RETHINKING THE FLOW-BASED GRADUAL DOMAIN ADAPTION: A SEMI-DUAL TRANSPORT PERSPECTIVE

Anonymous authors
Paper under double-blind review

ABSTRACT

Gradual domain adaptation (GDA) aims to mitigate domain shift by progressively adapting models from the source domain to the target domain via intermediate domains. However, real intermediate domains are often unavailable or ineffective, necessitating the synthesis of intermediate samples. Flow-based models have recently been used for this purpose by interpolating between source and target distributions; however, their training typically relies on sample-based log-likelihood estimation, which can discard useful information and thus degrade GDA performance. The key to addressing this limitation is constructing the intermediate domains via samples directly. To this end, we propose an Entropy-regularized Semi-dual Unbalanced Optimal Transport (E-SUOT) framework to construct intermediate domains. Specifically, we reformulate flow-based GDA as a Lagrangian dual problem and derive an equivalent semi-dual objective that circumvents the need for likelihood estimation. However, the dual problem leads to an unstable min–max training procedure. To alleviate this issue, we further introduce entropy regularization to convert it into a more stable alternative optimization procedure. Based on this, we propose a novel GDA training framework and provide theoretical analysis in terms of stability and generalization. Finally, extensive experiments are conducted to demonstrate the efficacy of the E-SUOT framework.

1 INTRODUCTION

Unsupervised Domain Adaptation (UDA) (Pan & Yang, 2010; Tzeng et al., 2017; Long et al., 2015; Courty et al., 2014; 2017a), which transfers knowledge from a well-trained source domain to a related yet unlabeled target domain, is of great importance across fundamental application areas. For example, in recommender systems (Liu et al., 2023; Zheng et al., 2024), a cold-start user has no interaction history with new items, so domain adaptation helps transfer user and item knowledge from an existing system to improve recommendations. Similar scenarios occur in machine translation, where a model trained on high-resource language pairs like English-French can be adapted to translate between English and low-resource languages with limited parallel data (Gazdieva et al., 2023). These scenarios highlight the importance of conducting UDA to bridge domain gaps and ensure reliable performance in real-world applications.

Despite these methodological advances, directly performing UDA can be brittle when the source–target shift is substantial or class overlap is weak. In such cases, one-shot alignment often degrades discriminability and amplifies pseudo-label errors during self-training. This challenge motivates a transition from the traditional UDA setting to the Gradual Domain Adaptation (GDA) setting (He et al., 2024), where adaptation proceeds through a sequence of intermediate distributions that progressively bridge the domain gap. A key aspect of generating intermediate domains in GDA is to interpolate between the source and target domains. Various methods have been proposed to construct such intermediate domains, among which flow-based approaches (Kobyzev et al., 2020; Papamakarios et al., 2021) have attracted increasing attention, primarily due to their property of preserving probability density along the transformation path, thereby enabling consistent and stable probability densities without distortion or loss of information. To drive the samples from the source domain towards those of the target domain, it is necessary to design an appropriate driving force, typically derived from a discrepancy metric. Among these metrics, f-divergence (Sason & Verdú, 2016) is most widely used due to its computational efficiency, empirical effectiveness, and principled formulation within the framework of geometry for probability distributions (Amari, 2016).

Despite the success of flow-based approaches in GDA (Sagawa & Hino, 2025; Zhuang et al., 2024; Zeng et al., 2025), we argue that directly applying standard flow-based models leads to suboptimal performance. Specifically, existing flow-based frameworks utilizing f-divergence often require the explicit estimation of target domain probability density functions (PDFs) from available target samples (Vincent, 2011; Santambrogio, 2017; Ambrosio et al., 2005). Consequently, the quality of the intermediate domain heavily depends on the accuracy of the estimated target PDF; if this estimation is inaccurate, the performance of the downstream task is likely to suffer significantly.

To address these limitations, we propose a novel flow-based GDA framework E-SUOT, which leverages the semi-dual formulation of gradient flows. Rather than explicitly estimating PDFs, we recast flow evolution as an optimization problem that combines an f-divergence term with a Wasserstein distance regularization term, enabling sample transport toward the target domain without reliance on PDF estimation. However, as the semi-dual reformulation inherently leads to an adversarial training paradigm that can compromise stability and performance, we introduce entropy regularization to the objective to guarantee the stability of the training process. Based on this, we summarize the algorithm for E-SUOT-based intermediate domain generation, prove the convergence of our E-SUOT framework, and empirically demonstrate its effectiveness on representative GDA tasks. Extensive experiments validate that E-SUOT achieves superior performance compared with existing methods.

Contributions. The main contributions of this paper are summarized as follows:

- We develop a semi-dual-formulation for intermediate domain generation in flow-based GDA, which eliminates the need for explicit PDF estimation in the target domain.
- We introduce an entropy regularization term to address the unstable issue inherent in the semi-dual formulation, resulting in the novel and stable E-SUOT framework.
- We conducted various experiments to demonstrate the superiority of the proposed E-SUOT approach compared to prevalent approaches.

2 Preliminaries

2.1 Settings and Notations

In GDA, we consider a labeled source domain, $T-1$ unlabeled intermediate domains, and an unlabeled target domain. Let the input space be $\mathcal{X}$ and the label space be $\mathcal{Y}$. We denote inputs as $x \in \mathcal{X}$ and labels as $y \in \mathcal{Y}$. We index the domains by $t \in \{0, 1, \ldots, T\}$, where $t = 0$ denotes the source domain and $t = T$ denotes the target domain. Each domain induces a marginal distribution p_t over $\mathcal{X}$. Let $\mathcal{H}$ be a hypothesis class of classifiers $h : \mathcal{X} \to \mathcal{Y}$. We assume that each domain admits a labeling function $q_t \in \mathcal{H}$. Given a loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$, the generalization error of h on domain t is defined as $\varepsilon_{p_t}(h) = \mathbb{E}_{p_t(x)}\big[\mathcal{L}\big(h(x), q_t(x)\big)\big]$. A source classifier $q_0 \in \mathcal{H}$ can be learned via supervised learning on the source domain with minimal error $\varepsilon_{p_0}(q_0)$. The objective of GDA is to evolve q_0 through the intermediate domains to a classifier h_T so as to minimize the target error $\varepsilon_{p_T}(h_T)$.

2.2 Flows for Intermediate Domain Generation

A flow describes the time-dependent evolution of particles induced by a smooth invertible (diffeomorphic) map. Based on this, the intermediate domains can be seen as a discretization of a continuous flow linking source and target distributions. This motivates flow-based models, which evolve a distribution over a fixed time horizon while preserving normalization, and are thus well-suited for GDA. From the flow perspective, intermediate domains are generated by the following ordinary differential equation:

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_t(x_t) = -\nabla \frac{\delta \mathbb{D}[p(x_t), p_T(x)]}{\delta p(x_t)}, \quad x_{t=0} = x_0, \tag{1}$$

where $p(x_t)$ is the (empirical) PDF induced by $\{x_{t,i}\}_{i=1}^{\mathrm{N}}$, and we desire the law $p(x_T)$ to approximate the target $p_T(x)$. Here $v_t : \mathcal{X} \to \mathcal{X}$ is the velocity field. The core design problem is to choose v_t so that $p(x_t) \xrightarrow[t \to T]{} p_T(x)$. A principled approach is to define v_t as the steepest descent direction of some discrepancy functional $\mathbb{D}[p(x_t), p_T(x)]$ between $p(x_t)$ and $p_T(x)$ as demonstrated in the

second equal sign in Eq. (1). Notably, $\delta/\delta p$ denotes the first variation, and the second equality sign is called "gradient flow".

Among various choices, f-divergences are favored in GDA for their task-aligned objectives, stable probability-preserving dynamics, and efficient computation when compared to alternatives such as Sinkhorn divergence and maximum mean discrepancy (Glaser et al., 2021). For an f-divergence,

$$\mathbb{D}_f[p(x_t), p_T(x)] = \int f\left(\frac{p(x_t)}{p_T(x)}\right) p_T(x)\,\mathrm{d}x, \tag{2}$$

with $f : (0, \infty) \to \mathbb{R}$ convex and $f(x) = 0$ if and only if $x = 1$. A canonical example is the Kullback–Leibler (KL) divergence with $f(u) = u \log u$. In this case,

$$v_t(x_t) = \nabla \log p_T(x) - \nabla \log p(x_t), \tag{3}$$

and, in the weak partial differential equation sense (Evans, 2022; Liu, 2017), the induced dynamics yield the classical *Langevin dynamic* (Welling & Teh, 2011; Santambrogio, 2017).

Intuitively, applying the forward Euler scheme with step size η to the gradient flow in Eq. (1) under an f-divergence yields a discrete-time generation for the intermediate domain, which is equivalent to solving a 2-Wasserstein-distance–regularized optimization problem as (see Section B.1):

$$x_{t+\eta} = x_t - \eta \nabla \frac{\delta \mathbb{D}_f[p(x_t), p_T]}{\delta p(x_t)} \Rightarrow p(x_{t+\eta}) = \underset{\rho(x)\in\mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \frac{1}{2\eta} \mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)], \tag{4}$$

where $\mathcal{P}_2(\mathbb{R}^D)$ denotes the Wasserstein space (Villani et al., 2009), which is the set of the distributions with finite second moment. Here $\mathcal{W}_2$ is the 2-Wasserstein distance, whose definition is given as follows:

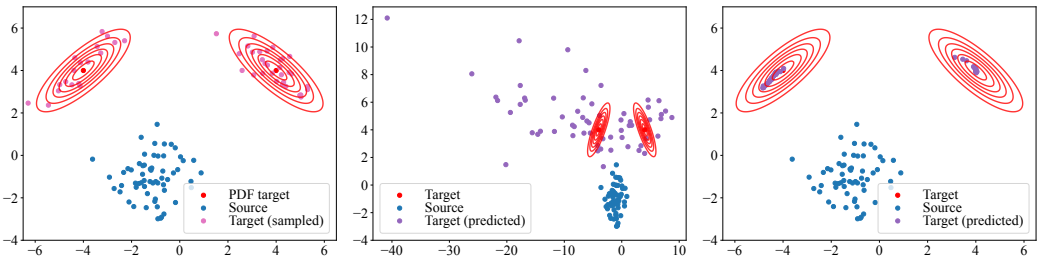
$$\mathcal{W}_2^2(\rho, \xi) = \inf_{\pi\in\Pi(\rho,\xi)} \iint \|x - y\|_2^2 \,\pi(x, y)\,\mathrm{d}x\,\mathrm{d}y, \tag{5}$$

and $\Pi(\rho, \xi)$ is the set of joint distribution on $\mathbb{R}^D \times \mathbb{R}^D$ with marginal distributions ρ and ξ.

3 Methodology

3.1 Motivation Analysis

Flow-based approaches, exemplified by gradient-flow methods, interpolate between the source and target distributions by gradually minimizing a discrepancy measure, typically an f-divergence, between the two domains. The success of these methods in GDA tasks critically depends on accurately estimating the target distribution's probability density function (PDF). Given a reliable estimate, one can construct a velocity field that progressively pushes source samples toward the target distribution.



(a) Ground Truth. (b) Langevin, $\mathcal{W}_2^2 \approx 9.7$. (c) E-SUOT, $\mathcal{W}_2^2 \approx 5.4$.

Figure 1: Illustrative Example: Comparison Between Lagevin Dynamic and E-SOUT.

However, directly estimating the PDF from target domain data is generally ill-posed (Vincent et al., 2010; Song et al., 2020). When the estimate is inaccurate, the induced velocity field can push samples into low-probability regions of the target distribution, causing a substantial shift between the generated and true target domains and degrading downstream task performance. To illustrate this issue, we compare ground-truth target samples with those obtained via Langevin dynamics and E-SUOT in Figs. 1(a) to 1(c). The PDF for the target domain is estimated using denoised score matching (Vincent, 2011). In addition, we also report the 2-Wasserstein distance between

the predicted and ground-truth samples (relative to Fig. 1(a)) in the captions of Figs. 1(b) and 1(c), which constituted the lower generalization bound for GDA tasks. From Figs. 1(b) and 1(c), it is evident that when the estimated log-likelihood function is inaccurate, the samples generated for the target distribution deviate substantially from the ground truth and yield a large Wasserstein distance, which may ultimately limits performance on GDA tasks. In summary, the key questions addressed in this paper can be summarized as follows: How can we generate intermediate domains without compromising the accuracy of the target domain? How can robust intermediate domain generation be achieved within this framework? Does this approach improve the performance for GDA task?

3.2 Dual-Form Transportation for Intermediate Domain Generation

As shown in Eq. (4), simulating the gradient flow to generate intermediate domains is precisely equivalent to solving a Wasserstein-distance-regularized optimization problem. This insight opens up a practical alternative: *instead of explicitly estimating the target domain's probability density, one can guide source samples by directly tackling this optimization formulation*. Thus, we have the following proposition regarding the solution property of the problem defined in Eq. (4):

Proposition 1. *Consider the following primal problem:*

$$\mathcal{L}^{\textit{Primal}} = \underset{\rho(x)\in\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})}{\arg\min} \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)]. \tag{6}$$

This problem is equivalent to the following semi-dual formulation:

$$\mathcal{L}^{\textit{SemiDual}} = \sup_{w} \mathbb{E}_{p(x_t)}\left[\inf_{\boldsymbol{T}}\left(\frac{1}{2\eta}\|\boldsymbol{T}(x_t) - x_t\|_2^2 - w(\boldsymbol{T}(x_t))\right)\right] - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{7}$$

where $w : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ is a measurable continuous function, $\boldsymbol{T} : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}^{\mathrm{D}}$ is the transport map, and $f^\star$ denotes the convex conjugate of f, defined as $f^\star(z) := \sup_{y\geq 0}(zy - f(y))$.

Importantly, the structure of the semi-dual problem ensures that both $p_t(x)$ and $p_T(x)$ are involved only through expectation operators, rather than through explicit density evaluations. This enables the use of Monte Carlo methods to approximate all necessary integrals, thereby eliminating the need for access to the density function—particularly for the target domain—when constructing intermediate distributions. Practically, following prior works (Korotin et al., 2023; Choi et al., 2023; 2024), we can parameterize both the dual potential w and the transport map $\boldsymbol{T}$ by neural networks, denoted as w_ϕ and $\boldsymbol{T}_\theta$ respectively. The models are trained in an alternating adversarial scheme to learn the sequence of maps $\{\boldsymbol{T}_{\theta,t}\}_{t=0}^{T-1}$, which can be applied to generate intermediate domains progressively.

