Rethinking the Flow-Based Gradual Domain Adaption: A Semi-Dual Optimal Transport Approach

Anonymous Authors[1]

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) (Courty et al., 2014; 2017a; Long et al., 2015; Pan & Yang, 2010; Tzeng et al., 2017), 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; Zeng et al., 2025; Zhuang et al., 2024), we argue that directly applying standard flow-based models leads to suboptimal performance. Specifically,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

For ICML 2026 reviewers: When reviewing this paper, please follow the policy for the use of LLMs as specified on your Reviewer Console. Note that the assigned LLM policy might differ from the policy you indicated during reviewer registration.

existing flow-based frameworks utilizing f-divergence often require the explicit estimation of target domain probability density functions (PDFs) from available target samples [1] (Ambrosio et al., 2005; Santambrogio, 2017; Vincent, 2011), whereas the subsequent GDA process relies on these estimated PDFs to drive the source-to-target transfer. For example, Zhuang et al. (2024) estimate the unnormalized target domain PDF in the score function form and generate intermediate domains via Langevin dynamics. 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:

1. We develop a semi-dual formulation for intermediate domain generation in flow-based GDA, which eliminates the need for explicit estimation of the target-domain PDF.
2. 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.
3. We conducted various experiments to demonstrate the superiority of the proposed E-SUOT approach compared to prevalent approaches.

2. Preliminaries

Preliminary Note. In this manuscript, we focus on analyzing the flow-based GDA *per se*, rather than establishing that GDA is superior to UDA. We inherit the standard assumptions commonly adopted in GDA and do not discuss when these assumptions hold in practice. A related unbalanced optimal transport (UOT) formulation will be derived as a direct consequence of standard definitions in the GDA setting; we use it as an analytical characterization, not as an additional assumption, and we do not study the inherent superiority of UOT compared with the vanilla optimal transport (OT) formulation.

[1] In this manuscript, we treat both logarithm PDF and its gradient, also known as score function, as forms of density estimation

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 $h : \mathcal{X} \to \mathcal{Y}$. The GDA task assumes 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 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:

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

where $p(x_t)$ is the 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. Notably, $\frac{\delta}{\delta p}$ denotes the first variation with-respect-to p, and the second equality sign is called "gradient flow". 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 Equation (1).

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 (Glaser

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

et al., 2021). For an f-divergence,

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

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,

$$v_t(x_t) = \nabla \log p_T(x) - \nabla \log p(x_t), \quad (3)$$

and, in the weak solution to the partial differential equations (Evans, 2022; Liu, 2017), the induced dynamics yield the classical *Langevin dynamic* (Santambrogio, 2017).

Intuitively, applying the forward Euler scheme with step size η to the gradient flow in Equation (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 follows, and the derivation is provided in Section C.1:

$$p(x_{t+\eta}) = \operatorname*{arg\,min}_{\rho \in \mathcal{P}_2(\mathbb{R}^D)} \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)], \quad (4)$$

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, $\mathcal{W}_2^2$ is the squared 2-Wasserstein distance, whose definition is given as follows:

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

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

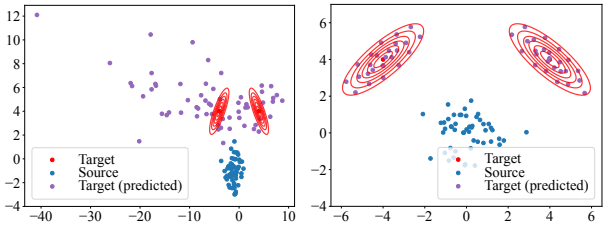
3. Methodology

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 (normalized/unnormalized) PDF. Given a reliable estimate, one can construct a velocity field that progressively pushes source samples toward the target distribution.

However, estimating the target PDF from samples alone is ill-posed (Song et al., 2020; Vincent et al., 2010). When the estimate is inaccurate, the resulting flow may drive samples into low-density regions, producing a noticeable mismatch between generated and true target samples and degrading downstream performance. To highlight this issue, we compare two transport strategies: (1) EstTrans, which first estimates the target PDF via denoised score matching (Vincent, 2011) and then applies gradient-flow transport, and (2) DirTrans, which directly transports samples using a learned optimal transport map. The qualitative results and the Wasserstein distance to ground-truth target samples are reported in Figures 1(a) and 1(b).

As shown, inaccurate PDF estimates lead EstTrans to produce samples that deviate from the target distribution and incur a larger Wasserstein distance, whereas DirTrans better recovers target samples and achieves a smaller distance. These observations motivate the following questions: *How can we generate intermediate domains without estimating the PDF of target domain? How can robust intermediate domain generation be achieved within this framework? Does this approach improve the performance for GDA task?*



(a) EstTrans, $\mathcal{W}_2^2 \approx 9.7\text{E}0$. (b) DirTrans, $\mathcal{W}_2^2 \approx 7.8\text{E} - 4$.

3.2. Dual-Form Optimal Transport for GDA Task

As shown in Equation (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 Equation (4):

Proposition 3.1. *Consider the following primal problem:*

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

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

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

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

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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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 (Choi et al., 2023; 2024; Korotin et al., 2023), 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 for Semi-Dual Formulation

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 Equation (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 3.2. *The semi-dual formulation in Equation (7) admits non-unique optimal solutions.*

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

Proposition 3.3. *Let $\kappa(x_t, x) := p(x_t)\, p_T(x)$ denote the reference joint PDF. The entropy-regularized primal problem can be formulated as follows:*

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

and is equivalent to the semi-dual optimization problem

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

where w and $f^\star$ are as defined in Proposition 3.1.

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

Proposition 3.4. *The semi-dual formulation in Equation (9) admits a unique optimal solution.*

Notably, as seen in Equation (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 Equation (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

Algorithm 1 Overall Workflow for E-SUOT

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 optimal transportation map: $\mathcal{T} = \{\boldsymbol{T}_{\theta,t}\}_{t=0}^{T-1}$.

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

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 E) and the corresponding illustration is given in Figure 1. As shown in the algorithm, the construction of w_ϕ and $\boldsymbol{T}_\theta$ are performed as separate steps, corresponding to Figure 1(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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

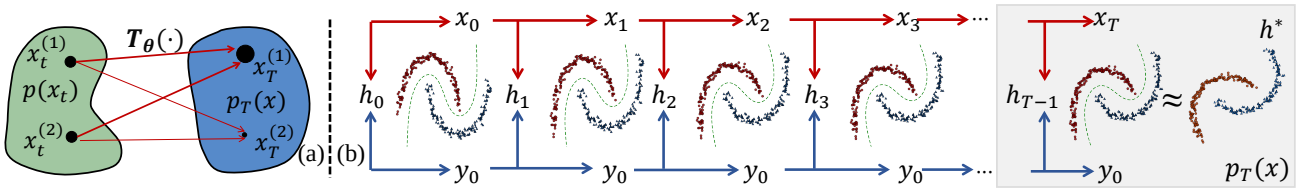


Figure 1. The illustration of the proposed E-SUOT: (a) the unbalanced OT formulation used to solve the transport map $\boldsymbol{T}_\theta(\cdot)$ at time t, where thicker arrows and larger points indicate higher mass flows, and (b) the evolution process from the source to the target domain. This figure is conceptually inspired by previous works on GDA and OT tasks (He et al., 2024; Wang et al., 2025; Zhuang et al., 2024).

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 Figure 1(b).

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.

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 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 qualitatively characterizes the discrepancy between $\rho(x)$ and $p_T(x)$:

Theorem 3.5. *The optimal solution $\rho^*(x)$ to problem defined in Equation (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 3.5, we observe that as t increases, the transported density $\rho(x)$ progressively approaches the target distribution $p_T(x)$. A detailed analysis for this statement is provided in Section C.6, , and additional discussions on the case where $\mathbb{D}_f$ is KL divergence are given in Section C.8.

Finally, we present the following theorem to demonstrate the model's performance on the target domain:

Theorem 3.6. *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 gradual domain adaption path, and $\mathcal{S}_{stat}$ is the statistical error term.

4. Main Experimental Results

4.1. Experimental Setup

We conduct case studies on four datasets. Specifically, for the GDA task, we use the Portraits dataset (Kumar et al., 2020) as well as MNIST 45° and MNIST 60° (LeCun, 1998). Following prior work (Sagawa & Hino, 2025; Zhuang et al., 2024), to ensure a fair comparison, we use the UMAP embeddings (McInnes et al., 2018) provided by Zhuang et al. (2024). In addition, to demonstrate the scalability of the proposed E-SUOT, we conduct experiments on the Office-Home dataset (Venkateswara et al., 2017). Other details about these datasets are provided in Section E.1. Due to space limitations, additional experimental results are provided in Section F.

4.2. Baseline Comparison Results

Table 1. Baseline comparison on the GDA task.

Method	Portraits		MNIST 45°		MNIST 60°	
	Acc. (%)	Δ	Acc. (%)	Δ	Acc. (%)	Δ
Source	71.2	-	58.4	-	36.8	-
Self Train	77.4	↑8.7%	58.7	↑0.5%	39.9	↑8.5%
GST (4)	76.1	↑6.9%	59.2	↑1.3%	39.9	↑8.5%
GOAT	74.9	↑5.3%	65.0	↑11.3%	37.2	↑1.1%
CNF	80.0	↑12.4%	57.6	↓1.4%	41.8	↑13.5%
GGF	83.4	↑17.2%	57.7	↓1.2%	40.8	↑11.0%
E-SUOT	**86.4**	↑21.5%	**72.1**	↑23.4%	**51.0**	↑38.6%

Kindly Note: **Bolded** results are the best results. Underlined results are the second best results.

GDA Task. We first compare our proposed approach with several existing GDA-based methods, including Self-training (Wei et al., 2021), GST (with 4 intermediate domains) (Kumar et al., 2020), GOAT (He et al., 2024), CNF (Sagawa & Hino, 2025), and GGF (Zhuang et al., 2024). The experimental results are listed in Table 1.

UDA Task. As shown in Table 1, our proposed E-SUOT framework consistently outperforms the current state-of-the-

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 2. UDA Acc. (%) comparison on the Office-Home dataset.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
DANN	45.6	59.3	70.1	47.0	58.5	60.9	46.1	43.7	68.5	63.2	51.8	76.8	57.6
MSTN	49.8	70.3	76.3	60.4	68.5	69.6	61.4	48.9	75.7	70.9	55.0	81.1	65.7
GVB-GD	57.0	74.7	79.8	64.6	74.1	74.6	65.2	55.1	81.0	74.6	59.7	84.3	70.4
RSDA	53.2	77.7	81.3	66.4	74.0	76.5	<u>67.9</u>	53.0	82.0	75.8	57.8	85.4	70.9
LAMDA	57.2	78.4	<u>82.6</u>	66.1	**80.2**	**81.2**	65.6	55.1	82.8	71.6	59.2	83.9	72.0
SENTRY	**61.8**	77.4	80.1	66.3	71.6	74.7	66.8	**63.0**	80.9	74.0	**66.3**	84.1	72.3
FixBi	58.1	77.3	80.4	67.7	79.5	<u>78.1</u>	65.8	57.9	81.7	<u>76.4</u>	62.9	**86.7**	72.7
CST	59.0	**79.6**	**83.4**	**68.4**	77.1	76.7	**68.9**	56.4	**83.0**	75.3	62.2	85.1	72.9
CoVi	58.5	78.1	80.0	<u>68.1</u>	<u>80.0</u>	77.0	66.4	60.2	82.1	**76.6**	<u>63.6</u>	<u>86.5</u>	<u>73.1</u>
GGF	59.4	75.6	81.7	67.6	77.6	78.0	67.4	61.0	82.7	75.9	62.4	85.4	72.9
E-SUOT	<u>61.6</u>	<u>79.3</u>	81.8	67.6	77.7	78.1	67.4	<u>61.2</u>	<u>82.9</u>	76.3	62.5	85.2	**73.5**
Win Counts	9	9	8	6	7	8	7	9	9	8	7	6	10

Kindly Note: **Bolded** and <u>underlined</u> results are the first and second best results, respectively.

art GDA approaches 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.

On this basis, we further evaluate the performance of E-SUOT under the Office-Home dataset under UDA setting to demonstrate the scalability of E-SUOT approach by comparing it to the DANN (Ganin & Lempitsky, 2015), MSTN (Xie et al., 2018), GVB-GD (Cui et al., 2020), RSDA (Gu et al., 2020; 2022), LAMBDA (Le et al., 2021), SENTRY (Prabhu et al., 2021), FixBi (Na et al., 2021), CST (Liu et al., 2021a), CoVi (Na et al., 2022), and GGF (Zhuang et al., 2024) on the Office-Home dataset (Venkateswara et al., 2017). In our UDA experiment, we use the CoVi algorithm to pretrain our backbone, and other detailed information on the experiments is provided in the Section E.2. The results are summarized in Table 2.

Table 2 shows that E-SUOT outperforms most existing methods on the majority of UDA tasks and achieves the highest overall average performance. Despite not achieving top performance on every individual task, E-SUOT attains the highest average result across the board, confirming its overall superiority in UDA task.

4.3. Ablation Studies

Ablation on T_θ Training. To ensure fair comparisons under the same feature representations (UMAP embeddings from Zhuang et al. (2024)), we conduct the ablation study in this part on the three datasets in the GDA task only. In this part, 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 Equation (7) to backup Theorem 3.4, 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) `SftPls` (softplus function), and also compare the 2) χ^2 divergence and the 3) identity function. More detailed information on these experiments' implementation is provided in Appendix E.3. The ablation study results are summarized in Table 3.

