat{\mathrm{OP}}_{\mathrm{ind}}$	$\widehat{\mathrm{IP}}_{\mathrm{W}}$	$1.2_{\pm 0.0}$	$15.9_{\pm 0.4}$	$586.7_{\pm 13.2}$	$2.5_{\pm 0.1}$	$34.3_{\pm 1.5}$	$1158.0_{\pm 13.4}$	$5.1_{\pm 0.4}$	$68.6_{\pm 0.6}$	$2303.2_{\pm 23.5}$

Table 7: Wall-clock time per training step (mean $\pm$ standard deviation over 5 runs, ms) for each (OP, IP) combination, varying batch size B and cloud size N, CPU only. LLW outer uses Euclidean distance on pre-ordered clouds with no inner resorting ($-$).

D Related Work

Wasserstein flow matching. WFM [Haviv et al., 2025] introduced flow matching for families of distributions and, for empirical point clouds, uses an optimal inner map with independent outer pairing. Its multisample variant performs outer matching with Fréchet/Bures distances between Gaussian approximations of the clouds. Our framework contains both regimes: $(\widehat{\mathrm{OP}}_{\mathrm{ind}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ recovers the former, while $(\widehat{\mathrm{OP}}_{\mathrm{BW}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ mirrors the latter. The distinction is that Theorem 2.5 proves validity directly on WoW space for arbitrary admissible inner and outer plans, including empirical measures with singularities, and therefore also covers the sliced and lazy linear constructions.

Not-so-optimal transport flows. NSOT [Hui et al., 2025] also admits an interpretation through our random-coupling framework. For each target surface measure, it precomputes a high-resolution coupling with a Gaussian source and then subsamples paired points. The resulting empirical plan has the sampled point clouds as its marginals, and its conditional expected transport cost equals the cost of the high-resolution plan. Its law therefore defines a measure over transport plans of the type considered here. In the population limit, the source metameasure is the Dirac measure at the Gaussian source, so the outer coupling is unique. The noiseless NSOT construction may consequently be viewed as an offline approximation of the corresponding optimal nested coupling. It does not solve a nontrivial outer matching problem, and finite-superset errors and the hybrid perturbation move it away from exact optimality. Unlike our online construction on observed finite measures, its superset approach is natural for resampled surfaces but does not directly apply to sets such as molecules.

For context, we adapted NSOT to MNIST by upsampling each 64-point digit to a 1024-point superset, using its standard-Gaussian source and offline coupling, and disabling superset augmentation to isolate the coupling. With the recommended hybrid noise $\beta = 0.2$, Chamfer-NNA at 5/25 steps is $0.813/0.694$ and OT-NNA is $0.717/0.626$. With $\beta = 0$, the respective values are $0.722/0.654$ and $0.643/0.600$. Exact WoW, BW, LLW, and the noiseless NSOT adaptation are broadly comparable on this saturated MNIST setting, but use different sources and incur different online/offline costs. For this reason, NSOT is reported here as contextual evidence rather than inserted into the controlled same-source table.

E Implementation Details

E.1 Computational Setup

All experiments are conducted on a single NVIDIA GeForce RTX 5090 (32 GB VRAM) or comparable. All transport plan computations are performed on the CPU using the Python Optimal Transport library [Flamary et al., 2021], while model forward and backward passes run on the GPU. Transport problems are computed fresh at every gradient step. We employ exact optimal transport solvers without entropic regularization, see Section C.5. MNIST, USPS and ShapeNet data are all normalized to $[0,1]^d$. Data loading and coupling recomputation run in the main training process.