3.3 Robust Training Procedure for Semi-Dual Form Transportation

While Section 3.2 provides a semi-dual form of the gradient flow problem that avoids explicit PDF estimation in target domain, naively training $\mathcal{L}^{\text{SemiDual}}$ in Eq. (7) is intrinsically unstable because of its composite 'sup–inf' structure. This instability is not merely algorithmic: the objective itself may be non-identifiable. We formalize this phenomenon by proving that the dual problem can have non-unique optima, as the following theorem shows:

Proposition 2. *The semi-dual formulation in Eq. (7) admits non-unique optimal solutions.*

To address this issue, we incorporate an entropy regularization term into the primal objective Eq. (6), which leads to the following proposition:

Proposition 3. *Let $\kappa(x_t, x) := p(x_t)\, p_T(x)$ denote the reference joint PDF. The entropy-regularized primal problem is*

$$\begin{aligned}\mathcal{L}^{\textit{E-Primal}} = \underset{\rho\in\mathcal{P}_2(\mathbb{R}^D)}{\arg\min} \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f\big[\rho(x), p_T(x)\big] \\ + \epsilon \iint \pi(x_t, x)\,[\log\frac{\pi(x_t, x)}{\kappa(x_t, x)} - 1]\,\mathrm{d}x_t\,\mathrm{d}x,\end{aligned} \tag{8}$$

and is equivalent to the semi-dual optimization problem

$$\mathcal{L}^{\textit{E-SemiDual}} = \sup_{w} -\epsilon\,\mathbb{E}_{p(x_t)}[\log \mathbb{E}_{p_T(x)}(\exp(\frac{w(x) - \frac{1}{2\eta}\|x - x_t\|_2^2}{\epsilon}))] \; - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{9}$$

where $w : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ *and* $f^\star$ *are as defined in Proposition 1.*

On this basis, we provide a theoretical guarantee of uniqueness for the semi-dual objective in Eq. (9):

Proposition 4. *The semi-dual formulation in Eq. (9) admits a unique optimal solution.*

Notably, as seen in Eq. (9), the semi-dual objective depends solely on the potential w. Consequently, we can optimize a single model, which lowers the computational burden. We therefore parameterize w by a neural network w_ϕ and carry out the optimization.

Finally, conditioned on the resulting w_ϕ, we subsequently optimize the transport map $\boldsymbol{T}_\theta(x)$ via the following objective based on Eq. (7):

$$\arg\min_{\theta} \quad \frac{1}{2\eta}\|x_t - \boldsymbol{T}_\theta(x_t)\|_2^2 - w_\phi(\boldsymbol{T}_\theta(x_t)). \tag{10}$$

Notably, we denote our approach as "E-SUOT", as the derivation of $\boldsymbol{T}_\theta$ is grounded in the Entropy-regularized Semi-dual Unbalanced Optimal Transport framework.

3.4 Overall Workflow for E-SUOT

Although Sections 3.2 and 3.3 have presented the E-SUOT framework for intermediate domain generation, they do not provide a unified view of the overall workflow for generating intermediate domains. To address this, we summarize the complete procedure in Algorithm 1 (Due to page limit, the complete algorithm and other detailed information are summarized in Appendix D) and the corresponding illustration is given in Fig. 2. As shown in the algorithm, the construction of w_ϕ and $\boldsymbol{T}_\theta$ are performed as separate steps, corresponding to Fig. 2(a), and are illustrated in *Lines 3–6* and *Lines 7–10*, respectively. By iteratively executing the procedure described in *Lines 3–10*, we obtain a sequence of transport maps, $\mathcal{T} = \{\boldsymbol{T}_{\theta,t}\}_{t=0}^{T-1}$, which progressively transport samples from the source domain to the target domain, as we demonstrate in Fig. 2(b).

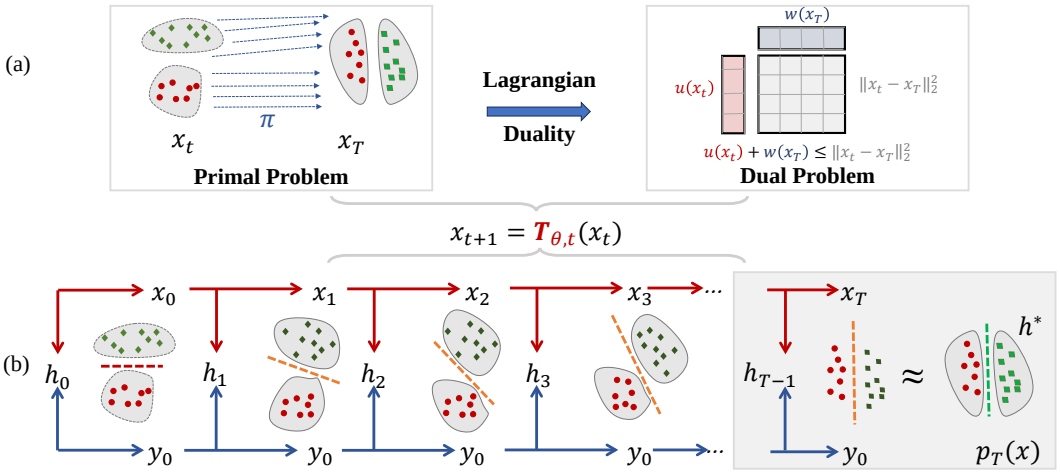


Figure 2: The illustration of the proposed E-SUOT: (a) the dual formulation for obtaining the transportation map, and (b) the evolution process from the source to the target domain.

Once the transport map sequence $\mathcal{T} = \{\boldsymbol{T}_{\theta,t}\}_{t=0}^{T-1}$ has been obtained, we proceed to train the classifier h in a stage-wise manner along the transport path. Specifically, at each intermediate step t, we first map samples x_t from the current domain to the next intermediate domain x_{t+1} using the corresponding transport map $\boldsymbol{T}_{\theta,t}$. We then update or train the model h_t using the mapped data x_{t+1} as input. By iteratively applying this procedure for $t = 0, \ldots, T-1$, the model is progressively adapted along the sequence of intermediate domains, ultimately bridging the source and target domains.

Algorithm 1 Overall Workflow for Construing E-SUOT-based Intermediate Domain Generation

Input: Source domain samples: $\{(x_0^{(i)}, y_0^{(i)})\}_{i=1}^{\mathrm{N}}$, target domain samples: $\{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\mathrm{N}}$, entropy regularization strength: ϵ, step size: η, number of intermediate domain $T-1$, neural network batch size $\mathcal{B}$, and neural network training epochs: $\mathcal{E}$.
Output: The set of transportation map: $\mathcal{T} = \{\boldsymbol{T}_{\theta,t}\}_{t=0}^{T-1}$.

$\mathcal{T} \leftarrow \varnothing$.
for $t = 0$ **to** $T-1$ **do**
  for $e = 1$ **to** $\mathcal{E}$ **do**
    Sample a batch $\{x_t^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_t^{(i)}, y_t^{(i)})\}_{i=1}^{\mathrm{N}}$ and $\{x_T^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\mathrm{N}}$.
    Update $w_{\phi,t}$ by: $\phi \leftarrow \arg\min_\phi \frac{\epsilon}{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} \log \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} [\exp(\frac{w_{\phi,t}(x_T^{(j)}) - \frac{1}{2\eta}\|x_t^{(j)} - x_T^{(i)}\|_2^2)}{\epsilon})] + \frac{1}{\mathcal{B}} \sum_{j=1}^{\mathcal{B}} f^\star(-w_{\phi,t}(x_T^{(j)}))$.
  end for
  for $e = 1$ **to** $\mathcal{E}$ **do**
    Sample a batch $\{x_t^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_t^{(i)}, y_t^{(i)})\}_{i=1}^{\mathrm{N}}$.
    Update $\boldsymbol{T}_{\theta,t}$ by: $\theta \leftarrow \arg\min_\theta \frac{1}{\mathcal{B}} \sum_{i=1}^{\mathcal{B}} \frac{1}{2\eta}\|x_t^{(i)} - \boldsymbol{T}_{\theta,t}(x_t^{(i)})\|_2^2 - w_{\phi,t}(\boldsymbol{T}_{\theta,t}(x_t^{(i)}))$.
  end for
  $x_{t+1}^{(i)} \leftarrow \boldsymbol{T}_{\theta,t}(x_t^{(i)}), \forall i \in \{1, \ldots, \mathrm{N}\}$.
  $\mathcal{T} \leftarrow \mathcal{T} \cup \{\boldsymbol{T}_{\theta,t}\}$
end for

3.5 Theoretical Analysis

Notably, our derivation sidesteps the explicit estimation of the PDF of the target domain by leveraging the semi-dual formulation. This naturally leads to two important questions: (1) Can the proposed E-SUOT framework transport the source domain sufficiently close to the target domain? (2) How does the model perform on the target domain after transport?

To address the first question, we present the following theorem, which quantitatively characterizes the discrepancy between $\rho(x)$ and $p_T(x)$:

Theorem 5. *The optimal solution $\rho^*(x)$ to problem defined in Eq. (8) satisfies the following bound:*

$$\mathbb{D}_f[\rho^*(x), p_T(x)] \leq \mathcal{W}_2(p(x_t), p_T(x)). \tag{11}$$

From Theorem 5, we observe that as t increases, the transported PDF $\rho(x)$ progressively becomes similar to $p_T(x)$. Based on this result, we present the following theorem, which provides a theoretical guarantee for the model's performance on the target domain:

Theorem 6. *Under mild assumptions, the E-SUOT-based GDA ensures that the target domain generalization error is upper-bounded by the following inequality:*

$$\varepsilon_{p_T}(h_T) \leq \varepsilon_{p_0}(h_0) + \varepsilon_{p_0}(h_T^*) + \iota\zeta\mathcal{C} + \mathcal{S}_{stat}, \tag{12}$$

where ι is the Lipschitz constant of the loss function, ζ is the Lipschitz constant bound for hypotheses in $\mathcal{H}$, $\mathcal{C}$ aggregates the cumulative domain transportation and label continuity costs along the adaptation path, and $\mathcal{S}_{stat}$ is the statistical error term.

4 Experimental Results

4.1 Experimental Setup

Datasets: We conduct case studies on three datasets. Specifically, the datasets are "Portraits" (Kumar et al., 2020), "MNIST 45°" and "MNIST 60°" (LeCun, 1998; Deng, 2012). For the last two datasets, we construct target domains by rotating vanilla images by 45° and 60°, thus referred to as MNIST 45° and MNIST 60°, respectively. Detailed information is given in Appendix D.1.

Implementation: Following prior work (Zhuang et al., 2024; Sagawa & Hino, 2025), we employ semi-supervised UMAP to produce low-dimensional embeddings while preserving class discriminability. Unless stated otherwise, we use the KL divergence in the implementation of the E-SUOT. Additional details are available in Appendix D.2.

4.2 Baseline Comparison Results

We first compare our proposed approach with several existing GDA-based methods, including Self-training, GST (4 intermediate domains) (Kumar et al., 2020), GOAT (He et al., 2024), CNF Sagawa & Hino (2025), and GGF (Zhuang et al., 2024).

Table 1: Baseline comparison on GDA setting.

Method	Portraits		MNIST 45°		MNIST 60°	
	Accuracy	Δ	Accuracy	Δ	Accuracy	Δ
Source	0.71	-	0.58	-	0.37	-
Self Train	0.77*	↑8.8%	0.59*	↑0.5%	0.40*	↑8.6%
GST (4)	0.76*	↑6.9%	0.59*	↑1.3%	0.40*	↑8.5%
GOAT	0.75*	↑5.3%	<u>0.65</u>*	↑11.3%	0.37*	↑1.1%
CNF	0.80*	↑12.4%	0.58	↓1.4%	<u>0.42</u>*	↑13.5%
GGF	<u>0.83</u>*	↑17.2%	0.58	↓1.2%	0.41*	↑11.0%
E-SUOT	**0.86***	↑21.5%	**0.72***	↑23.4%	**0.51***	↑38.6%

Kindly Note: "*" marks the variants that E-SUOT outperforms significantly at p-value < 0.05 over paired sample t-test. Δ denotes the performance change relative to the source classifier. **Bolded** and <u>underlined</u> results are the first and second best results, respectively.

As shown in Table 1, our proposed E-SUOT framework consistently outperforms the current state-of-the-art methods on all evaluated datasets. These results demonstrate the effectiveness and superiority of the E-SUOT framework. In addition, we observe that flow-based methods, such as CNF and GGF, generally achieve top-2 performance on most datasets, highlighting the potential of incorporating flow-based methods in GDA tasks. However, we also note that flow-based methods, occasionally underperform. This observation suggests that flow-based GDA, which requires explicit PDF estimation on target domain, may have inherent limitations, as discussed in Section 3.1.

4.3 Ablation Studies

We perform ablation studies from two perspectives: the training strategy for $\boldsymbol{T}_\theta$ and the choice of f-divergence. For the *training strategy*, we 1). examine the effect of removing the entropy regularization term—reducing the method to the adversarial training strategy in Eq. (7), and 2). evaluate a barycentric projection approach analogous to flow matching (Lipman et al., 2023), where the transport plan is first estimated and then used to project source samples toward the target, subsequently being refined during training. For the objective *functional*, we study different parameterizations of $f^\star$, such as employing non-decreasing convex functions like 1) `Softplus`, and also compare the 2) χ^2 divergence and the 3) identity function. More detailed information on these experiments' implementation is provided in Appendix D.3. The ablation study results are summarized in Table 2.

Table 2: Ablation study results.

