Amortized Optimal Transport from Sliced Potentials

Minh-Phuc Truong
Department of Computer Science
Hanoi Univeristy of Science and Technology
Hanoi, Vietnam
truongminhphuc08102005@gmail.com

Khai Nguyen*
Department of Statistics and Data Science
Texas A&M University
College Station, Texas, USA
kbn397@tamu.edu

Abstract

We propose a novel amortized optimization method for predicting optimal transport (OT) plans across multiple pairs of measures by leveraging Kantorovich potentials derived from sliced OT. We introduce two amortization strategies: regression-based amortization (RA-OT) and objective-based amortization (OA-OT). In RA-OT, we formulate a functional regression model that treats Kantorovich potentials from the original OT problem as responses and those obtained from sliced OT as predictors, and estimate these models via least-squares methods. In OA-OT, we estimate the parameters of the functional model by optimizing the Kantorovich dual objective. In both approaches, the predicted OT plan is subsequently recovered from the estimated potentials. As amortized OT methods, both RA-OT and OA-OT enable efficient solutions to repeated OT problems across different measure pairs by reusing information learned from prior instances to rapidly approximate new solutions. Moreover, by exploiting the structure provided by sliced OT, the proposed models are more parsimonious, independent of specific structures of the measures, such as the number of atoms in the discrete case, while achieving high accuracy. We demonstrate the effectiveness of our approaches on tasks including MNIST digit transport, color transfer, supply–demand transportation on spherical data, and mini-batch OT conditional flow matching[1].

1 Introduction

Optimal transport (OT) [73, 74] has been widely recognized as a fundamental and powerful tool in statistics, machine learning, and data science. It plays a key role across a broad range of applications. For instance, OT has improved the quality of generative models, including autoencoders [70, 52], generative adversarial networks [6, 33], and diffusion and flow-based models [39, 54, 71], as well as drifting models [35]. It is also widely used in domain adaptation, where it helps establish correspondences between source and target domains [19, 20, 23, 76]. Beyond these areas, OT has found numerous applications in computational biology [15, 62], chemistry [75], image processing [29], fairness [36], signal processing [38], computer graphics [67, 66, 12], statistical inference [8, 9], and dependency measurement [17, 16], among others.

Computational optimal transport (OT) [53] has emerged as a central yet challenging area within OT. Except for special cases such as OT between Gaussian measures, which admits a closed form solution [24], computing OT is generally expensive. In the discrete setting, where both measures are supported on finite sets, OT can be formulated as a linear programming problem with super cubic complexity in the number of support points [51]. Various approaches have been proposed to mitigate this computational burden, including stochastic approximation with mini-batches [68, 28, 27, 49] and low rank approximations [61, 60]. For general measures, OT is typically solved through its dual

*Corresponding author

[1]Code for the paper is available at https://github.com/tmp0810/Sliced-Amortized-OT.

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

formulation, which involves a functional optimization problem over Kantorovich potentials [74]. This dual problem is constrained and often requires parameterizing the potentials for practical computation [6, 43], which can be challenging.

One widely used approach is entropic regularization [21], which smooths the transport problem by adding an entropy penalty. In the discrete setting, this regularization ensures a unique transport plan that can be efficiently computed via matrix scaling algorithms such as Sinkhorn–Knopp [65], or, equivalently, through iterative Bregman projections [7]. Beyond reducing computational complexity to quadratic time, entropic regularization also improves the statistical rates for estimating both the transport cost [31] and the transport plan [44, 57]. In the general cases, it removes constraints in the dual formulation, enabling unconstrained optimization [32]. For these reasons, entropic regularized OT is widely used in practice as a surrogate for classical OT.

In many scenarios involving numerous pairs of measures, such as mini-batch generative models [33, 35], dataset comparison [3], 3D point-cloud autoencoders [1], point-cloud nearest neighbor classification and regression [58], seismic signal analysis [26], and single-cell perturbation modeling [14], optimal transport (OT) must be solved repeatedly. To accelerate computation in such settings, amortized optimization [64, 4] has been employed to predict OT solutions directly from input measures. Two primary targets arise for amortization: the transportation cost and the transportation plan. Predicting the transportation cost is relatively straightforward, as it reduces to estimating a nonnegative scalar. This direction has been widely explored, for example through learned embeddings or embedding networks [45, 18, 37, 34], as well as regression onto sliced optimal transport metrics [50]. In contrast, amortizing the transportation plan is significantly more challenging, as it involves predicting a joint measure over the input distributions and implicitly determines the cost. To the best of our knowledge, the only existing approach is Meta-OT [5], which learns to predict entropically regularized Kantorovich potentials from the input measures as a proxy for the plan.

In the discrete setting, Meta-OT constructs an amortized model, typically a neural network, that maps input measures, represented by their atomic supports and corresponding weights, to a vector whose dimension matches the number of atoms, thereby approximating one of the Kantorovich potentials. In the continuous setting, Meta-OT instead employs a hypernetwork as the amortized model to predict the parameters of a parametric Kantorovich potential. In both cases, Meta-OT adopts an objective-based amortization strategy, where the amortized models are learned by directly optimizing the Kantorovich dual objective. The main limitation of Meta-OT is that it directly takes the raw representation of input measures, hence, its amortized models contains many parameters. In this work, we propose a new approach to amortizing the transportation plan. The key idea is to use *Kantorovich potentials derived from sliced OT* [56, 46] as inputs to the amortized model, rather than the raw input measures. This representation substantially reduces the dimensionality of the input, and consequently lowers the number of parameters required for the amortized model.

Contribution. In summary, our main contributions are three-fold:

1. We introduce a new method to tackle amortized OT problem using sliced OT. In particular, we suggest to predict Kantorovich potentials of the original OT problem from Kantorovich potentials of the corresponding sliced OT problems. Due to the closed-form of sliced OT, the resulting method is computationally fast. Moreover, leveraging expressive sliced OT's potentials, we are able to use parsimonious amortized models with a small number of parameters and are independent of the structure of measures such as the number of atoms of measures in discrete cases. Finally, our proposal can also be seen as the first attempt to reconstruct OT plan from sliced OT as existing works only form a transporation plan from sliced OT without any notion of approximating the original OT one.

2. We consider two amortization strategies: regression-based amortization (RA-OT) and objective-based amortization (OA-OT). In RA-OT, we formulate a functional regression model, serving as the amortized model, where Kantorovich potentials from the original OT problem are treated as responses and those derived from sliced OT as predictors. The model is then estimated using least-squares methods. In OA-OT, we extend the Meta-OT framework by designing an amortized model that predicts one Kantorovich potential from the sliced OT potentials, and train it by optimizing the Kantorovich dual objective.

3. While the proposed method is well defined for both discrete and continuous measures, we focus on the discrete setting in this work. This choice is motivated by the fact that discrete OT is more commonly used in practice, whereas computing continuous OT has no exact solutions. We conduct

experiments on MNIST digit transport, color transfer, spherical supply–demand transportation, and mini-batch OT conditional flow matching to show that the proposed amortized OT approach leads to more favorable prediction of OT plan than the predicted plan from Meta-OT and sliced OT's plans.

2 Background on Optimal Transport and Entropic Regularization

We start this section by reviewing the definition of OT, Kantorovich duality, and Kantorovich potentials. Given two measures (μ, ν) supported on domains $(\mathcal{X}, \mathcal{Y})$ and a ground metric $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, the primal OT [74] problem is defined as follows:

$$\pi^\star \in \operatorname*{arg\,min}_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y), \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of admissible transportation plans (joint measures) between μ and ν, and $\pi^\star$ is the optimal transportation plan. (1) admits a dual problem which is known as Kantorovich duality:

$$(f^\star, g^\star) \in \operatorname*{arg\,max}_{(f,g)\in\mathcal{U}_c(\mathcal{X},\mathcal{Y})} \int_{\mathcal{X}} f(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y) \mathrm{d}\nu(y), \tag{2}$$

where $\mathcal{U}_c = \{(f, g) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) \mid \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, f(x) + g(y) \leq c(x, y)\}$ is the constraint set, $\mathcal{C}(\mathcal{X})$ and $\mathcal{C}(\mathcal{Y})$ are sets of continuous functions on $\mathcal{X}$ and $\mathcal{Y}$ respectively, and $f^\star, g^\star$ are known as Kantorovich potentials. The connection between $\pi^\star$ and $f^\star, g^\star$ is from complementary slackness condition: $f^\star(x) + g^\star(y) = c(x, y)$ for $\pi^\star$- almost everywhere (x, y).

Entropic regularization. Entropic OT [21] smoothens the OT problem and it can be conveniently written in a single convex optimization problem as follows:

$$\pi^\star_\epsilon = \operatorname*{arg\,min}_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) \mathrm{d}\pi(x,y) + \epsilon \mathrm{KL}(\pi \mid \mu \otimes \nu), \tag{3}$$

where $\mathrm{KL}(\pi \mid \mu \otimes \nu) = \int_{\mathcal{X}\times\mathcal{Y}} \left(\log\left(\frac{\mathrm{d}\pi}{\mathrm{d}\mu\otimes\nu}(x,y)\right)\right) \mathrm{d}\pi(x,y)$ and $\mu \otimes \nu$ is the product measure of μ and ν. The dual problem becomes:

$$\begin{aligned}(f^\star, g^\star) \in \operatorname*{arg\,max}_{f\in\mathcal{C}(\mathcal{X}), g\in\mathcal{C}(\mathcal{Y})} & \int_{\mathcal{X}} f(x) \mathrm{d}\mu(x) + \int_{\mathcal{Y}} g(y) \mathrm{d}\nu(y) \\ & - \epsilon \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{f(x) + g(y) - c(x,y)}{\epsilon}\right) \mathrm{d}\mu \otimes \nu(x,y)\right),\end{aligned} \tag{4}$$

For $\epsilon > 0$, the optimal transportation plan is recovered from $(f^\star, g^\star)$ by solving: $\mathrm{d}\pi^\star_\epsilon(x, y) = \exp\left(\frac{f^\star(x)+g^\star(y)-c(x,y)}{\epsilon}\right) \mathrm{d}\mu(x)\mathrm{d}\nu(y)$, and we can also infer one potential from the other i.e., $g(y) = -\epsilon \log\left(\int_{\mathcal{X}} \exp\left(\frac{f(x)-c(x,y)}{\epsilon}\mathrm{d}\mu(x)\right)\right)$ and $f(x) = -\epsilon \log\left(\int_{\mathcal{Y}} \exp\left(\frac{g(y)-c(x,y)}{\epsilon}\mathrm{d}\nu(y)\right)\right)$.

Discrete Case. When $\mu = \sum_{i=1}^n \alpha_i \delta_{x_i}$ and $\nu = \sum_{j=1}^m \beta_j \delta_{y_j}$ with $\sum_{i=1}^n \alpha_i = \sum_{j=1}^m \beta_j$ and $\alpha_i > 0, \beta_j > 0 \forall i, j$, the entropic OT problem becomes:

$$P^\star_\epsilon = \operatorname*{arg\,min}_{P\in\Gamma(\boldsymbol{\alpha},\boldsymbol{\beta})} \langle C, P\rangle - \epsilon H(P), \tag{5}$$

where $\Gamma(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \{P \in \mathbb{R}_+^{n\times m} \mid P\mathbf{1} = \boldsymbol{\alpha}, P^\top \mathbf{1} = \boldsymbol{\beta}\}$ is the set of discrete plans, $H(P) = -\sum_{i=1}^n \sum_{j=1}^m P_{ij} \log P_{ij}$ is the entropy, and $C_{ij} = c(x_i, y_j)$. The entropic OT dual becomes:

$$(\mathbf{f}^\star, \mathbf{g}^\star) \in \operatorname*{arg\,max}_{\mathbf{f}\in\mathbb{R}^n, \mathbf{g}\in\mathbb{R}^m} \langle \mathbf{f}, \boldsymbol{\alpha}\rangle + \langle \mathbf{g}, \boldsymbol{\beta}\rangle - \epsilon \sum_{i=1}^n \sum_{j=1}^m \exp\left(\frac{\mathbf{f}_i + \mathbf{g}_j - C_{ij}}{\epsilon}\right). \tag{6}$$

We can recover a discrete plan from any pairs $(\mathbf{f}, \mathbf{g})$ as follows:

$$P^\star_{ij} = \exp\left(\frac{\mathbf{f}^\star_i + \mathbf{g}^\star_j - C_{ij}}{\epsilon}\right). \tag{7}$$

The mapping between the dual can be rewritten as:

$$\mathbf{g} = \epsilon \log \boldsymbol{\alpha} - \epsilon \log \left(\exp(-C^\top/\epsilon) \exp(\mathbf{f}/\epsilon) \right), \tag{8}$$

$$\mathbf{f} = \epsilon \log \boldsymbol{\beta} - \epsilon \log \left(\exp(-C/\epsilon) \exp(\mathbf{g}/\epsilon) \right). \tag{9}$$

The Sinkhorn algorithm [21] is an efficient iterative method for solving the entropic optimal transport problem. It alternates between updating the dual potentials $\mathbf{f}$ and $\mathbf{g}$ according to (8) and (9) until the updates converge. A key property of entropic OT is that knowing only one of the optimal Kantorovich potentials, $\mathbf{f}^\star$ or $\mathbf{g}^\star$, is sufficient: the other potential can be directly computed from it, allowing the optimal transport plan to be fully determined. The time complexity of Sinkhorn algorithm is $\mathcal{O}(nm)$ [2] compared to $\mathcal{O}((n+m)^3 \log(n+m))$ [53] of the conventional OT.