E.2 Training Details of the Controlled Comparisons

ShapeNet (Table 2). All outer plans share one configuration: transformer, $N = 512$, $B = 8$, 500 epochs, Adam (step size $5 \cdot 10^{-4}$, cosine schedule), EMA with decay 0.999 (EMA weights are used for sampling), barycentric noise with $\sigma_0 = 0.3$ and $U_{i,j} \sim \mathrm{Unif}[0.5, 1.5]$, exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$, and seed 7. NNA compares 512 generated with 512 test clouds, both subsampled to 256 points, over 5 repeats.
QM9 (Table 3). We take the QM9 molecules provided by `torch_geometric` with exactly nine heavy atoms in $\{\mathrm{C}, \mathrm{N}, \mathrm{O}, \mathrm{F}\}$ after removing hydrogens, and split them by a fixed list of held-out indices. Training uses a random subset of 20,000 training molecules and evaluation draws from 12,000 held-out molecules. Each atom is the centred 3D position concatenated with its one-hot type (unit weight, no coordinate rescaling), and atom order is shuffled. Both outer plans use the QM9 baseline network (Appendix E), $B = 64$, 50 epochs, Adam (step size $5 \cdot 10^{-4}$, cosine schedule), EMA with decay 0.999, exact $\widehat{\mathrm{IP}}_{\mathrm{W}}$, barycentric noise with $\sigma_0 = 0.1$ and $U_{i,j} \equiv 1$, and seed 42. Before computing NNA, the type coordinates of generated atoms are projected to the nearest one-hot vector. NNA compares 512 generated with 512 held-out molecules over 3 repeats.
MNIST baselines. The BW, Euclidean, and NSOT runs (Table 1, Appendix D) use the MNIST setting of Section 4.3: baseline network, $B = 8$, 50 epochs, Adam (step size $5 \cdot 10^{-4}$, constant), no EMA, and seed 7. The $B = 64$ runs train for 150 epochs (≈140k steps) with otherwise identical settings.

E.3 Practical Procedure for Sampling from Discrete Plans

In practice, the Monte Carlo sampling of matched pairs from the outer and inner plans proceeds as follows. Each source measure $\hat{\mu}_b$ draws its partner $\hat{\nu}_b$ by multinomial sampling from the b-th row of the outer plan matrix $\hat{\Pi}^{\mathrm{OP}}$. Likewise, each source particle $x_{b,j}$ draws its partner $x'_{b,j}$ from the j-th row of the inner plan matrix $\hat{\pi}_b^{\mathrm{IP}}$, yielding the matrix-valued samples $(\mathbf{x}_b, \mathbf{x}'_b) \in \mathbb{R}^{d\times N} \times \mathbb{R}^{d\times N}$ used in the loss. For $\widehat{\mathrm{IP}}_{\mathrm{ind}}$, we draw a fresh uniformly random permutation at every step, i.e., we sample the product plan without replacement.

E.4 Sliced Wasserstein Estimator

In practice, the integral over the sphere in $\mathrm{SW}_2^2(\mu,\nu) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_2^2\big(\mathrm{proj}^\theta_\#\mu, \mathrm{proj}^\theta_\#\nu\big)\, \mathrm{d}S(\theta)$, is approximated via Monte Carlo sampling. In particular, let $\{\theta_\ell\}_{\ell=1}^{\sharp\,\mathrm{Slices}} \sim S := \mathrm{Unif}(\mathbb{S}^{d-1})$, then

$$\mathrm{SW}_2^2(\mu,\nu) \approx \frac{1}{\sharp\,\mathrm{Slices}} \sum_{\ell=1}^{\sharp\,\mathrm{Slices}} \mathrm{W}_2^2\big(\mathrm{proj}^{\theta_\ell}_\#\mu, \mathrm{proj}^{\theta_\ell}_\#\nu\big), \qquad \mathrm{proj}^{\theta_\ell}(x) = \langle \theta_\ell, x\rangle. \tag{18}$$

In all experiments, the Monte Carlo estimator in (18) uses $\sharp\,\mathrm{Slices} = 8$ uniformly sampled directions $\theta_\ell \sim \mathrm{Unif}(\mathbb{S}^{d-1})$, drawn fresh at each coupling recomputation step. We refer to Section C.4 for an ablation study.