Dataset			Portraits		MNIST 45°		MNIST 60°	
Metric			Accuracy	Δ	Accuracy	Δ	Accuracy	Δ
Training	Adversarial	KL	0.75*	↓9.4%	0.52*	↓27.8%	0.41	↓19.4%
	Barycentric	KL	0.84*	↓2.3%	0.62*	↓13.3%	0.41*	↓19.3%
Functional	Entropy	`Softplus`	0.80*	↓7.3%	0.60*	↓17.2%	0.38*	↓25.1%
	Entropy	χ^2	0.80	↓7.7%	0.60*	↓16.5%	0.42*	↓16.9%
	Entropy	Identity	0.81*	↓6.1%	0.60*	↓17.4%	0.40*	↓22.3%
	Entropy	KL	0.86	-	0.72	-	0.51	-

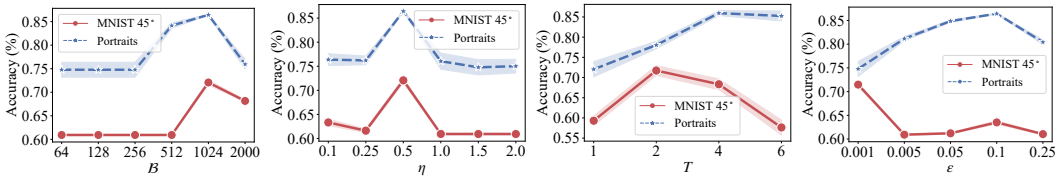
Kindly Note: "*" marks the variants that E-SUOT outperforms significantly at p-value $<$ 0.05 over paired sample t-test. Δ denotes performance change percentage compared to E-SUOT with entropy regularization and KL divergence.

From Table 2, we find that adversarial training performs the worst, underscoring the importance of entropy regularization for model training in Section 3.3. While barycentric mapping is competitive, it struggles on complex datasets such as MNIST 45 and MNIST 60, highlighting the need for the semi-dual formulation. Additionally, alternatives to KL divergence—especially `Softplus`—cause significant performance drops, emphasizing the importance of proper divergence selection. We also observe that replacing KL divergence with alternatives such as χ^2 divergence, the identity function, or particularly Softplus results in substantial performance degradation, further illustrating that choosing a suitable discrepancy to drive the evolution of source domain to target domain is critical for promising the performance of GDA.

4.4 Sensitivity Analysis

From Figs. 3(a) to 3(d), we systematically investigate the sensitivity of our E-SUOT model with respect to key hyperparameters, including batch size $\mathcal{B}$, discretization step size η, simulation steps T, and entropy regularization strength ϵ on the Portraits and MNIST 45° datasets.

Specifically, as shown in Fig. 3(a), we observe that increasing the batch size $\mathcal{B}$ initially improves model performance; however, after a certain point, further increasing the batch size leads to a performance decline. This pattern suggests that, in the simulation of WGF-based approaches (including ours), careful selection of batch size is crucial: if $\mathcal{B}$ is too small, stochastic sampling noise may dominate and degrade the results; conversely, excessively large $\mathcal{B}$ can cause the model to overfit and diminish its performance. A similar trend is found when varying the discretization step size η, as illustrated in Fig. 3(b). A small step size may prevent the simulation trajectory from adequately reaching the target distribution within a finite number of steps, limiting learning efficiency. On the other hand, a step size that is too large introduces significant discretization error, which again results in poor model performance. Furthermore, as demonstrated in Fig. 3(c), increasing the number of simulation steps T also produces a non-monotonic effect: beyond a certain threshold, more steps actually undermine performance. This is likely because aligning the feature/target distributions too strictly does not necessarily correspond to optimal performance in the target domain, thus further justifying our introduction of divergence-based regularization to relax strict alignment constraints compared to traditional OT-based methods. Finally, as shown in Fig. 3(d), the entropy regularization parameter ϵ also significantly influences results. We observe that varying ϵ can lead to diverse performance outcomes, highlighting the importance of properly investigating and tuning the entropy regularization strength in practical applications. In conclusion, our sensitivity study underscores the importance of carefully selecting the batch size $\mathcal{B}$, step size η, and end time T for E-SUOT performance, and further indicates that the entropy regularization strength ϵ is dataset-dependent and thus warrants systematic validation on the target dataset to achieve optimal E-SUOT performance.



(a) Accuracy (%) along $\mathcal{B}$. (b) Accuracy (%) along η. (c) Accuracy (%) along T. (d) Accuracy (%) along ϵ.

Figure 3: Sensitivity Analysis Results on Portrait and MNIST 45° Datasets.

4.5 Computational Time Comparison

In this subsection, we further analyze the empirical time complexity of the proposed E-SUOT approach in comparison with alternative methods on the GDA task. The computational time results are presented in Fig. 4.

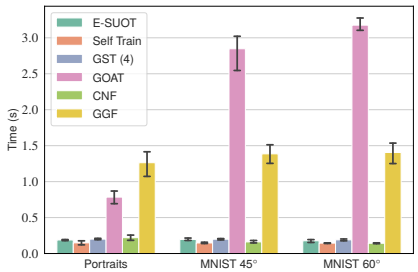


Figure 4: Computational time (s) comparison.

As shown in Fig. 4, the GOAT approach is the most time-consuming on larger datasets, while GGF takes more time on smaller datasets; both consistently rank among the top two in terms of computation cost. This can be attributed to their inherent algorithmic structures: GOAT involves solving the exact optimal transport problem, which becomes computationally prohibitive as the dataset size increases. In contrast, GGF relies on the forward Euler method, which requires a very small step size—and therefore a large number of iterations—to avoid significant simulation errors, resulting in higher computational overhead even on smaller datasets.Notably, the computational time of our proposed E-SUOT remains stable as dataset size grows. This efficiency stems from directly parameterizing the transport map using a single forward pass through a neural network and the JKO scheme, a variant of backward discretization approach, requiring only a few steps to achieve the desired performance.

5 Related Works

5.1 Gradual Domain Adaption

GDA seeks to bridge the distributional gap between source and target domains by leveraging a sequence of intermediate domains, thereby enabling more fine-grained adaptation. Early works have explored self-training strategies (Kumar et al., 2020), adversarial objectives (Wang et al., 2020), and provided generalization bounds under gradual distribution shifts (Kumar et al., 2020; Dong et al., 2022; Wang et al., 2022). However, these approaches often depend on the availability of discrete intermediate domains (Chen & Chao, 2021). To address this, optimal transport approaches (Abnar et al., 2021; He et al., 2024) have been leveraged to construct intermediate domains along the Wasserstein geodesic, ensuring minimal distributional discrepancy in the adaptation process. More recently, flow-based GDA has emerged, which explicitly models domain evolution and synthesizes continuous intermediate distributions via parametric flows. For instance, Sagawa & Hino (2025) uses continuous normalizing flows to parameterize domain trajectories as ODEs in the data space, while Zhuang et al. (2024) incorporates label information into this evolution and employs gradient flows to realize the steepest transformation from source to target domain. Nevertheless, flow-based methods still require explicit estimation of the target domain's PDF to guide the evolution, and inaccuracies in this estimation can lead to performance drops in target domain. To address this limitation, we reformulate the flow-based approach from a semi-dual formulation (see Proposition 1), which unifies the flow-based and optimal transport methods. Building on this, we further propose a convergence-guaranteed approach with the help of entropy regularization (Proposition 3) and analyze its generalization error (see Theorem 6). during the evolution of the flow and proposed gradient

5.2 Semi-Dual Formulation of Gradient Flows

Gradient flow (Santambrogio, 2017), which seeks to optimize a specified functional in the space of probability measures, has played a critical role in both sampling and optimization algorithm design. For gradient flows induced by f-divergences (with the KL divergence being the notable example), such as Langevin sampling (Welling & Teh, 2011), have been extensively explored to generate samples that progressively transition from the source domain toward the target domain. However, these methods typically assume access to an exact (unnormalized) PDF for the target distribution (Liu & Wang, 2016; Liu, 2017), which is often infeasible in practice when only samples are available. To overcome this, several approaches have explored dual formulations of f-divergence (Nguyen et al., 2007; 2010), which avoid explicit density estimation for the target domain and instead optimize primal formulation (Korotin et al., 2023; Rout et al., 2022; Fan et al., 2022; Gazdieva et al., 2023; Choi et al., 2023; 2024). These dual-formulation methods, however, generally require adversarial optimization characterized by a composite "$\sup\inf$" structure to in order to properly approximate the dual objective when implemented with neural networks (Nowozin et al., 2016; Arjovsky et al., 2017). Our work differs from these approaches in two key aspects. First, we provide a theoretical analysis from the perspective of the non-uniqueness of optimal solutions in Proposition 2, highlighting that such adversarial formulations can suffer from this issue, which may hinder training stability. Building upon this insight, we introduce the entropy regularization that transforms the adversarial game into an alternative paradigm in Proposition 3, and further prove that this regularization ensures the stability via the uniqueness of the optima in Proposition 4 and convergence in Theorem 5.

6 Conclusions

In this paper, we addressed the challenge in flow-based GDA, namely the reliance on explicit estimation of the target domain PDF inherited from traditional f-divergence formulations. To overcome this, we reformulated the flow simulation as an optimization problem augmented with a Wasserstein regularization term. Building on this, we derived a novel semi-dual formulation that avoids explicit estimation of the target density. However, we observed that the resulting semi-dual structure introduces instability due to its composite '$\sup\inf$' structure. To address this, we proposed an entropy regularization term that eliminates the inner $\inf$ operator, thereby restoring stability and ensuring uniqueness of the optimal solution. Based on these insights, we developed a new GDA framework called "E-SUOT" and provided theoretical guarantees for its convergence and generalization. Finally, extensive experiments validate the effectiveness and practical advantages of our approach.

Ethics Statement

The authors have read and comply with the ICLR Code of Ethics. This research does not involve human subjects or personally identifiable information, and uses only publicly available datasets under their respective licenses. We do not foresee harmful or dual-use implications from the proposed methods. There are no conflicts of interest or undisclosed sponsorship.

Reproducibility Statement

The anonymous downloadable source code is available at: https://anonymous.4open.science/r/E_SUOT_GDA-9240/. For theoretical results, the derivations proof of the claims are included in Appendix B. Based on this, a detailed overall workflow for the proposed E-SUOT is summarized in Appendix C. For datasets used in our experiments, we provide a complete description of the dataset statistics and processing work flow in Appendix D.

References

Samira Abnar, Rianne van den Berg, Golnaz Ghiasi, Mostafa Dehghani, Nal Kalchbrenner, and Hanie Sedghi. Gradual domain adaptation in the wild: When intermediate distributions are absent. *arXiv preprint arXiv:2106.06080*, pp. 1–15, 2021.

Ralph Abraham, Jerrold E Marsden, and Tudor Ratiu. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media, 2012.

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

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

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proc. Int. Conf. Mach. Learn.*, pp. 214–223. PMLR, 2017.

Heinz H Bauschke and Patrick L Combettes. Fenchel–rockafellar duality. In *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, pp. 247–262. Springer, 2017.

Lorenz T Biegler. *Nonlinear programming: concepts, algorithms, and applications to chemical processes*. SIAM, 2010.

Clément Bonet, Christophe Vauthier, and Anna Korba. Flowing datasets with wasserstein over wasserstein gradient flows. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–57. PMLR, 2025.

John Charles Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons, New York, USA, 2016.

Kenneth F. Caluya and Abhishek Halder. Gradient flow algorithms for density propagation in stochastic systems. *IEEE Trans. Autom. Control.*, 65(10):3991–4004, 2020.

Hong-You Chen and Wei-Lun Chao. Gradual domain adaptation without indexed intermediate domains. In *Proc. Adv. Neural inf. Process. Syst.*, volume 34, pp. 8201–8214, 2021.

Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 42433–42455, 2023.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Scalable wasserstein gradient flow for generative modeling through unbalanced optimal transport. In *Proc. Int. Conf. Mach. Learn.*, pp. 8629–8650. PMLR, 2024.

Jaemoo Choi, Jaewoong Choi, and Dohyun Kwon. Overcoming spurious solutions in semi-dual neural optimal transport: A smoothing approach for learning the optimal transport plan. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–21, 2025.

Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 274–289. Springer, 2014.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 30, pp. 1–10, 2017a.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(9):1853–1865, 2017b. doi: 10.1109/TPAMI.2016.2615921.

Li Deng. The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.*, 29(6):141–142, 2012. doi: 10.1109/MSP.2012.2211477.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 8780–8794, 2021.

Jing Dong, Shiji Zhou, Baoxiang Wang, and Han Zhao. Algorithms and theory for supervised gradual domain adaptation. *Trans. Mach. Learn. Res.*, pp. 1–14, 2022.

Lawrence C Evans. *Partial Differential Equations*, volume 19. American Mathematical Society, Amsterdam, Netherlands, 2022.

Jiaojiao Fan, Qinsheng Zhang, Amirhossein Taghvaei, and Yongxin Chen. Variational Wasserstein gradient flow. In *Proc. Int. Conf. Mach. Learn.*, pp. 6185–6215. PMLR, 2022.

Milena Gazdieva, Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Extremal domain translation with neural optimal transport. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 40381–40413, 2023.

Pierre Glaser, Michael Arbel, and Arthur Gretton. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 8018–8031, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

Yifei He, Haoxiang Wang, Bo Li, and Han Zhao. Gradual domain adaptation: Theory and algorithms. *J. Mach. Learn. Res.*, 25(361):1–40, 2024.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–8, 2015.

Ivan Kobyzev, Simon JD Prince, and Marcus A Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):3964–3979, 2020.

Alexander Korotin, Lingxiao Li, Aude Genevay, Justin M Solomon, Alexander Filippov, and Evgeny Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 14593–14605, 2021.

Alexander Korotin, Daniil Selikhanovych, and Evgeny Burnaev. Neural optimal transport. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–34, 2023.

Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pp. 5468–5479. PMLR, 2020.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–28, 2023.

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, and Jun Zhu. Understanding and accelerating particle-based variational inference. In *Proc. Int. Conf. Mach. Learn.*, pp. 4082–4092. PMLR, 2019.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Proc. Adv. Neural inf. Process. Syst.*, volume 30, pp. 1–15, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, pp. 1–9, 2016.

Weiming Liu, Jiajie Su, Chaochao Chen, and Xiaolin Zheng. Leveraging distribution alignment via stein path for cross-domain cold-start recommendation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 19223–19234, 2021.