3 Amortized Optimization of Optimal Transport

In this section, we first review the amortized optimal transport (OT) problem in Section 3.1, including the Meta-OT framework, as well as the definitions of amortized models and amortized losses. We then introduce two new amortization strategies based on sliced OT potentials: regression-based amortized OT (RA-OT) and objective-based amortized OT (OA-OT), and discuss their properties in Section 3.2.

3.1 Amortized Optimal Transport Problem

In the amortized OT problem, we observe multiple pairs of probability measures, potentially associated with different ground cost functions. We assume that these triplets are drawn from an unknown meta-distribution $\mathcal{D}$, yielding observations $(\mu_1, \nu_1, c_1), \ldots, (\mu_N, \nu_N, c_N)$ $(N > 1)$. Our objective is not only to compute entropic optimal transport for the given N pairs, but also to leverage these observations to accelerate computation for future pairs $(\mu'_1, \nu'_1, c'_1), \ldots, (\mu'_M, \nu'_M, c'_M)$ $(M > 1)$. Meta-OT [5] provides a solution by adopting amortized optimization. We now discuss how Meta-OT address the problem by reviewing the definition of the amortized model and amortized loss.

Amortized model. For each optimal transport (OT) problem, denoted (μ_i, ν_i, c_i), the goal is to obtain one of the corresponding Kantorovich potentials, e.g., $f^\star[\mu_i, \nu_i, c_i]$. An *amortized model* is a mapping that takes a pair of measures and the associated cost as input and predicts a Kantorovich potential. Formally, a parametric mapping $\hat{f}_\phi$ is considered in Meta-OT [5], where ϕ belongs to a parameter space Φ. The model is trained such that

$$\hat{f}_\phi[\mu, \nu, c] \approx f^\star[\mu, \nu, c] \quad \text{for } \mathcal{D}\text{-almost every } (\mu, \nu, c).$$

When μ and ν are discrete and c is fixed across all realizations from $\mathcal{D}$, the inputs to $\hat{f}_\phi$ can be written as $(\boldsymbol{\alpha}, X)$ and $(\boldsymbol{\beta}, Y)$, representing the weights and atoms of μ and ν, respectively. In Meta-OT, the authors consider cases where the atoms are fixed, so only the weights $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are used as inputs. In this setting, $\hat{f}_\phi$ can be implemented as a neural network (e.g., an MLP) that takes two vectors of sizes n and m as input and outputs a vector of size n. Note that Meta-OT requires the number of atoms, n and m, to be the same across all realizations of $\mathcal{D}$, since designing a neural network that accommodates variable input sizes is nontrivial.

Amortized loss. Meta-OT adapts an objective-based optimization approach to estimate the amortized model. In particular, it solves the following optimization problem:

$$\min_{\phi \in \Phi} \mathbb{E}_{(\mu,\nu,c)\sim\mathcal{D}}[J(\hat{f}_\phi[\mu, \nu, c]; \mu, \nu, c)], \tag{10}$$

where J is the Kantorovich dual objective in (6) with the plugged in form of the other potential using (8). By replacing the expectation with the finite sum across samples $(\mu_1, \nu_1, c_1), \ldots, (\mu_N, \nu_N, c_N)$, we obtain the empirical objective, which can be optimized using gradient descent algorithms or stochastic gradient algorithms for a large-scale dataset.

3.2 Amortized Optimal Transport from Sliced Potentials

We start by reviewing some essential definitions, such as one-dimensional OT plan, sliced potentials, and projection functions in sliced OT.

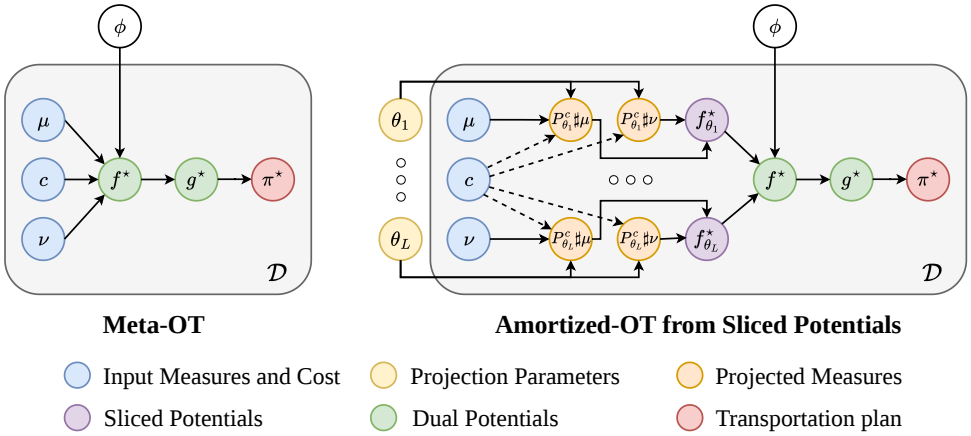

Figure 1: Amortized-OT uses sliced potentials to predict original dual potentials.

One-dimensional Optimal Transport. In one-dimension, when $c = h(x - y)$ with h is a strictly convex function, OT between μ and ν admits a closed-form [59]:

$$\pi^\star = (F_\mu^{-1}, F_\nu^{-1})\sharp\mathcal{U}([0,1]), \tag{11}$$

where $\mathcal{U}([0,1])$ is the uniform distribution on the interval $[0,1]$, F_μ^{-1} and F_ν^{-1} are quantile functions of μ and ν respectively. We then can access to the Kantorovich potential using complementary slackness condition:

$$f^\star(F_\mu^{-1}(u)) + g^\star(F_\nu^{-1}(u)) = h(F_\mu^{-1}(u) - F_\nu^{-1}(u)), \tag{12}$$

for $\mathcal{U}([0,1])$-almost every $u \in [0,1]$, this fact allows the OT plan and Kantorovich potentials in the discrete case to be computed extremely efficiently, with a time complexity of $\mathcal{O}((n+m)\log(n+m))$, where n and m denote the number of atoms in the two measures. A detailed algorithm for computing the OT plan and Kantorovich potentials in one-dimensional discrete settings can be found in Algorithm 1 of [63]. We also note that, alternatively, Kantorovich potentials in discrete cases can be obtained by taking the gradients of the optimal transport cost with respect to the weights of the two measures [22].

Sliced Optimal Transport. Sliced OT [56, 46] leverages the closed-form solution of optimal transport in one dimension through projections. Given the geometry of the space of atoms $\mathcal{X} \cup \mathcal{Y}$ of the measures and the ground metric c, a projection function $P_\theta^c : \mathcal{X} \cup \mathcal{Y} \to \mathcal{Z}$ maps the atoms to a one-dimensional space $\mathcal{Z}$, e.g., the real line $\mathbb{R}$, where $\theta \in \Theta$ denotes the projection parameters. Designing effective P_θ^c remains an active area of research. For instance, in Euclidean spaces ($\mathcal{X} \cup \mathcal{Y} \subset \mathbb{R}^d$) with the Euclidean ground metric $c(x,y) = \|x - y\|_2$, one can use linear projections $P_\theta^c(x) = \langle \theta, x \rangle$ with $\theta \in \mathbb{S}^{d-1}$ (the unit $(d-1)$-sphere). In spherical settings ($\mathcal{X} \cup \mathcal{Y} \subset \mathbb{S}^{d-1}$), spherical [10, 55] or stereographic [72] projections are applicable. Extensions to other domains, such as functions [30], manifolds, and products of manifolds [11, 48], have also been explored. Using the projection function, sliced OT projects μ and ν into $P_\theta^c\sharp\mu$ and $P_\theta^c\sharp\nu$, then obtains the 1D OT plan between them. Recent works have explored lifting the 1D OT plan back to the original space [42, 41, 69]. However, these methods focus on fast approximations and are not intended to recover the exact original OT. Furthermore, they do not operate in an amortized OT setting, as they are applied at the level of individual measure pairs.

Amortized model with sliced potentials. Given a set of projection parameters $\theta_1, \ldots, \theta_L$ ($L \geq 1$ is referred to as the number of projections), which can be created using Monte Carlo sampling [13] or quasi Monte Carlo sampling [47], we are able to obtain a set of sliced potentials $f_{\theta_1}^\star[\mu,\nu,c], \ldots, f_{\theta_L}^\star[\mu,\nu,c]$ using closed-form of one-dimensional OT problems between $P_{\theta_l}^c\sharp\mu$ and $P_{\theta_l}^c\sharp\nu$ for $l = 1, \ldots, L$. We then define an amortized model as follows:

$$\hat{f}_\phi[\mu,\nu,c] = \gamma_\phi(f_{\theta_1}^\star[\mu,\nu,c], \ldots, f_{\theta_L}^\star[\mu,\nu,c]), \tag{13}$$

where γ_ϕ is an operator that maps from the space of L functions in one dimension to the space of functions in the dimension of measures μ and ν. Since $f_{\theta_1}^\star[\mu,\nu,c], \ldots, f_{\theta_L}^\star[\mu,\nu,c]$ are "handcraft features" with a lot of information about transportation between μ and ν through one-dimensional slices, we propose to consider a simple linear model as follows: $\hat{f}_{\boldsymbol{\omega}}[\mu,\nu,c](x) = \sum_{l=1}^L \omega_l f_{\theta_l}^\star[\mu,\nu,c](P_{\theta_l}^c(x))$,

where $\boldsymbol{\omega} \in \mathbb{R}^L$ is the linear coefficients. We recall that the above construction of the amortized model is well-defined for both discrete and continuous cases. Importantly, the parameter $\boldsymbol{\omega}$ or ϕ of the proposed amortized model is independent of the input measures (e.g., their number of atoms), unlike in Meta-OT. Therefore, we are able to handle settings where the number of atoms varies.

Regression-based amortization. We now discuss the first amortization strategy: regression-based amortized OT (RA-OT). In this strategy, we assume that we can afford the computation of (entropic) OT during training. In particular, we solve the following optimization problem:

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^L} \mathbb{E}_{(\mu,\nu,c)\sim\mathcal{D}}[\|\hat{f}_{\boldsymbol{\omega}}[\mu,\nu,c] - f^\star[\mu,\nu,c]\|_2^2], \tag{14}$$

where $\|\cdot\|_2^2$ denotes the functional $\mathbb{L}_2$ norm. When μ is discrete with n atoms $x_1, \ldots, x_n$, the above optimization problem admits a closed-form. In particular, let $X_{\theta_l}[\mu,\nu,c] = (f^\star_{\theta_l}[\mu,\nu,c](P^c_{\theta_l}(x_1)), \ldots, f^\star_{\theta_l}[\mu,\nu,c](P^c_{\theta_l}(x_n)))^\top \in \mathbb{R}^n$, $X_\Theta[\mu,\nu,c] = (X_{\theta_1}[\mu,\nu,c], \ldots, X_{\theta_L}[\mu,\nu,c]) \in \mathbb{R}^{n\times L}$, and $Y[\mu,\nu,c] = (f^\star[\mu,\nu,c](x_1), \ldots, f^\star[\mu,\nu,c](x_n))^\top \in \mathbb{R}^n$, we can rewrite the optimization problem as:

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^L} \mathbb{E}_{(\mu,\nu,c)\sim\mathcal{D}}[\|X_\Theta[\mu,\nu,c]\boldsymbol{\omega} - Y[\mu,\nu,c]\|_2^2], \tag{15}$$

which has the optimal solution is the root of the following equation:

$$\mathbb{E}[X_\Theta[\mu,\nu,c]^\top X_\Theta[\mu,\nu,c]]\boldsymbol{\omega} = \mathbb{E}[X_\Theta[\mu,\nu,c]^\top Y[\mu,\nu,c]], \tag{16}$$

which has the closed-form solution:

$$\boldsymbol{\omega}^\star = \left(\mathbb{E}[X_\Theta[\mu,\nu,c]^\top X_\Theta[\mu,\nu,c]]\right)^{-1} \mathbb{E}[X_\Theta[\mu,\nu,c]^\top Y[\mu,\nu,c]], \tag{17}$$

where the expectation is under $\mathcal{D}$. When we observe a finite set of $(\mu_1,\nu_1,c_1), \ldots, (\mu_N,\nu_N,c_N)$, we can replace the expectation by the finite sum $\frac{1}{N}\sum_{i=1}^N$ as a Monte Carlo estimate for tractable computation. In practice, we can use other linear solvers of (16) rather than 17 for computational stability. The time complexity for training RA-OT involves the complexity for computing sliced potentials $\mathcal{O}(ML(n+m)(\log(n+m)+d))$ (P^c_θ has the time complexity of $\mathcal{O}(Ld(n+m))$), computing ground truth potential $\mathcal{O}(Nnm)$ for entropic OT or $\mathcal{O}((n+m)^3\log(n+m))$ for OT, and computing the optimal coefficient $\mathcal{O}(NnL^2 + L^3)$. For prediction on a new set $(\mu'_1,\nu'_1,c'_1), \ldots, (\mu'_M,\nu'_M,c'_M)$, the time complexity of RA-OT involves only the complexity for computing sliced potentials $\mathcal{O}(ML(n+m)(\log(n+m)+d))$ and applying linear prediction $\mathcal{O}(MnL)$. In the case of conventional OT, we can set up another regression problem for the other potential $g^\star$, while predicting $f^\star$ is sufficient for entropic OT as discussed in Section 2.