Table 3. Ablation study results on GDA setting.

Dataset			Portraits		MNIST 45°		MNIST 60°	
Metric			Acc. (%)	Δ	Acc. (%)	Δ	Acc. (%)	Δ
Training	Adversarial	KL	74.8	↓13.4%	52.0	↓27.8%	34.9	↓31.5%
	Barycentric	KL	83.9	↓3.0%	62.5	↓13.3%	38.3	↓24.8%
Functional	Entropy	`SftPls`	80.1	↓7.3%	59.7	↓17.2%	38.2	↓25.1%
	Entropy	χ^2	79.8	↓7.7%	60.2	↓16.5%	42.4	↓16.9%
	Entropy	Identity	81.2	↓6.1%	59.6	↓17.4%	39.6	↓22.3%
	Entropy	KL	86.4	-	72.1	-	51.0	-

Kindly Note: Δ denotes performance change percentage compared to E-SUOT with entropy regularization and KL divergence.

From Table 3, we find that adversarial training performs the worst, underscoring the importance of entropy regularization for robust 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 `SftPls`, cause

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Table 4. Accuracy (%) improvement over different UDA feature extractors.

Method	Ar→Cl	Ar→Pr	Ar→Rw	Cl→Ar	Cl→Pr	Cl→Rw	Pr→Ar	Pr→Cl	Pr→Rw	Rw→Ar	Rw→Cl	Rw→Pr	Avg.
MSTN	49.8	70.3	76.3	60.4	68.5	69.6	61.4	48.9	75.7	70.9	55.0	81.1	65.7
E-SUOT+MSTN	57.8	75.9	79.6	65.5	75.9	74.8	64.5	58.5	81.4	73.7	59.5	84.4	71.0
Δ	↑16.1%	↑8.0%	↑4.3%	↑8.4%	↑10.8%	↑7.5%	↑5.0%	↑19.6%	↑7.5%	↑3.9%	↑8.2%	↑4.1%	↑8.1%
RSDA	53.2	77.7	81.3	66.4	74.0	76.5	67.9	53.0	82.0	75.8	57.8	85.4	70.9
E-SUOT+RSDA	61.5	78.8	81.7	67.6	77.3	77.6	67.2	61.0	82.7	76.0	62.4	85.3	73.3
Δ	↑15.7%	↑1.4%	↑0.5%	↑1.8%	↑4.4%	↑1.5%	↓1.0%	↑15.2%	↑0.9%	↑0.3%	↑7.9%	↓0.1%	↑3.3%
FixBi	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7
E-SUOT+FixBi	61.7	79.1	81.7	67.6	77.6	78.2	67.3	61.3	82.7	76.0	62.5	85.3	73.4
Δ	↑6.2%	↑2.3%	↑1.6%	↓0.1%	↓2.4%	↑0.1%	↑2.3%	↑5.9%	↑1.2%	↓0.5%	↓0.6%	↓1.6%	↑1.0%
CoVi	58.5	78.1	80.0	68.1	80.0	77.0	66.4	60.2	82.1	76.6	63.6	86.5	73.1
E-SUOT+CoVi	61.6	79.3	81.8	67.6	77.7	78.1	67.4	61.2	82.9	76.3	62.5	85.2	73.5
Δ	↑5.3%	↑1.5%	↑2.2%	↓0.7%	↓2.9%	↑1.4%	↑1.5%	↑1.7%	↑1.0%	↓0.4%	↓1.7%	↓1.5%	↑0.5%

Kindly Note: Δ denotes performance change percentage of the vanilla UDA feature extractor compared to E-SUOT.

significant performance drops, emphasizing the importance of proper divergence selection for conducting GDA task. 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.

Ablation on UMAP Embedding. Since our E-SUOT uses UMAP embeddings obtained from backbones pretrained with different UDA algorithms. In this part, we further conduct ablations on the Office-Home dataset from two aspects: (1) the pretraining strategy and (2) the UMAP embedding dimensionality.

The ablation study on the pretraining strategy is reported in Table 2. We evaluate the proposed E-SUOT by integrating it with four representative UDA algorithms, namely MSTN, RSDA, FixBi, and CoVi, on the Office-Home dataset. For a fair comparison, we apply UMAP to obtain 8-dimensional embeddings. As shown in Table 4, E-SUOT consistently improves all baselines, with average gains of 8.1%, 3.3%, 1.0%, and 0.5% over MSTN, RSDA, FixBi, and CoVi, respectively. The largest improvement ($+16.1\%$) occurs on the challenging Ar→Cl transfer, suggesting that E-SUOT is particularly beneficial under large domain shifts. Notably, E-SUOT remains effective when combined with strong recent methods (e.g., FixBi and CoVi), indicating that it serves as a complementary module. Overall, these consistent gains across different pretrained feature extractors demonstrate the scalability and robustness of E-SUOT for UDA.

We further study the ablation of E-SUOT to the UMAP embedding dimension by varying the reduced dimension from 4 to 256, where 256 corresponds to using the original 256-dimensional backbone features (i.e., without UMAP).

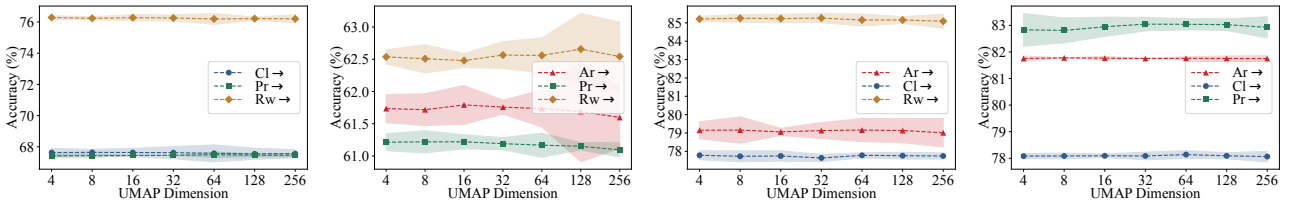
From Figure 2, we observe that as the embedding dimension changes, the classification accuracy tends to fluctuate within a relatively small range across all target domains. In detail, for each target domain, the performance remains quite stable as we vary the UMAP dimension from 4 to 256, it can be observed that no drastic drops are observed. In some domains (e.g., Ar in Figure 2(a), Pr in Figure 2(c) and Rw in Figure 2(d)), the accuracy curves are almost flat, indicating the proposed method is insensitive to UMAP dimension choices in these cases. For domain Cl in Figure 2(b), although the standard deviation is higher, the main trend is still relatively stable. It suggests that E-SUOT retains strong robustness to the UMAP embedding dimension and does not rely heavily on fine-tuning this hyperparameter. In summary, the ablation analysis shows that E-SUOT remains robust to variations in the UMAP embedding dimension.

5. Related Works

5.1. Gradual Domain Adaptation

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 (Dong et al., 2022; Kumar et al., 2020; 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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.



(a) Target Domain: Ar. (b) Target Domain: Cl. (c) Target Domain: Pr. (d) Target Domain: Rw.

Figure 2. Ablation study of E-SUOT performance vary the UMAP embedding dimension on the Office-Home dataset. For dimension 256, the vanilla backbone features are used without UMAP. The shaded area indicates the $\pm$ 5.0 times standarad deviation error.

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 3.1), 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 3.3) and analyze its generalization error (see Theorem 3.6). 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, 2017; Liu & Wang, 2016), 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 (Choi et al., 2023; 2024; Fan et al., 2022; Gazdieva et al., 2023; Korotin et al., 2023; Rout et al., 2022). These dual-formulation methods, however, generally require adversarial optimization characterized by a composite "sup-inf" structure in order to properly approximate the dual objective when implemented with neural networks (Arjovsky et al., 2017; Nowozin et al., 2016). In addition, recent works attempt to introduce entropy regularization into UOT-based generative modeling. For example, Choi & Choi (2024) addresses the resulting inner entropy-regularized problem via the Schrödinger bridge formulation (Chen et al., 2021). Our work differs from these approaches in three key aspects. First, we provide a theoretical analysis from the perspective of the non-uniqueness of optimal solutions in Proposition 3.2, 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 3.3, and further prove that this regularization ensures the stability via the uniqueness of the optima in Proposition 3.4 and convergence in Theorem 3.5. Third, our inner sup problem is derived from a Sinkhorn-iteration perspective, rather than from the Schrödinger bridge formulation.

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.

Limitations. Our current E-SUOT mainly focuses on marginal feature alignment and does not explicitly incorporate label/discriminator guidance, which may limit robustness under strong covariate or label shift. In addition, the entropic regularization used for computational efficiency can blur sparsity in the transport plan and may affect stability. Finally, the method relies on a multi-network, multi-stage offline training pipeline, leaving room for more parameter-efficient training designs. Detailed discussions and alleviation strategies are provided in Section G.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abnar, S., Berg, R. v. d., Ghiasi, G., Dehghani, M., Kalchbrenner, N., and Sedghi, H. Gradual domain adaptation in the wild: When intermediate distributions are absent. *arXiv preprint arXiv:2106.06080*, pp. 1–15, 2021.

Abraham, R., Marsden, J. E., and Ratiu, T. *Manifolds, tensor analysis, and applications*, volume 75. Springer Science & Business Media, 2012.

Amari, S.-i. *Information geometry and its applications*, volume 194. Springer, 2016.

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proc. Int. Conf. Mach. Learn.*, pp. 214–223. PMLR, 2017.

Bach, F. *Learning theory from first principles*. MIT Press, Cambridge, USA, 2024.

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

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

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

Brunzema, P., Jordahn, M., Willes, J., Trimpe, S., Snoek, J., and Harrison, J. Bayesian optimization via continual variational last layer training. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–30, 2025.

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

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

Caluya, K. F. and Halder, A. Wasserstein proximal algorithms for the Schrödinger bridge problem: Density control with nonlinear drift. *IEEE Trans. Autom. Control.*, 67 (3):1163–1178, 2022. doi: 10.1109/TAC.2021.3060704.

Chen, H.-Y. and Chao, W.-L. Gradual domain adaptation without indexed intermediate domains. In *Proc. Adv. Neural inf. Process. Syst.*, volume 34, pp. 8201–8214, 2021.

Chen, Y., Georgiou, T. T., and Pavon, M. Stochastic control liaisons: Richard sinkhorn meets gaspard monge on a schrodinger bridge. *SIAM Rev.*, 63(2):249–313, 2021.

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Unbalanced optimal transport: Dynamic and kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

Choi, J. and Choi, J. Scalable simulation-free entropic unbalanced optimal transport. *arXiv preprint arXiv:2410.02656*, 2024.

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

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

Choi, J., Choi, J., and Kwon, D. 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.

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

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

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

Cui, S., Wang, S., Zhuo, J., Su, C., Huang, Q., and Tian, Q. Gradually vanishing bridge for adversarial domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pp. 12455–12464, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Proc. Adv. Neural inf. Process. Syst.*, volume 26, pp. 1–9. Curran Associates, Inc., 2013.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Deng, L. 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.

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

Dong, J., Zhou, S., Wang, B., and Zhao, H. Algorithms and theory for supervised gradual domain adaptation. *Trans. Mach. Learn. Res.*, pp. 1–14, 2022.

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

Fan, J., Zhang, Q., Taghvaei, A., and Chen, Y. Variational Wasserstein gradient flow. In *Proc. Int. Conf. Mach. Learn.*, pp. 6185–6215. PMLR, 2022.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–10. PMLR, 2015.

Gazdieva, M., Korotin, A., Selikhanovych, D., and Burnaev, E. Extremal domain translation with neural optimal transport. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pp. 40381–40413, 2023.

Glaser, P., Arbel, M., and Gretton, A. 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.

Gu, X., Sun, J., and Xu, Z. Spherical space domain adaptation with robust pseudo-label loss. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020.

Gu, X., Sun, J., and Xu, Z. Unsupervised and semi-supervised robust spherical space domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, 2022. doi: 10.1109/TPAMI.2022.3158637.

Harrison, J., Willes, J., and Snoek, J. Variational Bayesian last layers. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–30, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 770–778, 2016.

He, Y., Wang, H., Li, B., and Zhao, H. Gradual domain adaptation: Theory and algorithms. *J. Mach. Learn. Res.*, 25(361):1–40, 2024.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. LoRA: Low-rank adaptation of large language models. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–13, 2022.

Johnson, R. and Zhang, T. Composite functional gradient learning of generative adversarial models. In *Proc. Int. Conf. Mach. Learn.*, pp. 2371–2379. PMLR, 2018.

Johnson, R. and Zhang, T. A framework of composite functional gradient methods for generative adversarial models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1): 17–32, 2021. doi: 10.1109/TPAMI.2019.2924428.

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

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

Kobyzev, I., Prince, S. J., and Brubaker, M. A. Normalizing flows: An introduction and review of current methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11):3964–3979, 2020.

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

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–34, 2023.

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

Le, T., Nguyen, T., Ho, N., Bui, H., and Phung, D. Lamda: Label matching deep domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pp. 6043–6054. PMLR, 2021.

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

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–28, 2023.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., and Zhu, J. Understanding and accelerating particle-based variational inference. In *Proc. Int. Conf. Mach. Learn.*, pp. 4082–4092. PMLR, 2019.