Weiming Liu, Xiaolin Zheng, Jiajie Su, Longfei Zheng, Chaochao Chen, and Mengling Hu. Contrastive proxy kernel stein path alignment for cross-domain cold-start recommendation. *IEEE Trans. Knowl. Data Eng.*, 35(11):11216–11230, 2023. doi: 10.1109/TKDE.2022.3233789.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–9. PMLR, 2015.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, pp. 1–63, 2018.

Kirill Neklyudov, Jannes Nys, Luca Thiede, Juan Carrasquilla, Qiang Liu, Max Welling, and Alireza Makhzani. Wasserstein quantum Monte Carlo: a novel approach for solving the quantum many-body schrödinger equation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 63461–63482, 2023.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 20, pp. 1–8, 2007.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory.*, 56(11):5847–5861, 2010.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, pp. 1–9, 2016.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22 (10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Min Sue Park, Cheolhyeong Kim, Hwijae Son, and Hyung Ju Hwang. The deep minimizing movement scheme. *J. Comput. Phys.*, 494:112518, 2023.

Michaël Perrot, Nicolas Courty, Rémi Flamary, and Amaury Habrard. Mapping estimation for discrete optimal transport. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, pp. 1–9, 2016.

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

Litu Rout, Alexander Korotin, and Evgeny Burnaev. Generative modeling with optimal transport maps. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–22, 2022.

Shogo Sagawa and Hideitsu Hino. Gradual domain adaptation via normalizing flows. *Neural Comput.*, 37(3):522–568, 2025.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154, 2017.

Igal Sason and Sergio Verdú. f-divergence inequalities. *IEEE Trans. Inf. Theory.*, 62(11):5973–6006, 2016.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, London, 2014.

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with mirrored Stein operators. *Proc. Int. Conf. Learn. Represent.*, pp. 1–26, 2022.

Stephen Simons. Minimax theorems and their proofs. In *Minimax and applications*, pp. 1–23. Springer, Berlin, Germany, 1995.

Maurice Sion. On general minimax theorems. *Pacific J. Math.*, 8(4):171–176, 1958.

Marta Skreta, Tara Akhound-Sadegh, Viktor Ohanesian, Roberto Bondesan, Alan Aspuru-Guzik, Arnaud Doucet, Rob Brekelmans, Alexander Tong, and Kirill Neklyudov. Feynman-Kac correctors in diffusion: Annealing, guidance, and product of experts. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–44. PMLR, 2025.

Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. Sliced score matching: A scalable approach to density and score estimation. In *Proc. Conf. Uncertainty in Artificial Intelligence*, pp. 574–584. PMLR, 2020.

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, New York, 2012.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *J. Mach. Learn. Res.*, 8(5):1–21, 2007.

Hugo Touchette. Legendre-fenchel transforms in a nutshell. *URL http://www. maths. qmul. ac. uk/ht/archive/lfth2. pdf*, pp. 25, 2005.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2962–2971, 2017. doi: 10.1109/CVPR.2017.316.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE Trans. Neural. Netw.*, 10(5): 988–999, 1999.

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

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010. ISSN 1532-4435.

Fangyikang Wang, Huminhao Zhu, Chao Zhang, Hanbin Zhao, and Hui Qian. GAD-PVI: A general accelerated dynamic-weight particle-based variational inference framework. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pp. 1–29, 2023.

Hao Wang, Hao He, and Dina Katabi. Continuously indexed domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pp. 9898–9907, 2020.

Haoxiang Wang, Bo Li, and Han Zhao. Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *Proc. Int. Conf. Mach. Learn.*, pp. 22784–22801. PMLR, 2022.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proc. Int. Conf. Mach. Learn.*, pp. 681–688, 2011.

Haoyun Yin, Yixuan Qiu, and Xiao Wang. Wasserstein coreset via sinkhorn loss. *Transactions on Machine Learning Research*, pp. 1–40, 2025.

Zhichen Zeng, Ruizhong Qiu, Wenxuan Bao, Tianxin Wei, Xiao Lin, Yuchen Yan, Tarek F Abdelzaher, Jiawei Han, and Hanghang Tong. Pave your own path: Graph gradual domain adaptation on fused gromov-wasserstein geodesics. *arXiv preprint arXiv:2505.12709*, 2025.

Chao Zhang, Zhijian Li, Xin Du, and Hui Qian. DPVI: A dynamic-weight particle-based variational inference framework. In Lud De Raedt (ed.), *Proc. Int. Joint Conf. Artif. Intell.*, pp. 4900–4906. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

Xiaolin Zheng, Weiming Liu, Chaochao Chen, Jiajie Su, Xinting Liao, Mengling Hu, and Yanchao Tan. Mining user consistent and robust preference for unified cross domain recommendation. *IEEE Trans. Knowl. Data Eng.*, 36(12):8758–8772, 2024. doi: 10.1109/TKDE.2024.3446581.

Jia-Jie Zhu. Inclusive KL minimization: A Wasserstein-Fisher-Rao gradient flow perspective. In *Proc. Int. Conf. Learn. Represent. Frontiers in Probabilistic Inference: Learning meets Sampling (Workshop)*, 2025.

Zhan Zhuang, Yu Zhang, and Ying Wei. Gradual domain adaptation via gradient flow. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–26, 2024.

Contents for Appendix

A Mathematical Background on Optimal Transport

We begin by reviewing the relevant background of optimal transport, based on references (Villani et al., 2009; Peyré et al., 2019). Assume continuous variables with densities: source $\rho(x)$ supported on $\mathcal{X}$, target $\xi(y)$ supported on $\mathcal{Y}$, and a cost $c(x, y) \geq 0$. We search for a joint probability density function which is called transport plan $\pi(x, y) \geq 0$ such that:

$$\int \pi(x, y)\, \mathrm{d}y = \rho(x), \tag{A.1a}$$

$$\int \pi(x, y)\, \mathrm{d}x = \xi(y), \tag{A.1b}$$

and minimize expected cost:

$$\inf_{\pi \geq 0} \iint c(x, y)\, \pi(x, y)\, \mathrm{d}y\, \mathrm{d}x, \tag{A.2}$$

where $c(x, y)$ is the cost function, for example, squared Euclidean norm: $c(x, y) = \|x - y\|_2^2$. Notably, when $c(x, y)$ is chosen as the squared Euclidean distance, the resulting optimal transport cost corresponds to the squared Wasserstein-2 distance between the two PDFs.

Introducing potentials $u(x)$ and $w(y)$ as Lagrange multipliers for the marginal constraints, we get:

$$\sup_{u,w} \Big[\int u(x)\, \rho(x)\, \mathrm{d}x + \int w(y)\, \xi(y)\, \mathrm{d}y \Big] \quad \text{s.t.} \quad u(x) + w(y) \leq c(x, y) \;\; \forall x, y. \tag{A.3}$$

Intuitively, u and w are "prices"; the constraint ensures the total price never exceeds the cost function. In addition, u and w are also called "(Kantorovich) potential" in optimal transport.

Based on this, we can eliminate one potential via the c-transform as follows:

$$w^c(x) := \inf_y c(x, y) - w(y). \tag{A.4}$$

Based on this, we get the semi-dual formulation of optimal transport problem (Korotin et al., 2021; 2023; Choi et al., 2023; 2024; 2025) which maximizes over one potential:

$$\sup_w \int w^c(x)\, \rho(x)\, \mathrm{d}x + \int w(y)\, \xi(y)\, \mathrm{d}y. \tag{A.5}$$

Notably, when total mass may differ or we allow creation/destruction of mass, we can relax marginal constraints using the f-divergence-based penalty terms (Chizat et al., 2018; Zhang et al., 2022). Specifically, we still want to optimize $\pi(x, y) \geq 0$, but we will penalize deviations of the induced marginals $\tilde{\rho}(x) := \int \pi(x, y)\, \mathrm{d}y$ and $\tilde{\xi}(y) := \int \pi(x, y)\, \mathrm{d}x$ from $\rho(x)$ and $\xi(y)$:

$$\min_{\pi \geq 0} \iint c(x, y)\, \pi(x, y)\, \mathrm{d}y\, \mathrm{d}x + \lambda_1\, \mathbb{D}_f(\tilde{\rho}(x), \rho(x)) + \lambda_2\, \mathbb{D}_f(\tilde{\xi}(y), \xi(y)), \tag{A.6}$$

where $\mathbb{D}_f(\tilde{\rho}(x), \rho(x)) = \int \rho(x)\, f\big(\frac{\tilde{\rho}(x)}{\rho(x)}\big)\, \mathrm{d}x$ and $\lambda_{1,2} > 0$.

In addition, using the convex conjugate $f^\star$, the dual problem becomes

$$\max_{u,w} - \int \rho(x)\, f_1^\star\big(-u(x)\big)\, \mathrm{d}x - \int \xi(y)\, f_2^\star\big(-w(y)\big)\, \mathrm{d}y \quad \text{s.t.} \quad u(x) + w(y) \leq c(x, y)\, \forall x, y, \tag{A.7}$$

where f_1, f_2 are the chosen divergences on each side.

Similarly, we can eliminate one potential via the c-transform as follows:

$$\max_w - \int \rho(x)\, f_1^\star\big(-w^c(x)\big)\, \mathrm{d}x - \int \xi(y)\, f_2^\star\big(-w(y)\big)\, \mathrm{d}y, w^c(x) = \inf_y \{c(x, y) - w(y)\}. \tag{A.8}$$

Based on this, we obtain the semi-dual formulation of the unbalanced optimal transport problem.

B THEORETICAL DERIVATION

B.1 DERIVATION OF EQ. (4)

In this subsection, we want to derive the following equivalent relationship in the main content to uphold the rigor of our manuscript:

$$x_{t+\eta} = x_t - \eta \nabla \frac{\delta \mathbb{D}_f[p(x_t), p_T]}{\delta p(x_t)} \Rightarrow p(x_{t+\eta}) = \underset{\rho(x)\in\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})}{\arg\min} \frac{1}{2\eta} \mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)]. \tag{B.1}$$

Notably, the optimization problem given by the right-hand-side of the abovementioned equation is also called Jordan-Kinderlehrer-Otto canonical form (Jordan et al., 1998; Caluya & Halder, 2020) or minimum movement scheme (Park et al., 2023). Before conducting the derivation, it is necessary to introduce the definition of Wasserstein distance. The squared 2-Wasserstein distance $\mathcal{W}_2^2$ can be defined by finding a transport map $\boldsymbol{T} : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}^{\mathrm{D}}$ that minimizes the average cost of transporting mass from $\rho(x)$ to $\xi(x)$ as follows:

$$\mathcal{W}_2^2(\rho, \xi) = \inf_{\boldsymbol{T}:\boldsymbol{T}_{\#}\rho(x)=\xi(x)} \int \|x - \boldsymbol{T}(x)\|_2^2 \, \rho(x) \mathrm{d}x, \tag{B.2}$$

where $\boldsymbol{T}_{\#}$ indicates the pushforward measure, and the expression for $\boldsymbol{T}(x)$ is defined as follows:

$$\boldsymbol{T}(x) = x + \eta v_t(x). \tag{B.3}$$

Meanwhile, during the transportation, the differential equation that delineates PDF of the evolution process driven by Eq. (1) is called *continuity equation*, defined as follows:

$$\frac{\partial \rho(x_t)}{\partial t} = -\nabla \cdot [v_t(x_t)\rho(x_t)]. \tag{B.4}$$

Building on Eqs. (B.3) and (B.4), and discretizing the continuity equation in the time domain using the forward Euler scheme (Butcher, 2016; Evans, 2022), we obtain:

$$\rho(x) = \rho(x_t) - \eta \nabla \cdot (\rho(x_t) v_t(x_t)) + \mathcal{O}(\eta^2). \tag{B.5}$$

Taking the functional derivative of $\mathbb{D}_f[\rho(x), p_T(x)]$ with respect to $\rho(x)$, we get:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta} \mathbb{D}_f[\rho(x), p_T(x)] &= \frac{\mathrm{d}}{\mathrm{d}\eta} \int p_T(x) \, f\left(\frac{\rho(x)}{p_T(x)}\right) dx \\
&= \int p_T(x) \, \frac{\mathrm{d}}{\mathrm{d}\eta} f\left(\tfrac{\rho(x)}{p_T(x)}\right) dx \\
&= \int \cancel{p_T(x)} \, f'\left(\tfrac{\rho(x)}{p_T(x)}\right) \cancel{\tfrac{1}{p_T(x)}} \frac{\partial \rho(x)}{\partial \eta} dx \\
&\overset{\text{(i)}}{=} \int \frac{\delta \mathbb{D}_f}{\delta \rho(x)} \frac{\partial \rho(x)}{\partial \eta} dx
\end{aligned} \tag{B.6}$$

Here, step (i) is based on comparing the first variation:

$$\begin{aligned}
\delta \mathbb{D}_f[\rho; \sigma] &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0} \int p_T(x) \, f\left(\frac{\rho(x) + \varepsilon\sigma(x)}{p_T(x)}\right) dx \\
&= \int p_T(x) \, f'\left(\tfrac{\rho(x)}{p_T(x)}\right) \tfrac{1}{p_T(x)} \, \sigma(x) \, dx \qquad \text{(chain rule, } p_T \text{ fixed)} \\
&= \int f'\left(\tfrac{\rho(x)}{p_T(x)}\right) \sigma(x) \, dx,
\end{aligned}$$

with the definition of functional derivative:

$$\delta \mathbb{D}_f[\rho; \sigma] = \int \frac{\delta \mathbb{D}_f}{\delta \rho(x)} \sigma(x) \, dx,$$

where $\sigma(x)$ denotes an arbitrary perturbation function. Inserting Eq. (B.5) into Eq. (B.6), we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\eta}\mathbb{D}_f[\rho(x),p_T(x)] &= \int \frac{\delta\mathbb{D}_f}{\delta\rho(x)}\left[-\nabla\cdot(\rho(x)v_t(x))\right]\mathrm{d}x \\ &= \int \frac{\delta\mathbb{D}_f}{\delta\rho(x)}\left[-v_t^\top(x)\nabla\rho(x)-\rho(x)\nabla\cdot v_t(x)\right]\mathrm{d}x \\ &\overset{\text{(ii)}}{=} \int\Big(-\nabla\cdot\big[\tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\rho(x)v_t(x)\big]+\rho(x)\,v_t^\top(x)\,\nabla\tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\Big)\mathrm{d}x \\ &\overset{\text{(iii)}}{=} \int\rho(x)\,v_t^\top(x)\,\nabla\frac{\delta\mathbb{D}_f(\rho(x),p_T(x))}{\delta\rho(x)}\,\mathrm{d}x.\end{aligned} \tag{B.7}$$