Objective-based amortization. For the objective-based amortized OT (OA-OT) strategy, we follow the same training scheme as Meta-OT with the new proposed amortized model. We consider the following optimization problem:

$$\min_{\boldsymbol{\omega} \in \mathbb{R}^L} \mathbb{E}_{(\mu,\nu,c)\sim\mathcal{D}}[J(\hat{f}_{\boldsymbol{\omega}}[\mu,\nu,c];\mu,\nu,c)], \tag{18}$$

where J is either the Kantorovich dual objective in (4) or in (6) in the plugged-in form of the other potential. Similar to Meta-OT, OA-OT is designed to address the entropic OT problem only. We also use a gradient-based algorithm for finding an estimate of $\boldsymbol{\omega}$ in OA-OT. For a set of N samples from $\mathcal{D}$, after computing sliced potentials with the time complexity of $\mathcal{O}(NL(n+m)(\log(n+m)+d))$, an estimate of gradient for $\boldsymbol{\omega}$ has $\mathcal{O}(Nmn)$ in time complexity. The time complexity is then scaled by $T > 0$ iterations of gradient updates. Compared to Meta-OT, OA-OT has a much smaller number of parameters; hence, it enjoys faster training. Given an estimate of $\boldsymbol{\omega}$, the time complexity of OA-OT during inference time is the same as RA-OT.

Remark 1 (Amortization Error and Amortization Gap) *As we considered linear amortized models, the amortization error* $\mathbb{E}_{(\mu,\nu,c)\sim\mathcal{D}}[\|\hat{f}_{\boldsymbol{\omega}}[\mu,\nu,c] - f^\star[\mu,\nu,c]\|_2^2]$ *and the amortization gap* $\mathbb{E}_{(\mu,\nu,c)\sim\mathcal{D}}[(J(\hat{f}_{\boldsymbol{\omega}}[\mu,\nu,c];\mu,\nu,c)] - J(f^\star[\mu,\nu,c];\mu,\nu,c))^2]$ *are minimized when* $f^\star[\mu,\nu,c]$ *lives in the linear spans of* $f^\star_{\theta_1}[\mu,\nu,c], \ldots, f^\star_{\theta_L}[\mu,\nu,c]$ *for* $\mathcal{D}$*-almost every* (μ,ν,c)*. Therefore, increasing the number of projections* L *can potentially enlarge the span generated by the sliced potentials, thereby improving the expressivity of the amortized model.*

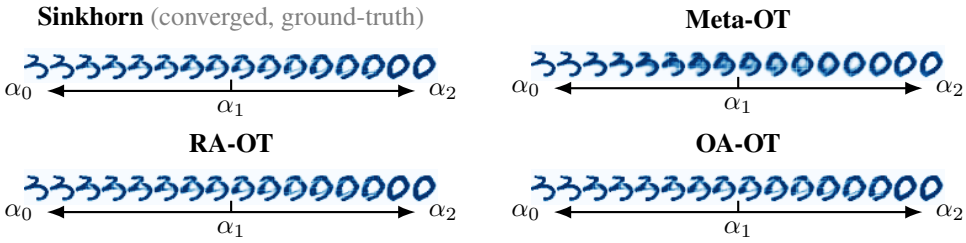

Figure 2: Wasserstein interpolations between MNIST test digits. (Top) Sinkhorn ground truth run to convergence and baseline Meta-OT. (Bottom) RA-OT and OA-OT (M=50, L=100). Both proposed methods produce interpolation sequences nearly identical to the ground truth.

4 Experiments

Following the experimental pipeline in [5], we evaluate the practical effectiveness of our proposed methods, RA-OT and OA-OT, across three representative settings that span a range of discrete optimal transport applications. First, in **MNIST grayscale transport**, we compute Wasserstein interpolations between digit images defined on a 2D Euclidean grid. Second, in **spherical supply–demand transport**, we consider a large-scale, non-Euclidean setting that maps global landmass distributions to population density over a sphere. Finally, in **color transfer**, we align image color palettes within a 3D RGB space. For these tasks, all methods are quantitatively assessed using the Root Mean Square Error (RMSE) of the predicted transport plan relative to the converged Sinkhorn ground truth, alongside total training and per-pair inference times. We compare our approaches against three recent baselines: Meta-OT [5], which predicts the plan via an MLP applied to raw measure weights, min-SWGG [42], a training-free baseline, and Min-STP [40], which learns an optimized projection direction. To ensure a rigorous evaluation, all trainable methods share an identical pool of 1,000 measure pairs per task with a 70/30 train-test split. Additionally, we evaluate the performance of our methods when applied to mini-batch OT conditional flow matching, focusing on low-dimensional 2D toy datasets to demonstrate accelerated pairing in generative modeling. Since this task deviates from the standard plan-prediction pipeline, we instead measure the training time and the NPE, benchmarking against Independent CFM (I-CFM) and OT-CFM [71]. Across all experiments, the number of sliced projections is fixed at L=100. For the first three standard OT tasks, we report results for varying training sizes $M \in \{10, 20, 50, 200\}$, focusing on M=50 for qualitative visualizations. In contrast, for the OT-CFM task, we maintain a fixed pre-training size of M=50. Experiment details are provided in Appendix A. Additional ablation results on the number of projections L are given in Appendix B, and further visualizations for all tasks are in Appendix E. We also provide additional experiments in Appendix D, including evaluations on measures with varying numbers of atoms, heterogeneous measure groups, Sinkhorn warm-starting, distribution shift, marginal feasibility and rounding, and the training-inference time trade-off.

4.1 MNIST

Pairs of MNIST digit images offer a standard test for amortized OT, sharing a common 28×28 grid geometry but varying in mass distributions. Each image is modeled as a discrete measure where pixel intensities dictate the weights of the n=784 atoms. The transport cost is the squared Euclidean distance between coordinates, and the regularization parameter is set to ε=0.1. As shown in Table 1, RA-OT and OA-OT remain accurate even with very limited training data ($M = 10, 20$), outperforming Meta-OT in this low-data regime. They also train significantly faster due to their lightweight parameterization. In contrast, single-projection baselines such as Min-STP and min-SWGG struggle to capture the spatial structure of image transport. This improvement is also reflected in Figure 8 (Appendix E), where the interpolations better match the Sinkhorn reference.

4.2 Spherical Transport

To evaluate the proposed methods in a non-Euclidean domain, we construct a supply-demand transportation problem on the unit sphere $\mathbb{S}^2 := \{x \in \mathbb{R}^3 : \|x\| = 1\}$. Supply capacities are assumed to be uniform across n=100 locations sampled from Earth's landmass, while demand weights are drawn from Earth's population density across m=10,000 locations based on the CC-licensed dataset

Table 1: RMSE, training and inference time on the **MNIST** grayscale transport task across varying number of training data M (ε=0.1, N=300 test pairs, L=100). The results are reported as the mean ± standard deviation. The label "no train" indicates that no prior training is required. **Bold** values denote the best performance in each block.

M	Method	RMSE ($\times 10^{-6}$, ↓)	Train (s)	Infer (ms,↓)	M	Method	RMSE ($\times 10^{-6}$, ↓)	Train (s)	Infer (ms,↓)
	Meta-OT [5]	16.16 ± 4.90	38.05	$\mathbf{2.46 \pm 0.20}$		Meta-OT [5]	15.54 ± 4.74	37.11	$\mathbf{2.39 \pm 0.29}$
	Min-STP [40]	93.51 ± 10.18	199.52	6.18 ± 0.37		Min-STP [40]	93.50 ± 10.18	198.01	6.25 ± 0.48
10	min-SWGG [42]	89.85 ± 8.64	no train	10.01 ± 0.73	50	min-SWGG [42]	89.85 ± 8.64	no train	9.72 ± 0.41
	RA-OT	8.23 ± 3.55	**1.36**	44.33 ± 2.91		**RA-OT**	7.77 ± 3.06	**3.03**	39.36 ± 3.61
	OA-OT	$\mathbf{6.16 \pm 2.53}$	16.65	42.54 ± 2.95		**OA-OT**	$\mathbf{6.02 \pm 2.52}$	15.78	38.92 ± 2.23
	Meta-OT [5]	16.68 ± 4.98	37.27	$\mathbf{2.44 \pm 0.39}$		Meta-OT [5]	14.12 ± 4.70	37.22	$\mathbf{2.37 \pm 0.40}$
	Min-STP [40]	93.51 ± 10.18	198.56	6.21 ± 0.41		Min-STP [40]	93.51 ± 10.18	197.86	6.17 ± 0.49
20	min-SWGG [42]	89.85 ± 8.64	no train	9.70 ± 0.47	200	min-SWGG [42]	89.85 ± 8.64	no train	9.73 ± 0.61
	RA-OT	7.49 ± 3.19	**1.44**	40.29 ± 2.74		**RA-OT**	7.74 ± 3.07	**11.12**	40.30 ± 4.39
	OA-OT	$\mathbf{6.04 \pm 2.50}$	15.60	40.65 ± 2.95		**OA-OT**	$\mathbf{6.11 \pm 2.54}$	16.39	40.48 ± 4.16

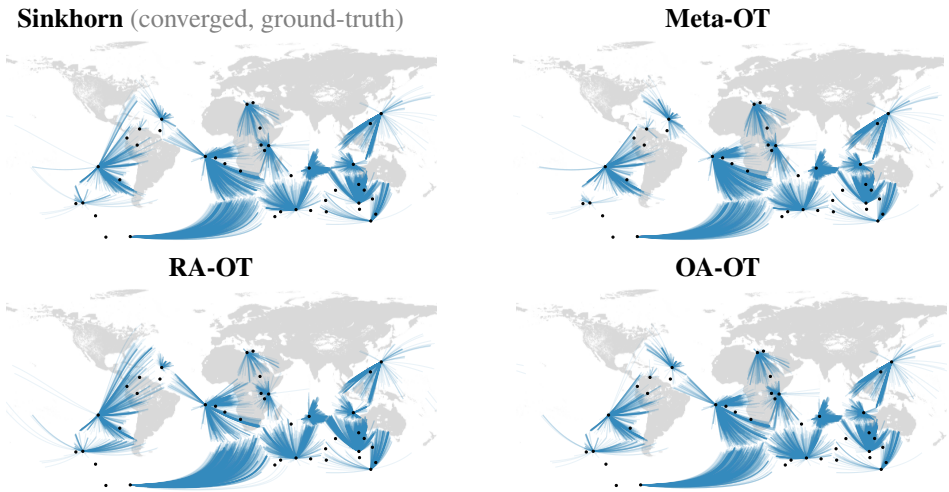

Figure 3: Prediction on a held-out spherical supply–demand instance (M=50, L=100, ε=0.5). (Top) Sinkhorn ground truth and baseline Meta-OT. (Bottom) our methods: RA-OT and OA-OT Black dots denote active supply locations; blue arcs show great-circle geodesics from each demand location to its assigned supply location under the predicted plan.

Table 2: RMSE, training and inference time on the **spherical supply-demand transport** task across varying number of training data M (ε=0.5, n=100, m=10,000, N=300, L=100). The results are reported as the mean ± standard deviation. The label "no train" indicates that no prior training is required. **Bold** values denote the best in each block.

M	Method	RMSE ($\times 10^{-7}$, ↓)	Train (s)	Infer (ms,↓)	M	Method	RMSE ($\times 10^{-7}$, ↓)	Train (s)	Infer (ms,↓)
	Meta-OT [5]	6.83 ± 2.49	52.08	15.94 ± 2.14		Meta-OT [5]	4.42 ± 1.55	52.07	16.09 ± 2.17
	Min-STP [40]	115.07 ± 0.54	46.06	$\mathbf{7.51 \pm 7.23}$		Min-STP [40]	115.10 ± 0.54	46.38	$\mathbf{7.88 \pm 1.22}$
10	min-SWGG [42]	112.15 ± 0.99	no train	34.62 ± 2.70	50	min-SWGG [42]	112.15 ± 0.99	no train	33.25 ± 1.71
	RA-OT	7.92 ± 2.77	**1.05**	41.16 ± 3.11		**RA-OT**	7.82 ± 1.89	**2.53**	41.96 ± 6.29
	OA-OT	$\mathbf{4.80 \pm 1.96}$	18.60	39.92 ± 3.12		**OA-OT**	$\mathbf{3.93 \pm 1.92}$	19.37	41.03 ± 2.76
	Meta-OT [5]	6.03 ± 2.20	51.99	15.57 ± 1.22		Meta-OT [5]	$\mathbf{3.62 \pm 1.14}$	52.12	15.17 ± 2.54
	Min-STP [40]	115.06 ± 0.54	46.39	$\mathbf{7.95 \pm 1.41}$		Min-STP [40]	115.05 ± 0.54	46.41	$\mathbf{7.93 \pm 1.22}$
20	min-SWGG [42]	112.15 ± 0.99	no train	33.85 ± 2.49	200	min-SWGG [42]	112.15 ± 0.99	no train	33.59 ± 1.99
	RA-OT	7.82 ± 2.25	**1.31**	40.70 ± 3.25		**RA-OT**	7.55 ± 1.88	**8.74**	39.40 ± 1.43
	OA-OT	$\mathbf{4.24 \pm 1.98}$	18.75	39.65 ± 3.26		**OA-OT**	3.66 ± 1.56	21.32	40.03 ± 3.83

[25]. The ground metric is defined by the geodesic cost $c(x, y) = \arccos(\langle x, y \rangle)$ with an entropic regularization parameter of ε=0.5. To properly accommodate the spherical geometry within our framework, we utilize stereographic projections [72] to define the slicing family for both RA-OT and OA-OT. As reported in Table 2, both methods generalize well with few training samples and converge quickly in practice. This task highlights the importance of geometry: while Min-STP and min-SWGG fail to respect the spherical structure, our approach remains close to the Sinkhorn solution. The transport patterns shown in Figure 9 (Appendix E) clearly illustrate this behavior.