E.5 Sliced Transport Plan Computation

For the sliced inner transport plan $\mathrm{IP}_{\mathrm{SW}}(\hat{\mu},\hat{\nu}) = \int_{\mathbb{S}^{d-1}} \mathrm{IP}_\theta(\hat{\mu},\hat{\nu})\, \mathrm{d}S(\theta)$, we employ the same Monte Carlo estimator as in (18). For each direction θ_ℓ, we project both measures onto $\mathbb{R}$ and compute the 1D optimal transport plan $\mathrm{IP}_{\theta_\ell}(\hat{\mu},\hat{\nu}) \in \mathrm{c}(\hat{\mu},\hat{\nu})$ via sorting of the projected support points, which lifts to a permutation on the original d-dimensional particles. Averaging across directions yields

$$\widehat{\mathrm{IP}}_{\mathrm{SW}}(\hat{\mu},\hat{\nu}) \approx \frac{1}{\sharp\,\mathrm{Slices}} \sum_{\ell=1}^{\sharp\,\mathrm{Slices}} \mathrm{IP}_{\theta_\ell}(\hat{\mu},\hat{\nu}),$$

with $\sharp\,\mathrm{Slices} = 8$ as above. In practice, most 1D plans $\mathrm{IP}_{\theta_\ell}(\hat{\mu}, \hat{\nu})$ are permutation matrices and our sliced plan is computed by averaging permutation matrices. We refer to [Liu et al., 2025] for further details.

E.6 Lazy Linear Inner Plan

Unlike $\widehat{\mathrm{IP}}_{\mathrm{W}}$ and $\widehat{\mathrm{IP}}_{\mathrm{SW}}$, which may yield fractional couplings during training, the lazy linear plan $\widehat{\mathrm{IP}}_{\mathrm{LLW}}(\hat{\mu}, \hat{\nu})$ always produces a permutation matrix by construction: both clouds are aligned to a shared barycenter $\hat{\rho}$. By construction of our barycentric metameasure, the source is trivially aligned with the barycenter. Each target point cloud $\hat{\nu}$ is aligned via a precomputed permutation $\sigma_{\hat{\nu}}$. Therefore, we skip the inner sampling step and simply employ the precomputed point-to-point correspondence $(x_{b,i},\, x'_{b,\sigma_{\hat{\nu}}(i)})$, which is exactly the plan $\widehat{\mathrm{IP}}_{\mathrm{LLW}}$ from Section 3.1.

E.7 Outer Plan Computation

Given a choice of pairwise divergence $D : \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}_{\geq 0}$ (e.g., W_2^2, SW_2^2, or LLW_2^2) and a batch of source and target measures $(\hat{\mu}_i)_{i=1}^B, (\hat{\nu}_{i'})_{i'=1}^B$, we compute the outer transport plan as follows. First, we assemble the $B \times B$ cost matrix

$$C \in \mathbb{R}^{B\times B}, \qquad C_{i,i'} \coloneqq D(\hat{\mu}_i, \hat{\nu}_{i'}).$$

We then solve the resulting discrete optimal transport problem

$$\hat{\Pi}^{\mathrm{OP}} \in \operatorname*{arg\,min}_{\Pi \in \mathbb{R}_{\geq 0}^{B\times B}} \sum_{i,i'=1}^{B} C_{i,i'}\,\Pi_{i,i'} \quad \text{s.t.} \quad \sum_{i'} \Pi_{i,i'} = \sum_{i} \Pi_{i,i'} = \tfrac{1}{B},$$

using an exact linear program solver from the Python Optimal Transport library [Flamary et al., 2021]. The resulting transport plan $\widehat{\mathrm{OP}}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}} = \sum_{i,i'=1}^{B} \hat{\Pi}^{\mathrm{OP}}_{i,i'}\,\delta_{(\hat{\mu}_i,\hat{\nu}_{i'})}$ is then used to sample matched pairs $(\hat{\mu}_b, \hat{\nu}_b) \sim \widehat{\mathrm{OP}}_{\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\nu}}}$ as described in Section 3. Depending on the divergence D, this procedure yields $\widehat{\mathrm{OP}}_{\mathbf{W}}$ (using W_2^2), $\widehat{\mathrm{OP}}_{\mathrm{SW}}$ (using SW_2^2), or $\widehat{\mathrm{OP}}_{\mathbf{LLW}}$ (using LLW_2^2). The BW baseline uses Gaussian Bures–Wasserstein costs computed in double precision from symmetrized, jittered empirical covariance matrices. The Euclidean baseline uses mean squared distance between vectorized clouds without point reordering when constructing C.