Step (ii) is based on the chain rule:

$$\begin{aligned}\nabla\cdot\Big[\tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\,\rho(x)v_t(x)\Big] &= \tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\,\rho(x)\,[\nabla\cdot v_t(x)] \\ &\quad+\tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\,v_t^\top(x)\,\nabla\rho(x) \\ &\quad+\big[\nabla\tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\big]^\top[\rho(x)v_t(x)].\end{aligned} \tag{B.8}$$

Step (iii) uses a mild regularity assumption (Abraham et al., 2012; Liu et al., 2019; Shi et al., 2022) on $\frac{\delta\mathbb{D}_f}{\delta\rho(x)}\rho(x)v_t(x)$, for example rapid decay as $x\to\infty$, so that

$$\int-\nabla\cdot\Big[\tfrac{\delta\mathbb{D}_f}{\delta\rho(x)}\,\rho(x)v_t(x)\Big]\mathrm{d}x=0. \tag{B.9}$$

Consequently, $\mathbb{D}_f[\rho(x),p_T(x)]$ can be expanded as follows when $\eta\to 0$:

$$\mathbb{D}_f[\rho(x),p_T(x)]=\mathbb{D}_f[p(x_t),p_T(x)]+\eta\int p(x_t)v_t^\top(x_t)\nabla\frac{\delta\mathbb{D}_f[p(x_t),p_T(x)]}{\delta p(x_t)}\mathrm{d}x. \tag{B.10}$$

For the squared 2-Wasserstein distance, we get:

$$\mathcal{W}_2^2(\rho(x),p(x_t))=\int p(x_t)\|x-\boldsymbol{T}^*(x_t)\|_2^2\mathrm{d}x=\eta^2\int p(x_t)\|v_t^*(x_t)\|_2^2\mathrm{d}x\le\eta^2\int p(x_t)\|v_t(x_t)\|_2^2\mathrm{d}x, \tag{B.11}$$

where $\boldsymbol{T}^*(x)$ and $v_t^*(x)$ are the optimal transportation map and optimal velocity field. Since $v_t(x)$ is not the optimal velocity filed, we obtain the last inequality. Based on Eqs. (B.10) and (B.11), we finally reach the following result:

$$\begin{aligned}&\mathbb{D}_f[\rho(x),p_T(x)]+\frac{1}{2\eta}\mathcal{W}_2^2(\rho(x),p(x_t))-\mathbb{D}_f[p(x_t),p_T(x)] \\ \le&\cancel{\mathbb{D}_f[\rho(x),p_T(x)]}+\frac{\eta}{2}\mathbb{E}_{p(x_t)}[\|v_t(x_t)\|_2^2]+\eta\int\cdot[p(x_t)v_t^\top(x_t)\nabla\frac{\delta\mathbb{D}_f[p(x_t),p_T(x)]}{\delta p(x_t)}]\mathrm{d}x-\cancel{\mathbb{D}_f[\rho(x),p_T(x)]} \\ \le&\underbrace{\frac{\eta}{2}\mathbb{E}_{p(x_t)}[\|\nabla\frac{\delta\mathbb{D}_f[p(x_t),p_T(x)]}{\delta p(x_t)}]\|_2^2]}_{\ge 0}+\frac{\eta}{2}\mathbb{E}_{p(x_t)}[\|v_t(x_t)\|_2^2]+\eta\int p(x_t)v_t^\top(x_t)\nabla\frac{\delta\mathbb{D}_f[p(x_t),p_T(x)]}{\delta p(x_t)}\mathrm{d}x \\ =&\frac{\eta}{2}\mathbb{E}_{p(x_t)}\{\|v_t(x_t)+\nabla\frac{\delta\mathbb{D}_f[p(x_t),p_T(x)]}{\delta p(x_t)}\|_2^2\}.\end{aligned} \tag{B.12}$$

Consequently, the optimal velocity field that reduces the upper bound of the optimization problem defined by the right-hand-side of Eq. (4) can be given as follows:

$$v_t^*(x_t)=-\nabla\frac{\delta\mathbb{D}_f[p(x_t),p_T(x)]}{\delta p(x_t)}, \tag{B.13}$$

which implies that the left-hand side of Eq. (B.1) is a sufficient condition for the optimality of its right-hand side.

B.2 Derivation of Proposition 1

Proposition (1). *Consider the following primal problem:*

$$\mathcal{L}^{Primal} = \underset{\rho(x)\in\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})}{\arg\min} \frac{1}{2\eta}\,\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)]. \tag{B.14}$$

This problem is equivalent to the following semi-dual formulation:

$$\mathcal{L}^{SemiDual} = \sup_{w} \mathbb{E}_{p(x_t)}\left[\inf_{\boldsymbol{T}}\left(\|\boldsymbol{T}(x_t) - x_t\|_2^2 - w(\boldsymbol{T}(x_t))\right)\right] - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{B.15}$$

where $w : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ is a measurable continuous function, $\boldsymbol{T} : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}^{\mathrm{D}}$ is the transport map, and $f^\star$ denotes the convex conjugate of f, defined as $f^\star(z) := \sup_{y\geq 0}(zy - f(y))$.

Proof. Eq. (B.14) can be reformulated as follows:

$$\inf_{\pi\in\mathbb{R}_+^{\mathrm{D}\times\mathrm{D}}} \quad \frac{1}{2\eta}\iint \|x_t - x\|_2^2\pi(x_t, x)\mathrm{d}x_t\mathrm{d}x + \int f\left(\frac{\rho(x)}{p_T(x)}\right)p_T(x)\mathrm{d}x, \tag{B.16a}$$

$$\text{s.t.} \quad p(x_t) = \int \pi(x_t, x)\mathrm{d}x, \quad \rho(x) = \int \pi(x_t, x)\mathrm{d}x_t. \tag{B.16b}$$

Based on this, we introduce the Lagrangian multiplier Biegler (2010) $u(x_t)$ and $w(x)$ to handle the equality constraints given by Eq. (B.16b) as follows:

$$\begin{aligned}
\mathcal{L} =& \frac{1}{2\eta}\iint \|x_t - x\|_2^2\pi(x_t, x)\mathrm{d}x_t\mathrm{d}x + \int f\left(\frac{\rho(x)}{p_T(x)}\right)p_T(x)\mathrm{d}x \\
&+ \int u(x_t)[p(x_t) - \int \pi(x_t, x)\mathrm{d}x]\mathrm{d}x_t + \int w(x)[\rho(x) - \int \pi(x_t, x)\mathrm{d}x_t]\mathrm{d}x \\
=& \iint [\frac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x)]\pi(x_t, x)\mathrm{d}x_t\mathrm{d}x \\
&+ \int u(x_t)p(x_t)\mathrm{d}x_t + \int w(x)\rho(x) + f\left(\frac{\rho(x)}{p_T(x)}\right)p_T(x)\mathrm{d}x.
\end{aligned} \tag{B.17}$$

On this basis, the dual function can be given as follows due to the linear independent structure of problem defined by Eq. (B.17):

$$\begin{aligned}
g(u, w) \;=\; & \inf_{\pi(x_t,x)} \iint \left[\tfrac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x)\right]\pi(x_t, x)\,\mathrm{d}x_t\,\mathrm{d}x \\
& + \int u(x_t)p(x_t)\,\mathrm{d}x_t + \inf_{\rho(x)} \int \left[w(x)\,\frac{\rho(x)}{p_T(x)} + f\left(\tfrac{\rho(x)}{p_T(x)}\right)\right]p_T(x)\,\mathrm{d}x \\
=\; & \inf_{\pi(x_t,x)} \iint \left[\tfrac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x)\right]\pi(x_t, x)\,\mathrm{d}x_t\,\mathrm{d}x \\
& + \int u(x_t)p(x_t)\,\mathrm{d}x_t - \int p_T(x)\,f^\star(-w(x))\,\mathrm{d}x.
\end{aligned} \tag{B.18}$$

where the last line uses the Legendre–Fenchel conjugate (Touchette, 2005; Caluya & Halder, 2020). Writing $y(x) = \rho(x)/p_T(x)$ and using separability, we have

$$\begin{aligned}
\inf_{\rho(x)} \int \left[w(x)\,\frac{\rho(x)}{p_T(x)} + f\left(\tfrac{\rho(x)}{p_T(x)}\right)\right]p_T(x)\,\mathrm{d}x &= \int \inf_{y\geq 0}\left(w(x)\,y + f(y)\right)p_T(x)\,\mathrm{d}x \\
&= -\int \sup_{y\geq 0}\left((-w(x))\,y - f(y)\right)p_T(x)\,\mathrm{d}x \\
&= -\int p_T(x)\,f^\star(-w(x))\,\mathrm{d}x.
\end{aligned} \tag{B.19}$$

Suppose that $\frac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x) < 0$ for some pair (x_t, x). In this case, concentrating all the mass of $\pi(x_t, x)$ at this point drives the Lagrangian in Eq. (B.18) to $-\infty$. To avoid such degenerate solutions, it is necessary to impose the condition $\frac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x) \geq 0$ almost everywhere. Consequently, the dual problem can be written as

$$\sup_{u(x_t)+w(x)\leq \frac{1}{2\eta}\|x_t - x\|_2^2 \ \pi\text{-a.e.}} \left\{ \int u(x_t)p(x_t)\,dx_t - \int p_T(x) f^\star(-w(x))\,dx \right\}. \tag{B.20}$$

Equivalently, introducing the convex indicator function ℓ, this becomes

$$\sup_{u,w} \left\{ \int u(x_t)p(x_t)\,dx_t - \int p_T(x) f^\star(-w(x))\,dx - \ell\left(u(x_t) + w(x) \leq \frac{1}{2\eta}\|x_t - x\|_2^2\right) \right\}. \tag{B.21}$$

Since $f^\star$ is convex, non-decreasing, and differentiable, and because $\|x_t - x\|_2^2 \geq 0$, the choice $u(x_t) \equiv -1$ and $w(x) \equiv -1$ ensures all terms in Eq. (B.21) are finite. By Fenchel–Rockafellar's theorem (Bauschke & Combettes, 2017), strong duality therefore holds. Moreover, by complementary slackness the optimal plan π^* assigns zero mass to pairs where $\frac{1}{2\eta}\|x_t - x\|_2^2 - u^*(x_t) - w^*(x) > 0$, implying that $\frac{1}{2\eta}\|x_t - x\|_2^2 = u^*(x_t) + w^*(x)$ π^*-almost everywhere. Hence,

$$u^*(x_t) = \inf_x \left(\frac{1}{2\eta}\|x_t - x\|^2 - w^*(x)\right).$$

Substituting this into the dual yields the semi-dual formulation

$$\sup_{w(x)} \left\{ \int \inf_x \left[\frac{1}{2\eta}\|x_t - x\|_2^2 - w(x)\right] p(x_t)\,dx_t - \int p_T(x) f^\star(-w(x))\,dx \right\}. \tag{B.22}$$

Defining the transport map via the c-transform as

$$\begin{aligned} &\boldsymbol{T}^*(x_t) \in \arg\min_x \left(\frac{1}{2\eta}\|x_t - x\|_2^2 - w(x)\right) \\ \iff\ &\inf_x \left(\frac{1}{2\eta}\|x_t - x\|_2^2 - w(x)\right) = \frac{1}{2\eta}\|x_t - \boldsymbol{T}^*(x_t)\|_2^2 - w(\boldsymbol{T}^*(x_t)), \end{aligned} \tag{B.23}$$

and substituting Eq. (B.23) into Eq. (B.22), we obtain the final semi-dual objective

$$\mathcal{L}^{\text{SemiDual}} = \sup_w \mathbb{E}_{p(x_t)} \left[\|\frac{1}{2\eta}\boldsymbol{T}^*(x_t) - x_t\|_2^2 - w(\boldsymbol{T}^*(x_t)) \right] - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{B.24}$$

It should be pointed out that there is no closed-form expression of the optimal $\boldsymbol{T}^*(x_t)$ for each $w(x)$ (Korotin et al., 2023; Choi et al., 2023). Hence, the optimization $\boldsymbol{T}(x_t)$ for each $w(x)$ is required, and we reach the final semi-dual objective as follows based on Eq. (B.24):

$$\mathcal{L}^{\text{SemiDual}} = \sup_w \mathbb{E}_{p(x_t)} \left[\inf_{\boldsymbol{T}} \left(\frac{1}{2\eta}\|\boldsymbol{T}(x_t) - x_t\|_2^2 - w(\boldsymbol{T}(x_t)) \right) \right] - \mathbb{E}_{p_T(x)}[f^\star(-w(x))].$$

□

B.3 Derivation of Proposition 2

Proposition (2). *The semi-dual formulation in Eq. (7) admits non-unique optimal solutions.*

Proof. Consider the discrete optimal transport setting with a single source point (x_t in Eq. (7)) and two symmetric target points (x in Eq. (7)). Augment the dual objective with an f-divergence term acting only on the target potential w, but not on the source potential u. Then the dual optimizer is not unique.

Specifically, let:

- **Source space:** $x_t = \{a\}$ with $p(x_t) \approx \delta_a$.

- **Target space:** $x = \{b_1, b_2\}$ with $\rho(x) = \frac{1}{2}\delta_{b_1} + \frac{1}{2}\delta_{b_2}$.
- **Cost constant on pairs:** $\|a - b_1\|_2^2 = \|a - b_2\|_2^2 = K$ for some fixed $K \in \mathbb{R}$.

The dual problem obtained from the primal with an additional term $\int f\left(\frac{\rho(x)}{p_T(x)}\right) p_T(x)\,\mathrm{d}x$ acting only on the target side admits multiple optimal solutions (u, w); in particular, uniqueness fails.