4.3 Color Transfer

Color transfer is formulated as an OT problem by representing images as weighted point clouds in the normalized RGB space $[0, 1]^3$. Following the experimental setup of Meta-OT, we collect approximately 200 public domain images from WikiArt to construct the dataset. To retain a discrete formulation, images are quantized into K=500 distinct color clusters using mini-batch k-means,

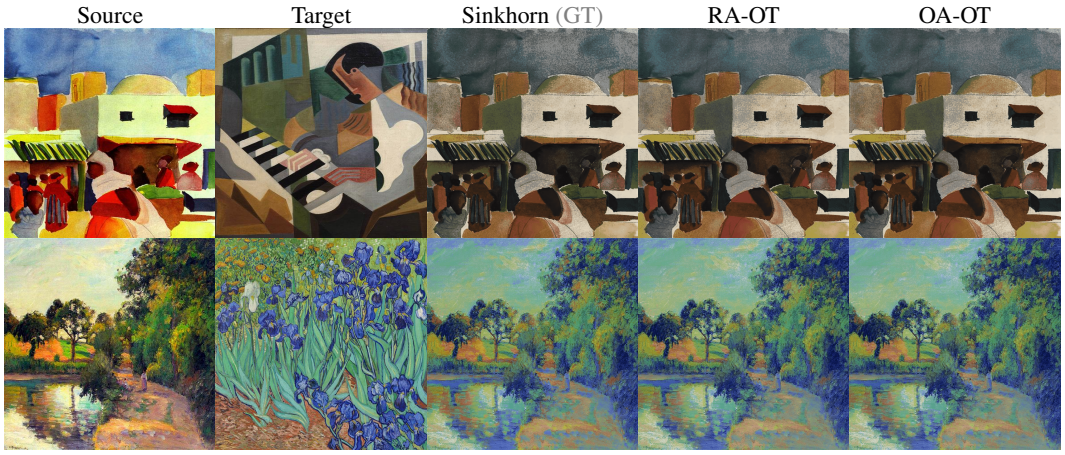

Figure 4: Color transfer results on held-out test pairs (M=50, L=100, ε=0.005, K=500 clusters). Each row corresponds to one source–target pair. Columns show: source α, target β, Sinkhorn ground truth, RA-OT, and OA-OT.

Table 3: RMSE, training and inference time on the **color transfer** task across varying number of training data M (ε=0.005, K=500 clusters, N=300 test pairs, L=100). The results are reported as the mean ± standard deviation. The label "no train" indicates that no prior training is required. **Bold** values denote the best performance in each block.

M	Method	RMSE ($\times 10^{-6}$, ↓)	Train (s)	Infer (ms,↓)	M	Method	RMSE ($\times 10^{-6}$, ↓)	Train (s)	Infer (ms,↓)
	Meta-OT [5]	36.83 ± 10.97	33.32	16.83 ± 1.33		Meta-OT [5]	33.16 ± 10.45	32.91	15.16 ± 0.64
	Min-STP [40]	80.93 ± 3.24	49.51	$\mathbf{4.02 \pm 0.27}$		Min-STP [40]	81.04 ± 3.42	50.00	$\mathbf{4.10 \pm 0.20}$
10	min-SWGG [42]	77.69 ± 3.19	no train	4.25 ± 0.21	50	min-SWGG [42]	77.69 ± 3.19	no train	4.34 ± 0.17
	RA-OT	10.73 ± 6.35	**1.95**	18.41 ± 1.21		**RA-OT**	9.99 ± 5.61	**7.40**	17.40 ± 0.83
	OA-OT	$\mathbf{9.78 \pm 5.78}$	17.52	18.77 ± 1.31		**OA-OT**	$\mathbf{9.00 \pm 5.02}$	18.13	17.76 ± 0.79
	Meta-OT [5]	34.73 ± 10.94	34.63	15.03 ± 0.70		Meta-OT [5]	33.23 ± 10.89	33.11	15.44 ± 1.20
	Min-STP [40]	80.97 ± 3.33	49.92	$\mathbf{4.07 \pm 0.23}$		Min-STP [40]	81.10 ± 3.38	49.68	$\mathbf{4.08 \pm 0.22}$
20	min-SWGG [42]	77.69 ± 3.19	no train	4.30 ± 0.17	200	min-SWGG [42]	77.69 ± 3.19	no train	4.27 ± 0.19
	RA-OT	9.74 ± 5.68	**3.11**	18.02 ± 1.34		**RA-OT**	10.21 ± 5.53	26.57	18.08 ± 1.40
	OA-OT	$\mathbf{9.31 \pm 5.24}$	17.92	18.45 ± 1.20		**OA-OT**	$\mathbf{9.00 \pm 5.05}$	**20.37**	18.25 ± 0.71

Table 4: Quantitative results on 2D Conditional Flow Matching tasks. We report both training and pretraining time.

	Gaussian → 8gaussians				**Gaussian → moons**				**Gaussian → scurve**			
Method	$\mathbf{W_2^2}$ ↓	**NPE** ↓	**Train (s)**	**Pretrain**	$\mathbf{W_2^2}$ ↓	**NPE** ↓	**Train (s)**	**Pretrain**	$\mathbf{W_2^2}$ ↓	**NPE** ↓	**Train (s)**	**Pretrain**
I-CFM	0.2800	0.2590	**168.7**	0.0s	0.3901	0.5380	**45.6**	0.0s	0.1113	1.9230	**44.5**	0.0s
OT-CFM	**0.1956**	**0.0108**	1054.7	0.0s	**0.2668**	**0.0142**	1513.1	0.0s	**0.0991**	**0.0243**	992.0	0.0s
RA-OT	0.5072	0.0144	411.8	**2.6s**	0.5569	**0.0619**	331.7	**1.0s**	0.4675	0.0400	299.3	**6.3s**
OA-OT	0.5041	0.0142	413.8	14.0s	0.5459	0.0631	319.7	14.6s	0.4672	0.0425	294.0	14.4s

where centroids act as atoms and normalized cluster occupancies as weights. The ground metric is the squared Euclidean distance with ε=0.005. Transferred images are generated by updating source pixels via the target centroids weighted by the row-normalized predicted transport plan. As shown in Table 3, RA-OT and OA-OT achieve strong performance across all data regimes, with clear advantages when training data is limited. They also require less training time compared to Meta-OT. The gap with Min-STP and min-SWGG suggests that single projections are insufficient for modeling color distributions. As seen in Figure 4, this leads to more faithful color reconstruction.

4.4 Mini-batch Optimal Transport Conditional Flow Matching

OT-CFM [71] leverages mini-batch OT to straighten flow trajectories in flow matching, thereby reducing inference cost. However, computing exact OT remains expensive and constitutes a major training bottleneck. To address this, we introduce a more efficient alternative by integrating RA-OT and OA-OT into the conditional flow matching framework, significantly accelerating training while preserving near-straight trajectories. Following [71], we map $\mathcal{N}(0, I)$ to 2D targets (8gaussians, moons, scurve) using a network with three 64-unit SELU layers and a linear head, trained for 20000 steps (batch size 512, $\sigma = 0.1$) with AdamW (learning rate 10^{-3}, weight decay 10^{-5}). We compare RA-OT and OA-OT against I-CFM and OT-CFM, where amortized methods use $M = 50$ pretraining batches and $L = 100$ projections. All methods employ fixed-step RK4 (101 steps); performance is evaluated using Sinkhorn W_2^2 at $t = 1$ and path straightness via NPE. As shown in Table 4, our methods achieve $2.5\times$–$4.5\times$ training speedups over OT-CFM while maintaining comparable

trajectory quality, with NPE close to OT-CFM and up to 1–2 orders of magnitude lower than I-CFM. Although slight degradation in Sinkhorn W_2^2 is observed in more complex settings, the overall trade-off highlights the scalability of our approach. The visualizations in Figures 5, 6, and 7 further illustrate the learned trajectories across the three 2D targets. Both RA-OT and OA-OT produce straighter and less entangled trajectories than I-CFM while closely resembling those of OT-CFM, consistent with the low NPE values in Table 4. To further examine scalability, we additionally evaluate our methods on high-dimensional CIFAR-10 image generation. The results in Appendix C show that RA-OT and OA-OT reduce the computational overhead of OT-based flow matching while maintaining competitive generation quality.

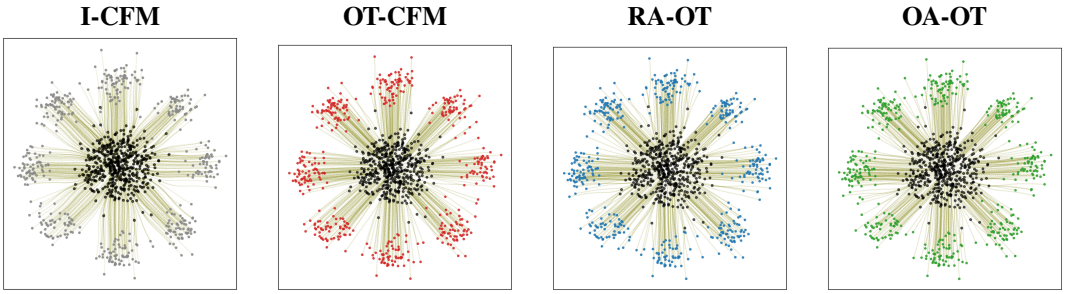

Figure 5: Generated flow trajectories moving from a 2D Gaussian prior to the **8-Gaussians** target. (Left to Right) Baseline I-CFM, OT-CFM, and our proposed amortized methods. Black dots signify source samples. The olive tracks map the predicted integration paths. Our solutions successfully reproduce the straight, uncrossed optimal transport paths of OT-CFM, yielding significantly lower collision energy than standard I-CFM.

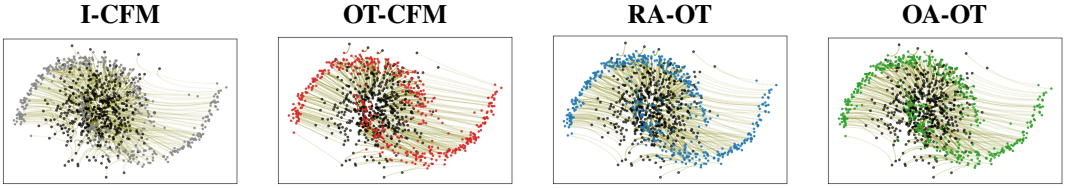

Figure 6: Flow trajectories transforming a 2D Gaussian prior to the **Moons** target distribution. (Left to Right) Baselines I-CFM and OT-CFM, followed by our proposed RA-OT and OA-OT methods. Both amortized variants effectively map unentangled optimal integration flows (olive paths) across the dual clusters, matching exact OT capabilities.

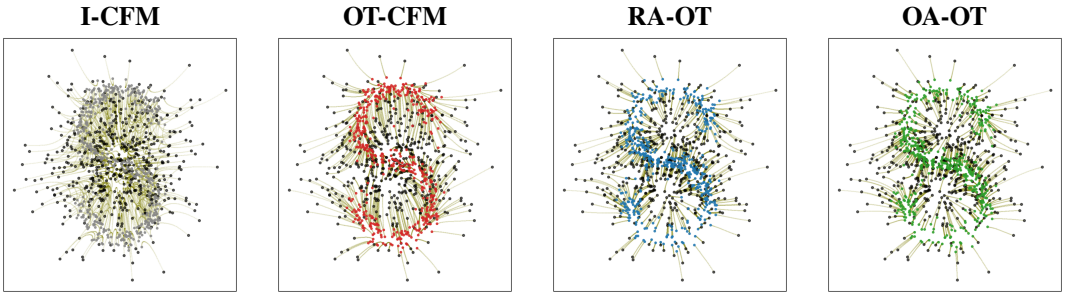

Figure 7: Visualized flow trajectories mapping a 2D Gaussian prior into the **S-Curve** distribution. (Left to Right) I-CFM, OT-CFM, RA-OT, and OA-OT. Our amortized approaches robustly straighten the generative pathways, practically bypassing the severe integration turbulence traditionally observed in uncoupled I-CFM trajectories.

5 Conclusion

The paper presents two novel amortized OT methods, RA-OT and OA-OT, which leverage sliced OT to efficiently predict OT plans across multiple measure pairs. RA-OT uses regression to map sliced OT potentials to original OT potentials, while OA-OT optimizes the Kantorovich dual within the span of sliced OT potentials. Both approaches enable rapid and accurate solutions for repeated OT problems, are parsimonious, and do not depend on the structure or size of the measures. Experiments on MNIST digit transport, color transfer, spherical supply–demand problems, and conditional flow matching demonstrate the practical effectiveness and generality of the methods.

Acknowledgement

Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing.

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning*, pages 40–49. PMLR, 2018.

[2] J. Altschuler, J. Niles-Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[3] D. Alvarez-Melis and N. Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

[4] B. Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023.