Liu, H., Wang, J., and Long, M. Cycle self-training for domain adaptation. *Proc. Adv. Neural Inf. Process. Syst.*, 34:22968–22981, 2021a.

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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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

Liu, W., Su, J., Chen, C., and Zheng, X. Leveraging distribution alignment via stein path for cross-domain cold-start recommendation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pp. 19223–19234, 2021b.

Liu, W., Zheng, X., Su, J., Zheng, L., Chen, C., and Hu, M. 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.

Long, M., Cao, Y., Wang, J., and Jordan, M. Learning transferable features with deep adaptation networks. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–9. PMLR, 2015.

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

Na, J., Jung, H., Chang, H. J., and Hwang, W. FixBi: Bridging domain spaces for unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1094–1103, 2021.

Na, J., Han, D., Chang, H. J., and Hwang, W. Contrastive vicinal space for unsupervised domain adaptation. In *Proc. Eur. Conf. Comput. Vis.*, pp. 92–110. Springer, 2022.

Neklyudov, K., Nys, J., Thiede, L., Carrasquilla, J., Liu, Q., Welling, M., and Makhzani, A. 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.

Nguyen, X., Wainwright, M. J., and Jordan, M. 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.

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

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

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

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57):1–64, 2021.

Park, M. S., Kim, C., Son, H., and Hwang, H. J. The deep minimizing movement scheme. *J. Comput. Phys.*, 494: 112518, 2023.

Perrot, M., Courty, N., Flamary, R., and Habrard, A. Mapping estimation for discrete optimal transport. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 29, pp. 1–9, 2016.

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

Prabhu, V., Khare, S., Kartik, D., and Hoffman, J. SENTRY: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 8558–8567, 2021.

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

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

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

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

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

Shi, J., Liu, C., and Mackey, L. Sampling with mirrored Stein operators. *Proc. Int. Conf. Learn. Represent.*, pp. 1–26, 2022.

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

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

Skreta, M., Akhound-Sadegh, T., Ohanesian, V., Bondesan, R., Aspuru-Guzik, A., Doucet, A., Brekelmans, R., Tong, A., and Neklyudov, K. Feynman-Kac correctors in diffusion: Annealing, guidance, and product of experts. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–44. PMLR, 2025.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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

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

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

Thomas, G. B., Weir, M. D., Hass, J., Heil, C., and Behn, A. *Thomas' calculus: early transcendentals*, volume 15. Pearson Boston, London, U.K., 2014.

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

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 2962–2971, 2017. doi: 10.1109/CVPR.2017.316.

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

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 5018–5027, 2017.

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

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

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., and Manzagol, P.-A. 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.

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

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

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

Wang, H., Chen, Z., Wang, H., Tan, Y., Pan, L., Liu, T., Chen, X., Li, H., and Lin, Z. Unbiased recommender learning from implicit feedback via weakly supervised learning. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–21, 2025.

Wei, C., Shen, K., Chen, Y., and Ma, T. Theoretical analysis of self-training with deep networks on unlabeled data. In *Proc. Int. Conf. Learn. Represent.*, pp. 1–30, 2021.

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

Xie, S., Zheng, Z., Chen, L., and Chen, C. Learning semantic representations for unsupervised domain adaptation. In *Proc. Int. Conf. Mach. Learn.*, pp. 1–10. PMLR, 2018.

Yin, H., Qiu, Y., and Wang, X. Wasserstein coreset via sinkhorn loss. *Transactions on Machine Learning Research*, pp. 1–40, 2025.

Zeng, Z., Qiu, R., Bao, W., Wei, T., Lin, X., Yan, Y., Abdelzaher, T. F., Han, J., and Tong, H. Pave your own path: Graph gradual domain adaptation on fused gromov-wasserstein geodesics. *arXiv preprint arXiv:2505.12709*, 2025.

Zhang, C., Li, Z., Du, X., and Qian, H. DPVI: A dynamic-weight particle-based variational inference framework. In Raedt, L. D. (ed.), *Proc. Int. Joint Conf. Artif. Intell.*, pp. 4900–4906. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

Zheng, X., Liu, W., Chen, C., Su, J., Liao, X., Hu, M., and Tan, Y. 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.

Zhu, J.-J. 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.

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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

A. Nomenclature

To facilitate reading our manuscript, we provide the major technical terminology we use during our derivation in Table 5.

Table 5. Technical Terminology Table.

Symbol	Description	Symbol	Description
T	Intermediate domain number	$\mathscr{A}$	Upper bound on the L_2 norm of the first variation of the KL divergence
Δ	Performance difference	$\mathscr{B}$	Upper bound on the L_2 norm of the gradient of the first variation of the KL divergence
$\|\|\|_2$	L_2 norm	$\mathscr{H}_0$	Light-tail constant
$\arg\min$	Argument of the minimum	∇	Gradient operator
$\boldsymbol{T}$	Transportation map	ω	Parameter of classifier
$\boldsymbol{T}^{-1}$	Inverse transformation of the transportation map	ϕ	Parameter of potential function
δ	Variation operator	π	Transportation plan
$\det$	Determinant	$\rho^*(x)$	Optimal PDF
ϵ	Entropy regularization coefficient	$\sup$	Supremum
η	Discretization stepsize	θ	Parameter of transportation map
$\exp$	Exponential function	δ	Dirac delta mass
$\inf$	Infimum	ε	Generalization error
ι	Lipschitz constant of the loss function	$\widehat{h}$	Logit layer of classifier
$\kappa(x, y)$	Reference joint PDF	$\widehat{y}$	Predicted label
λ_1	Coefficient of unbalanced optimal transport	ζ	Lipschitz constant bound for hypotheses in $\mathcal{H}$
λ_2	Coefficient of unbalanced optimal transport	c	Cost matrix
$\mathbb{D}_f[\rho(x), p_T(x)]$	f-divergence of distribution $q(x)$ with-respect-to distribution $p(x)$	$f^\star(x)$	Convex conjugate function of $f(x)$
$\mathbb{D}_{\mathrm{KL}}[\rho(x), p_T(x)]$	Kullback-Leibler divergence of distribution $q(x)$ with-respect-to distribution $p(x)$	h	Classifier
$\mathbb{E}_q(x)[f(x)]$	Expectation of function $f(x)$ with-respect-to distribution $q(x)$	t	Time index
$\mathbb{I}$	indicator function	u	Kantorovich potential function
$\mathcal{B}$	Batch size	v_t	Velocity field
$\mathcal{C}$	Cumulative domain transportation and label continuity costs along the adaptation path	w	Kantorovich potential function
$\mathcal{H}$	Hypothesis class for classifier	x	Input
$\mathcal{L}$	Loss function	y	Label
$\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})$	D-dimensional Wasserstein space	GDA	Gradual domain adaptation
$\mathcal{S}_{\text{stat}}$	Statistical error term.	JKO	Jordan-Kinderlehrer-Otto
$\mathcal{T}$	Set of transportation map	KL divergence	Kullback-Leibler divergence
$\mathcal{W}_p$	p-Wasserstein distance	OT	Optimal transport
$\mathcal{X}$	Input space	PDF	Probability density function
$\mathcal{Y}$	Label space	UDA	Unsupervised domain adaptation
N	Sample size	UOT	Unbalanced optimal transport
d	Differential operator		
$\mathrm{softmax}$	softmax function		

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

B. Mathematical Background on Optimal Transport

We begin by reviewing the relevant background of optimal transport, based on references (Peyré et al., 2019; Villani et al., 2009). 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:

$$\int \pi(x, y)\, \mathrm{d}y = \rho(x), \tag{13a}$$

$$\int \pi(x, y)\, \mathrm{d}x = \xi(y), \tag{13b}$$

and minimize expected cost:

$$\inf_{\pi \geq 0} \iint c(x, y)\, \pi(x, y)\, \mathrm{d}y\, \mathrm{d}x, \tag{14}$$

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:

$$\sup_{u,w} \Big[\int u(x)\, \rho(x)\, \mathrm{d}x + \int w(y)\, \xi(y)\, \mathrm{d}y \Big] \quad \text{s.t.} \quad u(x) + w(y) \leq c(x, y) \;\; \forall x, y. \tag{15}$$

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 (Peyré et al., 2019).

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

$$w^c(x) := \inf_y c(x, y) - w(y). \tag{16}$$

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

$$\sup_w \int w^c(x)\, \rho(x)\, \mathrm{d}x + \int w(y)\, \xi(y)\, \mathrm{d}y. \tag{17}$$

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 $\xi(y)$:

$$\min_{\pi \geq 0} \iint c(x, y)\, \pi(x, y)\, \mathrm{d}y\, \mathrm{d}x + \lambda_1\, \mathbb{D}_f(\tilde{\rho}(x), \rho(x)) + \lambda_2\, \mathbb{D}_f(\tilde{\xi}(y), \xi(y)), \tag{18}$$

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

$$\max_{u,w} - \int \rho(x)\, f_1^\star\big(-u(x)\big)\, \mathrm{d}x - \int \xi(y)\, f_2^\star\big(-w(y)\big)\, \mathrm{d}y \quad \text{s.t.} \quad u(x) + w(y) \leq c(x, y)\; \forall x, y, \tag{19}$$

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

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

$$\max_w - \int \rho(x)\, f_1^\star\big(-w^c(x)\big)\, \mathrm{d}x - \int \xi(y)\, f_2^\star\big(-w(y)\big)\, \mathrm{d}y, w^c(x) = \inf_y \{c(x, y) - w(y)\}. \tag{20}$$

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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

C. Theoretical Derivation

C.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:

$$x_{t+\eta} = x_t - \eta \nabla \frac{\delta \mathbb{D}_f[p(x_t), p_T]}{\delta p(x_t)} \Rightarrow p(x_{t+\eta}) = \underset{\rho(x)\in\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})}{\arg\min} \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)]. \tag{21}$$

Notably, the optimization problem given by the right-hand-side of the abovementioned equation is also called Jordan-Kinderlehrer-Otto canonical form (Caluya & Halder, 2020; 2022; Jordan et al., 1998) 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:

$$\mathcal{W}_2^2(\rho, \xi) = \inf_{\boldsymbol{T}:\boldsymbol{T}_{\#}\rho(x)=\xi(x)} \int \|x - \boldsymbol{T}(x)\|_2^2 \, \rho(x)\mathrm{d}x, \tag{22}$$

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

$$\boldsymbol{T}(x) = x + \eta v_t(x). \tag{23}$$

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

$$\frac{\partial \rho(x_t)}{\partial t} = -\nabla \cdot [v_t(x_t)\rho(x_t)]. \tag{24}$$

Building on Equations (23) and (24), and discretizing the continuity equation in the time domain using the forward Euler scheme (Butcher, 2016; Evans, 2022), we obtain:

$$\rho(x) = \rho(x_t) - \eta\nabla \cdot (\rho(x_t)v_t(x_t)) + \mathcal{O}(\eta^2). \tag{25}$$

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

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta}\mathbb{D}_f[\rho(x), p_T(x)] &= \frac{\mathrm{d}}{\mathrm{d}\eta}\int p_T(x)\, f\left(\frac{\rho(x)}{p_T(x)}\right) dx \\
&= \int p_T(x)\, \frac{\mathrm{d}}{\mathrm{d}\eta} f\left(\tfrac{\rho(x)}{p_T(x)}\right) dx \\
&= \int \cancel{p_T(x)}\, f'\left(\tfrac{\rho(x)}{p_T(x)}\right) \tfrac{1}{\cancel{p_T(x)}} \frac{\partial \rho(x)}{\partial \eta} dx \\
&\overset{\text{(i)}}{=} \int \frac{\delta \mathbb{D}_f}{\delta \rho(x)} \frac{\partial \rho(x)}{\partial \eta} dx
\end{aligned} \tag{26}$$

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

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

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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 Equation (25) into Equation (26), we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\eta}\,\mathbb{D}_f[\rho(x), p_T(x)] &= \int \frac{\delta \mathbb{D}_f}{\delta \rho(x)}\,\big[-\nabla\cdot(\rho(x)v_t(x))\big]\,\mathrm{d}x \\
&= \int \frac{\delta \mathbb{D}_f}{\delta \rho(x)}\,\big[-v_t^\top(x)\nabla\rho(x) - \rho(x)\nabla\cdot v_t(x)\big]\,\mathrm{d}x \\
&\overset{\text{(ii)}}{=} \int \Big(-\nabla\cdot\big[\tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\rho(x)v_t(x)\big] + \rho(x)\,v_t^\top(x)\,\nabla\tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\Big)\,\mathrm{d}x \\
&\overset{\text{(iii)}}{=} \int \rho(x)\,v_t^\top(x)\,\nabla\frac{\delta \mathbb{D}_f(\rho(x), p_T(x))}{\delta \rho(x)}\,\mathrm{d}x.
\end{aligned} \tag{27}$$

Step (ii) is based on the chain rule:

$$\nabla\cdot\Big[\tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\,\rho(x)v_t(x)\Big] = \tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\,\rho(x)\,[\nabla\cdot v_t(x)] + \tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\,v_t^\top(x)\,\nabla\rho(x) + \big[\nabla\tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\big]^\top[\rho(x)v_t(x)]. \tag{28}$$