The demonstration process can be summarized as follows:

1) At the beginning, let us recall the feasibility for the multipliers u and w:

$$u(a) + w(b_j) \le \|a - b_j\|_2^2 = K, \quad \forall j \in \{1, 2\}. \tag{B.25}$$

Based on this, we can define a shifted source potential $\tilde{u} := u - K$ and keep $\tilde{w} := w$. Hence, the feasibility in Eq. (B.25) can be given as follows:

$$\tilde{u}(a) + \tilde{w}(b_j) \le 0, \quad \forall j \in \{1, 2\}, \tag{B.26}$$

where the dual objective differs from the original by a global additive constant (independent of $(\tilde{u}, \tilde{w})$), hence the set of maximizers is unaffected by this normalization. As such, without loss of generality, it suffices to analyze the case $K = 0$. For notational simplicity we drop tildes and write

$$u + w_j \le 0, \quad \forall j \in \{1, 2\}. \tag{B.27}$$

2) Eliminating u and obtaining a piecewise-linear term Since $p(a) = 1$ and $\rho(b_1) = \rho(b_2) = \frac{1}{2}$, the dual objective function (up to an additive constant) can be reformulated as follows:

$$\max_{u, w_1, w_2} \quad u + \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 \; - \; \tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2), \tag{B.28}$$

subject to $u \le -w_1$ and $u \le -w_2$. At optimum the constraint in u is tight, hence we have the following result:

$$u = -\min\{w_1, w_2\}. \tag{B.29}$$

Substituting back yields an equivalent maximization over (w_1, w_2):

$$\Phi(w_1, w_2) \coloneqq -\min\{w_1, w_2\} + \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 - \tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2). \tag{B.30}$$

On this basis, we can define the "hinge" (V-shaped) linear part as follows:

$$L(w_1, w_2) := -\min\{w_1, w_2\} + \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 = \begin{cases} \frac{1}{2}(w_2 - w_1), & w_1 \le w_2, \\ \frac{1}{2}(w_1 - w_2), & w_2 \le w_1, \end{cases} \tag{B.31}$$

so that $L(w_1, w_2) = \frac{1}{2}|w_1 - w_2|$ and in particular $L(r, r) = 0$ for all r.

Consequently, we have:

$$\Phi(w_1, w_2) = \tfrac{1}{2}|w_1 - w_2| - \tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2). \tag{B.32}$$

3) Notably, on the diagonal $w_1 = w_2 = r$, we have the following result:

$$\Phi(r, r) = -f^\star(-r). \tag{B.33}$$

Since $f^\star$ is strictly convex, the one-dimensional problem $\max_t \Phi(r, r)$ has a unique maximizer r^*. Now let us consider antisymmetric perturbations around the diagonal:

$$w_1 = r^* + \delta, \qquad w_2 = r^* - \delta, \qquad \delta \in \mathbb{R}. \tag{B.34}$$

Then we obtain the following result:

$$\tfrac{1}{2}|w_1 - w_2| = \tfrac{1}{2}|2\delta| = |\delta|. \tag{B.35}$$

Using the second-order Taylor expansion of the strictly convex function $f^\star$ about $-r^*$, we have for the following equality for small $|\delta|$:

$$-\tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2) = -f^\star(-r^*) - \tfrac{1}{2}\left(f^{*\prime\prime}(-r^*)\right)\delta^2 + \mathcal{O}(\delta^2). \tag{B.36}$$

Based on this, we get:

$$\Phi(r^* + \delta,\, r^* - \delta) = |\delta| \;-\; \tfrac{1}{2} f^{*\prime\prime}(-r^*)\, \delta^2 \;-\; f^\star(-r^*) + \mathcal{O}(\delta^2). \tag{B.37}$$

It should be pointed out that, for any sufficiently small but nonzero δ, the linear gain term $|\delta|$ dominates the quadratic penalty term $\frac{1}{2} f^{*\prime\prime}(-r^*)\delta^2$, hence

$$\Phi(r^* + \delta,\, r^* - \delta) > \Phi(r^*, r^*).$$

Consequently, the diagonal point $(w_1, w_2) = (r^*, r^*)$ is not uniquely optimal; in fact, there exists a continuum of distinct maximizers in a neighborhood along the antisymmetric direction. The corresponding u is

$$u = -\min\{w_1, w_2\} = \begin{cases} -(r^* - \delta), & \delta \geq 0, \\ -(r^* + \delta), & \delta < 0, \end{cases}$$

yielding distinct optimal triples (u, w_1, w_2) for different $\delta \neq 0$.

4) If the original cost is $\|a - b_j\|_2^2 = K$, recall $u = \tilde{u} + K$. Thus each optimal $(\tilde{u}, w)$ constructed above gives an optimal (u, w) for the original problem by adding K to u. As the set of optimal w-pairs is already non-singleton, the full optimal dual variable pair (u, w) is non-unique.

In summary, our proof is based on the counter-example mentioned above. Specifically, in the symmetric two-target discrete setting, with the additional f-term acting only on the target potential w, the dual objective contains a V-shaped hinge $L(w_1, w_2) = \frac{1}{2}|w_1 - w_2|$ arising from eliminating u. This non-strict component competes with the strictly convex penalty $-\sum_j \rho(b_j)\, f^\star(-w(b_j))$. Along antisymmetric perturbations, the first-order increase from the hinge dominates the second-order decrease from the convex penalty, producing a continuum of maximizers. Hence the optimal dual variable pair is not unique. Consequently, the dual problem defined in Eq. (7) admits non-unique optimal solutions. □

B.4 Derivation of Proposition 3

Proposition (3). *Let $\kappa(x_t, x) := p(x_t)\, p_T(x)$ denote the reference joint PDF. The entropy-regularized primal problem is*

$$\begin{aligned} \mathcal{L}^{\text{E-Primal}} = \operatorname*{arg\,min}_{\rho \in \mathcal{P}_2(\mathbb{R}^D)} \frac{1}{2\eta}\, \mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f\big[\rho(x), p_T(x)\big] \\ + \epsilon \iint \pi(x_t, x)\, [\log \frac{\pi(x_t, x)}{\kappa(x_t, x)} - 1]\, \mathrm{d}x_t\, \mathrm{d}x, \end{aligned} \tag{B.38}$$

and is equivalent to the semi-dual optimization problem

$$\mathcal{L}^{\text{E-SemiDual}} = \sup_{w} -\epsilon\, \mathbb{E}_{p(x_t)}[\log \mathbb{E}_{p_T(x)}(\exp(\frac{w(x) - \frac{1}{2\eta}\|x - x_t\|_2^2}{\epsilon}))] \;- \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{B.39}$$

where $f^\star$ denotes the convex conjugate of f.

Proof. Define $c(x_t, x) := \frac{1}{2\eta}\, \|x_t - x\|_2^2$ as the quadratic transport cost. Introducing Lagrange multipliers $u(x_t) : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ (for the x_t-marginal) and $w(x) : \mathbb{R}^{\mathrm{D}} \to \mathbb{R}$ (for the x-marginal). The Lagrangian of Eq. (B.38) is

$$\begin{aligned} \mathcal{L}(\pi, \rho; u, w) = & \iint c(x_t, x)\, \pi(x_t, x)\, dx_t\, dx + \epsilon \iint \pi(x_t, x) \Big[\log \frac{\pi(x_t, x)}{\kappa(x_t, x)} - 1\Big] dx_t\, dx \\ & + \int f\Big(\tfrac{\rho(x)}{p_T(x)}\Big) p_T(x)\, dx + \int u(x_t) \Big[p(x_t) - \int \pi(x_t, x)\, dx\Big] dx_t \\ & + \int w(x) \Big[\rho(x) - \int \pi(x_t, x)\, dx_t\Big] dx. \end{aligned} \tag{B.40}$$

Grouping π-, ρ- and constant terms yields

$$\begin{aligned}\mathcal{L} = &\iint \big(c(x_t,x) - u(x_t) - w(x)\big)\,\pi(x_t,x)\,dx_t\,dx \\ &+ \epsilon \iint \pi(x_t,x)\Big[\log\frac{\pi(x_t,x)}{\kappa(x_t,x)} - 1\Big]dx_t\,dx \\ &+ \int \Big(w(x)\rho(x) + f\big(\tfrac{\rho(x)}{p_T(x)}\big)p_T(x)\Big)dx + \int u(x_t)p(x_t)\,dx_t.\end{aligned} \tag{B.41}$$

Define $a(x_t,x) := c(x_t,x) - u(x_t) - w(x)$. For each fixed (x_t,x), minimize

$$\phi(y) := a\,y + \epsilon\Big(y\log\tfrac{y}{\kappa} - y\Big), \qquad y \geq 0.$$

The first-order condition $a + \epsilon\log(y/\kappa) = 0$ gives

$$y^\star = \kappa\, e^{-a/\epsilon} = \kappa\, e^{(u+w-c)/\epsilon}. \tag{B.42}$$

Substituting back yields

$$\inf_{y\geq 0}\phi(y) = -\epsilon\,\kappa\, e^{-a/\epsilon} = -\epsilon\,\kappa\,\exp\Big(\tfrac{u+w-c}{\epsilon}\Big). \tag{B.43}$$

Hence

$$\inf_{\pi\geq 0}\Big\{\pi\text{-terms of Eq. (B.41)}\Big\} = -\epsilon\iint \kappa(x_t,x)\,\exp\Big(\tfrac{u(x_t)+w(x)-c(x_t,x)}{\epsilon}\Big)\,dx_t\,dx. \tag{B.44}$$

For ρ, by Legendre–Fenchel conjugate (Touchette, 2005; Caluya & Halder, 2020), we have:

$$\inf_{\rho(x)\geq 0}\Big\{w(x)\rho(x) + f\big(\tfrac{\rho(x)}{p_T(x)}\big)p_T(x)\Big\} = -p_T(x)\,f^\star\big(-w(x)\big). \tag{B.45}$$

Integrating over x gives

$$\inf_{\rho}\int [w\,\rho(x_t) + f\big(\frac{\rho(x_t)}{p_T(x)}\big)\,p_T(x)]\mathrm{d}x = -\int p_T(x)\,f^\star\big(-w(x)\big)\,\mathrm{d}x.$$

Combining Eq. (B.44) and Eq. (B.45), we obtain

$$\begin{aligned}g(u,w) = &-\epsilon\iint \kappa(x_t,x)\,\exp\Big(\tfrac{u(x_t)+w(x)-c(x_t,x)}{\epsilon}\Big)dx_t\,dx \\ &-\int p_T(x)\,f^\star\big(-w(x)\big)\,dx + \int u(x_t)p(x_t)\,dx_t.\end{aligned} \tag{B.46}$$

Using $\kappa = p(x_t)\cdot p_T(x)$, define

$$A(x_t) := \int \exp\Big(\tfrac{w(x)-c(x_t,x)}{\epsilon}\Big)\,p_T(x)\,\mathrm{d}x. \tag{B.47}$$

Then

$$\iint \kappa\,\exp[\frac{(u(x_t)+w(x)-c(x_t,x))}{\epsilon}]\mathrm{d}x_t\mathrm{d}x = \int p(x_t)\,A(x_t)\,e^{\frac{u(x_t)}{\epsilon}}\,\mathrm{d}x_t.$$

Thus, Eq. (B.46) can be reformulated as follows:

$$g(u,w) = \int\Big[p(x_t)\,u(x_t) - \epsilon\,p(x_t)\,A(x_t)\,e^{u(x_t)/\epsilon}\Big]\mathrm{d}x_t - \int p_T(x)\,f^\star(-w(x))\,\mathrm{d}x. \tag{B.48}$$

For each x_t, consider

$$\psi_{x_t}(u) := p(x_t)\,u - \epsilon\,p(x_t)\,A(x_t)\,e^{\frac{u(x_t)}{\epsilon}}.$$

The first-order condition

$$\frac{d}{du}\psi_{x_t}(u) = p(x_t) - p(x_t)A(x_t)e^{\frac{u(x_t)}{\epsilon}} = 0$$

gives

$$e^{u^\star(x_t)/\epsilon} = \frac{1}{A(x_t)} \quad \Longleftrightarrow \quad u^\star(x_t) = -\,\epsilon \log A(x_t). \tag{B.49}$$

Substituting back,

$$\sup_u \psi_{x_t}(u) = \epsilon\, p(x_t)\big(-\log A(x_t) - 1\big).$$

Summing over x_t and discarding the constant $-\epsilon \int p(x_t)\,\mathrm{d}x_t = -\epsilon$ (independent of $w(x)$), we obtain the semi-dual

$$\sup_w \; -\epsilon\, \mathbb{E}_{p(x_t)}\big[\log A(x_t)\big] - \mathbb{E}_{p_T(x)}\big[f^\star(-w(x))\big], \tag{B.50}$$

with $A(x_t)$ defined in Eq. (B.47). □

B.5 Derivation of Proposition 4

Proposition (4). *The semi-dual formulation in Eq. (9) admits a unique optimal solution.*

Proof. Let the entropy-regularized dual objective in Eq. (9) be

$$g(w) = -\epsilon\, \mathbb{E}_{p(x_t)}\{\log \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - \|x - x_t\|_2^2}{\epsilon})]\} - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{B.51}$$

where $f^\star$ is assumed to be strictly convex and proper, and $\epsilon > 0$.

We seek to show that $g(w)$ is a strictly concave functional on an appropriate space of measurable functions w, thus its maximizer (if it exists) is unique.

Our proof can be given by the following steps

1) Define for fixed x_t:

$$\Phi_\epsilon(w; x_t) := -\epsilon \log \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - \|x - x_t\|_2^2}{\epsilon})], \tag{B.52}$$

The mapping $w \mapsto \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - C(x, x_t)}{\epsilon})]$ is log-convex by Hölder's inequality, and therefore, $w \mapsto \Phi_\epsilon(w; x_t)$ is strictly concave, except in directions where w differs only by an additive constant almost everywhere. Taking the expectation over x_t preserves strict concavity unless w is constant almost everywhere.