[5] B. Amos, G. Luise, S. Cohen, and I. Redko. Meta optimal transport. In *International Conference on Machine Learning*, pages 791–813. PMLR, 2023.

[6] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[7] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[8] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019.

[9] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA*, 8(4):657–676, 2019.

[10] C. Bonet, P. Berg, N. Courty, F. Septier, L. Drumetz, and M.-T. Pham. Spherical sliced-Wasserstein. *International Conference on Learning Representations*, 2023.

[11] C. Bonet, L. Drumetz, and N. Courty. Sliced-Wasserstein distances and flows on Cartan-Hadamard manifolds. *Journal of Machine Learning Research*, 26(32):1–76, 2025.

[12] N. Bonneel and J. Digne. A survey of optimal transport for computer graphics and computer vision. In *Computer Graphics Forum*, volume 42, pages 439–460. Wiley Online Library, 2023.

[13] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 1(51):22–45, 2015.

[14] C. Bunne, L. Papaxanthos, A. Krause, and M. Cuturi. Proximal optimal transport modeling of population dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 6511–6528. PMLR, 2022.

[15] C. Bunne, S. G. Stark, G. Gut, J. S. Del Castillo, M. Levesque, K.-V. Lehmann, L. Pelkmans, A. Krause, and G. Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, 20(11):1759–1768, 2023.

[16] M. Catalano, H. Lavenant, A. Lijoi, and I. Prünster. A Wasserstein index of dependence for random measures. *Journal of the American Statistical Association*, 119(547):2396–2406, 2024.

[17] M. Catalano, A. Lijoi, and I. Prünster. Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics*, 49(5):2916–2947, 2021.

[18] N. Courty, R. Flamary, and M. Ducoffe. Learning Wasserstein embeddings. In *International Conference on Learning Representations*, 2018.

[19] N. Courty, R. Flamary, A. Habrard, and A. Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, pages 3730–3739, 2017.

[20] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[21] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[22] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning*, pages 685–693. PMLR, 2014.

[23] B. B. Damodaran, B. Kellenberger, R. Flamary, D. Tuia, and N. Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

[24] D. Dowson and B. Landau. The Fréchet distance between multivariate Normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982.

[25] E. Doxsey-Whitfield, K. MacManus, S. B. Adamo, L. Pistolesi, J. Squires, O. Borkovska, and S. R. Baptista. Taking advantage of the improved availability of census data: A first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3):226–234, 2015.

[26] B. Engquist and B. D. Froese. Application of the Wasserstein metric to seismic signals. *Communications in Mathematical Sciences*, 12(5):979–988, 2014.

[27] K. Fatras, T. Sejourne, R. Flamary, and N. Courty. Unbalanced minibatch optimal transport; applications to domain adaptation. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3186–3197. PMLR, 18–24 Jul 2021.

[28] K. Fatras, Y. Zine, R. Flamary, R. Gribonval, and N. Courty. Learning with minibatch Wasserstein: asymptotic and gradient properties. In *AISTATS 2020-23nd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 1–20, 2020.

[29] J. Feydy, B. Charlier, F.-X. Vialard, and G. Peyré. Optimal transport for diffeomorphic registration. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part I 20*, pages 291–299. Springer, 2017.

[30] R. C. Garrett, T. Harris, Z. Wang, and B. Li. Validating climate models with spherical convolutional Wasserstein distance. *Advances in Neural Information Processing Systems*, 37:59119–59149, 2024.

[31] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR, 2019.

[32] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. *Advances in neural information processing systems*, 29, 2016.

[33] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR, 2018.

[34] D. Haviv, R. Z. Kunes, T. Dougherty, C. Burdziak, T. Nawy, A. Gilbert, and D. Pe'er. Wasserstein wormhole: Scalable optimal transport distance with Transformer. In *Forty-first International Conference on Machine Learning*, 2024.

[35] P. He, O. Khangaonkar, H. Pirsiavash, Y. Bai, and S. Kolouri. Sinkhorn-drifting generative models. *arXiv preprint arXiv:2603.12366*, 2026.

[36] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence*, pages 862–872. PMLR, 2020.

[37] S. Kolouri, N. Naderializadeh, G. K. Rohde, and H. Hoffmann. Wasserstein embedding for graph learning. In *International Conference on Learning Representations*, 2021.

[38] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59, 2017.

[39] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[40] X. Liu, E. Akbari, R. Diaz Martin, N. NaderiAlizadeh, and S. Kolouri. Efficient transferable optimal transport via min-sliced transport plans. *arXiv e-prints*, pages arXiv–2511, 2025.

[41] X. Liu, R. D. Martin, Y. Bai, A. Shahbazi, M. Thorpe, A. Aldroubi, and S. Kolouri. Expected sliced transport plans. In *The Thirteenth International Conference on Learning Representations*, 2025.

[42] G. Mahey, L. Chapel, G. Gasso, C. Bonet, and N. Courty. Fast optimal transport through sliced Wasserstein generalized geodesics. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 35350–35385, 2023.

[43] A. Makkuva, A. Taghvaei, S. Oh, and J. Lee. Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR, 2020.

[44] T. Manole, S. Balakrishnan, J. Niles-Weed, and L. Wasserman. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024.

[45] C. Moosmüller and A. Cloninger. Linear optimal transport embedding: provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023.

[46] K. Nguyen. An introduction to sliced optimal transport: foundations, advances, extensions, and applications. *Foundations and Trends® in Computer Graphics and Vision*, 17(3-4):171–391, 2025.

[47] K. Nguyen, N. Bariletto, and N. Ho. Quasi-Monte Carlo for 3d sliced Wasserstein. In *The Twelfth International Conference on Learning Representations*, 2024.

[48] K. Nguyen and P. Mueller. Summarizing nonparametric Bayesian mixture posteriors–sliced optimal transport metrics for Gaussian mixtures. *Journal of Computational and Graphical Statistics*, (just-accepted):1–22, 2026.

[49] K. Nguyen, D. Nguyen, T. Pham, and N. Ho. Improving mini-batch optimal transport via partial transportation. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[50] K. Nguyen, H. Nguyen, and N. Ho. Fast estimation of Wasserstein distances via regression on sliced Wasserstein distances. In *The Fourteenth International Conference on Learning Representations*, 2026.

[51] J. B. Orlin. A polynomial time primal network simplex algorithm for minimum cost flows. *Mathematical Programming*, 78(2):109–129, 1997.

[52] G. Patrini, R. van den Berg, P. Forre, M. Carioni, S. Bhargav, M. Welling, T. Genewein, and F. Nielsen. Sinkhorn autoencoders. In *Uncertainty in Artificial Intelligence*, pages 733–743. PMLR, 2020.

[53] G. Peyré, M. Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[54] A.-A. Pooladian, H. Ben-Hamu, C. Domingo-Enrich, B. Amos, Y. Lipman, and R. T. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning*, pages 28100–28127. PMLR, 2023.

[55] M. Quellmalz, R. Beinert, and G. Steidl. Sliced optimal transport on the sphere. *Inverse Problems*, 39(10):105005, 2023.

[56] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, pages 435–446. Springer, One New York Plaza, Suite 4600, New York, NY 10004-1562, 2012.

[57] P. Rigollet and A. J. Stromme. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 53(1):61–90, 2025.

[58] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 59–66. IEEE, 1998.

[59] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[60] M. Scetbon and M. Cuturi. Low-rank optimal transport: Approximation, statistics and debiasing. *Advances in Neural Information Processing Systems*, 35:6802–6814, 2022.

[61] M. Scetbon, M. Cuturi, and G. Peyré. Low-rank Sinkhorn factorization. In *International Conference on Machine Learning*, pages 9344–9354. PMLR, 2021.

[62] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019.

[63] T. Séjourné, F.-X. Vialard, and G. Peyré. Faster unbalanced optimal transport: Translation invariant sinkhorn and 1-d frank-wolfe. In *International Conference on Artificial Intelligence and Statistics*, pages 4995–5021. PMLR, 2022.

[64] R. Shu. Amortized optimization `http://ruishu.io/2017/11/07/amortized-optimization/`, 2017.

[65] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

[66] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11, 2015.

[67] J. Solomon, G. Peyré, V. G. Kim, and S. Sra. Entropic metric alignment for correspondence problems. *ACM Transactions on Graphics (TOG)*, 35(4):72, 2016.

[68] M. Sommerfeld, J. Schrieber, Y. Zemel, and A. Munk. Optimal transport: Fast probabilistic approximation with exact solvers. *Journal of Machine Learning Research*, 20:105–1, 2019.

[69] E. Tanguy, L. Chapel, and J. Delon. Sliced optimal transport plans. *arXiv preprint arXiv:2508.01243*, 2025.

[70] I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

[71] A. Tong, K. FATRAS, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. Expert Certification.

[72] H. Tran, Y. Bai, A. Kothapalli, A. Shahbazi, X. Liu, R. D. Martin, and S. Kolouri. Stereographic spherical sliced Wasserstein distances. *International Conference on Machine Learning*, 2024.

[73] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[74] C. Villani. *Optimal transport: old and new*, volume 338. Springer, One New York Plaza, Suite 4600, New York, NY 10004-1562, 2009.

[75] F. Wu, N. Courty, S. Jin, and S. Z. Li. Improving molecular representation learning with metric learning-enhanced optimal transport. *Patterns*, 4(4), 2023.

[76] J. Zhu, A. Guha, D. Do, M. Xu, X. Nguyen, and D. Zhao. Functional optimal transport: regularized map estimation and domain adaptation for functional data. *Journal of Machine Learning Research*, 25(276):1–49, 2024.

A Experiment Details

In this section, we detail the architectures, hyperparameter settings, and optimization constraints for all methods utilized in the main experiments. All configuration values are dynamically loaded from a centralized configuration file to ensure exact reproducibility across multiple experimental runs. To guarantee a fair evaluation, methods requiring gradient-based optimization are allocated an equivalent training budget (5,000 steps). Optimizers (Adam) and learning rates werefixed per algorithm based on their respective convergence properties. The entropic regularization parameter (ε) defines the inherent Sinkhorn OT problem and is maintained uniformly across all baselines for consistent ground truth definition ($\varepsilon = 0.1$ for Grayscale, $\varepsilon = 0.5$ for spherical transport, and $\varepsilon = 0.005$ for color transfer). All experiments were conducted on a single Google Colab T4 GPU, except for the CIFAR-10 fine-tuning, which utilized a 40GB A100 GPU.

Table 5: Hyperparameters for **RA-OT** (Regression-based Amortized OT).

Parameter	**MNIST Grayscale**	**World Pair**	**Color Transfer**
Number of Projections (L)	100	100	100
Ridge Coefficient (λ)	1×10^{-3}	1×10^{-3}	1×10^{-3}

Table 6: Hyperparameters for **OA-OT** (Objective-based Amortized OT).

Parameter	**MNIST Grayscale**	**World Pair**	**Color Transfer**
Number of Projections (L)	100	100	100
Training Iterations (T)	5,000	5,000	5,000
Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}

Table 7: Hyperparameters for the **Meta-OT** baseline [5].

Parameter	**MNIST Grayscale**	**World Pair**	**Color Transfer**
Architecture Model	MLP	MLP	PointCloud Encoder
Hidden Layers / Units	3 Layers	3 Layers, 512 Units	256 Enc Dim, 512 Head
Training Iterations (T)	5,000	5,000	5,000
Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}

B Ablation Study: Effect of the Number of Projections L

To study the effect of the number of projections, we vary $L \in \{3, 5, 10, 20, 50, 100\}$. As shown in Table 9, both RA-OT and OA-OT benefit from increasing L, with clear improvements when moving away from very small values such as $L = 3$ or 5. With too few projections, the model cannot capture enough structural information, which leads to weaker performance. This is most noticeable in the Color Transfer task, where the data distribution is more complex and requires a richer representation.

Table 8: Hyperparameters for the amortized **Min-STP** baseline [40]. α denotes the inherent temperature parameter used during continuous soft-sorting.