E.8 Network Architectures

Baseline Model. The baseline model is a per-point MLP augmented with local and global set information as described in Section 4.1, i.e., each d-dimensional point feature is concatenated with (i) sorted pairwise distances to all other points and the correspondingly sorted points and (ii) the empirical mean and upper triangle of the covariance matrix of the set. Notably, the importance of pairwise distances for point cloud recovery has been empirically demonstrated in [Piening and Beinert, 2026a, Sato et al., 2020]. A sinusoidal time embedding of dimension 32 is concatenated to these features. The per-point features are processed by a multi-layer MLP, followed by one self-attention block, and a final linear output projection to $\mathbb{R}^d$. The architecture is shared across three experiments, with different capacities:

- **Circles (Section 4.2):** 3-layer MLP with hidden dimension 64, followed by 1 self-attention block with 4 heads on the hidden dimension. Total parameters: $\approx 71\,000$. Training uses EMA with decay 0.999.
- **MNIST generation (Section 4.3):** 5-layer MLP with hidden dimension 256, followed by 1 self-attention block with 4 heads on the hidden dimension. Total parameters: $\approx 1{,}179\,000$.
- **QM9 (Section 4.5):** a 4-layer MLP with hidden dimension 192, followed by one self-attention block with 4 heads on the hidden dimension. Total parameters: $\approx 623\,000$.

Point Transformer. The standard transformer processes each point cloud as an unordered set of tokens. An input linear projection maps each point $x_i \in \mathbb{R}^d$ ($d \in \{2, 3\}$) to a token of dimension 128. A sinusoidal time embedding of dimension 32 is projected to 128 and broadcast-added to all tokens. The model consists of 4 standard transformer blocks, each with 4 attention heads, LayerNorm, and an MLP with expansion ratio 4. A final linear layer projects back to $\mathbb{R}^d$. No positional embedding is used (the model is permutation-equivariant by design). We employ the same architecture with $\approx 798\,000$ parameters for ShapeNet (Section 4.4) and MNIST→USPS (Section B.1).

E.9 Kernel Density Estimation for Visualization

To visualize 2D empirical point cloud distributions in Experiment B.1, we estimate a continuous density from each point cloud $\{x_i\}_{i=1}^N \subset [0,1]^2$ via Gaussian kernel density estimation [Parzen, 1962], using the `scipy.stats.gaussian_kde` implementation. The bandwidth follows the $d = 2$ variant of Scott's rule [Scott, 2015], scaled by 0.9, giving $h = 0.9 \cdot N^{-1/6}$. The density is evaluated on a 64×64 grid over $[0,1]^2$. To prevent attenuation near the boundary, the grid is extended by 15% padding on each side and cropped back to $[0,1]^2$ before rendering.

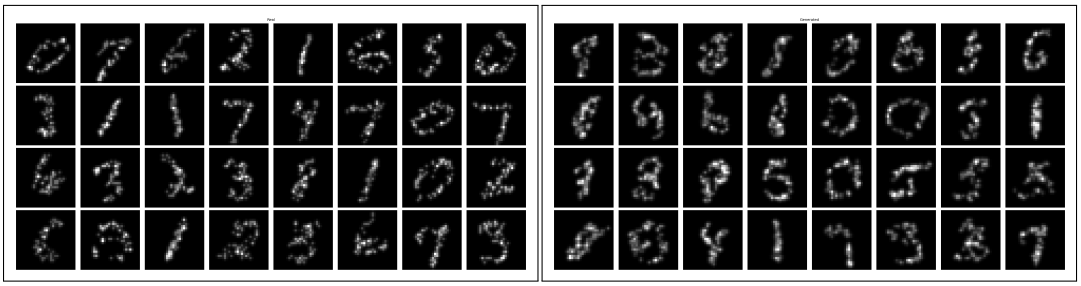
E.10 Practical Guide to Choosing Couplings

Exact WoW is preferable when the lowest prescribed transport energy and straightest interpolation justify its cost. Its B^2 inner OT solves limit scalability, particularly because minibatch OT generally benefits from large batches [Tong et al., 2024]. Gaussian Bures–Wasserstein outer costs are inexpensive when first and second moments describe the clouds well, but discard non-Gaussian shape information.