2) The term $-\mathbb{E}_x[f^\star(-w(x))]$ is strictly concave with respect to w because $f^\star$ is strictly convex. Specifically, for any distinct $w_1 \neq w_2$, strict convexity of $f^\star$ gives for all $\lambda \in (0, 1)$,

$$-\mathbb{E}_x[f^\star(-((1-\lambda)w_1(x) + \lambda w_2(x)))] > -(1-\lambda)\mathbb{E}_x[f^\star(-w_1(x))] - \lambda \mathbb{E}_x[f^\star(-w_2(x))]$$

provided $w_1(x) \neq w_2(x)$ on a set of positive measure.

3) Since the sum of a strictly concave function and a concave function is strictly concave, it follows that the full dual objective $g(w)$ is strictly concave on the set of admissible functions.

As a result, $g(w)$ admits at most one maximizer, and the proposition is proved. □

B.6 Derivation of Theorem 5

Theorem (5). *The optimal solution $\rho^*(x)$ to problem defined in Eq. (8) satisfies the following bound:*

$$\mathbb{D}_f[\rho^*(x), p_T(x)] \leq \mathcal{W}_2(p(x_t), p_T(x)). \tag{B.53}$$

Proof. To facilitate reading, we define the signal as follows:

$$\mathcal{W}_{2,\epsilon}^2(\rho,\xi) \coloneqq \inf_{\pi\in\Pi(\rho,\xi)} \iint \|x-y\|_2^2 \pi(x,y)\mathrm{d}x\mathrm{d}y \\ + \epsilon \iint \pi(\rho(x),p(x_t))[\log\pi(\rho(x),p(x_t))-1]\mathrm{d}x\mathrm{d}x_t, \tag{B.54}$$

The dual representation of the f-divergence based on the Legendre–Fenchel conjugate is:

$$\mathbb{D}_f[\rho(x),p_T(x)] = \sup_{v(x)} \left\{\mathbb{E}_{\rho(x)}[v(x)] - \mathbb{E}_{p_T(x)}\left[f^\star(v(x))\right]\right\}. \tag{B.55}$$

Thus, the problem defined in Eq. (8) can be written as:

$$\inf_{\rho(x)} \ \mathcal{W}_{2,\epsilon}(\rho(x),p(x_t)) + \sup_{v(x)}\{\mathbb{E}_{\rho(x)}[v(x)] - \mathbb{E}_{p_T(x)}\left[f^\star(v(x))\right]\} \tag{B.56}$$

Interchanging $\min_{\rho(x)}, \sup_{v(x)}$ by the convexity-concavity and Sion's theorem (Sion, 1958; Simons, 1995), we obtain the following result:

$$\sup_{v(x)} -\mathbb{E}_{p_T(x)}[f^\star(v(x))] + \inf_{\rho(x)}\{\mathcal{W}_{2,\epsilon}(\rho(x),p(x_t)) + \mathbb{E}_{\rho(x)}[v(x)]\} \tag{B.57}$$

The inner minimization with respect to $\rho(x)$ is precisely the entropic optimal transport problem in the semi-dual form for PDFs $\rho(x)$ and $p(x_t)$:

$$\min_{\rho(x)} \ \mathcal{W}_{2,\epsilon}(\rho(x),p(x_t)) + \mathbb{E}_{\rho(x)}\left[v(x)\right] \tag{B.58}$$

whose optimal value equals

$$\mathbb{E}_{p(x_t)}[-\epsilon\log\int\exp(\frac{v(x)-c(x_t,x)}{\epsilon})\mathrm{d}y]. \tag{B.59}$$

This follows from standard duality in entropic optimal transport.

Plug the expression above into the main problem:

$$\sup_{v(x)} \ \mathbb{E}_{p(x_t)}[-\epsilon\log\int\exp(\frac{v(x)-c(x_t,x)}{\epsilon})\mathrm{d}y] - \mathbb{E}_{p_T(x)}[f^\star(v(x))]. \tag{B.60}$$

This is the desired semi-dual form.

At optimality, plug in any variation $v = v^* + \delta\psi$ into $g(w)$ and take derivative w.r.t. δ at 0, then set to zero. The calculation is:

$$0 = \frac{\partial}{\partial\delta} g(v^*+\delta\psi)\Big|_{\delta=0} = \mathbb{E}_{p(x_t)}\left[\frac{\int\psi(x)\exp\left(\frac{v^*(x)-c(x_t,x)}{\epsilon}\right)\mathrm{d}x}{\int\exp\left(\frac{v^*(x)-c(x_t,x)}{\epsilon}\right)\mathrm{d}x}\right] - \mathbb{E}_{p_T(x)}\left[(f^\star)'(v^*(x))\psi(x)\right], \tag{B.61}$$

which for all test functions $\psi(x)$ implies

$$\underbrace{\int p(x_t)\frac{\exp\left(\frac{v^*(x)-c(x_t,x)}{\epsilon}\right)}{\int\exp\left(\frac{v^*(x)-c(x_t,x)}{\epsilon}\right)\mathrm{d}x}\mathrm{d}x_t}_{\coloneqq\tilde{p}_T(x)} = p_T(x)(f^\star)'(v^*(x)).$$

That is, the pushforward of $p(x_t)$ under the mapping:

$$\boldsymbol{T}^*(x|x_t) = \frac{\exp\left(\frac{v^*(x)-c(x_t,x)}{\epsilon}\right)}{\int\exp\left(\frac{v^*(x)-c(x_t,x)}{\epsilon}\right)\mathrm{d}x},$$

which indicates that

$$\tilde{p}_T(x) = p_T(x)(f^\star)'(v^*(x)). \tag{B.62}$$

So, $\rho^*(x) = \tilde{p}_T(x)$ is the marginal of the optimal transport π^* as claimed.

Since the value of the primal objective at $\rho(x) = p_T(x)$ gives an upper bound:

$$\mathbb{D}_f[\rho^*(x)\|p_T(x)] + \mathcal{W}_{2,\epsilon}(\rho^*(x), p(x_t)) \leq \mathcal{W}_{2,\epsilon}(p_T, p(x_t)). \tag{B.63}$$

So in particular, we get:

$$\mathbb{D}_f[\rho^*(x)\|p_T(x)] \leq \mathcal{W}_{2,\epsilon}(p(x_t), p_T(x)). \tag{B.64}$$

In addition, we notice that the following inequality holds for $\epsilon > 0$:

$$\epsilon \iint \pi(\rho(x), p(x_t))[\log \pi(\rho(x), p(x_t)) - 1]\mathrm{d}x\mathrm{d}x_t \leq 0. \tag{B.65}$$

Plugging Eq. (B.65) into Eq. (B.64), we arrive at the desired result. □

B.7 Derivation of Theorem 6

Theorem (6). *Under mild assumptions, the E-SUOT-based GDA ensures that the target domain generalization error is upper-bounded by the following inequality:*

$$\varepsilon_{p_T}(h_T) \leq \varepsilon_{p_0}(h_0) + \varepsilon_{p_0}(h_T^*) + \iota\zeta\mathcal{C} + \mathcal{S}_{stat}, \tag{B.66}$$

where ι is the Lipschitz constant of the loss function, ζ is the Lipschitz constant bound for hypotheses in $\mathcal{H}$, $\mathcal{C}$ aggregates the cumulative domain transportation and label continuity costs along the adaptation path, and $\mathcal{S}_{stat}$ is the statistical error term.

Before formally proving the theorem, we introduce the following assumptions, which are mild and commonly satisfied in practical domain adaptation scenarios:

- **(A. 1)** The loss function $\mathcal{L}(\cdot, y)$ is ι-Lipschitz with respect to its first argument; that is, for any a, a' and fixed y, we have:
$$|\mathcal{L}(a, y) - \mathcal{L}(a', y)| \leq \iota|a - a'|. \tag{B.67}$$
- **(A. 2)** Each hypothesis $h \in \mathcal{H}$ is ζ-Lipschitz, i.e., for any x, x', we have:
$$|h(x) - h(x')| \leq \zeta\|x - x'\|. \tag{B.68}$$
- **(A. 3)** The labeling function q_t along the adaptation path is such that $|q_t(x) - q_{t-1}(x)|$ is small for most x, to ensure local continuity.
- **(A. 4)** The sequence of domains $(p_0, p_1, \ldots, p_T)$ is induced by E-SUOT-based GDA transport, so that the total cumulative cost $\mathcal{C}$ as defined below is finite.
- **(A. 5)** At every step, empirical risk minimization over sufficient samples ensures a small empirical-to-expected error gap, leading to a statistical error term $\mathcal{S}_{\text{stat}}$.
- **(A. 6)** The sample size for each domain is large enough to make $\mathcal{S}_{\text{stat}}$ negligible in the asymptotic regime.

Notably, Assumptions (A.1), (A.2), (A.5) and (A.6) are standard and generally hold for commonly used loss functions and hypothesis classes. Unless the loss or model is exceptionally non-standard, these can be stated directly with the theorem and do not require additional justification. Assumption (A.3) holds in cases where the labeling function is changes smoothly along the adaptation path. For our construction, since the intermediate domains are generated by incremental, continuous transformations, we have $\mathbb{E}_{p_{t-1}(x)}|q_t(x) - q_{t-1}(x)|$ is small for every t. As for Assumption (A.4), in our E-SUOT-based GDA, each domain is generated via an iterative unbalanced optimal transport step that progressively reduces the transport cost as we proved in Theorem 6. This guarantees that the cumulative cost $\mathcal{C}$ is finite, as can be bounded analytically. In summary, all the above assumptions are justified in our setting. Based on these assumptions, we now proceed with the formal proof.

Proof. Our goal is to bound the target risk $\varepsilon_{p_T}(h_T)$. Consider the telescoping sum along the domain adaptation path:

$$\varepsilon_{p_T}(h_T) = \varepsilon_{p_0}(h_0) + [\varepsilon_{p_T}(h_T) - \varepsilon_{p_0}(h_0)]\,. \tag{B.69}$$

To make the recursion explicit, rewrite this as:

$$\varepsilon_{p_T}(h_T) = \varepsilon_{p_0}(h_0) + \sum_{t=1}^{T} \left[\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1})\right]. \tag{B.70}$$

For each $t \in \{0, \ldots, T-1\}$, we observe that

$$\begin{aligned} &\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1}) \\ &= \underbrace{\left[\varepsilon_{p_t}(h_t) - \varepsilon_{p_t}(h_{t-1})\right]}_{\text{optimization error}} + \underbrace{\left[\varepsilon_{p_t}(h_{t-1}) - \varepsilon_{p_{t-1}}(h_{t-1})\right]}_{\text{domain shift term}} + \underbrace{\left[\varepsilon_{p_{t-1}}(h_{t-1}) - \varepsilon_{p_{t-1}}(h_t)\right]}_{\leq 0 \text{ by ERM}} \end{aligned} \tag{B.71}$$

In practice, the last term is non-positive since 'empirical risk minimization' (Vapnik, 1999; Shalev-Shwartz & Ben-David, 2014; Zhuang et al., 2024) ensures moving toward lower risk, so we can drop it for an upper bound.

By the Lipschitz property of $\mathcal{L}$ and h,

$$|\varepsilon_{p_t}(h) - \varepsilon_{p_{t-1}}(h)| \leq \iota\zeta \cdot \mathcal{W}_1(p_{t-1}, p_t). \tag{B.72}$$

Suppose the true label function q_t changes along the path. Following standard analysis, this gives an additional cost due to the label discrepancy:

$$\iota\, \mathbb{E}_{p_t(x)}|q_t(x) - q_{t-1}(x)|. \tag{B.73}$$

Therefore, each step can be bounded by

$$|\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1})| \leq \iota\zeta\mathcal{W}_1(p_{t-1}, p_t) + \iota\, \mathbb{E}_{p_t(x)}|f_t(x) - f_{t-1}(x)| + s_t \tag{B.74}$$

where s_t denotes the statistical error at step t.

Let

$$\mathcal{C} := \sum_{t=0}^{T-1} \left[\mathcal{W}_1(p_{t-1}, p_t) + \frac{1}{\zeta}\mathbb{E}_{p_t(x)}|q_t(x) - q_{t-1}(x)|\right] \tag{B.75}$$

and

$$\mathcal{S}_{\text{stat}} := \sum_{t=1}^{T-1} s_t. \tag{B.76}$$

Sum these bounds for all $t \in \{0, \ldots, T-1\}$, we get:

$$\sum_{t=0}^{T-1} |\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1})| \leq \iota\zeta\mathcal{C} + \mathcal{S}_{\text{stat}}. \tag{B.77}$$

As the final classifier h_T may not be optimally trained with respect to p_0, include the approximation gap:

$$\varepsilon_{p_0}(h_0) + \varepsilon_{p_0}(h_T^*) - \varepsilon_{p_0}(h_0) \tag{B.78}$$

where h_T^* is the risk minimizer in $\mathcal{H}$ for p_0.

Finally,

$$\varepsilon_{p_T}(h_T) \leq \varepsilon_{p_0}(h_0) + \varepsilon_{p_0}(h_T^*) + \iota\zeta\mathcal{C} + \mathcal{S}_{\text{stat}},$$

as desired. □

C Detailed Algorithm of E-SUOT Framework

While Algorithm 1 outlines the general workflow for generating the intermediate domain, it does not specify how E-SUOT can be applied to the GDA task. To bridge this gap, we first present the complete workflow for E-SUOT-based GDA in Algorithm 2.

Before detailing this workflow, we emphasize that our focus is on the classification setting. Specifically, we denote the classifier's output as $\hat{y}$ and the ground-truth label as y. The loss function for our

classifier $h_{\omega,t}$, parameterized by ω at time t, is defined as follows. In our implementation, we adopt cross-entropy as the loss function:

$$\mathcal{L}_{\text{CE}}(x_t, h_{\omega,t}, y_t) = -\sum_{i=1}^{\mathcal{B}} y_t^{(i)} \log \hat{y}_t^{(i)} = -\sum_{i=1}^{\mathcal{B}} y_t^{(i)} \log h_{\omega,t}(x_t^{(i)}). \tag{C.1}$$

Building on this foundation, the complete workflow for E-SUOT-based gradual domain adaptation is summarized in Algorithm 2 based on Algorithm 1. Notably, our algorithm decouples the training of the transport function $\boldsymbol{T}_\theta$ from the fine-tuning of the classifier h_ω. This separation allows the intermediate domain to be generated offline and subsequently used for online inference, potentially reducing overall computation time comparable to traditional GDA approaches.