Parameter	**MNIST Grayscale**	**World Pair**	**Color Transfer**
Encoder/Context Dim	32	32	32
Hidden Dim	128	128	64
Training Iterations (T)	5,000	5,000	5,000
Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}
Soft-Sort Temperature (α)	0.05	0.05	0.05

Table 9: Ablation study on the effect of the number of sliced projections $L \in \{3, 5, 10, 20, 50, 100\}$. The performance of RA-OT and OA-OT is evaluated across three tasks: **MNIST** grayscale transport, **Spherical** supply-demand, and **Color Transfer**. Results are reported in terms of Plan RMSE ($\times 10^{-6}$), training time, and inference time. **Bold** indicates the best metric between the two methods

L	Task	Method	RMSE ($\times 10^{-6}$, ↓)	Train (s)	Infer (ms)
3	MNIST	RA-OT	9.25 ± 3.73	$\mathbf{3.22}$	37.47 ± 3.91
		OA-OT	$\mathbf{8.50 \pm 3.55}$	15.35	$\mathbf{37.38 \pm 2.65}$
	Spherical	RA-OT	0.71 ± 0.17	$\mathbf{2.17}$	$\mathbf{30.51 \pm 4.43}$
		OA-OT	$\mathbf{0.41 \pm 0.16}$	18.62	34.20 ± 5.70
	Color	RA-OT	25.60 ± 10.42	$\mathbf{6.68}$	$\mathbf{16.48 \pm 0.96}$
		OA-OT	$\mathbf{23.96 \pm 6.79}$	16.86	17.29 ± 1.15
5	MNIST	RA-OT	8.80 ± 3.53	$\mathbf{3.62}$	40.89 ± 3.02
		OA-OT	$\mathbf{8.01 \pm 3.41}$	16.10	$\mathbf{40.76 \pm 3.06}$
	Spherical	RA-OT	0.70 ± 0.18	$\mathbf{2.37}$	$\mathbf{31.90 \pm 4.62}$
		OA-OT	$\mathbf{0.41 \pm 0.16}$	18.60	33.11 ± 6.03
	Color	RA-OT	$\mathbf{13.16 \pm 6.95}$	$\mathbf{7.25}$	18.69 ± 1.75
		OA-OT	14.83 ± 6.22	18.93	$\mathbf{18.67 \pm 0.89}$
10	MNIST	RA-OT	7.85 ± 3.22	$\mathbf{3.20}$	$\mathbf{37.56 \pm 2.80}$
		OA-OT	$\mathbf{6.34 \pm 2.63}$	14.44	37.90 ± 2.74
	Spherical	RA-OT	0.67 ± 0.18	$\mathbf{2.19}$	$\mathbf{32.01 \pm 4.35}$
		OA-OT	$\mathbf{0.42 \pm 0.17}$	18.54	35.13 ± 5.99
	Color	RA-OT	9.63 ± 5.47	$\mathbf{6.48}$	17.14 ± 1.59
		OA-OT	$\mathbf{9.38 \pm 5.23}$	17.14	$\mathbf{17.08 \pm 1.03}$
20	MNIST	RA-OT	7.62 ± 3.11	$\mathbf{3.18}$	38.11 ± 2.45
		OA-OT	$\mathbf{6.05 \pm 2.53}$	15.35	$\mathbf{37.83 \pm 2.23}$
	Spherical	RA-OT	0.74 ± 0.19	$\mathbf{2.33}$	32.06 ± 4.97
		OA-OT	$\mathbf{0.38 \pm 0.16}$	18.23	$\mathbf{30.89 \pm 4.24}$
	Color	RA-OT	9.39 ± 5.50	$\mathbf{7.03}$	17.64 ± 1.28
		OA-OT	$\mathbf{9.11 \pm 5.11}$	17.31	$\mathbf{17.48 \pm 1.37}$
50	MNIST	RA-OT	7.66 ± 3.05	$\mathbf{2.98}$	37.73 ± 3.34
		OA-OT	$\mathbf{6.03 \pm 2.52}$	15.19	$\mathbf{37.71 \pm 2.40}$
	Spherical	RA-OT	0.78 ± 0.19	$\mathbf{2.46}$	35.53 ± 4.06
		OA-OT	$\mathbf{0.39 \pm 0.19}$	18.47	$\mathbf{33.04 \pm 3.95}$
	Color	RA-OT	9.76 ± 5.62	$\mathbf{6.91}$	17.52 ± 1.51
		OA-OT	$\mathbf{9.04 \pm 5.02}$	17.48	$\mathbf{17.51 \pm 1.17}$
100	MNIST	RA-OT	7.77 ± 3.06	$\mathbf{3.03}$	39.36 ± 3.61
		OA-OT	$\mathbf{6.02 \pm 2.52}$	15.78	$\mathbf{38.92 \pm 2.23}$
	Spherical	RA-OT	0.78 ± 0.19	$\mathbf{2.53}$	41.96 ± 6.29
		OA-OT	$\mathbf{0.39 \pm 0.19}$	19.37	$\mathbf{41.03 \pm 2.76}$
	Color	RA-OT	9.99 ± 5.61	$\mathbf{7.40}$	$\mathbf{17.40 \pm 0.83}$
		OA-OT	$\mathbf{9.00 \pm 5.02}$	18.13	17.76 ± 0.79

As L increases, the error decreases quickly at first, and the results also become more stable across runs. However, after around $L = 20$, the improvements start to level off, and adding more projections only brings small gains. This suggests that a moderate number of projections is already enough to capture most of the important structure in the data. Importantly, increasing L does not significantly affect efficiency. Both training and inference times remain nearly the same across all values of L, showing that the method scales well with the number of projections. In practice, this means we can use a reasonably large L to ensure good performance without worrying about computational cost.

C High-dimensional Flow Matching Fine-tuning on CIFAR-10

In addition to 2D experiments, we investigate the scalability and practical advantages of our amortized methods on high-dimensional image generation. Specifically, we focus on fine-tuning a converged Continuous Normalizing Flow model on the CIFAR-10 dataset using an large batch setting.

We employ a widely adopted U-Net architecture containing approximately 35M parameters. Following the standard protocol [71], we utilize a converged Independent CFM (I-CFM) checkpoint natively pretrained from scratch for 400,000 steps with a learning rate of 2×10^{-4} and a batch size of 128. Rather than training all configurations from scratch (which incurs extreme computational cost), we establish a rigorous fine-tuning protocol. All comparative methods natively load the equivalent I-CFM converged checkpoint (400000 steps training) and fine-tune for strictly 10 epochs. This setup establishes an absolutely fair ground where every method originates from an identical prior capacity and uniquely differs strictly by the assigned mini-batch coupling strategies (I-CFM, OT-CFM, OA-OT, and RA-OT). To avoid catastrophic forgetting, we reduce the learning rate to 5×10^{-5}, apply a short 100-step warmup, tighten gradient clipping to 0.5, and lower the EMA decay from 0.9999 to 0.999 for faster adaptation. For amortized methods, predictors are constructed from $M = 50$ mini-batches with $L = 100$ sliced projections. All runtime results are reported under a large batch size ($\mathcal{B} = 2048$). About evaluation, we sample 50,000 outputs per model and evaluate the FID

and NFE via the adaptive `dopri5` numerical solver (with tolerance 10^{-5}). As shown in Table 10, directly applying the same optimal transport setup (OT-CFM) to large-batch training only increases the computational burden. In contrast, our precomputation strategies (RA-OT and OA-OT) effectively bypass this bottleneck by generating coupling plans much more efficiently, significantly reducing fine-tuning time. More interestingly, amortizing optimal transport does more than just speed things up, it also improves performance. OA-OT produces straighter trajectories, achieving the best NFE of 146.00, while RA-OT delivers the strongest perceptual quality, reaching an FID of 3.543.

Table 10: Adaptive sampling generation results (`dopri5`) fine-tuning Flow Matching models on CIFAR-10. Despite operating on an extreme scaling batch ($\mathcal{B} = 2048$), our amortized paradigms (OA-OT, RA-OT) drastically accelerate iterative epochs relative to exact OT-CFM while yielding highly superior FID scores and tightly constrained NFE parameters.

Method	**FID** $\downarrow$	**NFE/sample** $\downarrow$	**Training Time (s)** $\downarrow$	**Pretrain Time (s)** $\downarrow$
ICFM	3.638	146.61	**481.9**	0.0
OT-CFM	3.630	146.86	744.2	0.0
OA-OT	3.575	**146.00**	607.0	**12.0**
RA-OT	**3.543**	147.10	618.1	16.4

D Additional Experiments

In this section, we provide additional experiments to further examine several properties of the proposed amortized OT framework. In particular, we study its ability to handle measures with varying numbers of atoms, the expressivity of using a global coefficient vector across heterogeneous measure pairs, its use for warm-starting iterative Sinkhorn solvers, robustness under distribution shift, the marginal feasibility of the predicted transportation plans, and the trade-off between training and inference time across different numbers of test problems.

Transportation between measures with different numbers of atoms. A key property of our amortized model is that its parameters are independent of the number of atoms in the input measures. To empirically evaluate this property, we construct a multi-resolution MNIST benchmark in which each image is randomly resized to one of three resolutions, namely 28×28, 20×20, and 14×14, corresponding to measures with 784, 400, and 196 atoms, respectively. For both the training set ($M = 50$) and the test set ($N = 300$), the resolutions of the source and target images are sampled independently. As a result, the benchmark contains both same-resolution and cross-resolution transportation problems, including, for example, transportation between a 784-atom measure and a 196-atom measure. We train a single RA-OT or OA-OT model on the resulting mixed-resolution training set and evaluate the same model on all test pairs without any architectural modification. This setting is particularly suitable for our construction since the coefficient vector $\omega \in \mathbb{R}^L$ does not depend on the dimensions of the input measures. In contrast, the fixed-dimensional MLP architecture used by Meta-OT requires a fixed number of input atoms and therefore cannot be directly applied to this heterogeneous setting.

Table 11 reports the prediction errors separately for same-resolution and cross-resolution pairs. Both RA-OT and OA-OT remain stable across the two cases, indicating that the proposed representation can naturally accommodate transportation problems with different support sizes. Table 12 further reports the overall performance. Both methods use only $L = 100$ learned coefficients while operating on all combinations of resolutions within the same amortized model.

Expressivity of a global coefficient vector on heterogeneous measures. The linear amortized model considered in this work uses a single coefficient vector ω across measure pairs drawn from the underlying meta-distribution. We further investigate whether sharing this coefficient vector introduces an expressivity bottleneck when the measure pairs are heterogeneous. To this end, we partition MNIST digits into two disjoint groups: Group A containing digits 0-4 and Group B containing digits 5-9. For each group, we use $M = 20$ training pairs and $N = 25$ test pairs.

We first estimate two group-specific RA-OT models, yielding coefficient vectors ω_A and ω_B, using only training pairs from their corresponding groups. We then estimate a global coefficient vector

Table 11: Plan RMSE on the multi-resolution MNIST benchmark, separated into same-resolution and cross-resolution source-target pairs.

Method	Same-resolution RMSE	Cross-resolution RMSE
RA-OT	$7.37 \times 10^{-5} \pm 8.33 \times 10^{-5}$	$4.20 \times 10^{-5} \pm 3.02 \times 10^{-5}$
OA-OT	$4.37 \times 10^{-5} \pm 4.69 \times 10^{-5}$	$2.50 \times 10^{-5} \pm 1.55 \times 10^{-5}$

Table 12: Overall performance on the multi-resolution MNIST benchmark. A single amortized model is used for measure pairs with different numbers of atoms.

Method	Plan RMSE	Train (s)	Infer (ms)	# Params
RA-OT	$5.18 \times 10^{-5} \pm 5.50 \times 10^{-5}$	2.3	15.50 ± 11.79	100
OA-OT	$3.08 \times 10^{-5} \pm 3.00 \times 10^{-5}$	17.1	15.31 ± 11.01	100

ω_{global} using the combined training set of $M = 40$ pairs from both groups. Table 13 compares the resulting plan prediction errors. Despite the substantial heterogeneity between the two subsets, the global coefficient performs comparably to, and in some cases slightly better than, the corresponding group-specific coefficients. To further examine the learned representations, Table 14 reports the cosine similarities between the coefficient vectors. Interestingly, ω_A and ω_B are nearly orthogonal, with a cosine similarity of -0.0294, suggesting that the two subsets favor substantially different combinations of sliced potentials. Nevertheless, ω_{global} maintains good prediction accuracy on both groups. These results indicate that, in this experiment, using a shared global coefficient does not introduce a substantial expressivity bottleneck and provides a parsimonious model across heterogeneous measures. We emphasize that the amortized formulation in Section 3.2 is not restricted to linear operators, and more expressive choices of γ_ϕ can be considered when needed.

Table 13: Plan RMSE obtained using a global RA-OT coefficient vector and group-specific coefficient vectors on heterogeneous MNIST subsets. The global model is trained on the combined training pool from both groups.

Test set	ω_{global} (Mixed)	ω_A (Digits 0-4)	ω_B (Digits 5-9)
Test A (Digits 0-4)	$8.11 \times 10^{-6} \pm 3.40 \times 10^{-6}$	$8.95 \times 10^{-6} \pm 3.09 \times 10^{-6}$	$9.65 \times 10^{-6} \pm 6.05 \times 10^{-6}$
Test B (Digits 5-9)	$7.14 \times 10^{-6} \pm 2.73 \times 10^{-6}$	$9.13 \times 10^{-6} \pm 3.91 \times 10^{-6}$	$7.40 \times 10^{-6} \pm 3.13 \times 10^{-6}$
Test Combined	$7.62 \times 10^{-6} \pm 3.12 \times 10^{-6}$	-	-

Warm-starting Sinkhorn with predicted Kantorovich potentials. Although our main experiments directly recover transportation plans from the predicted potentials, the amortized potentials can also be used to initialize iterative OT solvers. We investigate this use case by using the potentials predicted by RA-OT, OA-OT, and Meta-OT to warm-start the standard Sinkhorn iterations. We compare these initializations against the conventional cold start using zero potentials.

Using $N = 100$ test pairs, we track the L_1 marginal error at fixed iteration checkpoints up to $1{,}000$ Sinkhorn iterations. The results are reported in Table 15. The predicted potentials provide substantially better initializations than a cold start. In particular, after one Sinkhorn iteration, OA-OT obtains a marginal error of 1.41×10^{-1}, compared with 2.55×10^{-1} for Meta-OT and 5.35×10^{-1} for the cold start. This advantage remains visible during the early iterations: at iteration 10, OA-OT reaches a marginal error of 1.63×10^{-2}, which is below the errors obtained by the other initializations at the same iteration. As expected, all initialization schemes eventually converge to machine precision when sufficiently many Sinkhorn iterations are performed. These results show that the proposed amortized potentials can also serve as useful warm-starts when an exact or more tightly converged Sinkhorn solution is desired.