Step (iii) uses a mild regularity assumption (Abraham et al., 2012; Johnson & Zhang, 2018; 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

$$\int -\nabla\cdot\Big[\tfrac{\delta \mathbb{D}_f}{\delta \rho(x)}\,\rho(x)v_t(x)\Big]\mathrm{d}x = 0. \tag{29}$$

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

$$\mathbb{D}_f[\rho(x), p_T(x)] = \mathbb{D}_f[p(x_t), p_T(x)] + \eta \int p(x_t)v_t^\top(x_t)\nabla\frac{\delta \mathbb{D}_f[p(x_t), p_T(x)]}{\delta p(x_t)}\mathrm{d}x. \tag{30}$$

For the squared 2-Wasserstein distance, we get:

$$\mathcal{W}_2^2(\rho(x), p(x_t)) = \int p(x_t)\|x - \boldsymbol{T}^*(x_t)\|_2^2\mathrm{d}x = \eta^2\int p(x_t)\|v_t^*(x_t)\|_2^2\mathrm{d}x \le \eta^2\int p(x_t)\|v_t(x_t)\|_2^2\mathrm{d}x, \tag{31}$$

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 Equations (30) and (31), we finally reach the following result:

$$\begin{aligned}
&\mathbb{D}_f[\rho(x), p_T(x)] + \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x), p(x_t)) - \mathbb{D}_f[p(x_t), p_T(x)] \\
\le&\cancel{\mathbb{D}_f[\rho(x), p_T(x)]} + \frac{\eta}{2}\mathbb{E}_{p(x_t)}[\|v_t(x_t)\|_2^2] + \eta\int\cdot[p(x_t)v_t^\top(x_t)\nabla\frac{\delta \mathbb{D}_f[p(x_t), p_T(x)]}{\delta p(x_t)}]\mathrm{d}x - \cancel{\mathbb{D}_f[\rho(x), p_T(x)]} \\
\le&\frac{\eta}{2}\underbrace{\mathbb{E}_{p(x_t)}[\|\nabla\frac{\delta \mathbb{D}_f[p(x_t), p_T(x)]}{\delta p(x_t)}]\|_2^2]}_{\ge 0} + \frac{\eta}{2}\mathbb{E}_{p(x_t)}[\|v_t(x_t)\|_2^2] + \eta\int p(x_t)v_t^\top(x_t)\nabla\frac{\delta \mathbb{D}_f[p(x_t), p_T(x)]}{\delta p(x_t)}\mathrm{d}x \\
=&\frac{\eta}{2}\mathbb{E}_{p(x_t)}\{\|v_t(x_t) + \nabla\frac{\delta \mathbb{D}_f[p(x_t), p_T(x)]}{\delta p(x_t)}\|_2^2\}.
\end{aligned} \tag{32}$$

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

$$v_t^*(x_t) = -\nabla\frac{\delta \mathbb{D}_f[p(x_t), p_T(x)]}{\delta p(x_t)}, \tag{33}$$

which implies that the left-hand side of Equation (21) is a sufficient condition for the optimality of its right-hand side.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

C.2. Derivation of Proposition 3.1

Proposition (3.1). Consider the following primal problem:

$$\mathcal{L}^{\text{Primal}} = \underset{\rho(x)\in\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})}{\arg\min} \frac{1}{2\eta}\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)]. \tag{34}$$

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

$$\mathcal{L}^{\text{SemiDual}} = \sup_{w} \mathbb{E}_{p(x_t)}\left[\inf_{\boldsymbol{T}}\left(\|\boldsymbol{T}(x_t) - x_t\|_2^2 - w(\boldsymbol{T}(x_t))\right)\right] - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{35}$$

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. Equation (34) can be reformulated as follows:

$$\inf_{\pi\in\mathbb{R}_+^{\mathrm{D}\times\mathrm{D}}} \quad \frac{1}{2\eta}\iint \|x_t - x\|_2^2\pi(x_t, x)\mathrm{d}x_t\mathrm{d}x + \int f\left(\frac{\rho(x)}{p_T(x)}\right)p_T(x)\mathrm{d}x, \tag{36a}$$

$$\text{s.t.} \quad p(x_t) = \int \pi(x_t, x)\mathrm{d}x, \quad \rho(x) = \int \pi(x_t, x)\mathrm{d}x_t. \tag{36b}$$

Based on this, we introduce the Lagrangian multiplier (Biegler, 2010) $u(x_t)$ and $w(x)$ to handle the equality constraints given by Equation (36b) as follows:

$$\begin{aligned}\mathcal{L} =& \frac{1}{2\eta}\iint \|x_t - x\|_2^2\pi(x_t, x)\mathrm{d}x_t\mathrm{d}x + \int f\left(\frac{\rho(x)}{p_T(x)}\right)p_T(x)\mathrm{d}x \\ &+ \int u(x_t)[p(x_t) - \int \pi(x_t, x)\mathrm{d}x]\mathrm{d}x_t + \int w(x)[\rho(x) - \int \pi(x_t, x)\mathrm{d}x_t]\mathrm{d}x \\ =& \iint [\frac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x)]\pi(x_t, x)\mathrm{d}x_t\mathrm{d}x \\ &+ \int u(x_t)p(x_t)\mathrm{d}x_t + \int w(x)\rho(x) + f\left(\frac{\rho(x)}{p_T(x)}\right)p_T(x)\mathrm{d}x.\end{aligned} \tag{37}$$

On this basis, the dual function can be given as follows due to the linear independent structure of problem defined by Equation (37):

$$\begin{aligned} g(u, w) \;=\; & \inf_{\pi(x_t,x)} \iint \left[\tfrac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x)\right]\pi(x_t, x)\,\mathrm{d}x_t\,\mathrm{d}x \\ & + \int u(x_t)p(x_t)\,\mathrm{d}x_t + \inf_{\rho(x)} \int \left[w(x)\,\frac{\rho(x)}{p_T(x)} + f\Big(\tfrac{\rho(x)}{p_T(x)}\Big)\right]p_T(x)\,\mathrm{d}x \\ =\; & \inf_{\pi(x_t,x)} \iint \left[\tfrac{1}{2\eta}\|x_t - x\|_2^2 - u(x_t) - w(x)\right]\pi(x_t, x)\,\mathrm{d}x_t\,\mathrm{d}x \\ & + \int u(x_t)p(x_t)\,\mathrm{d}x_t - \int p_T(x)\,f^\star(-w(x))\,\mathrm{d}x. \end{aligned} \tag{38}$$

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

$$\begin{aligned} \inf_{\rho(x)} \int \left[w(x)\,\frac{\rho(x)}{p_T(x)} + f\Big(\tfrac{\rho(x)}{p_T(x)}\Big)\right]p_T(x)\,\mathrm{d}x &= \int \inf_{y\geq 0}\big(w(x)\,y + f(y)\big)\,p_T(x)\,\mathrm{d}x \\ &= -\int \sup_{y\geq 0}\big((-w(x))\,y - f(y)\big)\,p_T(x)\,\mathrm{d}x \\ &= -\int p_T(x)\,f^\star\big(-w(x)\big)\,\mathrm{d}x. \end{aligned} \tag{39}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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 Equation (38) 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

$$\sup_{u(x_t)+w(x)\leq \frac{1}{2\eta}\|x_t-x\|_2^2 \ \pi\text{-a.e.}} \left\{ \int u(x_t)p(x_t)\,dx_t - \int p_T(x) f^\star(-w(x))\,dx \right\}. \tag{40}$$

Equivalently, introducing the convex indicator function ℓ, this becomes

$$\sup_{u,w} \left\{ \int u(x_t)p(x_t)\,dx_t - \int p_T(x) f^\star(-w(x))\,dx - \ell\left(u(x_t) + w(x) \leq \frac{1}{2\eta}\|x_t - x\|_2^2 \right) \right\}. \tag{41}$$

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 Equation (41) 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,

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

Substituting this into the dual yields the semi-dual formulation

$$\sup_{w(x)} \left\{ \int \inf_x \left[\frac{1}{2\eta}\|x_t - x\|_2^2 - w(x)\right] p(x_t)\,dx_t - \int p_T(x) f^\star(-w(x))\,dx \right\}. \tag{42}$$

Defining the transport map via the c-transform as

$$\begin{aligned} &\boldsymbol{T}^*(x_t) \in \arg\min_x \left(\frac{1}{2\eta}\|x_t - x\|_2^2 - w(x) \right) \\ \iff\ &\inf_x \left(\frac{1}{2\eta}\|x_t - x\|_2^2 - w(x) \right) = \frac{1}{2\eta}\|x_t - \boldsymbol{T}^*(x_t)\|_2^2 - w(\boldsymbol{T}^*(x_t)), \end{aligned} \tag{43}$$

and substituting Equation (43) into Equation (42), we obtain the final semi-dual objective

$$\mathcal{L}^{\text{SemiDual}} = \sup_w \mathbb{E}_{p(x_t)} \left[\| \frac{1}{2\eta}\boldsymbol{T}^*(x_t) - x_t\|_2^2 - w(\boldsymbol{T}^*(x_t)) \right] - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{44}$$

It should be pointed out that there is no closed-form expression of the optimal $\boldsymbol{T}^*(x_t)$ for each $w(x)$ (Choi et al., 2023; Korotin 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 Equation (44):

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

□

C.3. Derivation of Proposition 3.2

Proposition (3.2). The semi-dual formulation in Equation (7) admits non-unique optimal solutions.

Proof. Consider the discrete optimal transport setting with a single source point (x_t in Equation (7)) and two symmetric target points (x in Equation (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 $p(x_t) \approx \delta_a$.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

- **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:

$$u(a) + w(b_j) \le \|a - b_j\|_2^2 = K, \quad \forall j \in \{1, 2\}. \tag{45}$$

Based on this, we can define a shifted source potential $\tilde{u} := u - K$ and keep $\tilde{w} := w$. Hence, the feasibility in Equation (45) can be given as follows:

$$\tilde{u}(a) + \tilde{w}(b_j) \le 0, \quad \forall j \in \{1, 2\}, \tag{46}$$

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

$$u + w_j \le 0, \quad \forall j \in \{1, 2\}. \tag{47}$$

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:

$$\max_{u, w_1, w_2} \quad u + \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 \; - \; \tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2), \tag{48}$$

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:

$$u = -\min\{w_1, w_2\}. \tag{49}$$

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

$$\Phi(w_1, w_2) \coloneqq -\min\{w_1, w_2\} + \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 - \tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2). \tag{50}$$

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

$$L(w_1, w_2) := -\min\{w_1, w_2\} + \tfrac{1}{2}w_1 + \tfrac{1}{2}w_2 = \begin{cases} \frac{1}{2}(w_2 - w_1), & w_1 \le w_2, \\ \frac{1}{2}(w_1 - w_2), & w_2 \le w_1, \end{cases} \tag{51}$$

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:

$$\Phi(w_1, w_2) = \tfrac{1}{2}|w_1 - w_2| - \tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2). \tag{52}$$

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

$$\Phi(r, r) = -f^\star(-r). \tag{53}$$

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:

$$w_1 = r^* + \delta, \qquad w_2 = r^* - \delta, \qquad \delta \in \mathbb{R}. \tag{54}$$

Then we obtain the following result:

$$\tfrac{1}{2}|w_1 - w_2| = \tfrac{1}{2}|2\delta| = |\delta|. \tag{55}$$

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

$$-\tfrac{1}{2}f^\star(-w_1) - \tfrac{1}{2}f^\star(-w_2) = -f^\star(-r^*) - \tfrac{1}{2}\big(f^{*\prime\prime}(-r^*)\big)\,\delta^2 + \mathcal{O}(\delta^2). \tag{56}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Based on this, we get:

$$\Phi(r^* + \delta,\, r^* - \delta) = |\delta| \;-\; \tfrac{1}{2} f^{*\prime\prime}(-r^*)\, \delta^2 \;-\; f^\star(-r^*) + \mathcal{O}(\delta^2). \tag{57}$$

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 Equation (7) admits non-unique optimal solutions. □

C.4. Derivation of Proposition 3.3

Proposition (3.3). Let $\kappa(x_t, x) := p(x_t)\, p_T(x)$ denote the reference joint PDF. The entropy-regularized primal problem is

$$\begin{aligned} \mathcal{L}^{\text{E-Primal}} = \underset{\rho \in \mathcal{P}_2(\mathbb{R}^D)}{\arg\min}\; & \frac{1}{2\eta}\, \mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f\big[\rho(x), p_T(x)\big] \\ & + \epsilon \iint \pi(x_t, x)\, [\log \frac{\pi(x_t, x)}{\kappa(x_t, x)} - 1]\, \mathrm{d}x_t\, \mathrm{d}x, \end{aligned} \tag{58}$$

and is equivalent to the semi-dual optimization problem

$$\mathcal{L}^{\text{E-SemiDual}} = \sup_{w} -\epsilon\, \mathbb{E}_{p(x_t)}[\log \mathbb{E}_{p_T(x)}(\exp(\frac{w(x) - \frac{1}{2\eta}\|x - x_t\|_2^2}{\epsilon}))] \;- \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{59}$$

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 Equation (58) is

$$\begin{aligned} \mathcal{L}(\pi, \rho; u, w) = & \iint c(x_t, x)\, \pi(x_t, x)\, dx_t\, dx + \epsilon \iint \pi(x_t, x) \Big[\log \frac{\pi(x_t, x)}{\kappa(x_t, x)} - 1\Big] dx_t\, dx \\ & + \int f\Big(\tfrac{\rho(x)}{p_T(x)}\Big) p_T(x)\, dx + \int u(x_t) \Big[p(x_t) - \int \pi(x_t, x)\, dx\Big] dx_t \\ & + \int w(x) \Big[\rho(x) - \int \pi(x_t, x)\, dx_t\Big] dx. \end{aligned} \tag{60}$$