Sliced Wasserstein is a generally applicable choice for low-dimensional, non-Gaussian clouds. It is a metric, supports arbitrary source and target metameasures, and is substantially cheaper than exact W in $d = 2$ or 3 [Nguyen et al., 2021]. Its projection cost grows in higher dimension [Nadjahi et al., 2020].

Lazy linear Wasserstein is preferable when scalable training and accurate few-step generation are priorities and a barycentric source is appropriate. Its validity as a coupling does not depend on exactness, but its fidelity to W depends on source concentration: too little noise can make learning difficult, while too much breaks the local linearization. We therefore recommend comparing LLW and exact Wasserstein costs and rankings over several noise levels on a small subset before training, as in Appendix C.2.

F Visualization of Additional Samples



(a) True MNIST digits. (b) Generated MNIST digits.

Figure 8: 2D histogram visualization based on true and generated MNIST point clouds ($N = 64$) based on the lazy linear flow learned via ($\widehat{\mathrm{OP}}_{\mathbf{LLW}}$, $\widehat{\mathrm{IP}}_{\mathrm{LLW}}$) for the barycentric noise→MNIST generation experiment in Section 4.3. Pixel intensities correspond to number of aggregated point counts.

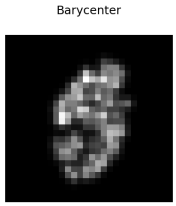


Figure 9: 2D histogram visualization of the computed MNIST barycenter employed as the center of the barycentric noise for the experiment in Section 4.3.

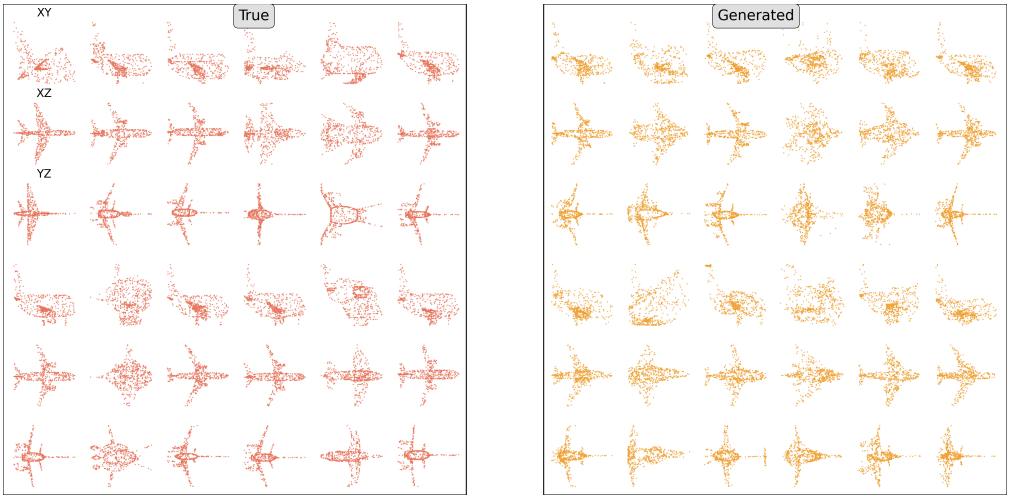
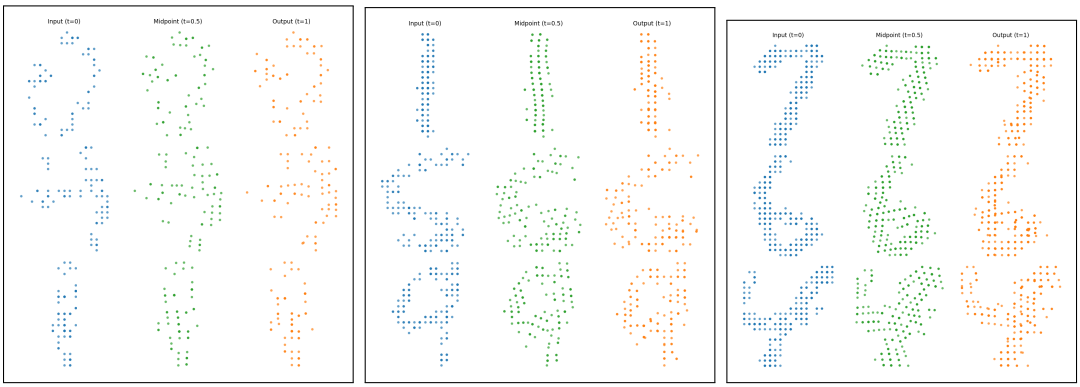
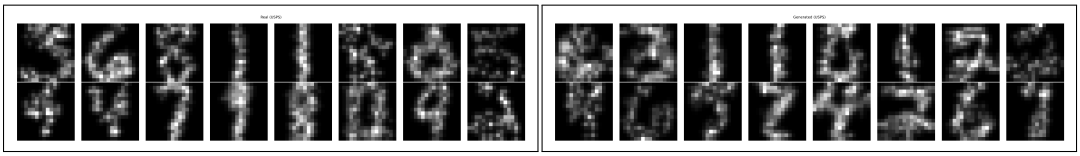