Robustness under rotational distribution shift. The amortized OT formulation assumes that future measure pairs are sampled from a meta-distribution related to that observed during training. We do not assume that the learned amortized model is invariant to arbitrary distribution shifts. Nevertheless, we conduct an additional experiment to examine its behavior under a simple geometric shift. Following the Rotated MNIST setting, we train all amortized models using only unrotated

Table 14: Cosine similarity between the RA-OT coefficient vectors learned from the two MNIST subsets and the combined training pool.

Vector pair	Cosine similarity
(ω_A, ω_B)	-0.0294
$(\omega_A, \omega_{\text{global}})$	0.7520
$(\omega_B, \omega_{\text{global}})$	0.5762

Table 15: L_1 marginal error during Sinkhorn iterations under different initialization strategies. The predicted Kantorovich potentials from Meta-OT, RA-OT, and OA-OT are used as warm-starts, while the cold-start baseline initializes the potentials at zero. **Bold** indicates the best performance.

Init. method	Iter. 0	Iter. 1	Iter. 2	Iter. 5	Iter. 10	Iter. 20	Iter. 50	Iter. 100	Iter. 200	Iter. 1000
Cold Start	1.61×10^{4}	5.35×10^{-1}	2.08×10^{-1}	6.93×10^{-2}	2.81×10^{-2}	9.78×10^{-3}	8.72×10^{-4}	2.83×10^{-5}	$\mathbf{1.22 \times 10^{-7}}$	4.54×10^{-16}
Meta-OT	4.18×10^{-1}	2.55×10^{-1}	1.52×10^{-1}	6.23×10^{-2}	2.69×10^{-2}	9.42×10^{-3}	8.18×10^{-4}	2.76×10^{-5}	1.63×10^{-7}	6.98×10^{-16}
RA-OT	3.24×10^{-1}	1.79×10^{-1}	1.09×10^{-1}	5.23×10^{-2}	2.62×10^{-2}	9.70×10^{-3}	8.23×10^{-4}	2.97×10^{-5}	2.88×10^{-7}	4.44×10^{-16}
OA-OT	$\mathbf{2.84 \times 10^{-1}}$	$\mathbf{1.41 \times 10^{-1}}$	$\mathbf{7.95 \times 10^{-2}}$	$\mathbf{3.46 \times 10^{-2}}$	$\mathbf{1.63 \times 10^{-2}}$	$\mathbf{5.96 \times 10^{-3}}$	$\mathbf{5.32 \times 10^{-4}}$	$\mathbf{2.09 \times 10^{-5}}$	1.95×10^{-7}	$\mathbf{4.29 \times 10^{-16}}$

images ($0°$) and evaluate them without further adaptation on images rotated by $30°$, $60°$, and $90°$. Table 16 reports the resulting plan RMSE. Both RA-OT and OA-OT remain relatively stable across the evaluated rotations. For example, the RMSE of OA-OT varies from 6.02×10^{-6} at $0°$ to 6.45×10^{-6} at $90°$. RA-OT exhibits a similarly moderate change. Both methods also retain lower reconstruction errors than Meta-OT across all evaluated rotation angles. These observations suggest that the sliced potentials preserve useful geometric information under this particular rotational shift. However, this experiment only considers one structured form of distribution shift, and a systematic study of OOD generalization for amortized OT remains an interesting direction for future work.

Table 16: Plan RMSE under rotational distribution shift on MNIST. All amortized models are trained using only images at $0°$ and evaluated zero-shot at the indicated rotation angles.

Method	$0°$	$30°$	$60°$	$90°$
RA-OT	7.77×10^{-6}	7.21×10^{-6}	7.31×10^{-6}	8.34×10^{-6}
OA-OT	$\mathbf{6.02 \times 10^{-6}}$	$\mathbf{5.19 \times 10^{-6}}$	$\mathbf{5.31 \times 10^{-6}}$	$\mathbf{6.45 \times 10^{-6}}$
Meta-OT	1.84×10^{-5}	1.63×10^{-5}	1.63×10^{-5}	1.84×10^{-5}

Marginal violations and rounding to valid transportation plans. The transportation plan reconstructed from an amortized prediction is obtained from a predicted Kantorovich potential together with the corresponding update of the other potential. Consequently, the resulting matrix is not necessarily guaranteed to satisfy both marginal constraints exactly before further normalization. We explicitly quantify this effect by measuring the L_1 marginal violations with respect to the source and target measures, denoted by $\mathrm{MargErr}_a$ and $\mathrm{MargErr}_b$, respectively.

We additionally apply the standard rounding procedure of [2] to project each predicted matrix onto the transportation polytope with the prescribed marginals. As shown in Table 17, before rounding, each method satisfies one marginal up to numerical precision, while the other may exhibit a noticeable

Table 17: Plan RMSE and L_1 marginal violations before and after applying the rounding procedure of [2]. Results are reported for the MNIST experiment with $M = 50$, $N = 300$, and $\epsilon = 10^{-2}$.

Method	Stage	Plan RMSE	$\mathrm{MargErr}_a$	$\mathrm{MargErr}_b$
RA-OT	Before	$7.77 \times 10^{-6} \pm 3.1 \times 10^{-6}$	1.818×10^{-1}	4.225×10^{-9}
	After	$6.52 \times 10^{-6} \pm 2.5 \times 10^{-6}$	1.358×10^{-16}	2.941×10^{-16}
OA-OT	Before	$6.02 \times 10^{-6} \pm 2.5 \times 10^{-6}$	1.441×10^{-1}	4.171×10^{-9}
	After	$\mathbf{4.91 \times 10^{-6} \pm 2.0 \times 10^{-6}}$	$\mathbf{1.340 \times 10^{-16}}$	2.998×10^{-16}
Meta-OT	Before	$1.55 \times 10^{-5} \pm 4.7 \times 10^{-6}$	8.503×10^{-16}	4.300×10^{-1}
	After	$1.11 \times 10^{-5} \pm 3.0 \times 10^{-6}$	1.350×10^{-16}	$\mathbf{2.514 \times 10^{-16}}$

violation due to the single-potential update. After rounding, both marginal errors are reduced to approximately 10^{-16} for all methods, yielding valid transportation plans up to machine precision. Importantly, enforcing the marginal constraints does not compromise prediction quality: the plan RMSE decreases after rounding, and RA-OT and OA-OT continue to achieve lower reconstruction errors than Meta-OT.

Training-inference time trade-off. Although RA-OT and OA-OT require substantially less training time than Meta-OT, their per-pair inference time is higher due primarily to the computation of the sliced potentials. This additional inference cost comes with several properties of the sliced-potential representation: the resulting amortized model is independent of the number of atoms in the input measures, can naturally accommodate measures with varying support sizes, and uses only a small number of learned parameters. We further investigate the resulting trade-off between the one-time training cost and the per-pair inference cost through a break-even analysis.

In particular, we consider the MNIST setting with $M = 50$ and $\epsilon = 0.01$ and measure all runtimes sequentially on the GPU, without parallelizing inference across test pairs. For an amortized method, we define the total computational cost for processing N test pairs as

$$\mathrm{Cost}(N) = T_{\mathrm{train}} + N t_{\mathrm{infer}}, \tag{19}$$

where T_{train} denotes the one-time training cost and t_{infer} is the average inference time per test pair. We compare this cost with solving every OT problem independently from scratch using Sinkhorn, whose average runtime is $t_{\mathrm{scratch}} = 158.10$ ms per pair. We denote by $N^\star$ the number of test pairs at which the total costs of two methods become equal.

Table 18 shows that RA-OT, OA-OT, and Meta-OT break even with solving OT from scratch after 34, 137, and 238 test pairs, respectively. Comparing amortized methods directly, Table 19 shows that RA-OT and OA-OT remain cheaper than Meta-OT up to approximately 1,013 and 599 test pairs, after which Meta-OT benefits from its lower per-pair inference cost.

Table 18: Training time, per-pair inference time, and break-even point relative to independently solving each OT problem using Sinkhorn. Results are measured on MNIST with $M = 50$ and $\epsilon = 0.01$. The timing is sequential and does not exploit parallelization across test pairs.

Method	T_{train} (s)	t_{infer} (ms/pair)	$N^\star$ vs. scratch
RA-OT	4.24	34.48	34
OA-OT	16.68	36.23	137
Meta-OT	37.23	1.92	238
Sinkhorn from scratch	-	158.10	-

Table 19: Break-even points between the proposed methods and Meta-OT under sequential inference. A method in the left column has a lower total computational cost below the corresponding $N^\star$.

Comparison	$N^\star$	Interpretation
RA-OT vs. Meta-OT	1013.5	RA-OT lower cost for $N < 1013$
OA-OT vs. Meta-OT	599.2	OA-OT lower cost for $N < 599$

Table 20: Total computational cost, in seconds, at representative numbers of test pairs N. The cost of each amortized method includes its one-time training cost.

N	RA-OT	OA-OT	Meta-OT	Sinkhorn from scratch
1	4.27	16.71	37.24	0.16
100	7.68	20.30	37.42	15.81
1,000	38.71	52.90	39.15	158.10

These results highlight a clear training–inference trade-off. Meta-OT offers faster sequential inference at large scale, whereas RA-OT and OA-OT require less training time and achieve lower plan

reconstruction errors at a higher per-pair inference cost. The preferred regime therefore depends on the balance between training cost, throughput, and accuracy. **Since test pairs can be processed independently, parallel inference may further shift the practical crossover points**; we leave a detailed throughput study to future work.

E Additional Visualizations

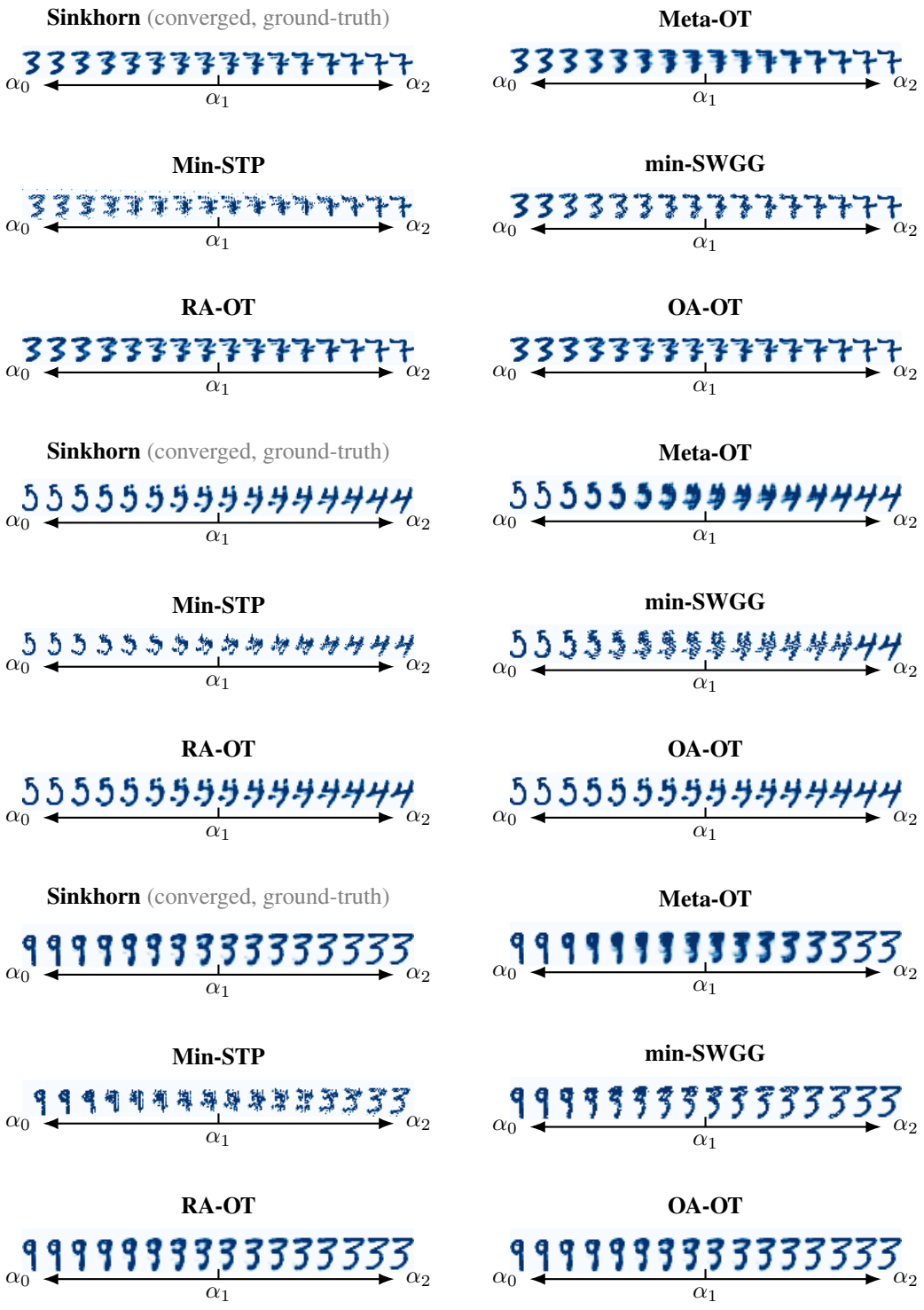

Figure 8: Additional qualitative visualizations of the predicted optimal transport plans on the MNIST dataset ($M = 50, L = 100$). The figure compares the structural alignment and digit matching quality across different amortized and sliced OT solvers (RA-OT, OA-OT, Meta-OT, Min-STP, and min-SWGG). Our proposed methods consistently preserve the topological structure of the digits.