Grouping π-, ρ- and constant terms yields

$$\begin{aligned} \mathcal{L} = & \iint \big(c(x_t, x) - u(x_t) - w(x)\big)\, \pi(x_t, x)\, dx_t\, dx \\ & + \epsilon \iint \pi(x_t, x) \Big[\log \frac{\pi(x_t, x)}{\kappa(x_t, x)} - 1\Big] dx_t\, dx \\ & + \int \Big(w(x)\rho(x) + f\big(\tfrac{\rho(x)}{p_T(x)}\big) p_T(x)\Big) dx + \int u(x_t) p(x_t)\, dx_t. \end{aligned} \tag{61}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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

$$y^\star = \kappa\, e^{-a/\epsilon} = \kappa\, e^{(u+w-c)/\epsilon}. \tag{62}$$

Substituting back yields

$$\inf_{y \geq 0} \phi(y) = -\epsilon\, \kappa\, e^{-a/\epsilon} = -\epsilon\, \kappa\, \exp\Big(\tfrac{u+w-c}{\epsilon}\Big). \tag{63}$$

Hence

$$\inf_{\pi \geq 0} \Big\{\pi\text{-terms of Equation (61)}\Big\} = -\epsilon \iint \kappa(x_t, x)\, \exp\Big(\tfrac{u(x_t)+w(x)-c(x_t,x)}{\epsilon}\Big)\, dx_t\, dx. \tag{64}$$

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

$$\inf_{\rho(x) \geq 0} \Big\{w(x)\rho(x) + f\big(\tfrac{\rho(x)}{p_T(x)}\big)p_T(x)\Big\} = -p_T(x)\, f^\star\big(-w(x)\big). \tag{65}$$

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 Equation (64) and Equation (65), we obtain

$$\begin{aligned} g(u, w) = &-\epsilon \iint \kappa(x_t, x)\, \exp\Big(\tfrac{u(x_t)+w(x)-c(x_t,x)}{\epsilon}\Big) dx_t\, dx \\ &- \int p_T(x)\, f^\star\big(-w(x)\big)\, dx + \int u(x_t) p(x_t)\, dx_t. \end{aligned} \tag{66}$$

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

$$A(x_t) := \int \exp\Big(\tfrac{w(x)-c(x_t,x)}{\epsilon}\Big)\, p_T(x)\, \mathrm{d}x. \tag{67}$$

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, Equation (66) can be reformulated as follows:

$$g(u, w) = \int \Big[p(x_t)\, u(x_t) - \epsilon\, p(x_t)\, A(x_t)\, e^{u(x_t)/\epsilon}\Big]\mathrm{d}x_t - \int p_T(x)\, f^\star(-w(x))\, \mathrm{d}x. \tag{68}$$

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

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

gives

$$e^{u^\star(x_t)/\epsilon} = \frac{1}{A(x_t)} \quad \Longleftrightarrow \quad u^\star(x_t) = -\,\epsilon \log A(x_t). \tag{69}$$

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

$$\sup_w\ -\epsilon\, \mathbb{E}_{p(x_t)}\big[\log A(x_t)\big] - \mathbb{E}_{p_T(x)}\big[f^\star(-w(x))\big], \tag{70}$$

with $A(x_t)$ defined in Equation (67). □

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

C.5. Derivation of Proposition 3.4

Proposition (3.4). The semi-dual formulation in Equation (9) admits a unique optimal solution.

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

$$g(w) = -\epsilon\,\mathbb{E}_{p(x_t)}\{\log \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - \|x - x_t\|_2^2}{\epsilon})]\} - \mathbb{E}_{p_T(x)}[f^\star(-w(x))], \tag{71}$$

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:
$$\Phi_\epsilon(w; x_t) \coloneqq -\epsilon \log \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - \|x - x_t\|_2^2}{\epsilon})], \tag{72}$$
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. □

C.6. Derivation of Theorem 3.5

Theorem (3.5). The optimal solution $\rho^*(x)$ to problem defined in Equation (8) satisfies the following bound:

$$\mathbb{D}_f[\rho^*(x), p_T(x)] \leq \mathcal{W}_2(p(x_t), p_T(x)). \tag{73}$$

According to the definition of the JKO proximal recursion Equations (6) and (8), each update $p(x_{t+1})$ satisfies the following inequality (see Theorem 4.0.4 of reference (Ambrosio et al., 2005)):

$$\frac{1}{2\eta}\mathcal{W}_2^2(p(x_{t+1}), p(x_t)) + \mathbb{D}_f[p(x_{t+1}), p_T(x)] \leq \mathbb{D}_f[p(x_t), p_T(x)],$$

where $\mathbb{D}_f[p(x_t), p_T(x)] \geq 0$ and attains its minimum at the target distribution $p_T(x)$. This inequality implies that the total energy decreases at every step, thereby reducing the Wasserstein distance to $p_T(x)$ when $\mathbb{D}_f[p(x_{t+1}), p_T(x)]$ is geodesically convex:

$$\mathcal{W}_2\big(p(x_{t+1}), p_T(x)\big) \leq \mathcal{W}_2\big(p(x_t), p_T(x)\big) - \Delta_t, \quad \Delta_t \geq 0. \tag{74}$$

Hence, as t increases, based on Equations (73) and (74) the upper bound on $\mathbb{D}_f[\rho^*(x_t), p_T(x)]$ is gradually tightened:

$$\mathbb{D}_f[\rho^*(x_{t+1}), p_T(x)] \leq \mathbb{D}_f[\rho^*(x_t), p_T(x)] - \Delta_t,$$

which shows that the transported probability density $\rho(x)$ progressively becomes more similar to the target $p_T(x)$.

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

$$\begin{aligned}\mathcal{W}_{2,\epsilon}^2(\rho, \xi) \coloneqq \inf_{\pi \in \Pi(\rho,\xi)} \iint \|x - y\|_2^2 \pi(x, y)\mathrm{d}x\mathrm{d}y \\ + \epsilon \iint \pi(\rho(x), p(x_t))[\log \pi(\rho(x), p(x_t)) - 1]\mathrm{d}x\mathrm{d}x_t,\end{aligned} \tag{75}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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

$$\mathbb{D}_f[\rho(x), p_T(x)] = \sup_{v(x)} \left\{ \mathbb{E}_{\rho(x)}[v(x)] - \mathbb{E}_{p_T(x)}\left[f^\star(v(x))\right] \right\}. \tag{76}$$

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

$$\inf_{\rho(x)} \quad \mathcal{W}_{2,\epsilon}(\rho(x), p(x_t)) + \sup_{v(x)} \{ \mathbb{E}_{\rho(x)}[v(x)] - \mathbb{E}_{p_T(x)}\left[f^\star(v(x))\right] \} \tag{77}$$

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

$$\sup_{v(x)} -\mathbb{E}_{p_T(x)}[f^\star(v(x))] + \inf_{\rho(x)} \{ \mathcal{W}_{2,\epsilon}(\rho(x), p(x_t)) + \mathbb{E}_{\rho(x)}[v(x)] \} \tag{78}$$

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 $p(x_t)$:

$$\min_{\rho(x)} \quad \mathcal{W}_{2,\epsilon}(\rho(x), p(x_t)) + \mathbb{E}_{\rho(x)}\left[v(x)\right] \tag{79}$$

whose optimal value equals

$$\mathbb{E}_{p(x_t)}[-\epsilon \log \int \exp(\frac{v(x) - c(x_t, x)}{\epsilon}) \mathrm{d}y]. \tag{80}$$

This follows from standard duality in entropic optimal transport.

Plug the expression above into the main problem:

$$\sup_{v(x)} \quad \mathbb{E}_{p(x_t)}[-\epsilon \log \int \exp(\frac{v(x) - c(x_t, x)}{\epsilon}) \mathrm{d}y] - \mathbb{E}_{p_T(x)}[f^\star(v(x))]. \tag{81}$$

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:

$$0 = \frac{\partial}{\partial \delta} g(v^* + \delta\psi)\Big|_{\delta=0} = \mathbb{E}_{p(x_t)} \left[\frac{\int \psi(x) \exp\left(\frac{v^*(x) - c(x_t,x)}{\epsilon}\right) \mathrm{d}x}{\int \exp\left(\frac{v^*(x) - c(x_t,x)}{\epsilon}\right) \mathrm{d}x} \right] - \mathbb{E}_{p_T(x)}\left[(f^\star)'(v^*(x))\psi(x) \right], \tag{82}$$

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

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

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

$$\tilde{p}_T(x) = p_T(x)(f^\star)'(v^*(x)). \tag{83}$$

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:

$$\mathbb{D}_f[\rho^*(x) \| p_T(x)] + \mathcal{W}_{2,\epsilon}(\rho^*(x), p(x_t)) \leq \mathcal{W}_{2,\epsilon}(p_T, p(x_t)). \tag{84}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

So in particular, we get:

$$\mathbb{D}_f[\rho^*(x)\|p_T(x)] \leq \mathcal{W}_{2,\epsilon}(p(x_t), p_T(x)). \tag{85}$$

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

$$\epsilon \iint \pi(\rho(x), p(x_t))[\log \pi(\rho(x), p(x_t)) - 1]\mathrm{d}x\mathrm{d}x_t \leq 0. \tag{86}$$

Plugging Equation (86) into Equation (85), we arrive at the desired result. □

C.7. Derivation of Theorem 3.6

Theorem (3.6). Under mild assumptions, the E-SUOT-based GDA ensures that the target domain generalization error is upper-bounded by the following inequality:

$$\varepsilon_{p_T}(h_T) \leq \varepsilon_{p_0}(h_0) + \varepsilon_{p_0}(h_T^*) + \iota\zeta\mathcal{C} + \mathcal{S}_{\text{stat}}, \tag{87}$$

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 $\mathcal{S}_{\text{stat}}$ 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:

$$|\mathcal{L}(a, y) - \mathcal{L}(a', y)| \leq \iota|a - a'|. \tag{88}$$

(A. 2) Each hypothesis $h \in \mathcal{H}$ is ζ-Lipschitz, i.e., for any x, x', we have:

$$|h(x) - h(x')| \leq \zeta\|x - x'\|. \tag{89}$$

(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, Assumption (A.1) is standard and well-justified for classification tasks employing the categorical cross-entropy loss. More specifically, the gradient of $\mathcal{L}_{\text{CE}}$ with respect to the logits a is bounded as

$$\left|\frac{\partial \mathcal{L}_{\text{CE}}}{\partial a_j}\right| = \left|[\text{softmax}(a)]_j - \mathbb{I}(j = y)\right| \leq 1,$$

for any class index j, since each softmax component lies within $[0, 1]$, and the indicator function $\mathbb{I}(j = y)$ takes values in $\{0, 1\}$. By the classical Lagrange mean value theorem (see, Theorem 4 in (Thomas et al., 2014)), this bounded-gradient property implies that $\mathcal{L}_{\text{CE}}$ is globally Lipschitz continuous with respect to its first argument on any bounded input domain, with a Lipschitz constant of at most 1. In practice, this assumption can be further facilitated by applying weight normalization or spectral-norm regularization, which help maintain bounded network outputs and thereby make the Lipschitz condition more readily satisfied during optimization.

Furthermore, Assumptions (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 changes smoothly along the adaptation path. For our construction, since the intermediate domains are generated by incremental, continuous transformations (e.g., gradual style or environmental shifts), the underlying semantics of inputs remain stable, and thus $\mathbb{E}_{p_{t-1}(x)}\left[q_t(x) - q_{t-1}(x)\right]$ is small for

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

every t. This situation typically occurs under covariate shift, where only the input distribution evolves while class definitions stay fixed. However, we acknowledge that this assumption may not hold in fine-grained recognition settings or tasks with rapidly changing label semantics, where small variations in features can lead to distinct class assignments. In such cases, the theoretical guarantees derived under Assumption (A.3) would apply only locally, within regions where the labeling function remains approximately smooth. However, the assumption may not hold in tasks with abrupt semantic boundaries—such as fine-grained classification—where visually similar samples can belong to distinct categories. Our analysis therefore applies when the domain evolution does not induce significant concept shift.

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 3.6. 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:

$$\varepsilon_{p_T}(h_T) = \varepsilon_{p_0}(h_0) + \left[\varepsilon_{p_T}(h_T) - \varepsilon_{p_0}(h_0)\right]. \tag{90}$$

To make the recursion explicit, rewrite this as:

$$\varepsilon_{p_T}(h_T) = \varepsilon_{p_0}(h_0) + \sum_{t=1}^{T} \left[\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1})\right]. \tag{91}$$

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

$$\begin{aligned} &\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1}) \\ &= \underbrace{\left[\varepsilon_{p_t}(h_t) - \varepsilon_{p_t}(h_{t-1})\right]}_{\text{optimization error}} + \underbrace{\left[\varepsilon_{p_t}(h_{t-1}) - \varepsilon_{p_{t-1}}(h_{t-1})\right]}_{\text{domain shift term}} + \underbrace{\left[\varepsilon_{p_{t-1}}(h_{t-1}) - \varepsilon_{p_{t-1}}(h_t)\right]}_{\leq 0 \text{ by ERM}} \end{aligned} \tag{92}$$

In practice, the last term is non-positive since 'empirical risk minimization' (Shalev-Shwartz & Ben-David, 2014; Vapnik, 1999; 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,

$$|\varepsilon_{p_t}(h) - \varepsilon_{p_{t-1}}(h)| \leq \iota\zeta \cdot \mathcal{W}_1(p_{t-1}, p_t). \tag{93}$$

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

$$\iota\, \mathbb{E}_{p_t(x)} |q_t(x) - q_{t-1}(x)|. \tag{94}$$

Therefore, each step can be bounded by

$$|\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1})| \leq \iota\zeta \mathcal{W}_1(p_{t-1}, p_t) + \iota\, \mathbb{E}_{p_t(x)} |f_t(x) - f_{t-1}(x)| + s_t \tag{95}$$

where s_t denotes the statistical error at step t.