Figure 10: True and generated ShapeNet airplanes for discretization $N = 512$ for the experiment in Section 4.4, where we show generated samples for the model trained with $\widehat{\mathrm{OP}}_{\mathbf{SW}}$. We display the point cloud as seen from different viewpoints.



(a) $N = 64$ (b) $N = 128$ (c) $N = 256$

Figure 11: 2D scatter plot visualization of the point cloud flow learned using the WoW setup $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ for $N = 64, 128, 256$ for the MNIST→USPS experiment in Section B.1. Left is starting point ($t = 0$, blue), middle is midpoint ($t = 0.5$, green), and right is endpoint ($t = 1$, orange).



(a) True USPS digits. (b) Generated USPS digits.

Figure 12: 2D histogram visualization based on true and generated USPS point clouds for $N = 256$ based on the WoW flow learned via $(\widehat{\mathrm{OP}}_{\mathbf{W}}, \widehat{\mathrm{IP}}_{\mathrm{W}})$ for the MNIST→USPS experiment in Section B.1. Pixel intensities correspond to number of aggregated point counts.

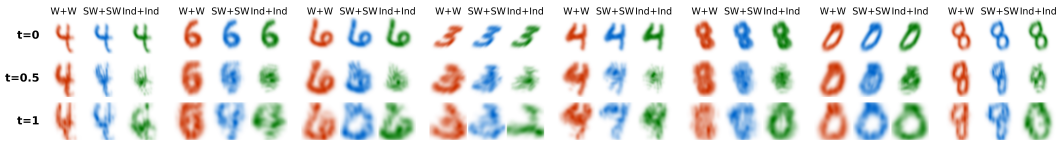


Figure 13: Same visualization as in Figure 4 with more samples for $N = 4096$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Theory: Section 2, Practical Implementation: Section 3, Actual Performance: Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the limitations in the Conclusion, Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All theoretical results are stated precisely. Novel proofs are given in Appendix A, known results are cited appropriately.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix E on implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code for the circle and MNIST experiments, including all configurations, is available at `https://github.com/MoePien/wow-flow-matching`. Data is online.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Appendix E on implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, NNA standard deviations are reported and described.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E contains hardware information.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Code confirmed. Relatively theoretical work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Theoretical work on generative modelling with no *direct* consequences. Generally, generative AI should be employed carefully.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No such risks identified.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We reimplemented main algorithms. OT toolbox has been cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: No new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs as writing + editing tool.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.