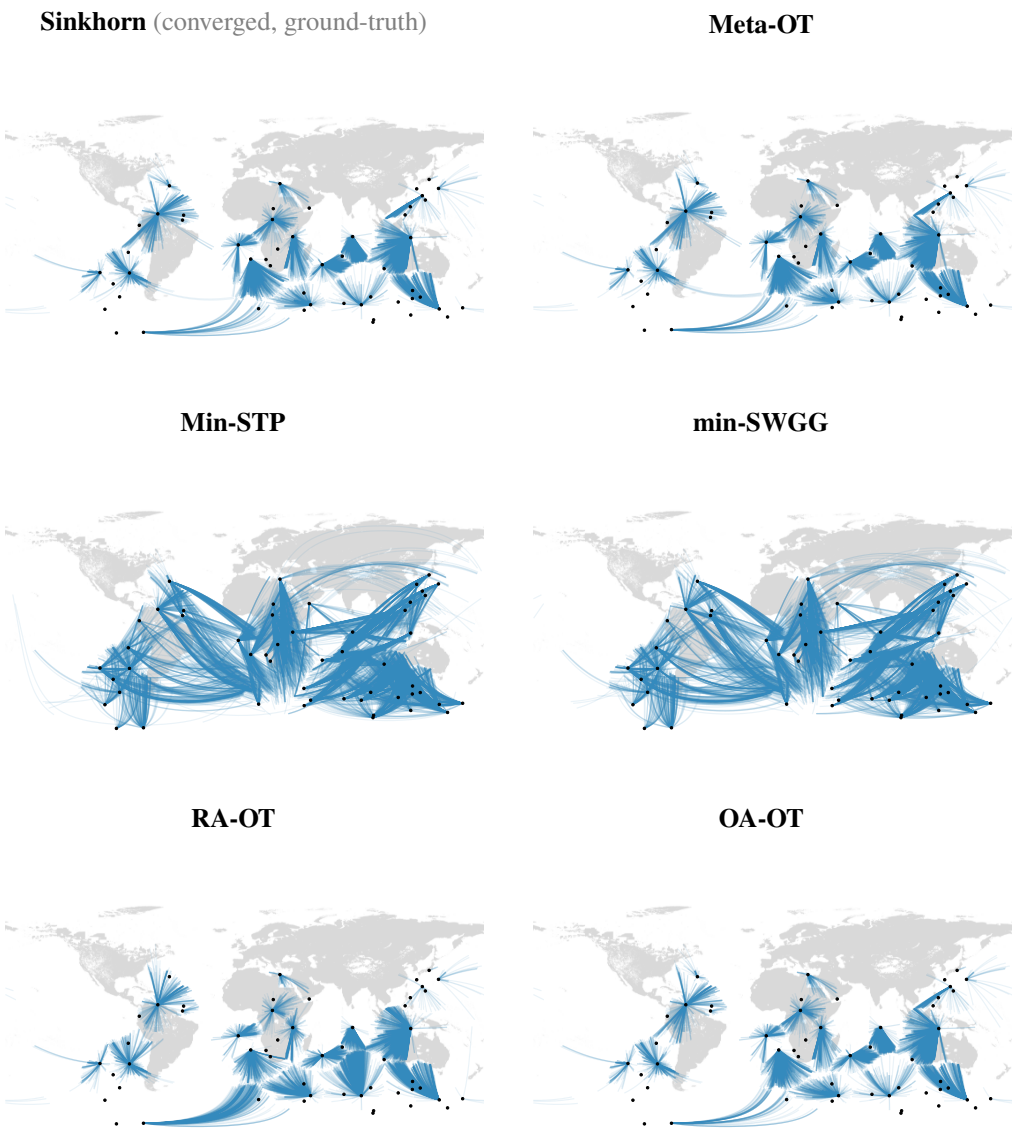

Figure 9: Extended qualitative results on the spherical world supply-demand transport task ($M = 50, L = 100$). The visualizations illustrate the transport assignments from dense global population distributions (demand) to uniformly scattered supply points across the globe. Observations indicate that RA-OT and OA-OT generate smoother and more globally coherent matching patterns, effectively avoiding the localized structural artifacts often observed in standard linear sliced estimation methods.

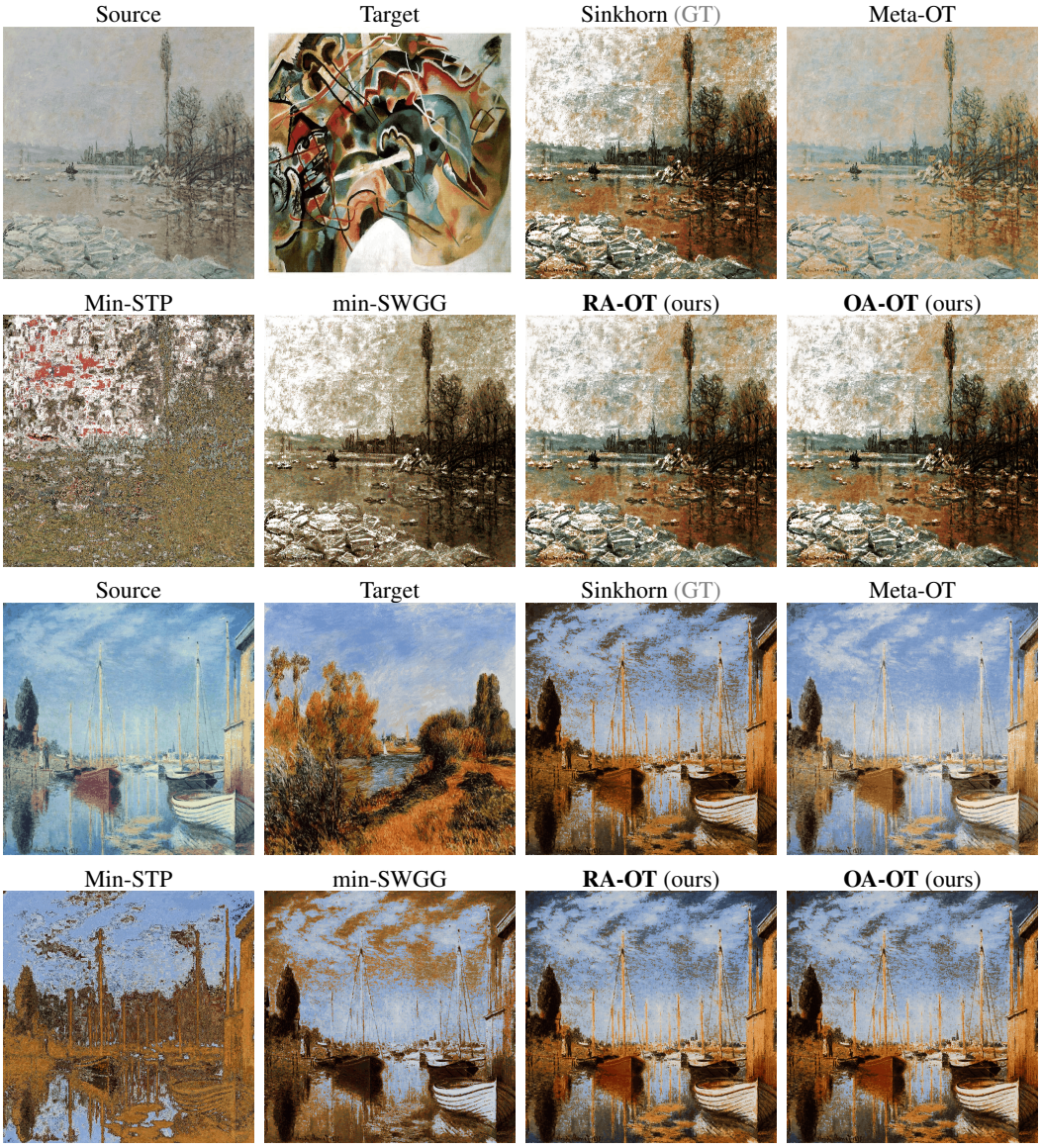

Figure 10: Further qualitative examples of the color transfer task between natural images using different OT solvers ($M = 50, L = 100$). The results demonstrate the capability of our amortized solvers to robustly align high-dimensional RGB color histograms. The images transported by RA-OT and OA-OT strictly adhere to the target color palettes while faithfully preserving the underlying textural and geometric details of the source images, exhibiting highly competitive visual fidelity against established exact and amortized baselines.

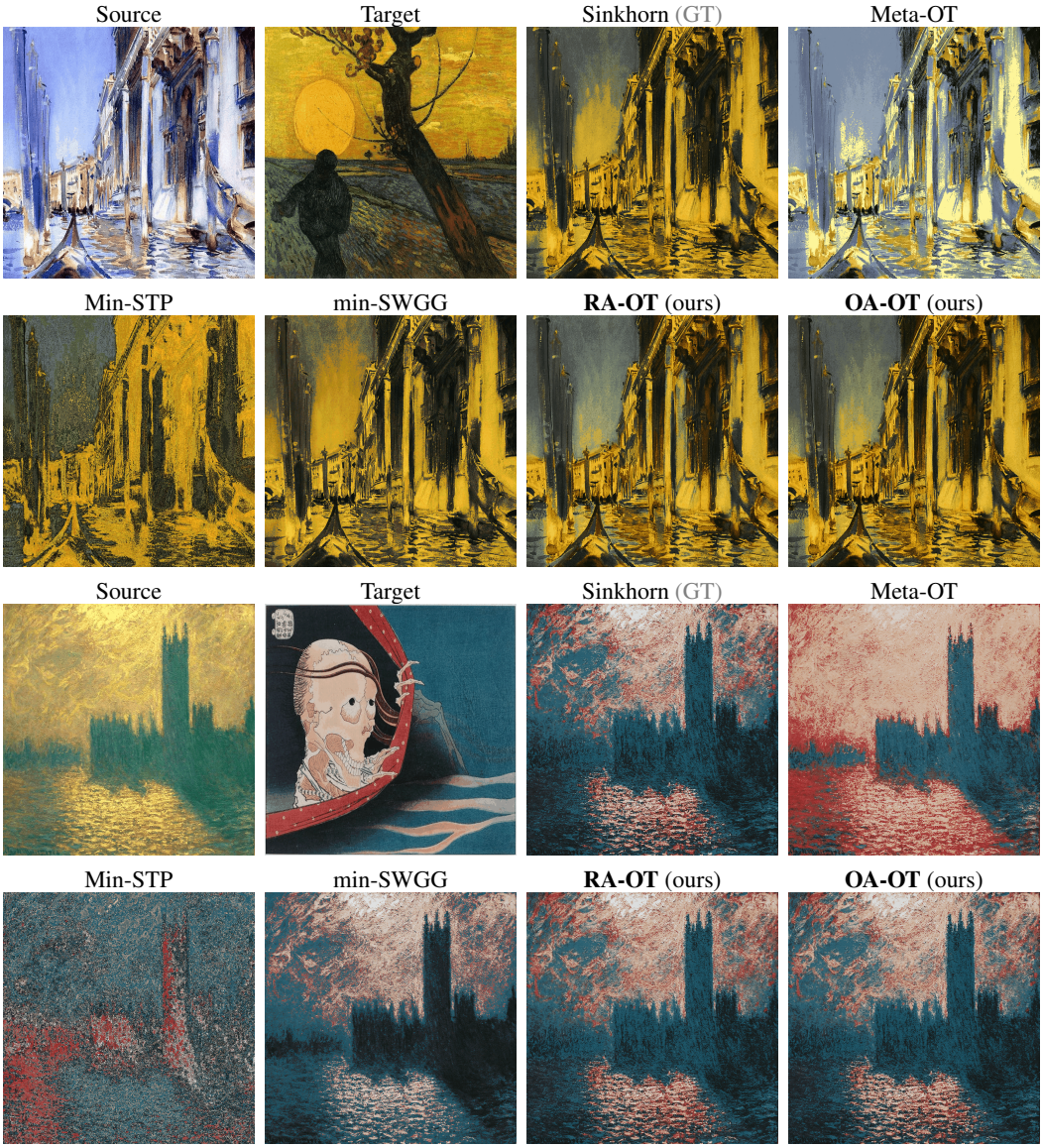

Figure 11: Additional qualitative examples of the color transfer task ($M = 50, L = 100$). The figure presents further comparisons of the visual outputs generated by the evaluated methods when mapping the color palette of the target images onto the source images.

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: We believe that the abstract and introduction accurately reflect the paper's contributions and scope.

   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: While the proposed methods reduce the training time, the inference time is slightly increased.

   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: [N/A]

Justification: The paper does not include theoretical results.

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: We provide all details of experimental settings and provide the code for all experiments.

   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: All datasets in the paper are published and we provide the code for all experiments.

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: We specify all the training and test 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: We provide the error bars in experiments.

   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: We provide the detail of the computer resources that we use in Appendix A.

   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: We follow the NeurIPS Code of Ethics.

   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: There is no societal impact of the work performed.

    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: The paper poses no such risks.

    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: [N/A]

    Justification: The paper does not use existing assets.

    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: [Yes]

    Justification: We provide a new code for implementation of our proposed methods.

    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: 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.
    - 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: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

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.