Algorithm 2 Overall Workflow for Construing E-SUOT-based Gradual Domain Adaption

Input: Source domain samples: $\{(x_0^{(i)}, y_0^{(i)})\}_{i=1}^{\text{N}}$, target domain samples: $\{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\text{N}}$, entropy regularization strength: ϵ, step size: η, number of intermediate domain $T-1$, neural network batch size $\mathcal{B}$, and neural network training epochs: $\mathcal{E}$.
Output: Classifier in target domain $h_{\omega,T}$.

1: Initialize the classifier $h_{\omega,0}$: $h_{\omega,0} \leftarrow \arg\min_\omega \mathcal{L}_{\text{CE}}(x_0, h_{\omega,t}, y_0)$.
2: Train $\mathcal{T} = \{\boldsymbol{T}_{\theta,t}\}_{t=1}^{T-1}$: $\mathcal{T} \leftarrow$Algorithm 1.
3: **for** $t = 0$ **to** $T-1$ **do**
4: Obtain the intermediate domain data $\{(x_{t+1}^{(i)}, y_{t+1}^{(i)})\}$: $x_{t+1}^{(i)} \leftarrow \boldsymbol{T}_{\theta,t}(x_t^{(i)})$ and $y_{t+1}^{(i)} \leftarrow y_t^{(i)}$ for all $i \in \{1, \ldots, \text{N}\}$.
5: Finetune the classifier $h_{\omega,t+1}$: $h_{\omega,t+1} \leftarrow \arg\min_\omega \mathcal{L}_{\text{CE}}(x_{t+1}, h_{\omega,t}, y_{t+1})$.
6: **end for**

D Detailed Information for Experiments

D.1 Dataset Descriptions

- **Portraits:** Portraits is a binary gender classification dataset comprising 37,921 front-facing portrait images collected between 1905 and 2013. Following the chronological split protocol of (Kumar et al., 2020), we divide the data into a source domain (the earliest 2,000 images), intermediate domains (14,000 images not utilized in this work), and a target domain (the subsequent 2,000 images), similar to the setting in reference (Zhuang et al., 2024).
- **Rotated MNIST:** Rotated MNIST is a variant of the standard MNIST dataset Deng (2012) in which images are rotated to create domain adaptation challenges. As described in He et al. (2024); Kumar et al. (2020), we use 4,000 source images and 4,000 target images, with the target images rotated by 45° to 60°.

D.2 Experimental Settings

The official implementations of GOAT (He et al., 2024) and CNF (Sagawa & Hino, 2025) are used in our experiments. Additionally, we employ UMAP (McInnes et al., 2018) to reduce the dimensionality of the three GDA datasets to 8. The experiments are conducted on a workstation equipped with two NVIDIA RTX 4090 GPUs under five different random seeds at least three times. The overall hyper-parameters we use in our GDA task are summarized in Table D.1.

Table D.1: Hyperparameters for E-SUOT on GDA task.

Datasets	η	$\mathcal{B}$	ϵ	T
Portraits	0.5	1024	0.1	5
MNIST 45°	0.5	1024	0.01	5
MNIST 60°	0.5	2048	0.005	5

In all experiments, we parameterize the classifier h_ϕ as a three-layer multi-layer perceptron (MLP) at each step, utilizing ReLU activation functions and a hidden dimension of 100 for each layer. For both $\boldsymbol{T}_\theta$ and w_ϕ, we employ a two-layer MLP with the SiLU activation function and incorporate a skip connection to enable a residual structure (He et al., 2016). All models are optimized by the Adam optimizer (Kingma & Ba, 2015) with learning rate at 0.0001. For all three GDA datasets, we apply UMAP (McInnes et al., 2018) to reduce their dimensionality to eight. We use the official implementation of the baseline models in our experiment.

D.3 Detailed Information for Ablation Studies

For the ablation study, we ablate two module namely the training strategy of $\boldsymbol{T}_\theta$ and the objective functional. The detailed information are elaborated in this part.

For "Training Strategy", the detailed experimental protocols are given as follows:

- **Adversarial Training:** In our adversarial training scheme, we optimize Eq. (7). Building on (Korotin et al., 2021; 2023; Choi et al., 2023; 2024), the training of $\boldsymbol{T}_\theta$ is formulated adversarially, as summarized in Algorithm 3. In *Line 5*, the penalty term $\frac{1}{2\eta}\|x_{t-1} - \boldsymbol{T}_{\theta,t-1}(x_{t-1})\|_2^2$ is omitted since it is constant with respect to $w_{\phi,t-1}$.
- **Barycentric-based Training:** We propose the algorithm for barycentric-based training in Algorithm 4. For barycentric-based training, rather than first compute the transport map, we attempt to compute the optimal transport map π^* between $\rho(x_{t-1})$ and $p_T(x)$ as we demonstrate in *Line 4*. Based on this, we make barycentric projection (Courty et al., 2017b; Perrot et al., 2016) using this π^* to obtain the proxy points (Liu et al., 2021; 2023) for transport map learning as we demonstrate in *Line 5*. Finally, the transport map $\boldsymbol{T}_{\theta,t-1}$ is constructed based on these points, similar to the flow matching (Lipman et al., 2023), as we demonstrate in *Line 6*.

Algorithm 3 Adversarial Training for $\{\boldsymbol{T}_{\theta,t}\}_{t=1}^{T-1}$.

Input: Intermediate domain samples: $\{(x_{t-1}^{(i)}, y_{t-1}^{(i)})\}_{i=1}^{\mathrm{N}}$ for all $t \in \{1, \dots, \mathrm{T}\}$, target domain samples: $\{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\mathrm{N}}$, entropy regularization strength: ϵ, step size: η, neural network batch size $\mathcal{B}$, and neural network training epochs: $\mathcal{E}$.
Output: The transportation map at $t-1$: $\boldsymbol{T}_{\theta,t-1}$.

1: Initialize
2: **for** $e = 1$ **to** $\mathcal{E}$ **do**
3: Sample a batch $\{x_{t-1}^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_{t-1}^{(i)}, y_{t-1}^{(i)})\}_{i=1}^{\mathrm{N}}$ and $\{x_T^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\mathrm{N}}$.
4: Update $w_{\phi,t-1}$ by: $\phi \leftarrow \arg\min_\phi \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} -\frac{1}{2\eta}\|x_{t-1} - \boldsymbol{T}_{\theta,t-1}(x_{t-1})\|_2^2 + w_{\phi,t-1}(\boldsymbol{T}_\theta(x_t^{(i)})) + \frac{1}{\mathcal{B}}\sum_{j=1}^{\mathcal{B}} f^\star(-w_{\phi,t-1}(x_T^{(j)}))$.
5: Sample a batch $\{x_{t-1}^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_{t-1}^{(i)}, y_{t-1}^{(i)})\}_{i=1}^{\mathrm{N}}$.
6: Update $\boldsymbol{T}_{\theta,t-1}$ by: $\theta \leftarrow \arg\min_\theta \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \frac{1}{2\eta}\|x_{t-1}^{(i)} - \boldsymbol{T}_{\theta,t-1}(x_{t-1}^{(i)})\|_2^2 - w_{\phi,t-1}(\boldsymbol{T}_{\theta,t-1}(x_{t-1}^{(i)}))$.
7: **end for**

Algorithm 4 Barycentric-based training for $\{\boldsymbol{T}_{\theta,t}\}_{t=1}^{T-1}$.

Input: Intermediate domain samples: $\{(x_{t-1}^{(i)}, y_{t-1}^{(i)})\}_{i=1}^{\mathrm{N}}$ for all $t \in \{1, \dots, \mathrm{T}\}$, target domain samples: $\{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\mathrm{N}}$, entropy regularization strength: ϵ, step size: η, neural network batch size $\mathcal{B}$, and neural network training epochs: $\mathcal{E}$.
Output: The transportation map at $t-1$: $\boldsymbol{T}_{\theta,t-1}$.

1: Initialize
2: **for** $e = 1$ **to** $\mathcal{E}$ **do**
3: Sample a batch $\{x_{t-1}^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_{t-1}^{(i)}, y_{t-1}^{(i)})\}_{i=1}^{\mathrm{N}}$ and $\{x_T^{(i)}\}_{i=1}^{\mathcal{B}} \sim \{(x_T^{(i)}, y_T^{(i)})\}_{i=1}^{\mathrm{N}}$.
4: Obtain the optimal transport map $\pi^*(x_{t-1}, x_T)$ by: $\pi^*(x_{t-1}, x_T) \leftarrow \inf_\pi \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x_{t-1}), p_T(x)) + \epsilon\iint \pi(x_{t-1}, x_T)[\log \pi(x_{t-1}, x_T) - 1]\mathrm{d}x_{t-1}\mathrm{d}x_T + \mathbb{D}_f[\rho(x_{t-1}), p_T(x)]$.
5: Obtain the projected samples $\tilde{x}_t$ via $\pi^*(x_{t-1}, x_T)$: $\tilde{x}_t = x_{t-1}\pi^*(x_{t-1}, x_T)$:
6: Update $\boldsymbol{T}_{\theta,t-1}$ by: $\theta \leftarrow \frac{1}{\mathcal{B}}\sum_{i=1}^{\mathcal{B}} \|\tilde{x}_t^{(i)} - \boldsymbol{T}_{\theta,t-1}(x_{t-1}^{(i)})\|_2^2$
7: **end for**

For "Objective Functional", the detailed experimental protocols are given as follows:

- χ^2 **Divergence:** The expression for χ^2 divergence can be given as follows:

$$\mathbb{D}_{\chi^2}[\rho(x_t), p_T(x)] \int p_T(x)[\frac{\rho(x_t)}{p_T(x)} - 1]^2 \mathrm{d}x_t, \quad \text{where} \quad f(x) = (x-1)^2. \tag{D.1}$$

  Based on this, the corresponding conjugate function $f^\star$ can be given as follows:

$$f^\star(x) = \begin{cases} \frac{1}{4}x^2 + x, \text{if} x \geq -2 \\ -1, \text{if} x < -2 \end{cases}. \tag{D.2}$$

- **Identity:** For the identity function, we remove the f-divergence-based regularization term during the construction of E-SUOT framework. Based on this, the training objective for w_ϕ is reformulated as follows:

$$\mathcal{L}^{\text{E-SemiDual}}_{\text{Identity}} = \sup_w -\epsilon \mathbb{E}_{p(x_t)}\{\log \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - \frac{1}{2\eta}\|x - x_t\|_2^2}{\epsilon})]\} + \mathbb{E}_{p_T(x)}[w(x)], \tag{D.3}$$

- **`Softplus`:** We directly parameterize the $f^\star$ using the smooth, convex, and non-decreasing softplus function as follows:

$$f^\star = \log(1 + \exp(x)). \tag{D.4}$$

E Limitations & Future Directions and Broader Impact

E.1 Limitations & Future Directions

The limitations and future research directions of this work can be summarized as follows:

- **Consideration of Label Information:** In this work, we focused primarily on feature adaptation and did not explicitly incorporate label or discriminator information into the adaption process. As a result, the performance of the proposed E-SUOT framework may degrade under scenarios involving significant covariate shift (Sugiyama et al., 2007; Sugiyama & Kawanabe, 2012). An important direction for future research is to integrate label information into the transportation process, for example, classifier guidance approach (Courty et al., 2017a; Dhariwal & Nichol, 2021; Bonet et al., 2025; Zhuang et al., 2024), which could further enhance model robustness and adaptation performance.
- **Regularization for Transport Plan:** To facilitate computation, we introduced entropy regularization on the transport plan; however, this may introduce potential instability or blur sparsity in the map (Yin et al., 2025). Future work may explore alternative regularization strategies (Courty et al., 2014; 2017b), such as group sparsity (to better incorporate label priors) or Laplacian regularization (to preserve local relationships), in order to further stabilize training and improve the properties of the learned potential function w.
- **Exploration of Other Discrepancy:** In this work, we adopted the Wasserstein distance as the primary metric for measuring domain discrepancy. However, other discrepancy measures, such as the Fisher-Rao distance (Zhang et al., 2022; Wang et al., 2023; Zhu, 2025), could also be explored to enable more flexible or principled adaptation approaches. Future work may investigate the use of alternative metrics (Neklyudov et al., 2023; Skreta et al., 2025) to further improve the effectiveness of the quality of intermediate domain thereby improving the performance of GDA task.

E.2 Broader Impact Statement

GDA addresses a critical challenge in machine learning: transferring knowledge from a labeled source domain to an unlabeled target domain when there is a substantial gap between the two. Rather than relying on abrupt, one-shot shifts—which are often brittle in the face of large distributional discrepancies—GDA interpolates through a series of intermediate domains, allowing for a smoother and more effective adaptation process. This paradigm has direct implications for many real-world applications. For example, in recommender systems, GDA enables knowledge transfer to

serve cold-start users or to integrate new items, and in language processing it allows models trained on high-resource languages to adapt more robustly to low-resource languages. By constructing and navigating intermediate distributions, GDA provides a principled foundation for bridging domain gaps and ensuring stable model performance under challenging conditions. Our work advances the field of GDA by unifying flow-based methods and optimal transport within the semi-dual formulation, identifying fundamental issues of stability and generalization that have limited previous approaches. We further propose theoretically-grounded regularization strategies that improve the robustness and reliability of the adaptation process. These advances not only deepen the theoretical understanding of GDA but also offer practical benefits for deploying adaptable machine learning systems in diverse settings. We believe our findings will help catalyze the development of more general, stable, and information-preserving domain adaptation methods, with impact across fields ranging from recommendation and computational linguistics to broader AI applications.

F LLM USAGE STATEMENT

In accordance with the conference guidelines, we disclose our use of Large Language Models (LLMs) in the preparation of this paper as follows:

We used LLMs (specifically, OpenAI GPT-4.1, GPT-5 and Google Gemini 2.5) *solely for checking grammar errors and improving the readability of the manuscript*. The LLMs *were not involved in research ideation, the development of research contributions, experiment design, data analysis, or interpretation of results*. All substantive content and scientific claims were created entirely by the authors. The authors have reviewed all LLM-assisted text to ensure accuracy and originality, and take full responsibility for the contents of the paper. The LLMs are not listed as an author.