Let

$$\mathcal{C} := \sum_{t=0}^{T-1} \left[\mathcal{W}_1(p_{t-1}, p_t) + \frac{1}{\zeta} \mathbb{E}_{p_t(x)} |q_t(x) - q_{t-1}(x)|\right] \tag{96}$$

and

$$\mathcal{S}_{\text{stat}} := \sum_{t=1}^{T-1} s_t. \tag{97}$$

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

$$\sum_{t=0}^{T-1} |\varepsilon_{p_t}(h_t) - \varepsilon_{p_{t-1}}(h_{t-1})| \leq \iota\zeta\mathcal{C} + \mathcal{S}_{\text{stat}}. \tag{98}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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

$$\varepsilon_{p_0}(h_0) + \varepsilon_{p_0}(h_T^*) - \varepsilon_{p_0}(h_0) \tag{99}$$

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. □

While Theorem 3.6 provides a clean decomposition, the last two terms are not directly computable from data. To bridge this gap between theory and practice, we attempt to estimate them using the following strategies:

- **Loss Lipschitz constant ι:** According to the analysis of Assumption (A.1), we conclude that $\iota < 1$ with the help of applying weight normalization or spectral-norm regularization.
- **Hypothesis Lipschitz constant ζ.** This bounds how sensitively hypotheses $h \in \mathcal{H}$ react to input perturbations. In our implementation process, we use the multi-layer-perceptron with ReLU activated function. Thus, the Lipschitz constant ζ for the classfier can be estimated as follows:

$$\zeta \;\leq\; \prod_{\ell=1}^{L} \|W_\ell\|_2, \tag{100}$$

  where W_ℓ is the weight of the linear layer, $\|W_\ell\|_2$ is the spectral norm of W_ℓ. Similarily, ζ can be controlled by enforcing spectral normalization or weight normalization on layers.
- **Cumulative cost $\mathcal{C}$:** Recall Equation (96), we observe that the 1-Wasserstein distance term $\mathcal{W}_1(p_{t-1}, p_t)$ can be approximated using sample-based optimal transport distances, for example, Sinkhorn distance (Cuturi, 2013), computed on intermediate feature representations. Besides, the inter-step labeling-function shift term $\mathbb{E}_{p_t(x)}\big|q_t(x) - q_{t-1}(x)\big|$ in Equation (96) can be approximated in practice via pseudo-labels predicted by the models at steps $t-1$ and t, effectively quantifying how much pseudo-labels change along the adaptation path.
- **Statistical error $\mathcal{S}_{\text{stat}}$:** The term $\mathcal{S}_{\text{stat}} = \sum_{t=1}^{T-1} s_t$ collects the statistical deviations between empirical and population risks at each adaptation step. Under mild assumptions, including G-Lipschitz losses and norm-constrained neural networks, standard uniform-convergence bounds based on Rademacher complexity yield the following result, adapted from Section 9.4 of (Bach, 2024):

$$s_t = \mathcal{O}\Big(\frac{1}{\sqrt{\mathrm{N}}}\Big), \tag{101}$$

  Consequently, for all $t \in \{0, 1, \ldots, T-1\}$, we can estimate $\mathcal{S}_{\text{stat}}$ as follows:

$$\mathcal{S}_{\text{stat}} = \mathcal{O}(\frac{T}{\sqrt{\mathrm{N}}}). \tag{102}$$

C.8. Discussions on the selection of step size η

Let us recall Equations (6) and (18) as follows:

$$\begin{cases} \mathcal{L}^{\text{Primal}} = \arg\min_{\rho(x)\in\mathcal{P}_2(\mathbb{R}^{\mathrm{D}})} \frac{1}{2\eta}\,\mathcal{W}_2^2(\rho(x), p(x_t)) + \mathbb{D}_f[\rho(x), p_T(x)], \\ \min_{\pi\geq 0} \iint c(x,y)\,\pi(x,y)\,\mathrm{d}y\,\mathrm{d}x + \lambda_1\,\mathbb{D}_f(\tilde{\rho}(x), \rho(x)) + \lambda_2\,\mathbb{D}_f(\tilde{\xi}(y), \xi(y)). \end{cases} \tag{103}$$

Since Equation (6) is a variant of Equation (18), where $\lambda_1 \equiv 0$. Based on this, we may raise one question: How to select the λ_2, i.e. η?

Since our target is decreasing the functional $\mathbb{D}_f[\rho(x), p_T(x)]$ along the simulation process, thus one of the key factor is that the selection of the η can decrease the functional $\mathbb{D}_f[\rho(x), p_T(x)]$. Take the KL divergence, the f-divergence we consider in the proposed E-SUOT approach, we have the following proposition for selecting the η:

Proposition C.1. *Suppose that the $\|\frac{\delta\mathbb{D}_{\mathrm{KL}}[\rho(x),p_T(x)]}{\delta\rho(x)}\| \leq \mathscr{A}$ and $\|\nabla\frac{\delta\mathbb{D}_{\mathrm{KL}}[\rho(x),p_T(x)]}{\delta\rho(x)}\| \leq \mathscr{B}$, there exists a constant $\mathscr{H}_0$ control the tailness of $p_T(x)$, and let $\{\rho_t(x)|t = 1, \ldots, T\}$ denote the sequence of empirical PDF of the intermediate*

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

domain generated by the JKO recursion. When η satisfies the following condition, the sequence of KL divergence $\{\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)]|t = 1, \ldots, T\}$ converges to a finite value as $t \to \infty$:

$$0 < \eta < \min(\frac{1}{\mathscr{B}}, \frac{\mathscr{H}_0}{\mathscr{A}}). \tag{104}$$

Before proposing the proof, we should introduce the light-tailness property on the target distribution $p_T(x)$ in order to ensure the validity of our Taylor expansion and to control higher-order discretization errors during the proof process. Specifically, we say that $p_T(x)$ is light-tailed (Ambrosio et al., 2005; Johnson & Zhang, 2018; 2021) if there exists a universal constant $\mathscr{H}_0 < \infty$ such that

$$\int \|\nabla \log p_T(x)\| \, p_T(x) \mathrm{d}x < \mathscr{H}_0. \tag{105}$$

We call $\mathscr{H}_0$ the "light-tail constant" of $p_T(x)$. This condition requires that the expectation (under $p_T(x)$) of the norm of the score function $\nabla p_T(x)$ is finite. Intuitively, this ensures that $p_T(x)$ decays sufficiently rapidly in the tails so that the gradients do not blow up at infinity. On this basis, the proof is articulated as follows:

Proof. Suppose at time t and time $t + \eta$, we have:

$$x_{t+\eta} = \boldsymbol{T}(x) \coloneqq x_t + \eta v_t(x_t). \tag{106}$$

We denote the probability distributions, before and after applying Equation (106) as $\rho_t(x)$ and $\rho_{t+\eta}(x)$, respectively. The aim is to Taylor expand the evolution of

$$\mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] = \int \rho_{t+\eta}(x) \log \frac{\rho_{t+\eta}(x)}{p_T(x)} \mathrm{d}x, \tag{107}$$

with respect to η around $\eta = 0$. The new probability distribution, for small η, can be given as follows according to the Liouville's theorem:

$$\rho_{t+\eta}(x) = \rho_t(\boldsymbol{T}^{-1}(x)) \cdot |\det \boldsymbol{J}_{\boldsymbol{T}^{-1}}(z)| \tag{108}$$

where $\boldsymbol{J}_{\boldsymbol{T}^{-1}}(z)$ is the Jacobian matrix of the inverse map, and when η is small enough, the inverse function $\boldsymbol{T}^{-1}(x)$ of function $\boldsymbol{T}(x)$ can be given as follows:

$$\boldsymbol{T}^{-1}(x) \approx x - \eta v_t(x). \tag{109}$$

Hence, expanding to the first order in η can be given as follows:

$$\rho_{t+\eta}(x) \approx \rho_t(x - \eta v_t(x)) \left[1 - \eta \nabla \cdot \phi(z)\right] \approx \rho_t(x) - \eta \nabla \left[\rho_t(x) v_t(x)\right]. \tag{110}$$

Define $F(\eta)$:

$$F(\eta) \coloneqq \mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] = \int \rho_{t+\eta}(x) \log \frac{\rho_{t+\eta}(x)}{p_T(x)} \mathrm{d}x. \tag{111}$$

Applying Taylor's expansion at $\eta = 0$, we can obtain the following result:

$$F(\eta) = F(0) + \eta F'(0) + \mathcal{O}(\eta^2). \tag{112}$$

Now, we start deriving $F'(0)$. When we take the derivative inside the integral, we have:

$$F'(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta} \int \rho_{t+\eta}(x) \log \frac{\rho_{t+\eta}(x)}{p_T(x)} \mathrm{d}x = \int \frac{\mathrm{d}}{\mathrm{d}\eta} \rho_{t+\eta}(x) [1 + \log \frac{\rho_{t+\eta}(z)}{p_T(x)}] \mathrm{d}x \tag{113}$$

At $\eta = 0$, $\rho_{t+\eta}(x) = \rho_t(x)$:

$$F'(0) = \int \frac{\mathrm{d}}{\mathrm{d}\eta} \rho_{t+\eta}(x) \Big|_{\eta=0} [1 + \log \frac{\rho_t(x)}{p_T(x)}] \mathrm{d}x. \tag{114}$$

Now, using the result from the calculus of variations:

$$\frac{\mathrm{d}}{\mathrm{d}\eta} \rho_{t+\eta}(x) \Big|_{\eta=0} = -\nabla \cdot (\rho_t(x) v_t(x)) \tag{115}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Thus,

$$F'(0) = -\int \nabla \cdot (\rho_t(x)v_t(x))[1 + \log \frac{\rho_t(x)}{p_T(x)}]\mathrm{d}x. \tag{116}$$

Now, use integration by parts:

$$\int -\nabla \cdot [\rho_t(x)v_t(x)]g(x)\mathrm{d}z = \int [\rho_t(x)v_t(x)]^\top \nabla g(x)\mathrm{d}x. \tag{117}$$

Set $g(x) = 1 + \log \frac{\rho_t(x)}{p_T(x)}$. Its gradient is:

$$\nabla g(x) = \nabla \log \rho_t(x) - \nabla \log p_T(x). \tag{118}$$

Hence,

$$F'(0) = \int [\rho_t(x)v_t(x)]^\top [\nabla \log \rho_t(x) - \nabla \log p_T(x)]\mathrm{d}x = \mathbb{E}_{\rho_t(x)}\left[v_t^\top(x)(\nabla \log \rho_t(x) - \nabla \log p_T(x))\right]. \tag{119}$$

But with a negative sign because the original derivative is minus divergence:

$$F'(0) = -\mathbb{E}_{\rho_t(x)}\left[v_t^\top(x)(\nabla \log \rho_t(x) - \nabla \log p_T(x))\right]. \tag{120}$$

Putting all together, we get:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] \\ =&\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \eta\mathbb{E}_{\rho_t(x)}\left[v_t^\top(x)(\nabla \log p_T(x) - \nabla \log \rho_t(x))\right] + \mathcal{O}(\eta^2). \end{aligned} \tag{121}$$

Notably, the optimal velocity field $v_t^*(x)$ for KL divergence is $-\nabla \frac{\delta \mathbb{D}_{\mathrm{KL}}[\rho_t(x)]}{\delta \rho_t(x)}$. Thus, we have:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] \\ =&\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] \underbrace{-\eta\mathbb{E}_{\rho_t(x)}\left[(\nabla \log p_T(x) - \nabla \log \rho_t(x))^\top v_t^*(x)\right]}_{\leq 0} + \mathcal{O}(\eta^2). \end{aligned} \tag{122}$$

Since $\|\nabla \frac{\delta \mathbb{D}_{\mathrm{KL}}[\rho(x), p_T(x)]}{\delta \rho(x)}\| \leq \mathscr{B}$, there exists a positive constant C such that:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \eta\mathbb{E}_{\rho_t(x)}\left[(\nabla \log p_T(x) - \nabla \log \rho_t(x))^\top v_t^*(x)\right] + \mathcal{O}(\eta^2) \\ \leq&\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \eta\mathbb{E}_{\rho_t(x)}\left[(\nabla \log p_T(x) - \nabla \log \rho_t(x))^\top v_t^*(x)\right] + C\eta^2, \end{aligned} \tag{123}$$

where constant C satisfies the following condition:

$$C \propto \mathscr{B}^2. \tag{124}$$

To avoid $C\eta^2$ dominating the right-hand-side of Equation (123), we should satisfy the following condition:

$$\frac{1}{\eta^2} \gg \mathscr{B}^2 \Rightarrow \eta \ll \frac{1}{\mathscr{B}} \Rightarrow \eta < \frac{1}{\mathscr{B}}. \tag{125}$$

According to the log–Sobolev inequality (Ambrosio et al., 2005; Villani et al., 2009), for the target distribution $p_T(x)$ satisfying the curvature condition controlled by $\mathscr{H}_0 > 0$ (i.e., strong log–concavity or equivalent tailness control), there exists a log–Sobolev constant $\mathscr{H}_0$ such that for any smooth density $\rho_t(x)$,

$$\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] \leq \frac{\mathscr{H}_0}{2}\,\mathbb{E}_{\rho_t(x)}\left[\|\nabla \log \tfrac{\rho_t(x)}{p_T(x)}\|^2\right]. \tag{126}$$

Equivalently, the following lower-bound form holds:

$$\mathbb{E}_{\rho_t(x)}\left[\|v_t^*(x)\|^2\right] \geq \frac{1}{\mathscr{H}_0}\,\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)], \tag{127}$$

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

where we used $v_t^*(x) = \nabla \log p_T(x) - \nabla \log \rho_t(x)$.

When the variational derivative $\frac{\delta \mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)]}{\delta \rho_t(x)}$ is bounded by $\mathscr{A}$, and the spatial gradient of this functional derivative is bounded by $\mathscr{B}$, the log–Sobolev inequality admits a perturbation-corrected version:

$$\mathbb{E}_{\rho_t(x)}[\|v_t^*(x)\|^2] \geq \frac{1}{\mathscr{H}_0}\{\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \frac{\mathscr{A}}{\mathscr{H}_0}\}, \tag{128}$$

where the correction term $\frac{\mathscr{A}}{\mathscr{H}_0}$ compensates for the bounded $\|\nabla \frac{\delta \mathbb{D}_{\mathrm{KL}}}{\delta \rho}\|$ and ensures dimensional consistency of the energy inequality. Plugging Equation (128) into Equation (123) gives:

$$\begin{aligned} &\mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] \\ \leq &\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \eta\, \mathbb{E}_{\rho_t}[\|v_t^*(x)\|^2] + C\eta^2 \\ \leq &\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \frac{\eta}{\mathscr{H}_0}\{\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] - \frac{\mathscr{A}}{\mathscr{H}_0}\} + C\eta^2, \end{aligned} \tag{129}$$

Rearranging Equation (129), we have:

$$\mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] \leq (1 - \tfrac{\eta}{\mathscr{H}_0})\, \mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] + \tfrac{\eta \mathscr{A}}{\mathscr{H}_0^2} + C\eta^2. \tag{130}$$

To promise the iteration will gradually reduce $\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)]$, we should require:

$$(1 - \tfrac{\eta}{\mathscr{H}_0}) \in (0, 1) \quad \Rightarrow \quad 0 < \eta < \mathscr{H}_0. \tag{131}$$

According to Equation (125), ignoring $C\eta^2$, we obtain the equilibrium point, corresponding to the steady state as $t \to \infty$:

$$\mathbb{D}_{\mathrm{KL}}[\rho_\infty(x), p_T(x)] = \frac{\mathscr{A}}{\mathscr{H}_0}. \tag{132}$$

Next, to ensure that each discrete update indeed decreases the KL divergence, we impose

$$\mathbb{D}_{\mathrm{KL}}[\rho_{t+\eta}(x), p_T(x)] < \mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)]. \tag{133}$$

Substituting Equation (130) into Equation (133) yields the following result:

$$-\frac{\eta}{\mathscr{H}_0}\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] + \frac{\eta \mathscr{A}}{\mathscr{H}_0^2} + C\eta^2 < 0. \tag{134}$$

Dividing both sides by $\eta > 0$ and rearranging terms gives the following equation:

$$\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)] > \frac{\mathscr{A}}{\mathscr{H}_0} + C\eta\, \mathscr{H}_0. \tag{135}$$

In the late stage of GDA task, $\mathbb{D}_{\mathrm{KL}}[\rho_t(x), p_T(x)]$ approaches its equilibrium value $\mathbb{D}_{\mathrm{KL}}[\rho_\infty(x), p_T(x)] = \frac{\mathscr{A}}{\mathscr{H}_0}$, so that:

$$\mathbb{D}_{\mathrm{KL}}[\rho_t, p_T] - \mathbb{D}_{\mathrm{KL}}[\rho_\infty, p_T] \approx \mathbb{D}_{\mathrm{KL}}[\rho_\infty, p_T] = \frac{\mathscr{A}}{\mathscr{H}_0}. \tag{136}$$

Hence, the typical contraction strength per update is of order:

$$\frac{\eta}{\mathscr{H}_0}\left(\mathbb{D}_{\mathrm{KL}}[\rho_t, p_T] - \mathbb{D}_{\mathrm{KL}}[\rho_\infty, p_T]\right) \;\approx\; \frac{\eta}{\mathscr{H}_0}\frac{\mathscr{A}}{\mathscr{H}_0}. \tag{137}$$

To guarantee that the quadratic residual $C\eta^2$ does not dominate the contraction term in Equation (137), we require:

$$C\eta^2 \ll \frac{\eta}{\mathscr{H}_0}\frac{\mathscr{A}}{\mathscr{H}_0} \quad \Longrightarrow \quad \eta \ll \frac{\mathscr{H}_0}{\mathscr{A}}. \tag{138}$$

That is, the discretization step must satisfy the stability condition since $\mathscr{A} > 0$:

$$0 < \eta < \frac{\mathscr{H}_0}{\mathscr{A}} < \mathscr{H}_0, \tag{139}$$

which ensures that the numerical update is dominated by the contraction term rather than by the additive bias or high-order error. Based on Equations (125) and (139), we arrive at the desired result.

□

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

D. Detailed Algorithm of the 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.

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 Adaptation

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: Classifier in target domain $h_{\omega,T}$.

1: Initialize the classifier $h_{\omega,0}$: $h_{\omega,0} \leftarrow \arg\min_\omega \mathcal{L}_{\mathrm{CE}}(x_0, h_{\omega,t}, y_0)$.
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 $i \in \{1, \ldots, \mathrm{N}\}$.
5: Finetune the classifier $h_{\omega,t+1}$: $h_{\omega,t+1} \leftarrow \arg\min_\omega \mathcal{L}_{\mathrm{CE}}(x_{t+1}, h_{\omega,t}, y_{t+1})$.
6: **end for**

E. Detailed Information for Experiments

E.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°.
- **Office-Home:** Office-Home is a domain adaptation benchmark dataset consisting of approximately 15,500 images categorized into 65 object classes commonly found in office and home environments (Venkateswara et al., 2017). The dataset encompasses four visually distinct domains—*Artistic (Ar)*, *Clipart (Cl)*, *Product (Pr)*, and *Real-World (Rw)*. Following common domain adaptation protocols, one domain is selected as the source domain while another serves as the target domain, resulting in a total of 12 domain transfer tasks (e.g., Ar→Rw, Cl→Pr, etc.)

E.2. Experimental Settings

E.2.1. Preliminaries of Self-Training method

Self-training is a classical semi-supervised learning strategy that leverages the model's own predictions on unlabeled data to iteratively improve its performance. Given a model h_ω trained on labeled source data, we use it to generate pseudo-labels for unlabeled samples in a target or auxiliary dataset $\mathcal{D}_{\text{aux}}$. Each unlabeled input $x_i \in \mathcal{D}_{\text{aux}}$ is assigned a pseudo-label $\tilde{y}_i = \mathrm{sign}(h_\omega(x_i))$, indicating a positive or negative prediction. A new model $h_{\omega'}$ is then trained to minimize the empirical loss on this pseudo-labeled dataset:

$$\mathrm{ST}(\theta, \mathcal{D}_{\text{aux}}) \;=\; \arg\min_{h'_\omega \in \mathcal{H}} \frac{1}{\mathrm{N}_{\text{aux}}} \sum_{x_i \in \mathcal{D}_{\text{aux}}} \mathcal{L}(h'_\omega(x_i), \mathrm{sign}(h_\omega(x_i)))\,, \tag{140}$$

where ST is the abbreviation of self-training, N_{aux} denotes the sample size of auxiliary dataset $\mathcal{D}_{\text{aux}}$. This procedure can be iteratively repeated, replacing θ with the newly optimized θ' to refine the pseudo-labels over time.

Intuitively, self-training alternates between

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

1) Producing pseudo-labels using the current classifier.
2) Retraining the model on these pseudo-labels, thereby progressively refining the decision boundary.

In practice, confidence thresholding or pseudo-label sharpening can be incorporated to reduce noise accumulation and improve stability.

E.2.2. Training & Evaluation Protocols

For GDA and UDA task, we follow the standard domain adaptation protocol, where model training and hyperparameter tuning are performed using labeled source data merely since the validation on the target domain is infeasible in unsupervised adaptation setting. All results are reported on the target domian dataset without using target labels for validation or early stopping.

For classifier training under this protocol, let $\widehat{h}_{\omega,t}(x_t)$ denote the logits produced by the classifier parameterized by ω at time step t. We define

$$h_{\omega,t}(x_t) = \mathrm{softmax}\big(\widehat{h}_{\omega,t}(x_t)\big) \tag{141}$$

as the corresponding class-probability vector. Given the ground-truth label y_t, the categorical cross-entropy loss is formulated as follows:

$$\mathcal{L}_{\mathrm{CE}}(x_t, \omega, y_t) = -\sum_{i=1}^{\mathcal{B}} y_t^{(i)} \log h_{\omega,t}^{(i)}(x_t) = -\sum_{i=1}^{\mathcal{B}} y_t^{(i)} \log[\mathrm{softmax}\big(\widehat{h}_{\omega,t}(x_t)\big)^{(i)}]. \tag{142}$$

On this basis, we use the classification accuracy (denoted as "Accuracy") as the evaluation metric, defined as the proportion of correctly predicted target samples:

$$\text{Accuracy} = [\frac{1}{\mathrm{N_{tgt}}} \sum_{i=1}^{\mathrm{N_{tgt}}} \mathbb{I}(\hat{y}_i = y_i)] \times 100\%, \tag{143}$$

where $\mathrm{N_{tgt}}$ is the number of target test samples, $\hat{y}_i$ denotes the predicted label of the i-th sample, y_i is the corresponding ground-truth label, and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 when the condition holds and 0 otherwise. A larger accuracy value corresponds to better adaptation performance.

E.2.3. GDA Task

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 in order to align with the experimental tuple provided by (Zhuang et al., 2024). 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 6.

Table 6. 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. For the classifier h_ω, we use a two-layer MLP with ReLU activations and 128 hidden units. All baseline models are trained on features embedded by the UMAP.

E.2.4. UDA Task

For the UDA task, we adopt the Office-Home dataset (Venkateswara et al., 2017) as the benchmark to evaluate the performance of the proposed E-SUOT framework. Following the standard unsupervised domain adaptation (UDA) protocol, model training and hyperparameter tuning are performed solely using the labeled source data, without access to target labels for validation or early stopping. All results are reported on the target domain dataset.

We compare E-SUOT with a diverse set of representative UDA approaches, including DANN (Ganin & Lempitsky, 2015), MSTN (Xie et al., 2018), GVB-GD (Cui et al., 2020), RSDA (Gu et al., 2020; 2022), LAMBDA (Le et al., 2021),

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

SENTRY (Prabhu et al., 2021), FixBi (Na et al., 2021), CST (Liu et al., 2021a), CoVi (Na et al., 2022), and GGF (Zhuang et al., 2024). For baselines including DANN, MSTN, GVB-GD, RSDA, LAMBDA, SENTRY, FixBi, CST, and CoVi, we directly report the publicly available results from their original papers under identical experimental settings (i.e., the same dataset and evaluation protocol). For the GGF method, as the public release of GGF provides only code for GDA task, we re-implemented the missing components according to the paper's description and ran the experiments locally to maintain consistency with the original experimental protocol.

In addition, similar to GGF, our E-SUOT framework builds upon CoVi as the backbone feature extractor. The extracted features are embedded into an eight-dimensional space using UMAP. The classifiers h_ω for GGF and E-SUOT are implemented as two-layer ReLU MLPs with 256 hidden units. We set the η, $\mathcal{B}$, ϵ, and T as 0.5, 1024, 0.001, and 4, respectively. Both GGF and E-SUOT models are trained under the same conditions for fair comparison.

E.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 Equation (7). Building on (Choi et al., 2023; 2024; Korotin et al., 2021; 2023), 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., 2021b; 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 $t \in \{1, \ldots, \mathrm{T}\}$, 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**

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:

$$\mathbb{D}_{\chi^2}[\rho(x_t), p_T(x)] \int p_T(x)[\frac{\rho(x_t)}{p_T(x)} - 1]^2 \mathrm{d}x_t, \quad \text{where} \quad f(x) = (x-1)^2. \tag{144}$$

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

$$f^\star(x) = \begin{cases} \frac{1}{4}x^2 + x, \text{if} x \geq -2 \\ -1, \text{if} x < -2 \end{cases}. \tag{145}$$

- **Identity:** For the identity function, we remove the f-divergence-based regularization term during the construction of

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

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 $t \in \{1, \ldots, \mathrm{T}\}$, 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**

E-SUOT framework. Based on this, the training objective for w_ϕ is reformulated as follows:

$$\mathcal{L}_{\text{Identity}}^{\text{E-SemiDual}} = \sup_w -\epsilon \mathbb{E}_{p(x_t)}\{\log \mathbb{E}_{p_T(x)}[\exp(\frac{w(x) - \frac{1}{2\eta}\|x - x_t\|_2^2}{\epsilon})]\} + \mathbb{E}_{p_T(x)}[w(x)], \tag{146}$$

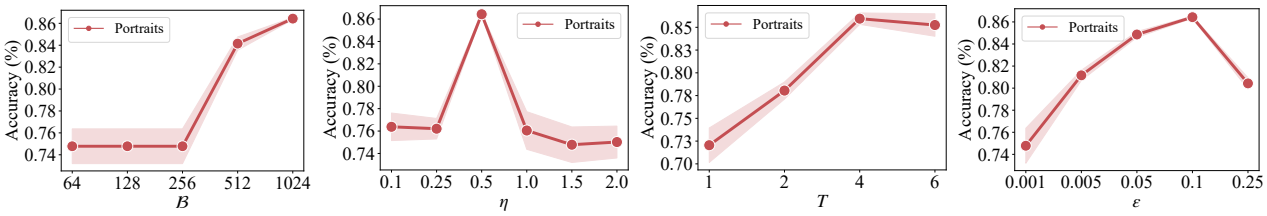
- **`SftPls`:** We directly parameterize the $f^\star(x)$ using the smooth, convex, and non-decreasing softplus function as follows:

$$f^\star(x) = \log(1 + \exp(x)). \tag{147}$$

F. Additional Experimental Results

F.1. Sensitivity Analysis

From Figures 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 dataset.



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

Figure 3. Sensitivity analysis results on Portrait Dataset.

Specifically, as shown in Figure 3(a), increasing the batch size $\mathcal{B}$ consistently improves model performance. This suggests that, in simulations of WGF-based approaches (including ours), using a sufficiently large batch size is beneficial, as it reduces stochastic sampling noise and yields a more stable approximation of the underlying distributions.

A non-monotonic trend is observed when varying the discretization step size η, as illustrated in Figure 3(b). When η is too small, the simulation trajectory may not reach the target distribution within a finite number of steps, limiting learning efficiency. In contrast, an overly large η introduces substantial discretization error, which also degrades performance. These empirical observations are consistent with our theoretical guidance on step-size selection provided in Proposition XX.

Moreover, as demonstrated in Figure 3(c), increasing the number of simulation steps T likewise yields a non-monotonic effect: after a certain threshold, additional steps actually deteriorate performance. A plausible explanation is that enforcing excessively strict alignment between feature and target distributions does not necessarily correspond to optimal performance in the target domain. This observation further supports our introduction of divergence-based regularization, which relaxes strict alignment constraints compared with traditional OT-based methods.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

Finally, as shown in Figure 3(d), the entropy regularization parameter ϵ has a pronounced impact on the results. Different values of ϵ can lead to markedly different performance, underscoring the need to carefully examine and tune the entropy regularization strength in practice.

In summary, this sensitivity analysis highlights the importance of properly choosing the batch size $\mathcal{B}$, step size η, and end time T for achieving good E-SUOT performance, and indicates that the optimal entropy regularization strength ϵ is dataset-dependent, calling for systematic validation on the target dataset.

F.2. Impact of Early Transport Steps on Adaptation Performance

The multi-step structure of E-SUOT involves a series of learned transport maps $\{T_{\theta,0}, T_{\theta,1}, \ldots\}$, where each step aims to progressively align intermediate feature distributions between domains. While this design promotes smooth domain alignment, it also raises an important question: how sensitive the overall adaptation performance is to the quality of early transport maps, and whether suboptimal early mappings introduce cumulative errors that affect subsequent steps.

To investigate this, we conduct an experiment on the Portraits dataset, where we selectively disable the training of certain transport maps to simulate incomplete or inaccurate early-stage optimization. Two complementary training strategies are designed:

- **Forward strategy:** Progressively remove the training of early transport maps $(T_{\theta,0}, T_{\theta,1}, \ldots)$ while keeping the later ones active, thereby testing whether missing early steps hinder later GDA performance.
- **Backward strategy:** progressively disable the training of later maps $(T_{\theta,4}, T_{\theta,3}, \ldots)$ while retaining the trained early steps, examining whether well-trained initial stages are sufficient to sustain strong performance.

Table 7 summarizes the results. We observe that in the "Forward" direction, excluding early transport steps causes a substantial accuracy drop (up to 17.7%), indicating that early-stage mappings are crucial for forming a reliable transport foundation. In contrast, the "Backward" experiments show that once these early steps are properly optimized, subsequent refinements yield consistent performance gains, suggesting that later transport maps mainly provide fine adjustments on top of an already well-aligned feature space.

Table 7. Performance of the E-SUOT vary different training stage on the Portraits dataset.

Direction	Forward							Backward						
Time Index	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	Accuracy (%)	Δ	$t=0$	$t=1$	$t=2$	$t=3$	$t=4$	Accuracy (%)	Δ
Training Status	✗	✓	✓	✓	✓	$76.4_{\pm 2.06\text{E-}2}$	↑7.2%	✓	✗	✗	✗	✗	$81.5_{\pm 1.70\text{E-}2}$	↑14.4%
	✗	✗	✓	✓	✓	$75.0_{\pm 2.48\text{E-}2}$	↑5.3%	✓	✓	✗	✗	✗	$83.2_{\pm 1.13\text{E-}2}$	↑16.8%
	✗	✗	✗	✓	✓	$74.4_{\pm 1.92\text{E-}2}$	↑4.4%	✓	✓	✓	✗	✗	$83.9_{\pm 8.10\text{E-}3}$	↑17.8%
	✗	✗	✗	✗	✓	$74.0_{\pm 1.64\text{E-}2}$	↑3.9%	✓	✓	✓	✓	✗	$84.0_{\pm 6.56\text{E-}3}$	↑17.9%
	✗	✗	✗	✗	✗	$58.6_{\pm 1.87\text{E-}2}$	↓17.7%	✓	✓	✓	✓	✓	$86.4_{\pm 8.72\text{E-}2}$	↑21.5%

Kindly Note: Δ denotes performance change percentage of the initial classifier accuracy.

Overall, the results highlight that the early transport steps are the key drivers of successful adaptation. When the early mappings are well trained, the rest of the chain benefits from a stabilized feature representation, leading to larger and more consistent improvements. This behavior parallels diffusion-like processes, where the early transport transformations largely determine the shape and quality of the final distribution, as also illustrated in Fig. 3 in reference (Caluya & Halder, 2020) and Fig. 1 in reference (Liu & Wang, 2016).

F.3. Investigation of the UOT formulation

While our main contribution lies in introducing the semi-dual UOT formulation to analyze and improve the flow-based GDA approach, we further investigate "why the UOT-based method performs better than the vanilla OT formulation". To this end, we conduct experiments under label-shift and missing-class scenarios using the Portrait dataset (binary classification task). Specifically, we resample the target domain to vary the class prior $p(y=1)$; when $p(y=1)=0.0$ or $p(y=1)=1.0$, a missing-class situation is realized. The comparison results between vanilla OT and UOT are reported in Table 8. Here, "E-SOT" denotes entropy-regularized semi-dual optimal transport. From the table, we observe that in the missing-class cases

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

($p(y=1)=0.0$ or $p(y=1)=1.0$), the vanilla OT formulation not only fails to improve but even degrades the classifier's performance in the target domain. Moreover, under moderate label shift ($p(y=1)$ between 0.3 and 0.9), the performance of vanilla OT fluctuates strongly and lacks stability, whereas UOT consistently improves performance. These observations demonstrate that incorporating the unbalanced OT formulation provides a more robust and effective approach for handling domain adaptation under label distribution mismatch.

Table 8. Comparison of vanilla and unbalanced OT formulations for GDA task on Portraits dataset.

Method	$p(y=1)=0.0$		$p(y=1)=0.1$		$p(y=1)=0.2$		$p(y=1)=0.3$		$p(y=1)=0.4$	
	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ
Initial	35.4	-	41.0	-	47.1	-	53.2	-	59.6	-
E-SOT	$55.4_{\pm 3.15E-1}$	$\uparrow 56.41\%$	$61.1_{\pm 2.01E-1}$	$\uparrow 48.94\%$	$67.1_{\pm 2.45E-1}$	$\uparrow 42.55\%$	$56.7_{\pm 2.16E-1}$	$\uparrow 6.54\%$	$57.7_{\pm 1.66E-1}$	$\downarrow 3.22\%$
E-SUOT	$64.5_{\pm 9.09E-2}$	$\uparrow 82.32\%$	$78.2_{\pm 6.55E-2}$	$\uparrow 90.50\%$	$74.5_{\pm 1.02E-1}$	$\uparrow 58.28\%$	$79.8_{\pm 8.24E-2}$	$\uparrow 49.92\%$	$77.7_{\pm 3.39E-2}$	$\uparrow 30.42\%$
Method	$p(y=1)=0.6$		$p(y=1)=0.7$		$p(y=1)=0.8$		$p(y=1)=0.9$		$p(y=1)=1.0$	
	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ	Accuracy (%)	Δ
Initial	73.8	-	80.0	-	85.9	-	91.4	-	97.1	-
E-SOT	$75.2_{\pm 3.08E-2}$	$\uparrow 1.95\%$	$74.1_{\pm 4.80E-2}$	$\downarrow 7.33\%$	$80.0_{\pm 2.21E-2}$	$\downarrow 6.89\%$	$89.9_{\pm 1.92E-2}$	$\downarrow 1.65\%$	$96.0_{\pm 2.93E-2}$	$\downarrow 1.08\%$
E-SUOT	$79.1_{\pm 2.56E-2}$	$\uparrow 7.24\%$	$84.0_{\pm 3.46E-2}$	$\uparrow 5.05\%$	$87.8_{\pm 1.12E-2}$	$\uparrow 2.18\%$	$91.8_{\pm 1.71E-3}$	$\uparrow 0.45\%$	$97.6_{\pm 3.23E-3}$	$\uparrow 0.54\%$

Kindly Note: Δ denotes performance change percentage compared to E-SUOT with entropy regularization and KL divergence. The accuracy is reported as mean (%) $\pm$ 1 $\times$ standard deviation error. For source domain, $p(y=1)=0.63$. The acronym E-SOT stands for entropy-regularized semi-dual optimal transport.

F.4. 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 Figure 4.

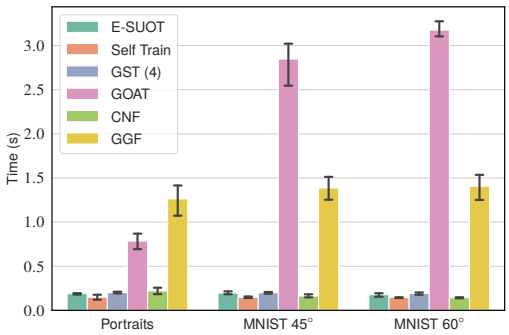


Figure 4. Computational time (s).

As shown in Figure 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.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.

G. Discussions on 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 adaptation process. As a result, the performance of the proposed E-SUOT framework may degrade under scenarios involving significant covariate shift (Sugiyama & Kawanabe, 2012; Sugiyama et al., 2007). An important direction for future research is to integrate label information into the transportation process, for example, classifier guidance approach (Bonet et al., 2025; Courty et al., 2017a; Dhariwal & Nichol, 2021; 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 (Wang et al., 2023; Zhang et al., 2022; 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.
- **Assumption of Label Invariance along the Transport Path:** The current formulation assumes that labels remain invariant during adaptation, i.e., $y_{t+1} \leftarrow y_t$, and thus primarily focuses on aligning the marginal feature distributions $p(x_t)$. This assumption may limit performance under pronounced label shift scenarios, where the conditional relationship $p(y|x)$ varies across domains, a case often encountered in unbalanced or fine-grained settings. Although our framework can be extended by incorporating classifier uncertainty or pseudo-label refinement (drawing inspiration from self-training schemes such as Eq. (3)–(4) in reference (Kumar et al., 2020)), handling substantial concept drift remains an open challenge. Future work may consider integrating adaptive label transport (Courty et al., 2017a) or uncertainty-aware pseudo-labeling (Kumar et al., 2020; Zhuang et al., 2024) to explicitly account for label-shift dynamics along the adaptation trajectory.
- **Higher Efficiency Utilization of Neural Networks:** In our current design, each stage requires training three separate networks, namely w_ϕ, $\boldsymbol{T}_\theta$, and h_ω, to generate each intermediate domain. Although this strategy can save computation time compared to existing approaches that perform intermediate-domain generation online during the domain adaptation stage, it may still be suboptimal in terms of overall training efficiency in the offline stage. A promising future direction is to reformulate the training of $\boldsymbol{T}_\theta$ into a more parameter-efficient form, such as adopting a LoRA-style adaptation (Hu et al., 2022) or using reparameterization trick to parameterize the difference between different stage (Choi et al., 2024). For w_ϕ and h_ω, one possible improvement is to fine-tune only the last layer (Brunzema et al., 2025; Harrison et al., 2024), which could further reduce the offline training cost.

Confidential reviewer copy. This manuscript is under double-blind review by ICML 2026. Unauthorized sharing, redistribution, or disclosure is strictly prohibited.